MARKETING ANALYTICS ASSIGNMENT-1

By:

TATHAGATA CHAKRABORTY

(MASTER OF SCIENCE, BUSINESS ANALYTICS)

STUDENT NUMBER: 40379173

MODULE CODE: MGT_7215

MODULE NAME: STATISTICS FOR BUSINESS

QUEEN'S UNIVERSITY BELFAST | MANAGEMENT SCHOOL

INDEX

INTRODUCTION AND BACKGROUND

The chain restaurant sector has experienced rapid expansion in recent years as businesses go global to satisfy consumer demand. The difficulty of meeting the diversity of consumers' requirements and tastes, meanwhile, comes with a growth in the number of clients. Customer heterogeneity is the buzzword for the variations in client desires, preferences, and behaviours that make it challenging for businesses to deliver a consistent product or service to every customer.

Due to the necessity for a wide range of services and products, consumer heterogeneity can have a substantial negative influence on organisations, resulting in decreased customer satisfaction, fewer sales, and higher costs. Also, as market competition grows, it is crucial for firms to set themselves apart from rivals by meeting the personalized demands of their clientele and forging a sustainable competitive advantage.

In recent years, using analytics to handle customer heterogeneity has grown in popularity. Large datasets and sophisticated analytics tools are now readily available, allowing organisations to analyse customer data more effectively and efficiently than before. In order for enterprises to retain a sustainable competitive advantage over its rivals and to deliver high-quality goods and services that satisfy their customers' varied needs, they must be able to address client heterogeneity.

LITERATURE REVIEW

Several studies have examined the use of analytics in addressing customer heterogeneity in the chain restaurant industry. A study by Zhang et al. (2019) used customer data analytics to segment customers based on their dining preferences and behaviours. The study found that by segmenting customers, businesses could create targeted marketing campaigns that were more effective in reaching specific customer segments.

Similarly, a study by Chen et al. (2018) used analytics to predict customer behaviour and preferences. The study found that by using customer data to predict future behaviour, businesses could make better decisions about product development and marketing campaigns.

Another study by Kim et al. (2017) used analytics to personalize menu recommendations for customers. The study found that by using customer data to recommend personalized menu items, businesses could increase customer satisfaction and loyalty.

Overall, these studies highlight the importance of analytics in addressing customer heterogeneity in the chain restaurant industry. By analysing customer data, businesses can create targeted marketing campaigns, predict customer behaviour and preferences, and personalize products and services to meet the needs of specific customer segments.

METHODOLOGY

The analysis for this market segmentation problem at hand has been carried out in four progressive and distinct phases:

- **[PHASE1]- TABLEAU EXPLORATION OF RAW DATASET:** The restaurant dataset is loaded into **tableau** to obtain an elementary understanding of the distribution of the customers. A couple of visualizations were conducted across the dataset to identify the purchase characteristic of the customers and their geographic distribution.

- **[PHASE 2]- FEATURE SELECTION:** The dataset has 14 bases variables and 22 descriptor variables related to customer behavior and feedback. **Principal component analysis** is used as a dimension reduction technique to select a subset of variables that efficiently explain significant variation in the data. By reducing the dimensionality of the data, PCA helps identify a set of predictors that can be used to group the data more efficiently and analyze patterns and relationships.

- **[PHASE 3]- SEGMENTATION OF CUSTOMERS:** To better understand customer behavior and characteristics, it's necessary to group them into clusters with similar features. Hierarchical clustering was used with four different linkage methods to determine the optimal number of clusters (k-value). Once the k-value was obtained, k-means clustering was used to create k-number of clusters, and the total within-cluster variation was calculated by using different starting points. The objective was to **minimize the total within-cluster variation** or the sum of squared Euclidean distances between the data points and the cluster center. Each customer observation was assigned to the cluster it belonged to, and this information was bound to the dataset. The goal was to have little variation within each cluster and considerable difference between clusters.

- **[PHASE 4]- PROFILING AND TARGETING CUSTOMERS:** Once customer segments are identified through clustering, a chain restaurant may use **Linear Discriminant Analysis** (LDA) to predict which segment a new customer belongs to. LDA is a supervised classification algorithm used when there are multiple response classes. It identifies a linear boundary that maximizes the distance between class means and minimizes within-class variance. The LDA model's predictive accuracy can be tested to assign classes to prospective customers based on their demographic and geographic features, which can help the restaurant make changes to cater to customer demands.

ANALYSIS AND RESULTS

[PHASE 1]

Tableau visualization of raw dataset.

