Customer Segmentation and Targeting using Cluster Analysis and Linear Discriminant Analysis

Chukwuekwu Musei

1.0 Introduction

1.1 Background

Market segmentation is a major area of interest within the field of Marketing. It helps companies to identify and focus on a niche within the market where it can easily compete. Historically, marketing strategies concentrated on mass marketing, which did not yield much result because some of the heterogeneous behaviour and preference of customers were not met (Palmatier and Crecelius, 2019). Today, in the era of big data, it has become even more difficult as large volume and variety of customer data continue to emerge (Erevelles *et al*., 2015), due to technological innovations such as social media. As such, there is the need for companies to adopt advanced marketing techniques such as cluster analysis that can discover hidden insights in data (Erevelles *et al*., 2015), for an appropriate segmentation strategy that meets the heterogeneous needs of customers in order to remain competitive in the ever-dynamic market (Wijekoon *et al*., 2021).

The purpose of this paper is to solve customer heterogeneity problem of a chain restaurant company in its expansion plan, using cluster analysis to segment the market as well as linear discriminant analysis and classification technique to target the best customer segment based on revenue, profit and other descriptor variables. The paper has been organised as follows: Session 2 will discuss the methods for carrying out the task. The interpretation of results and findings will be discussed in session 3. The paper concludes in session 4 by identifying the main implication of the task in practice and theory, including the limitations associated with it.

1.2 Related Work

1.2.1 Big Data Analytics and Marketing Analytics

In the field of marketing, several definition of big data analytics have been found. Chen *et al*. (2012) define it as huge datasets that are unstructured in nature, which require innovative technologies for effective storage, analysis and visualization, as well as discovering the behavioral pattern of customers to aid marketing strategies (Erevelles *et al*., 2015).

According to Hetrakul and Cirillo (2014), companies who considered customer heterogeneity in its marketing segmentation strategy gained about 20% growth in revenue, contrary to some who did not (Mithas *et al*., 2013). For instance, Amazon has continued to generate profit yearly (Statista 2022; Lilien *et al*., 2013), using collaborative filtering, an analytics method (*Lilien et al*., 2013) deployed in marketing that enables it to track customer behaviour and preference when they shop online. It can be claimed that when companies adopt advanced analytics method in marketing decisions, it should first, consider the right customers and at the same time recognizing its strengths and weaknesses that will enable it to take on new opportunities while recognizing the external threat.

2.0 Methodology

The Cross Industry Standard Procedure for Data Mining (CRISP-DM) framework was used to carry out this task. This method was adopted because it follows a sequence of a project's life cycle from understanding a business problem to when it is finally deployed (Schröer *et al*.,2021)(see Figure 1).

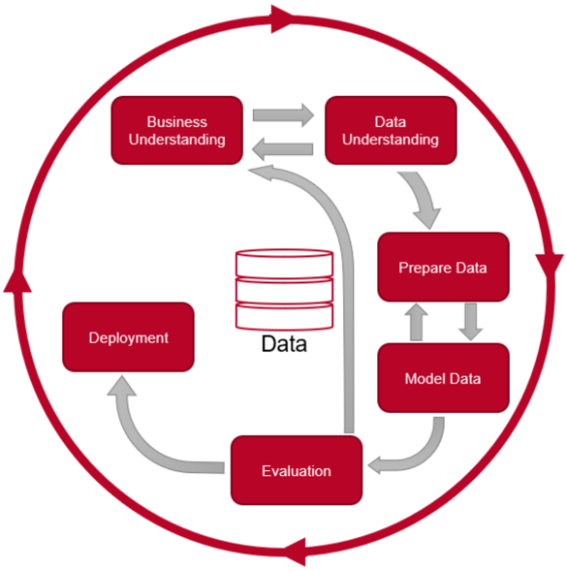


Figure 1: The CRISP-DM stages of a projects lifecycle Source: Schröer *et al*., 2021

The exploratory data analysis (EDA) of the restaurant dataset was carried out using Tableau to understand the customer characteristics, behaviour and attributes. First, a logical sequence of segmentation, targeting and positioning (STP) was deployed because customers differ in attributes and preferences in order to develop a marketing mix for new product development (NPD) and effective customer loyalty (Xu *et al*., 2015).

Hierarchical and K-means clustering was performed in R to define customers into different homogenous groups using only the base variables. The Euclidian distance was used as the dissimilarity measure with a complete linkage. The global optimum number of k with the shortest distance withinss clusters was selected after several random numbers of k values.

Linear discriminant analysis (LDA) technique was used to classify each customer segments using descriptor variables. LDA was considered above other classification methods like logistic regression because it is best at describing which descriptor variables that help differentiate between two or more customer segments and it remains stable when the classes are separated, unlike logistics regression that becomes unstable (Hastie *et al*., 2021). After which, the number of discriminant function was obtained from minimum {(h-1), number of descriptors}, leading to the computation of the discriminant score, which was used to assign the observations to a class that has the largest discriminant score. Analysis of variance (ANOVA) was conducted to check if the discriminant function is statistically significant, thereafter, the confusion matrix was used to test for the performance of the algorithm, including details on the sensitivity and specificity of the model model.

3.0 Results and Discussions

3.1 Data Exploration and Visualisation

In Figure 1, it can be seen that the highest number of customer order size are in the region where the zip code is coloured green in the map. This will enable the manager of the restaurant company to get a hint of the likelihood of the location to set up its new market.

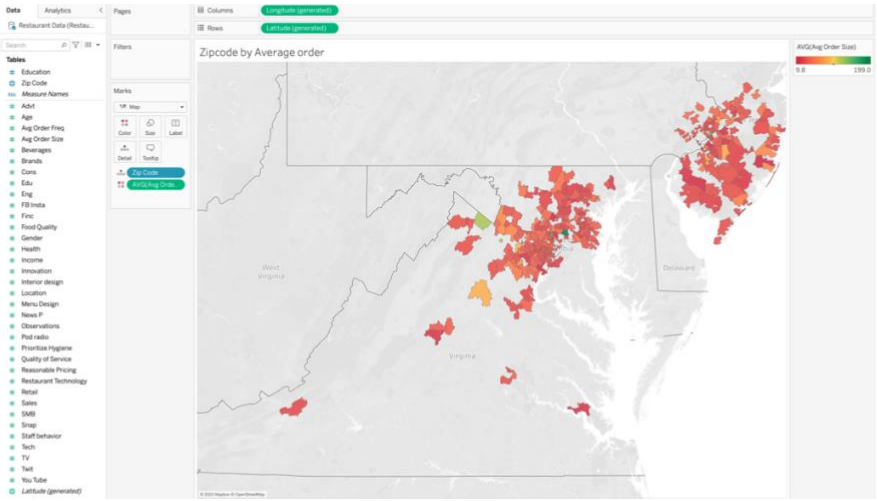
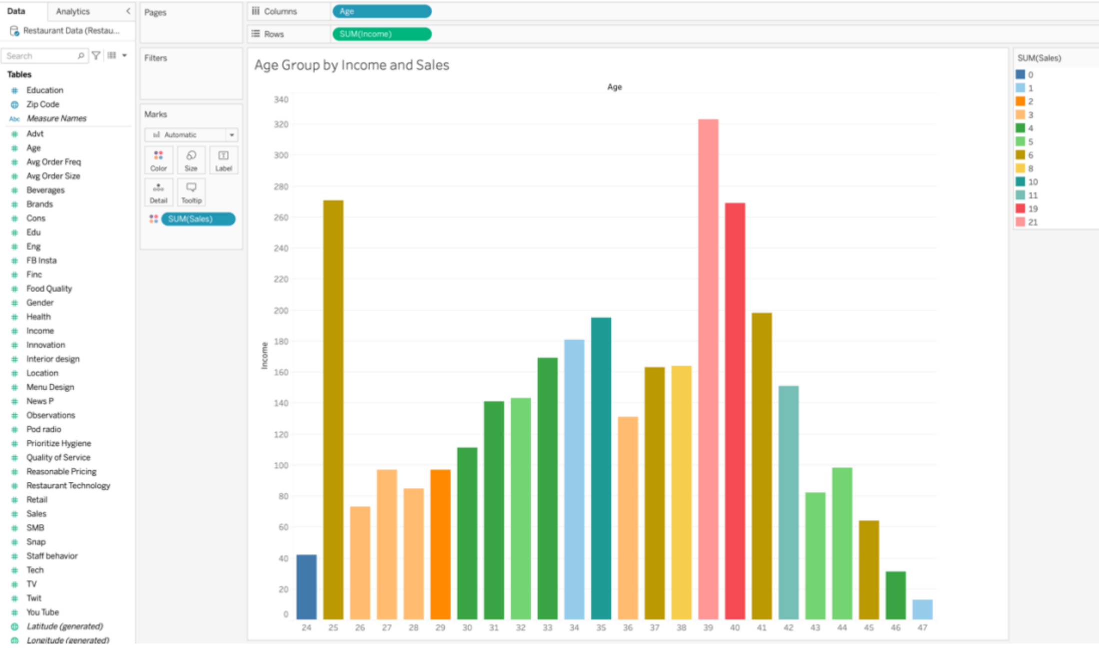


Figure 1: Average Order Size by Zip code.



