

An Approach to Improving Marketing Campaign Effectiveness and Customer Experience Using Geospatial Analytics

By

Michael Philippus Brink

*Thesis presented in fulfilment of the requirements for the degree of
Master of Engineering in Industrial Engineering in the Faculty of
Engineering at Stellenbosch University*



Supervisors:
Dr A. van Rensburg, Mr J. van Eeden

March 2017

Stellenbosch University

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights, and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2017

Copyright © 2017 Stellenbosch University
All rights reserved.

Abstract

This thesis discusses a case study in which a South African furniture and household goods retailer wishes to improve its marketing campaigns by employing location-based marketing insights, and also to prioritise customer satisfaction. This thesis presents two methods of achieving these improvements to the retailer's business. The first method uses customer delivery addresses and population data (for a sample area) to identify the location-based profiles of customers. The locations are restricted to regions within Gauteng, and key variables such as age, race, income, and family size are used to create the customer profiles. The second method builds on the intelligence produced by the customer profiles by presenting an option for improving location-based marketing campaigns. This is achieved by identifying customer clusters based on the home addresses to which purchased goods were delivered. A grid-based clustering method is applied using the sample area contained in Gauteng. This thesis shows how spatial data can be used to solve the business problems presented by the furniture retailer. The findings show how the dwelling types of customers can be used to explain why some areas are more clustered than others. This study summarises how customer profiles and location-based density clusters can be used to improve the retailer's strategic marketing strategies, and also improve the customer experience by enhancing customer-product association logic. Several recommendations are made to improve on the results produced in this study.

Uitreksel

Hierdie tesis bespreek 'n gevallestudie waarin 'n Suid Afrikaanse meubel-en-huishoudelikegoedere handelaar beoog om hul bemarkingsveldtogte te verbeter deur plek-gebaseerde bemarkingsinsigte en verder, om kliënt-tevredenheid te prioriteer. Die tesis stel twee metodes voor wat mik om die verbeteringe te bereik. Die eerste metode maak gebruik van 'n steekproef kliënte se huisaddresse waarheen afleverings plaasgevind het vir meubels en ander huisgoedere wat gekoop is. Bevolkingsdata is gebruik om die administratiewe areas te identifiseer waarin die verskeie kliënte se addresse geleë is. Profiele is geskep vir al die geografiese segmente van die kliënte. Die word bepaal deur die grense van al die munisipale distrikte binne Gauteng. Veranderlikes soos ouderdom, inkomste, gesinsgrootte, en ras word gebruik om die segmente te klassifiseer. Die tweede metode bou op die intelligensie wat in die eerste metode geskep is deur kliëntebondels te identifiseer. 'n Roosterbondelings metode is toegepas op die steekproefruimte wat omskryf is deur die area van Gauteng. Hierdie tesis wys hoe die gebruik van ruimtelike data gebruik kan word om die besigheidsprobleme, wat voorkom in die gevallestudie, op te los. Die resultate wys verder hoe die woning tipes van sekere bondels gebruik kan word om te verstaan waarom sekere bondels digter voorkom as ander. Die studie som op hoe kliënteprofile en plek-gebaseerde kliëntebondels waarde kan toevoeg deur die kleinhandelaar se bemarkingsstrategi te verbeter asook die kliënte-tevredenheid. Verskeie aanbevelings word voorgestel om die resultate in die tesis te verbeter en die studie te vergroot.

Stellenbosch University

Acknowledgements

I would like to thank my wife, Jeanne Brink, for her endless support and encouragement, and my supervisor, Dr Antonie van Rensburg, for his thought leadership in the fields of industrial engineering and data science.

Contents

Declaration	i
Abstract	ii
Uitreksel	iii
Acknowledgements	iv
List of Figures	vii
List of Tables.....	viii
List of Abbreviations.....	ix
Introduction	1
1.1 Introduction	2
1.2 Retail Business Environment: Case Study	4
1.3 Problem Statement and Objectives	7
1.4 Thesis Layout	8
Literature Review	10
2.1 Statistical Methods for Understanding Data Relationships.....	11
2.1.1 Distribution Variance	11
2.1.2 Comparison of Multiple Variance.....	12
2.1.3 Linear Regression and Correlation.....	13
2.2 Statistical Inference	19
2.3 Data Aggregation	21
2.4 Cluster Techniques	22
2.5 Mapping and Data Visualization.....	24
Data Management	26
3.1 Data Handling Concepts.....	27
3.2 Data Sources.....	28
3.2.1 Internal Data.....	28
3.2.2 External Data.....	29
3.3 Data Samples.....	30
Methodology	32

Stellenbosch University

4.1 Methodological Framework	33
4.2 Customer Profiling	35
4.2.1 Determining Administrative Boundaries	35
4.2.2 Profiling Variables.....	37
4.2.3 Determining Customer Characteristics.....	38
4.3 Area Segmentation	39
4.3.1 Grid Dimensioning	39
4.3.2 Bin Characteristics and Variance Significance.....	43
4.4 Cluster Inference	44
4.4.1 Data Transformation.....	45
4.4.2 Multivariate Analysis	46
4.4.3 Correlation Analysis.....	50
4.4.4 Test Sample Analysis	51
Results	52
5.1 Customer Profiles	53
5.2 Customer Cluster Inference	54
5.2.1 Linear Regression and Correlation Models	56
Closure.....	61
6.1 Discussion.....	62
6.2 Limitations.....	63
6.3 Conclusion.....	63
6.4 Recommendations	64
List of References.....	65
Appendix 1	72
Appendix 2	85
Appendix 3	93
Appendix 4	97

List of Figures

Figure 1: Analytics Maturity framework	2
Figure 2: The Retailer's supply chain and distribution network	5
Figure 3: Scree plot graph showing eigenvalues.....	18
Figure 4: Customer addresses overlaid on a static map	25
Figure 5: ETL process diagram.....	27
Figure 6: Data transformation approach.....	28
Figure 7: Plot of Gauteng, RSA (R Core Team, 2016) using shapefile data	30
Figure 8: Methodological framework	33
Figure 9: Provinces of South Africa.....	36
Figure 10: Municipal districts of Gauteng	36
Figure 11: Electoral wards of Gauteng	36
Figure 12: Multi-dimensional histogram showing bins of the sample area	40
Figure 13: Raster plot of customers in the sample data	41
Figure 14: Variance of bins in the train sample data.....	44
Figure 15: An example of data transformation logic	45
Figure 16: Scree plot of the PCA object containing <i>Dwellings</i> variables	49
Figure 17: Relationship between average income and family size	53
Figure 18: Box plot of bin variances for the municipal districts of Gauteng	54
Figure 19: Number of bins (≥ 40 customers) in each municipal region of Gauteng	55
Figure 20: Train sample linear regression results	56
Figure 21: Variance of bins in the test sample data	57
Figure 22: Test sample linear regression results	58
Figure 23: Residuals plot of test sample	58
Figure 24: Test sample improved linear regression model results	59

List of Tables

Table 1: Example of POD information.....	29
Table 2: Parameters of internal data used.....	30
Table 3: Sample constraints.....	30
Table 4: Detailed summary of methodological framework	34
Table 5: Example of shapefile fields	37
Table 6: Reduction in bins and customers of the train sample from applying the threshold.....	43
Table 7: Census data considered in the PCA study	47
Table 8: Eigenvalues of the PCA object.....	48
Table 9: Correlation values of standardised variables for each principal component	50
Table 10: Customer profiles for all municipal districts	53
Table 11: Top 10 lowest variance bins	55
Table 12: Correlation results of top 10 lowest variance bins in the train sample.....	56
Table 13: Correlation results of top 10 lowest variance bins in the test sample.....	59

List of Abbreviations

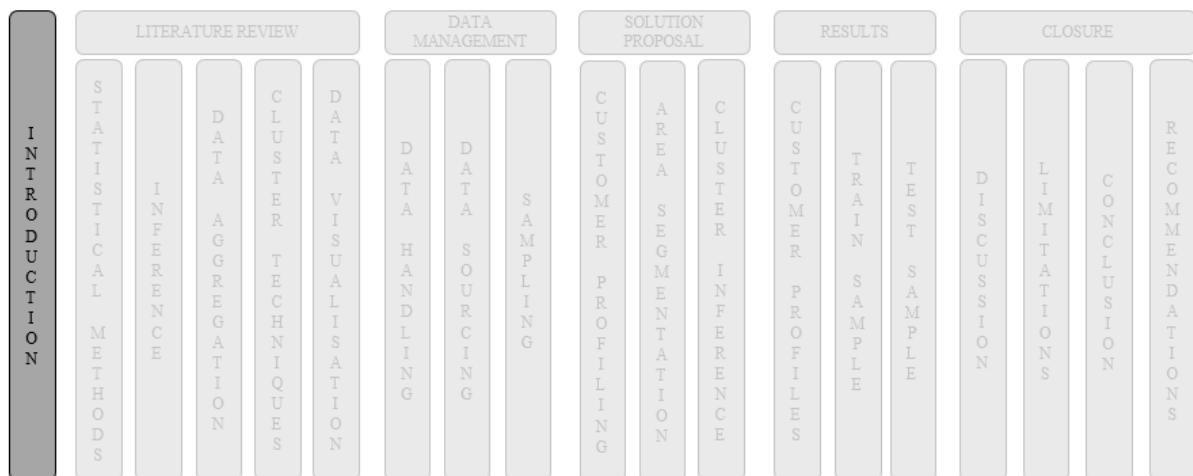
3D	Three-dimensional
API	Application Programming Interface
ASH	Average Shifted Histogram
DC	Distribution Centre
ERP	Enterprise Resource Planning
ESRI	Environmental Systems Research Institute, Inc
GIS	Geographical Information System
GPS	Geographic Positioning System
JSON	Java Script Object Notation
PCA	Principal Component Analysis
POD	Proof of delivery
VAS	Value Added Services
XD	Cross Dock

Chapter 1

Introduction

Chapter Aim:

The aim of this chapter is to introduce the business problem that is presented in this thesis, as well as the research objectives of this study. The introduction discusses the drivers of a retail sales environment and how these drivers affect the value chain. A view is presented on the relationship between spatial data and the business value drivers that can be influenced by interpreting spatial data (e.g., patterns of customer dispersion and population data that intersect these patterns). Finally, a case study is discussed that puts the problem statement in context.



Chapter Outcomes:

- Delineation of the research domain and research problem
- Presentation of the case study
- Presentation of the problem statement and research objectives
- Development of the thesis structure

1.1 Introduction

The high level of competition and the speed of the South African business environment makes for a challenging task not only of retaining customers but also of gaining market share. Applying the knowledge gained by customer insights can be the differentiating factor in gaining market share over competitors. In an environment where data are readily available, and in large quantities, the use of business intelligence, defined by customer insights, becomes an important asset. According to Daniel (2007), a global short-coming in business intelligence is that there is so much data but too little insight. Daniel (2007) supports this statement by quoting Bill Hostmann, a Gartner research analyst: "Everything we use and buy is becoming a source of information and companies must be able to decipher how to harness that". One reason to harness business intelligence is to understand how this intelligence influences customer satisfaction. In their White Paper, Frost & Sullivan (2015) state that in regular interactions with customers via multiple communication channels, there is ample opportunity to reduce costs and enhance customer satisfaction significantly. In addition, Frost & Sullivan (2015) state that a top industry trend is the prioritisation of customer satisfaction, retention, and loyalty. While these trends remain important, consideration should also be given to reducing the cost of customer acquisition and the cost of serving these customers.

In order to influence customer satisfaction, retention, and loyalty, knowledge about the customer needs to be extracted from vast amounts of data. New York University (2013) says that data science involves using automated methods to analyse massive amounts of data and to extract knowledge from them. Knowledge or 'intelligence' may be produced at various levels of maturity and consequently define different attributes of the customer. Figure 1 shows the four levels of maturity in analytics as defined by Chandler *et al.* (2011).

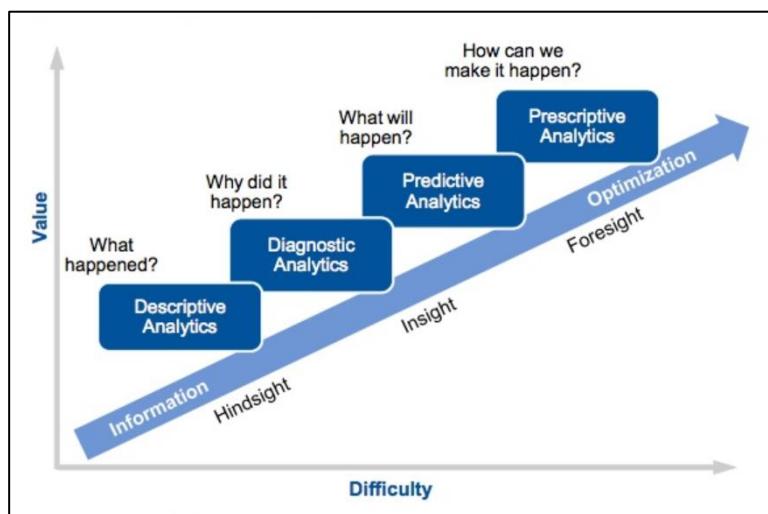


Figure 1: Analytics Maturity Framework

*Adapted from Chandler *et al.* (2011).*

Stellenbosch University

Figure 1 shows the relationship between the value achieved and the complexity of performing analytical techniques on some data. Although prescriptive analytics would always be desirable from a value point of view, the complexity of analysis is dependent on the maturity of the information that is used in performing the analysis. This point is reinforced by Cai and Zhu (2015), who state that “high quality data are the precondition for analysing and using meta data and for guaranteeing the value of the data”. In this thesis, both the *quality* and *quantity* of data that are investigated are determinants in scoping the objectives of the study with regard to the maturity framework shown in Figure 1. The focus of this study is to produce hindsight and insight about customer behaviour. This requires the use of both descriptive and diagnostic analytics.

According to Vega *et al.* (2015), the consumer market is a spatial reality that is defined by two influencing factors: the geographic component of the market, and the distribution system. In addition to introducing these two factors, Vega *et al.* (2015) state that the main drivers of market integration are the supply and demand components, as well as all the elements of the geographical surroundings that affect them. Customers determine demand, and businesses compete to supply this demand. The *supplier* environment is expanded upon in the section that follows. With regard to demand, however, the potential of the local market and the tendency of customers to purchase goods depends largely on the demographic characteristics of the market area (Grewal *et al.* 1999; Mulhern and Williams 1994; Johnson 1989). According to Johnson (1989), the geo-demographic characteristics constitute the classification of the people according to the type of neighbourhood they reside in, as opposed to the conventional socioeconomic criteria such as income or social class (Bearden *et al.* 1978; Bawa and Shoemaker 1987; Kalyanam and Putler 1997; Ailawadi *et al.* 2001). These geo-demographic characteristics are defined by Sleight (1995) as demographic information that can be obtained from various sources such as population census surveys.

While descriptive and diagnostic analytics may interpret data in many forms, mapping is a key component of data visualisation, used in this study to show interrelationships between customer locations and population data. In *Statistical Analysis & Dissemination of Census Data*, Palma (2007) suggests six reasons for using maps to display population data: to communicate a concept or idea; to support textual information; to aggregate large amounts of data; to illustrate comparisons in densities, trends, patterns, etc.; to describe, explore, and tabulate; and finally, to appeal to the viewer’s curiosity.

Chapter 2 discusses further the spatial analysis techniques that can be applied to population data using the geospatial attributes of maps.

Mention has been made of ‘customers’, ‘sales’, and ‘supplier’ in the context of a consumer market. Customer information is unique to the supplier from whom customers have purchased goods or services.

Stellenbosch University

A case study is presented in the next section, which defines the market conditions of a South African furniture and household goods retailer. Customer information about this retailer has been provided¹ and will be used to perform descriptive and diagnostic analyses to produce hindsight and insight about their customers, and thereby solve the business problems the retailer faces.

The next two sub-sections introduce the case study and the problem statement.

1.2 Retail Business Environment: Case Study

A large South African furniture and household goods retailer, hereafter referred to as ‘the retailer’, boasts an extensive national supply chain network that services a large customer base (*see Figure 2*). Although the retailer primarily trades in furniture, its product range includes general household appliances and electronics. The products range from low-price, low-quality items to high-value, good-quality items. However, the majority of the items that are sold appeal to a market of customers who are attracted to low-price and thus to low-quality items. This would suggest that the retailer’s target market is primarily, but is not restricted to, lower income earning individuals in South Africa. The retailer’s supply chain network that supports their demand is characterised by five provincially-based distribution centres (DCs), twenty-seven cross docks (XDs) and over five hundred stores located in all nine South African provinces. The distribution network that services this supply chain includes three transport channels: supplier, primary, and secondary transport.

The retailer purchases all of its goods from foreign manufacturers. The goods are shipped directly to the various DCs by the supplier. The retailer therefore only takes over the storage and distribution of these goods to cross docks, to stores, or directly to customers who have made a purchase. Additionally, goods may be transferred between two or more of the retailer’s facilities on an *ad hoc* basis, in order to balance demand. Figure 2 illustrates the integrated functions of the supply chain and distribution network.

¹ This customer information has been provided by a South African logistics company whose identity will not be disclosed for confidentiality reasons.

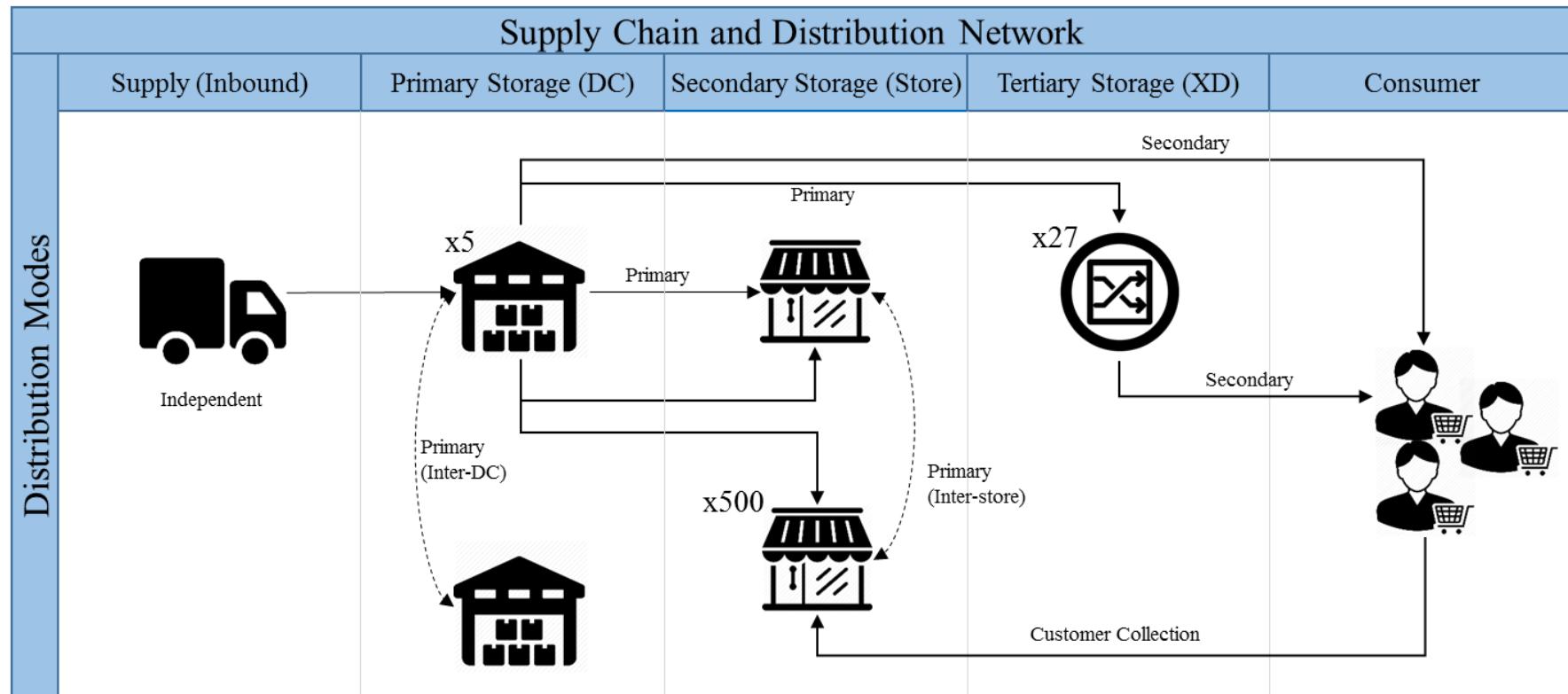


Figure 2: The Retailer's supply chain and distribution network

Stellenbosch University

The distribution network model in Figure 2 shows the various channels-to-market from a distribution perspective. Given the market that the retailer attracts, a physical footprint is an essential part of the retailer's business model, as these types of customers are more attracted to a 'bricks-and-mortar' shopping experience than to online shopping. This business model is evident in the high number of stores in the retailer's supply chain network. The retailer therefore aims to attract customers through exhibition² and in-store products that are stocked in the branches, as opposed to virtual catalogues and online sales. However, stores only aim to *display* products, not to stock them. Although branches are permitted to sell their display stock, purchases are generally made in branches and the stock is delivered to the customer's home from the DC. Furniture and large home appliances are bulky and so, given the limited space in a store, are seldom sold out-of-store. Other goods such as electronics and smaller appliances or textiles may be more readily sold in branches. The retailer also offers value added services (VAS) to its customers. These include furniture protection (applied to furniture with a vaporised chemical), in-house assembly of certain products, and exchanges of damaged goods.

Figure 2 only shows the forward flow of distribution and storage. Reverse logistics is also an essential part of the retailer's business, as it caters for the collection and replacement of damaged goods or goods that need to be returned to shelf for cancelled sales. The retailer's reverse logistics supply chain is not elaborated on, as it is not within the scope of this study. When goods are purchased, the branches promise customers a lead time to delivery. This lead time is dependent on stock availability and the customer's location (address to which the delivery is made). Fulfilling this promise is an important driver of customer satisfaction. The retailer ensures that delivery lead times are met by employing customer liaison agents who confirm orders, manage payments, and ensure that the delivery information is correct. Customer delivery addresses are captured as free-text fields in the branch at the point of sale, and this information is stored in the retailer's ERP system. Addresses are then geocoded³ in order for the transporter to locate the address using a GPS device. The successful conversion of a text address to a geocode is dependent on the quality and completeness of the text address. The manual process of address capturing therefore poses a risk to ensuring quality geocoded data. Missing fields (address lines) may result in conversion errors, which then need to be manually determined by the transporter. When a delivery of purchased goods is made, a proof-of-delivery (POD) is captured electronically by the driver. The POD contains the status, date, time and address for every customer delivery that is made. This information is fed back into the ERP system and archived. A large data set of POD information is provided (for a two-year period) for all deliveries made to customers for goods purchased from the retailer. This information is used in this thesis as *customer data*.

² Products that are stocked only for display purposes.

³ The process of converting a text address to a latitude and longitude coordinate.

1.3 Problem Statement and Objectives

The retailer introduces two problems that are faced in their business. The objective of this study is to address these problems and provide a solution based on mining customer data. The first problem arises from the retailer's lack of understanding of *who* their customer is. Given their limited customer data that are acquired at the point of sale (as described in Section 1.2), the retailer would like to gain insight into their customers and in so doing, to create a better customer experience by using these insights to enhance customer-product association. Customer insights can be defined by identifying and interpreting key characteristics of the customer, such as age, income, ethnicity, etc. The characteristics of a customer can be used to market products more appropriately – that is, a low income earning individual would most likely not be attracted to high-end, expensive products, but rather to a range of products that are more affordable. The second problem is the inability of the retailer's marketing team to develop specific location-based marketing campaigns. This problem arises as a result of limited data and thus a limited understanding of the relationship between *who* the customers are (i.e., the customer characteristics) and *where* customers are located. The value of this relationship for location-based marketing campaigns is the knowledge of dense and sparse customer clusters and the needs of the customers in these clusters. This enables the retailer to target key locations and deploy appropriate marketing campaigns for the customers in those locations.

These two problems scope the objectives of this thesis: first, to profile certain customer segments; and second, to inform a location-based marketing strategy by identifying customer clusters and insights into these clusters. Given that this study focuses on revealing information about underlying patterns in data, there is an expectation that additional insights might be produced while exploring the data. These will be recorded as auxiliary insights, and a qualitative interpretation of the results will be discussed.

1.4 Thesis Layout

This document is logically organised to enable the reader to comprehend the flow of the research most easily. Given that this study relies heavily on handling digital information, the acquisition, transformation, and visualisation of data is clearly articulated throughout the chapters that follow.

Chapter 1: Introduction

Chapter 1, this introductory section, describes the research problem statement and objectives of the study. It also introduces the research domain (data science) and provides a case study that serves as a test sample for meeting the objectives by applying a research methodology. Finally, the chapter sets out the logical structure of the thesis' content.

Chapter 2: Literature Review

Chapter 2 introduces several fundamental concepts, principles and methods that are required to produce insights from raw data. The chapter discusses statistical methods and probability indicators, as well as methods of data mining and data transformation. This chapter paves the way for the application of statistical methods in complex data structures that convert raw data into intelligence.

Chapter 3: Data Management

Chapter 3 presents several fundamental concepts of data management. These concepts not only support the logic of the methodology described in Chapter 4, but also illustrate the approach to validating the results presented in this thesis. The chapter presents the data samples that are used in this study, together with the assumptions and limitations of the sample data. Finally, the software tool used to perform the data modelling is introduced.

Chapter 4: Methodology

Chapter 4 presents the methodologies employed to produce hindsight and insight from the sample data, and thereby to meet the objectives of this study by providing solutions to the problem statements. Although numerous statistical and data mining techniques are applied, this chapter groups these into three focused, logical methods that are aligned to achieving the objectives of this study. This chapter ensures that iterative output results are validated and that the integrity of the data is maintained.

Chapter 5: Results

Chapter 5 presents the results of the study in both a graphical and a tabular view. The significance of these results and the key statistical indicators are discussed in order to provide the context for how the results answer the problem statements.

Chapter 6: Closure

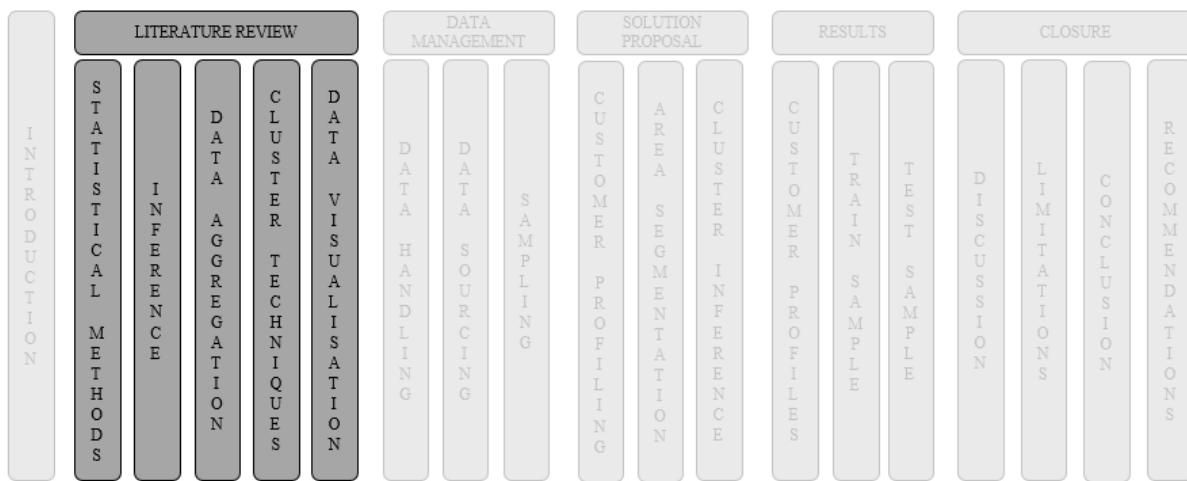
Chapter 6 discusses the results that are produced from modelling the data in such a way that the business value achieved (for the retailer) from the data insights is understood. The limitations of the study are recorded, and several recommendations are made for further study. The findings are summarised in a conclusion.

Chapter 2

Literature Review

Chapter Aim:

This chapter aims to present the literature that supports the use of statistical methods in performing spatial analytics, and paves the way for the applied methodology that is performed using a sample data set from the retailer.



Chapter Outcomes:

- Understanding of statistical methods applied to data.
- Understanding of statistical inference techniques.
- Understanding of key data aggregation principles.
- Understanding of cluster techniques.
- Understanding of mapping capabilities for data visualisation.

2.1 Statistical Methods for Understanding Data Relationships

Customer and spatial data are used to deduce customer locations and characteristics. While geocoded addresses provide point locations of where customers reside, little is known about the distribution of many customers across a geographic area or how the distribution patterns – i.e., customer clusters – might influence location-based marketing campaigns. The distribution of customers across a sample area will exhibit some patterns defined by the dense and sparse variations of customer clusters.

The scope of the analysis that will be applied in this thesis is restricted to descriptive and diagnostic analytics. Descriptive statistics is described by Lane *et al.* (2015) as numbers that are used to summarise and describe data. Descriptive statistics are *only* descriptive, and do not make inferences beyond the data at hand. Inferential statistics, however, is used to generalise from the data to hand (Lane *et al.*, 2015), and will therefore help to support the diagnostics analytics – the discovery of data insights. The research objectives of this study will require this maturity of analysis. The sections that follow introduce techniques used in descriptive and inferential statistics. These include variance, linear regression, correlation, clustering methods, data aggregation, and statistical inference.

2.1.1 Distribution Variance

Variance is defined by Montgomery and Runger (2007) as a measure of variability defined as the expected value of the square of the random variable around its mean. The mean refers either to the expected value of a random variable, or to the arithmetic average of a set of data (Montgomery and Runger, 2007). Variance is an important measure when analysing spatial data, as it indicates how clustered spatial objects – *e.g.*, *customer addresses* – are in relation to one another for some mean spatial coordinate. The mean of the data contained in a sample area must therefore be computed in order to know the measure of variance of the data. The equations shown below are adapted from Montgomery and Runger (2007), and show the formula for calculating these measures of descriptive statistics.

The mean of some discrete random variable X, denoted as μ or $E(X)$, is

$$\mu = E(X) = \sum_x xf(x) \quad (1)$$

The variance of X, denoted as σ^2 or $V(X)$, is

$$\sigma^2 = V(X) = E(X - \mu)^2 = \sum_x (X - \mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2 \quad (2)$$

The standard deviation of X is

$$\sigma = \sqrt{\sigma^2} \quad (3)$$

Stellenbosch University

The variance of some random variable X uses weight $f(x)$ as the multiplier of each possible squared deviation $(X - \mu)^2$ (Montgomery and Runger, 2007). The deviations defined in this thesis are the physical distances from customers residing in a sample area to the mean customer.⁴ Calculating $X - \mu$ is therefore not a simple arithmetic computation: it requires the computation of distance across a geographic coordinate system. Using the theorem of Pythagoras, Apparicio *et al.* (2008) define the formula for calculating distance d as

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (4)$$

where:

x_i and y_i are X and Y coordinates of point i with a plane projection.

2.1.2 Comparison of Multiple Variance

Variance may be calculated for multiple independent experiments, where each experiment might have unique features such as sample size or mean (Montgomery and Runger, 2007). When the variance of two or more independent experiments is compared, these unique features need to be considered before drawing conclusions about the comparison – e.g., *which experiment has the most or the least variance*. According to Manoukian *et al.* (1986), much research has been done on Bartlett's (1937) test of the homogeneity of variances. However, Bartlett's (1937) test is sensitive to departures from normality, and thus presupposes a normally-distributed sample for effective results (Manoukian *et al.*, 1986). In the case of non-normal data, or when the distribution profile is unknown, Allingham and Rayner (2012) suggest using the nonparametric Levene test, which is known to be more robust than Bartlett's test, but is less powerful when the data are approximately normal. Allingham and Rayner (2012) state that, when normality is in doubt, it is common practice to use Levene's test. Given that the normality of customer samples used in this study will be unknown, Levene's test is an appropriate method of testing for variance homogeneity. Levene's testing procedure, summarised by Scott-Street (2001), is shown below.

Assumptions

1. The samples from the population under consideration are independent.
2. The populations under consideration are approximately normally distributed.

Hypotheses

$$\text{Null} \qquad H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_t^2$$

⁴ The mean customer is defined by the coordinate points of a *virtual* customer location, computed as the arithmetic average of all customer coordinate points in that sample.