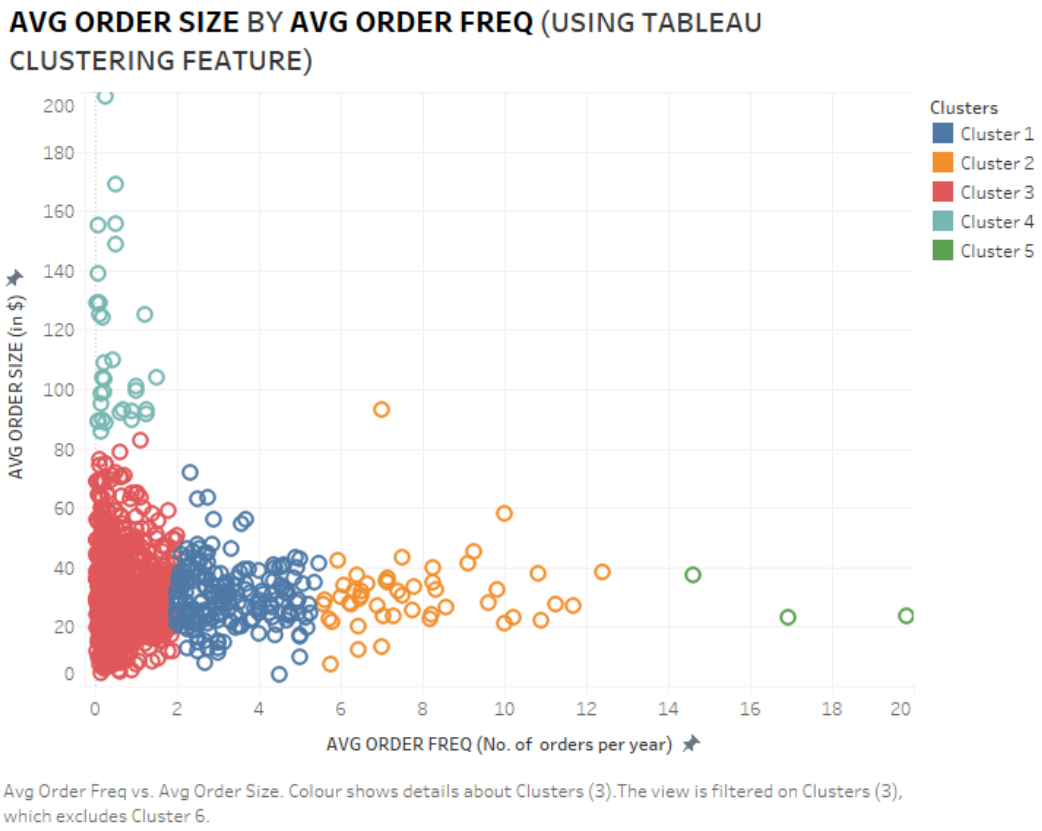


Figure 1

Average Order Size was plotted against **Average Order Frequency** in **figure 1** and the distribution was broken down into 5 clusters using the "**cluster**" feature of Tableau. The axis was scaled to represent the actual distribution of the data. The average order size was restricted to 200, to avoid the influence of outliers.

Insights:

- It can be inferred from this graph that customers with high order frequencies have low transaction amount per transaction (average order size).
- So broadly there could be a set of – "**less frequent high value customers**" and "**more frequent low value customers**".

The average income brackets of male and female customers were visualized on a symbol map (Figure 2) by applying **a calculated field and a gender filter**. This analysis provides insights into the income distribution among male and female customers, which can be used to tailor business strategies to meet the needs of this customer segment.

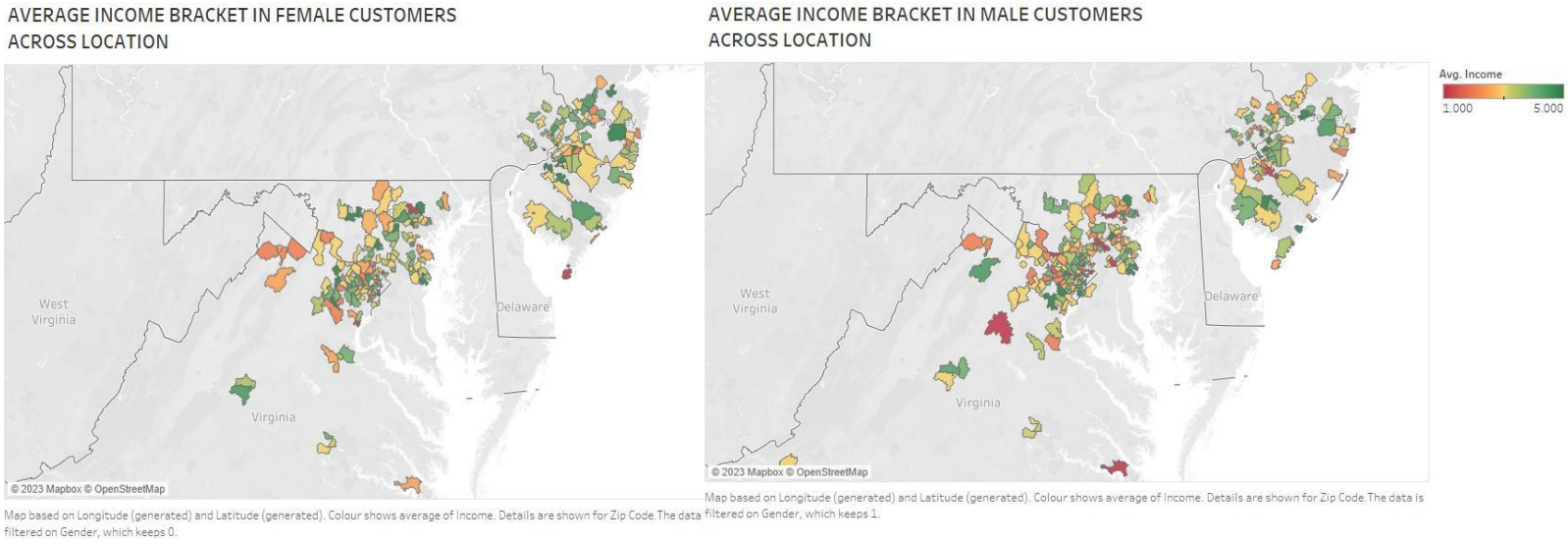


Figure : 2

Insights:

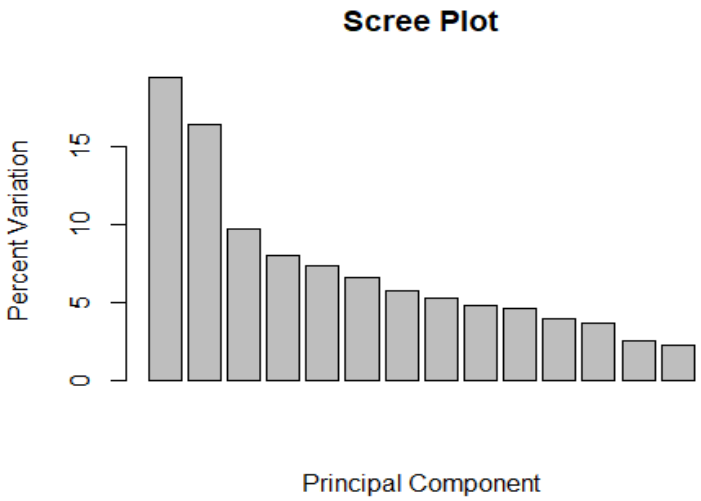
- For income levels less than 3 there are **more female customers** than male.
- Considering the orders placed from **New Jersey area**, there have been **more female** customers than male customers in that region.
- The proportion of **male customers** belonging to an **income bracket of 4 or above** is much **more than that of female** customers.