Chat 1:Data exploration and visualisation

Further preliminary exploration and visualisation of the restaurant dataset in Chat 2 show that customers with an undergraduate degree place more orders than those with master's degree. It can be claimed that the low number of master's degree holders in order frequency could be that the latter is fewer compared to the former.

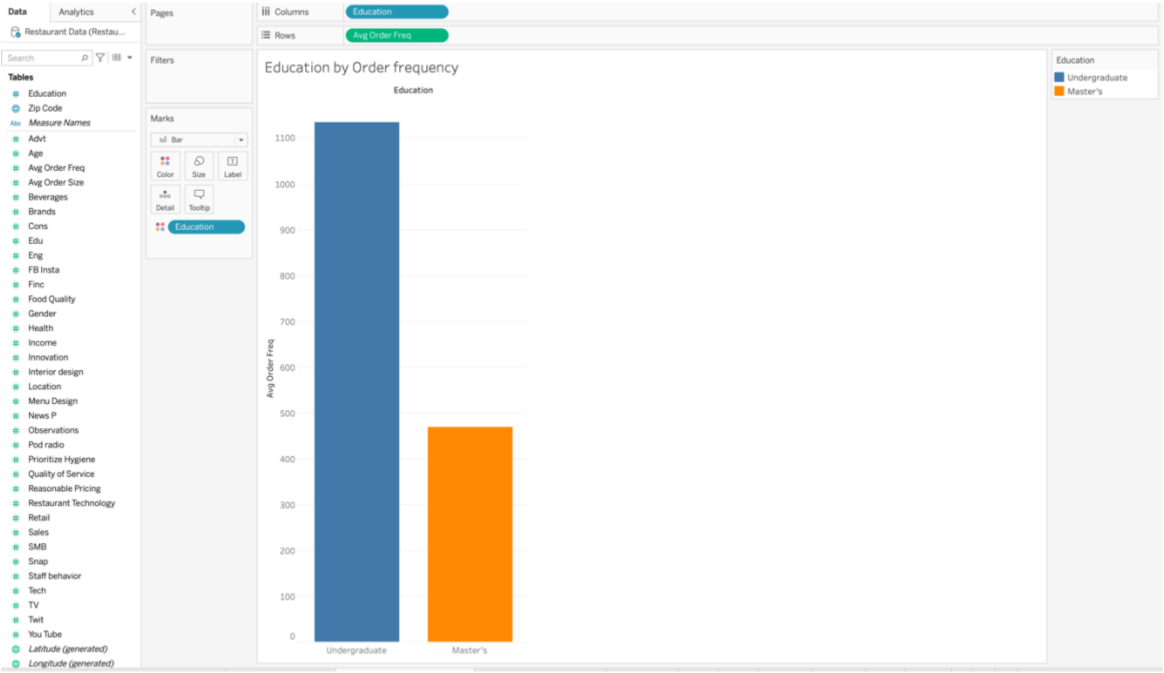


Chart 2: Visualisation showing Educational qualification by order frequency.

It can be seen in Chat 3 that customers in age bracket of 39 and 40 have remained consistent for placing orders with respect to increase in income, unlike other age group that remained inconsistent.

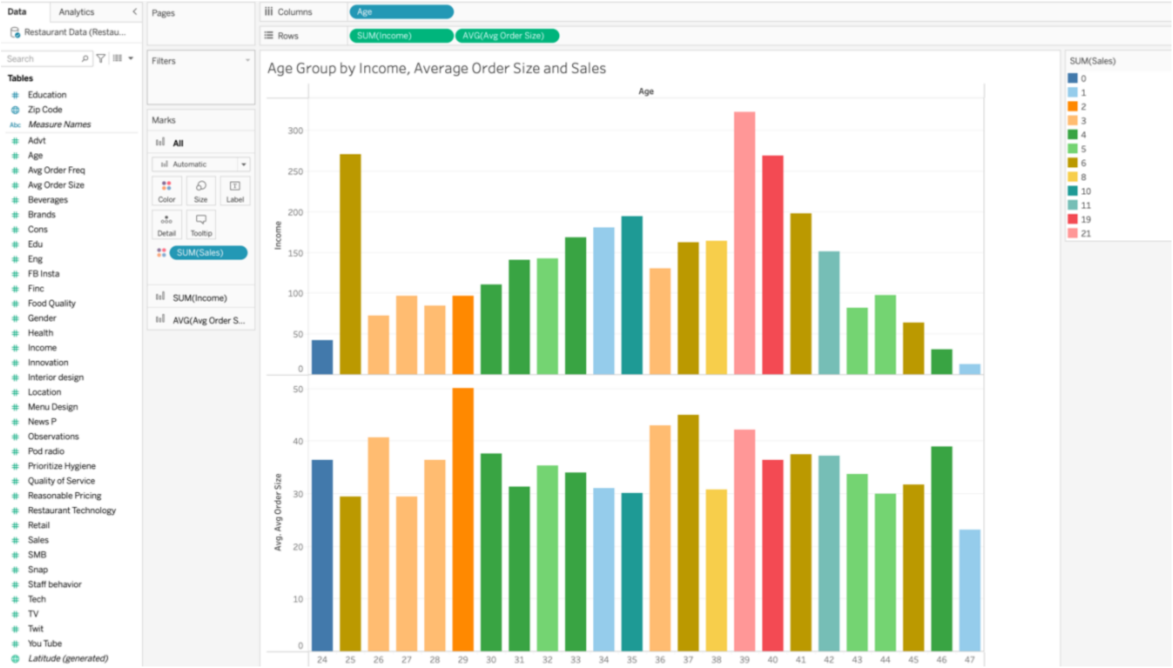


Chart 3: Age Group by Income, Average Order Size and sales

3.2 Cluster Analysis

3.2.1 Hierarchical and K-means Clustering

The results obtained from the preliminary analysis of hierarchical clustering performed on the bases variables are shown in the dendrogram and elbow plot in Figure 3 below. It can be observed that the global optimum number of k with the shortest distance withinss clusters of 20471.84 was achieved with k=3, with nstart of 30, after running several k numbers of clusters.
Additionally, after the hierarchical clustering procedure, the K-means clustering was run on the bases variables. Findings show that K-means assigned 398, 367, and 235 number of homogeneous customers to clusters 1, 2, and 3 respectively (see Appendix B).

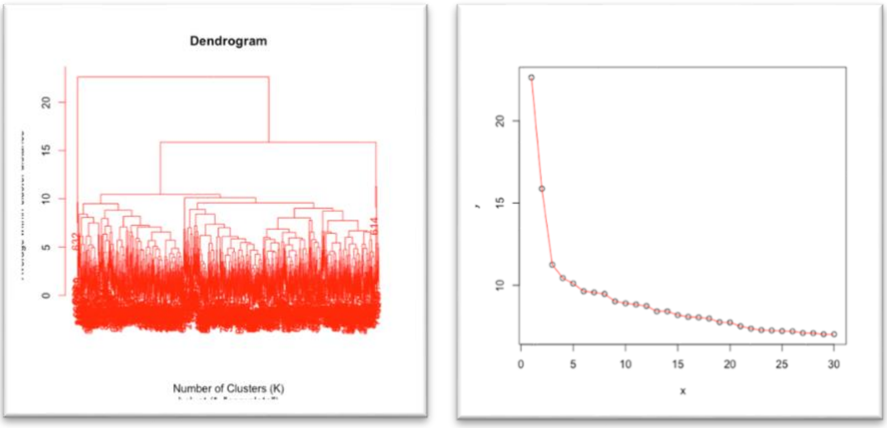
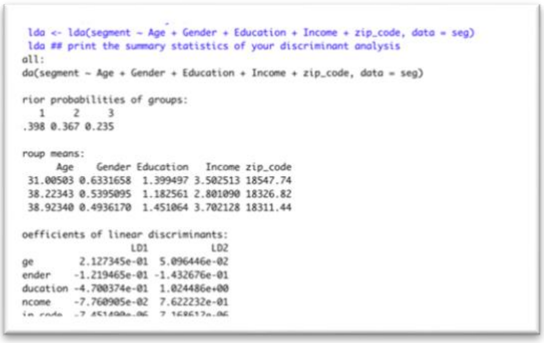


Figure 3: Diagram showing the global optimum of k with the shortest distance within clusters.

3.3 Linear Discriminant Analysis(LDA)

The result of the LDA function was carried out with the descriptor variables . It can be seen that the prior probabilities of groups shown in Table 1 below indicate that out of 1000 observations, 398 customers were assigned to cluster 1, as well as 367 and 235 observations to clusters 2 and 3 respectively.

Table 1: Result showing the output of LDA on descriptor variables.