Stellenbosch University

Alternative H_1 : Not all variances are equal.**Critical value and rejection criteria**

	Test Statistic Evaluation	<i>p</i> -Value Evaluation
Critical Value	$F_{\alpha, (df_1=t-1, df_2=N-t)}$	N/A
Rejection Region	$F_{Levene} \geq F_{\alpha, (df_1=t-1, df_2=N-t)}$	$p < \alpha$

Levene's Statistic, F_{Levene}

$$F_{Levene} = \frac{\frac{\sum_{i=1}^t n_i (\bar{D}_i - \bar{D})^2}{(t-1)}}{\frac{\sum_{i=1}^t \sum_{j=1}^{n_i} (D_{ij} - \bar{D}_i)^2}{(N-t)}} \quad (5)$$

where:

 t = number of populations y_{ij} = sample observation j from population i ($j = 1, 2, \dots, n_i$ and $i = 1, 2, \dots, t$) n_i = number of observations from population i (at least one n_i must be 3 or more) $N = n_1 + n_2 + \dots + n_t$ = total number of pieces of data (overall size of combined samples) \bar{y}_i = mean of sample data from population i $D_{ij} = y_{ij} - \bar{y}_i$ = absolute deviation of observation j from population i mean \bar{D}_i = average of the n_i absolute deviations from population i \bar{D} = average of all N absolute deviations $\alpha = 0.95^5$

Based on the value of the test statistic, F_{Levene} , the null hypothesis is either accepted or rejected, and thus the variance significance of multiple populations can be determined.

2.1.3 Linear Regression and Correlation

Linear regression and correlation are methods of identifying relationships between variables. Bewick *et al.* (2003) define correlation as the strength of the linear relationship between two variables, while regression expresses this relationship in the form of an equation. In the context of the research conducted in this thesis, the relationships between customer data and population data need to be understood in

⁵ For a 95% confidence interval

Stellenbosch University

order to understand how demographic factors might influence customer sales. Hamburg (1985) provides the derivation of formulas that show how correlation and linear regression are measured.

2.1.3.1 Correlation

Correlation can also be defined as the degree of association between two continuous variables (Cahusac and De Winter, 2014). Therefore, if X and Y represent two continuous variables, the measure of the amount of correlation or ‘association’ between X and Y can be calculated in terms of the relative variation of the dependent Y values around the regression line, and the corresponding variation around the mean of the Y variable. The term ‘variation’ refers to the sum of squared deviations. The variation of Y values around the regression line is given by Hamburg (1985) as

$$\sum(Y - \hat{Y})^2 \quad (6)$$

Likewise, the variation of Y values around the mean of Y is given by

$$\sum(Y - \bar{Y})^2 \quad (7)$$

The relationship between these two equations indicates the degree of association between X and Y . This relationship is defined by the sample coefficient of determination, and the equation that shows this relationship is given by

$$r^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} \quad (8)$$

Therefore, r^2 shows the percentage of variation in the dependent variable Y that has been accounted for by the relationship between Y and X expressed in the regression line. The percentage variation that remains unaccounted for is thus shown by

$$\frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2} \quad (9)$$

The degree of association is derived from the r^2 value calculated in Equation (8), by taking the square root of the coefficient of determination (Hamburg, 1985). The correlation coefficient of a population is therefore defined by r as

$$r = \sqrt{r^2} \quad (10)$$

2.1.3.2 Linear Regression

For two variables (X and Y) that have a strong degree of correlation to one another, a linear regression model may be used to fit the values of the dependent variable to the independent variable (Hamburg, 1985). The independent variable is typically called the *predictor*, and the dependent is known as the *regressor*.

The expected value of Y for each value of X is given by

$$E(Y|x) = \beta_0 + \beta_1 x \quad (11)$$

where β_0 and β_1 are unknown regression coefficients. The linear regression model can therefore be defined as

$$Y = \beta_0 + \beta_1 x + \epsilon \quad (12)$$

where ϵ denotes a random error value with a mean of zero and an unknown variance σ^2 . The estimates of β_0 and β_1 should produce a line that best fits the data characterised by n pairs of (x_n, y_n) observations. Karl Gauss (1777 – 1855), a German scientist, suggested estimating the values of β_0 and β_1 to minimise the sum of squares of vertical deviations between the observed values and the regression line (Hamburg, 1985). This criterion for estimating regression coefficients is known as ‘the method of least squares’.

The sum of squares of the deviations of the observations from the true regression line is given by

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad i = 1, 2, \dots, n \quad (13)$$

The least square estimates the intercept and the slope in the simple linear regression model given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (14)$$

where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} \quad (15)$$

and

$$\begin{aligned} \bar{y} &= (1/n) \sum_{i=1}^n y_i, \\ \bar{x} &= (1/n) \sum_{i=1}^n x_i \text{ and} \\ i &= 1, 2, \dots, n. \end{aligned}$$

The fitted (estimated) regression line is therefore represented by the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (16)$$

and each pair of observations in the sample data satisfies the relationship

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \quad i = 1, 2, \dots, n \quad (17)$$

where $e_i = y_i - \hat{y}_i$ is called the *residual*. The denominator numerator of Equation 15 can be denoted by

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (18)$$

and

$$S_{xy} = \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) \quad (19)$$

2.1.3.3 Multivariate Linear Regression

When determining the relationship between a predictor variable and multiple regressor variables, multivariate linear regression might be required. Renchen (2002) defines a multivariate analysis as one consisting of several methods that can be used when multiple measurements are made on each individual or object in one or more samples. In his book, *Methods of Multivariate Analysis*, Renchen (2002) presents an approach to analysing a single sample with several variables measured on each sampling unit. The steps are shown below.

1. Test the hypothesis that the means of the variables have specified values.
2. Test the hypothesis that the variables are uncorrelated and have a common variance.
3. Find a small set of linear combinations of the original variables that summarises most of the variation in the data (principal components).
4. Express the original variables as linear functions of a smaller set of underlying variables that account for the original variables and their intercorrelations (factor analysis).

In the fourth step of his multivariate analysis, Renchen (2002) mentions the application of a factor analysis. Rahn (2012) defines a factor analysis or *Principal Component Analysis* (PCA) as a tool for exploring variable relationships for complex concepts. Factor analysis is therefore useful when there are multiple variables for a single predictor variable that need to be investigated in order to find the most significant ones. The fundamental concept of factor analysis is that multiple variables exhibit similar patterns of responses, as they are all associated with a latent variable (Rahn, 2012). The aim of factor analysis, therefore, is to identify the latent variable in order to reduce the group of variables to a smaller

Stellenbosch University

subset of latent variables that are most influential. An important metric in factor analysis is deciding the number of factors to be used in the analysis. Eigenvalues and scree plots are useful methods of factor selection (Rahn, 2012; Garrett-Mayer, 2016; Renchen, 2002).

The following approach, suggested by Renchen (2002), presents a guideline that shows how eigenvalues are used in determining which principal components or *factors* to use in the factor analysis.

1. Retain sufficient components to account for a specified percentage of the total variance – say, 80%.
2. Retain the components whose eigenvalues are greater than the average of the eigenvalues, $\sum_{i=1}^p \lambda_i / p$. For a correlation matrix, this average is 1.
3. Use the scree graph, a plot of λ_i versus i , and look for a natural break between the ‘large’ eigenvalues and the ‘small’ eigenvalues.
4. Test the significance of the ‘larger’ components – that is, the components corresponding to the larger eigenvalues.

Eigenvalues can best be explained using an example presented by Dahyot (2006) that calculates the eigenvalues for the matrix

$$A = \begin{matrix} 1 & -3 & 3 \\ 3 & -5 & 3 \\ 6 & 6 & 4 \end{matrix}$$

To do this, the values of λ are found that satisfy the characteristic equation of the matrix A , namely those values of λ for which

$$\det(A - \lambda I) = 0 \quad (20)$$

where I is the 3×3 identity matrix. The matrix $A - \lambda I$ is defined as

$$A - \lambda I = \begin{matrix} 1 - \lambda & -3 & 3 \\ 3 & -5 - \lambda & 3 \\ 6 & 6 & 4 - \lambda \end{matrix}$$

and $\det(A - \lambda I)$ is therefore computed by Dahyot (2006) as follows:

$$\det(A - \lambda I) = (1 - \lambda) \begin{vmatrix} -5 - \lambda & 3 \\ -6 & 4 - \lambda \end{vmatrix} - (-3) \begin{vmatrix} 3 & 3 \\ 6 & 4 - \lambda \end{vmatrix} + 3 \begin{vmatrix} 3 & -5 - \lambda \\ 6 & -6 \end{vmatrix}$$

Solving Equation (20) produces a set of integer value roots called the *eigenvalues* of the matrix (Dahyot, 2006). Renchen’s (2002) principal component guideline suggests the use of a scree plot in step 3 to

Stellenbosch University

differentiate between large and small eigenvalues. A scree plot graphs the eigenvalue against the component number, and serves as a useful visual aid in determining an appropriate number of principal components (OriginLab, 2016). Renchen (2002) suggests that the eigenvalues exhibiting a steep slope should be maintained, while the ‘tail’ of the slope should be tested for significance using a test statistic. An example of a scree graph is shown in Figure 3. In this example, only the first two components would be retained, as values 3 to 6 visually identify the tail of the graph.

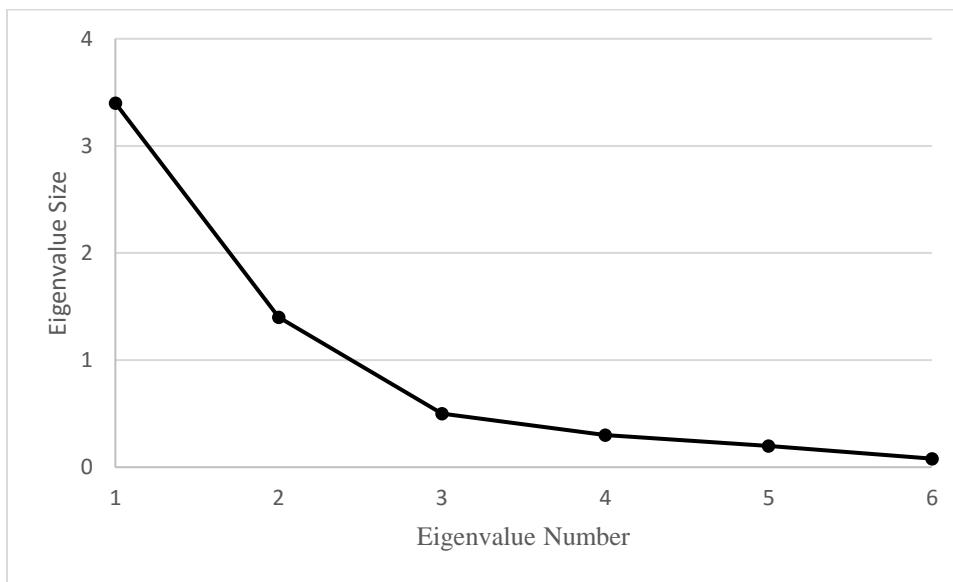


Figure 3: Scree plot graph showing eigenvalues

Adapted from Renchen (2002)

Figure 3 can therefore be used to visually identify the break between large and small eigenvalues. Visual inspection is useful for a quick evaluation of the obvious principal components that should be selected for evaluation. There are, however, alternative methods for selecting principal components. Jackson (1993) and Peres-Neto *et al.* (2005) present an approach with several selection criteria for picking an appropriate number of principal components. Their summary of several different published approaches is shown as follows:

1. Look for a ‘knee in the curve’ using a scree plot (similar to Renchen’s (2002) approach)
2. For data that have been transformed to unit variance: retain the components corresponding to single values greater than 1.
3. Select the number of components that are sufficient to cover some fixed fraction (generally 95%) of the observed variance.
4. Perform a statistical test to see which single values are larger than we would expect from an appropriate null hypothesis or noise process.

Once a PCA has been conducted using one of the methods presented by Jackson (1993) and Peres-Neto *et al.* (2005), the task remains of identifying the variables within the selected principal components that have the greatest influence on the principal components. The measure of influence of a variable on its principal component is characterised by the coefficients of correlation between variables and components; this measure is known as *loading* (Abdi and Williams, 2010; Suhr, 2012). Abdi and Williams (2010) state that the sum of the squared coefficients of correlation between a variable and all the components is equal to 1. Therefore, the *squared loadings* are easier to interpret than the loadings themselves (due to the fact that squared loadings give the *proportion of the variance* of the variables explained by the components). Although there are no formal criteria for selecting a cut-off loading value, Clark (2009) deems loading values of > 0.5 to be significant, while other sources suggest that values > 0.4 are significant (*Factor Analysis: A Short Introduction*, Part 4, 2014). However, this might not always be an appropriate method; in the event that there are several significant variables with loading values < 0.5 , none would be selected. Abdi and Williams (2010) present an alternative selection method whereby variables whose loading score is above the average loading score of all the variables contained in a given principal component are selected. Each PCA study should therefore be evaluated independently in order to apply the correct method of variable selection.

2.2 Statistical Inference

Section 2.1 shows how the degree of association between variables can be measured. Statistical hypothesis is a basic yet important component of mathematical statistics (Shi and Tao, 2008). Montgomery and Runger (2007) state that many problems require a decision about whether a statement about some parameter should be accepted or rejected. This statement is called ‘the hypothesis’, and the decision-making procedure about the hypothesis is called ‘hypothesis testing’. Essentially, the hypothesis aims to minimise a Type II error by controlling the Type I error (Shi and Tao, 2008). A Type I error occurs when the null hypothesis should be accepted, but is rejected; a Type II error is exactly the opposite (Montgomery and Runger, 2007). Depending on the test that was performed, various test statistics can be computed to perform hypothesis testing.

Hamburg (1985) suggests the following test statistics for calculating the significance of the correlation coefficient, r .

Hypotheses

Null	$H_0: r = 0$
Alternative	$H_1: r \neq 0$

Stellenbosch University

Critical value and rejection criteria

	Test Statistic Evaluation	<i>p</i> -Value Evaluation
Critical Value	$t_{\alpha/2,n-2}$	N/A
Rejection Region	$t_0 > t_{\alpha/2,n-2}$ or $t_0 < t_{1-\alpha/2,n-2}$	$p < \alpha$

Test Statistic, t_0

$$t_0 = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} \quad (21)$$

where:

 r = correlation coefficient n = sample size $\alpha = 0.95^6$

When deciding whether to accept or reject the parameters of a linear regression model, Hamburg (1985) introduces a hypothesis test to determine whether or not the slope, β_1 (see Equation (16)) is equal to zero. Accepting the null hypothesis (H_0) would imply that there is no linear relationship between the variables X and Y . Rejecting H_0 (and thus accepting H_1) indicates that the straight-line model is adequate, or that – in addition to the linear effect of X – better results could be obtained with the addition of higher order polynomial terms in X (Hamburg, 1985). The hypothesis test is shown by Hamburg (1985) below.

Hypotheses

$$\text{Null} \quad H_0: \beta_1 = 0$$

$$\text{Alternative} \quad H_1: \beta_1 \neq 0$$

Critical value and rejection criteria

	Test Statistic Evaluation	<i>p</i> -Value Evaluation
Critical Value	$t_{\alpha/2,n-2}$	N/A
Rejection Region	$ T_0 > t_{\alpha/2,n-2}$	$p < \alpha$

Test Statistic, t_0

$$T_0 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \quad (22)$$

⁶ For a 95% confidence interval.

where α denotes the upper tail of the confidence interval and $se(\hat{\beta}_1)$ is the computed standard error of the slope given by

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad (23)$$

The parameters of Equation (23) are described as follows:

The unbiased estimator, $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{SS_E}{n - 2} \quad (24)$$

The error sum of squares, SS_E :

$$SS_E = \sum_{i=1}^n e_i^2 \quad (25)$$

2.3 Data Aggregation

GIS data are represented in two dimensions when considering a planar space, characterised by latitude and longitude. As the analysis performed in this study relies heavily on the use of GIS data, a core statistical concept of this study is the aggregation of GIS data. GIS data may often be required to be aggregated in order to draw inferences about a sample or ‘segment’ of data. Aggregation can therefore be performed in a logical manner to create sample sizes that are reasonable for obtaining results for spatial analysis. Scott (1979) introduces the histogram as a nonparametric density estimator that is an important statistical tool for displaying and summarising data. Aggregation parameters are an important component of summarising data. Data may be aggregated accordingly to upper and lower limits that define each group with which unique data observations are associated. Scott (1979) confirms the importance of choosing the correct aggregation parameters by stating: “Choosing the correct number of groupings or ‘bins’ in a histogram is important, as too few may dilute the data and too many produces a result which may be too granular”. According to He & Meeden (1997), there are (unfortunately) only limited explicit guidelines (based on statistical theory) for choosing the optimal number of bins that should appear in a histogram. This study considers three methods of bin size estimation: Sturges’ (1926) rule, Scott’s (1979) rule, and Freedman and Diaconis’s (1981) rule.

In his article, Hyndman (1995) suggests that Sturges’ rule only produces reasonable results for small to moderate sample sizes. Given the large customer sample size considered in this research, Sturges’ rule

Stellenbosch University

is disregarded as an inappropriate method for bin size estimation. Both Scott (1979) and Freedman and Diaconis (1981) present a formula for the optimal bin width that aims at asymptotically minimising the integrated mean squared error. When observing a data sample, the underlying density of the data set is often unknown, as is the case with the sample data used in this study. Scott (1979) suggests using the Gaussian density as a reference standard to overcome this. This approach leads to a data-based choice for the bin width. Scott's (1979) formula is given by

$$\text{bin width}_{\text{Scott}} = a \times s \times n^{-1/3} \quad (26)$$

where $a = 3.49$, s is an estimate of the standard deviation, and n is the sample size (Scott, 1979). Freedman and Diaconis's rule is similar to Scott's rule; however, the bin width is calculated as

$$\text{bin width}_{\text{Freedman \& Diaconis}} = 2 \times IQR \times n^{-1/3} \quad (27)$$

where IQR denotes the interquartile range of the sample (Freedman and Diaconis, 1981). Hyndman (1995) states that both these methods are well-founded in statistical theory, and are conducted by assuming that the data are close to normally-distributed.

2.4 Cluster Techniques

The literature presented above on *Data Aggregation* eluded to the grouping of data using the principle of a histogram. Although data might appear aggregated or grouped in any form of cluster, various methods of clustering can be used to identify dense or sparsely distributed data. Chauhan *et al.* (2010) introduce the importance of clustering within spatial mining in their article, stating that spatial data mining includes the discovery of interesting and valuable patterns from spatial data by grouping the objects into clusters. Murray (1998) substantiates the importance of clustering in spatial analysis by noting that the application of conventional clustering methods is being pursued as an exploratory approach for the analysis of spatial data.

In its web publication, the University of Toronto (2002) distinguishes between two basic clustering techniques: *partitional* and *hierarchical*. These techniques are defined by Han and Kamber (2001) as follows:

Partitional: For a given data set of n objects, a partitional clustering algorithm constructs k *partitions* or *clusters* where each cluster optimises some cluster criterion – e.g., the minimisation of *sum of squared distance from the mean* within each cluster.

Stellenbosch University

Hierarchical: Hierarchical algorithms generate a hierarchical decomposition of the spatial objects contained in a data set. The decomposition may be either *divisive (top-down)* or *agglomerative (bottom-up)*.

Apart from the two basic methods of clustering described above, Han and Kamber (2001) introduce several other methods of cluster analysis that are mainly focused on specific problems or problems that have specific data sets available. These include:

- Density-based clustering,
- grid-based clustering,
- model-based clustering, and
- categorical data clustering.

Given the specific data sets that are studied in this research, the appropriate clustering analysis method should be applied. The primary application of grid-based clustering is with spatial data – i.e., data that model the geometric structure of objects in space, their relationships, properties, and operations (University of Toronto, 2002). The objective of grid-based clustering is to quantise the data set into a fixed number of bins or *cells*, and then to work with objects⁷ belonging to these bins. The construction of the grid is not dependent on variable distance measures, but is determined by a fixed predefined parameter (University of Toronto, 2002). The attribute of a fixed-size parameter that is used in grid-based clustering is what differentiates this method from the other three listed above. Grid-based clustering qualifies as a logical clustering analysis method to be used on the spatial data provided in this research.

Murray (1998) defines an approach to performing grid-based clustering by using the centre points of bins. In the context of spatial data, the centre point is not defined as the geometrical centre point, but rather as an artificial point in space that identifies the most central location of all the observations contained within that bin (Murray, 1998). The application of a fixed-size grid makes the observations contained in each bin independent and thus exclusive of the rest of the population. While Murray (1998) suggests optimising some objective function that evaluates the distances of *all* objects in a sample to several artificial means, the application of a *grid* simplifies Murray's approach. Given that clusters are already identified (by the application of a fixed parameter grid), it only remains to calculate how clustered the observations are within each cluster or *bin*. The calculation of variance, defined by Equation (2), is used to calculate the degree of clustering in each bin.

⁷ In this study, ‘objects’ are represented by customer locations (addresses).

2.5 Mapping and Data Visualization

In its White Paper, Oracle (2010) conveys the importance of data visualisation by stating that the ability to display data is paramount when providing insights into business intelligence. While Oracle (2010) lists several methods of visualising data, it states that maps allow statistical measures to be displayed for an area or a region. This becomes particularly useful for visualising large amounts of data, such as those from a population census. Using electronic maps to overlay spatial data requires the integration of a mapping service provider with a software tool that can access interactive electronic maps via an application programming interface (API). For the research conducted in this study, Google – the publicly-accessible mapping interface hosted by Map Data® 2016 AfriGIS (Pty) Ltd – is used. The statistical programming software R (R Core Team, 2016) is used to integrate map data with spatial data. R contains the library *RgoogleMaps*, which serves the following two purposes or functions (Loecher and Ropkins, 2015): it provides a comfortable R interface for querying the Google™ server for static maps, and it uses the map as a background image for overlaying plots within R.

R's *RgoogleMaps* library performs these two purposes by integrating with the map data stored in the graphical information system (GIS) that is hosted by Map Data® 2016 AfriGIS (Pty) Ltd, Google (Google). The integration is achieved by using java script object notation (JSON), a readable format for structuring data that is used primarily to transmit data between a server (such as R), and a web application (such as Google) (Squarespace, 2016). In this way, static maps can be accessed and overlaid with spatial information such as customer address locations. Figure 4 illustrates how several customer addresses are overlaid on a static map accessed from Google.

Stellenbosch University

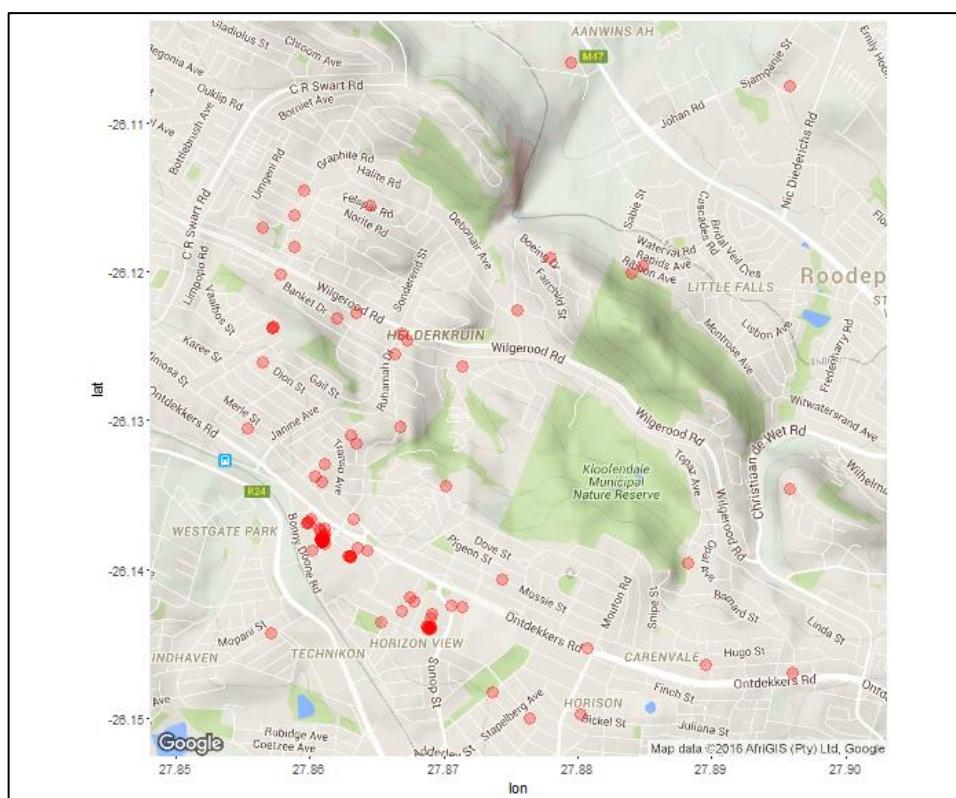


Figure 4: Customer addresses overlaid on a static map

Chapter 3

Data Management

Chapter Aim:

The aim of this chapter is to introduce several techniques that are applied in data management. These techniques ensure that data are preserved and that results can be validated. In addition, the data set that is used in this study is presented together with sample selections. Finally, this chapter presents the validation rules and methods for ensuring that spatial data are independent and can therefore be used for cluster analysis.



Chapter Outcomes:

- Understanding of key data handling concepts.
- Presentation of both internal and external data.
- Presentation of data samples.

3.1 Data Handling Concepts

This chapter discusses the ETL (extraction, loading and transformation) process with data, and presents a flow diagram of how the data used in this study are segmented and selected for experimentation. Theodorou *et al.* (2014) state that ETL processes play an important role in supporting modern business operations that are centred around artefacts (data) that exhibit high variability and diverse lifecycles. In this research, spatial data are transformed at multiple stages of their life cycle in order to create clusters, relationships, and a host of measures that define these relationships. Figure 5 offers a simplified view of the ETL process.

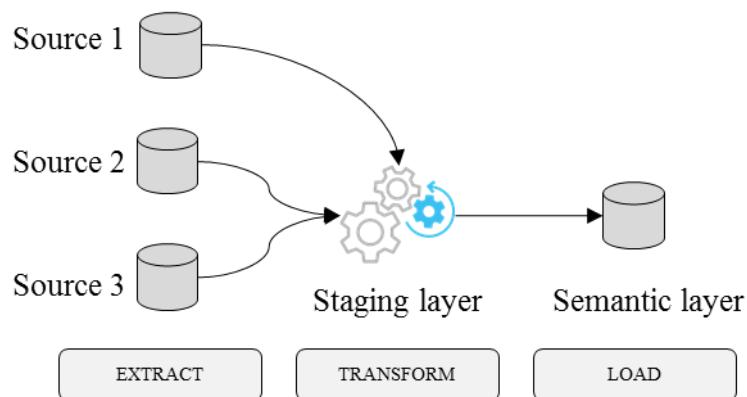


Figure 5: ETL process diagram

Adapted from Theodorou *et al.* (2014)

ETL is a key process for data management and control, as it facilitates the process of data storage (Vassiliadis, 2009). However, this study evaluates a static snapshot of data that have been extracted from a system. An important data quality control is the validation of geocoded customer address data. The manual process of capturing addresses on a system is described in Section 1.2. Address data are stored as master data on one of the retailer's source systems. The master data are updated at a set time once in every twenty-four-hour period. Updates include the addition of new addresses or changes to current addresses. Customer address data associated with an order that has been placed are extracted from the source systems when the order is ready to be processed, and the transformation from a text-based address to a geocoded address takes place using GIS software. Successful conversions (determined by reverse-geocoding quality checks) are loaded into a transport planning system, where the data are stored until the order is planned for delivery to the customer. Customer address data may be updated at any point in the order life cycle by means of manual intervention; but when the delivery has been successfully made to the customer, the order is closed and the all data pertaining to that order are loaded back on to the source system of the retailer and stored for historical record-keeping. The data used in this study are accessed from this historical data storage. In conducting statistical analysis on the data, numerous transformations are required in order to produce insights that can be interpreted and understood.

Horn (2016) and Manjunath *et al.* (2012) present an approach to data transformation when conducting an analytical study that requires data interrogation. This approach is shown below.

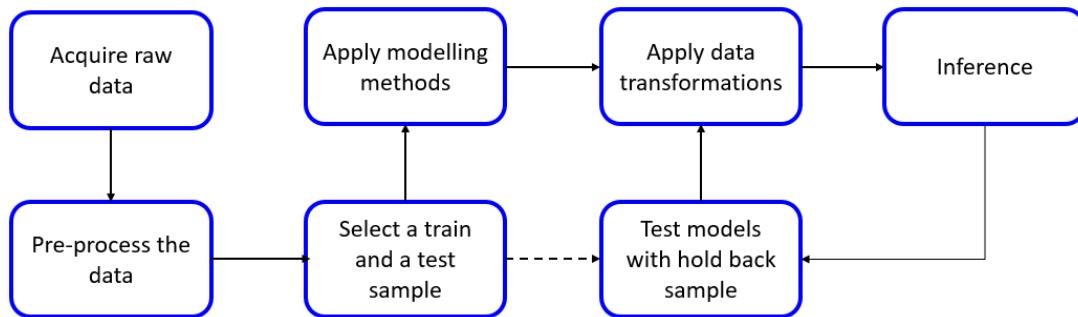


Figure 6: Data transformation approach

Adapted from Horn (2016)

Several steps are then iteratively applied to the data in order to achieve the required transformations. One of these steps, *Select a train and a test sample*, splits the data into two samples in order to produce two semantic layers of data that undergo transformations independently. The reason for this is two-fold. The first is that the test sample is the validation sample, and is used to test how a new set of data performs under the same model parameters. The second reason is to ensure that the model that is used is robust and reliable. In their study on *Sampling Techniques for Data Testing*, Manjunath *et al.* (2012) present a case for cluster sampling, in which train and test samples are selected by virtue of subgroups such as geographic locations, over a consistent period of time. Data transformations are applied to the sample data in order to fit the models more accurately and to make sense of the data by eliminating ‘noise’ (Manikandan, 2010).

The next two sections of this chapter will define what data are extracted and used in this research, and how the train and test samples are identified in order to achieve the objectives stated in this thesis.

3.2 Data Sources

3.2.1 Internal Data

Section 1.2 introduced the *retailer* and gave a brief overview of the operational processes that drive the retailer’s business. A large international logistics company and the custodian of the retailer’s delivery information provided a sample of the retailer’s customer data that are used to conduct this study. The customer data consist of proof-of-delivery (POD) information for deliveries of goods purchased to a host of customers throughout South Africa. Table 1 shows an example of the POD information provided.

Table 1: Example of POD information

Date of delivery	Latitude	Longitude	Region
31 March 2014	-26.523	27.771	Gauteng

The data in Table 1 are regarded as *internal data*, as they are acquired from the retailer's ERP system and contain sensitive information about the retailer's business. Due to the sensitivity of the complete internal data set (customer identity, transaction information, contact numbers, etc.), only the fields shown in Table 1 were provided for the research conducted in this thesis. Given that the information in Table 1 is somewhat limited, it is necessary to supplement this data with external data in order to replace sensitive customer information with alternative information.

3.2.2 External Data

In this study, 'external data' refers to any data that are not directly related to the retailer's customers. The purpose of using external data is ultimately to understand the internal data better by using properties of external data that can be matched to the customer information. Two external data sources are used in this study: population data and geospatial data.

Population data: Population data are acquired from Statistics South Africa (2011), and consist of data that describe living conditions in South Africa. This information is collected every five years through a survey that aims to identify and profile poverty in South Africa, and gives policy-makers information about who is poor, where the poor are located, and what drives poverty in the country (Statistics South Africa, 2011).

Geospatial data: York University Libraries (2016) define geospatial data as data that identify the geographic location of features and boundaries on Earth: natural features, oceans, rivers, etc. Spatial data are generally stored as coordinates (latitude and longitude points) and topology. One such example of geospatial data is shapefiles, which have been developed for numerous countries by the Environmental Systems Research Institute, Inc. (ESRI) (ESRI, 1998). A shapefile is an ESRI vector data storage format used for storing the shape, locations, and attributes of geospatial data (ArcGIS, 2016). This storage format enables the data contained in shapefiles to be used for plotting maps. Figure 7 illustrates the use of shapefiles by showing the municipal boundaries of Gauteng, South Africa.

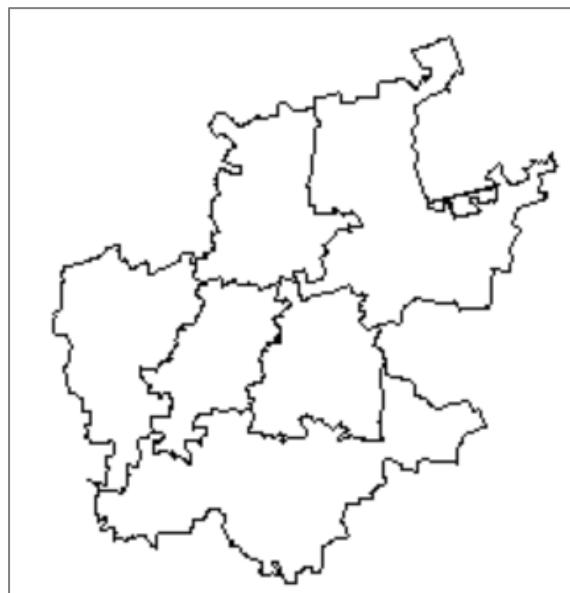


Figure 7: Plot of Gauteng, RSA (R Core Team, 2016) using shapefile data

The ESRI shapefile used to plot Figure 7 was acquired from the Municipal Demarcation Board (2016).