[PHASE 2]

Subset Selection using Principal Component Analysis

PCA is used to simplify the process and identify the variables that explain a significant proportion of the variance in the dataset.

The adjacent scree plot and the variance proportion output below, shows that the first four principal components account for more than 50% of the variance.




```
>pca.var.per      #output for proportion of variance explained by each component
[1] 19.4 16.4  9.7  8.0  7.3  6.6  5.7  5.2  4.8  4.6  3.9  3.6  2.5  2.2
```

Upon analyzing the weighted averages of the variables across each of the four principal components, **we were able to settle on a list of 11 bases variables** as our feature for furthering our classification model:

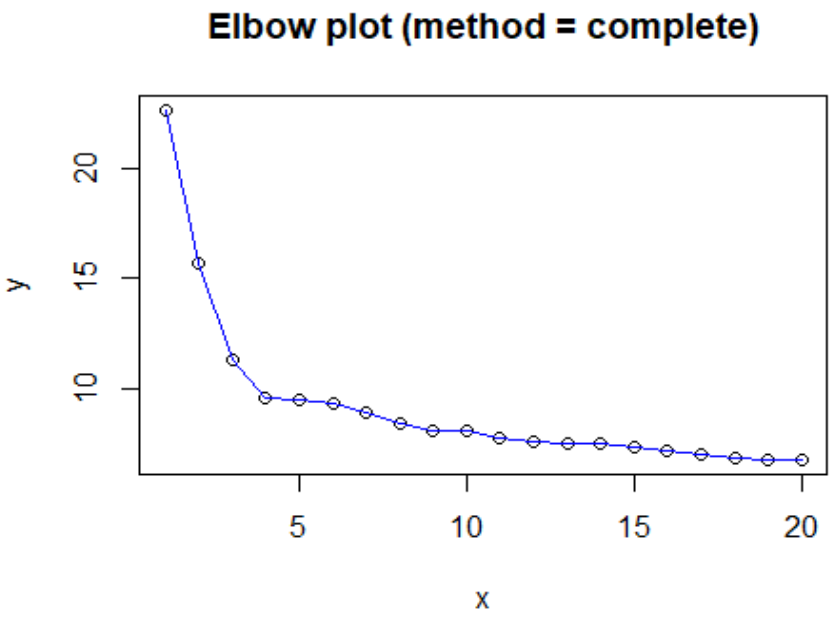
"**Food_Quality, Beverages, Location, innovation, Quality_of_Service, Menu_Design, Prioritize_Hygiene, Interior_design, Staff_behavior, avg_order_size, avg_order_freq**".

[PHASE 3]

Finding Optimal K-value using Hierarchical Clustering

We use hierarchical clustering to find the optimal K-value and ensure accurate distance calculations **by scaling the variables**. Four linkage methods are used to perform hierarchical clustering on the selected bases variables.

For instance, for method=complete, we have obtained the following elbow plot:



From, the adjacent elbow plot we can find the first kink at **k=4.**

Thus, we repeat this process **for all four linkage methods** and obtain all possible k-values.

The **possible k-values** that we have observed against each linkage methods from their respective elbow plots are tabulated below:

Linkage	**K-value(s) (from respective elbow plots)**
Complete	4
Single	7
Average	6
Centroid	5

K-means Clustering using observed K-values

We use the k-values obtained from hierarchical clustering (4,5,6,7) to ideally determine for which **k-value we have the lowest total within cluster variation (WCV)** in the clusters generated.

Upon performing k-means clustering with **k-values=4,5,6,7** and **nstart ranging from 20-50 (to obtain a global optimal solution)**, it was observed that for **k-value = 6** and **nstart=50**, we have obtained the **lowest within cluster variation** of **52933.05** Thus the optimal number of well-defined clusters that can be formed is **6**.

[PHASE 4]

Targeting and Profiling customers using Linear Discriminant Analysis

In order to achieve this objective, we perform **Linear Discriminant Analysis**, which is an important statistical technique used for supervised classification problems with multiple response classes.

It can be interpreted from the diagonal elements of the confusion matrix, which represents the number of observations for each class that the LDA model correctly predicts. The row number represents the original class of the data while the column number represents the predicted class.

From the predictive accuracy output it can be seen that the model has a **predictive accuracy of 11.8%**.

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	1	104	57	69	16	17
3	0	14	14	6	2	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

It is also evident from ANOVA testing that that the **LDA model is not statistically significant**. The main reasons behind this are:

- LDA assumes that the descriptor variables would be normally distributed (Tibshirani et al, 2013) whereas in this scenario, most of the descriptor variables are of a binary nature while the continuous ones are **NOT** normally distributed.

- **LDA also assumes that the prior probabilities of the classes are similar**(Tibshirani et al., 2013), however we can see that the prior probabilities of the classes are very different from each other which makes LDA an unsuitable approach for this problem.

```
##
## Prior probabilities of groups:
##           1           2           3           4           5           6
## 0.002857143 0.390000000 0.288571429 0.202857143 0.091428571 0.024285714
##
```

DISCUSSION AND INSIGHTS

From the customer attributes per segment, some key insights can be drawn :

- Average age of customers across all segments is around 35 years.
- Average income bracket is around 3 for all segments.
- Segments 3,4, and 6 has the highest number of customers among all segments.