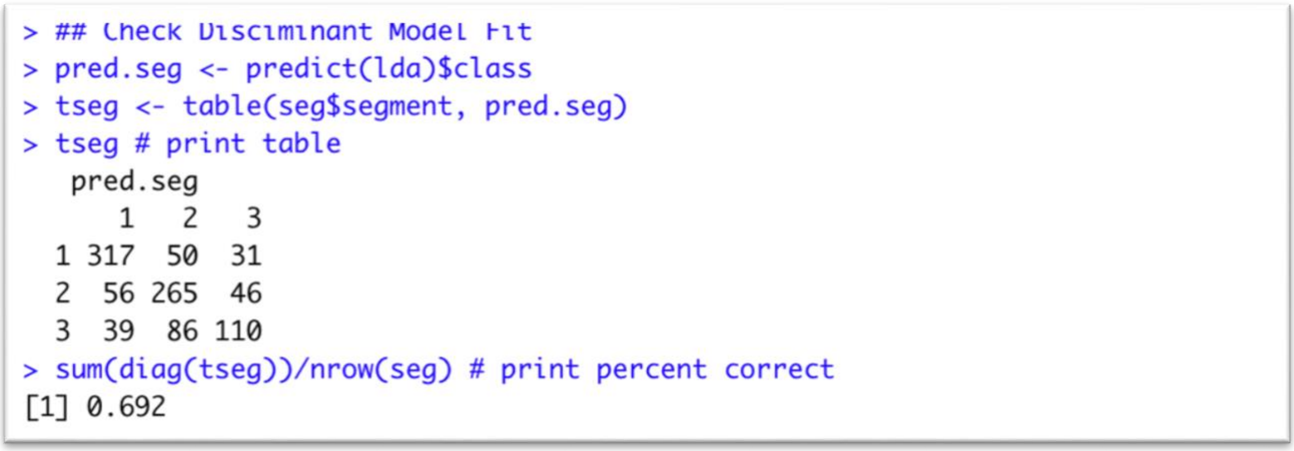
```
 lda <- lda(segment ~ Age + Gender + Education + Income + zip_code, data = seg)
 lda ## print the summary statistics of your discriminant analysis
all:
da(segment ~ Age + Gender + Education + Income + zip_code, data = seg)

rior probabilities of groups:
   1     2     3
.398 0.367 0.235

roup means:
      Age    Gender Education   Income zip_code
 31.00503 0.6331658  1.399497 3.502513 18547.74
 38.22343 0.5395095  1.182561 2.801090 18326.82
 38.92340 0.4936170  1.451064 3.702128 18311.44

oefficients of linear discriminants:
                   LD1           LD2
ge         2.127345e-01  5.096446e-02
ender     -1.219465e-01 -1.432676e-01
ducation -4.700374e-01  1.024486e+00
ncome     -7.760905e-02  7.622232e-01
```

The average age of customers in cluster 1 stands at 31 years compared to 38years and approximately 39 years of customers in clusters 2 and 3 respectively. It can be suggested that cluster 1 would be the best option for the restaurant company if their meals are mainly for those with the age bracket of 31 years. In addition, a closer inspection of Table 1 above shows that one unit increase in *Age* variable lead to an increase in the value of the discriminant score by 2.123. Further analysis also show that the discriminant functions are statistically significant when a one-way ANOVA test was carried out, leading to p-value < 0.05 (see Appendix B).

3.4 Confusion Matrix

The result of the confusion matrix used for checking the accuracy of the LDA model is shown below in Table 2. This table is quite revealing in several ways. First, the level of accuracy of the model is 69.2% . This means that out of the 1000 observation from the restaurant data, the LDA model was able to predict accurately that 317 customers belong to segment 1, as well as predicting 265 and 110 customers accurately in segments 2 and 3.

Table 2: Result showing confusion matrix of the LDA model.



```
> ## Check Discriminant Model Fit
> pred.seg <- predict(lda)$class
> tseg <- table(seg$segment, pred.seg)
> tseg # print table
   pred.seg
      1   2   3
  1 317  50  31
  2  56 265  46
  3  39  86 110
> sum(diag(tseg))/nrow(seg) # print percent correct
[1] 0.692
```

Further analysis on the multi-class confusion matrix in Table 2 can be computed for the total sensitivity and specificity of the 3 segments.

$$Sensitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative(FN)}$$

$$= \frac{TP}{TP+FN}$$

$$= \frac{317+265+110}{(317+265+110)+(50+31+56+46+39+86)} = 0.692$$

The result of 0.692 indicate that the LDA model is able to predict accurately the number of customers in each segment with an accuracy of 69.2%. However, if the sensitivity of the LDA model is small, for instance, less than 50%, it means that the model has a high tendency to assign wrongly(False Positive) the number of customers to each clusters in the case of the restaurant data. This scenario is also known as a Type 1 error in statistics.

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + False\ Positive (FP)}$$

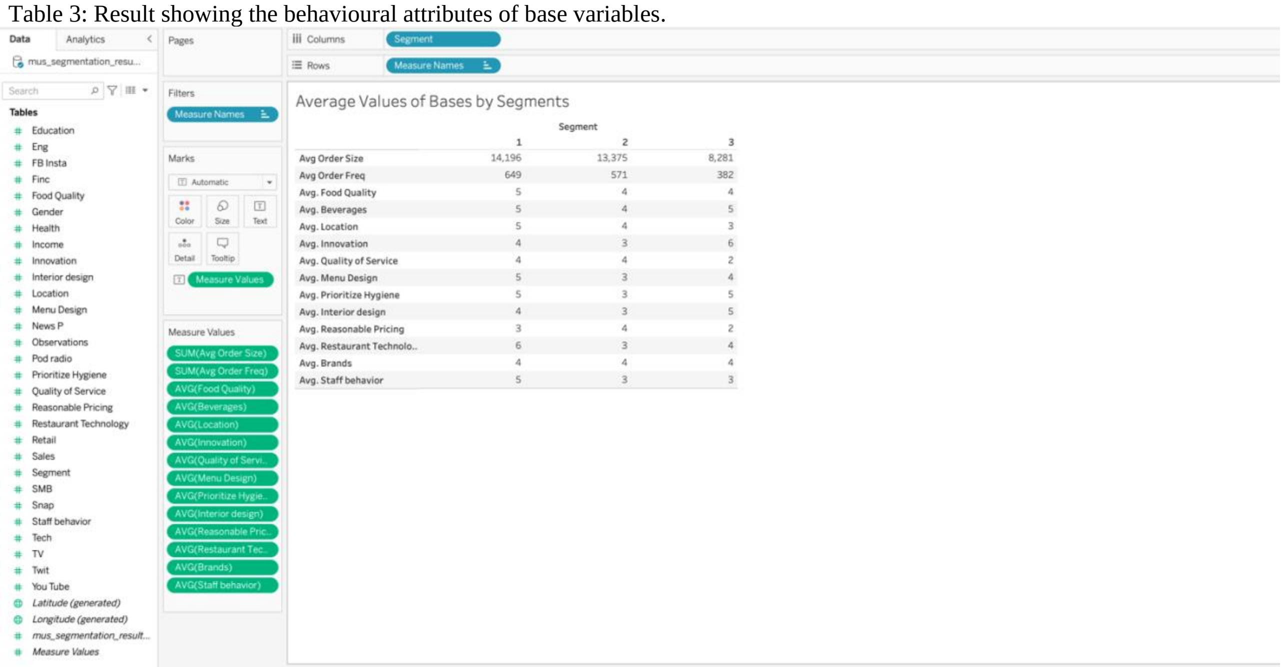
$$= \frac{TN}{TN+FP}$$

$$= \frac{39 + 265 + 31}{(39 + 265 + 31) + (317 + 31 + 56 + 46 + 86 + 110)} = 0.503$$

The result of the specificity analysis of 0.503 suggests a moderate rate at which the LDA model can predict clusters that are not attractive (i.e. true negative). In a situation where the specificity of the LDA model is small, it means that resources will be spent so much on clusters that are not profitable because the model falsely predicted that such segment will be attractive for the restaurant's proposed plan to target a segment in a new market. This scenario is also known as a Type 2 error. The sensitivity and specificity of individual segment in the multiclass confusion matrix can be seen in Appendix C.

3.5 Behavioural Components of Base Variables in Tableau.

Table 3: Result showing the behavioural attributes of base variables.



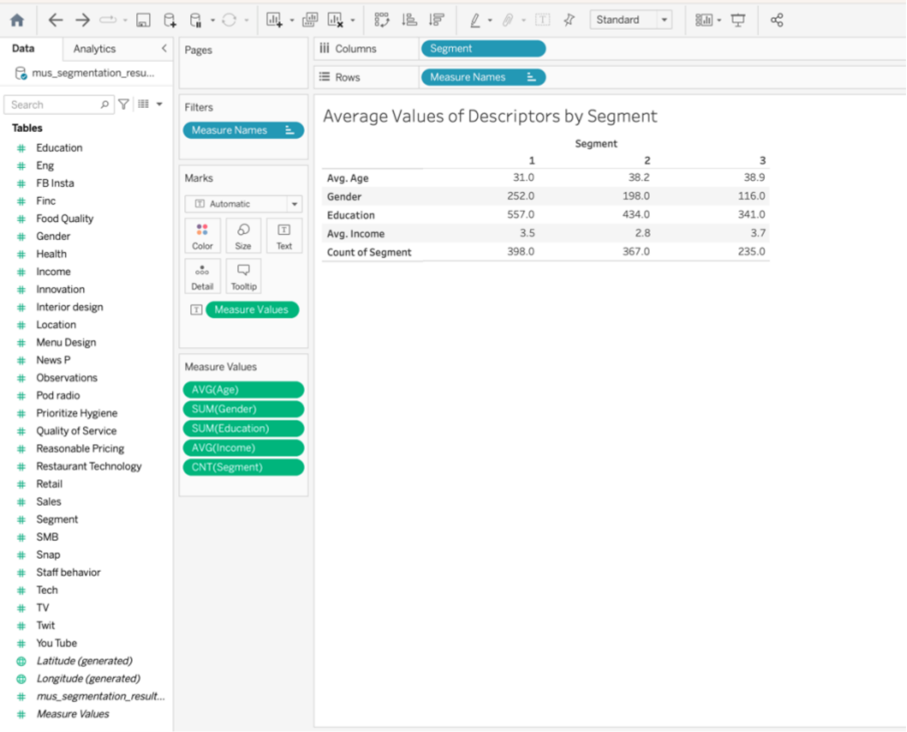
Average Values of Bases by Segments

	Segment		
	1	2	3
Avg Order Size	14,196	13,375	8,281
Avg Order Freq	649	571	382
Avg. Food Quality	5	4	4
Avg. Beverages	5	4	5
Avg. Location	5	4	3
Avg. Innovation	4	3	6
Avg. Quality of Service	4	4	2
Avg. Menu Design	5	3	4
Avg. Prioritize Hygiene	5	3	5
Avg. Interior design	4	3	5
Avg. Reasonable Pricing	3	4	2
Avg. Restaurant Technolo..	6	3	4
Avg. Brands	4	4	4
Avg. Staff behavior	5	3	3

