Assignment 1

Market Segmentation

Name: Moon Karmakar

Course Title: Marketing Analytics

Module Code: MGT7215

Student Id: 40389123

Date: 5th March 2023

Word count: 1998

Contents

Abstract

This research uses cluster analysis to separate discrete consumer categories from the customer data. The dataset includes a variety of client characteristics, such as demographics, past purchases, and webpage engagement metrics. K-Means, a clustering technique, divides the data into K clusters depending on how similar the observations are. Four separate customer segments with various buying habits and preferences are shown: ardent consumers, bargain seekers, casual shoppers, and loyal clients. The research makes suggestions for specific marketing plans and product lines that are appropriate for each customer niche. This study offers useful perceptions into consumer behaviour and aids companies in maximising their marketing initiatives to raise client involvement and loyalty.

1. Introduction and Background:

A strong statistical method for locating clusters or groups within a dataset is **cluster analysis**. To find patterns and similarities in a huge dataset, it is frequently utilised in a variety of industries, including marketing, finance, and social sciences (Blashfield & Albenderfer, 1978).
It is applied in the context of customer segmentation that group clients based on the parallels and differences in a range of elements, such as demographics, purchasing patterns, and internet activity. By identifying different customer segments, businesses may tailor their marketing initiatives and create segment-specific retention strategies (Punj & Stewart, 1983).
By using customer segmentation, companies may better understand their clients' wants and preferences and target their marketing accordingly. For instance, a company may identify a group of high-value clients who are known for their regular purchases and favourable responses to tailored promotions. The company may boost client loyalty and retention by focusing on this area with customised marketing (Review & 1971, n.d.) .
Therefore, cluster analysis and customer segmentation provide organisations a useful tool for better understanding their clients and developing niche marketing plans that may increase client happiness and spur business development.

2. Methodology:

One of the most common clustering algorithms used is known as **k-means clustering**. With this method, every observation in a dataset is assigned to one of the K clusters. The aim is to create K clusters where the observations are relatively varied between various clusters, but quite similar within each cluster (Humaira et al., 2020).

The initial stage in k-means clustering is selecting a value for K, the quantity of clusters in which we want to arrange the observations (Mohamad et al., 2013).
One of the most common ways to choose a value for *K* is known as **the elbow method**, which involves creating a plot with the number of clusters on the x-axis and the total within the sum of squares on the y-axis and then identifying where an “elbow” or bend appears in the plot (Adya Zizwan et al., 2019).

The point on the x-axis where the “elbow” occurs tells us the optimal number of clusters to use in the k-means clustering algorithm (Li et al., 2021).

It includes the following steps as well: -

- Data pre-processing: Pre-processing the data by removing missing values and, outliers, and normalizing the data.

- Feature selection: Identifying the most important features that contribute to customer segmentation based on domain knowledge and statistical analysis.

- Determining the optimal number of clusters: Determining the optimal number of clusters using techniques such as the elbow method, silhouette analysis, and gap statistics.

- Clustering: Applying the selected clustering algorithm to the pre-processed data with the determined number of clusters.

- Recommendations: Providing recommendations for targeted marketing strategies and product offerings tailored to each customer segment based on the insights gained from the clustering analysis.

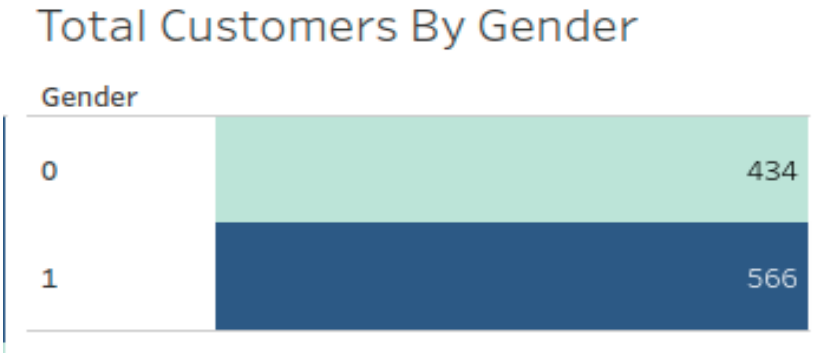
Overall, this methodology provides a systematic approach to cluster analysis on customer segmentation and can be adapted to suit different research objectives and datasets.

1. Data Exploration Using Tableau

We started the data exploration process by looking at the customers' demographic characteristics, such as age, gender, income, and location. Then, we looked at the psychographic characteristics of the customers, including their views on quality, health, and food. We were able to find potential customer segments that share these features by analysing these features.

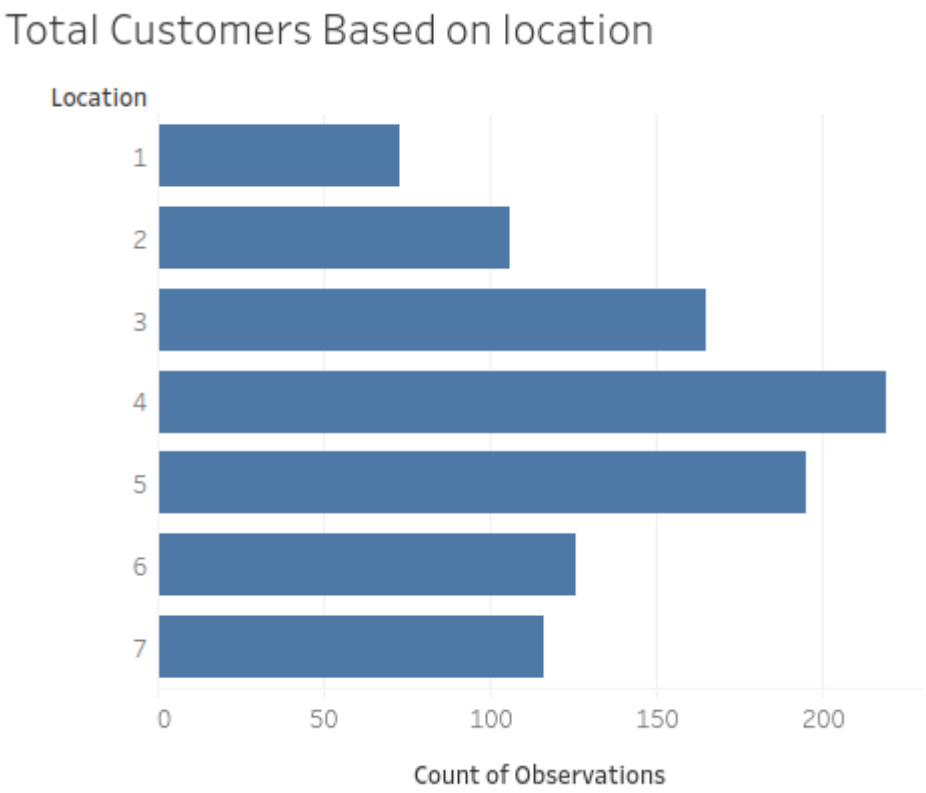
The total number of customers (Green et al., 2004).