3.3 Data Samples

Following the data selection approach proposed by Manjunath *et al.* (2012), an appropriate train and test sample is selected. However, before selecting samples from the data, the parameters that constrain this study need to be defined. The parameters of the internal data set used in this study are given below.

Table 2: Parameters of internal data used

Parameter	Description
Geography	South Africa, by province
Date Range	1 Jan 2013 – 31 Dec 2013

The constraints of the test and train samples are defined in the table below.

Table 3: Sample constraints

	Train Sample	Test Sample
Geography	Gauteng, South Africa	Gauteng, South Africa
Date Range	1 Jan 2013 – 30 June 2013	1 July 2013 – 31 Dec 2013
Objects (customer addresses)	25 312	28 443

Stellenbosch University

The constraints in Table 3 show that only customer deliveries in Gauteng, South Africa are considered in this research. In addition, a date range of six months has been used for *both* samples (50:50). The reason for the geographical constraint is to simplify the results of this study by evaluating customer deliveries in one province rather than in several. A 50:50 split has been selected, based on the logic that an uneven split of the data could result in the test sample not having sufficient customer data to match to population data attributes⁸. This is an initial assumption, and is verified in the results.

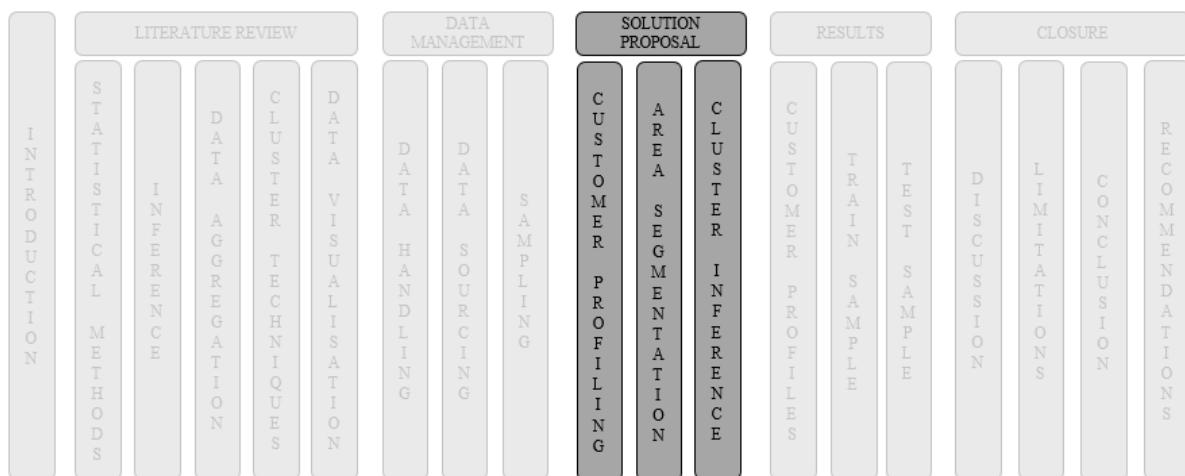
⁸ Sufficient customer data need to be presented in order to identify *significant* customer clusters.

Chapter 4

Methodology

Chapter Aim:

This aim of this chapter is to introduce the methodology that has been applied in modelling the data presented in this study. The methodology is split into three sequential parts that show the progression from producing hindsight, using descriptive statistical techniques, to insight where inferential techniques have been applied.



Chapter Outcomes:

- Presentation of the methodology applied in profiling customers.
- Presentation of the methodology applied in geographically segmenting customers.
- Presentation of the methodology applied in determining which variables account for high-density customer areas (clusters).

4.1 Methodological Framework

The high-level methodology applied in this research is presented in the form of a flow diagram in Figure 8. This diagram shows the three primary tasks required to achieve the objectives set out in this thesis.

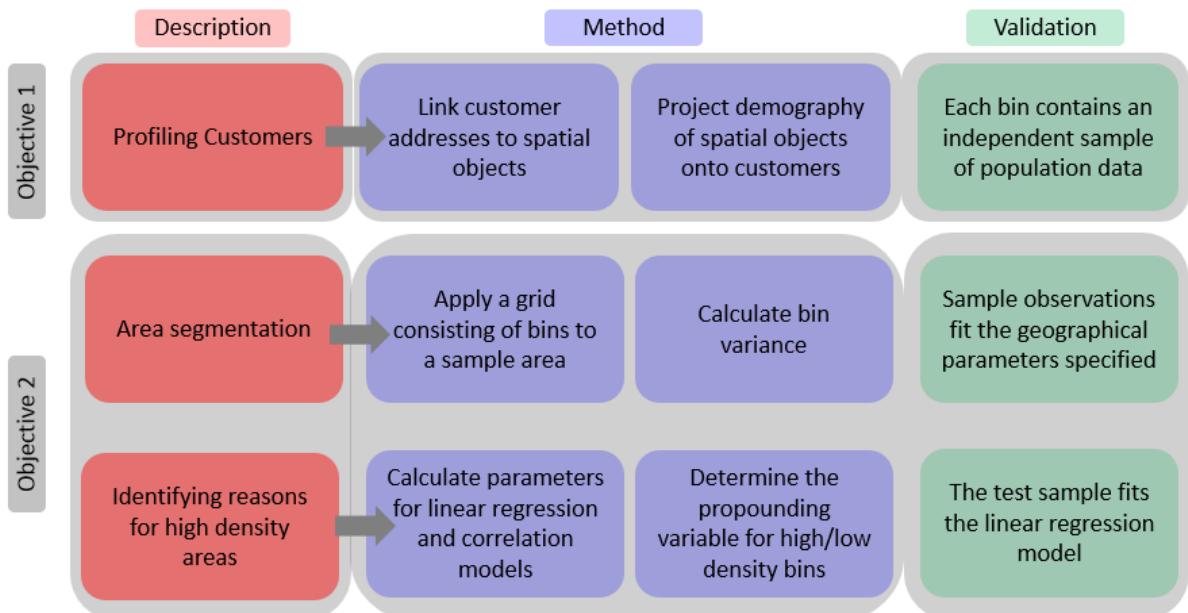


Figure 8: Methodological framework

In Chapter 3, the necessity of a software tool to handle spatial data is mentioned. All computations required by the methodology proposed in Figure 8 are performed using R software (version 3.2.5) (R Core Team, 2016). R is a statistical programming language that supports numerous libraries containing embedded functions that support the computation of complex algorithms. Many of these libraries are used in this study, and are explained in the sections to follow. Although the original R code is shown in Appendix 1, pseudocode – which demonstrates the application of library functions – is included in this chapter. Table 4 expands on the flow diagram of Figure 8 by presenting a detailed summary of the methods, techniques, and supporting R libraries that are used in this chapter.

Table 4: Detailed summary of methodological framework

Method Description	Tasks	Research Method	R Libraries (R Core Team, 2016)
Customer Profiling	Data Transformation	N/A	<i>dplyr, stringr, rgdal, RColorBrewer, grDevices, gdata</i>
	Multivariate Analysis	<i>Principle Component Analysis</i>	<i>FactoMineR, factoextra, psych</i>
	Goodness-of-Fit Test	<i>Linear Regression</i>	<i>dplyr, stats, ggplot2</i>
	Test Sample Analysis	N/A	<i>stats, ggplot2</i>
Area Segmentation	Cluster Method	<i>Grid-based Clustering</i>	<i>maptools, rgeos, ash2, dplyr, geosphere, raster, stringr, plyr, ggplot2, tidyr, ggmap</i>
	Grid Dimensioning	<i>Scott's (1979) Rule</i>	<i>stats, grDevices, cloud, lattice</i>
	Distribution Characteristics	<i>Variance</i>	<i>stats, dplyr</i>
Density Inference	Data Transformation	N/A	<i>dplyr, stringr, splitstackshape</i>
	Variance Indicators	<i>Correlation</i>	<i>stats, ggplot2, RgoogleMaps, corrgram, corrplot, dplyr</i>

4.2 Customer Profiling

This section discusses how profiles are created for customers in Gauteng. The profiles are based on population census data that include a host of descriptive fields about the population of Gauteng. The objective of profile creation is to produce a view of what the ‘average’ customer looks like for certain administrative regions of Gauteng. Section 3.3 mentioned the use of shapefiles, acquired from the Municipal Demarcation Board (2016), that can be used to identify the administrative boundaries of a country or region. The shapefile for South Africa contains boundaries at various degrees of granularity, ranging from provinces to electoral wards. Given that the scope of the customer data used in this research is limited to Gauteng, customers will be profiled at a district level within the province of Gauteng. Therefore, for each of the six municipal districts of Gauteng, a profile will be created that best depicts the attributes of that district’s population. In order to produce accurate depictions, the *most granular* administrative boundary containing a *customer location* should be used. Therefore, the electoral ward in which each customer resides will be used to identify individual customer attributes. The population data of these wards will be projected on to the customers, such that the following statement can be made about each customer: *customers are assumed to exhibit the attributes of the population of the electoral ward in which they reside*. Following this assumption, a second statement can then be made about a group of customers residing in a larger area; customers are assumed to exhibit the average attributes of *all* the customers⁹ in that region. The steps applied to producing customer profiles (based on the two assumptions stated above) are given below:

1. Determine both the district and the most granular administrative boundary (i.e., electoral ward) in which each of the customers in the sample area resides.
2. Identify the characteristics of the *average* customer for each electoral ward.
3. Determine the characteristics of the *average* customer for each municipal district of Gauteng.

The sub-sections below show how each of the three steps of the methodology is achieved.

4.2.1 Determining Administrative Boundaries

The requirement for the first step in the methodology for customer profiling is to determine in which administrative area they live, at both a district and an electoral ward level. To do this, the shapefile containing all the administrative areas for South Africa needs to be used. The *rgdal* library in R is used to import the shapefile and read its contents, which consist primarily of lists of boundary coordinates at province, municipal district, and electoral ward levels. Figures 9, 10 and 11 illustrate these three levels of South Africa’s administrative boundaries.

⁹ These customers’ attributed are determined by those of the ward in which they reside.

Stellenbosch University



Figure 9: Provinces of South Africa



Figure 10: Municipal districts of Gauteng

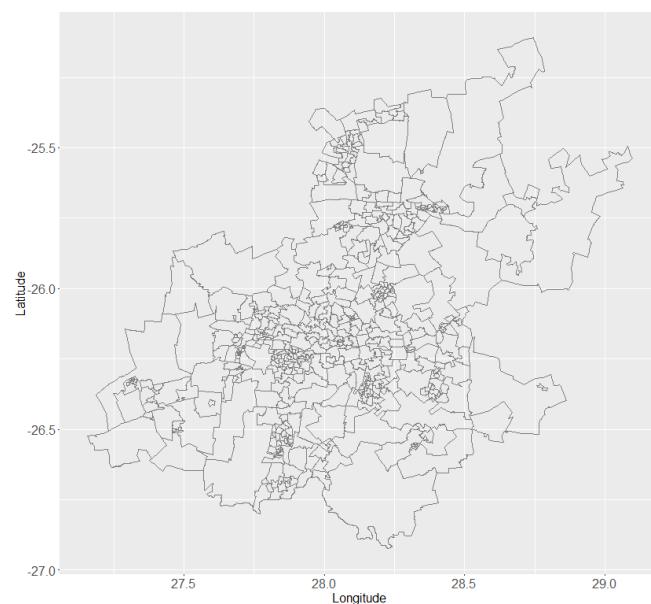


Figure 11: Electoral wards of Gauteng

This shapefile is imported as follows:

```
require(rgdal)
Shapefile = readOGR (read in shapefile from root directory)
```

This shape contains the fields shown in Table 5.

Table 5: Example of shapefile fields

	Province	Category	Municipality name	Ward ID
<i>example</i>	Gauteng	GT423	Lesedi Local Municipality	74203013

A function is required that assigns a ward identity to each customer by computing whether or not each customer address is contained within the boundary of each ward shown in Figure 11. This is computed using the *point.in.polygon* function from the *sp* library (Pebesma and Bivand, 2005; Bivand *et al.*, 2013) in R. The function verifies for one or more coordinate points whether they fall in a given polygon or *boundary* (R Core Team, 2016).

```
for (loop through each customer address in the sample) {
  for (loop through each ward in the shapefile) {
    Var1 = [store each customer address in a temporary variable]
    Test = polypoint.in.polygon (Test whether the customer address exists
      within each ward's boundary)
    Assign = ifelse (assign a binary output for the test)
    Var2 = which (Select the ID of the ward which scored a positive test)

  Var3 = unlist (Store each ward ID in a vector)
```

The output of this function is a vector containing a ward identity (ID) corresponding to each customer address in the sample. Wards, which are associated with municipalities (Table 5), are now linked to individual customer addresses. A frequency count is performed on the number of customers within each ward, and this number is associated with each unique ward ID.

```
Customers = data.frame (table (Count the number of customers per wards))
UniqueWards = merge (Merge the number of customers per wards to each unique
  ward ID)
```

4.2.2 Profiling Variables

Statistics South Africa (2011) provides several fields of information about the population of South Africa. For the objectives of this research, the following four fields are used to create customer profiles of the retailer's customer base in Gauteng:

1. Age
2. Monthly income
3. Family size
4. Population group

These four indicators have been chosen for this study because they will provide a good representation of the type of product that the customer will be interested in, based on age, population group, and family size, as well as an affordability indicator of products in different parts of Gauteng. The values associated with these indicators are shown in Appendix 4. Note that the mean values for each variable (per municipal district in Gauteng) have been deduced from the population data set from Statistics South Africa (2011). A summary of the population data per electoral ward is given in Appendix 2. Equation (1) has been applied to achieve the mean of each variable.

4.2.3 Determining Customer Characteristics

This section shows how population characteristics are projected on to customers in proportionate measures. That is, given a sample of 200 customers residing in various wards within a municipal district, if 50 of these customers reside in Ward A, then 25% ($50/200$) of Ward A's population data are attributed to the profile of the average customer for that municipal district. This proportion of customers per ward versus total number of customers per district is multiplied by the number of customers in each ward. This calculation is shown below:

$$X_{jk} = \sum_j^i \frac{c_i}{d_j} \times w_{ik} \quad (28)$$

where:

$$i = [1, 508]$$

$$j = [1, 9]$$

$$k = [1, 4]$$

X_{jk} = average proportionate population of field, k for district, j

c_i = customers in ward i

d_j = customers in district j

w_{ik} = population values of field k for ward i

The *ddply* function from the *plyr* library in R, for each subset of a data frame or data table, applies a function (in this case the 'sum' function), and then combines the results into a new data frame. This function is an effective way of performing the computation of Equation (28) across multiple variables.

```
Xjk = data.frame(ddply (Aggregate population data based on the calculated
proportions of customers in wards, for each district))
```

Stellenbosch University

From the new data table containing the segmented results, the values of all four population data fields are attributed to the average customer of the retailer who resides in each respective municipal district.

4.3 Area Segmentation

This next part of the methodology addresses the problem of supporting a location-based marketing strategy for the retailer. The sub-sections that follow describe the sequence of steps that are applied in order to identify, first, where customer clusters exist, and second, *why* clusters exist. Given the fact that spatial data about customers are available, the grid-based cluster method described in the literature study in Chapter 2 is the most appropriate method for achieving the desired outcome. The grid-based clustering method is recommended when using spatial data on a point coordinate system (University of Toronto, 2002), and relies on the application of a $m \times n$ grid that is applied to a sample area or *window*. The grid therefore overlays the sample area and records the number of objects contained within each element or *bin* of the grid. Bins with higher object counts are more densely-populated (in terms of objects per unit area) than bins with lower object counts. This statement can be made, given that all the bins are exactly the same size.

4.3.1 Grid Dimensioning

The information about customer clusters that is produced in this methodology needs to be interpretable, and so bins cannot be too large or too small in size and number. Logically, if a large area – e.g., 20,000 km^2 consisting of more than 20,000 objects – was segmented by a very small number of bins – e.g., ten bins – then the objects contained in each of those ten bins would not add significant value to the information about the density of the area. Conversely, having a large number of bins – e.g., 10,000 bins – would result in data that are too granular, leaving little inference to be made about the density difference of the bins. A large number of bins, each containing a few or no objects, would distort the identification of densely-populated areas by flooding the information with a micro-segmented area. In the context of this study, the aim is to segment the sample area of *Gauteng* in order to produce location-based marketing information that is of practical use to the retailer. Therefore, an optimal bin size needs to be selected in order to produce areas of the retailers' customers that can be penetrated by marketing campaigns. Scott's (1979) rule determines an optimal bin size through the use of a multi-dimensional histogram. This method places a grid over the sample area of Gauteng with bins of appropriate dimensions. The *nclass* function from the *grDevices* library in R provides the different methods of bins size calculation presented in Chapter 2. This function is used as follows to determine the optimal number of bins, both vertically and horizontally, across the sample area:

```
Number of latitude bins = nclass.scott(customer latitude points)
Number of longitude bins = nclass.scott(customer longitude points)
```

Stellenbosch University

Scott's formula produces a matrix that contains 55 bins in the latitudinal (vertical) direction and 47 bins in the longitudinal (horizontal) direction. This equates to 2,585 bins in total, each having a surface area of 9.84 km^2 . The grid is applied to the sample area by specifying the bin dimensions in the *ash2* function from the *ash* (*Average Shifted Histograms*) library in R, which computes a bivariate ASH estimate or *product polynomial kernel*. The function requires several inputs to compute the estimates:

```
mat = matrix (create a 2x2 matrix of the max and min sample area coordinate
points)
nbin = (create a 2x1 matrix of the bin sizes calculated by Scott's formula)
bins = bin2(create an object using the bin2 function that stores all the
above variables with customer coordinates)
grid = ash2(create an object using the ash2 function that stores all the grid
data in a list)
```

The final output contained in the variable `data` is represented by a list that contains multiple embedded data structures, including:

- The coordinate points of each bin.
- The number of customers from the sample contained in each bin.
- The probability density function for the customers contained in each bin.

Figure 12 shows the grid's bins in both horizontal and vertical dimensions, with the z-axis indicating the number of customers in each bin. This plot is constructed using the *cloud* function from the *lattice* library in R, which contains generic functions that are used to draw 3D scatter plots and surfaces.

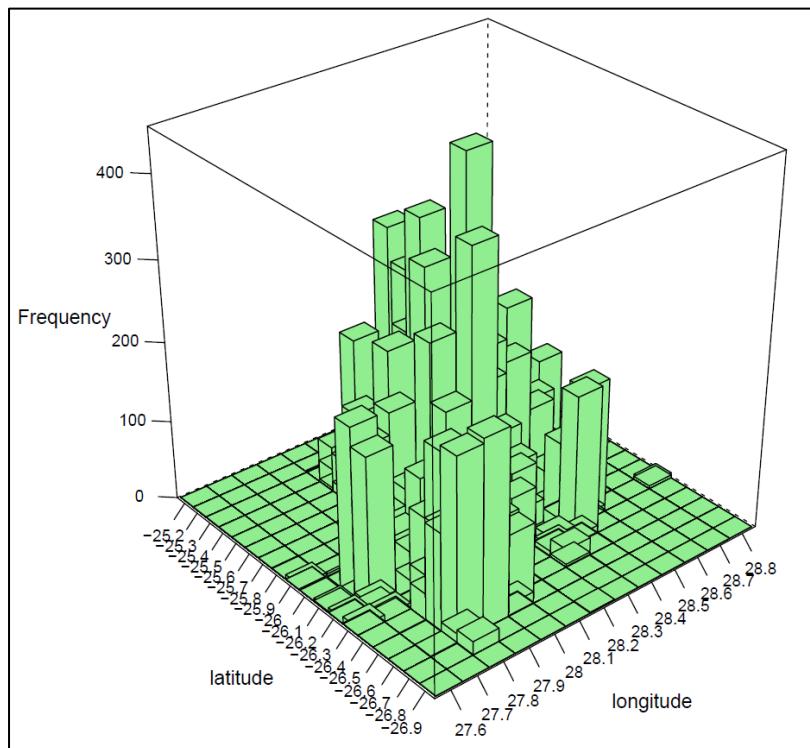


Figure 12: Multi-dimensional histogram showing bins of the sample area

Stellenbosch University

A density plot can now be made using the *ash2* object, which presents a view of those bins containing many customers and of those that do not. To do this, the *raster* function from the *raster* library is used. The data contained in the `data` variable need to be converted to raster data. Raster data divides space into cells (rectangles or pixels) of equal size according to the units of a coordinate reference system (the *ash2* function produced such coordinates from the bin dimensions provided). Such continuous spatial data are also referred to as *grid* data, and can be contrasted with discrete (object-based) spatial data (points, lines, polygons, etc.) (R Core Team, 2016). Raster data are useful for showing heat maps of spatial data by grouping observations – or in this case, customers – into equally-sized bins.

```
est.raster = raster (create a raster object by calling the bin coordinates  
and probability density function previously calculated)  
plot (create a plot of the raster object and include a colour scale to  
indicate density)  
con tour (add con tour lines to compliment density indicators)
```

The plot of the raster object is shown in Figure 13. A colour scale (blue to red) indicates bins that contain few (blue) to many (red) customers. This image gives an indicative view of where customer clusters are expected to be found.

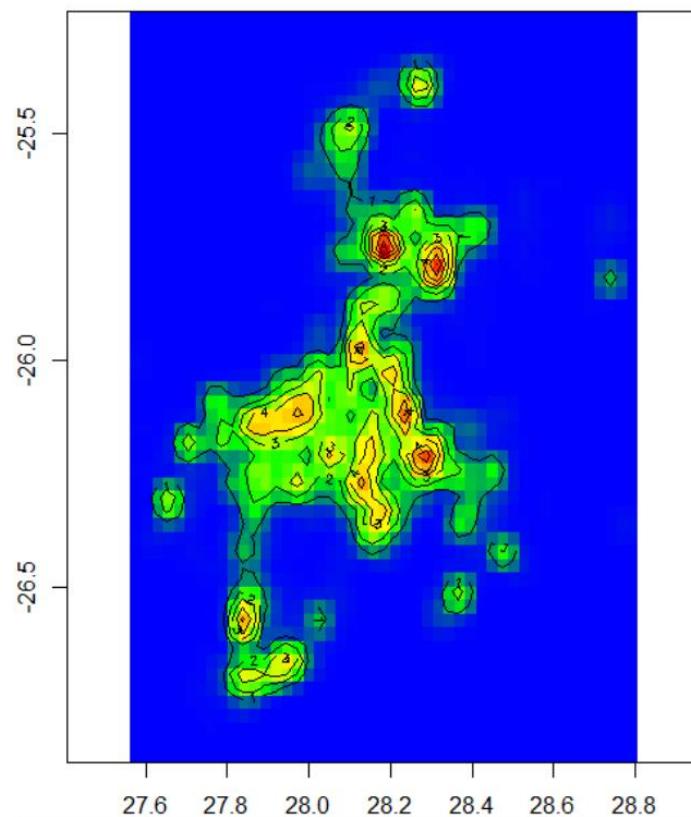


Figure 13: Raster plot of customers in the sample data

Stellenbosch University

While the customer density view of Figure 13 is a useful visual representation of where the retailer's customers are located, the view presents the data at a high level, and so omits detail. A quantitative analysis of high-density areas at a more granular level will enable the retailer to pinpoint key areas of interest. While customer density per unit area is a useful measure of density, determining the measure of how clustered customers are requires additional computation above simply counting the number of customers per bin. Equations (2) and (4) from Chapter 2 suggest a method of calculating measures of variance using distance as a determinant. The literature presented in Section 2.1 makes further use of the location of a mean 'virtual' customer in each bin, representing the average weighted position of all the customers contained within a bin. The variance of the distances from all the customers in a bin to that bin's mean customer presents quantitative measures of low- and high-variance bins. These measures illustrate a granular view of the raster plot in Figure 13. The steps followed in determining the variance of each bin are summarised below.

1. Calculate the mean customer coordinates for each bin.
2. Calculate the distance from each customer coordinate to the mean coordinate of each bin.
3. Calculate the variance of each bin by using the distance measures.

While the steps listed above appear to be computationally taxing, R provides several libraries that accommodate these calculations. The mean coordinates of each bin are calculated by applying the *mean* function in R's base libraries. This function calculates the mean value of a set of numbers according to Equation (1). Given that there are 2,585 bins and more than 25,000 customers in the test sample area, the *sapply* and *split* functions from the R base library enable the application of the *mean* function to groups – which, in the case of the customer data base, are bins. The mean coordinates are therefore computed as follows:

```
Mean latitude per bin ( $\mu_1$ ) = sapply (split (all latitudes, Bins), mean)
Mean longitude per bin ( $\mu_2$ ) = sapply (split (all longitudes, Bins), mean)
```

The mean latitude and longitude points for each bin are now used in a function that calculates the distance from each customer coordinate to the mean coordinates, and calculates the variance using the base *var* function in R.

```
for (loop through all horizontal bins) {
  for (loop through all vertical bins) {

    f1 = function (calculates variance of each bin)
    test = function (calculates customer distance to mean)
    dcalc = function (stores bin and customer coordinates)
    d = assign the distance formula from Equation (4)

    Distance [i] = store each distance value in a vector
    varVec[i] = var(store each bin variance in a vector)
  }
}
```

Stellenbosch University

The output of the function above is a variance measure for each bin. This variance measured is organised from the smallest to the greatest variance in order to identify the most- to least-clustered bins.

4.3.2 Bin Characteristics and Variance Significance

Section 4.3.1 showed how varying degrees of customer clusters can be identified by applying Equations (1) to (4). Now this section presents a view of the statistical characteristics of the bins that contain the customer addresses from the train sample. These characterises are a useful way of summarising granular data so that the retailer can make sense of where clusters are located and how significant the clusters appear to be.

All the bins created in this methodology are uniformly-sized, but contain varying numbers of customers. Bins containing a small number of customers – e.g., *three customers* – might show significantly low variance measures, as there is little variation between three observations. A logical model parameter needs to be set in place that serves as a threshold for the number of customers in a bin that will be considered insignificant. A threshold of 40 customers per bin is set in place in this study, and thus only bins of 40 customers or more are considered for significant customer density. Table 6 shows the effects of applying this threshold.

Table 6: Reduction in bins and customers of the train sample from applying the threshold

	Number of bins	Number of customers
Original sample	2,585	25,312
Customers / bin ≥ 40	175	18,820

After removing bins with an insignificant number of customers, the remaining 175 bins can now be considered to have a significant variance measure, even though the number of customers per bin might still vary. The fact that variance has been calculated for varying bin sizes still implies that no conclusions may be drawn as yet. For example, a bin with a low variance and small size might be statistically equivalent to a bin with a higher variance measure and a larger size. Levene's test of variance equality is performed to determine whether or not the 175 bins have an equal variance. The *leveneTest* function from the *car* library in R uses Levene's test statistic in Equation (5) to test for homogeneity of variance across all the bins. The hypothesis for this test states:

$$\begin{array}{ll} \text{Null} & H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_t^2 \\ \text{Alternative} & H_1: \text{Not all variances are equal.} \end{array}$$

The test is performed in R as follows:

Stellenbosch University

```
leveneTest (perform Levene's Test on bin having >= 40 customers)
```

The result of the test indicates a significantly small p-value, and so suggests a rejection of the null hypothesis that variances across all bins are homogeneous.

```
➤ Output:
Levene's Test for Homogeneity of Variance
  Df F value    Pr(>F)
group   175  4.2425 < 2.2e-16 ***
                18820
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1
```

The group of 175 bins can now be said to have unequal variances of significant magnitude. In addition, the variances of these bins provide a *true* representation of areas where there is a significant degree of clustering of customer locations. The variance from least to most for all bins with ≥ 40 customers is shown in Figure 14.

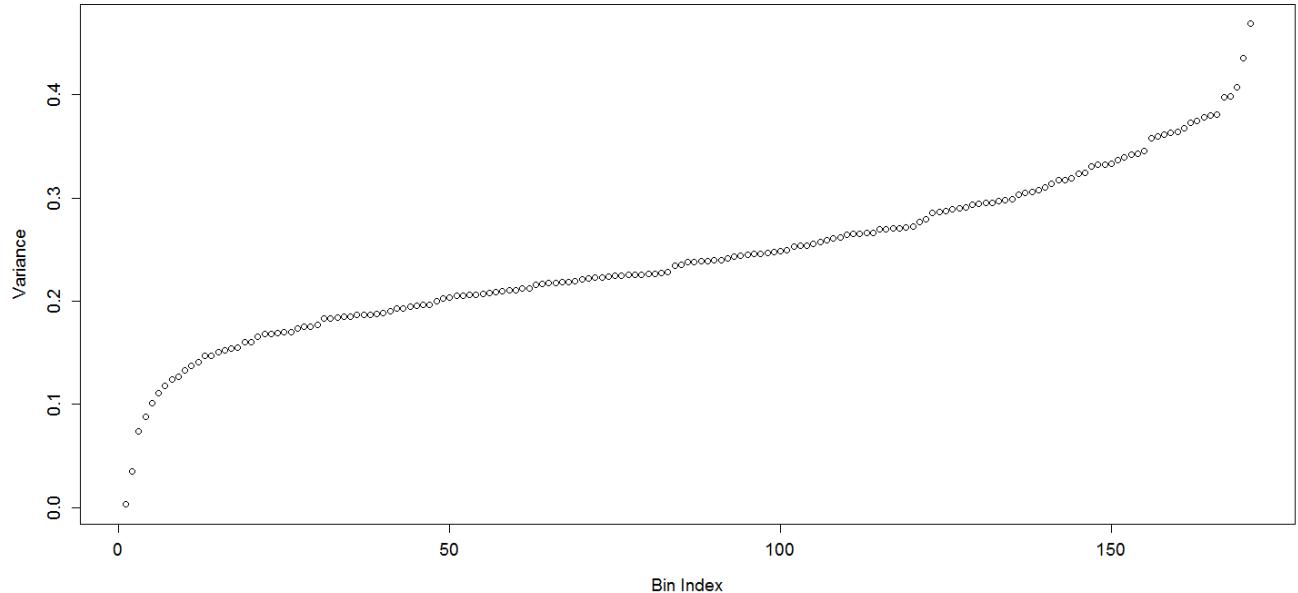


Figure 14: Variance of bins in the train sample data

4.4 Cluster Inference

Section 4.3 showed how bin variance can be used to determine cluster significance in various parts of Gauteng. This next section evaluates several population variables with the aim of determining why customers are significantly clustered in some bins and not particularly in others. The variable(s) that account for clustering are called the *propounding variables*, and these are identified by means of a multivariate analysis whereby the principal components of the sample data set are extracted and

Stellenbosch University

evaluated for their contribution to low- or high-variance bins. Once the propounding variables are identified in the multivariate model, the *test* data sample is plugged into the same model to test how the model performs, given a new data set. In order to build a model that can evaluate the effect of population data variables on the variance of bins (which contain customer locations), the population data need to be transformed such that each bin contains a unique (mutually-exclusive) set of population data.

4.4.1 Data Transformation

Census data are stored in groups of uniquely-shaped polygons that are determined by administrative geographical boundaries – i.e., electoral wards, municipal districts, provinces, etc. Bins, however, are defined by different parameters: those that are created in the grid-cluster method. The population data need to be projected on to the bins in order to assess what the population characteristics are of the customers within each bin. The population data therefore need to be transformed from their current grouping of variables (by administrative areas) to each bin's parameters. This can be achieved by assigning population data to bins in exactly the same proportion in which actual customers were assigned to bins in the grid-cluster method. The administrative boundaries of electoral wards are used for population data projection, as these are the most granular form of boundary available in the census data, and are therefore most useful in projecting accurate proportions at the most granular level. Figure 15 explains this illustratively.