Table 3 presents the summary statistics of the base variables. In terms of average order size, and quality of service, cluster 1 can be seen to have more prospect than clusters 2 and 3. It can be claimed that customers in cluster 1 are particular about the quality of food and service, which may indicate the willingness to pay without reasonable pricing.

3.6 Demographic Characteristics of Descriptor Variables in Tableau

Table 4: Result showing the demographic and geographic characteristics variables.



Columns: Segment

Rows: Measure Names

Filters: Measure Names

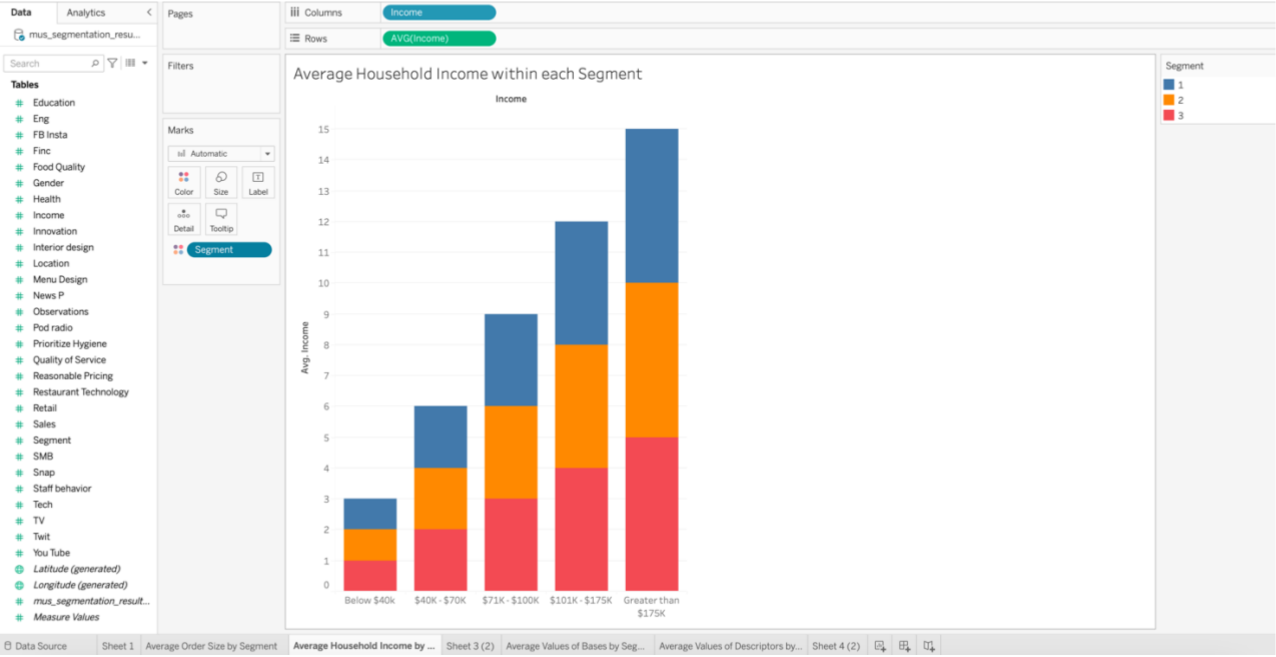
Marks: Measure Values

Measure Values: AVG(Age), SUM(Gender), SUM(Education), AVG(Income), CNT(Segment)

Average Values of Descriptors by Segment

	Segment		
	1	2	3
Avg. Age	31.0	38.2	38.9
Gender	252.0	198.0	116.0
Education	557.0	434.0	341.0
Avg. Income	3.5	2.8	3.7
Count of Segment	398.0	367.0	235.0

The result from Table 4 indicate that segment 1 has the highest number of customers that are educated. This insight is appealing because in Chat 4, segment 1 belongs to the customers that receive the highest income.



Columns: Income

Rows: AVG(Income)

Marks: Segment

Chat 4: Household income with by segments

3.7 Targeting

Results from preliminary exploration, visualisation, segmentation and LDA analysis indicate that segment 1 is the most attractive segment to target because it has the highest revenue and order frequency, including the highest number of customers. Also, customers in segment 1 are high income earners who are not particular about pricing but willing to pay any price for quality brand and services. Other factors to consider for targeting may include education level and age group. These approaches recognise that customer needs differ (Wang and Seidle, 2017) and segmenting customers with similar needs and preferences (Adner, 2002) can lead to effective targeting.

4.0 Conclusions

The present study was designed to ascertain the effect of customer heterogeneity problem as a crucial factor in market segmentation, targeting and positioning, particularly on how big data analytics help to shape the appropriate decision-making based on the insights from the captured data in marketing. Clearly, the implication of this task in practice is acknowledging customer difference in terms of behaviour, attributes and preference, leading to a high return on investment for any organization. It can be argued that the success of any market segmentation strategy is not hinged alone using analytic approach to identify customer heterogeneity problem, but also the ability to recognise the right customers, the strength of the company as well as its competitors. It can be claimed that the limitation of cluster analysis stems from certain approaches to be adopted such as standardization of observations, the type of linkage and dissimilarity measure, have an impact in producing different results. This is why in practice there are various choices available to implement by looking at the best solution that can be interpreted.

References

Adner, R. (2002) ‘When are technologies disruptive? a demand-based view of the emergence of *competition*’, *Strategic Management Journal, 23(8), pp. 667–688.* doi:10.1002/smj.246

Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., & Weaven, S. (2019) ‘Market segmentation and travel choice prediction in spa hotels through TripAdvisor’s online reviews’, *International Journal of Hospitality Management,* 80, pp.52–77. https://doi/10.1016/j.ijhm.2019.01.003.

Barney, J. B. (1991) ‘Firm resources and sustained competitive advantage’, *Journal of Management*, 17(1), 99–120.

Chen, H., Chang, R. H., & Storey, V. C. (2012) ‘Business intelligence and analytics: Frombig data to big impact. *MIS Quarterly*, 36(4), pp.1165–1188.

Erevelles, S., Fukawa, N. & Swayne, L. (2015) ‘Big Data consumer analytics and the transformation of marketing’, *Journal of Business Research,* 69, pp.897–904. Available at: http://dx.doi.org/10.1016/j.jbusres.2015.07.001.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) ‘An introduction to statistical learning. Springer Science and Business Media, New York.

Hetrakul, P., & Cirillo, C. (2014) ‘Customer heterogeneity in revenue management for railway services’, *Journal of Revenue and Pricing Management, 14(1), pp. 28–49. Available at:* https://doi/10.1057/rpm.2014.27

Lee, R. P., & Grewal, R. (2004) ‘Strategic responses to new technologies and their impact on firm performance’, *The Journal of Marketing*, 68(4), pp. 157–171.

Lilien, G. L., Roberts, J. H., & Shankar, V. (2013) ‘Effective marketing science applications: insights from the ISMS-MSI practice prize finalist papers and projects*’, Marketing Science*, 32(2), pp. 229–245.

Mithas, S., Lee, M. R., Earley, S., & Murugesan, S. (2013) ‘Leveraging big data and business analytics’, *IT Professional*, 15(6), 18–20.

Palmatier, R. W., and A. T. Crecelius. (2019), ‘The “first principles” of marketing strategy. *AMS Review* 9(1-2), pp. 5-26. https://doi.org/10.1007/s13162-019-00134-y

Schröer, C., Kruse, F., & Gómez, J. M. (2021) ‘A systematic literature review on applying CRISP-DM process model*’, Procedia Computer Science*, 181, pp. 526–534. https://doi/10.1016/j.procs.2021.01.199.