SEGMENT WISE CUSTOMER ATTRIBUTES (Rounded to nearest whole number)

	Segment					
	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6
Avg. Age	37	38	35	35	36	35
Avg Order Size (in $)	95	144	28	41	61	16
Avg Order Frequency (Rounded)	1	0	2	1	1	1
Aggregate order size (in $)	2,376	1,728	10,960	11,223	4,995	3,396
Avg. Prioritize Hygiene	4	3	4	4	4	4
Avg. Food Quality	4	4	4	4	4	4
Avg. Income Bracket	3	3	3	3	3	3
Avg. Innovation Score	4	4	4	4	4	4
Avg. Quality of Service Score	4	3	4	3	4	4
Avg Menu Design Score	4	3	4	4	4	4
Number of Customers	25	12	390	271	82	217

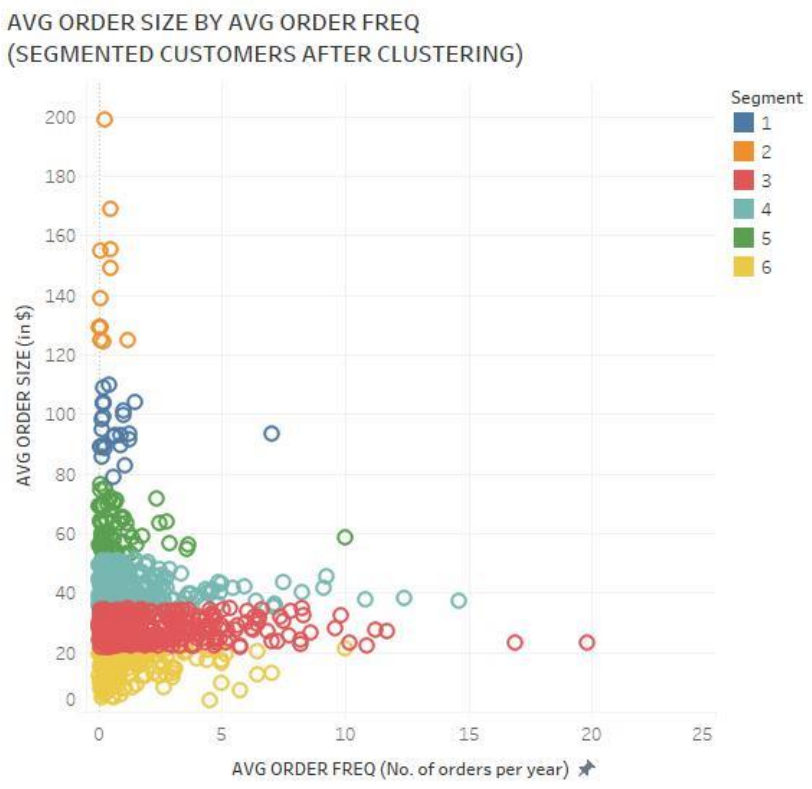
Avg. Age, Avg Order Size (in $), Avg Order Frequency (Rounded), Aggregate order size (in $), Avg. Prioritize Hygiene, Avg. Food Quality, Avg. Income Bracket, Avg. Innovation Score, Avg. Quality of Service Score, Avg Menu Design Score and Number of Customers broken down by Segment.

- Segment 2 has the highest average order size, however the aggregate order sizes of segments 3,4,5 and 6 are higher than the remaining 2.
- Segment 3 on an average has lower priority scores on all restaurant services (Quality of Service, Menu Design etc.) than all other segments.

The business should ensure to target customers who have either of the following qualities:

- A high order frequency and moderate order size.
  OR,
- A high order size and low to moderate order frequency.

From, the distribution of segments in the adjacent chart, however it is evident that the number of customers in the high order size segments (1 and 2) is very few and have low order frequencies compared to segments 3 and 4, which have a moderate order size and significant order frequencies.



Avg Order Freq vs. Avg Order Size. Colour shows details about Segment.

From the customer attributes table, it is also clear that despite having a higher average order size than other segments, segment 2 has much less aggregate order transaction value than segment 3 or 4. This is due to the fact that the number of customers in segments 3 and 4 is much higher than segment 2.

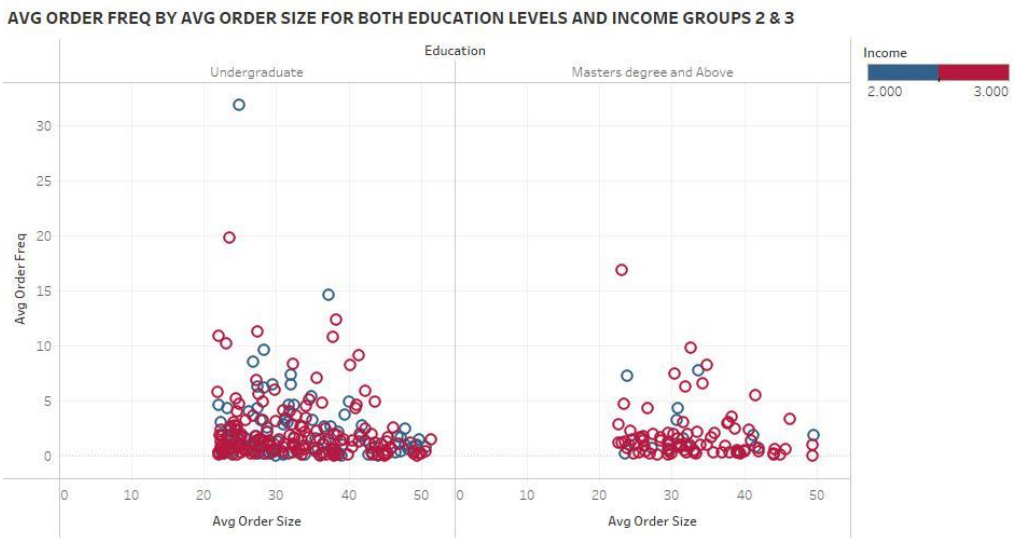
A scenario similar to this was obtained in Chen's et al. (2018) where eventually the target segment with the largest number of customers were chosen to customize their business strategy.

