MARCH 5, 2023

# MGT7215- Marketing Analytics

## ASSIGNMENT 1: MARKET SEGMENTATION

NAME : KANIKA KUSHWAHA
Student ID : 40366629
Count : 1863

# Index

## 1.0 Introduction

It is necessary to comprehend customer-level heterogeneity or variations to grasp consumers (Hahn *et al.,* 2002). In much of the literature on applied econometrics, heterogeneity is often viewed as an incidental nuisance that needs to be resolved but is not the focus of the investigation. Instead, the focus is on assessing the average impact of policy variables (Allenby and Rossi, 1998). Models that ignore heterogeneity could provide very skewed results about the nature of correlations between variables and, as a result, lead to misleading conclusions (Ansari et al., 2000). Decisions to measure customer satisfaction at the segment level are only possible by understanding individual or market segment differences in the formation of overall satisfaction ratings and the consequent heterogeneity in the role of these various determinants, although aggregated market-level research to understand the determinants of customer satisfaction abounds (Allenby and Rossi, 1998; Wu *et al.,* 2006).

Methods for market research that worked well in a relationship-based environment with repeated choices would not work so well in a transaction-based environment with initial decisions, so these approaches should be complemented with explicitly dynamic techniques (Rust and Huang, 2014). The same IT elements that fuel the service economy are also changing the appropriate toolbox for marketing analytics, with a focus on computationally intensive data analysis of consumer databases (Rust and Huang, 2014).

This report aims to solve the challenge of capturing a new market by a restaurant chain by identifying the right demographic and psychographic features of the customers to increase profitability in a new market and find the best approaches using visualisations on Tableau and by performing cluster analysis.

## 2.0 Methodology

The survey dataset contains 1000 observations and 37 variables of the data stored in the marketing system which is related to the information about the attributes and characteristics of the restaurants as well as the demographics of the customers. The data set was checked for missing data and outliers using summary statistics and appropriate visualisations. Further, an initial exploration of the dataset was conducted by creating gantt chart representing a Likert Scale of the average rating of important characteristics a restaurant should have on Tableau. Further, bar charts and tables were used to analyse the occupations and social engagement of the customers.

By using a preset selection criterion, cluster analysis is a technique for grouping things that are remarkably similar together (Hair, 2009). The most popular method for developing embedded categorization methods is agglomerative hierarchical clustering (Murtagh and Contreras, 2012). As segmented models usually outperform aggregated models of customer behaviour, clustering analysis is used to identify relevant characteristics for creating targeting

segments(Allenby and Rossi, 1998). Hierarchical algorithms are of great interest for several application domains and are thought to produce better quality (Nazari *et al.,* 2015).

The data were scaled in order to get uniform measurements while calculating the Euclidean distance that determines the distance from an instance to a cluster centre (Davidson, 2002). All of this was performed along with the hierarchical clustering. Different linkage methods like complete, single, average and centroid  Nielsen (2016) were performed along with k-means clustering that included a different number of clusters determined by elbow plots that help determine the suitable no. of clusters by examining the kink of the elbow plots (see in Appendix – R Code). Because of its ease of development and quick execution, the K-means algorithm is a common method for identifying clusters (Davidson, 2002). Each method was examined by the total within-cluster variation that determines the total variation among clusters to find the best method with the least total within-cluster variation. The method with the least total within-cluster variation was selected and later, these cluster segments were separated into an excel sheet to carry out further customer segmentation analysis using Tableau. These clusters were then interpreted using appropriate visualisations to determine the base attributes of each cluster segment along with the demographic attributes or descriptors using visualizations from Tableau.

## 3.0 Results

### 3.1 Exploration of data using Tableau

When survey data was visualised using Tableau, it was discovered that the beverages and restaurant technology were the most important characteristics of the restaurant, with an average rating score of 4.7 and 4.3, respectively, followed by the brands (4.3), innovation (4.3), and location (4.2), as shown in the Gantt Chart in Figure 1. Figure 1 shows that TV, Facebook, and Instagram generated the maximum revenue with an average order size of 793 and 743, respectively, while Snapchat and Twitter generated the least revenue with 491 and 365, respectively. It can also be observed that those in the news industry spent the most, while those in the health industry spent the least.