Statista (2022) ‘Amazon: statistics and facts’. Available at: https://www.statista.com/topics/846/amazon/#topicOverview (Accessed: 20 February 2023)

Wang, I. K., & Seidle, R. (2017) ‘The degree of technological innovation: a demand heterogeneity perspective’, *Technological Forecasting and Social Change,* 125, pp. 166–177. https://doi/10.1016/j.techfore.2017.07

Wijekoon, A., Salunke, S., & Athaide, G. A. (2021) ‘Customer heterogeneity and innovation-based competitive strategy: a review, synthesis, and research agenda, *Journal of Product Innovation Management,* 38(3), pp.315–333. https://doi/10.1111/jpim.12576.

Xu, Z., Gary, Frankwick, G.L. & Ramirez, E. (2015) Effects of big data analytics and traditional marketing analytics on new product success: a knowledge fusion perspective, *Journal of Business Research* 69, pp. 1562–1566. Available at: http://dx.doi.org/10.1016/j.jbusres.2015.10.017.

Zhang, H., & Xiao, Y. (2020) ‘Customer involvement in big data analytics and its impact on B2B innovation’, *Industrial Marketing Management,* 86(1), pp. 99–108. https://doi/10.1016/j.indmarman.2019.02.020

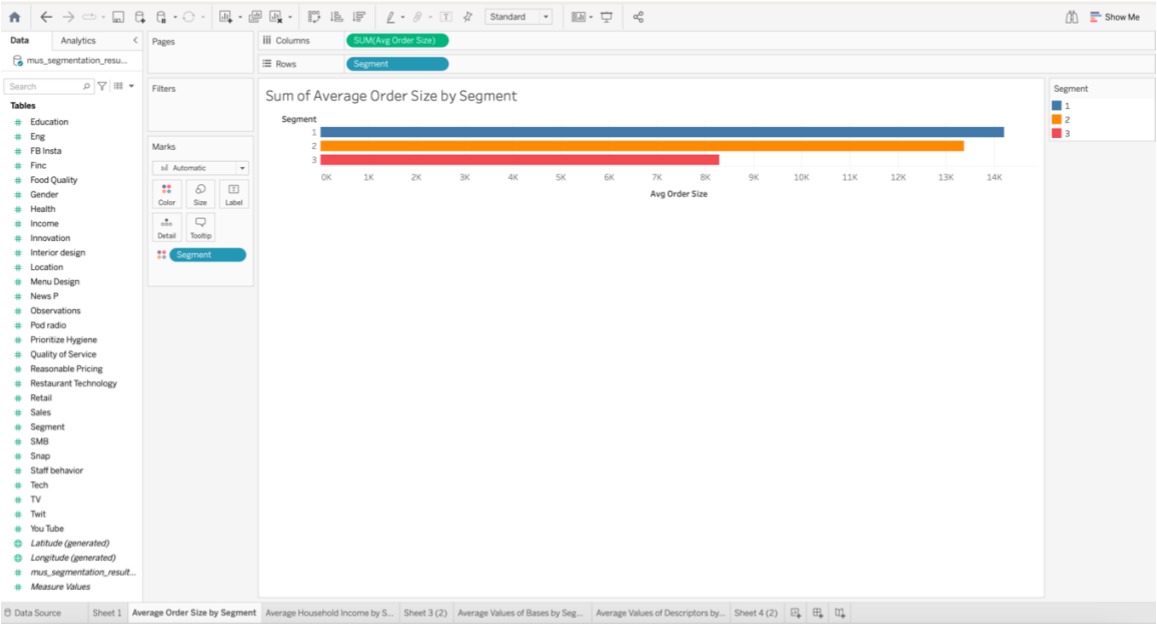
Zhou, K. Z., Brown, J. R., & Dev, C. S. (2009) ‘Market orientation, competitive advantage, and performance: a demand-based perspective, *Journal of Business Research,* 62(11), 1063–1070. https://doi/10.1016/j.jbusres.2008.10.001.

R Code

```
####################################################
## Cluster Analysis in R #
####################################################

set.seed(1845)

library(readxl)


#Loading in the Restaurant data
musei <- read.csv("Restaurant data.csv")


#Run hierarchical clustering with bases variables
mus_hclust <- hclust(dist(scale(cbind(musei$Food_Quality, musei$Beverages,
musei$Location, musei$innovation, musei$Quality_of_Service,
                    musei$Menu_Design, musei$Prioritize_Hygiene,
musei$Interior_design, musei$Reasonable_Pricing,
                    musei$Restaurant_Technology, musei$Brands, musei$Staff_behavior,
musei$avg_order_size, musei$avg_order_freq))), method="complete")



# Elbow plot for first 30 segments
x <- c(1:30)
sort_height <- sort(mus_hclust$height,decreasing=TRUE)
y <- sort_height[1:30]
plot(x,y) ; lines(x,y, col= "red")
plot(mus_hclust, main = "Dendrogram", xlab = "Number of Clusters (K)", y="Average
within cluster distance", col="red")

# Run k-means with 3 segments
mus_kmeans <- kmeans(x = data.frame(musei$Food_Quality, musei$Beverages,
musei$Location, musei$innovation, musei$Quality_of_Service,
                   musei$Menu_Design, musei$Prioritize_Hygiene,
musei$Interior_design, musei$Reasonable_Pricing,
                   musei$Restaurant_Technology, musei$Brands, musei$Staff_behavior),
3)


segment = mus_kmeans$cluster
mus_kmeans
mus_kmeans$tot.withinss


# Add segment number back to original data
mus_segmentation_result4 <- cbind(musei, segment)

# Export data to a CSV file
```

```
write.csv(mus_segmentation_result4, file = file.choose(new=TRUE), row.names = FALSE) ## Name file mus_segmentation_result.csv


###################################################
## Discriminant Analysis and Classification in R #
###################################################


set.seed(1845)
library(MASS)

## Read in Segment Data and Classification Data
seg <- read.csv(file.choose()) ## Choose mus_segmentation_result4.csv file
class <- read.csv(file.choose()) ## Choose classification_Data.csv file

## Run Discriminant Analysis
lda <- lda(segment ~ Age + Gender + Education + Income + zip_code, data = seg)
lda ## print the summary statistics of your discriminant analysis

## Check which Discriminant Functions are Significant
ldaPred <- predict(lda, seg)
ld <- ldaPred$x
anova(lm(ld[,1] ~ seg$segment))
anova(lm(ld[,2] ~ seg$segment))


## Check Disciminant Model Fit
pred.seg <- predict(lda)$class
tseg <- table(seg$segment, pred.seg)
tseg # print table
sum(diag(tseg))/nrow(seg) # print percent correct

## Run Classification Using Discriminant Function


pred.class <- predict(lda, class)$class
tclass <- table(pred.class)
tclass # print table

## Add Predicted Segment to Classification Data
class.seg <- cbind(class, pred.class)
write.csv(class.seg, file = file.choose(new=TRUE), row.names = FALSE) ## Name file mus_classification_pred.csv
```

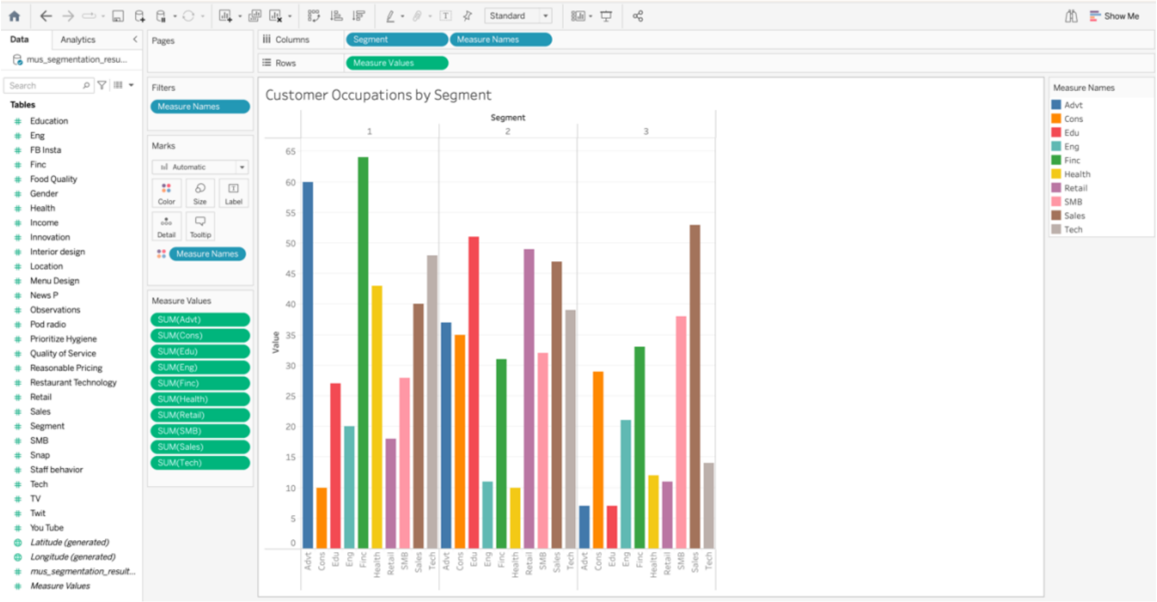
Appendix A: Preliminary exploration and visualization of Restaurant Dataset



Columns: SUM(Avg Order Size)