Another important aspect to look at is the combination of education level and income bracket of customers and how the average order sizes vary across these categories.

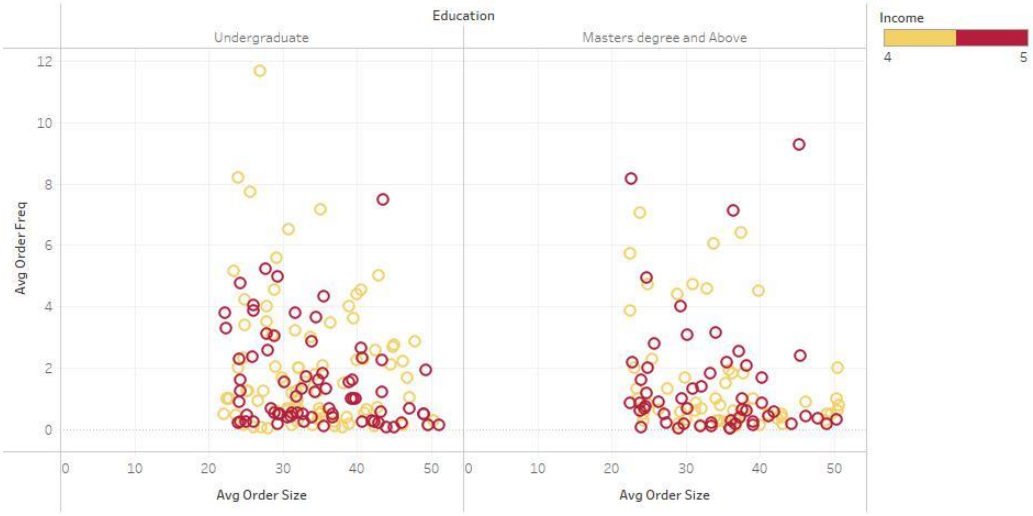
Since segments 3 and 4 are the most populated segments with high order revenue, a distribution of customer order size and frequency for these two segments has been plotted below across two education levels of the customers and different income brackets to get insights on their purchase behaviour.

In the adjacent chart, the customers with income levels 2 and 3 are depicted across the two education levels.

It is observed that for segments 3 and 4 more customers with an undergraduate degree have high order sizes than customers with a masters degree or above.



AVG ORDER FREQ BY AVG ORDER SIZE FOR BOTH EDUCATION LEVELS AND INCOME GROUPS 4 & 5



Avg Order Size vs. Avg Order Freq broken down by Education. Colour shows Income.The data is filtered on IncomeandSegment.The Income filter keeps 4and5.The Segment filter keeps Segment 3andSegment 4.

As is evident from the adjacent chart here, for income groups 4 and 5, the distribution pattern is quite similar as the previous chart.

However, it can be inferred from both these plots that in segments 3 and 4 there is a high population of customers in the income brackets of 3 and 4, who have high order sizes.

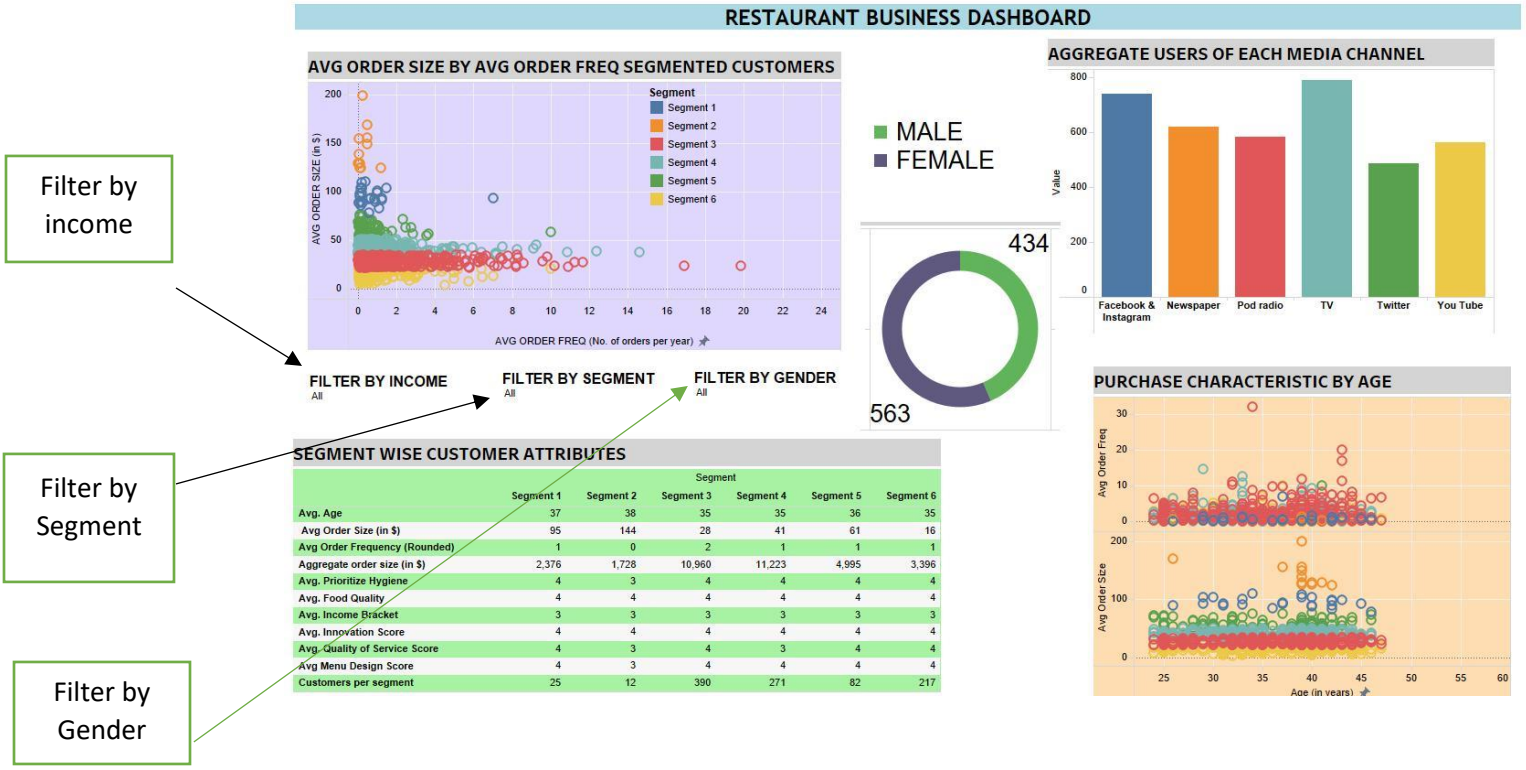
The business must also consider efficient marketing channels for advertising purpose for their target customers. From the aggregate media usage per segment tabulated below, it is clear that for segments 3 and 4, there are more subscribers to Facebook/Instagram and television compared to other media channels.