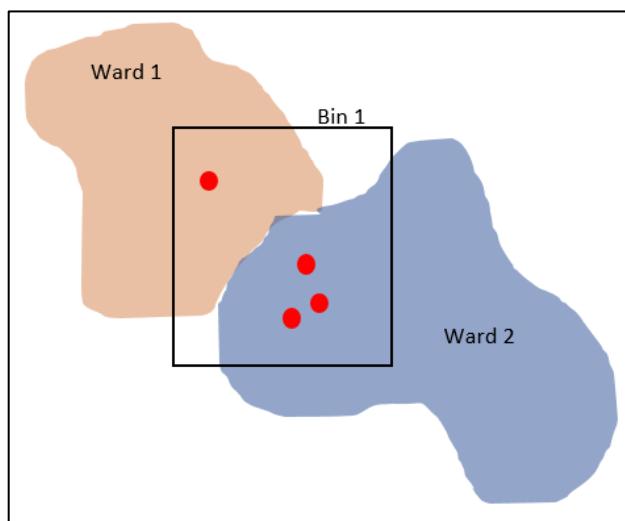


Figure 15: An example of data transformation logic

Figure 15 shows an example of multiple customers found in one bin. These customers live in two different electoral wards, each with its own population information (contained in the census data). The boundaries of the wards and bins intersect at several points, and portions of both wards are found within the bin. The aim is to project the population data from the wards on to the bins according to some logical approach. The customers can be used to select a proportional subset of both wards' population data for

Stellenbosch University

projection on to the bin. In this example, one-quarter of all the customers in Bin 1 live in Ward 1, and the remaining three-quarters *live* in Ward 2. Therefore, 25% of the population data of Ward 1 and 75% of that of Ward 2 will be projected on to Bin 1. This method of apportioning the population data ensures that each bin contains the population data from one or more wards, attributed by mutually-exclusive customers. This method can be tedious, given that the census data contain 509 electoral wards in Gauteng that need to be overlaid with 2,585 bins. For this reason, a function is created in R which simplifies the calculation:

```

#Count number of customers in each ward
Var1 = data.frame(table (count number of unique customer-ward identities))

#Calculate the number of wards intersecting each bin
Var2 = paste (associate unique bin/ward identities)
Var3 = data.frame(table (Count the number of bin/ward identities))

#Calculate percentage of customers per ward per bin
Var4 = merge (merge Var1 and Var3 to obtain customer/bin/ward)
Percentage = Frequency of wards per bin/frequency of customers per ward

#Apply the above-calculated percentage to census (population) data

for (loop through all census data fields) {
  for (loop through all unique bin/ward identities) {
    a = proportion of customers per ward per bin
    proportion = multiply percentage by each population data field
  }
}

#Aggregate the proportions of data from various wards for each unique bin
Aggregate = data.frame(ddply (aggregate data by unique bin ID))

```

The output of this code produces a table containing the aggregated proportion of population data from all wards for each unique bin. Each bin now has several fields of population data assigned to it, as well as a variance measure that was calculated in Section 4.3.2. The next section involves data exploration and inference to test whether or not any of the population data are responsible for varying measures of variance in the bins.

4.4.2 Multivariate Analysis

In this section, a PCA is performed to identify one or more propounding variables in the data that account for customer clusters (high-density bins). The census data that are considered in this study contain over 100 fields of descriptive living conditions of the population of South Africa. The PCA is performed on a subset of these variables, based on logical selection; this is discussed later in this section. The multivariate analysis (performed by means of a PCA) tests the validity of the following research question: *The factors that cause customer clusters are those pertaining to the infrastructure of customers' residences – i.e., the classification of the residence.* This research question is based on the logic that customers who reside in large, free-standing properties are more likely to be geographically dispersed than customers who reside in more confined residences such as housing clusters, complexes,

Stellenbosch University

or apartments. To answer the research question posed in this study, all population variables pertaining to the category of *dwelling* will be considered in the PCA. These variables are shown in Table 7.

Table 7: Census data considered in the PCA study

Category	Sub-category	Field
Dwellings	Type of main dwelling	House or brick/concrete block structure on a separate stand or yard
Dwellings	Type of main dwelling	Traditional dwelling/hut/structure made of traditional materials
Dwellings	Type of main dwelling	Flat or apartment in a block of flats
Dwellings	Type of main dwelling	Cluster house in complex
Dwellings	Type of main dwelling	Townhouse (semi-detached house in a complex)
Dwellings	Type of main dwelling	Semi-detached house
Dwellings	Type of main dwelling	House/flat/room in backyard
Dwellings	Type of main dwelling	Informal dwelling (shack; in backyard)
Dwellings	Type of main dwelling	Informal dwelling (shack; not in backyard)
Dwellings	Type of main dwelling	Room/flatlet on a property or larger dwelling/servant's quarters/granny flat
Dwellings	Type of main dwelling	Caravan or tent
Dwellings	Type of main dwelling	Other dwelling

Two libraries in R cater for the steps required in a PCA, as discussed in Section 2.1.3.3. These are *FactoMineR* and *Factoextra*. Lê *et al.* (2008) describe *FactoMineR* as an effective library for partitioning variables in a multivariate analysis, placing variables in hierarchical order. It is useful, they state, when working with different data structures. *Factoextra* is useful for extracting and visualising the output of multivariate data analyses such as PCA (Kassambara and Mundt, 2016). These two libraries are used, as shown below, to perform a PCA using the population variables shown in Table 7.

```
#Subset the population data
data = subset (subset the Dwellings-related variables)

#Transform the data - calculate the ratio of customers per bin to population
per bin for ascending order of bin variance
for (loop through all Dwellings variables) {
  Ratio = (calculate ratio of sales to population per bin for each variable)
}

#Principal component Analysis using FactoMineR
res.pca = PCA (apply the PCA function to create an object of class PCA)

#Show the eigenvalues and percentage of variance
eigenvalues = [extract the eigenvalues from the PCA object]
eigenvalues [print the eigenvalues]

➤ Output:
```

Stellenbosch University

Table 8: Eigenvalues of the PCA object

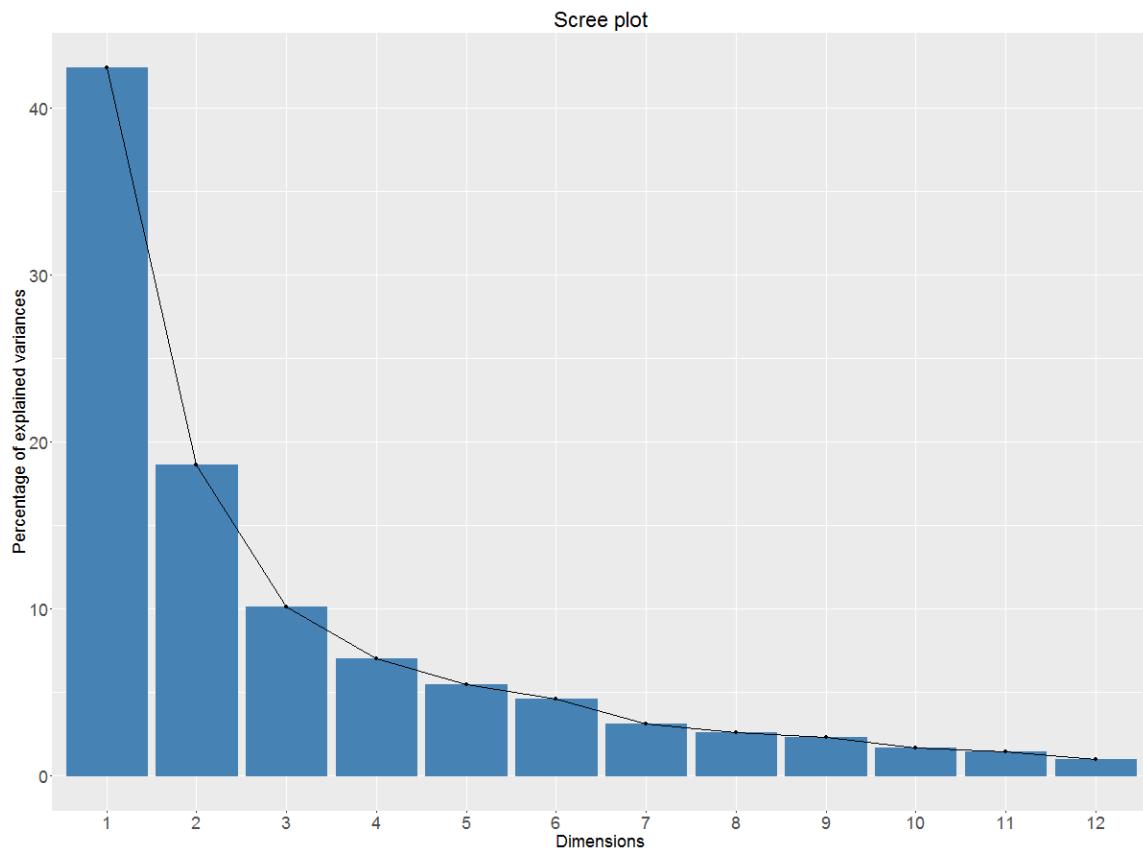
Principal components	Eigenvalue	Percentage of variance
comp 1	5.09	42.40
comp 2	2.23	18.60
comp 3	1.21	10.08
comp 4	0.84	7.01
comp 5	0.65	5.42
comp 6	0.55	4.58
comp 7	0.37	3.09
comp 8	0.31	2.55
comp 9	0.27	2.25
comp 10	0.20	1.65
comp 11	0.17	1.40
comp 12	0.12	0.97

The output from the PCA object is shown in Table 8, and presents the eigenvalues and percentage of variance that each of the 12 calculated principal components (PCs) accounts for. The next step is to select and inspect the ‘significant’ components that the PCA has produced. The non-quantitative method of component selection, presented by Jackson (1993) and Peres-Neto *et al.* (2005), is that of looking for a natural break between the ‘large’ eigenvalues and the ‘small’ eigenvalues using a scree plot. The *factoextra* library is used to create a scree plot using the *fviz_screeplot* function.

```
fviz_screeplot (create a scree plot of the PCA object)
```

Figure 16 is a graphical representation of the components and their respective percentage of variance for which each accounts. By visual inspection, the natural break in the curve appears somewhere between PC 3 and PC 4.

Stellenbosch University

Figure 16: Scree plot of the PCA object containing *Dwellings* variables

As a more quantitative selection method, Jackson (1993) and Peres-Neto *et al.* (2005) propose retaining the components whose eigenvalues are greater than the average of all the eigenvalues, which is 1. Based on this section method, the PCs with an eigenvalue greater than 1 in Table 8 are PC 1, PC 2, and PC 3. These three principal components account for 71.08% of the total variance of population data for variables relating to dwelling type. The standardised values of the original 12 variables are evaluated, based on their correlation with the three respective principal components. The standardised values are embedded in the PCA object, and are extracted as follows:

```
PCA_Object$var$cos2
```

These values are shown in Table 9. From the values in this table, non-significant correlation values are disregarded; a threshold limit is required to determine this. Several statistical tests and criteria for eigenvalue assessment are proposed by Peres-Neto *et al.* (2003), one of which suggests using the cut-off rule of 0.5, as proposed by Richman (1998) in his published study. This method is less accurate than the more scientific methods proposed by Peres-Neto *et al.* (2003); but, given that this study entails exploratory analysis of a research question rather a set of hypotheses, this threshold will be used. Streiner and Norman (2011) support this method when conducting exploratory analysis.

Stellenbosch University

Table 9: Correlation values of standardised variables for each principal component

Variables	PC 1	PC 2	PC 3
House or brick concrete block structure	0.69	0.05	0.01
Traditional dwelling hut	0.70	0.01	0.00
Flat or apartment in a block of flats	0.07	0.74	0.04
Cluster house in complex	0.14	0.68	0.02
Townhouse semi-detached house in a complex	0.12	0.53	0.11
Semi-detached house	0.62	0.00	0.10
House flat room in backyard	0.29	0.02	0.44
Informal dwelling shack in backyard	0.50	0.09	0.01
Informal dwelling shack not in backyard	0.48	0.07	0.27
Room flatlet on a property or larger dwelling	0.51	0.00	0.13
Caravan or tent	0.23	0.00	0.07
Other dwelling	0.75	0.04	0.00

The values above the threshold limit of 0.5 are shown in Table 9. Extracting these variables results in eight variables that show significance, out of the original 12 that were analysed. The PCA therefore has reduced the set of *Dwelling type* variables by 33.34% (from 12 to eight). These eight variables are sufficiently significant in their relationship to the number of customers contained in each bin. In the next section, these variables are used in a correlation study to test which of them accounts for low- and high-bin variance.

4.4.3 Correlation Analysis

Each of the eight variables identified in the PCA has an associated population size for each bin. This section aims at identifying which of the eight variables correlates with the high- and low-variance bins respectively. The variables are also fitted into a linear regression model where bin variance is used as the *regressor* and the ratio of customers per bin to population size (for each of the variables) is used as the *predictor*. This is to compute the goodness-of-fit of the variables, and to identify those that are significant.

In R's *stats* library, the *lm* function provides the framework for a linear regression model, as shown below:

```
LinearModel = lm (Regressor [Bin Variance] ~ Predictor(s) [Variables from
PCA])
summary(LinearModel)
plot(LinearModel)
```

The detailed results from this model show the goodness-of-fit by testing the calculated *p-value* at a 95% confidence in interval. If the data fit the model, it could be expected that there would be some significant correlation between bin variance and the variables. However, when trying to identify the correlation

Stellenbosch University

between low- and high-variance bins and the eight selected variables, a different approach is required. Recall the profile of bin variance shown in Figure 14: this image shows that the most extreme cases of high- and low-variance bins are found at the head and tail of the profile respectively. The further away from the two ends of the profile, the more linear the variance profile becomes. The method applied in this study therefore tests for correlation only for the extreme cases of low and high variance – which can be identified by the top and bottom 10 bins in the sample area respectively. A useful function from the *stats* library is the *cor.test* function, which performs a correlation test for a specified confidence interval. A function is required that tests for correlation between bin variance and variable ratio¹⁰ for the 10 lowest- and highest-variance bins.

```
#Top 10 Bins
for (Loop through the PCA variables) {
    print (cor.test (print results from correlation test for top 10 bins))
}

#Bottom 10 Bins
for (Loop through the PCA variables) {
    print (cor.test (print results from correlation test for bottom 10 bins))
}
```

The population variables relating to *Dwelling type* that have a significant correlation with low- and high-variance bins are now identified. The research question that asked whether this type of population category can be used to explain clustering can therefore be answered with statistically justified reasoning. In addition, other variables within the category of *Dwelling type* have been identified that show varying degrees of significance in a goodness-of-fit test defined by a linear regression model. In order to validate the confidence of the findings produced through this methodology, a new customer data set should be used to test the robustness of the model.

4.4.4 Test Sample Analysis

The *test* sample defined in Chapter 3 is used to test how the results of a new data set compare with the results of the *train* sample. The same procedures followed for Area Segmentation and Cluster Significance are applied; however, the parameters of the bins are kept constant in order to obtain comparable results – i.e., a grid of equal dimensions is applied to the sample space of the *test* sample. The linear regression models of the train and test samples respectively are compared, as well as the population variables that show significance in the correlation study.

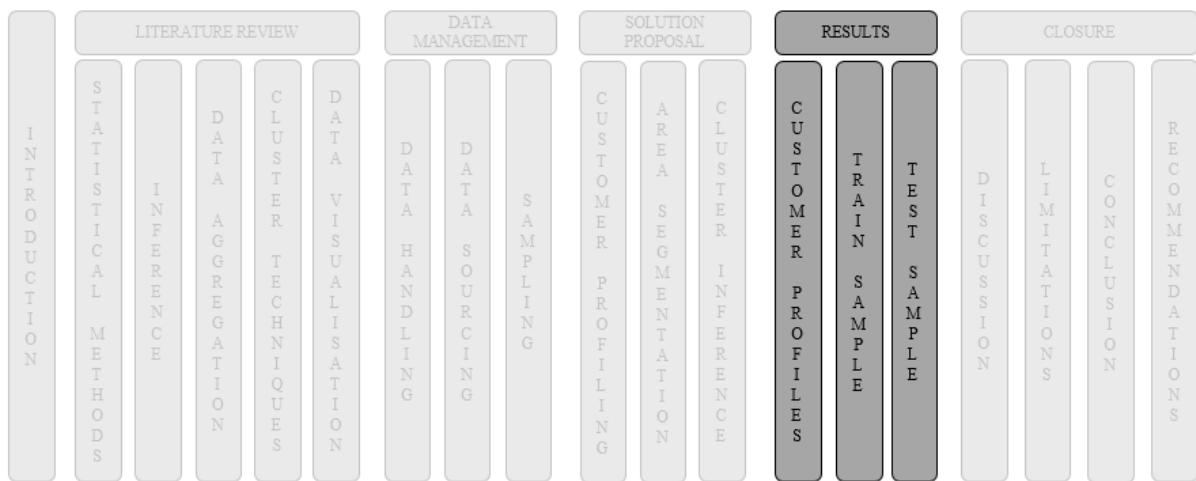
¹⁰ Ratio of customers per bin to population size per bin (for each population variable).

Chapter 5

Results

Chapter Aim:

This aim of this chapter is to present the results of the customer profiles and the customer clusters. The findings are discussed and explained, using a range data of visualisation libraries in R.



Chapter Outcomes:

- Presentation of customer profiles for each municipal district
- Presentation of cluster inference
- Presentation of the *test* sample results

5.1 Customer Profiles

This section presents a descriptive view of the retailer's *average* customer for different areas in Gauteng. The average customer is defined by the four variables that were selected in the methodology, and a profile is shown for the average customer of each municipal district of Gauteng in Table 10 below.

Table 10: Customer profiles for all municipal districts

	Age	Monthly Income	Family Size	Population group	
				Black	White
City of Johannesburg	38.6	R 26 729.00	6	0.57	0.28
City of Tshwane	39	R 16 179.00	5	0.58	0.37
Ekurhuleni	39	R 13 031.00	5	0.59	0.33
Emfuleni	38	R 4 882.00	3	0.71	0.25
Lesedi	41	R 3 204.00	1	0.45	0.5
Midvaal	45.6	R 1 653.00	1	0.33	0.64
Mogale City	39.7	R 3 626.00	2	0.64	0.34
Randfontein	40.3	R 1 747.00	1	0.6	0.27
Westonaria	37.9	R 2 690.00	1	0.69	0.29

The profiles show that there is a near-direct proportional relationship between income and family size for the profiles shown. Figure 17 illustrates this relationship of proportionality where a high average monthly income relates to a larger family size, and likewise for the opposite case.

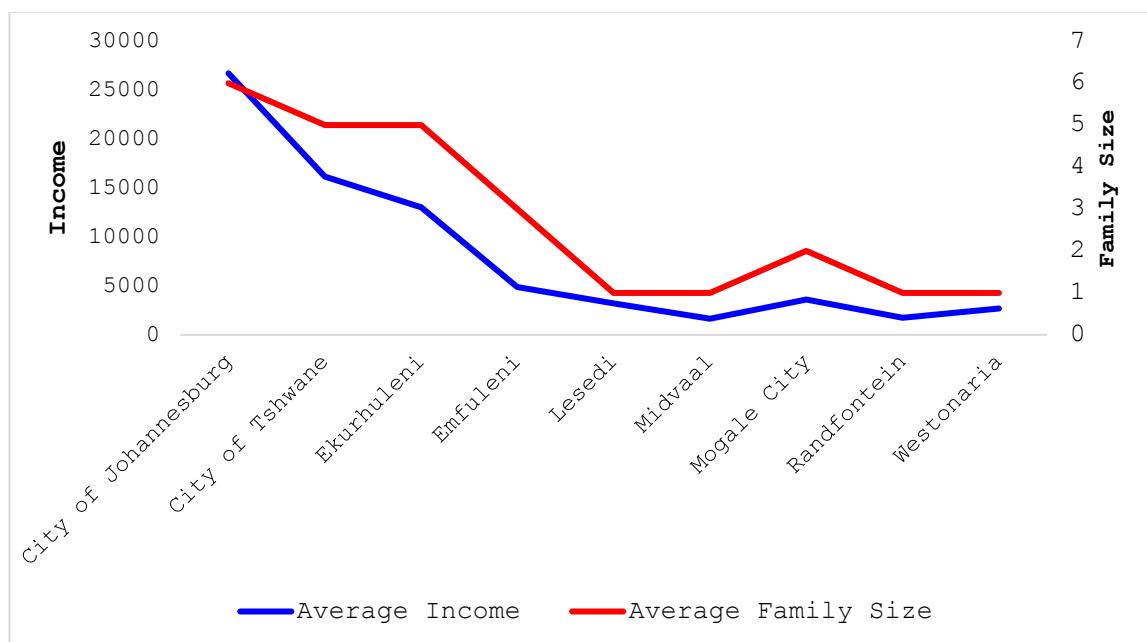


Figure 17: Relationship between average income and family size

Stellenbosch University

Average family size values have been rounded up or down to the nearest whole numerical value, and range from one to six across the various municipal districts. The average age spans a six-year range, and indicates that the average customer in Gauteng is a middle-aged individual. The two primary population groups, *Black* and *White*, are shown in the profiles. (The detailed results, which include the proportions of all population groups for each profile, are shown in Appendix 4.) These profiles enable the retailer to align marketing campaigns, based on the target market characteristics defined in this study. For example, the average customer living in the *City of Johannesburg* is able to purchase either more expensive, better quality furniture, or furniture that caters for a large family, typically with children. A *Randfontein* customer might be interested in smaller products in a more affordable price range, rather than in the kinds of product that appeal to the average customer in the *City of Johannesburg*. The data used in this study exclude product descriptions; but the population group might impact the *style* of furniture that is preferred by each customer profile.

5.2 Customer Cluster Inference

The second section of the results provides insights into customer density, shown by bins in the grid-based clustering method that was applied to the train sample area. Following the segmentation of customer profiles (by municipal district), the same administrative boundaries are used to illustrate customer clusters by various geographical regions of Gauteng. Figure 18 shows the mean variance of the bins determined for each municipal district of Gauteng.

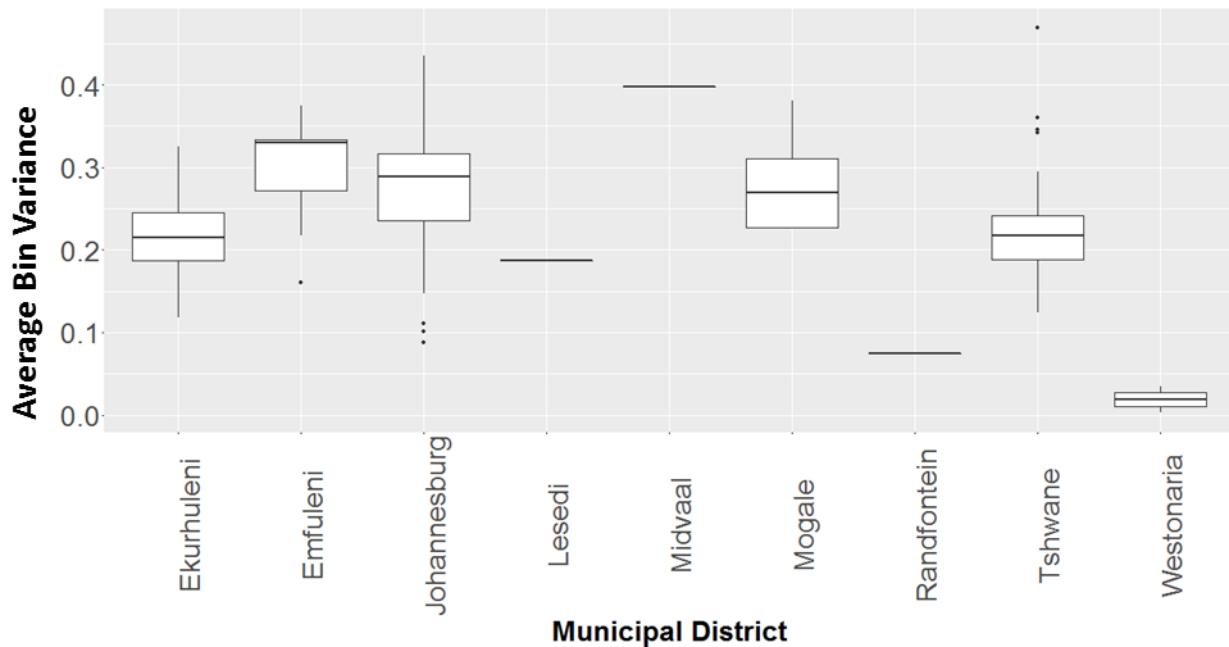


Figure 18: Box plot of bin variances for the municipal districts of Gauteng

Stellenbosch University

The number of bins containing 40 customers or more is shown in Figure 19. The train sample produced 172 bins that contain more than 40 customers, out of the initial 2,585 bins that were created by applying the grid-cluster method.

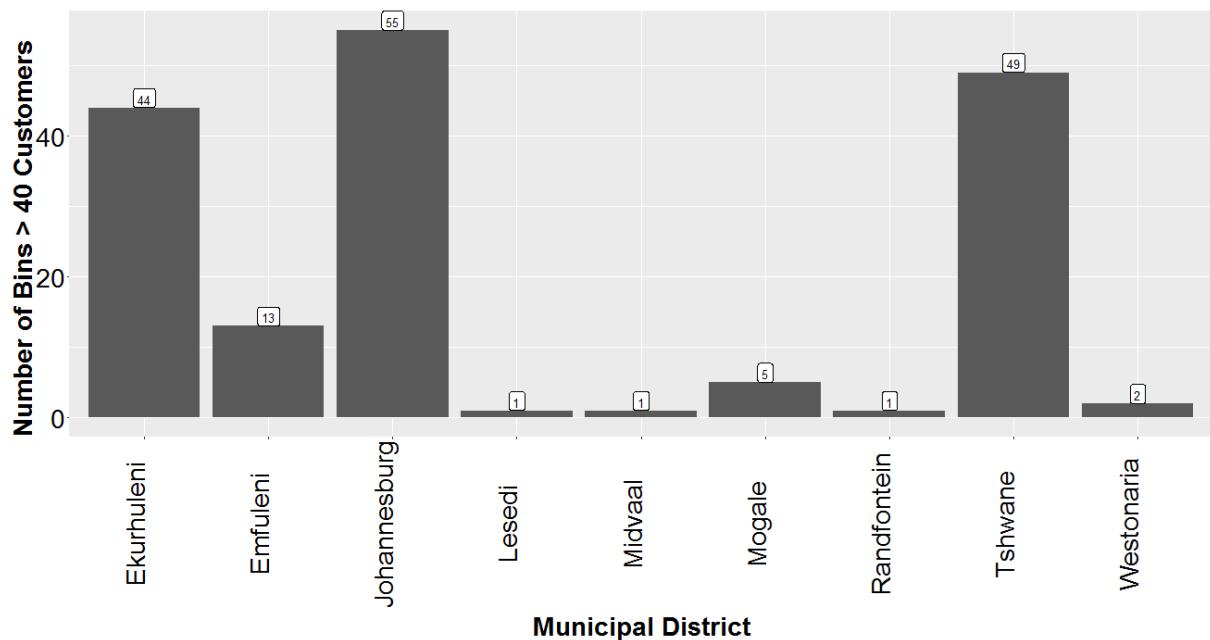


Figure 19: Number of bins (≥ 40 customers) in each municipal region of Gauteng

Figures 18 and 19 show that *City of Tshwane* and *Ekurhuleni* have a lower mean value for bin variance than *City of Johannesburg*, *Emfuleni*, and *Mogale City*. The remaining districts contain a negligible number of bins from which to draw comparisons. Table 11 shows the top 10 bins that exhibit the highest degree of clustering, based on their low variance measures. (The entire list of bins, with their corresponding variance measures and locations, is given in Appendix 3.)

Table 11: Top 10 lowest variance bins

Variance	Customers	Municipal district
0.003	187	Westonaria Local Municipality
0.035	87	Westonaria Local Municipality
0.074	213	Randfontein Local Municipality
0.088	277	City of Johannesburg Metropolitan Municipality
0.101	63	City of Johannesburg Metropolitan Municipality
0.111	233	City of Johannesburg Metropolitan Municipality
0.118	446	Ekurhuleni Metropolitan Municipality
0.124	358	City of Tshwane Metropolitan Municipality
0.127	178	City of Tshwane Metropolitan Municipality
0.133	101	City of Tshwane Metropolitan Municipality

5.2.1 Linear Regression and Correlation Models

This section reveals and discusses the results of the linear regression model that fits the values of the test and train samples respectively, as well as the correlation tests conducted for each sample.

5.2.1.1 Train Sample

The results of the linear regression model for the train sample are shown in Figure 20.

```
Residuals:
    Min      1Q   Median     3Q    Max 
-0.168189 -0.052347  0.000539  0.043788  0.199771 

Coefficients:
            Variable        Estimate Std. Error t-value p-value    
House.or.brick -0.7502080  0.4034276 -1.860  0.0648 .  
Traditional.dwelling -0.0012573  0.0017854 -0.704  0.4823 
Flat.or.apartment. -0.0102622  0.0081621 -1.257  0.2105 
Cluster.in.complex 0.0005301  0.0024523  0.216  0.8291 
Townhouse          0.0010886  0.0034653  0.314  0.7538 
Semi.detached.house 0.0030503  0.0014014  2.177  0.0310 *  
Flatlet.on.a.property -0.0079921  0.0043253 -1.848  0.0665 .  
OtherDwelling       0.0015057  0.0032025  0.470  0.6389 
---| 
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.07249 on 162 degrees of freedom
Multiple R-squared:  0.1303, Adjusted R-squared:  0.08734 
F-statistic: 3.034 on 8 and 162 DF,  p-value: 0.003332
```

Figure 20: Train sample linear regression results

The output shows that, for the eight significant *Dwelling type* variables selected in the PCA, the linear model passes the goodness-of-fit test at the 95% confidence interval (*p*-value < 0.05). Three variables were found to fit the model to a significant degree. These include: *House or Brick*, *Semi-detached house*, and *Flatlet on property*. These three variables show an individual goodness-of-fit for the 95%, 99%, and 95% confidence intervals respectively, account for *most* of the variance in the linear regression model, and are indicative of the most probable dwelling types that describe the retailers' customers' place of residence in the train sample. The variable(s) most responsible for high areas of clusters are shown in the results of the correlation study conducted in Section 4.4.3. The results are shown in Table 12.

Table 12: Correlation results of top 10 lowest variance bins in the train sample

Variables	Correlation coefficient	p - value	t - value
House or brick	-0.615	0.0586	-2.204
Traditional dwelling	-0.857	0.0015	-4.695
Flat or apartment	-0.843	0.0021	-4.441
Cluster in complex	-0.483	0.1573	-1.560

Stellenbosch University

Townhouse	-0.526	0.1177	-1.752
Semi-detached house	-0.208	0.5628	-0.604
Flatlet on property	-0.638	0.0469	-2.346
Other dwelling	-0.590	0.0725	-2.068

Table 12 shows that *Traditional dwellings*, *Flats or apartments*, and *Flatlet on a property* have a significant correlation with low-variance bins at a 95% confidence interval. These three variables are therefore the most probable dwelling types of the retailers' customers *in areas where there is a significant degree of clustering*.

5.3.1.2 Test Sample

The test sample produced 240 bins that contain 40 or more 40 customers. This is somewhat more than those produced by the train sample (172). The variance curve of these bins is shown in Figure 21, in which a number of bins appear to be outliers, as they deviate quite significantly from the S-curve profile. This profile differs somewhat from the S-curve produced by the train sample in Figure 14. The outlying variance values need to be removed in order for the test sample data to fit the linear model (created by the train sample) with some significance.

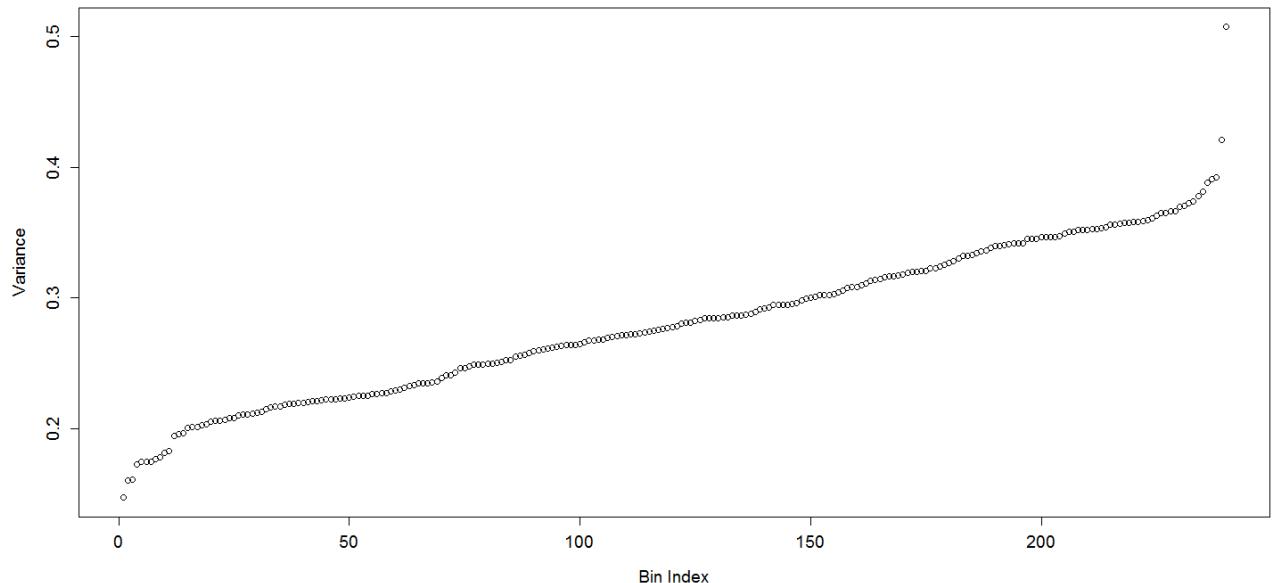


Figure 21: Variance of bins in the test sample data

The results of the linear regression model for the test sample (without removing outliers) are shown in Figure 22.