Rows: Segment

Sum of Average Order Size by Segment



Columns: Segment, Measure Names

Rows: Measure Values

Filters: Measure Names

Measure Values: SUM(Advt), SUM(Cons), SUM(Edu), SUM(Eng), SUM(Finc), SUM(Health), SUM(Retail), SUM(SMB), SUM(Sales), SUM(Tech)

Customer Occupations by Segment

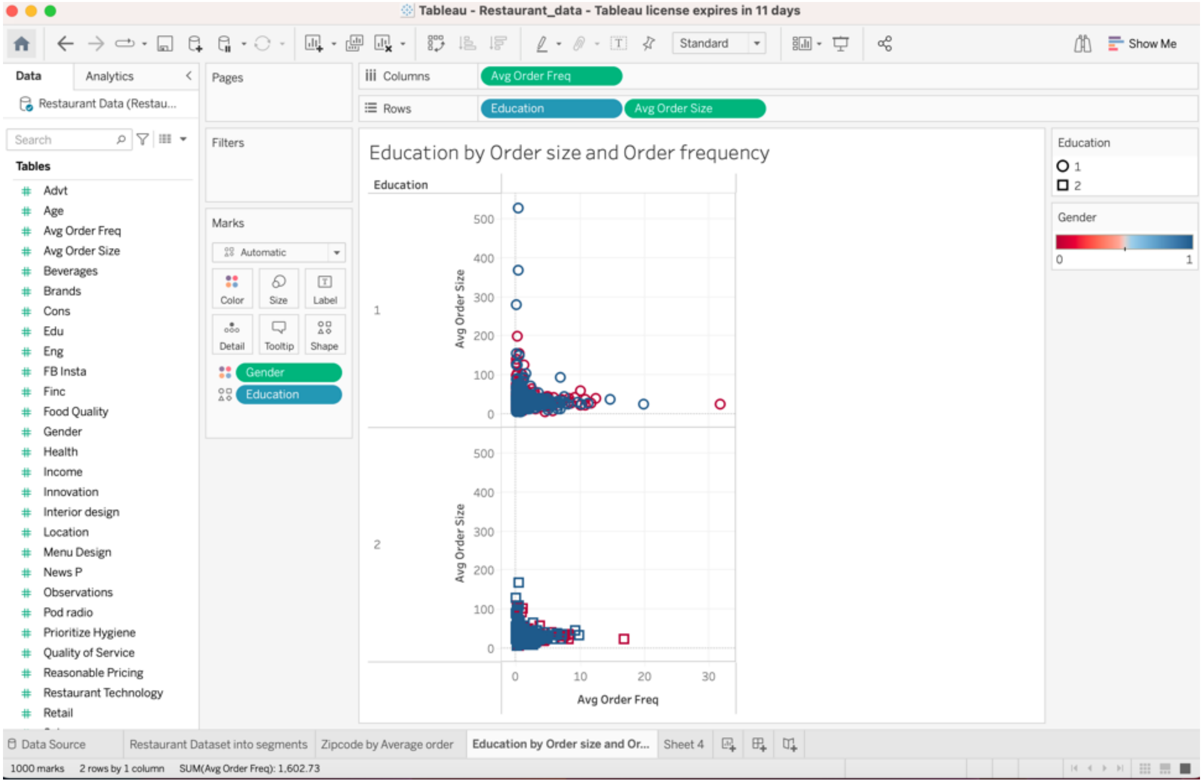


Tableau - Restaurant_data - Tableau license expires in 11 days
Standard
Show Me
Data
Analytics
Restaurant Data (Restau...
Search
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Cons
Edu
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Innovation
Interior design
Location
Menu Design
News P
Observations
Pod radio
Prioritize Hygiene
Quality of Service
Reasonable Pricing
Restaurant Technology
Retail
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
Gender
Education
Columns
Avg Order Freq
Rows
Education
Avg Order Size
Education by Order size and Order frequency
Education
1
2
Avg Order Size
500
400
300
200
100
0
0
10
20
30
Avg Order Freq
Education
1
2
Gender
0
1
Data Source
Restaurant Dataset into segments
Zipcode by Average order
Education by Order size and Or...
Sheet 4
1000 marks
2 rows by 1 column
SUM(Avg Order Freq): 1,602.73