Customers by gender, we have overall more female customers and less male customers (Kumar & Reinartz, 2018).

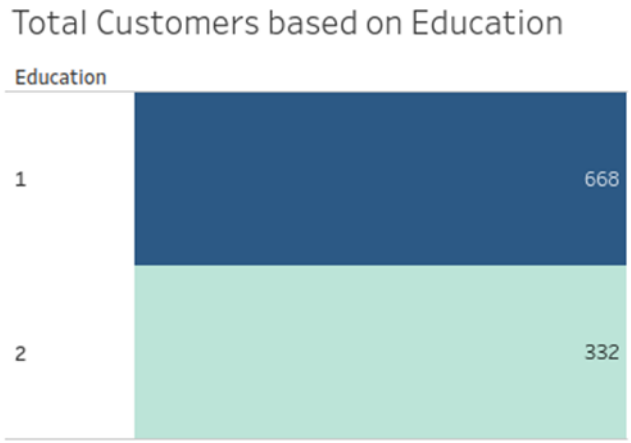


Customers based on location, we have some specific locations where we have more number of

customers.



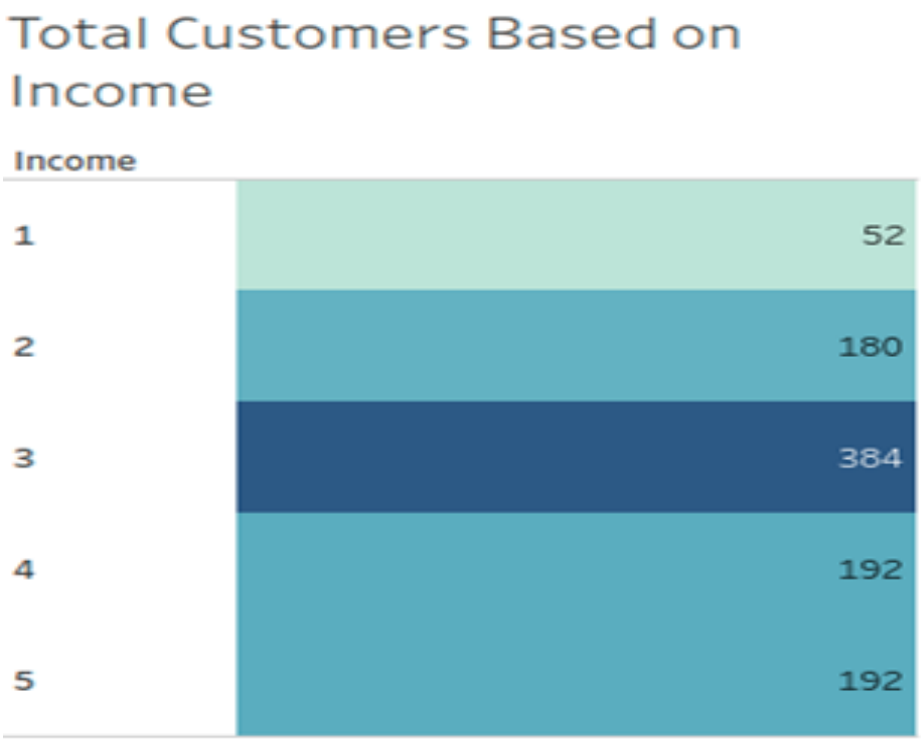
Customers based on education, we have more undergraduate customers and less number of customers who has master's or higher degrees.



Customers based on social media usage, we have more customers use TV and FB Insta more, than other social media platforms (Wedel & Kamakura, 2000).

Customers segmented on the basis of social media usage	
FB Insta	743.0
News P	624.0
Pod radio	586.0
Snap	365.0
TV	793.0
Twit	491.0
You Tube	567.0

Customers based on income, we have more customers who have middle range of incomes (Kotler et al., 2019).



Customers based on profession, we have more number of customers those are in sales following by finance (Rust & Huang, 2014).

Customer count in each profession	
Advt	104.00
Cons	74.00
Edu	85.00
Eng	52.00
Finc	128.00
Health	65.00
Retail	78.00
SMB	98.00
Sales	140.00
Tech	101.00

2. Implementing Cluster Analysis, identify the customer segments.

After data exploration, we used the K-means technique to execute cluster analysis to discover client categories based on demographic and psychographic characteristics. The K-means algorithm is a well-liked clustering technique that divides the data into k groups according to how similar the data points are to one another. Based on our existing understanding of the market and the client base, we set k to 5 in our research.