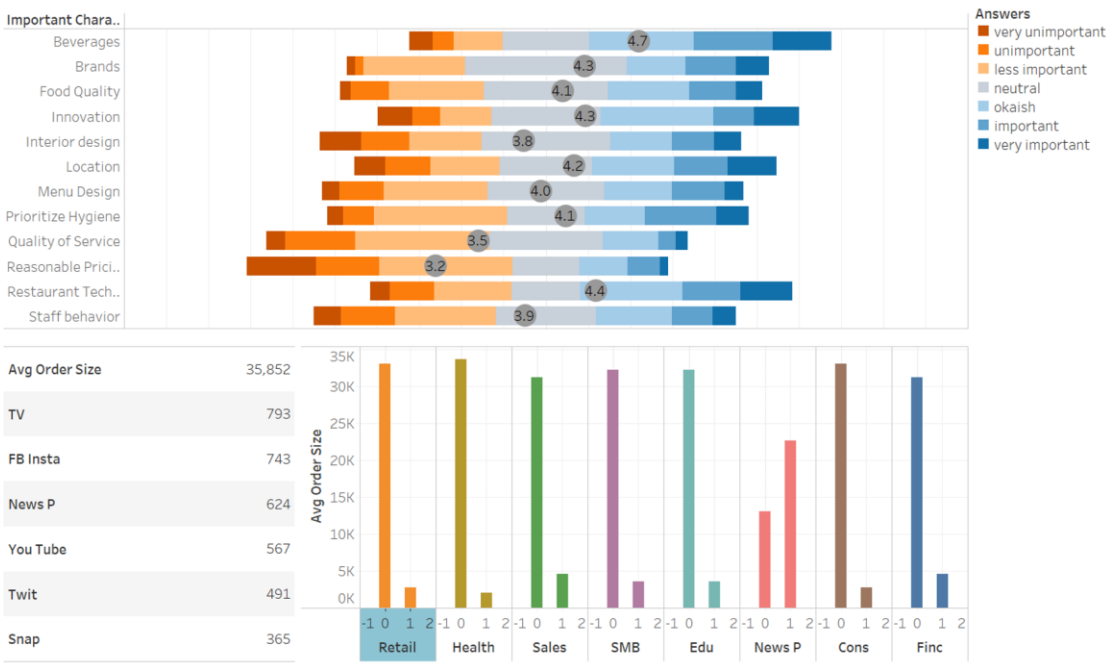



Figure 1: Dashboard for Summary Statistics

## 3.2 Model Interpretation

The survey_hclust_centr model with the centroid linkage method was proven to be the best among other linkage methods as it has the lowest total within-cluster variation (approx. 93,975) leading to the formation of 7 cluster segments.

Table 1 shows the overall summary of all methods.

| Model | Linkage Method | No. of clusters | Total Within Cluster Variation |
|---|---|---|---|
| survey_hclust_comp | Complete | 3 | 256052 |
| survey_hclust_single | Single | 6 | 112861.4 |
| survey_hclust_avg | Average | 3 | 256052 |
| survey_hclust_centr | Centroid | 7 | 93975.45 |

Table 1: Overall summary of methods

The result summary of the survey_hclust_centr  model is as follows:

```
Within cluster sum of squares by cluster:
[1] 13950.351  9276.457  8753.842 16502.787 13021.782 22923.079  9547.149
 (between_SS / total_SS =  89.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
>
>
> centr_segments = km_centr.out$cluster
> segment_centr_result <- cbind(cus_data, centr_segments)
>
> ##shows within cluster variation for every cluster
>
> km_centr.out$withinss
[1] 13950.351  9276.457  8753.842 16502.787 13021.782 22923.079  9547.149
>
> ##shows the total within cluster variations for different clusters
>
> km_centr.out$tot.withinss
[1] 93975.45
>
```

Figure 2: Result Summary of centroid linkage method(survey_hclust_centr)

## 3.3 Bases Variables Interpretation

While performing the attitude and behaviour analysis of the customer segments (bases variables), as shown in Figure 3, it was found that the average order frequency is highest for cluster 3 followed by clusters 2, 6,4 and 7. Clusters 5 and 1 have the least order frequency. On the other hand, the average order size for clusters 5 and 1 is the highest followed by clusters 7 and 4 with clusters 2 and 3 having the least average order size.