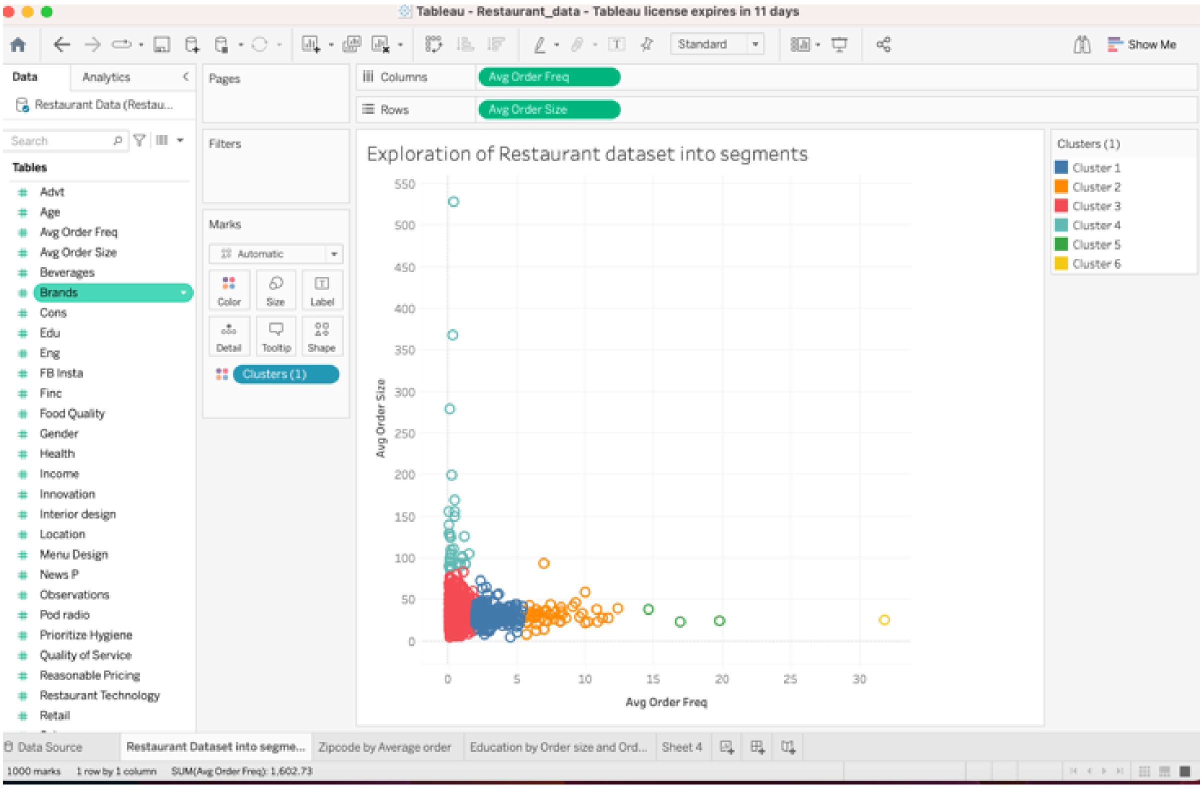
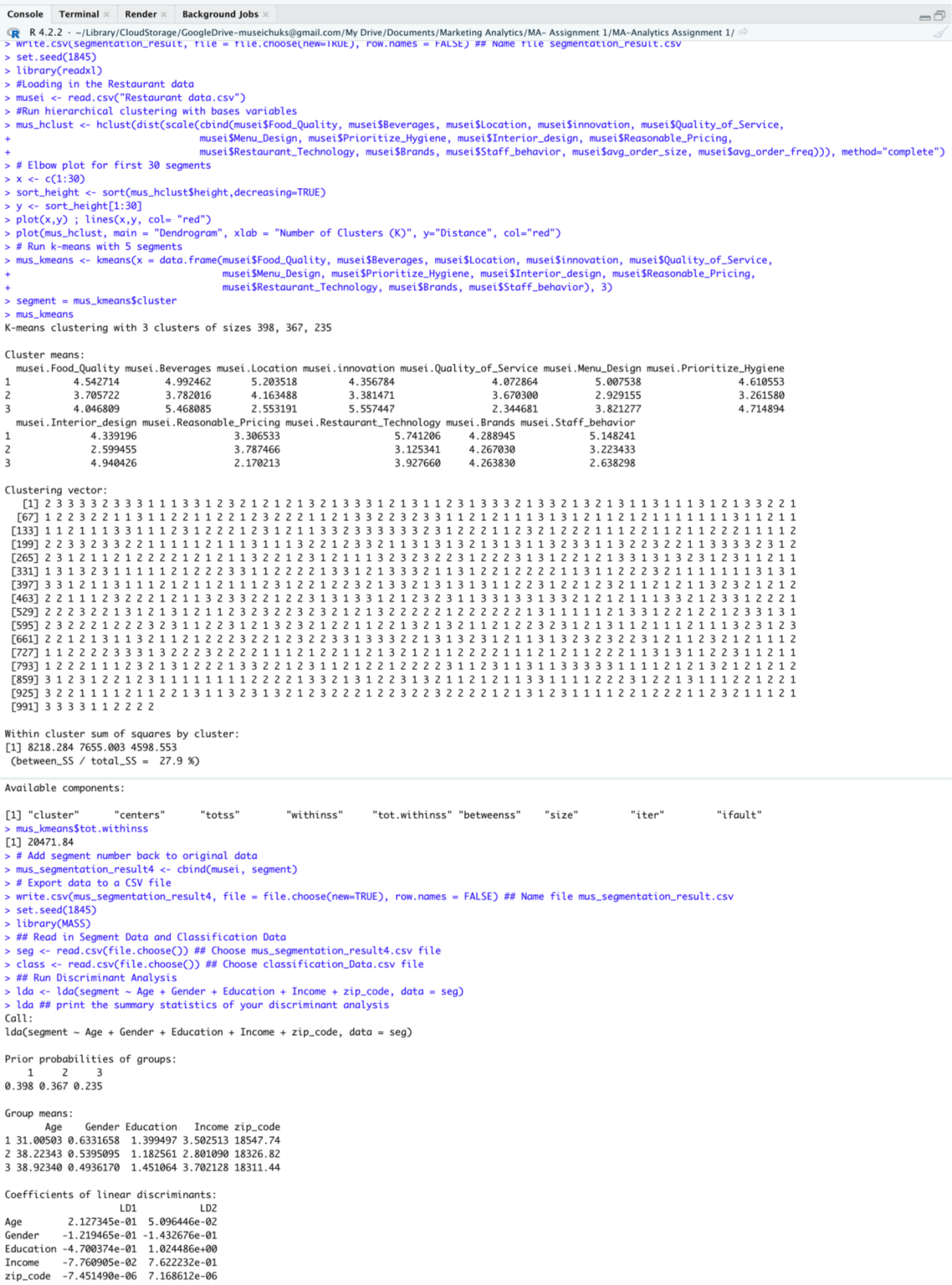
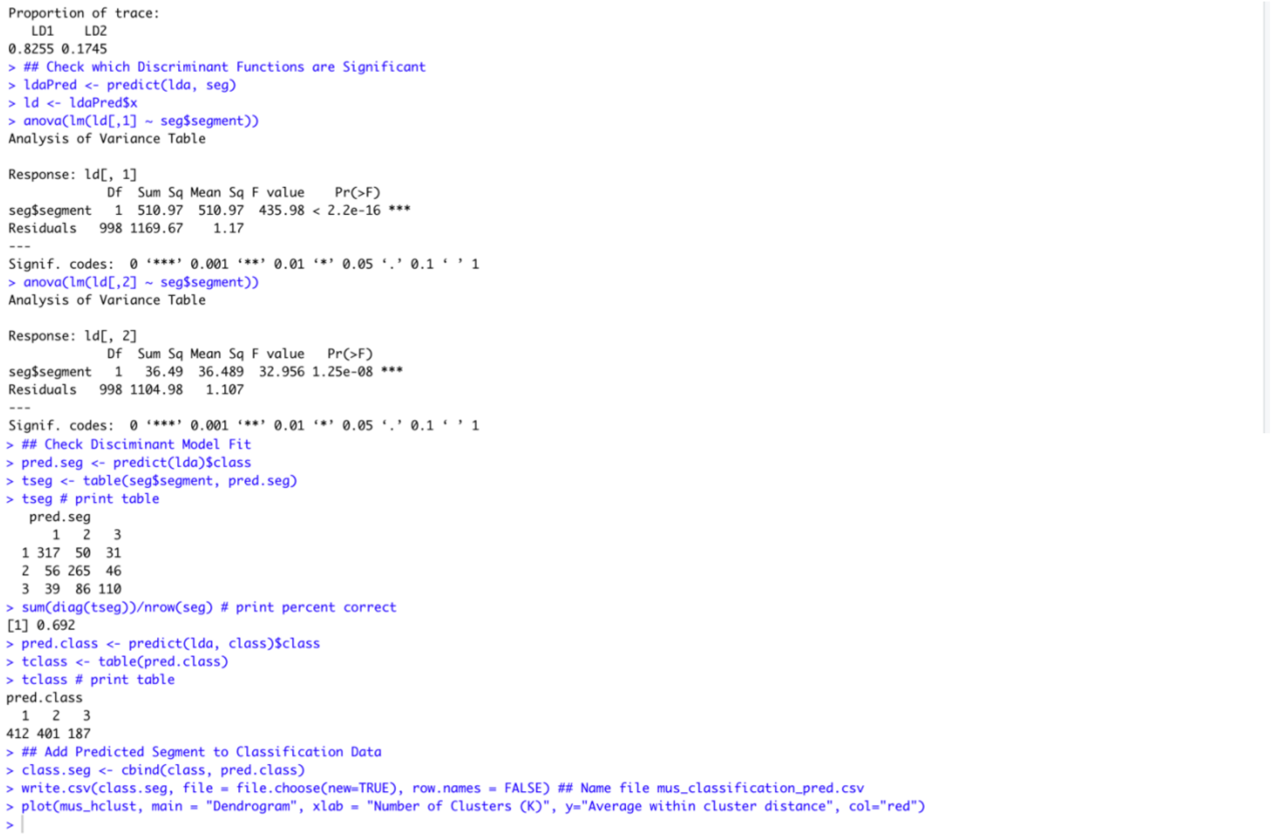


Tableau - Restaurant_data - Tableau license expires in 11 days
Standard
Show Me
Data
Analytics
Restaurant Data (Restau...
Search
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Cons
Edu
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Innovation
Interior design
Location
Menu Design
News P
Observations
Pod radio
Prioritize Hygiene
Quality of Service
Reasonable Pricing
Restaurant Technology
Retail
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
Clusters (1)
Columns
Avg Order Freq
Rows
Avg Order Size
Exploration of Restaurant dataset into segments
Avg Order Size
550
500
450
400
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30
Avg Order Freq
Clusters (1)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Data Source
Restaurant Dataset into segme...
Zipcode by Average order
Education by Order size and Ord...
Sheet 4
1000 marks
1 row by 1 column
SUM(Avg Order Freq): 1,602.73