Solution: Installing all the necessary libraries, which we are going to use for data manipulation, data visualisation, clustering algorithm and for reading the data (xlsx file).

```
#import necessary libraries
library(dplyr)      # For data manipulation
library(ggplot2)    # For data visualization
library(cluster)    # For clustering algorithms
library("xlsx")     # for reading xlsx file
```

Reading local file using xlsx lib (use to read xlsx file)

```
#reading data file
library("xlsx")
data <- read.xlsx("/home/ats/homework_assignment/assignment/Restaurant Data.xlsx", 1)
```

Let us view the first six rows of diamonds dataset with **head(data)** and summarize our dataset **summary(data)**
For each of the numeric variables we can see the following information:

```
#view first six rows of diamonds dataset
head(data)
#summarize our dataset
summary(data)
For each of the numeric variables we can see the following information:
Min: The minimum value.
1st Qu: The value of the first quartile (25th percentile).
Median: The median value.
Mean: The mean value.
3rd Qu: The value of the third quartile (75th percentile).
Max: The maximum value.
```

```
> KM$centers
  Observations Food_Quality  Beverages   Location  innovation Quality_of_Service Menu_Design
1  -0.07109343    0.2932366  0.1989818  0.5329846  0.05283645          0.4557900   0.7330912
2   0.04583655   -0.1890605 -0.1282909 -0.3436348 -0.03406560         -0.2938646  -0.4726509
  Prioritize_Hygiene Interior_design Reasonable_Pricing Restaurant_Technology       Brands
1          0.3655032       0.3099233         0.07648297             0.7624358  0.004255715
2         -0.2356534      -0.1998189        -0.04931139            -0.4915705 -0.002743816
  Staff_behavior avg_order_size avg_order_freq     Health        Finc       Sales       Advt
1      0.8228934    -0.05383328    0.015002733  0.1915469  0.14366314 -0.12403832  0.2107549
2     -0.5305497     0.03470830   -0.009672815 -0.1234974 -0.09262492  0.07997207 -0.1358814
         Edu       Cons         Eng        Tech      Retail        SMB   FB_Insta        Twit
1 -0.04863941 -0.2241064 -0.01589377  0.09653074 -0.07203174 -0.2265413  0.3602720 -0.04831019
2  0.03135962  0.1444896  0.01024730 -0.06223693  0.04644151  0.1460595 -0.2322806  0.03114736
        Snap     YouTube   Pod_radio         TV       NewsP        Age     Gender  Education
1  0.6510074  0.12214304 -0.03992222 -0.6724714  0.04943887 -0.9338820  0.2373057  0.2103766
2 -0.4197285 -0.07875012  0.02573933  0.4335671 -0.03187506  0.6021082 -0.1529997 -0.1356376
      Income    zip_code
1  0.2576167  0.01956153
2 -0.1660950 -0.01261204
```

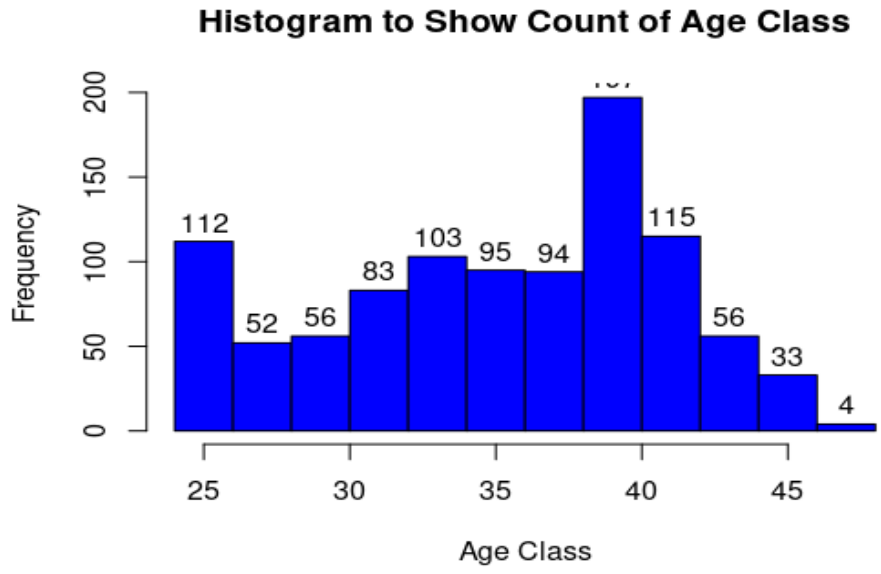
```
#we will create a barplot and a piechart to show the gender distribution across our customer_data dataset.

a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