Stellenbosch University

```

Residuals:
    Min      1Q   Median      3Q     Max 
-0.146582 -0.050413 -0.002528  0.046474  0.227837 

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
House.or.brick          -0.2412498  0.3271248 -0.737   0.4616    
Traditional.dwelling     0.0006597  0.0015984  0.413   0.6802    
Flat.or.apartment        0.0099323  0.0070727  1.404   0.1616    
Cluster.house            -0.0029072  0.0016064 -1.810   0.0716 .  
Townhouse                -0.0003460  0.0016984 -0.204   0.8388    
Semi.detached.house     -0.0007407  0.0008464 -0.875   0.3824    
Room.flatlet.on.a.property -0.0042297  0.0043668 -0.969   0.3338    
Other                     0.0011038  0.0025990  0.425   0.6714    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.05877 on 231 degrees of freedom
Multiple R-squared:  0.03068, Adjusted R-squared:  -0.002894 
F-statistic: 0.9138 on 8 and 231 DF,  p-value: 0.5057

```

Figure 22: Test sample linear regression results

As expected, the test sample data do not fit the linear model ($p > 0.05$). The outliers responsible for the test results are shown in the residuals plot in Figure 23.

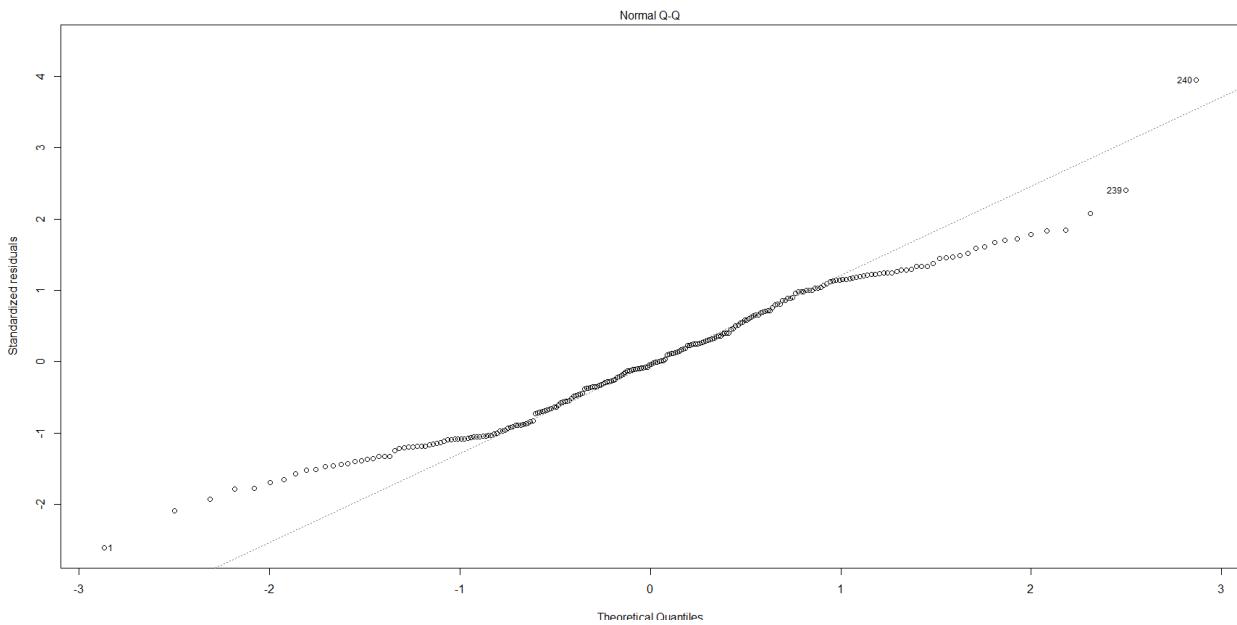


Figure 23: Residuals plot of test sample

The outliers are removed in a step-by-step process until the data fit the linear regression model with a 95% confidence interval. This results in the following nine bins being removed from the test sample:

- Bins 1, 2, 3, 7, 229, 232, 233, 239, and 240.

The results of the improved linear regression model (outliers removed) are shown in Figure 24.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.2904015 0.0065959 44.027 <2e-16 ***
dfRat$House.or.brick -0.0914888 0.2972638 -0.308 0.7585  
dfRat$Traditional.dwelling 0.0001151 0.0014687 0.078 0.9376  
dfRat$Flat.or.apartment 0.0078869 0.0065062 1.212 0.2267  
dfRat$Cluster.house -0.0034391 0.0015018 -2.290 0.0230 *  
dfRat$Townhouse 0.0002988 0.0015632 0.191 0.8486  
dfRat$Semi.detached.house -0.0001784 0.0007716 -0.231 0.8174  
dfRat$Room.flatlet.on.a.property -0.0091719 0.0040162 -2.284 0.0233 *  
dfRat$Other 0.0017313 0.0023854 0.726 0.4687  
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0529 on 222 degrees of freedom
Multiple R-squared: 0.06728, Adjusted R-squared: 0.03367 
F-statistic: 2.002 on 8 and 222 DF, p-value: 0.04735
```

Figure 24: Test sample improved linear regression model results

The *p-value* of this model has drastically decreased from 0.508 to 0.047, rendering a good fit result of the test sample data. Two variables are deemed significant (both at a 99% confidence interval): *Cluster house* and *Flatlet on property*. Only *Flatlet on property* is a common significant variable between the two sample data sets. The results do show, however, that the train sample linear regression model comfortably passes a goodness-of-fit test for a new data sample, given the removal of several outlying bins. Before a correlation study is performed on the train sample data, the linear model should be further enhanced to produce a more accurate fit to the model outputs shown in Section 5.2.1.1. Given the significant outliers shown in Figure 14, additional bins are removed from the head and tail of the S-curve in order to produce a goodness-of-fit that matches the train sample's model output to within a 0.5% tolerance. This results in the removal of the following additional bins from the test sample:

- Bins 4, 5, 6, 8, 202, 214, 221, 223, 225, 227, 228.

Fitting the test sample data with the above outlying bins removed results in a linear regression model that fits the test sample data with a confidence interval that is within a 0.5% tolerance of the confidence interval of the train sample model. A correlation study can now be performed to identify the variables that are responsible for *low* variance bins, and that can be accurately compared with the results of the train sample. The correlation results for the top 10 low-variance bins are shown in Table 13.

Table 13: Correlation results of top 10 lowest-variance bins in the test sample

Variables	Correlation coefficient	p - value	t - value
House or brick	-0.425	0.2197	-1.3315

Stellenbosch University

Traditional dwelling	-0.593	0.0708	-2.0832
Flat or apartment	-0.399	0.2523	-1.2337
Cluster in complex	0.008	0.9826	0.0225
Townhouse	-0.854	0.0018	-4.6518
Semi-detached house	-0.851	0.0017	-4.5885
Flatlet on property	-0.812	0.0044	-3.9143
Other dwelling	-0.1831	0.6124	-0.5271

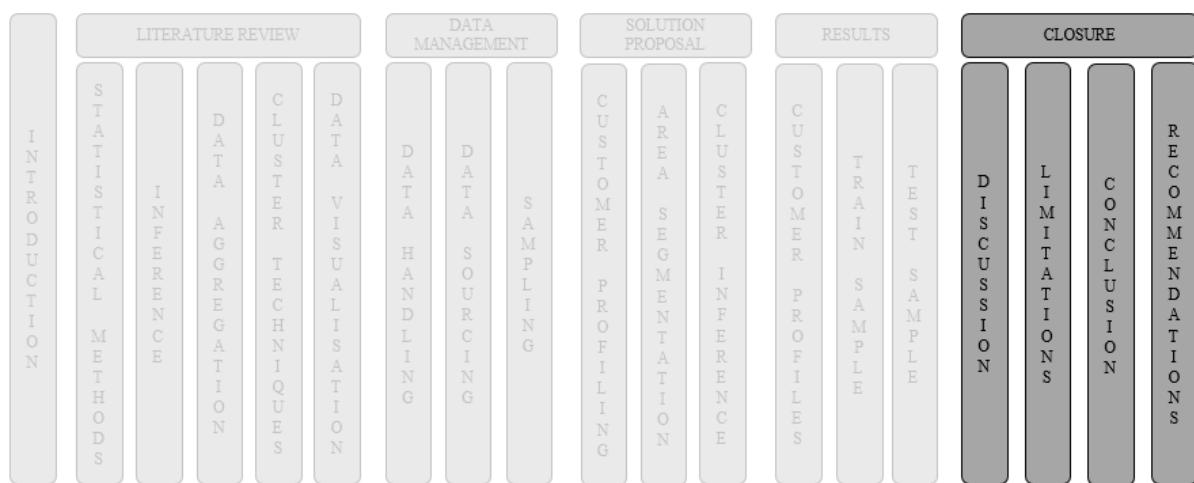
Those variables that show a correlation significance (within a 95% confidence interval) are highlighted in Table 13. *Flatlet on property* and *Traditional dwelling* are common to both train and test data samples as having a strong association with low-variance bins. These results support the answer to the research question that *a customer's dwelling type influences how clustered customers are dispersed*, by showing consistent measures of correlation for several variables across two independent data sets.

Chapter 6

Closure

Chapter Aim:

This chapter discusses the key outcomes, and how the insights produced in this study address the problem statements. In addition to key outcomes, this chapter presents a view on how the findings can be applied to the retailer's environment to add business value.



Chapter Outcomes:

- Discussion of the results produced in the study
- Discussion of the limitations of the study
- Conclusion
- Recommendations

6.1 Discussion

Customer profiles can be derived by overlaying population data with customer address locations and using population attributes contained in specific administrative boundaries. The results produced in this study showed a significant range of values for each of the four population variables used to create customer profiles. Customers who are more likely to purchase costly items and a broader range of goods reside in *City of Johannesburg* and *City of Tshwane*, while regions such as *Midvaal* and *Randfontein* reflect lower-income earning individuals with smaller families, on average. According to the *Local Government Handbook* (2016), the spatial structure of the *Midvaal* municipal area is predominantly that of a rural area, while *City of Johannesburg* boasts a world class infrastructure and is considered the first choice for job seekers across the country. *City of Johannesburg*'s population is predominantly young (*Local Government Handbook*, 2016), which explain why this profile group has the one of the lowest average ages of all nine profiles. The profiles enable *the retailer* to use information about its customers that it did not have before to improve the experience of the customer. This can be done by marketing more appropriate goods to different profiles of customers, based on their age, income, family size, and ethnicity.

The question of *where* to focus marketing campaigns that attract the customer profiles created in this study is answered by the results of the customer cluster locations. This study showed how significant areas of customer clusters can be identified, and exactly where these areas exist geographically, shown by the centre points of bins, which each span an area of 9.84 km². Appendix 3 has a complete list of all the bins, their locations, and their relative measures of variance ranked from lowest to highest. The retailer is armed with a location-based strategy for where their customers are either densely or sparsely gathered. Dense areas of customers are the prime focus for marketing campaigns, given that less marketing collateral is needed to attract a large number of customers. However, 'dense' areas of customers are still only defined by area segments of 9.84 km². In order further to enable specific location-based marketing strategies, certain dwelling types that are prominent in dense (high-clustered) areas have been identified. These include *Traditional dwellings*, *Flats or apartments*, and *Flatlet on a property*. Marketing material should be tailored to attract the profiles of people who reside in these dwelling types. For example, residents staying in a block of flats in a high customer cluster area could be attracted by visual marketing collateral, such as a poster that is placed at the entrance to the premises. Potential customers residing in traditional dwelling types could be attracted by focusing marketing collateral on public transport hubs such as taxi ranks, given that many people who reside in traditional dwellings (such as huts) make use of public transport.

To summarise the results obtained through this study, it is evident that, in order to achieve effective marketing strategies, the intelligence produced from customer profiles should be used in collaboration with the locations of customer clusters and the variables that support the reason for high density clusters – e.g., dwelling types.

6.2 Limitations

This study of customer profiling and cluster identification was subject to several limitations. These are made explicit below. The assumptions about the models constructed in this study are not repeated at this point.

1. The customer information used in this study was limited to geocoded customer delivery addresses for goods purchased over a period of time. This limited the customer profiles to segmentation by geographical location and the use of population data to deduce customer characteristics.
2. The latest population data issued by Statistics South Africa (2016) are for the year 2011. This limits the study to using population data that are updated by means of half-year estimates. However, the most recent South Africa census data set was used in this study.

6.3 Conclusion

The application of descriptive analytics using spatial data has been shown to produce insights that add strategic value to the retailer's business environment. With a specific focus on creating customer profiles and identifying densely-populated customer clusters, this study has discussed how these two objectives could be used to enhance location-based marketing campaigns in a retail sales environment. The use of geocoded customer delivery addresses, together with population data, has revealed the key attributes of customers, based on specific geographical segments within Gauteng, South Africa. The application of proven statistical methods was conducted using R. Several reproducible models were built that enabled data transformations, and the handling of complex data structures and spatial data. The method of grid-based clustering has been developed and applied in a reproducible R script to effectively identify customer clusters based on customer locations. The use of a Principal Component Analysis proved to be an effective method of identifying key variables that account for customer clusters and the good performance of a test sample showed how well the population variables, that were identified in the PCA performed against the original train sample data. This substantiates the validity of the findings and shows how the approach applied in this study can be used to produce accurate results for further investigations.

The study has produced a baseline for customer intelligence that can be further enhanced by overlaying the customer data with more sophisticated customer information such as product (order) details, the value of goods purchased, and detailed delivery information such as dispatch information, order life-cycle information, and the shelf life of products.

6.4 Recommendations

During the research process, some considerations emerged that might be worth investigating further. There are five specific recommendations for future research:

1. Extend the maturity of the analysis by performing a customer segmentation based on more sophisticated customer information. This would require a broader set of customer data that includes product information, delivery schedules, and detailed order information. A customer segmentation study may identify that physical customer location is not the only characteristic to consider when targeting customer sales. Segment sizes will also be dependent on several variables as opposed to only physical area.
2. Develop customer profiles based on psychographic factors that relate more to customer behaviour than to demographic information. This entails developing variables that relate to the behavioural aspects of customers, such as status, loyalty, and spending habits.
3. Perform a cost-benefit study on the effect that the insights produced in this study have on enhancing location-based marketing strategies in a retail sales environment. This would require additional information, such as marketing spend and revenue gained as a result of the campaign effectiveness. Determining the success of a campaign requires the development of quantitative measures to ascertain whether or not a sale can in fact be attributed to a campaign. A cost-benefit study provides valuable input for prioritising which customer clusters should be prioritised for location based marketing campaigns.
4. Extend the methods used in this study to include machine learning logic to identify customer clusters. For example, the use of Self Organising Maps (SOMs) may be used to identify customer clusters by using artificial neural networks. The advantage of using machine learning techniques is that several outputs may be achieved based on the ability of an algorithm to improve on the output achieved in each iteration. In this study, only one iteration of a single clustering method, namely grid-based clustering, was used to produce a set of bins from which clusters were identified.
5. Extend the methods used in this study to include spatial point process analysis and logistic regression in order to identify key relationships between customer locations and external spatial data. For example, customers could be densely populated due to the presence of schools, hospitals, shopping malls etc. By identifying key points of interest and their proximity to customers, more intelligence can be applied to marketing campaigns over above the census data used in this study.

The above recommendations could provide interesting opportunities for future research projects that involve statistical methods and data modelling.

List of References

- Abdi, H. and Williams, L.J. 2010. ‘Principal component analysis’, *WIREs Comp Stat*, 2(4), pp. 433-459.
- Ailawadi, K.L., Neslin, S.A. and Gedenk, K. 2001. ‘Pursuing the Value-Conscious Consumer: Store Brands Versus National Brand Promotions’, *Journal of Marketing*, 65 (1), pp. 71-89.
- Allingham, D. and Rayner, J.C. 2012. ‘Testing equality of variances for multiple univariate normal populations’, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 4-12, 2012, 12. <http://ro.uow.edu.au/cssmwp/91>.
- Apparicio, P., Abdelmajid, M., Riva, M. and Shearmur, R. 2008. ‘Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues’, *International Journal of Health Geographics*, [Online]. 7(1). Available from: <https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-7-7>. [Accessed: 12 June 2016]
- ArcGIS. 2016. ArcGIS Online Help. [Online] Available from: <https://doc.arcgis.com/en/arcgis-online/reference/shapefiles.htm>. [Accessed: 27 June 2016]
- Bartlett, M.S. 1937. ‘Properties of sufficiency and statistical tests’, *Proceedings of the Royal Society of London Series A* 160 (901), pp. 268-282.
- Bawa, K. and Shoemaker, R.W. 1987. ‘The Coupon-Prone Consumer: Some Findings Based on Purchase Behaviour Across Product Classes’, *Journal of Marketing*, 51(4), pp. 99-110.
- Bearden, W.O., Teel, J.E. and Durand, R.M. 1978. ‘Media Usage, Psychographic, and Demographic Dimension of Retail Shoppers’, *Journal of Retailing*, 54(1), pp. 65-74.
- Bewick, V., Cheek, L. and Ball, J. 2003. ‘Statistics review 7: Correlation and regression’, *Critical Care* [Online]. 7(6), pp. 451-459. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC374386/>. [Accessed: 17 May 2016]

Stellenbosch University

Cahusac, P.M.B and De Winter, P. 2014. *Starting out in Statistics: An Introduction for Students of Human Health, Disease, and Psychology*, John Wiley & Sons.

Cai, L. and Zhu, Y. 2015. ‘The Challenges of Data Quality and Data Quality Assessment in the Big Data Era’, *Data Science Journal*, 14(10), p. 2.

Chandler, N., Hostmann, B., Rayner, N. and Herschel, G. 2011. ‘Gartner’s Business Analytics Framework’ (ID: G00219420). Retrieved from Gartner database. [Accessed: 13 March 2015]

Chauhan, R., Kaur, H. and Alam, M. 2010. ‘Data Clustering Method for Discovering Clusters in Spatial Cancer Databases’, *International Journal of Computer Applications* [Online], 10(6), pp. 9-14. Available from: <http://www.ijcaonline.org/volume10/number6/pxc3872004.pdf>. [Accessed: 12 April 2016]

Clark, M. 2009. Principal Component Analysis. Psych. 6810. Class Notes. University of North Texas.

Dahyot, R. 2006. *Finding Eigenvalues and Eigenvectors*. CS1BA1 Mathematics. Lloyd Institute.

Daniel, D. 2007. ‘Five Key Business Intelligence Trends You Need to Know’ [Online]. CIO. Available from: <http://www.cio.com/article/2437743/business-intelligence/five-key-business-intelligence-trends-you-need-to-know.html>. [Accessed: 24 June 2016]

ESRI. 2010. ‘ESRI Shapefile Technical Description: An ESRI White Paper’. 1998 [Online]. Available from: <https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf>. [Accessed: 27 June 2016]

Factor Analysis: A Short Introduction, Part 4. 2014, October 20. ‘The Analysis Factor’. [Web blog post]. Available: <http://www.theanalysisfactor.com/factor-analysis-how-many-factors/>. [Accessed: June 15 2016]

Freedman, D. and Diaconis, P. 1981. ’On the histogram as a density estimator: L₂ theory’, *Probability Theory and Related Fields*, 57 (4), pp. 453-476.

Frost & Sullivan. 2015. Customer Intelligence is the New Black: A Frost and Sullivan White Paper. [Online]. Available from: http://www.sourcingfocus.com/uploaded/documents/Firstsource_-customer_interaction_analytics.pdf. [Accessed: 27 May 2015]

- Garrett-Mayer, E. 2016. Statistics in Psychological Research: Lecture 8 Factor Analysis I. [Online] Johns Hopkins Bloomberg School of Public Health, USA. Available from: <http://ocw.jhsph.edu/courses/statisticspsychosocialresearch/pdfs/lecture8.pdf>. [Accessed: 21 June 2013]
- Jagdale, G.B., Kamoun, S. and Grewal, P.S. 2009. ‘Entomopathogenic nematodes induce components of systemic resistance in plants: Biochemical and molecular evidence’, *Biological Control*, 51(1), pp. 102-109.
- Hamburg, M. 1985. *Basic statistics: A modern approach*, San Diego: Harcourt Brace Jovanovich.
- Han, J. and Kamber, M. 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- He, K. and Meeden, G. 1997. ‘Selecting the number of bins in a histogram: A decision theoretic approach’, *Journal of Statistical Planning and Inference*, 61(1), pp. 59 -59.
- Horn, X. 2016. Machine Learning: Process, Model validation & Feature engineering [Online]. 2 June. Available from: <http://rusers.co/meetups/RUsersXanderHorn>. [Accessed: 1 July 2016]
- Hyndman, R.J. 1995. *The problem with Sturges' rule for constructing histograms. I*. Melbourne, Australia: Monash.
- Jackson, D.A. 1993. ‘Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches’, *Ecology*, 74(8), pp. 2204-2214.
- Johnson, M. 1989. ‘The Application of Geodemographics to Retailing – Meeting the Needs of the Catchment’, *Journal of the Market Research Society*, 31(1), pp. 7-36.
- Kalyanam, K. and Putler, D.S. 1997. ‘Incorporating Demographic Variables in Brand Choice Models: An Indivisible Alternatives Framework’, *Marketing Science*, 16(2), pp. 166-181.
- Kassambara, A. and Mundt, F. ‘factoextra: Extract and Visualise the Results of Multivariate Data Analyses’, <https://CRAN.R-project.org/package=factoextra>, r package version 1.0.3, 2016.

Stellenbosch University

Lane, D.M., Hebl, M., Guerra, R., Osherson, D., Scott, D. and Zimmer, H. 2015. ‘Online Statistics Education: A Multimedia Course of Study’. Rice University [Online]. Available from: http://onlinestatbook.com/Online_Statistics_Education.pdf/. [Accessed: 12 June 2016]

Lê, S., Josse, J. and Husson, F. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* [Online]. 25(1), pp. 1-18. Available from: <https://cran.r-project.org/web/packages/FactoMineR/vignettes/FactoMineR.pdf>. [Accessed: 12 May 2016]

Loecher, M. and Ropkins, K. 2015. ‘RgoogleMaps and loa: Unleashing R Graphics Power on Map Tiles’, *Journal of Statistical Software*, 63(4), pp. 1-18.

Manikandan, S. 2010. ‘Data transformation’, *Journal of Pharmacol & Pharmacother* [Online]. 1(2), pp. 126-127. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3043340/#CIT1>, [Accessed: 21 June 2016]

Manoukian, E.B., Maurais, J. and Ouimet, R. 1986. ‘Exact Critical Values of Bartlett’s test of Homogeneity of Variances of Unequal Samples Sizes for Two Populations and Power of the Test’, *Mertika* [Online]. 33(1), pp. 275-289. Available from: <http://link.springer.com/article/10.1007/BF01894757>. [Accessed: 2 June 2016]

Montgomery, D.C. and Runger, G.C. 2007. *Applied Statistics and Probability for Engineers*, 4th ed. USA: Wiley.

Mulhern, F.J. and Williams, J.D. 1994. ‘A comparative analysis of shopping behaviour in Hispanic and non-Hispanic market areas’, *Journal of Retailing*, 70(3), pp. 231-251.

Municipal Demarcation Board. 2016. Boundary Data [Online]. Available from: <http://www.demarcation.org.za/index.php/downloads/boundary-data/boundary-data-main-files/wards/11457-wards/file>. [Accessed: 2 June 2016]

Murray, A.T. 1998. ‘Assessing Clustering Methods for Exploratory Spatial Data Analysis’ [Online]. Available from: <http://www-sre.wu.ac.at/ersa/ersaconsf/ersa98/papers/346.pdf>. [Accessed: 5 October 2015]

New York University. 2013. ‘What is data science?’ [Online]. Available from: <http://datascience.nyu.edu/what-is-data-science/>. [Accessed: 14 June 2016]

Stellenbosch University

- Oracle. 2010. ‘An Oracle White Paper: Value of Spatial Analytics in Business Intelligence’. [Online]. Available from: <http://www.oracle.com/technetwork/middleware/bi-foundation/value-of-spatial-analytics-in-bi-ag-1-130184.pdf>. [Accessed: 14 October 2015]
- OriginLab. 2016. ‘Interpreting Results of Principal Component Analysis’ [Online]. Available from: <http://www.originlab.com/doc/Origin-Help/PCA-Result>. [Accessed: 1 July 2016]
- Palma, O. 2011. ‘Statistical Analysis & Dissemination of Census Data’. UNSD-CELADE Regional Workshop on Census Cartography for the 2010 Latin America’s census round. United Nations Statistics Division [Online]. Available from: <http://www.cepal.org/celade/noticias/paginas/8/35368/pdfs/8unsd.pdf>. [Accessed: 2 February 2015]
- Pebesma, E.J. and Bivand, R.S. 2005. ‘Classes and methods for spatial data in R’, *R News*, 5(2), pp. 9-13. Available: <http://cran.r-project.org/doc/Rnews/>. [Accessed: 3 June 2015]
- Peres-Neto, P., Jackson, D.A. and Somers, K.M. 2003. ‘Giving Meaningful Interpretation to Ordination Axes: Assessing Loading Significance in Principal Component Analysis’, *Ecology* [Online], 84(9), pp. 2347-2363. Available from: <http://publications.chestnet.org/pdfAccess.ashx?url=%2Fdata%2FJournals%2FCHEST%2F22098%2F110523.pdf>. [Accessed: 6 July 2016]
- Peres-Neto, P., Jackson, D.A. and Somers, K.M. 2005. ‘Stopping rules for determining the number of non-trivial axes revisited’, *Computational Statistics & Data Analysis*, 49(4), pp. 974-997.
- R Core Team. 2016. ‘R: A language and environment for statistical computing’. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. [Accessed: 14 April 2015]
- Rahn, M. 2012. ‘The Analysis Factor’ [Online]. Available from: <http://www.theanalysisfactor.com/factor-analysis-1-introduction/>. [Accessed: 29 June 2016]
- Rencher, A.C. 2002. *Methods of Multivariate Analysis*, 2nd ed. Canada: John Wiley & Sons, Inc.
- Richman, M.B. 1988. ‘A cautionary note concerning a commonly applied eigen analysis procedure’, *Tellus*, 40(B), pp. 50-58.

Stellenbosch University

Scott, D.W. 1979. ‘On optimal and data-based histograms’, *Biometrika*, 66(3), pp. 605-610.

Scott-Street, W. 2003. Levene’s Test for Equality of Variances [Online]. Virginia Commonwealth University, USA. 1 December. Available from:
http://www.people.vcu.edu/~wsstreet/courses/314_20033/Handout.Levene.pdf. [Accessed: 21 June 2016]

Shi, N. and Tao, J. 2008. *Statistical Hypothesis Testing Theory and Methods*, 1st ed. Singapore: World Scientific Publishing.

Sleight, P. 1995. ‘Neighbourhood watch: Geodemographic and lifestyle data in the UK GIS marketplace’, *Mapping Awareness*, 9(6), pp. 18-21.

Squarespace. 2016. What is JSON? [Online]. Available from:
<http://developers.squarespace.com/what-is-json/>. [Accessed: 3 June 2016].

Statistics South Africa. ‘Living Conditions Survey, 2011’ [dataset].

Streiner, D. and Norman, G.R. 2011. ‘Correction for Multiple Testing’, *CHEST* [Online], 140(1), pp. 16-18. Available from:
<http://publications.chestnet.org/pdfAccess.ashx?url=%2Fdata%2FJournals%2FCHEST%2F22098%2F110523.pdf>. [Accessed: 6 July 2016]

Sturges, H.A. 1926. ‘The choice of a class interval’, *Journal of the American Statistical Association*, 21(1), pp. 65-66.

Suhr, D.D. 2012. *Exploratory Factor Analysis with the World Values Survey*. SAS Global Forum. University of Northern Colorado.

The Local Government Handbook. 2016. ‘A complete guide to the municipalities in South Africa’ [Online]. Available from: <http://www.localgovernment.co.za/metropolitans/view/2/City-of-Johannesburg-Metropolitan-Municipality>. [Accessed: 1 August 2016]

Theodorou, V., Abelló, A. and Lehner, W. 2014. ‘Quality Measures for ETL Processes’. In *Proceedings of the 16th International Conference on Data Warehousing and Knowledge Discovery*, DaWaK 2014, edited by Bellatreche, L. and Mohania, M.K., pp. 9-22. Lecture Notes in Computer Science, number 8646. Munich, Germany: Springer.

Stellenbosch University

University of Toronto. 2002. ‘Data Clustering Techniques’. [Online] Available from:
<http://www.cs.toronto.edu/~periklis/pubs/depth.pdf>. [Accessed: 2 June 2016]

Vassiliadis, P. 2009. ‘A Survey of Extract–Transform– Load Technology’, *International Journal of Data Warehousing & Mining* [Online], 5(3), pp. 1-27. Available from:
https://cs.brown.edu/courses/cs227/papers/IJDWM_2009.pdf. [Accessed: 10 August 2016]

York University Libraries. 2016. ‘Geospatial Data’. [Online] Available from:
<http://researchguides.library.yorku.ca/content.php?pid=245987&sid=2176375>. [Accessed: 1 July 2016]

Appendix 1

Data Cleaning Script

```
#The purpose of this script is to read in raw data and get it onto a useful format
require (gdata)
require(stringr)
require(dplyr)

#Set the required wd
setwd("D:/Users/Michael.Brink/Desktop/Masters/Masters/EHL")

#Set root directories accordingly
#get the data into simplified format i.e. lats, longs and associated zip code
raw <- read.csv ("D:/Users/Michael.Brink/Desktop/Masters/Masters/EHL/Mike.csv")
# MasterCodes <-
read.csv("D:/Users/Michael.Brink/Desktop/Masters/R_Workspace/Distances/zipGeos.csv")
# Towns_imp <- read.csv("D:/Users/Michael.Brink/Desktop/Masters/Masters/EHL/Towns_master.csv")

raw <- raw %>% distinct(X2012.06.22.10.43.17) #remove multiple POD msgs on exactly the same
time

#rename the raw file columns headings
names (raw)[1] = 'Zip'
names (raw)[2] = 'Date'
names (raw)[3] = 'Lat'
names (raw)[4] = 'Long'

Sample <- raw
#Sample <- raw

#Remove zeros/faults in the dataset
Sample <- Sample[Sample[, 3]<0, ]
#Sample <- Sample[Sample$Zip != 0, ]
Sample <- Sample %>% distinct(Lat)

#Now, ensure that each of the sample falls within Gauteng
require(rgeos)
require(rgdal)
require(maptools)

#1st validation test: do all points occur within the boundary of Gauteng?
border <- readOGR(dsn = "D:/Users/Michael.Brink/Desktop/Masters/External/Provinces", layer =
"SOU-level_1" )
plot(border[border$ID == 'Gauteng',])
mike <- data.frame(border@data)
poly2 <- data.frame()

#Create a df to store associated ward boundaries with a unique unique ID that is assigned in
b_test5$unID
tempfile <- border[border$ID == 'Gauteng', ]
poly2 <- as.data.frame(tempfile@polygons[[1]]@Polygons[[1]]@coords)

store <- vector() #Create a vector in which to store the UnID - identifier (note we note
storing the unID value)
checkVec <- vector()

#for (i in 1:nrow(Sample)) {
fpounds <- function(i){
  polypoint.x.y <- Sample[i,c(3, 4)]
  zeroTest <- point.in.polygon(polypoint.x.y[, 1],polypoint.x.y[, 2], poly2[, 2], poly2[, 1],
mode.checked = FALSE)
  var21 <- ifelse('0' %in% zeroTest, 0, 1)
  checkVec[i] <- var21
}

K <- unlist(lapply(1:nrow(Sample), fpounds))
Sample$check <- K
Sample1 <- Sample[Sample$check != 0,]

#For choropleth Map
# chlor <- merge(Towns_imp, Sample0, by = "City.Town")
# chlor <- data.frame(chlor[, c(10,11)])
# names(chlor)[2] = "id"
# GT_chlor <- chlor[, c(1, 3, 10)]
```

Stellenbosch University

```

# GT_chlor <- GT_chlor[which(!duplicated(GT_chlor[,2])),]
# names(GT_chlor)[2] = "Zip"
# GT_chlor$Zip<-str_pad(GT_chlor$Zip, 4, pad = "0")
# write.csv(GT_chlor,"C:/Users/mbrink/Desktop/R_Workspace/Distances/GT_Chlor.csv" )