DIFFERENT MEDIA CHANNEL USERS PER SEGMENT

	Segment					
	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6
Facebook & Instagram	21.0	7.0	294.0	204.0	55.0	160.0
Twitter	17.0	1.0	190.0	146.0	35.0	99.0
Snapchat	13.0	2.0	143.0	99.0	30.0	77.0
You Tube	13.0	7.0	217.0	165.0	46.0	117.0
Pod radio	19.0	6.0	238.0	170.0	40.0	110.0
TV	21.0	11.0	303.0	213.0	63.0	180.0
Newspaper	19.0	6.0	247.0	179.0	42.0	128.0
Customers per segment	25.0	12.0	390.0	271.0	82.0	217.0

Facebook & Instagram,Twitter,Snapchat,You Tube,Pod radio,TV,Customers per segmentandNewspaper broken down by Segment.

Finally, the business dashboard designed below summarises the findings of the analysis.



	Segment					
	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6
Avg. Age	37	38	35	35	36	35
Avg Order Size (in $)	95	144	28	41	61	16
Avg Order Frequency (Rounded)	1	0	2	1	1	1
Aggregate order size (in $)	2,376	1,728	10,960	11,223	4,995	3,396
Avg. Prioritize Hygiene	4	3	4	4	4	4
Avg. Food Quality	4	4	4	4	4	4
Avg. Income Bracket	3	3	3	3	3	3
Avg. Innovation Score	4	4	4	4	4	4
Avg. Quality of Service Score	4	3	4	3	4	4
Avg Menu Design Score	4	3	4	4	4	4
Customers per segment	25	12	390	271	82	217

INSIGHTS AND STRATEGICAL ADVICE :

- **The restaurant should target customers in segments 3 and 4 to grow their revenue.**
- **The customers belonging to income brackets of 3 and 4, i.e earning between $75K-$175K should be targeted to grow the restaurants business.**
- **The restaurant should also target segment 6, as it shows great potential in terms of population, but their order sizes are not proportionally large. This segment can help grow the business further.**
- **Facebook/Instagram along with Television should be used as primary mode of advertising channel to expand the outreach of the business and target their customers efficiently.**
- **Customers in segments 3 and 4 have an above average requirement of Quality of Service, Restaurant Technology, Menu Design and Food Quality. Thus, the business must maintain the standards in the aforementioned areas to attract prospective customers who have similar profile.**

CONCLUSION

The analysis for this project is theoretically sound but as we have encountered, it was not possible to formulate a targeting model using LDA, given the nature of the dataset. However, methods like quadratic discriminant analysis or decision tree would be more appropriate to form a profiling model. It is also difficult to train a model accurately with just 1000 data-points having moderate diversity. However, it was possible to gain meaningful insights through clustering and visualisation of the data and understanding the categorical characteristics of each segment. This shall help to solve the problem of customer heterogeneity to an extent.

To address the problem of customer heterogeneity, businesses must collect and analyse customer data to better understand their preferences, behaviours, and needs. This information can be used to segment customers into groups based on shared characteristics, such as age, income, or dining habits. By doing so, businesses can create targeted marketing campaigns, develop products and services that meet the needs of specific customer segments, and make better-informed decisions about product development and pricing.

REFERENCES

- Chen, Y., Huang, L., & Wang, Y. (2018). Prediction of customer preferences in the restaurant industry: A big data analytics approach. International Journal of Hospitality Management, 72, 16-27.

- James G., Witten D., Hastie T., Tibshirani R., (2013). An Introduction to Statistical Learning : with Applications in R. New York :Springer.