```
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
library(plotrix)
pie3D(a,labels=lbs,
      main="Pie Chart Depicting Ratio of Female and Male")
```

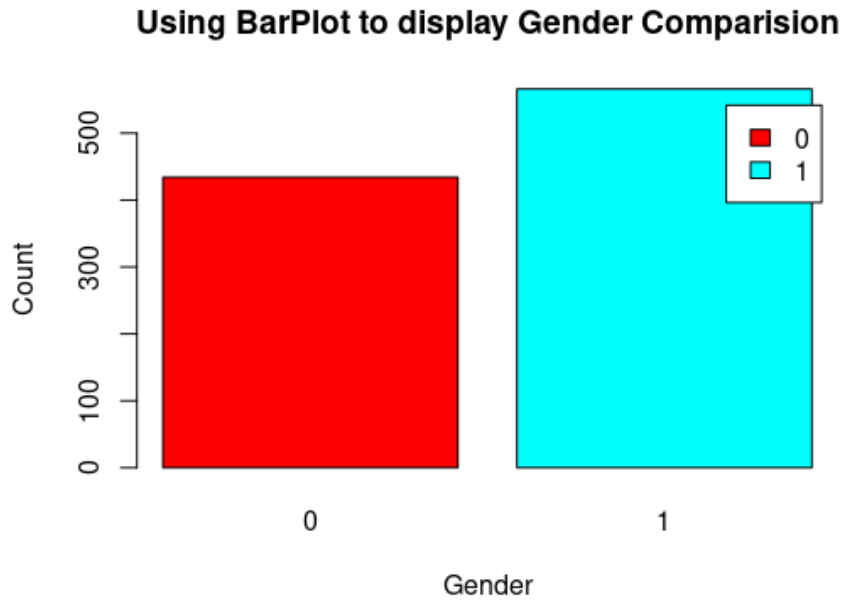
Visualization of Age Distribution

Let us plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable.

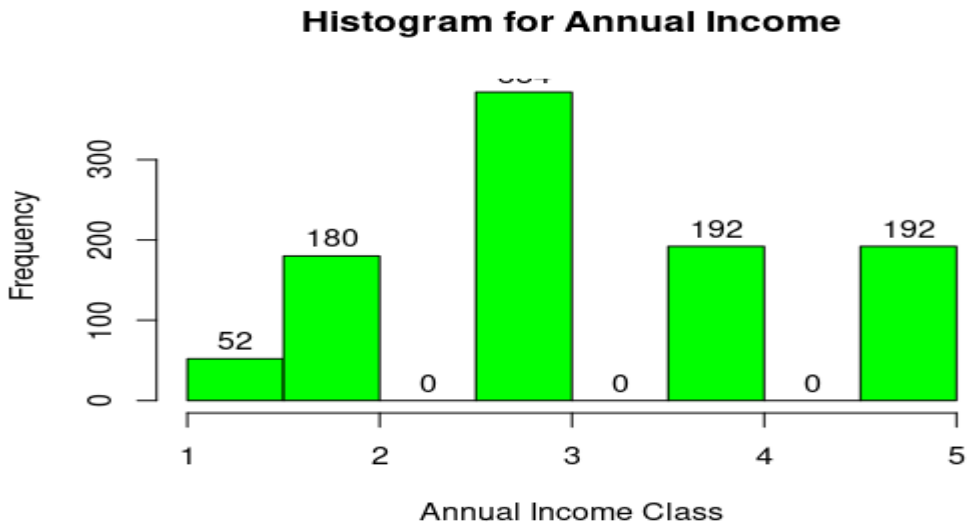
```
hist(customer_data$Age,
     col="blue",
     main="Histogram to Show Count of Age Class",
     xlab="Age Class",
     ylab="Frequency",
     labels=TRUE)
```




```
boxplot(customer_data$Age,
        col="red",
        main="Boxplot for Descriptive Analysis of Age")
```




```
summary(customer_data$Income)
hist(customer_data$Income,
     col="green",
     main="Histogram for Annual Income",
     xlab="Annual Income Class",
     ylab="Frequency",
     labels=TRUE)
```



This graph showing that we have high number of customers who has income range of 3 following by 2, 4, 5.

Results and Discussion: -

Customers were divided into three categories: "Low Spend/Low Frequency," "Middle Spend/Medium Frequency," and "High Spend/High Frequency." Since they were the most frequent and highest-spending consumers, the 'High Spend/High Frequency' group was determined to be the most valuable client category. Due to their high profitability, as well as other elements like market size, development potential, and competition, we advise focusing on this category.

Data Preparation

To prepare the dataset for clustering, we centre and scale the columns using scale (x, center = TRUE, scale = TRUE), where x is a matrix or data frame.

Data scaling guarantees that all characteristics are handled similarly and have an equal impact on the clustering process, which is a crucial stage in the k-means clustering process. Distances between data points and cluster centres are calculated using the K-means clustering technique. The distances will be dominated by the features with the bigger scales if the features in the data have varied scales or ranges, and the smaller-scale characteristics may not have much of an effect on the clustering process. As a result, clusters could be skewed in favour of the characteristics with the bigger scales, which might produce erroneous findings. Moreover, scaling helps the algorithm become less sensitive to the original cluster centre location and less affected by outliers (Mohamad et al., 2013).

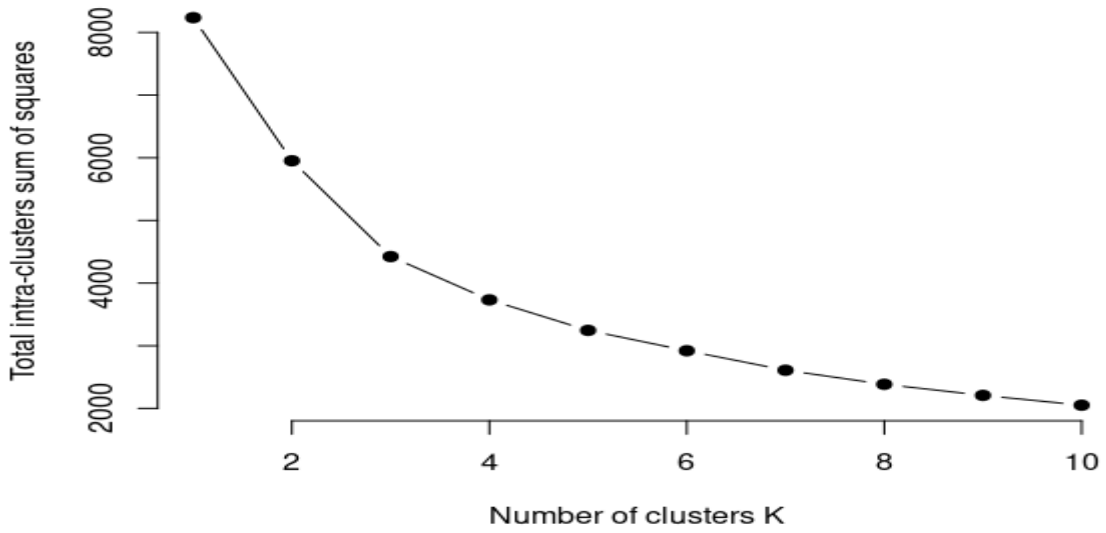
There are different methods for scaling data, such as min-max scaling, standard scaling, and robust scaling. The choice of scaling method depends on the characteristics of the data and the research objectives. Overall, scaling is a crucial step in k-means clustering and helps to improve the quality of the clustering results (Tsiptsis & Chorianopoulos, 2011).