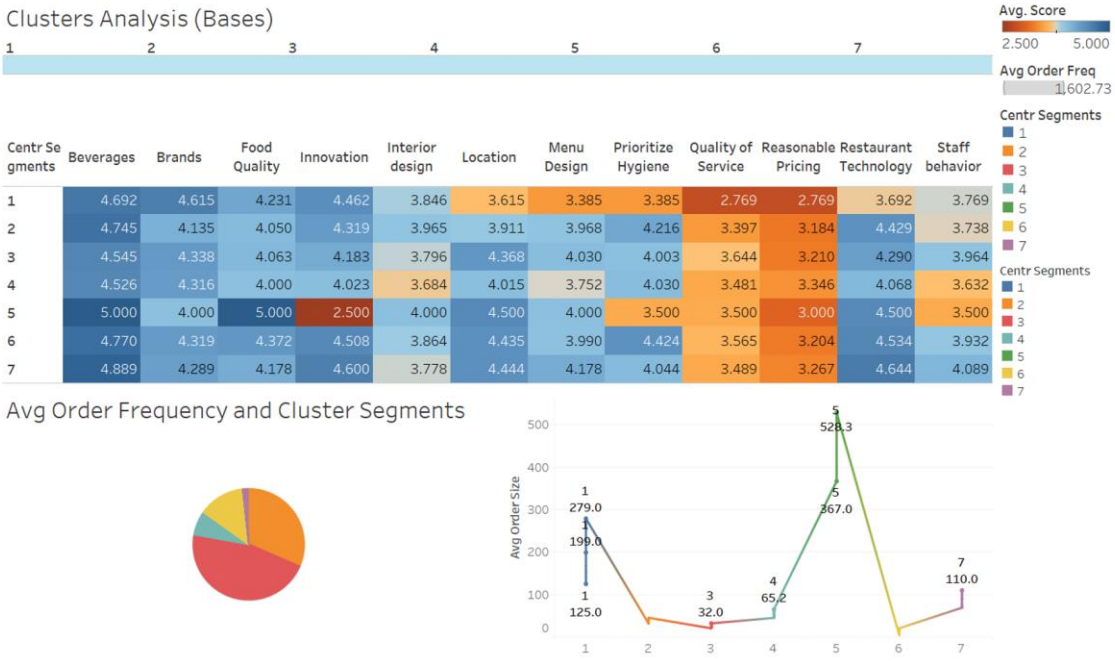



Figure 3: Dashboard for bases attributes of each cluster segment

Table 2 shows the priority characteristics and the least priority characteristics of each cluster segment.

| Cluster Name | Priority Characteristics | Least Priority Characteristics |
|---|---|---|
| Cluster 1 | beverages, innovation, brand and food quality, least order frequency, highest order size | quality of service and reasonable pricing |
| Cluster 2 | beverages, innovation, restaurant, hygiene , brand, high order frequency | quality of service and reasonable pricing |
| Cluster 3 | beverages, innovation, brand and location, high order frequency, low order size | quality of service and reasonable pricing |
| Cluster 4 | beverages, brands, medium order frequency, low order size | interior design, quality of service, staff behaviour and reasonable pricing |
| Cluster 5 | beverages, food quality, location and restaurant technology, least order frequency, highest order size | innovation and reasonable pricing |
| Cluster 6 | beverages, innovation, hygiene, restaurant technology, medium order frequency, least order size | quality of service and reasonable pricing |

| | | |
|---|---|---|
| Cluster 7 | beverages, innovation, restaurant technology, location, low order frequency, medium order size | quality of service and reasonable pricing |

Table 2: Bases Attributes of each cluster segment

## 3.4 Demographic Variables Interpretation

While performing the demographic analysis of the customer segments (descriptor variables), as shown in Figure 4, customers were discovered to be located in the United States, with clusters 2,3,4, and 6 indicating the most populous groups, while clusters 5 and 1 reflect consumers with a relatively small population.

In terms of spending habits, cluster 5 and cluster 1 were the group of customers with the highest expenditure and they belonged to age groups 36 – 38 and 25 – 29 years respectively.

Further, in terms of education, customers with Masters's degree were more likely to spend than people with a bachelor's degree. This is in line with (Adhikari et al., 2013) as they state that consumers in the hospitality industry are significantly more willing to spend money on their experiential consumption. This seems to be true given that their study research group included upper-class people with higher education and higher incomes.