# -----
Sample1$Date <- as.Date(Sample1$Date)
min(Sample1$Date)
max(Sample1$Date)

#check delivery frequencies for this period

SalesFreq <- table(cut(Sample1$Date, breaks="month"))
plot(SalesFreq, type="l", xlab=" ", ylab=" ")

Sample <- Sample1

#Select Sample 1 as: 1 Jan 2013 - 30 June 2013
#Select Sample 2 as: 1 July 2013 - 31 Dec 2013

# ====== Subset x2 samples ======
#Select a time period for this Data:

Sample1 <- Sample[Sample$Date > "2013-01-01",]
Sample1 <- Sample1[Sample$Date < "2013-06-30",]

Sample2 <- Sample[Sample$Date > "2013-07-01",]
Sample2 <- Sample2[Sample$Date < "2013-12-31",]

# ===== END =====

#write the 'sample' file to use in GeoMatrix
Sample_GeoMat1 <- data.frame("lat" = Sample1$Lat, "lng" = Sample1$Long)
Sample_GeoMat2 <- data.frame("lat" = Sample2$Lat, "lng" = Sample2$Long)

#write sample 1
write.csv
(Sample_GeoMat1,"D:/Users/Michael.Brink/Desktop/Masters/R_Workspace/R_excelexamples/sample1.csv"
")
#write sample 2
write.csv
(Sample_GeoMat2,"D:/Users/Michael.Brink/Desktop/Masters/R_Workspace/R_excelexamples/sample2.csv"
")

```

Grid-Clustering Script

```

setwd("D:/Users/Michael.Brink/Desktop/Masters/R_Workspace/R_excelexamples")
require(ash)
require(ggmap)
require(tidyr)
require(geosphere)
require(grDevices)
require(ggplot2)

#Read in the data source that was saved in the shrub.EHL.R file
# Choose between Sample, sample2 or sample 3 here:
data <- read.csv
("D:/Users/Michael.Brink/Desktop/Masters/R_Workspace/R_excelexamples/sample2.csv")

#Order: Long and Lat
data<- data[,c(3, 2)] #*#
#Select only the good values, remove the #NAs
good <- data[complete.cases(data),]
allGeo = gather(good, ColumnName)
x <- matrix (allGeo[,2], nrow(good) , 2)

#Max/min coordinates of sample area
maxLat <- max(good[, 2])
maxLng <- max(good[, 1])
minLat <- min(good[, 2])
minLng <- min(good[, 1])

```

Stellenbosch University

```

#Calculate optimal bin size:
x. <- good[, 1]
y.<- good[, 2]

hist(x.,breaks="scott") #

x2. <-nclass.scott(x.)#
y2. <-nclass.scott(y.)#

dfbins <- data.frame(scott=c(x2., y2.))
rownames(dfbins)[1] = "x-Lat"
rownames(dfbins)[2] = "y-long"
View(dfbins)#

#Results = x: 55, y: 47
#Now use these results to set the bin specs for sample 2:

dfbins[1, 1] <- 55 #grid dims based on Scott's formula calculated in sample1
dfbins[2, 1] <- 47

#-----
#CONFIGURE ab:
ab <- matrix (c(minLng,minLat,maxLng,maxLat), 2, 2) # *#
nbin <- c(dfbins[2, 1], dfbins[1, 1]) # bi * bi bins
bins <- bin2(x,ab, nbin) # bin counts,ab,nskip
class(bins)
#----- Contour Plot -----
m <- c(2,2)
f <- ash2(bins,m) #m:(input integer vector of length 2) x and y direction smoothing
parameters. Default is 5 by 5
image(f$x,f$y, f$z)
#plot(f$x, f$y, f$z)

#Create a raster
require (raster)
require(grDevices)
est.raster = raster(list(x=f$x,y=f$y,z=f$z))
plot(est.raster, col=colorRampPalette(c("blue", "green", "yellow", "orange", "red"))(250),
main = 'Raster plot of bins')
contour(est.raster,add=TRUE)

#----- Contour Plot -----

#here we assign the mesh grids values per block to a dataframe called "mike"
var33 <- bins[[1]] #*#
var33 <- as.data.frame(var33) #*#
var33 <- as.data.frame(t(var33)[ncol(var33):1,]) #*#

#Now i need a formula to test into which block each of the geocodes in 'codes' falls into and
allocate some reference
require(stringr)
#read in the unique, complete sample list created in scrubEHL.R ... CompleteList_4618.csv
#Codes <-
read.csv("D:/Users/Michael.Brink/Desktop/Masters/R_Workspace/Distances/CompleteList_4618.csv")
Codes <- good
#Codes$Zip<-str_pad(Codes$Zip, 4, pad = "0")
#Codes <- Codes[-1]

#assign variables to lat and long number of bins
latbin <- dfbins[1, 1]
longbin <- dfbins[2, 1]

head(Codes)

#Latitude referencing
f2 <- function (x= x:nrow(Codes)){
  Codes$LatRef[x] <- latbin +1 - which(abs(f[[3]]- Codes[x,2]) ==min (abs(f[[3]]- Codes[x,2])))
  #return(Codes$LatRef)
}

k <- vector("list", nrow(Codes))
Codes$LatRef <- 1: nrow(Codes)
#Codes$LatRef <- unlist(k) #New code line to fix the break
k <- lapply(1:nrow(Codes), f2)
Codes$LatRef <- unlist(k)
head(Codes)

```

Stellenbosch University

```

#Longitude referencing
#Write a function that references all coorindate points
f3 <- function (x = x:nrow(Codes)){
  Codes$LngRef[x] <- which(abs(f[[2]] - Codes[x,1]) == min(abs(f[[2]] - Codes[x,1])))
  #return(Codes$LngRef)
}
g <- vector("list", nrow(Codes))
Codes$LngRef <- 1: nrow(Codes) #New code line to fix the break
#Codes$LngRef <- unlist(g) #New code line to fix the break
g <- lapply(1:nrow(Codes), f3)
Codes$LngRef <- unlist(g)

#Combine lat/long refs to create a new variable 'uniqueRef'
Codes$Index <- 1: nrow(Codes)
Codes$UniqueRef <- paste(Codes$LatRef,"-", Codes$LngRef)

head(Codes)
# ----- END -----
#FUNCTION L1
#i.e. the purpose of the following functions is to count customer addresses for each bin &
populate a df, 'var33' with the count values

for (i in 1:ncol(var33)){
  names(var33)[i] <- i
}

for (i in 1:nrow(var33)){
  row.names(var33)[i] <- i
}

sko <- data.frame()

#Here I assign a lat and long for the centre point of each square
#Now this WORKS! (21 July 2015)

#Function L2
j <- 1
i <- 1

for (i in 1:ncol(var33)){
  for (j in 1:nrow(var33)){
    sko[j, i] <- paste(f[[3]][latbin + 1 -j], ',', f[[2]][i]) #f[[2]] = lat and
f[[3]] = long      sko[j, i] <- paste(f[[3]][longbin-j], ',', f[[2]][i]) #f[[2]] = lat and
f[[3]] = long

    }
} #f[[3]][[latbin-j]] ... check
#-----
#test the results
#f[[2]][[1]] #F returns the co-ords of the centrepoint of the bin
#f[[3]][[1]] #F[[2]] = Longitude ; F[[3]] = Latitude
#-----
var33Long <- gather(var33)

var33Long$Latref <- seq(1,latbin, 1)
var33Long$Lngref <- rep(1:longbin, each = latbin)
var33Long$UniqueRef <- paste(var33Long$Latref,"-", var33Long$Lngref)

#This code will now sum the number of geocodes (column = value) per unique ref
library (plyr)
#aggregate(var33Long$value, by=list(var33Long$UniqueRef), "sum")
#-----
skoLong <- gather(sko)
firstRun <- cbind(var33Long, skoLong)

Use <- firstRun[, c(2, 7)]

names(Use) [1] <- "codesQty"
names(Use) [2] <- "geocode"

#Use <- Use[order(-Use[, 1]),]

require(splitstackshape)
Use1 <- data.frame(cSplit(Use, 'geocode', sep=" ,", type.convert=FALSE))
names(Use1) [2] = "Lat"

```

Stellenbosch University

```

names(Use1)[3] = "Lng"

Use1[, 2] <- as.numeric(Use1[, 2])
Use1[, 3] <- as.numeric(Use1[, 3])

#test...
Use1[1, 2] + Use1[1, 2]

#Data frame now used to link Bins and Zip code data by the customer Geocodes:
Use1$BinIndex <- 1:nrow(Use1)
GT_bins <- cbind(Use1, var33Long)
head(GT_bins)
GT_bins <- GT_bins[, c(2, 3, 4, 9)]
names(GT_bins)[1] = "BinLat"
names(GT_bins)[2] = "BinLng"
GT <- merge(Codes, GT_bins, by = "UniqueRef")
head(GT)
GT <- GT[-c(4,5)]

View(GT)
#the following are all output files to be used in the distances.R script that follows
write.csv (GT,"D:/Users/Michael.Brink/Desktop/Masters/R_Worksplace/Distances/GTsampel2.csv")
write.csv (Use1,"D:/Users/Michael.Brink/Desktop/Masters/R_Worksplace/Distances/Usesampel2.csv")
write.csv
(var33,"D:/Users/Michael.Brink/Desktop/Masters/R_Worksplace/Distances/CountMatrixsample2.csv")
write.csv (sko,"D:/Users/Michael.Brink/Desktop/Masters/R_Worksplace/Distances/skosampel2.csv")
# ----- END -----

```

Customer Profiling Script

```

#This script is to associate the number of people per ward to census data per wards and and
thereby find the average customer per ward and per district

df <- read.csv ("D:/Users/Michael.Brink/Desktop/Masters/31Aug/CensusData/Averages.csv") #read
in population data
df$UnID <- 2907:3414 #Rename the ward IDs to match those in the original shapefile
names(df)[2] = 'unID'#Rename unID coloumn to match that in the original shapefile as per
sampleWards.R script

#Count the number of people per ward:
#First, call the sampleWards workspace that contains the GT table with all customer-ward
associations
Customers <- data.frame(table(GT$unID))
names(Customers)[1] = 'unID'
#From this we see that only 471 wards even contain customers at all!

#Merge the count of customers per ward to the ward/census data df:
dfmrg <- merge(df, Customers, x.all =TRUE, by = 'unID')
names(dfmrg)[11] <- 'Customers'

#Link the Municipal district to each ward (unID)
districts <- WardsDF[WardsDF$PROVINCE == 'Gauteng', c(3, 10)]
dfmrg <- merge(dfmrg, districts, x.all =T, by = 'unID')

#Calculate total custoemrs for each district
tmp <- data.frame(ddply(dfmrg,~MUNICNAME,summarise,ttl=sum(Customers)))
dfmrg <- merge(dfmrg, tmp, x.all =T, by = 'MUNICNAME')

#Now calculate the proportion of population data for each variable per district
dfmrg$PFamSize <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$Average.Family.Size
dfmrg$PIncome <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$Average.Monthly.Income
dfmrg$Age <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$Average.Adult.Age
dfmrg$Black <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$Black.African
dfmrg$White <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$White
dfmrg$Indian <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$Indian.or.Asian
dfmrg$Coloured <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$Coloured
dfmrg$Other <- (dfmrg$Customers /dfmrg$ttl)*dfmrg$Other
dfmrg$Race <-

#Summarise the proportions by district aggregation
fam <- data.frame(ddply(dfmrg,~MUNICNAME,summarise,ttl=sum(PFamSize)))
income <- data.frame(ddply(dfmrg,~MUNICNAME,summarise,ttl=sum(PIncome)))
age <- data.frame(ddply(dfmrg,~MUNICNAME,summarise,ttl=sum(Age)))
blc <- data.frame(ddply(dfmrg,~MUNICNAME,summarise,ttl=sum(Black)))
white <- data.frame(ddply(dfmrg,~MUNICNAME,summarise,ttl=sum(White)))
indian <- data.frame(ddply(dfmrg,~MUNICNAME,summarise,ttl=sum(Indian)))

```

Stellenbosch University

```

coloured <- data.frame(ddply(dfmrg, ~MUNICNAME, summarise, ttl=sum(Coloured)))
other <- data.frame(ddply(dfmrg, ~MUNICNAME, summarise, ttl=sum(Other)))

cor.test(fam[, 2], income[, 2])

```

Bin Variance Script

```

require(gdata)
require(stringr)
require(geosphere)
library(dplyr)

#Read in data sources created in either scrubEHL.R or GeoMatrixv1.1.R or external data

GT <- read.csv ("D:/Users/Michael.Brink/Desktop/Masters/R_Workspace/Distances/GTsample1.csv")
#Codes <- Codes # Created as shown in GeoMatrixv1.1_sample2

means <- vector()

tstlat <- sapply(split(GT$lat, GT$BinIndex), mean)
tstlon <- sapply(split(GT$lng, GT$BinIndex), mean)

dfM <- data.frame(tstlat, tstlon)
dfM$BinIndex <- rownames(dfM)
dfM <- dfM[, c(3, 1, 2)]

GT <- merge(GT, dfM, by = "BinIndex")

#Now calculate the distances from each bin to the mean coordinates calculated in dfM above
#distance formula:

splitBinslat <- split(GT$lat, GT$BinIndex)
splitBinslon <- split(GT$lng, GT$BinIndex)

#Get number of addresses in each bin in the list
n <- unlist(lapply(splitBinslat, function(x) length(x)))
class(n)

#test
#splitBinslat[[430]][[1]]
#length(splitBinslat[[430]])

#Create a list of unique bin median coordinates
binMeans <- GT[, c(3, 9, 10)]
binMeans <- binMeans %>% distinct(UniqueRef)

distVec <- vector()
varVec <- vector()
countObs <- vector()
store <- matrix()

R <- 6371 #Earth circum

for (i in 1: length(splitBinslat)){
  for (j in 1: length(splitBinslat[[i]])){
    dcalc <- function(long1, lat1, long2, lat2) {
      d <- acos(sin(lat2*pi/180)*sin(lat1*pi/180) + cos(lat2*pi/180)*cos(lat1*pi/180) * 
        cos(long2*pi/180-long1*pi/180)) * R
    }

    d2 <- dcalc(splitBinslon[[i]][[j]], splitBinslat[[i]][[j]], binMeans$tstlon[i],
    binMeans$tstlat[i])
    distVec[j] <- d2
    #dcalc(Codes[1:nrow(Codes), 2], Codes[1:nrow(Codes), 1], Bubbles[i, 3], Bubbles[i, 2])
  }
  varVec[i] <- var(distVec)
  # countObs[i] <- length(distVec)
}

varVec #returns the variance for each bin in the sample!
max(varVec, na.rm = TRUE)
min(varVec, na.rm = TRUE)

binMeans$Variance <- varVec
binMeans$n = n

```

Stellenbosch University

```

binMeans40 <- binMeans[binMeans$n > 40,]

#dfZip <- GT[, c(3, 8)]
#dfZip <- dfZip %>% distinct(UniqueRef)

# _____ BARTLETT TEST _____

#Create distance parameters for Bartlett test
GT$dist <- acos(sin(GT$tstlat*pi/180)*sin(GT$lat*pi/180) +
cos(GT$tstlat*pi/180)*cos(GT$lat*pi/180) * cos(GT$tstlon*pi/180-GT$lng*pi/180)) * R

intBart <- merge(int, GT, by = "UniqueRef")
intBart <- intBart[, c(1, 4, 5, 7, 8, 18)]

table(Test1$BinIndex)

Test1 <- intBart[intBart$Municipality == "Ehurhuleni", ]
#Test2 <- int[int$Municipality == "Ehurhuleni", ] #does not work, require distance (granular)
#data

bartlett.test(dist ~ BinIndex, data = Test1 )
#bartlett.test(Variance ~ UniqueRef, data = Test2 ) #does not work, require distance
#(granular) data

#perform Levene's Test on the data:
library (car)

data <- merge(GT, binMeans, X.ALL= T, by = 'UniqueRef')
data <- data[data$n >= 40, ]
data <- data[, c(1, 2, 11)]

leveneTest(dist ~ UniqueRef, data=data)

#Perform some graphs and plots for the results:
tmp <- merge(GT, WardsDF[, c(3, 10)], by = 'unID')
tmp$MUNICNAME <- as.character(tmp$MUNICNAME)
tmp$unID <- as.numeric(tmp$unID)

require(dplyr)
tmp2 <- tmp %>% group_by(UniqueRef) %>% mutate(n = n())
tmp2 <- tmp2[tmp2$n >= 40, ]
tmp2 <- tmp2 %>% distinct(UniqueRef)
tmp3 <- data.frame(table(tmp2$MUNICNAME))
names(tmp3)[1] = 'DistrictMunicipalies'

tmp3$lab <- as.character(tmp3$Freq)

dev.off()

require(ggplot2)
ggplot(data=tmp3, aes(x=DistrictMunicipalies, y=Freq, fill = DistrictMunicipalies, label =
lab, las = 2)) +
  geom_bar(stat="identity")+
  geom_text(size=6)+
  theme(text = element_text(size=20),axis.text.x=element_text(angle=90,
size=20),axis.text.y=element_text(size=20))+
  ggtitle("Bins in Gauteng containing more than or equal to 40 customers")+
  coord_fixed(ratio = 0.2)+ 
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())

plot (temptop$Variance)

#Now, create a table showing top 10 variance bins
tmp4 <- merge(temptop[, c(1, 2, 3)], tmp2[, c( 4, 13)], by = 'UniqueRef')

k <- unique(tmp4$MUNICNAME)
munCode <- c("Tshwane", "Johannesburg", "Ekurhuleni", "Mogale", "Randfontein", "Westonaria",
"Emfuleni", "Lesedi", "Midvaal")
munID <- data.frame(MUNICNAME = k, MUNCode = munCode)
intNew <- merge(tmp4, munID, by = 'MUNICNAME')

dev.off()

```

Stellenbosch University

```

means <- aggregate(Variance ~ MUNCode, intNew, mean)
means$Variance <- round(means$Variance, digits = 2)

ggplot(data=intNew, aes(x=MUNCode, y=Variance, fill=MUNCode, show.legend = FALSE)) +
  geom_boxplot() +
  stat_summary(fun.y=mean, colour="darkred", geom="point", shape=18, size=3, show.legend =
  FALSE) +
  geom_text(data = means, aes(label = Variance, y = Variance + 0.016), size = 6, show.legend =
  FALSE) +
  xlab("Municipal Districts") +
  ylab("Variance") +
  theme(text = element_text(size=20), axis.text.x=element_text(angle=45,
  size=20), axis.text.y=element_text(size=20))

```

Bin/Ward Matching Script

```

#First create a table showing total population of customers per ward
wardCust <- data.frame(table(GT$unID))
names(wardCust)[1] = 'unID'

#Next, create a table showing bin/ward relationships this will show how many bins per ward
wardBin <- GT[, c(1, 12)]
wardBin$unWB <- paste(wardBin$BinIndex, wardBin$unID)
wardBin <- data.frame(table(wardBin$unWB))

require (stringr)
#now add the ward ID into a seperate column in the wardBin df
wardBin$unID <- str_sub(wardBin$Var1,-4,-1)
#rename column headings
names(wardBin)[1] = 'binWard'
names(wardBin)[2] = 'binWardFreq'
#Now merge the 2 dfs to get totol customer per ward
wardBin <- merge(wardBin, wardCust, by = 'unID')
names(wardBin)[4] = 'WardFreq'
#calculate the % of ward customers in each bin
wardBin$Percentage <- wardBin$binWardFreq/wardBin$WardFreq
#Add Bin unique ID back
wardBin$BinIndex <- str_sub(wardBin$binWard,1,-5)

#Check how many (customers within...) wards fall within a bin:
wardperBin <- data.frame (table(wardBin$BinIndex))

Census <- read.csv("D:/Users/Michael.Brink/My
Documents/Sales/GautengSegmentation/Census2.csv")
names(Census)[1] = 'WARD_ID'

mike <- as.data.frame(WardsRSA)
temp <- mike[, c(5, 10)]
wardBin <- merge(wardBin, temp, by = 'unID')
temp <- merge(wardBin, Census, by = 'WARD_ID')
tempRe <- temp[, c(8:ncol(temp))]
tempRe <- data.frame(apply(tempRe, 2, function(x) as.numeric(as.character(x)))) 

#Loop parameters
temp2 <- data.frame()
b <- temp[, 6]
-----
for (i in 1:ncol(tempRe)){
  for (j in 1:nrow(tempRe)){
    a <- tempRe[, i]
    temp2[j, i] <- a[j] * b[j]
  }
}
names(temp2) = names(temp[, c(8:ncol(temp))])
temp2$BinIndex <- temp[, 7]
temp2$unID <- temp[, 1]
temp2 <- temp2[,c(43, 42, 1:42)] #up until here everything is still legit!

idx <- split(1:nrow(temp2), temp2$BinIndex)
a2 <- data.frame()
head(idx)
tail(idx)
require(plyr)
pmet <- data.frame(ddply(temp2,"BinIndex",numcolwise(sum)))

```

Stellenbosch University

```

#Calculate Bin parameters based on the data projected from the ward

require(rgdal)
require(maptools)
unzip("Wards.zip")
WardsRSA <- readOGR(dsn = "D:/Users/Michael.Brink/Desktop/Masters/Zipped Files/Wards", layer =
"Wards2011" )
plot(WardsRSA)

#1st Shape File
WardsRSA <- readOGR(dsn = "D:/Users/Michael.Brink/Desktop/Masters/Zipped Files/SOU-
8_boundaries_SHP", layer = "SOU-8_boundaries" )
par(mar = rep(2, 4))
plot(WardsRSA)
WardsDF <- data.frame(WardsRSA@data)
#2nd SF
# WardsRSA <- readOGR(dsn = "D:/Users/Michael.Brink/Desktop/Masters/Zipped Files/SOU-
8_admin_SHP", layer = "SOU-8" )
# par(mar = rep(2, 4))
# plot(WardsRSA)
# WardsDF <- data.frame(WardsRSA@data)

#use this df to reference ward IDs
WardsDF <- data.frame(WardsRSA@data)
poly2 <- data.frame()
WardsRSA <- data.frame(WardsRSA@data)

#levels(as.character(WardsDF$PROVINCE == 'Gauteng', 3)))

#Now, create a loop to construct a dataframe for boundary coordinates only of required wards.
#Use the Var@data[[4]] access code to find the ward ID/ or use the ward description

#assign a unique id to track wards in Gauteng Only
WardsDF$unID <- 1:4277 #Check the shapefile contents to see how many wards in Gauteng = 4277
WardsRSA$unID <- 1:4277 #Check the shapefile contents to see how many wards in Gauteng = 4277

#id the range of the unique identifier for Gauteng:
unmax <- max(WardsDF[WardsDF$PROVINCE == 'Gauteng', 10])
unmin <- min(WardsDF[WardsDF$PROVINCE == 'Gauteng', 10])

#Create a df to store associated ward boundries with a unique unique ID that i assigned in
WardsRSA$unID
for (i in unmin:unmax) {
  tempfile <- WardsRSA[WardsRSA$unID == i, ]
  poly1 <- as.data.frame(tempfile@polygons[[1]]@Polygons[[1]]@coords)
  poly1$ward <- tempfile@data[[10]][1] #the [[10]] selects the index in the #data fields that
  points to the ward ID
  poly2 <- rbind(poly2, poly1)
  #poly1 <- data.frame()
}

poly1$ward <- tempfile@data[[10]] #the [[10]] selects the index in the #data fields that
points to the ward_ID

#We now have adf showing ward boundaries and corresponding unique identifier to match any
fields in the original poly1@data df
splitWardsLat <- split(poly2$V2, poly2$ward) #split ward coords for referencing purposes in
the point.in.poly function
splitWardsLng <- split(poly2$V1, poly2$ward)

#Check
#1. splitWardsLat[[508]]
#2. length(splitWardsLat)

store <- vector() #Create a vector in which to store the UnID - identifier (note we note
storing the unID value)

checkVec <- vector()
store <- vector()

for (i in 1:nrow(GT)) {

  for (j in 1 :length(splitWardsLat) ){


```

Stellenbosch University

```

polypoint.x.y <- GT[i,c(5, 4)]
zeroTest <- point.in.polygon(polypoint.x.y[, 1],polypoint.x.y[, 2], splitWardsLat[[j]],
splitWardsLng[[j]], mode.checked = FALSE)
king <- ifelse('0' %in% zeroTest, 0, 1)
checkVec[j] <- king
wardSp <- which(checkVec==1)
}
if(1 %in% checkVec)
{
  store[i] <- wardSp
  print (store[i])
} else{
  store[i] <- 0
}
}

system.time(
  K <- unlist(lapply(1 :nrow(GT),polypoint )))

#Now assign actual ward unID to each customer address in the GT df
GT$ward <- unmin + store -1
GT$ward <- gsub(unmin -1, 'NA', GT$ward)
names(GT)[12] = 'unID'

```

Linear Regression Script

```

#Merge with another df to include frequency or variance or both of bins...
require(dplyr)
#first run of this model will only include bins with > 40 observations
pmet$BinIndex <- as.numeric(pmet$BinIndex)
temp <- GT[, c(1, 3)]
temp4 <- merge(pmet, temp, by = 'BinIndex')
temp4 <- temp4 %>% distinct(BinIndex)
temp4 <- temp4[ ,c(43, 1, 2:42)]
temp <- binMeans[, c(1, 4, 5)]
temp4 <- merge(temp4, temp, by = 'UniqueRef')
temp4 <- temp4[ ,c(45, 44, 1, 2:43)]

temp <- temp4[complete.cases(temp4$Variance),] # this gives us the bins woth > 40 obs and
hence have a valid variance value
temp <- temp[temp$n > 40, ]

#top 50 lowest variance bins -test 1
temptop <- temp[order(temp$Variance),]
#temptop <- temptop[temptop$n > 150, ]

#Final multi-variate linear model:
test <- temptop[1:171, c(1,2,4,10:13, 14, 15, 19, 21)]
#test2 <- temptop[151:170, c(1, 12, 13, 14)]

lmFin <- lm(test$Variance ~ test[, 4]+test[, 5]+test[, 6]+test[, 7]+test[, 8]+test[, 9]+test[, 10]+test[, 11] )

dfRat <- matrix(0, ncol = ncol(test)-3, nrow = nrow(test))
dfRat <- data.frame(dfRat)
names(dfRat) <- names(test)[4:ncol(test])

#Using ratios
for (i in 1:(ncol(test)-3)){
  dfRat[, i] <- test$n/test[,i+3]
  #print(i+2)
}
dfRat <- data.frame(apply(dfRat, 2, function(x) as.numeric(as.character(x))))
dfRat <- data.frame(do.call(data.frame,lapply(dfRat, function(x) replace(x, is.na(x),0)))))

#Now redo the lm
lmFinRat <- lm(test$Variance ~
dfRat$House.or.brick.concrete.block.structure.on.a.separate.stand.or.yard.or.on.a.farm +
dfRat$Traditional.dwelling.hut.structure.made.of.traditional.materials +
dfRat$Flat.or.apartment.in.a.block.of.flats + dfRat$Cluster.house.in.complex +
dfRat$Townhouse..semi.detached.house.in.a.complex. +
dfRat$Semi.detached.house+

```

Stellenbosch University

```

dfRat$Room.flatlet.on.a.property.or.larger.dwelling.servants.quarters.granny.flat +
dfRat$OtherDwelling )
summary(lmFinRat)
plot(lmFinRat)

#other lm
summary(lmFin)
plot(lmFin)

View(test)

for (i in 1:8){
  print(cor.test(test[1:10, 2], dfRat[1:10, i]))
}

for (i in 1:8){
  print(cor.test(test[162:171, 2], dfRat[162:171, i]))
}

cor.test(test$Variance[1:10], test$ratio[1:10])
cor.test(test$Variance[161:170], test$ratio[161:170])

#1st Var test
cor.test(test$Variance[1:10], test$wtf[1:10])
cor.test(test$Variance[161:170], test$wtf[161:170])

#2nd Var test
cor.test(test$Variance[1:10], test$cluster[1:10])
cor.test(test$Variance[161:170], test$cluster[161:170])

#3rd Var test
cor.test(test$Variance[1:10], test$town[1:10])
cor.test(test$Variance[161:170], test$town[161:170])

#4th : Brick
cor.test(test$Variance[1:10], test$brick [1:10])
cor.test(test$Variance[161:170], test$brick [161:170])

#5th : Traditional
cor.test(test$Variance[1:10], test$strad [1:10])
cor.test(test$Variance[161:170], test$strad [161:170])

```

Principle Component Analysis

```

#install.packages('FactoMineR')
#install.packages('factoextra')

require(FactoMineR)
require(factoextra)

#Scrub script
df<- temptop
df<- df[, c(1, 2, 5:ncol(df))]
#df <- df[, colSums(is.na(df)) < nrow(df)]
df[is.na(df)] <- 0
df <- data.frame(apply(df, 2, function(x) as.numeric(as.character(x)))

dfRat <- matrix(0, ncol = ncol(df)-2, nrow = nrow(df))
dfRat <- data.frame(dfRat)
names(dfRat) <- names(df)[3:ncol(df)]

for (i in 1:(ncol(df)-2)){
  dfRat[, i] <- df$n/df[,i+2]
  #print(i+2)
}

dfRat <- data.frame(do.call(data.frame,lapply(dfRat, function(x) replace(x,
is.infinite(x),0)))) 

mydata <- temptop[, 10:21] #My data
mydata <- dfRat[, 6:ncol(dfRat)] #My data
mydata <- dfRat[, 1:5] #My data
mydata <- dfRat #My data
mydata[is.na(mydata)] <- 0
rownames(mydata) <- 1:nrow(dfRat)

```

Stellenbosch University

```

decathlon2.active <- mydata
class(res.pca)

res.pca <- PCA(decathlon2.active, graph = FALSE)
print(res.pca)

#Get eigenvalues and relative explained variance from the EVs
eigenvalues <- res.pca$eig
head(eigenvalues[, 1:2])

#Scree plot
dev.off()
fviz_screepplot(res.pca, ncp=12) +
  theme(text = element_text(size=18), axis.text.x=element_text(size=18),
axis.text.y=element_text(size=18))
head(res.pca$var$coord)

names <- data.frame(names = names(dfRat))

#Variables factor map
fviz_pca_var(res.pca) #Variables factor map

head(res.pca$var$cos2) #Shows split of loading per variables (across PC1 ... PC10)
loading <- data.frame(res.pca$var$cos2)

fviz_pca_var(res.pca, col.var="cos2") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=0.5) + theme_minimal()

#Variables that are correlated with PC1 and PC2 are the most important in explaining the
variability in the data set.
#The contribution of variables can be extracted as follow :
head(res.pca$var$contrib)

#The larger the value of the contribution, the more the variable contributes to the component.
#The most important variables associated with a given PC can be visualized, using the function
fviz_pca_contrib()[factoextra package], as follow :
#(factoextra >= 1.0.1 is required)
dev.off()
fviz_pca_contrib(res.pca, choice = "var", axes = 1) #The red dashed line on the graph above
indicates the expected average contribution. For a given component, a variable with a
contribution larger than this cutoff could be considered as important in contributing to the
component.

# Contributions of variables on PC1
fviz_pca_contrib(res.pca, choice = "var", axes = 2)

# Total contribution on PC1 and PC2
fviz_pca_contrib(res.pca, choice = "var", axes = 1:2)

#The total contribution of a variable, on explaining the variations retained by PC1 and PC2, is
calculated as follow : (C1 * Eig1) + (C2 * Eig2)
fviz_pca_contrib(res.pca, choice = "var", axes = 1, top = 7)

# Control variable colors using their contributions
fviz_pca_var(res.pca, col.var="contrib")

# Change the gradient color
fviz_pca_var(res.pca, col.var="contrib") +
  scale_color_gradient2(low="white", mid="blue",
                        high="red", midpoint=50) + theme_minimal()

#This is helpful to highlight the most important variables in explaining the variations
retained by the principal components.
#Visual inspection of variable contributions on PCs is nice. But, How to extract the most
significantly associated variables with a given principal component?