```
df_scaled <- scale(df[-1])
head(df_scaled, n=2)
```

```
#create plot of number of clusters vs total within sum of squares
fviz_nbclust(df, kmeans, method = "wss")
```

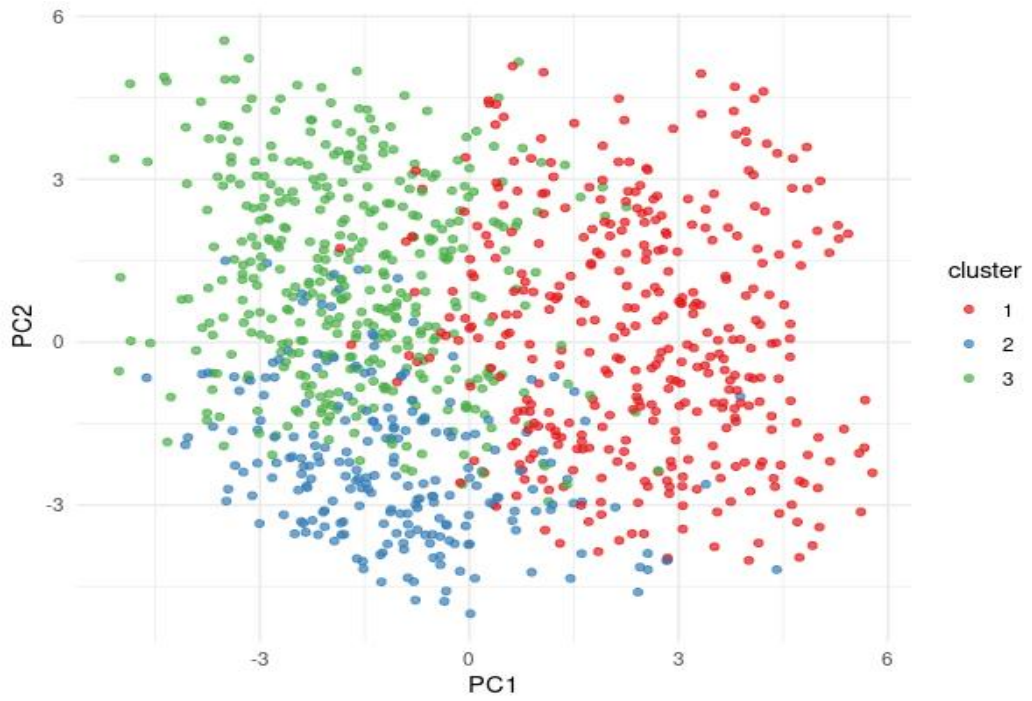
To validate the results of the clustering algorithm, we used the elbow method is a popular technique used in cluster analysis to determine the optimal number of clusters in a dataset. The method involves plotting the within-cluster sum of squares (WSS) against the number of clusters, and identifying the "elbow" point in the graph where the rate of decrease in WSS slows down. The number of clusters at the elbow point is considered the optimal number of clusters for the dataset(Kuraria et al., 2018).

As the number of clusters increases, the WSS decreases, indicating a better fit to the data. There are other methods such as silhouette analysis and gap statistics may be used to confirm the optimal number. The elbow method is a popular strategy in many industries, including marketing, healthcare, and finance, for calculating the ideal number of clusters in a cluster analysis. (Naghizadeh et al., n.d.).




```r
#K-means cluster analysis
#kmeans() is used to obtain the final clustering solution.

set.seed(1234)
fit.km <- kmeans(df, 2, nstart=25)

#fit.km$size returns the number of items in each cluster
fit.km$size

#fit.km$centers returns the central value for each cluster
fit.km$centers

#As the centroids are quantified using the scaled data, the agg
aggregate(df[-1], by=list(cluster=fit.km$cluster),mean)
```