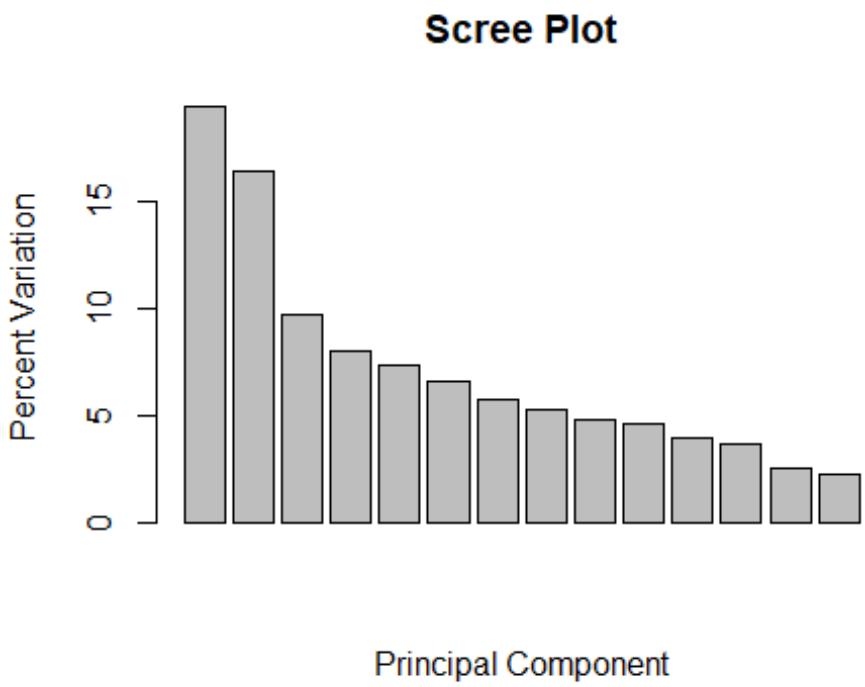
- Jolliffe, I. T. (2002). Principal component analysis. Wiley Online Library.

- Kim, S., Kim, J., & Kwon, Y. (2017). Personalized menu recommendation system for restaurants using big data analysis. Journal of Hospitality and Tourism Technology, 8(1), 94-107.

- Zhang, Y., Wang, D., Yu, Z., & Liu, C. (2019). Customer segmentation and relationship management in the Chinese restaurant industry: A data-driven approach. International Journal of Hospitality Management, 80, 23-34.

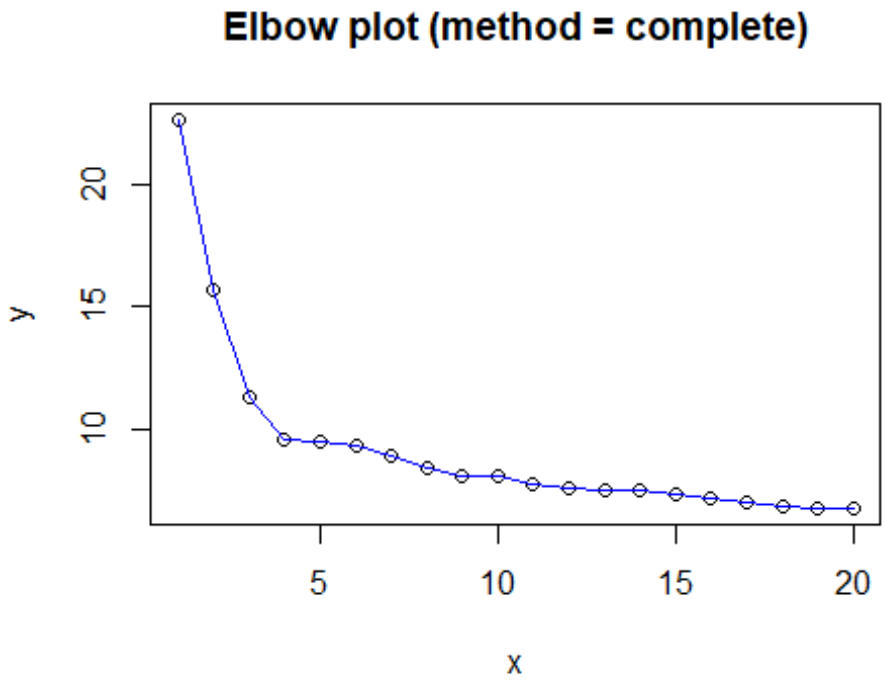
APPENDIX

```r
#reading and summarising the data
data_res <- read.csv(file.choose())

head(data_res)

str(data_res)

summary(data_res)

#checking for missing values
sum(is.na(data_res))

## [1] 0

#correlation matrix of the basis variables
cor(data_res[,c(2:15)])

#Perform PCA on the basis variables of the dataset
pca <- prcomp(data_res[c(2:15)], center = TRUE, scale. = TRUE)

#extract the variances explained by each component
pca.var <- pca$sdev^2  #also known as eigen value

#Calculate the proportion of variance explained by each component
pca.var.per <- round(pca.var/sum(pca.var)*100, 1)
pca.var.per

##  [1] 19.4 16.4  9.7  8.0  7.3  6.6  5.7  5.2  4.8  4.6  3.9  3.6  2.5  2.2

#build a scree plot
barplot(pca.var.per, main="Scree Plot", xlab="Principal Component", ylab="Per
cent Variation")
```




```
# get the name and correlation of the top 10 predictors that contribute most
to pc1,pc2,pc3 and pc4.
loading_scores <- pca$rotation[,c(1:4)]
abs.loading <- abs(loading_scores) ## get the magnitudes

#Sort the variables with high correlation in each of the three principal comp
onents in descending order.
lscore.pc1 <- sort(abs.loading[,1], decreasing=TRUE)
lscore.pc2 <- sort(abs.loading[,2], decreasing=TRUE)
lscore.pc3 <- sort(abs.loading[,3], decreasing=TRUE)
lscore.pc4 <- sort(abs.loading[,4],decreasing = TRUE)

#Extract the names of the top 10 variables in each PC.
top10.pc1 <- names(lscore.pc1[1:10])
top10.pc2 <- names(lscore.pc2[1:10])
top10.pc3 <- names(lscore.pc3[1:10])
top10.pc4 <- names(lscore.pc4[1:10])
top10.pc1

##  [1] "Restaurant_Technology" "Menu_Design"           "Staff_behavior"
##  [4] "Location"              "Prioritize_Hygiene"    "Interior_design"
##  [7] "Beverages"             "Food_Quality"          "Quality_of_Service"
## [10] "innovation"

set.seed(1)
#dataset is already loaded, so we can avoid loading it again, rather attach i
t for better coding.
```

```r
attach(data_res)  #attach dataset to avoid repetitive mentioning of dataset.
library(dplyr)

#remove outliers from data
data_res <- data_res %>% filter(avg_order_size<200)
#perform hierarchical clustering on the basis variables obtained from PCA in
the previous block.

##performing h-clustering using different linkage methods
hclust.res <- hclust(dist(scale(cbind(Menu_Design,Staff_behavior,Location,Pri
oritize_Hygiene,Interior_design,Beverages,Food_Quality,Quality_of_Service,inn
ovation,avg_order_freq,avg_order_size))), method="complete")   ##we have chan
ged method to single/average/centroid to obtain respective k-values

#generate an elbow plot to determine the K-value
x<-c(1:20)
height.sorted <- sort(hclust.res$height, decreasing=TRUE)
y<-height.sorted[1:20]
plot(x,y,main="Elbow plot (method = complete)") ; lines(x,y,col="blue")  #elb
ow plot
```



```
##Perform k-means clustering. change k-values to 4,5,6,7 and assign different
nstart values to compare
##the WCV, and then finalise on the k-value for the minimum WCV

kmeans.res <- kmeans(x=data.frame(Menu_Design,Staff_behavior,Location,Priorit
ize_Hygiene,Interior_design,Beverages,Food_Quality,Quality_of_Service,innovat
ion,avg_order_freq,avg_order_size),6,nstart=50)

#find the total within cluster variation, select the k-value for which this i
s minimum
kmeans.res$tot.withinss

## [1] 54835.88

table(kmeans.res$cluster)

##
##   1   2   3   4   5   6
## 194 268  45 121 357  12

#store the assigned cluster numbers of the data in a variable and join the co
lumn to the original dataset
segment <- kmeans.res$cluster
restaurant_segmented <- cbind(data_res,segment)
detach(data_res)
#write the final dataset to a file for further visualisation
#write.csv(restaurant_segmented, file = file.choose(new=TRUE), row.names = FA
LSE)
```