#Dimension description
#The function dimdesc() [in FactoMineR] can be used to identify the most correlated variables
with a given principal component.
#A simplified format is :

dimdesc(res.pca, axes = 1:3, proba = 0.05)

#where

```

Stellenbosch University

```
#res : an object of class PCA
#axes : a numeric vector specifying the dimensions to be described
#prob : the significance level

res.desc <- dimdesc(res.pca, axes = c(1,2))
# Description of dimension 1
res.desc$Dim.1
res.desc$Dim.2

#Graph of individuals
#The coordinates of the individuals on the principal components are :
head(res.pca$ind$coord)
fviz_pca_ind(res.pca)

#Cos2 : quality of the representation for individuals on the principal components
#The squared cosine shows the importance of a component for a given observation.

head(res.pca$ind$cos2)
```

Appendix 2**Census data by electoral ward**

<i>Electoral Ward</i>	<i>Average Family Size</i>	<i>Average Monthly Income</i>	<i>Average Adult Age</i>	<i>Proportion of Population Group</i>				
				Black African	Coloured	Indian or Asian	White	Other
74201001: Ward 1	2.57	R 11 154.33	46.65	18%	1%	1%	79%	1%
74201002: Ward 2	2.80	R 1 811.85	37.52	99%	0%	0%	0%	0%
74201003: Ward 3	2.98	R 4 057.10	37.99	79%	1%	0%	20%	0%
74201004: Ward 4	2.20	R 9 715.91	36.21	49%	1%	1%	48%	1%
74201005: Ward 5	2.69	R 12 311.37	44.89	13%	1%	2%	83%	1%
74201006: Ward 6	4.20	R 1 829.32	35.83	99%	0%	0%	0%	0%
74201007: Ward 7	3.19	R 2 201.25	38.41	88%	0%	0%	11%	0%
74201008: Ward 8	2.21	R 1 864.82	37.33	94%	0%	0%	5%	0%
74201009: Ward 9	2.87	R 7 265.94	36.49	47%	1%	1%	50%	1%
74201010: Ward 10	2.64	R 6 636.92	39.06	44%	2%	1%	51%	2%
74201011: Ward 11	3.31	R 2 837.11	36.99	89%	1%	0%	9%	0%
74201012: Ward 12	1.85	R 1 131.94	40.93	99%	0%	0%	0%	0%
74201013: Ward 13	2.49	R 1 450.37	40.37	99%	0%	0%	0%	0%
74201014: Ward 14	1.98	R 1 039.14	39.03	100%	0%	0%	0%	0%
74201015: Ward 15	3.24	R 5 296.31	36.40	69%	3%	1%	25%	2%
74201016: Ward 16	2.46	R 5 768.64	40.21	45%	23%	3%	28%	1%
74201017: Ward 17	1.71	R 1 064.85	35.40	99%	0%	0%	0%	0%
74201018: Ward 18	2.54	R 929.67	37.23	99%	0%	0%	0%	0%
74201019: Ward 19	2.90	R 1 340.55	37.47	99%	0%	0%	0%	0%
74201020: Ward 20	2.89	R 2 009.78	38.65	99%	0%	0%	0%	0%
74201021: Ward 21	3.18	R 2 864.77	39.06	74%	1%	24%	1%	1%
74201022: Ward 22	2.95	R 1 272.70	37.19	99%	0%	0%	0%	0%
74201023: Ward 23	2.47	R 2 245.32	37.65	91%	1%	0%	8%	0%
74201024: Ward 24	1.98	R 780.70	37.22	99%	0%	0%	0%	0%
74201025: Ward 25	3.28	R 3 315.75	39.65	74%	1%	0%	25%	0%
74201026: Ward 26	4.49	R 1 769.78	36.97	99%	0%	0%	0%	0%
74201027: Ward 27	1.89	R 718.41	38.19	99%	0%	0%	0%	0%
74201028: Ward 28	4.23	R 2 169.06	37.22	99%	0%	0%	1%	0%
74201029: Ward 29	2.70	R 962.59	38.30	99%	0%	0%	0%	1%
74201030: Ward 30	2.03	R 1 085.78	39.84	100%	0%	0%	0%	0%
74201031: Ward 31	1.92	R 2 341.11	39.25	100%	0%	0%	0%	0%
74201032: Ward 32	1.94	R 946.12	41.14	99%	0%	0%	0%	0%
74201033: Ward 33	2.64	R 1 334.11	39.72	99%	0%	0%	0%	0%
74201034: Ward 34	1.96	R 1 137.48	40.00	99%	0%	0%	0%	0%
74201035: Ward 35	1.87	R 1 000.64	40.58	99%	0%	0%	0%	0%
74201036: Ward 36	3.73	R 2 834.91	37.28	99%	1%	0%	0%	1%
74201037: Ward 37	1.95	R 917.27	40.72	99%	0%	0%	0%	0%
74201038: Ward 38	2.54	R 1 781.57	38.84	99%	0%	0%	0%	0%
74201039: Ward 39	2.25	R 937.59	38.33	96%	3%	0%	0%	0%
74201040: Ward 40	2.12	R 1 247.90	40.12	99%	0%	0%	0%	0%
74201041: Ward 41	2.21	R 1 598.26	38.44	99%	0%	0%	0%	0%
74201042: Ward 42	1.48	R 702.91	38.91	99%	1%	0%	0%	0%
74201043: Ward 43	2.84	R 1 377.27	37.27	99%	1%	0%	0%	0%
74201044: Ward 44	3.17	R 1 162.60	37.47	99%	0%	0%	0%	0%
74201045: Ward 45	2.07	R 5 382.34	42.89	37%	3%	1%	58%	0%
74202001: Ward 1	1.12	R 1 581.46	40.22	67%	1%	0%	31%	1%
74202002: Ward 2	0.79	R 3 126.60	44.12	17%	1%	1%	81%	1%
74202003: Ward 3	0.86	R 2 414.10	43.01	14%	1%	1%	84%	0%
74202004: Ward 4	1.15	R 3 910.28	42.73	47%	2%	1%	49%	0%
74202005: Ward 5	1.57	R 3 850.21	41.20	50%	2%	1%	47%	1%
74202006: Ward 6	1.59	R 492.06	36.71	99%	1%	0%	0%	0%
74202007: Ward 7	1.02	R 2 184.07	40.00	61%	3%	1%	35%	1%
74202008: Ward 8	1.46	R 753.63	34.26	98%	1%	0%	1%	0%
74202009: Ward 9	0.11	R 337.30	51.73	13%	2%	1%	83%	0%
74202010: Ward 10	1.43	R 1 132.45	35.68	95%	1%	0%	3%	0%
74202011: Ward 11	1.06	R 1 902.30	40.56	66%	6%	3%	25%	1%
74202012: Ward 12	0.55	R 1 073.42	39.12	67%	1%	2%	29%	0%
74202013: Ward 13	1.10	R 2 061.19	42.03	32%	1%	0%	67%	0%
74202014: Ward 14	0.96	R 4 161.25	44.25	5%	0%	0%	94%	1%

Stellenbosch University

74203001: Ward 1	0.92	R	950.64	40.10	89%	1%	0%	10%	0%
74203002: Ward 2	1.36	R	558.12	36.58	99%	0%	0%	0%	0%
74203003: Ward 3	1.16	R	1 049.02	38.77	99%	0%	0%	0%	0%
74203004: Ward 4	1.13	R	766.22	38.88	99%	0%	0%	0%	0%
74203005: Ward 5	1.16	R	544.14	37.50	99%	1%	0%	0%	0%
74203006: Ward 6	0.94	R	1 003.75	39.45	80%	1%	0%	19%	0%
74203007: Ward 7	1.52	R	716.40	34.20	97%	1%	0%	0%	1%
74203008: Ward 8	1.44	R	3 707.88	43.82	30%	2%	8%	58%	2%
74203009: Ward 9	0.98	R	4 515.21	43.15	17%	3%	3%	76%	0%
74203010: Ward 10	1.11	R	3 446.74	41.08	54%	1%	1%	43%	0%
74203011: Ward 11	1.04	R	731.07	36.61	90%	1%	0%	8%	1%
74203012: Ward 12	1.34	R	1 631.44	39.39	65%	2%	1%	32%	1%
74203013: Ward 13	1.45	R	624.33	37.30	97%	0%	1%	2%	0%
74801001: Ward 1	2.23	R	821.27	35.00	99%	0%	0%	0%	1%
74801002: Ward 2	2.66	R	916.93	35.17	99%	0%	0%	0%	0%
74801003: Ward 3	1.99	R	2 474.90	37.86	49%	1%	49%	0%	1%
74801004: Ward 4	2.17	R	1 111.09	36.39	100%	0%	0%	0%	0%
74801005: Ward 5	1.32	R	644.12	35.92	99%	0%	0%	0%	0%
74801006: Ward 6	1.21	R	1 224.27	39.30	99%	0%	0%	0%	0%
74801007: Ward 7	1.61	R	2 621.84	37.64	99%	0%	0%	0%	0%
74801008: Ward 8	1.82	R	1 958.92	37.34	99%	0%	0%	0%	0%
74801009: Ward 9	1.07	R	2 599.63	40.73	61%	2%	1%	36%	0%
74801010: Ward 10	1.44	R	817.83	37.80	100%	0%	0%	0%	0%
74801011: Ward 11	2.21	R	1 905.99	38.20	99%	0%	0%	0%	0%
74801012: Ward 12	0.97	R	635.16	39.55	100%	0%	0%	0%	0%
74801013: Ward 13	1.66	R	1 765.63	36.32	99%	0%	0%	0%	0%
74801014: Ward 14	1.83	R	2 555.79	37.45	76%	1%	0%	23%	0%
74801015: Ward 15	1.28	R	848.44	37.53	99%	0%	0%	0%	0%
74801016: Ward 16	1.87	R	2 244.13	36.61	74%	2%	1%	23%	1%
74801017: Ward 17	1.34	R	5 228.14	45.22	17%	1%	0%	80%	1%
74801018: Ward 18	1.53	R	6 808.75	43.94	13%	1%	1%	84%	1%
74801019: Ward 19	1.09	R	990.74	39.50	99%	0%	0%	0%	1%
74801020: Ward 20	1.91	R	3 857.86	40.18	40%	2%	1%	57%	1%
74801021: Ward 21	1.30	R	7 544.17	46.25	15%	1%	1%	82%	1%
74801022: Ward 22	1.64	R	8 863.52	45.13	14%	1%	1%	82%	1%
74801023: Ward 23	1.78	R	2 284.86	35.17	87%	1%	0%	12%	1%
74801024: Ward 24	2.08	R	1 208.86	37.35	99%	1%	0%	0%	0%
74801025: Ward 25	1.15	R	680.15	36.48	97%	1%	0%	2%	0%
74801026: Ward 26	2.01	R	3 597.37	35.06	51%	1%	1%	46%	0%
74801027: Ward 27	2.19	R	1 396.96	36.12	88%	1%	0%	11%	0%
74801028: Ward 28	1.52	R	12 819.52	41.87	20%	2%	2%	75%	1%
74801029: Ward 29	1.21	R	5 522.82	41.38	34%	1%	1%	62%	2%
74801030: Ward 30	2.61	R	2 013.93	36.76	84%	1%	0%	14%	0%
74801031: Ward 31	1.61	R	1 810.01	38.87	83%	1%	1%	15%	1%
74801032: Ward 32	1.08	R	1 267.32	37.78	87%	1%	0%	11%	0%
74801033: Ward 33	1.21	R	2 286.51	37.99	77%	1%	0%	21%	1%
74801034: Ward 34	2.50	R	819.11	34.83	99%	0%	0%	0%	0%
74802001: Ward 1	1.54	R	1 462.85	37.91	80%	1%	0%	19%	0%
74802002: Ward 2	1.90	R	2 620.66	39.25	69%	3%	1%	26%	1%
74802003: Ward 3	0.89	R	1 715.88	41.43	39%	2%	1%	57%	1%
74802004: Ward 4	0.53	R	1 235.38	42.13	32%	3%	1%	64%	0%
74802005: Ward 5	0.70	R	1 434.41	40.43	63%	10%	2%	23%	3%
74802006: Ward 6	1.05	R	3 384.65	41.81	26%	3%	1%	70%	1%
74802007: Ward 7	1.25	R	4 413.39	43.14	20%	2%	1%	77%	0%
74802008: Ward 8	1.29	R	1 664.52	39.62	54%	34%	1%	10%	1%
74802009: Ward 9	1.35	R	4 399.42	40.39	45%	3%	1%	50%	0%
74802010: Ward 10	0.68	R	426.11	38.79	26%	74%	0%	0%	0%
74802011: Ward 11	0.65	R	800.11	37.64	89%	1%	0%	10%	0%
74802012: Ward 12	1.10	R	331.02	35.61	99%	1%	0%	0%	0%
74802013: Ward 13	0.89	R	1 063.47	37.79	98%	1%	1%	1%	0%
74802014: Ward 14	1.05	R	612.72	37.13	99%	1%	0%	0%	0%
74802015: Ward 15	1.12	R	503.82	36.85	98%	1%	0%	0%	0%
74802016: Ward 16	1.26	R	1 064.51	38.88	99%	0%	0%	0%	0%
74802017: Ward 17	0.68	R	519.38	39.71	99%	0%	0%	0%	1%
74802018: Ward 18	1.19	R	804.00	40.09	100%	0%	0%	0%	0%
74802019: Ward 19	0.77	R	435.34	38.03	98%	2%	0%	0%	0%
74802020: Ward 20	1.47	R	1 187.34	37.90	32%	66%	0%	1%	1%
74802021: Ward 21	0.65	R	421.24	40.68	99%	0%	0%	1%	0%
74802022: Ward 22	0.75	R	528.21	38.04	99%	1%	0%	0%	0%
74803001: Ward 1	1.06	R	1 548.90	37.04	88%	0%	0%	10%	1%

Stellenbosch University

74803002: Ward 2	1.23	R	3 166.01	36.45	84%	1%	0%	15%	0%
74803003: Ward 3	0.07	R	873.99	40.58	100%	0%	0%	0%	0%
74803004: Ward 4	0.80	R	1 480.38	37.33	73%	1%	1%	24%	0%
74803005: Ward 5	0.75	R	1 421.53	38.94	71%	1%	0%	28%	0%
74803006: Ward 6	1.22	R	3 483.82	38.58	61%	1%	1%	37%	0%
74803007: Ward 7	1.42	R	714.36	35.67	99%	1%	0%	0%	0%
74803008: Ward 8	1.65	R	624.70	35.87	99%	1%	0%	0%	0%
74803009: Ward 9	1.07	R	276.58	34.31	99%	0%	0%	0%	0%
74803010: Ward 10	1.32	R	562.29	34.95	99%	0%	0%	0%	1%
74803011: Ward 11	0.85	R	241.94	34.19	99%	0%	0%	0%	1%
74803012: Ward 12	1.34	R	946.90	36.91	99%	0%	0%	0%	1%
74803013: Ward 13	1.05	R	609.33	36.67	96%	2%	0%	0%	2%
74803014: Ward 14	1.22	R	373.39	34.00	96%	0%	0%	0%	3%
74803015: Ward 15	1.13	R	528.89	35.63	99%	0%	0%	0%	0%
74803016: Ward 16	1.44	R	1 038.73	36.70	97%	1%	0%	1%	0%
74804001: Ward 1	1.73	R	1 191.90	37.06	92%	1%	0%	7%	0%
74804002: Ward 2	1.31	R	556.71	36.53	99%	1%	0%	0%	0%
74804003: Ward 3	1.09	R	325.70	35.88	99%	1%	0%	0%	0%
74804004: Ward 4	1.54	R	465.36	35.52	99%	0%	0%	0%	0%
74804005: Ward 5	0.45	R	566.55	36.66	88%	0%	0%	12%	0%
74804006: Ward 6	0.95	R	497.58	37.62	99%	0%	0%	0%	0%
74804007: Ward 7	1.25	R	885.36	38.77	99%	0%	0%	0%	0%
74804008: Ward 8	1.03	R	380.47	37.56	99%	1%	0%	0%	0%
74804009: Ward 9	1.04	R	373.99	37.64	100%	0%	0%	0%	0%
74804010: Ward 10	1.11	R	550.92	39.85	99%	0%	0%	0%	0%
74804011: Ward 11	0.98	R	1 603.04	39.13	95%	0%	0%	4%	0%
74804012: Ward 12	0.78	R	1 186.58	41.75	53%	1%	0%	45%	1%
74804013: Ward 13	0.90	R	1 174.66	37.18	97%	1%	0%	2%	0%
74804014: Ward 14	1.25	R	2 419.77	39.79	89%	1%	0%	10%	0%
74804015: Ward 15	0.81	R	1 349.72	35.07	86%	1%	0%	13%	0%
74804016: Ward 16	1.15	R	2 477.29	39.36	68%	1%	1%	29%	0%
74804017: Ward 17	0.90	R	2 417.86	42.75	40%	1%	1%	58%	0%
74804018: Ward 18	0.92	R	1 869.56	37.42	73%	1%	2%	23%	1%
74804019: Ward 19	0.00	R	-	41.34	99%	0%	0%	1%	0%
74804020: Ward 20	1.64	R	1 004.87	34.62	97%	0%	0%	3%	0%
74804021: Ward 21	1.72	R	3 542.83	38.12	62%	7%	0%	30%	1%
74804022: Ward 22	1.10	R	799.76	37.70	95%	1%	0%	4%	0%
74804023: Ward 23	1.15	R	855.67	35.36	99%	0%	0%	0%	1%
74804024: Ward 24	0.77	R	1 670.28	42.08	69%	1%	0%	30%	0%
74804025: Ward 25	1.24	R	598.30	36.34	97%	0%	0%	2%	1%
74804026: Ward 26	1.77	R	1 150.59	36.02	90%	1%	0%	8%	1%
74804027: Ward 27	0.81	R	1 424.92	37.13	85%	1%	0%	13%	1%
74804028: Ward 28	1.10	R	2 453.07	40.06	56%	1%	1%	41%	0%
79700001: Ward 1	9.23	R	27 085.77	34.47	82%	1%	1%	16%	1%
79700002: Ward 2	5.19	R	4 097.86	34.61	99%	0%	0%	0%	1%
79700003: Ward 3	4.34	R	2 979.51	33.16	99%	0%	0%	0%	1%
79700004: Ward 4	1.87	R	1 622.68	34.82	100%	0%	0%	0%	0%
79700005: Ward 5	4.52	R	2 847.78	36.12	98%	0%	0%	0%	1%
79700006: Ward 6	4.98	R	3 566.60	35.27	99%	0%	0%	0%	0%
79700007: Ward 7	4.28	R	3 413.40	35.54	99%	0%	0%	0%	0%
79700008: Ward 8	5.24	R	3 513.16	34.84	99%	0%	0%	0%	1%
79700009: Ward 9	3.77	R	3 589.14	35.81	99%	0%	0%	0%	0%
79700010: Ward 10	3.80	R	3 057.35	36.17	99%	0%	0%	0%	1%
79700011: Ward 11	3.77	R	3 019.69	34.89	98%	0%	0%	0%	1%
79700012: Ward 12	6.13	R	4 387.75	32.99	98%	0%	0%	0%	1%
79700013: Ward 13	4.78	R	15 396.82	36.64	72%	2%	3%	22%	0%
79700014: Ward 14	5.05	R	3 260.83	35.44	99%	0%	0%	0%	0%
79700015: Ward 15	4.63	R	22 198.81	42.89	23%	1%	3%	71%	1%
79700016: Ward 16	3.98	R	17 267.09	41.65	34%	2%	2%	61%	1%
79700017: Ward 17	5.13	R	20 012.09	37.29	58%	4%	5%	32%	1%
79700018: Ward 18	4.29	R	25 832.66	42.69	21%	3%	10%	64%	2%
79700019: Ward 19	4.17	R	28 291.90	44.54	20%	2%	6%	69%	2%
79700020: Ward 20	4.39	R	28 578.91	44.13	27%	4%	10%	56%	4%
79700021: Ward 21	7.01	R	13 226.71	38.28	57%	2%	2%	38%	1%
79700022: Ward 22	4.48	R	19 052.71	40.11	25%	3%	6%	65%	1%
79700023: Ward 23	4.45	R	20 483.34	43.12	18%	1%	4%	75%	1%
79700024: Ward 24	7.15	R	18 967.44	39.69	65%	1%	2%	32%	0%
79700025: Ward 25	6.85	R	6 076.25	36.85	92%	1%	0%	7%	0%
79700026: Ward 26	5.58	R	2 169.93	35.87	99%	0%	0%	0%	0%
79700027: Ward 27	3.65	R	20 062.61	45.12	13%	1%	4%	81%	1%

Stellenbosch University

79700028: Ward 28	3.96	R	22 304.21	43.98	19%	2%	19%	60%	1%
79700029: Ward 29	2.64	R	3 094.07	38.71	48%	6%	45%	0%	1%
79700030: Ward 30	3.30	R	2 127.48	35.48	99%	0%	0%	0%	0%
79700031: Ward 31	5.14	R	13 870.55	39.46	53%	4%	7%	36%	1%
79700032: Ward 32	3.73	R	15 609.46	42.13	26%	8%	3%	62%	1%
79700033: Ward 33	6.81	R	10 003.95	36.12	74%	2%	1%	22%	1%
79700034: Ward 34	3.73	R	3 432.97	37.69	32%	65%	1%	0%	1%
79700035: Ward 35	4.77	R	5 793.66	34.39	80%	5%	1%	14%	1%
79700036: Ward 36	4.22	R	13 469.60	39.00	51%	3%	3%	42%	1%
79700037: Ward 37	4.55	R	21 152.09	43.01	16%	2%	4%	76%	1%
79700038: Ward 38	5.32	R	26 963.45	42.87	22%	4%	7%	66%	1%
79700039: Ward 39	5.51	R	16 749.83	39.77	47%	3%	4%	46%	1%
79700040: Ward 40	5.11	R	7 276.72	36.41	98%	1%	0%	1%	0%
79700041: Ward 41	7.08	R	7 832.87	34.85	99%	0%	0%	0%	0%
79700042: Ward 42	5.81	R	9 900.85	36.42	71%	7%	1%	22%	0%
79700043: Ward 43	4.39	R	16 950.96	39.14	61%	13%	1%	25%	0%
79700044: Ward 44	5.32	R	7 358.57	36.74	99%	0%	0%	0%	0%
79700045: Ward 45	5.05	R	4 470.94	36.30	94%	2%	3%	1%	0%
79700046: Ward 46	4.56	R	4 994.41	37.80	100%	0%	0%	0%	0%
79700047: Ward 47	4.89	R	4 066.64	37.02	100%	0%	0%	0%	0%
79700048: Ward 48	3.66	R	2 292.14	37.40	99%	0%	0%	0%	0%
79700049: Ward 49	4.08	R	2 295.05	36.87	100%	0%	0%	0%	0%
79700050: Ward 50	4.57	R	2 909.98	37.08	99%	0%	0%	0%	0%
79700051: Ward 51	4.97	R	2 591.13	34.27	99%	0%	0%	0%	1%
79700052: Ward 52	5.39	R	2 748.22	33.93	99%	0%	0%	0%	0%
79700053: Ward 53	5.91	R	3 600.00	35.55	73%	25%	0%	0%	1%
79700054: Ward 54	4.15	R	2 219.06	36.16	99%	0%	0%	0%	1%
79700055: Ward 55	4.74	R	2 830.48	36.30	99%	0%	0%	0%	0%
79700056: Ward 56	3.92	R	2 428.54	37.06	99%	0%	0%	0%	0%
79700057: Ward 57	5.39	R	3 104.25	34.58	99%	1%	0%	0%	0%
79700058: Ward 58	7.17	R	4 877.22	35.61	79%	15%	6%	0%	0%
79700059: Ward 59	4.93	R	3 282.41	35.81	99%	0%	0%	0%	0%
79700060: Ward 60	5.04	R	3 559.18	36.37	99%	0%	0%	0%	0%
79700061: Ward 61	6.55	R	1 935.05	34.97	99%	0%	0%	0%	0%
79700062: Ward 62	3.89	R	1 260.00	35.52	99%	0%	0%	0%	0%
79700063: Ward 63	5.24	R	2 304.92	35.31	99%	0%	0%	0%	0%
79700064: Ward 64	5.73	R	2 990.02	36.25	99%	1%	0%	0%	0%
79700065: Ward 65	6.56	R	2 538.87	36.02	99%	0%	0%	0%	0%
79700066: Ward 66	6.90	R	2 893.76	36.93	99%	0%	0%	0%	0%
79700067: Ward 67	5.60	R	3 592.46	36.98	97%	0%	1%	2%	0%
79700068: Ward 68	5.20	R	2 421.34	36.50	99%	0%	0%	0%	0%
79700069: Ward 69	3.55	R	2 111.11	37.75	98%	0%	0%	0%	1%
79700070: Ward 70	3.68	R	2 171.79	37.51	98%	0%	0%	0%	1%
79700071: Ward 71	5.30	R	4 590.33	35.46	97%	1%	1%	1%	1%
79700072: Ward 72	5.25	R	9 764.34	37.90	64%	1%	13%	21%	1%
79700073: Ward 73	4.97	R	7 850.96	37.11	69%	2%	21%	8%	1%
79700074: Ward 74	4.22	R	9 200.43	41.01	60%	1%	1%	38%	0%
79700075: Ward 75	5.06	R	11 753.86	41.21	42%	2%	1%	54%	1%
79700076: Ward 76	4.29	R	12 643.43	42.10	33%	1%	1%	64%	1%
79700077: Ward 77	5.43	R	4 198.71	37.14	99%	0%	0%	0%	0%
79700078: Ward 78	3.75	R	3 918.62	39.37	96%	0%	0%	3%	0%
79700079: Ward 79	5.29	R	2 604.52	38.01	99%	0%	0%	0%	1%
79700080: Ward 80	3.62	R	2 573.33	40.82	99%	0%	0%	0%	0%
79700081: Ward 81	7.42	R	2 918.76	35.30	99%	0%	0%	0%	0%
79700082: Ward 82	4.43	R	3 286.44	38.24	99%	1%	0%	0%	0%
79700083: Ward 83	4.25	R	2 437.04	39.14	99%	0%	0%	0%	0%
79700084: Ward 84	6.62	R	3 303.08	37.15	99%	0%	0%	0%	1%
79700085: Ward 85	5.03	R	3 022.66	37.45	98%	0%	0%	1%	1%
79700086: Ward 86	6.66	R	2 586.04	36.09	99%	0%	0%	0%	0%
79700087: Ward 87	5.01	R	2 367.46	36.84	99%	1%	0%	0%	0%
79700088: Ward 88	4.95	R	9 012.12	40.68	42%	20%	5%	33%	1%
79700089: Ward 89	8.39	R	7 439.46	33.07	97%	0%	0%	2%	1%
79700090: Ward 90	4.66	R	3 212.19	32.81	99%	0%	0%	0%	1%
79700091: Ward 91	5.27	R	17 993.39	37.56	74%	2%	2%	22%	0%
79700092: Ward 92	4.31	R	19 997.51	41.44	33%	2%	6%	57%	2%
79700093: Ward 93	4.04	R	4 227.82	33.16	88%	9%	1%	2%	1%
79700094: Ward 94	4.17	R	23 482.50	40.71	33%	3%	8%	55%	1%
79700095: Ward 95	4.88	R	3 726.77	35.59	99%	0%	0%	0%	0%
79700096: Ward 96	4.45	R	2 265.14	36.67	99%	0%	0%	0%	0%
79700097: Ward 97	5.28	R	11 792.24	41.06	33%	2%	2%	63%	0%

Stellenbosch University

79700098: Ward 98	3.61	R	4 381.85	40.77	87%	1%	0%	11%	1%
79700099: Ward 99	6.96	R	4 766.96	35.14	81%	16%	0%	2%	1%
79700100: Ward 100	4.61	R	17 317.64	39.05	55%	1%	1%	42%	1%
79700101: Ward 101	7.09	R	3 019.77	35.33	99%	0%	0%	0%	0%
79800001: Ward 1	7.52	R	2 605.98	36.56	99%	0%	0%	0%	0%
79800002: Ward 2	6.91	R	2 687.31	37.28	99%	0%	0%	0%	0%
79800003: Ward 3	5.22	R	2 079.97	36.55	99%	0%	0%	0%	0%
79800004: Ward 4	6.88	R	2 750.61	36.87	99%	0%	0%	0%	0%
79800005: Ward 5	8.30	R	2 830.28	35.72	99%	1%	0%	0%	0%
79800006: Ward 6	6.38	R	2 207.34	35.73	98%	1%	0%	0%	1%
79800007: Ward 7	6.38	R	6 546.86	36.82	58%	41%	1%	0%	1%
79800008: Ward 8	6.37	R	4 698.37	36.15	74%	2%	24%	0%	0%
79800009: Ward 9	4.39	R	8 525.11	42.01	17%	4%	78%	0%	1%
79800010: Ward 10	5.21	R	7 003.43	39.34	57%	1%	41%	0%	1%
79800011: Ward 11	5.21	R	4 538.40	37.07	71%	28%	0%	0%	0%
79800012: Ward 12	4.44	R	4 078.34	37.09	100%	0%	0%	0%	0%
79800013: Ward 13	6.27	R	9 639.42	35.52	99%	0%	0%	0%	0%
79800014: Ward 14	4.98	R	8 374.38	38.65	100%	0%	0%	0%	0%
79800015: Ward 15	4.22	R	2 186.54	38.38	100%	0%	0%	0%	0%
79800016: Ward 16	4.46	R	3 255.21	38.34	99%	0%	0%	0%	0%
79800017: Ward 17	5.40	R	4 568.14	37.43	34%	63%	1%	0%	2%
79800018: Ward 18	5.92	R	6 348.45	39.39	7%	90%	1%	0%	1%
79800019: Ward 19	4.46	R	3 820.70	37.17	97%	2%	0%	0%	0%
79800020: Ward 20	5.02	R	3 051.34	39.31	100%	0%	0%	0%	0%
79800021: Ward 21	4.81	R	3 440.73	39.33	100%	0%	0%	0%	0%
79800022: Ward 22	4.59	R	4 510.74	38.37	99%	1%	0%	0%	0%
79800023: Ward 23	5.09	R	31 082.25	42.30	32%	5%	12%	49%	2%
79800024: Ward 24	4.54	R	2 781.14	33.74	99%	0%	0%	0%	1%
79800025: Ward 25	5.61	R	4 626.48	37.91	100%	0%	0%	0%	0%
79800026: Ward 26	3.64	R	2 576.16	38.17	99%	0%	0%	0%	0%
79800027: Ward 27	3.85	R	2 958.84	39.20	99%	0%	0%	0%	0%
79800028: Ward 28	4.06	R	3 007.36	38.18	99%	0%	0%	0%	0%
79800029: Ward 29	4.12	R	3 161.52	38.03	77%	22%	0%	0%	0%
79800030: Ward 30	5.66	R	3 085.76	36.29	99%	0%	0%	0%	1%
79800031: Ward 31	4.33	R	2 726.56	37.04	99%	0%	0%	0%	0%
79800032: Ward 32	6.54	R	29 884.71	36.01	66%	2%	10%	20%	1%
79800033: Ward 33	4.39	R	3 397.03	39.86	99%	0%	0%	0%	0%
79800034: Ward 34	4.79	R	3 081.44	37.57	99%	0%	0%	0%	0%
79800035: Ward 35	3.85	R	2 305.33	39.13	99%	0%	0%	0%	0%
79800036: Ward 36	3.50	R	2 630.47	39.55	99%	0%	0%	0%	0%
79800037: Ward 37	4.16	R	2 529.97	36.81	99%	0%	0%	0%	0%
79800038: Ward 38	3.33	R	3 123.32	39.06	100%	0%	0%	0%	0%
79800039: Ward 39	4.69	R	3 614.34	38.57	99%	0%	0%	0%	0%
79800040: Ward 40	4.01	R	2 216.99	36.71	100%	0%	0%	0%	0%
79800041: Ward 41	3.68	R	2 446.46	39.44	99%	0%	0%	0%	0%
79800042: Ward 42	3.92	R	2 382.24	38.00	99%	0%	0%	0%	0%
79800043: Ward 43	4.25	R	2 753.08	39.28	100%	0%	0%	0%	0%
79800044: Ward 44	9.03	R	6 286.40	35.13	99%	0%	0%	0%	0%
79800045: Ward 45	4.52	R	3 150.87	38.40	100%	0%	0%	0%	0%
79800046: Ward 46	4.69	R	2 828.12	37.72	99%	0%	0%	0%	0%
79800047: Ward 47	4.59	R	3 229.47	37.62	99%	0%	0%	0%	0%
79800048: Ward 48	5.10	R	5 323.72	38.16	99%	0%	0%	0%	0%
79800049: Ward 49	8.33	R	4 395.44	35.92	99%	0%	0%	0%	0%
79800050: Ward 50	3.55	R	1 492.33	36.81	99%	0%	0%	0%	0%
79800051: Ward 51	4.56	R	2 220.90	37.92	100%	0%	0%	0%	0%
79800052: Ward 52	4.59	R	2 782.28	38.42	100%	0%	0%	0%	0%
79800053: Ward 53	10.60	R	13 978.13	36.28	99%	1%	0%	0%	0%
79800054: Ward 54	6.76	R	30 985.64	39.82	44%	12%	23%	19%	3%
79800055: Ward 55	4.73	R	9 051.24	38.03	52%	15%	13%	18%	1%
79800056: Ward 56	4.91	R	12 499.53	39.22	58%	7%	4%	29%	2%
79800057: Ward 57	5.94	R	12 526.64	38.66	56%	11%	5%	26%	1%
79800058: Ward 58	4.36	R	8 206.66	37.56	37%	8%	43%	7%	6%
79800059: Ward 59	0.78	R	880.85	31.28	96%	2%	1%	0%	0%
79800060: Ward 60	5.47	R	9 851.78	30.50	91%	2%	4%	2%	1%
79800061: Ward 61	2.13	R	2 879.67	33.31	91%	3%	2%	4%	1%
79800062: Ward 62	2.40	R	3 584.60	31.63	98%	1%	0%	0%	0%
79800063: Ward 63	4.12	R	4 409.12	31.06	99%	1%	0%	0%	0%
79800064: Ward 64	5.02	R	6 140.53	32.52	96%	1%	1%	1%	0%
79800065: Ward 65	3.49	R	2 580.30	31.82	97%	1%	1%	1%	1%
79800066: Ward 66	6.61	R	14 701.24	37.00	72%	6%	7%	14%	2%

Stellenbosch University

79800067: Ward 67	2.98	R	7 192.96	33.35	84%	2%	7%	6%	1%
79800068: Ward 68	4.89	R	6 877.43	37.08	56%	42%	1%	0%	1%
79800069: Ward 69	4.68	R	13 078.67	35.69	41%	26%	15%	16%	2%
79800070: Ward 70	4.89	R	16 576.40	40.65	36%	27%	7%	28%	2%
79800071: Ward 71	4.89	R	14 418.48	40.13	42%	17%	2%	39%	0%
79800072: Ward 72	2.69	R	15 873.17	47.78	26%	1%	1%	70%	1%
79800073: Ward 73	4.37	R	27 753.51	43.00	45%	2%	13%	37%	2%
79800074: Ward 74	4.07	R	27 277.76	44.07	42%	2%	5%	48%	2%
79800075: Ward 75	3.66	R	2 549.17	33.87	99%	0%	0%	0%	0%
79800076: Ward 76	3.00	R	1 779.73	36.28	99%	0%	0%	0%	0%
79800077: Ward 77	8.33	R	4 360.32	32.70	99%	0%	0%	0%	1%
79800078: Ward 78	7.45	R	3 839.86	32.89	98%	0%	0%	0%	1%