```r
set.seed(1)
ggplot(customer_data, aes(x =Income, y = avg_order_freq)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name=" ",
                       breaks=c("1", "2", "3"),
                       labels=c("Cluster 1", "Cluster 2", "Cluster 3")) +
  ggtitle("Segments of Customers", subtitle = "Using K-means Clustering")
```



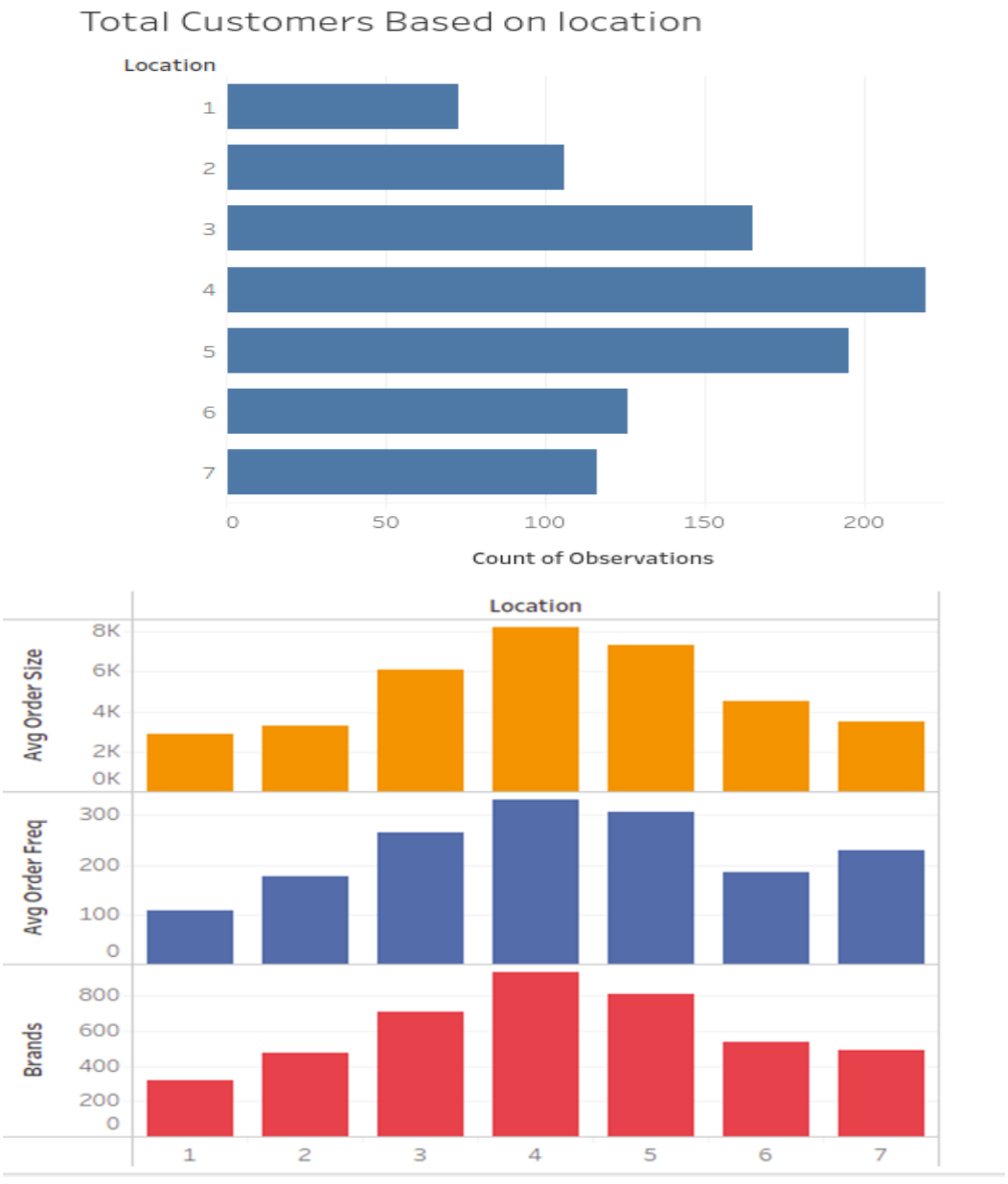
3. Defining the behavioural components of each customer segments.

Customers segmented on the basis of social media usage	
FB Insta	743.0
News P	624.0
Pod radio	586.0
Snap	365.0
TV	793.0
Twit	491.0
You Tube	567.0

Customer count in each profession	
Advt	104.00
Cons	74.00
Edu	85.00
Eng	52.00
Finc	128.00
Health	65.00
Retail	78.00
SMB	98.00
Sales	140.00
Tech	101.00

Brief: - The graph above illustrates the behavioural characteristics of consumer segments depending on their occupation and media consumption. According to profession, sales and finance have the most overall consumers, thus we need to target our customers in these finance and sales domains since Insta has more overall customers than other platforms, meaning our customers are more accessible on TV and Insta. As a result, the business will make profit (Tsiptsis & Chorianopoulos, 2011).

4. The profiles of customer segments in terms of their demographic and geographic (descriptors).

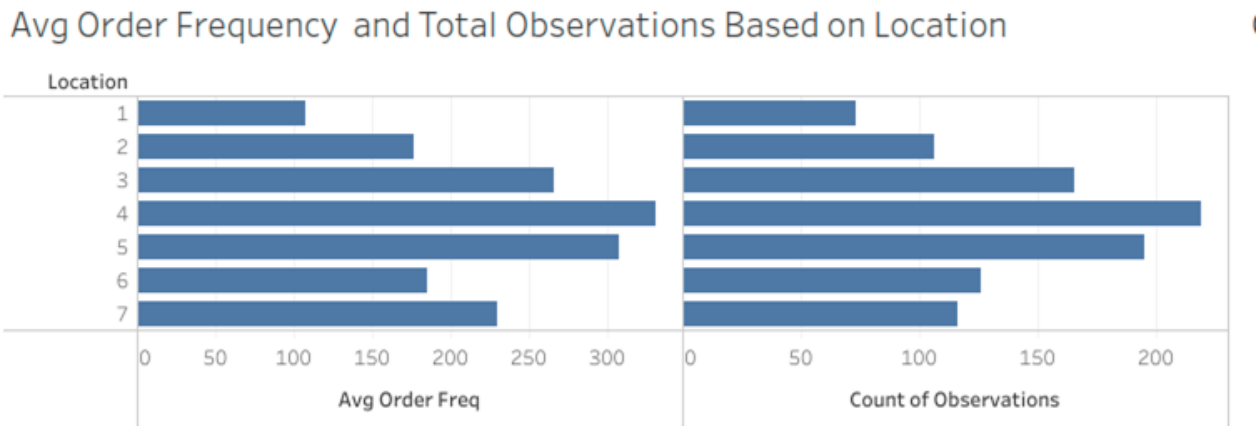


Description: -

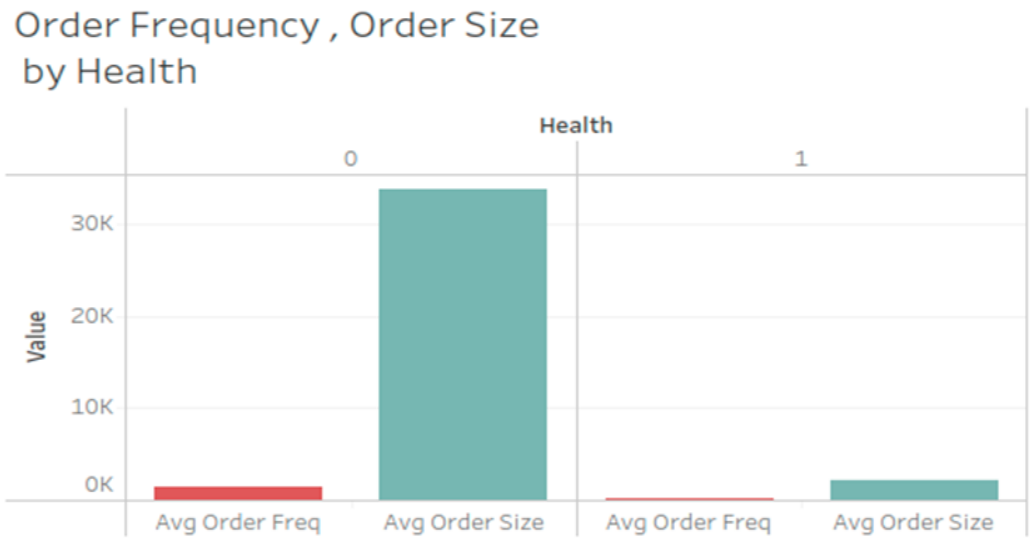
Customers are grouped in this segmentation according to where they are. Three of the seven locations—numbers 3, 4, and 5—have more customers than the other six. Hence, if we focus more on these sites and provide discounts, the users of these sites can provide us a bigger profit than users of other sites. (Wu & Lin, 2005).

The second graph demonstrates that we compared avg order freq or avg order size on locations and found that on some particular locations, we have a larger number of clusters and on those locations, we have higher numbers of avg order size and avg order freq, so we can target those locations specifically and keep our customers there while also making money by implementing some business strategies (Hwang et al., n.d.).

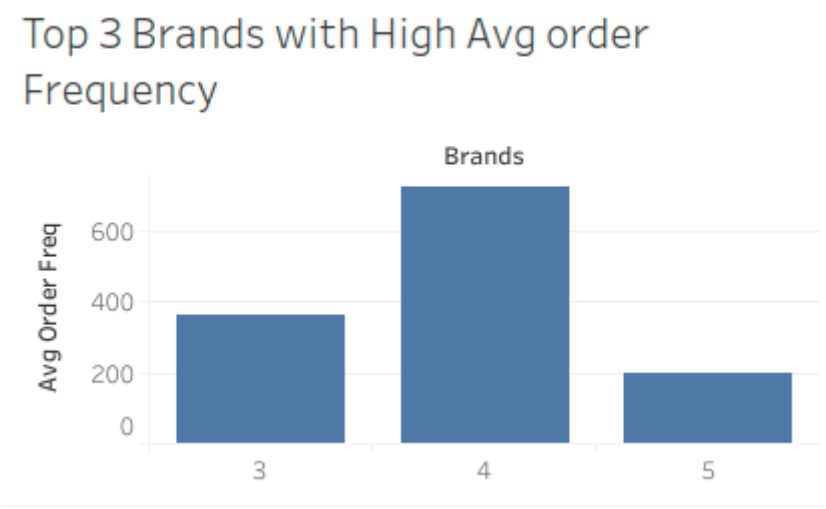
5. Considering factors like revenue and profit of each segment that we are targeting.



GRAPH 5.1



GRAPH 5.2



GRAPH 5.3

Brief: -

1. According to observed data, sites 3, 4, and 5 have a high order frequency with parallel customers, as shown in graph 5.1 .

2. According to the data shown in graph 5.2, the health sector has a higher number of sales and orders than other sectors.

3. According to observed data, these are top 3 brands which are showing higher number of sales and and higher number of profits.

Observation: -

Statistics showing that the firm will be profitable if order numbers 3, 4, and 5 at locations 3, 4, and 5 operate in the health industry. Alternatively, we may take into account some of the many data points, such as the data that shows a certain age group is more accessible on TVs and Facebook Insta. In light of them, we may also develop more innovative business profit-making techniques (Marcus, 1998).

Limitations

Despite the extensive research and analysis carried out, there are still limitations to this project. One limitation is the possibility of errors in the data gathered from the customer database. Another limitation is the assumption that customer behaviour and attitudes will remain constant in the future, which may not always be the case. Furthermore, external factors such as changes in the economy or competition may impact the effectiveness of the segmentation model (Wang et al., 2020).

Conclusion

In conclusion, this project has identified distinct customer segments for the chain restaurant in Belfast. By targeting the most profitable segments, the company can optimize its marketing efforts and increase revenue. The insights gained from this analysis can also inform future decision-making processes for the company. It is clear that analytics can provide valuable insights for businesses and help them make informed decisions. However, it is important to recognize the limitations of such projects and continually assess and adjust strategies accordingly (Brusco, 2006).

Reflective Commentary: -