Appendix B: R Code Outputs



```
R 4.2.2 · ~/Library/CloudStorage/GoogleDrive-museichuks@gmail.com/My Drive/Documents/Marketing Analytics/MA- Assignment 1/MA-Analytics Assignment 1/
> write.csv(segmentation_result, file = file.choose(new=TRUE), row.names = FALSE) ## Name file segmentation_result.csv
> set.seed(1845)
> library(readxl)
> #Loading in the Restaurant data
> musei <- read.csv("Restaurant data.csv")
> #Run hierarchical clustering with bases variables
> mus_hclust <- hclust(dist(scale(cbind(musei$Food_Quality, musei$Beverages, musei$Location, musei$innovation, musei$Quality_of_Service,
+                                 musei$Menu_Design, musei$Prioritize_Hygiene, musei$Interior_design, musei$Reasonable_Pricing,
+                                 musei$Restaurant_Technology, musei$Brands, musei$Staff_behavior, musei$avg_order_size, musei$avg_order_freq))), method="complete")
> # Elbow plot for first 30 segments
> x <- c(1:30)
> sort_height <- sort(mus_hclust$height,decreasing=TRUE)
> y <- sort_height[1:30]
> plot(x,y) ; lines(x,y, col= "red")
> plot(mus_hclust, main = "Dendrogram", xlab = "Number of Clusters (K)", y="Distance", col="red")
> # Run k-means with 5 segments
> mus_kmeans <- kmeans(x = data.frame(musei$Food_Quality, musei$Beverages, musei$Location, musei$innovation, musei$Quality_of_Service,
+                                     musei$Menu_Design, musei$Prioritize_Hygiene, musei$Interior_design, musei$Reasonable_Pricing,
+                                     musei$Restaurant_Technology, musei$Brands, musei$Staff_behavior), 3)
> segment = mus_kmeans$cluster
> mus_kmeans
K-means clustering with 3 clusters of sizes 398, 367, 235

Cluster means:
  musei.Food_Quality musei.Beverages musei.Location musei.innovation musei.Quality_of_Service musei.Menu_Design musei.Prioritize_Hygiene
1           4.542714        4.992462       5.203518         4.356784                 4.072864          5.007538                 4.610553
2           3.705722        3.782016       4.163488         3.381471                 3.670300          2.929155                 3.261580
3           4.046809        5.468085       2.553191         5.557447                 2.344681          3.821277                 4.714894
  musei.Interior_design musei.Reasonable_Pricing musei.Restaurant_Technology musei.Brands musei.Staff_behavior
1              4.339196                 3.306533                    5.741206     4.288945             5.148241
2              2.599455                 3.787466                    3.125341     4.267030             3.223433
3              4.940426                 2.170213                    3.927660     4.263830             2.638298

Clustering vector:
   [1] 2 3 3 3 3 2 3 3 3 1 1 1 3 3 1 2 3 2 1 2 1 2 1 3 2 1 3 3 3 1 2 1 3 1 1 2 3 1 3 3 3 2 1 3 3 2 1 3 2 1 3 1 1 3 1 1 1 3 1 2 1 3 3 2 2 1
  [67] 1 2 2 3 2 2 1 1 3 1 1 2 2 1 1 2 2 1 2 3 2 2 2 1 1 2 1 3 3 2 2 3 2 3 3 1 1 2 1 2 1 1 1 3 1 3 1 2 1 1 2 1 2 1 1 1 1 1 1 1 3 1 1 2 1 1
 [133] 1 1 2 1 1 1 3 3 1 1 1 2 3 1 2 2 2 1 2 3 1 2 1 1 3 3 2 3 3 3 3 3 3 2 3 1 2 2 2 1 1 2 3 2 1 2 2 2 1 1 1 2 2 1 1 2 1 1 2 2 2 1 1 1 1 2
 [199] 2 2 3 3 2 3 3 2 2 1 1 1 1 1 2 1 1 1 3 1 1 1 3 2 2 1 2 3 3 2 1 1 3 1 3 1 3 2 1 3 1 3 1 1 3 2 3 3 1 1 3 2 2 3 2 2 1 1 3 3 3 3 2 3 1 2
 [265] 2 3 1 2 1 1 2 1 2 2 2 2 1 2 1 2 1 1 3 2 2 1 2 3 1 2 1 1 1 3 2 3 2 3 2 2 3 1 2 2 2 3 1 3 1 2 2 1 2 1 3 3 1 3 1 3 2 3 1 2 3 1 1 2 1 1
 [331] 1 3 1 3 2 3 1 1 1 1 1 2 1 2 2 2 3 3 1 1 2 2 2 2 1 3 3 1 2 1 3 3 3 2 1 1 3 1 2 2 1 2 2 2 2 1 1 3 1 1 2 2 2 3 2 1 1 1 1 1 1 1 3 1 3 1
 [397] 3 3 1 2 1 1 3 1 1 1 2 1 2 1 1 2 1 1 1 2 3 1 2 2 1 2 2 3 2 1 3 3 2 1 3 1 3 1 3 1 1 2 2 3 1 2 2 1 2 3 2 1 1 2 1 2 1 1 3 2 3 2 1 2 1 2
 [463] 2 2 1 1 1 2 3 2 2 2 1 2 1 1 3 2 3 3 2 2 1 2 2 3 1 3 1 3 3 1 2 1 2 3 2 3 1 1 3 3 1 3 3 1 3 3 2 1 2 1 2 1 1 1 3 3 2 1 2 3 3 1 2 2 2 1
 [529] 2 2 2 3 2 2 1 3 1 2 1 3 1 2 1 1 2 3 2 3 2 2 3 2 3 2 1 2 1 3 2 2 2 2 2 1 2 2 2 2 2 2 1 3 1 1 1 1 1 2 1 3 3 1 2 2 1 2 2 1 2 3 3 1 3 1
 [595] 2 3 2 2 2 1 2 2 2 3 2 3 1 1 2 2 3 1 2 1 3 2 3 2 1 2 2 1 1 2 2 1 3 2 1 3 2 1 1 2 1 2 2 3 2 3 1 2 1 3 1 1 2 1 1 1 2 1 1 1 3 2 3 1 2 3
 [661] 2 2 1 2 1 3 1 1 3 2 1 1 2 1 2 2 2 3 2 2 1 2 3 2 2 3 3 1 3 3 3 2 2 1 3 1 3 2 3 1 2 1 1 3 1 3 2 3 2 3 2 2 3 1 2 1 1 2 3 2 1 2 1 1 1 2
 [727] 1 1 2 2 2 2 3 3 3 1 3 2 2 2 3 2 2 2 2 1 1 1 2 1 2 2 1 1 2 1 3 2 1 2 1 1 2 2 2 2 1 1 1 2 1 2 1 1 2 2 2 1 1 3 1 3 1 1 2 2 3 1 1 2 1 1
 [793] 1 2 2 2 1 1 1 2 3 2 1 3 1 2 2 2 1 3 3 2 2 1 2 3 1 1 2 1 2 2 1 2 2 2 2 3 1 1 2 3 1 1 3 1 1 3 3 3 3 3 1 1 1 1 2 1 2 1 3 2 1 2 1 2 1 2
 [859] 3 1 2 3 1 2 2 1 2 3 1 1 1 1 1 1 1 1 2 2 2 2 1 3 3 2 1 3 1 2 2 3 1 3 2 1 1 2 1 2 1 1 3 3 1 1 1 1 2 2 2 3 1 2 2 1 3 1 1 1 2 2 1 2 2 1
 [925] 3 2 2 1 1 1 1 2 1 1 2 2 1 3 1 1 3 2 3 1 3 2 1 2 3 2 2 2 1 2 2 3 2 2 3 2 2 2 2 1 2 1 3 1 2 3 1 1 1 1 2 2 1 2 2 2 1 1 2 3 2 1 1 1 2 1
 [991] 3 3 3 3 1 1 2 2 2 2

Within cluster sum of squares by cluster:
[1] 8218.284 7655.003 4598.553
 (between_SS / total_SS =  27.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
> mus_kmeans$tot.withinss
[1] 20471.84
> # Add segment number back to original data
> mus_segmentation_result4 <- cbind(musei, segment)
> # Export data to a CSV file
> write.csv(mus_segmentation_result4, file = file.choose(new=TRUE), row.names = FALSE) ## Name file mus_segmentation_result.csv
> set.seed(1845)
> library(MASS)
> ## Read in Segment Data and Classification Data
> seg <- read.csv(file.choose()) ## Choose mus_segmentation_result4.csv file
> class <- read.csv(file.choose()) ## Choose classification_Data.csv file
> ## Run Discriminant Analysis
> lda <- lda(segment ~ Age + Gender + Education + Income + zip_code, data = seg)
> lda ## print the summary statistics of your discriminant analysis
Call:
lda(segment ~ Age + Gender + Education + Income + zip_code, data = seg)

Prior probabilities of groups:
    1     2     3 
0.398 0.367 0.235 

Group means:
       Age    Gender Education   Income zip_code
1 31.00503 0.6331658  1.399497 3.502513 18547.74
2 38.22343 0.5395095  1.182561 2.801090 18326.82
3 38.92340 0.4936170  1.451064 3.702128 18311.44

Coefficients of linear discriminants:
                    LD1           LD2
Age        2.127345e-01  5.096446e-02
Gender    -1.219465e-01 -1.432676e-01
Education -4.700374e-01  1.024486e+00
Income    -7.760905e-02  7.622232e-01
zip_code  -7.451490e-06  7.168612e-06
```



```
Proportion of trace:
   LD1    LD2
0.8255 0.1745
> ## Check which Discriminant Functions are Significant
> ldaPred <- predict(lda, seg)
> ld <- ldaPred$x
> anova(lm(ld[,1] ~ seg$segment))
Analysis of Variance Table

Response: ld[, 1]
             Df  Sum Sq Mean Sq F value    Pr(>F)
seg$segment   1  510.97  510.97  435.98 < 2.2e-16 ***
Residuals   998 1169.67    1.17
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(lm(ld[,2] ~ seg$segment))
Analysis of Variance Table

Response: ld[, 2]
             Df  Sum Sq Mean Sq F value   Pr(>F)
seg$segment   1   36.49  36.489  32.956 1.25e-08 ***
Residuals   998 1104.98   1.107
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> ## Check Disciminant Model Fit
> pred.seg <- predict(lda)$class
> tseg <- table(seg$segment, pred.seg)
> tseg # print table
   pred.seg
      1   2   3
  1 317  50  31
  2  56 265  46
  3  39  86 110
> sum(diag(tseg))/nrow(seg) # print percent correct
[1] 0.692
> pred.class <- predict(lda, class)$class
> tclass <- table(pred.class)
> tclass # print table
pred.class
  1   2   3
412 401 187
> ## Add Predicted Segment to Classification Data
> class.seg <- cbind(class, pred.class)
> write.csv(class.seg, file = file.choose(new=TRUE), row.names = FALSE) ## Name file mus_classification_pred.csv
> plot(mus_hclust, main = "Dendrogram", xlab = "Number of Clusters (K)", y="Average within cluster distance", col="red")
> 
```

Appendix C: Sensitivity and Specificity for individual Segment.

Segment 1: TP = 317; FN=50+31=81; FP=56+39=95; TN=265+46+86+110=507

$$Sensitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative(FN)}$$

$$= \frac{TP}{TP+FN}$$

$$= \frac{317}{317+81} = 0.796$$

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + False\ Positive(FP)}$$

$$= \frac{TN}{TN+FP}$$

$$= \frac{507}{507+95} = 0.842$$

Segment 2: TP = 265; FN=56+46=102; FP=50+86=136; TN=317+31+39+110=497

$$Sensitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative(FN)}$$

$$= \frac{TP}{TP+FN}$$

$$= \frac{265}{265+102} = 0.722$$

Appendix C: Sensitivity and Specificity for individual Segment.

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + False\ Positive (FP)}$$

$$= \frac{TN}{TN+FP}$$

$$= \frac{497}{497 + 136} = 0.785$$

Segment 3: TP = 110; FN=39+86=125; FP=31+46=77; TN=317+50+56+265=688

$$Sensitivity = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative (FN)}$$

$$= \frac{TP}{TP+FN}$$

$$= \frac{110}{110 + 125} = 0.468$$

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + False\ Positive (FP)}$$

$$= \frac{TN}{TN+FP}$$

$$= \frac{688}{688 + 77} = 0.899$$