Linear Discriminant Analysis Section :

```
## Load Packages and Set Seed
library(MASS)

set.seed(1)

## Read in Segment Data and Classification Data
 #intital data is already stored in data_res and segmented data is stored in
restaurant_segmented
attach(restaurant_segmented)  # attaching dataset to avoid repetition

## The following object is masked _by_ .GlobalEnv:
##
##     segment

index <- sample(1:nrow(restaurant_segmented), round(0.7*nrow(restaurant_segme
nted))) # 70% training data
train <- restaurant_segmented[index, ]
test <- restaurant_segmented[-index, ]

## Run Discriminant Analysis
fit <- lda(segment ~ Age + zip_code + Gender + FB_Insta + Snap + Education +
```

```
Twit + Health + Income + YouTube + Pod_radio + NewsP , data = train)

fit ## print the summary statistics of your discriminant analysis

## Check which Discriminant Functions are Significant
ldaPred <- predict(fit, test)
#print(ldaPred)
ld <- ldaPred$x
anova(lm(ld[,1] ~ test$segment))  #ANOVA to check significance of LDA

## Analysis of Variance Table
##
## Response: ld[, 1]
##               Df  Sum Sq Mean Sq F value Pr(>F)
## test$segment   1   0.939 0.93907  1.0527 0.3057
## Residuals    297 264.933 0.89203

anova(lm(ld[,2] ~ test$segment))

## Analysis of Variance Table
##
## Response: ld[, 2]
##               Df Sum Sq Mean Sq F value Pr(>F)
## test$segment   1   0.00 0.00008   1e-04 0.9925
## Residuals    297 278.98 0.93931

anova(lm(ld[,3] ~ test$segment))

## Analysis of Variance Table
##
## Response: ld[, 3]
##               Df  Sum Sq Mean Sq F value Pr(>F)
## test$segment   1   0.815 0.81464  0.7772 0.3787
## Residuals    297 311.290 1.04811

anova(lm(ld[,4] ~ test$segment))

## Analysis of Variance Table
##
## Response: ld[, 4]
##               Df  Sum Sq Mean Sq F value Pr(>F)
## test$segment   1   0.033 0.03309  0.0316  0.859
## Residuals    297 310.956 1.04699

anova(lm(ld[,5] ~ test$segment))

## Analysis of Variance Table
##
## Response: ld[, 5]
##               Df  Sum Sq Mean Sq F value Pr(>F)
## test$segment   1   0.121 0.12139  0.1259  0.723
## Residuals    297 286.406 0.96433
```

```
## Check Discriminant Model Fit

tseg <- table(ldaPred$class, test$segment)
tseg # print table

##
##      1  2  3  4  5  6
##   1  2  0  0  1  3  0
##   2 12 10  8  2 23  0
##   3  0  0  0  0  0  0
##   4  1  0  0  0  0  0
##   5 42 76 12 23 78  6
##   6  0  0  0  0  0  0

sum(diag(tseg))/nrow(restaurant_segmented) # print percent correct

## [1] 0.09027081

## Run Classification Using Discriminant Function
pred.class <- predict(fit, restaurant_segmented)$class
tclass <- table(pred.class)
tclass # print table

## pred.class
##   1   2   3   4   5   6
##  13 197   1   1 785   0

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

detach(data_res) ##detaching dataset

detach(restaurant_segmented) ##detaching dataset
```