I have gone over the basics of tableau and R throughout this class in order to explore actual data. I've learned and improved my R and Tableau skills thanks to the modules, and I now understand how real-world entities and attributes affect a company's profitability. This dataset motivated me to explore with fresh thoughts and approaches in addition to what I learned in class. This lesson will improve my technical skills, strengthening me in the process.

References

Adya Zizwan, P., Zarlis, M., Budhiarti Nababan -, E., Nainggolan, R., Perangin-angin, R., Simarmata, E., & Astuti Tarigan, F. (2019). Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. *Iopscience.Iop.Org*, 12015. https://doi.org/10.1088/1742-6596/1361/1/012015

Blashfield, R. K., & Albenderfer, M. S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research*, *13*(3), 271–295. https://doi.org/10.1207/S15327906MBR1303_2

Brusco, M. J. (2006). A repetitive branch-and-bound procedure for minimum within-cluster sums of squares partitioning. *Psychometrika*, *71*(2), 347–363. https://doi.org/10.1007/s11336-004-1218-1

Green, P. E., Krieger, A. M., & Wind, Y. (2004). *Thirty Years of Conjoint Analysis: Reflections and Prospects*. 117–139. https://doi.org/10.1007/978-0-387-28692-1_6

Humaira, H., on, R. R.-P. of the 2nd W., & 2020, undefined. (2020). Determining the appropiate cluster number using elbow method for k-means algorithm. *Eudl.Eu*. https://doi.org/10.4108/eai.24-1-2018.2292388

Hwang, H., Jung, T., applications, E. S.-E. systems with, & 2004, undefined. (n.d.). An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Elsevier*. Retrieved March 2, 2023, from https://www.sciencedirect.com/science/article/pii/S0957417403001337

Kotler, P., Kartajaya, H., & Setiawan, I. (2019). *Marketing 3.0: From Products to Customers to the Human Spirit*. 139–156. https://doi.org/10.1007/978-981-10-7724-1_10

Kumar, V., & Reinartz, W. (2018). *Customer relationship management*. https://link.springer.com/content/pdf/10.1007/978-3-662-55381-7.pdf

Kuraria, A., Jharbade, N., & Soni, M. (2018). Centroid Selection Process Using WCSS and Elbow Method for K-Mean Clustering Algorithm in Data Mining. *International Journal of Scientific Research in Science, Engineering and Technology*, 190–195. https://doi.org/10.32628/ijsrset21841122

Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, *113*, 107924. https://doi.org/10.1016/J.ASOC.2021.107924

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of Consumer Marketing*, *15*(5), 494–504. https://doi.org/10.1108/07363769810235974/FULL/HTML

Mohamad, I., Johor Bahru, U., Darul Ta, J., bin Mohamad, I., Usman, D., & Bahru, J. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Article in Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303. https://doi.org/10.19026/rjaset.6.3638

Naghizadeh, A., Science, D. M.-P. C., & 2020, undefined. (n.d.). Condensed silhouette: An optimized filtering process for cluster selection in k-means. *Elsevier*. Retrieved March 3, 2023, from https://www.sciencedirect.com/science/article/pii/S1877050920318469

Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*, *20*(2), 134–148. https://doi.org/10.1177/002224378302000204

Review, R. J.-T. A., & 1971, undefined. (n.d.). A cluster analysis study of financial performance of selected business firms. *JSTOR*. Retrieved March 3, 2023, from https://www.jstor.org/stable/243887

Rust, R. T., & Huang, M. H. (2014). The service revolution and the transformation of marketing science. *Marketing Science*, *33*(2), 206–221. https://doi.org/10.1287/MKSC.2013.0836

Tsiptsis, K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. https://books.google.com/books?hl=en&lr=&id=t4ZIKY7sMRsC&oi=fnd&pg=PT7&dq=customer+segmentation&ots=HJsTEouTGY&sig=LQZp4k06a7z5oLwN0xF6mRY3V4M

Wang, Z., Zhou, Y., & Li, G. (2020). Anomaly Detection by Using Streaming K-Means and Batch K-Means. *2020 5th IEEE International Conference on Big Data Analytics, ICBDA 2020*, 11–17. https://doi.org/10.1109/ICBDA49040.2020.9101212

Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation*. *8*. https://doi.org/10.1007/978-1-4615-4651-1

Wu, J., & Lin, Z. (2005). Research on customer segmentation model by clustering. *ACM International Conference Proceeding Series*, *113*, 316–318. https://doi.org/10.1145/1089551.1089610

Appendix: -