79800079: Ward 79	7.91	R	4 155.78	32.94	99%	0%	0%	0%	1%
79800080: Ward 80	4.50	R	2 556.46	32.93	99%	0%	0%	0%	0%
79800081: Ward 81	5.38	R	15 419.91	37.75	79%	2%	5%	13%	0%
79800082: Ward 82	5.62	R	5 990.88	39.00	26%	57%	2%	12%	3%
79800083: Ward 83	3.34	R	19 088.22	45.64	18%	4%	2%	76%	1%
79800084: Ward 84	5.01	R	16 497.50	41.11	35%	20%	5%	39%	1%
79800085: Ward 85	3.64	R	20 193.50	42.62	39%	7%	5%	49%	1%
79800086: Ward 86	5.30	R	20 726.44	40.95	33%	17%	6%	43%	2%
79800087: Ward 87	2.47	R	24 222.58	44.24	30%	2%	10%	56%	3%
79800088: Ward 88	3.78	R	29 707.81	43.64	26%	4%	14%	55%	2%
79800089: Ward 89	3.47	R	25 679.13	43.45	22%	6%	7%	63%	1%
79800090: Ward 90	3.74	R	39 737.14	44.38	33%	2%	5%	58%	2%
79800091: Ward 91	4.20	R	25 727.83	40.89	62%	2%	6%	29%	2%
79800092: Ward 92	7.50	R	20 467.86	34.70	86%	1%	4%	7%	1%
79800093: Ward 93	3.70	R	39 873.38	38.36	32%	3%	17%	46%	2%
79800094: Ward 94	3.36	R	34 563.89	42.73	26%	1%	5%	66%	2%
79800095: Ward 95	7.49	R	4 051.53	32.78	99%	0%	0%	0%	1%
79800096: Ward 96	9.07	R	53 904.04	36.26	62%	2%	5%	30%	1%
79800097: Ward 97	8.55	R	57 756.00	37.99	34%	4%	6%	55%	1%
79800098: Ward 98	4.53	R	26 537.35	40.95	46%	4%	7%	42%	1%
79800099: Ward 99	3.17	R	22 215.12	43.23	24%	3%	8%	63%	1%
79800100: Ward 100	8.97	R	17 199.33	33.23	92%	1%	1%	4%	1%
79800101: Ward 101	5.03	R	37 829.75	42.26	25%	3%	10%	60%	1%
79800102: Ward 102	4.93	R	38 626.91	42.27	36%	3%	9%	51%	1%
79800103: Ward 103	5.78	R	65 897.48	42.71	29%	2%	13%	53%	2%
79800104: Ward 104	4.73	R	34 316.16	41.33	37%	3%	9%	49%	2%
79800105: Ward 105	8.13	R	5 857.90	35.60	99%	0%	0%	0%	0%
79800106: Ward 106	5.56	R	56 858.51	41.77	29%	3%	9%	57%	2%
79800107: Ward 107	3.19	R	1 809.72	36.35	98%	1%	0%	0%	1%
79800108: Ward 108	5.35	R	3 102.22	35.01	98%	1%	0%	0%	1%
79800109: Ward 109	4.07	R	16 718.91	38.12	63%	2%	16%	18%	1%
79800110: Ward 110	8.89	R	10 809.60	33.24	88%	7%	1%	3%	1%
79800111: Ward 111	9.79	R	7 899.95	32.92	99%	1%	0%	0%	1%
79800112: Ward 112	9.43	R	75 901.23	36.03	54%	3%	20%	22%	1%
79800113: Ward 113	15.27	R	8 833.81	32.19	98%	0%	0%	0%	2%
79800114: Ward 114	6.15	R	20 883.95	34.34	78%	1%	3%	17%	1%
79800115: Ward 115	5.98	R	53 579.35	40.13	25%	2%	8%	63%	1%
79800116: Ward 116	3.49	R	1 937.06	35.29	99%	0%	0%	0%	0%
79800117: Ward 117	3.20	R	32 294.03	44.83	28%	2%	6%	63%	2%
79800118: Ward 118	4.11	R	14 316.92	39.77	57%	6%	10%	26%	2%
79800119: Ward 119	7.27	R	8 480.23	34.57	93%	6%	0%	0%	1%
79800120: Ward 120	6.18	R	9 022.04	37.16	58%	4%	38%	0%	1%
79800121: Ward 121	8.80	R	8 578.86	36.63	75%	17%	7%	0%	0%
79800122: Ward 122	8.41	R	6 556.45	36.17	85%	2%	11%	1%	1%
79800123: Ward 123	6.23	R	6 679.46	30.23	98%	1%	0%	0%	0%
79800124: Ward 124	6.82	R	11 121.08	34.53	79%	6%	6%	9%	1%
79800125: Ward 125	5.88	R	21 987.15	36.22	81%	9%	3%	7%	1%
79800126: Ward 126	4.51	R	31 320.14	41.82	24%	4%	7%	63%	1%
79800127: Ward 127	4.65	R	2 610.31	33.86	98%	1%	0%	0%	1%
79800128: Ward 128	9.00	R	3 771.66	35.09	99%	0%	0%	0%	1%
79800129: Ward 129	6.09	R	3 159.44	36.96	100%	0%	0%	0%	0%
79800130: Ward 130	3.55	R	2 586.21	38.29	97%	1%	1%	1%	0%
79900001: Ward 1	3.77	R	10 806.25	43.03	12%	2%	1%	85%	1%
79900002: Ward 2	3.07	R	12 367.61	43.74	25%	2%	1%	71%	1%
79900003: Ward 3	5.08	R	14 865.43	37.78	64%	3%	5%	27%	1%
79900004: Ward 4	6.12	R	27 886.55	37.10	81%	1%	1%	16%	0%
79900005: Ward 5	4.34	R	23 991.31	43.13	25%	2%	2%	72%	1%
79900006: Ward 6	2.82	R	2 538.82	38.99	99%	0%	0%	0%	0%

Stellenbosch University

79900007: Ward 7	6.97	R	10 128.65	36.15	88%	1%	4%	6%	0%
79900008: Ward 8	4.84	R	2 083.09	38.49	99%	0%	0%	0%	0%
79900009: Ward 9	6.05	R	1 783.56	36.49	100%	0%	0%	0%	0%
79900010: Ward 10	6.87	R	3 154.16	33.61	99%	0%	0%	0%	1%
79900011: Ward 11	3.93	R	2 294.21	37.65	99%	0%	0%	0%	0%
79900012: Ward 12	3.64	R	2 577.70	37.28	99%	0%	0%	0%	0%
79900013: Ward 13	4.39	R	1 794.73	37.62	99%	0%	0%	0%	0%
79900014: Ward 14	4.34	R	1 646.94	37.30	99%	0%	0%	0%	0%
79900015: Ward 15	3.18	R	4 188.23	38.33	96%	0%	3%	0%	1%
79900016: Ward 16	3.46	R	2 254.51	36.80	99%	0%	0%	0%	0%
79900017: Ward 17	6.58	R	6 041.01	34.80	99%	0%	0%	0%	0%
79900018: Ward 18	3.89	R	4 454.24	38.80	99%	0%	0%	0%	0%
79900019: Ward 19	4.81	R	2 245.73	36.83	99%	0%	0%	0%	0%
79900020: Ward 20	4.48	R	4 729.98	39.35	99%	0%	0%	0%	0%
79900021: Ward 21	4.99	R	6 329.68	39.59	100%	0%	0%	0%	0%
79900022: Ward 22	5.49	R	5 451.49	36.98	99%	0%	0%	0%	0%
79900023: Ward 23	3.49	R	2 516.25	38.22	99%	0%	0%	0%	0%
79900024: Ward 24	7.82	R	2 444.46	36.93	99%	0%	0%	0%	0%
79900025: Ward 25	4.19	R	1 812.56	36.73	99%	0%	0%	0%	0%
79900026: Ward 26	4.17	R	2 521.13	36.94	99%	0%	0%	0%	0%
79900027: Ward 27	4.62	R	1 953.21	37.11	99%	0%	0%	0%	0%
79900028: Ward 28	2.93	R	2 923.27	39.61	99%	0%	0%	0%	0%
79900029: Ward 29	4.72	R	3 236.78	37.17	99%	0%	0%	0%	0%
79900030: Ward 30	6.37	R	4 621.03	37.42	99%	0%	0%	0%	0%
79900031: Ward 31	3.97	R	3 560.72	39.34	99%	0%	0%	0%	1%
79900032: Ward 32	4.76	R	6 484.27	37.79	99%	0%	0%	0%	0%
79900033: Ward 33	2.77	R	2 341.86	40.59	99%	0%	0%	0%	0%
79900034: Ward 34	3.29	R	2 623.99	37.81	99%	0%	0%	0%	0%
79900035: Ward 35	2.88	R	2 997.79	38.09	99%	0%	0%	0%	0%
79900036: Ward 36	3.27	R	3 044.93	36.17	99%	0%	0%	0%	0%
79900037: Ward 37	7.27	R	3 580.03	34.62	99%	0%	0%	0%	1%
79900038: Ward 38	1.11	R	1 542.51	38.46	99%	1%	0%	0%	0%
79900039: Ward 39	5.14	R	2 540.47	36.14	99%	0%	0%	0%	1%
79900040: Ward 40	10.53	R	6 060.10	34.38	99%	1%	0%	0%	0%
79900041: Ward 41	2.88	R	14 618.73	41.95	32%	7%	2%	58%	1%
79900042: Ward 42	3.25	R	27 878.37	45.64	24%	1%	3%	69%	3%
79900043: Ward 43	4.17	R	6 763.15	39.36	19%	74%	1%	4%	1%
79900044: Ward 44	2.80	R	19 251.32	46.53	16%	1%	2%	79%	2%
79900045: Ward 45	1.75	R	11 140.83	44.09	18%	2%	3%	76%	1%
79900046: Ward 46	3.25	R	21 438.79	44.88	23%	2%	3%	69%	3%
79900047: Ward 47	2.64	R	17 696.85	42.40	19%	1%	2%	77%	1%
79900048: Ward 48	5.56	R	5 466.29	34.99	95%	0%	0%	4%	0%
79900049: Ward 49	5.64	R	4 528.08	37.11	91%	0%	0%	8%	1%
79900050: Ward 50	3.29	R	16 040.48	44.33	18%	1%	2%	78%	1%
79900051: Ward 51	4.18	R	4 820.88	37.83	87%	0%	11%	0%	1%
79900052: Ward 52	3.26	R	14 824.82	44.77	11%	2%	1%	86%	1%
79900053: Ward 53	3.49	R	15 252.11	44.01	24%	2%	1%	71%	1%
79900054: Ward 54	3.23	R	12 252.63	44.43	13%	2%	1%	82%	1%
79900055: Ward 55	3.97	R	11 365.33	41.21	30%	2%	1%	67%	1%
79900056: Ward 56	2.38	R	10 479.75	31.81	39%	2%	4%	53%	2%
79900057: Ward 57	4.43	R	24 674.97	41.08	33%	3%	5%	58%	1%
79900058: Ward 58	4.23	R	14 000.99	33.00	75%	3%	1%	21%	0%
79900059: Ward 59	3.48	R	12 659.80	33.28	73%	2%	2%	22%	1%
79900060: Ward 60	3.86	R	9 971.32	31.26	86%	3%	2%	8%	2%
79900061: Ward 61	6.58	R	11 357.55	37.44	50%	2%	45%	2%	1%
79900062: Ward 62	3.28	R	3 089.68	40.72	99%	0%	0%	0%	0%
79900063: Ward 63	2.69	R	2 968.24	37.06	99%	0%	0%	1%	0%
79900064: Ward 64	5.60	R	37 838.67	38.82	28%	3%	5%	64%	1%
79900065: Ward 65	3.35	R	26 771.74	42.43	25%	2%	3%	68%	1%
79900066: Ward 66	3.24	R	16 460.83	41.65	30%	5%	4%	60%	1%
79900067: Ward 67	3.11	R	2 778.93	37.69	99%	1%	0%	0%	0%
79900068: Ward 68	4.67	R	4 123.03	37.33	99%	0%	0%	0%	0%
79900069: Ward 69	3.94	R	27 097.12	42.50	12%	1%	4%	82%	1%
79900070: Ward 70	4.98	R	34 198.78	39.90	27%	2%	16%	54%	1%
79900071: Ward 71	5.27	R	2 794.35	32.95	99%	0%	0%	0%	1%
79900072: Ward 72	3.70	R	2 309.39	36.77	100%	0%	0%	0%	0%
79900073: Ward 73	6.31	R	3 072.43	36.85	98%	0%	0%	1%	0%
79900074: Ward 74	3.90	R	2 615.27	39.65	99%	0%	0%	0%	0%
79900075: Ward 75	4.85	R	4 830.12	38.11	99%	0%	0%	0%	0%
79900076: Ward 76	3.82	R	1 726.91	39.26	99%	0%	0%	0%	0%

Stellenbosch University

79900077: Ward 77	12.46	R	27 354.65	32.93	92%	1%	1%	5%	1%
79900078: Ward 78	3.73	R	28 157.14	37.89	29%	2%	5%	63%	1%
79900079: Ward 79	4.53	R	30 532.19	42.05	18%	2%	2%	77%	1%
79900080: Ward 80	3.79	R	8 539.19	30.42	92%	2%	1%	5%	1%
79900081: Ward 81	1.99	R	4 243.46	30.05	92%	2%	1%	5%	0%
79900082: Ward 82	2.66	R	21 487.51	39.70	25%	2%	2%	69%	2%
79900083: Ward 83	1.85	R	13 198.97	43.58	24%	2%	2%	71%	2%
79900084: Ward 84	4.03	R	16 540.28	43.49	28%	6%	2%	63%	2%
79900085: Ward 85	5.53	R	39 155.07	43.11	23%	3%	2%	70%	2%
79900086: Ward 86	6.77	R	6 328.37	34.54	94%	3%	0%	1%	2%
79900087: Ward 87	3.65	R	9 336.53	39.67	39%	26%	1%	32%	1%
79900088: Ward 88	4.57	R	4 030.38	37.74	100%	0%	0%	0%	0%
79900089: Ward 89	4.65	R	4 896.06	36.23	99%	0%	0%	0%	0%
79900090: Ward 90	7.24	R	8 820.25	34.22	99%	0%	0%	0%	1%
79900091: Ward 91	6.40	R	46 296.78	39.79	30%	2%	2%	65%	1%
79900092: Ward 92	4.28	R	13 246.08	33.62	73%	3%	2%	20%	2%
79900093: Ward 93	2.67	R	2 438.10	37.19	93%	1%	0%	5%	0%
79900094: Ward 94	3.37	R	3 282.83	38.36	99%	0%	0%	0%	0%
79900095: Ward 95	3.86	R	1 471.22	39.12	99%	0%	0%	0%	0%
79900096: Ward 96	4.71	R	15 518.02	40.39	49%	1%	1%	49%	1%
79900097: Ward 97	3.56	R	1 804.33	33.50	99%	0%	0%	0%	0%
79900098: Ward 98	4.08	R	15 176.29	40.75	45%	1%	1%	53%	0%
79900099: Ward 99	5.33	R	7 314.49	36.50	88%	0%	0%	11%	1%
79900100: Ward 100	4.50	R	6 485.75	37.91	74%	2%	0%	23%	1%
79900101: Ward 101	4.72	R	33 293.22	39.95	38%	1%	2%	57%	1%
79900102: Ward 102	5.05	R	5 503.52	37.12	88%	1%	0%	9%	1%
79900103: Ward 103	4.28	R	2 889.90	36.78	99%	0%	0%	0%	0%
79900104: Ward 104	4.31	R	1 842.72	35.65	99%	0%	0%	0%	0%
79900105: Ward 105	3.81	R	7 579.57	40.81	74%	1%	1%	23%	0%

Appendix 3**Train sample bins ranked from lowest to highest variance**

<i>Bin Index</i>	<i>Variance</i>	<i>Population</i>	<i>Latitude</i>	<i>Longitude</i>	<i>Municipal District</i>
201	0.002919	187	-	26.299	27.652
202	0.035181	87	-	26.329	27.652
307	0.073955	213	-	26.179	27.705
1023	0.088304	277	-	26.209	28.049
1133	0.100865	63	-	26.209	28.102
1128	0.110576	233	-	26.058	28.102
1405	0.117986	446	-	26.119	28.235
1283	0.124337	358	-	25.757	28.182
1338	0.126581	178	-	25.757	28.208
1342	0.133325	101	-	25.878	28.208
1435	0.137487	77	-	25.366	28.261
1132	0.140531	112	-	26.179	28.102
1291	0.146672	97	-	25.998	28.182
1164	0.147162	90	-	25.486	28.129
1407	0.1506	199	-	26.179	28.235
1243	0.152736	207	-	26.209	28.155
1518	0.154203	272	-	26.209	28.288
1130	0.154904	41	-	26.119	28.102
1572	0.160224	213	-	26.179	28.314
596	0.160542	228	-	26.601	27.837
1287	0.165276	51	-	25.878	28.182
1286	0.168354	110	-	25.848	28.182
1136	0.168553	50	-	26.299	28.102
1025	0.168903	72	-	26.269	28.049
1341	0.170158	58	-	25.848	28.208
1346	0.170398	72	-	25.998	28.208
1910	0.173852	176	-	26.420	28.473
1684	0.175179	150	-	26.239	28.367
1504	0.175675	169	-	25.787	28.288
1126	0.177081	117	-	25.998	28.102
1177	0.182996	133	-	25.878	28.129
1776	0.183141	102	-	25.697	28.420
636	0.184024	270	-	26.149	27.864
1228	0.184879	105	-	25.757	28.155
1135	0.185001	102	-	26.269	28.102
1693	0.186679	187	-	26.510	28.367
590	0.186912	76	-	26.420	27.837
1358	0.187071	81	-	26.360	28.208
2440	0.18797	116	-	25.817	28.738
1505	0.188922	43	-	25.817	28.288
593	0.190631	55	-	26.510	27.837
530	0.192735	80	-	26.269	27.811
1137	0.193053	51	-	26.329	28.102
1347	0.194859	112	-	26.028	28.208
1445	0.195738	200	-	25.667	28.261
1436	0.196283	168	-	25.396	28.261

Stellenbosch University

1400	0.196353	76	-	25.968	28.235	Ekurhuleni Metropolitan Municipality
1112	0.200013	74	-	25.576	28.102	City of Tshwane Metropolitan Municipality
1354	0.202512	60	-	26.239	28.208	Ekurhuleni Metropolitan Municipality
1190	0.203878	338	-	26.269	28.129	Ekurhuleni Metropolitan Municipality
1736	0.205174	53	-	26.149	28.394	Ekurhuleni Metropolitan Municipality
1390	0.205446	47	-	25.667	28.235	City of Tshwane Metropolitan Municipality
1403	0.206376	63	-	26.058	28.235	Ekurhuleni Metropolitan Municipality
1559	0.206433	258	-	25.787	28.314	City of Tshwane Metropolitan Municipality
1777	0.20745	45	-	25.727	28.420	City of Tshwane Metropolitan Municipality
1109	0.208243	107	-	25.486	28.102	City of Tshwane Metropolitan Municipality
1466	0.208669	62	-	26.299	28.261	Ekurhuleni Metropolitan Municipality
1491	0.209283	153	-	25.396	28.288	City of Tshwane Metropolitan Municipality
1401	0.210163	48	-	25.998	28.235	Ekurhuleni Metropolitan Municipality
1301	0.210378	76	-	26.299	28.182	Ekurhuleni Metropolitan Municipality
1615	0.212481	51	-	25.817	28.341	City of Tshwane Metropolitan Municipality
1240	0.212758	212	-	26.119	28.155	City of Johannesburg Metropolitan Municipality
1557	0.216044	141	-	25.727	28.314	City of Tshwane Metropolitan Municipality
1722	0.216994	60	-	25.727	28.394	City of Tshwane Metropolitan Municipality
595	0.217244	199	-	26.570	27.837	Emfuleni Local Municipality
1561	0.217295	53	-	25.848	28.314	City of Tshwane Metropolitan Municipality
1393	0.21843	81	-	25.757	28.235	City of Tshwane Metropolitan Municipality
1357	0.218599	58	-	26.329	28.208	Ekurhuleni Metropolitan Municipality
1449	0.219391	41	-	25.787	28.261	City of Tshwane Metropolitan Municipality
1115	0.221092	66	-	25.667	28.102	City of Tshwane Metropolitan Municipality
1688	0.222325	117	-	26.360	28.367	Ekurhuleni Metropolitan Municipality
1464	0.222615	168	-	26.239	28.261	Ekurhuleni Metropolitan Municipality
1232	0.223013	70	-	25.878	28.155	City of Tshwane Metropolitan Municipality
1057	0.223644	43	-	25.576	28.076	City of Tshwane Metropolitan Municipality
638	0.224953	53	-	26.209	27.864	City of Johannesburg Metropolitan Municipality
1404	0.224965	63	-	26.089	28.235	Ekurhuleni Metropolitan Municipality
599	0.225173	235	-	26.691	27.837	Emfuleni Local Municipality
1194	0.225677	44	-	26.390	28.129	Ekurhuleni Metropolitan Municipality
472	0.225847	134	-	26.179	27.784	Mogale City Local Municipality
1054	0.226283	108	-	25.486	28.076	City of Tshwane Metropolitan Municipality
1560	0.226763	219	-	25.817	28.314	City of Tshwane Metropolitan Municipality
471	0.227524	89	-	26.149	27.784	Mogale City Local Municipality
1849	0.227949	139	-	26.239	28.447	Ekurhuleni Metropolitan Municipality
1125	0.234274	104	-	25.968	28.102	City of Johannesburg Metropolitan Municipality
1247	0.235631	133	-	26.329	28.155	Ekurhuleni Metropolitan Municipality
1179	0.237775	45	-	25.938	28.129	City of Johannesburg Metropolitan Municipality
1614	0.238108	57	-	25.787	28.341	City of Tshwane Metropolitan Municipality
1558	0.238399	135	-	25.757	28.314	City of Tshwane Metropolitan Municipality
1110	0.239157	66	-	25.516	28.102	City of Tshwane Metropolitan Municipality
1180	0.239571	282	-	25.968	28.129	City of Johannesburg Metropolitan Municipality
1574	0.240084	130	-	26.239	28.314	Ekurhuleni Metropolitan Municipality
1613	0.241597	44	-	25.757	28.341	City of Tshwane Metropolitan Municipality
1348	0.2435	115	-	26.058	28.208	Ekurhuleni Metropolitan Municipality
1463	0.243858	119	-	26.209	28.261	Ekurhuleni Metropolitan Municipality
1245	0.24493	66	-	26.269	28.155	Ekurhuleni Metropolitan Municipality
1227	0.245415	114	-	25.727	28.155	City of Tshwane Metropolitan Municipality

Stellenbosch University

581	0.245459	67	-	26.149	27.837	City of Johannesburg Metropolitan Municipality
1188	0.246438	73	-	26.209	28.129	Ekurhuleni Metropolitan Municipality
1193	0.247531	47	-	26.360	28.129	Ekurhuleni Metropolitan Municipality
1280	0.248453	79	-	25.667	28.182	City of Tshwane Metropolitan Municipality
1072	0.248984	44	-	26.028	28.076	City of Johannesburg Metropolitan Municipality
1349	0.252856	72	-	26.089	28.208	Ekurhuleni Metropolitan Municipality
1741	0.253593	81	-	26.299	28.394	Ekurhuleni Metropolitan Municipality
1298	0.254104	52	-	26.209	28.182	Ekurhuleni Metropolitan Municipality
1231	0.255432	51	-	25.848	28.155	City of Tshwane Metropolitan Municipality
1122	0.257381	41	-	25.878	28.102	City of Tshwane Metropolitan Municipality
1181	0.259397	141	-	25.998	28.129	City of Johannesburg Metropolitan Municipality
1612	0.261223	76	-	25.727	28.341	City of Tshwane Metropolitan Municipality
1302	0.261629	150	-	26.329	28.182	Ekurhuleni Metropolitan Municipality
580	0.264577	53	-	26.119	27.837	City of Johannesburg Metropolitan Municipality
1241	0.26524	67	-	26.149	28.155	Ekurhuleni Metropolitan Municipality
1337	0.265479	45	-	25.727	28.208	City of Tshwane Metropolitan Municipality
914	0.266386	41	-	26.239	27.996	City of Johannesburg Metropolitan Municipality
1178	0.266417	91	-	25.908	28.129	City of Tshwane Metropolitan Municipality
1519	0.26946	120	-	26.239	28.288	Ekurhuleni Metropolitan Municipality
469	0.269684	92	-	26.089	27.784	Mogale City Local Municipality
1242	0.270162	147	-	26.179	28.155	Ekurhuleni Metropolitan Municipality
635	0.270615	104	-	26.119	27.864	City of Johannesburg Metropolitan Municipality
799	0.271138	62	-	26.089	27.943	City of Johannesburg Metropolitan Municipality
594	0.272084	133	-	26.540	27.837	Emfuleni Local Municipality
800	0.276708	50	-	26.119	27.943	City of Johannesburg Metropolitan Municipality
600	0.278931	50	-	26.721	27.837	Emfuleni Local Municipality
911	0.285538	47	-	26.149	27.996	City of Johannesburg Metropolitan Municipality
693	0.286272	69	-	26.209	27.890	City of Johannesburg Metropolitan Municipality
1292	0.287466	175	-	26.028	28.182	City of Johannesburg Metropolitan Municipality
801	0.289015	43	-	26.149	27.943	City of Johannesburg Metropolitan Municipality
639	0.289772	85	-	26.239	27.864	City of Johannesburg Metropolitan Municipality
1075	0.290656	106	-	26.119	28.076	City of Johannesburg Metropolitan Municipality
915	0.293621	69	-	26.269	27.996	City of Johannesburg Metropolitan Municipality
798	0.294626	73	-	26.058	27.943	City of Johannesburg Metropolitan Municipality
1063	0.294834	78	-	25.757	28.076	City of Tshwane Metropolitan Municipality
804	0.294882	87	-	26.239	27.943	City of Johannesburg Metropolitan Municipality
746	0.296741	124	-	26.149	27.917	City of Johannesburg Metropolitan Municipality
1249	0.297781	124	-	26.390	28.155	Ekurhuleni Metropolitan Municipality
910	0.298565	50	-	26.119	27.996	City of Johannesburg Metropolitan Municipality
909	0.302972	198	-	26.089	27.996	City of Johannesburg Metropolitan Municipality
747	0.304494	111	-	26.179	27.917	City of Johannesburg Metropolitan Municipality
860	0.305824	195	-	26.269	27.970	City of Johannesburg Metropolitan Municipality
959	0.307546	42	-	25.938	28.023	City of Johannesburg Metropolitan Municipality
414	0.310386	101	-	26.089	27.758	Mogale City Local Municipality
855	0.313963	228	-	26.119	27.970	City of Johannesburg Metropolitan Municipality
857	0.316952	119	-	26.179	27.970	City of Johannesburg Metropolitan Municipality
695	0.317086	142	-	26.269	27.890	City of Johannesburg Metropolitan Municipality
690	0.318621	57	-	26.119	27.890	City of Johannesburg Metropolitan Municipality
764	0.323771	52	-	26.691	27.917	Emfuleni Local Municipality
1248	0.324537	120	-	26.360	28.155	Ekurhuleni Metropolitan Municipality

Stellenbosch University

650	0.330295	50	-	26.570	27.864	Emfuleni Local Municipality
649	0.332318	51	-	26.540	27.864	Emfuleni Local Municipality
763	0.332366	62	-	26.661	27.917	Emfuleni Local Municipality
873	0.332976	105	-	26.661	27.970	Emfuleni Local Municipality
745	0.336414	184	-	26.119	27.917	City of Johannesburg Metropolitan Municipality
648	0.339165	44	-	26.510	27.864	City of Johannesburg Metropolitan Municipality
1118	0.342352	67	-	25.757	28.102	City of Tshwane Metropolitan Municipality
856	0.342809	54	-	26.149	27.970	City of Johannesburg Metropolitan Municipality
892	0.345867	47	-	25.576	27.996	City of Tshwane Metropolitan Municipality
818	0.35784	189	-	26.661	27.943	Emfuleni Local Municipality
1000	0.359517	65	-	25.516	28.049	City of Tshwane Metropolitan Municipality
819	0.361267	98	-	26.691	27.943	Emfuleni Local Municipality
1022	0.363195	85	-	26.179	28.049	City of Johannesburg Metropolitan Municipality
1124	0.363863	46	-	25.938	28.102	City of Johannesburg Metropolitan Municipality
966	0.367632	54	-	26.149	28.023	City of Johannesburg Metropolitan Municipality
642	0.373146	125	-	26.329	27.864	City of Johannesburg Metropolitan Municipality
709	0.374672	183	-	26.691	27.890	Emfuleni Local Municipality
965	0.378043	118	-	26.119	28.023	City of Johannesburg Metropolitan Municipality
742	0.380291	46	-	26.028	27.917	Mogale City Local Municipality
854	0.381016	77	-	26.089	27.970	City of Johannesburg Metropolitan Municipality
962	0.397509	209	-	26.028	28.023	City of Johannesburg Metropolitan Municipality
980	0.398152	107	-	26.570	28.023	Midvaal Local Municipality
744	0.407056	48	-	26.089	27.917	City of Johannesburg Metropolitan Municipality
853	0.435456	58	-	26.058	27.970	City of Johannesburg Metropolitan Municipality
1282	0.468664	201	-	25.727	28.182	City of Tshwane Metropolitan Municipality

Appendix 4

Detailed customer profile information (by municipal district)

Municipality	Average Family Size	Average Monthly Income	Average Age	Population Group %				
				Black	White	Indian or Asian	Coloured	Other
City of Johannesburg Metropolitan Municipality	6.00	R 26 729.16	38.61	57%	28%	8%	6%	1%
City of Tshwane Metropolitan Municipality	5.00	R 16 179.68	39.04	58%	37%	2%	2%	1%
Ekurhuleni Metropolitan Municipality	5.00	R 13 031.48	38.85	59%	33%	3%	4%	1%
Emfuleni Local Municipality	3.00	R 4 884.20	38.27	71%	25%	1%	1%	1%
Lesedi Local Municipality	2.00	R 3 204.99	41.26	45%	50%	2%	2%	0%
Midvaal Local Municipality	1.00	R 1 653.82	45.60	33%	64%	1%	2%	0%
Mogale City Local Municipality	2.00	R 3 626.73	39.67	64%	34%	1%	1%	1%
Randfontein Local Municipality	1.00	R 1 747.14	40.32	61%	27%	2%	10%	2%
Westonaria Local Municipality	2.00	R 2 689.54	37.89	69%	29%	1%	1%	0%