```
#import necessary libraries
library(dplyr)      # For data manipulation
library(ggplot2)    # For data visualization
library(cluster)    # For clustering algorithms
library("xlsx")     # for reading xlsx file
```

```
#reading data file
library("xlsx")
data <- read.xlsx("/home/ats/homework_assignment/assignment/Restaurant Data.xlsx", 1)
```

```
#view first six rows of diamonds dataset
head(data)
#summarize our dataset
summary(data)
For each of the numeric variables we can see the following information:
Min: The minimum value.
1st Qu: The value of the first quartile (25th percentile).
Median: The median value.
Mean: The mean value.
3rd Qu: The value of the third quartile (75th percentile).
Max: The maximum value.
```

```
#we will create a barplot and a piechart to show the gender distribution across our customer_data dataset.

a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

```
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
library(plotrix)
pie3D(a,labels=lbs,
      main="Pie Chart Depicting Ratio of Female and Male")
```

```
hist(customer_data$Age,
     col="blue",
     main="Histogram to Show Count of Age Class",
     xlab="Age Class",
     ylab="Frequency",
     labels=TRUE)
```

```
boxplot(customer_data$Age,
        col="red",
        main="Boxplot for Descriptive Analysis of Age")
```

```
summary(customer_data$Income)
hist(customer_data$Income,
     col="green",
     main="Histogram for Annual Income",
     xlab="Annual Income Class",
     ylab="Frequency",
     labels=TRUE)
```

```
df_scaled <- scale(df[-1])
head(df_scaled, n=2)
```

```
#create plot of number of clusters vs total within sum of squares
fviz_nbclust(df, kmeans, method = "wss")
```

```
#K-means cluster analysis
#kmeans() is used to obtain the final clustering solution.

set.seed(1234)
fit.km <- kmeans(df, 2, nstart=25)

#fit.km$size returns the number of items in each cluster
fit.km$size

#fit.km$centers returns the central value for each cluster
fit.km$centers

#As the centroids are quantified using the scaled data, the agg
aggregate(df[-1], by=list(cluster=fit.km$cluster),mean)
```

```
set.seed(1)
ggplot(customer_data, aes(x =Income, y = avg_order_freq)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name=" ",
                       breaks=c("1", "2", "3"),
                       labels=c("Cluster 1", "Cluster 2", "Cluster 3")) +
  ggtitle("Segments of Customers", subtitle = "Using K-means Clustering")
```