It was also examined that the females are more likely to spend than males which can be seen in all cluster segments as female representation is higher in all cluster segments as compared to men. This is corroborated by (Koo et al., 1999) who report that in their research, female respondents rated the best restaurants higher than male respondents, thereby determining their fondness of restaurants. From Figure 4, it can be seen that cluster 5 has only females and is also the cluster segment with the highest average order size.

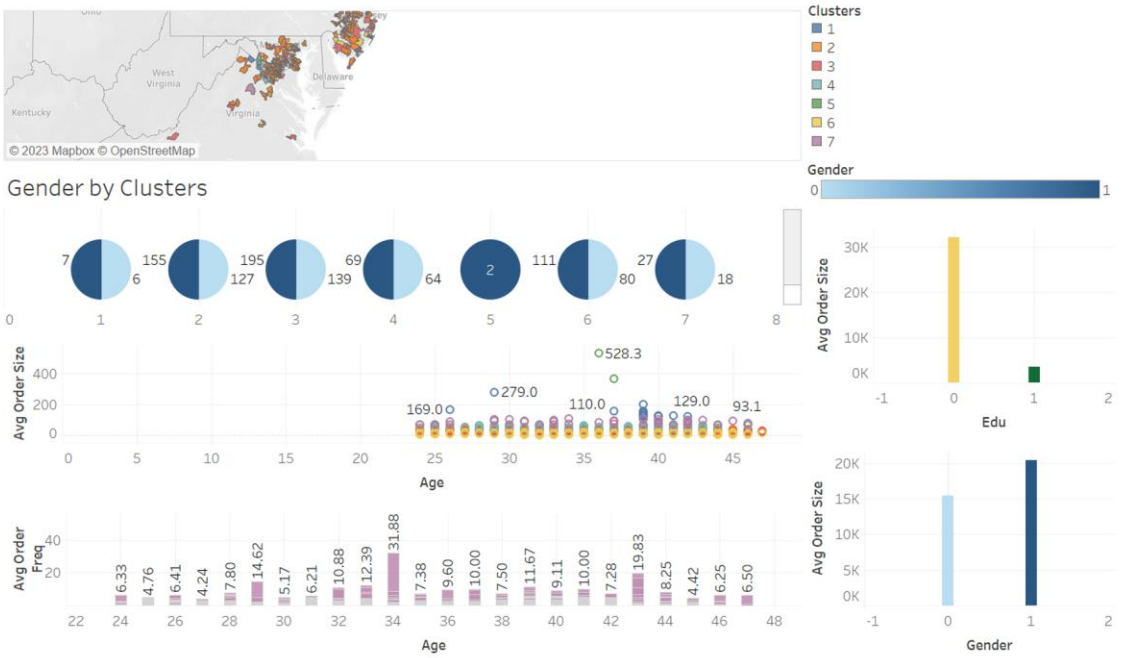



Figure 4: Dashboard for descriptor attributes of each cluster segment

## 3.5 Findings

As cluster 2 and 3 represents the group of people that prefer beverages, brand, restaurant hygiene, innovation and location and are the ones with the highest spending frequency, marketers should consider these segments for targeting. They also represent the major population of the surveyed customers. This is supported by Cullen (2004) who claims that the main decision factors consumers make when choosing restaurants are brand, restaurant cleanliness and location.

Although the average order size of these cluster segments(2 & 3) is less as compared to other cluster segments, the average order frequency for these segments is the highest which makes it comparable. Further, as stated by Kim et al. (2006), customers are less inclined to visit a restaurant if they think the costs are excessively expensive. Therefore, prices should be feasible and fair for customers to visit the restaurant often. This highest average order frequency attribute demonstrates the customer segment's likeness towards the restaurant chain as they are frequently ordering from the same restaurant and have the chance of becoming loyal customers. These cluster segments (2 & 3), therefore, make them more important and can help in increasing the revenue as well as becoming profitable for the restaurant chain in Belfast. According to Kim et al., (2006), other factors like food quality and staff behaviour are also considered important when taking into consideration the dining experience of a restaurant and can be used for enhancing customer experience.

## 4.0 Conclusion

Nowadays, the main emphasis of many statistical marketing applications is the modelling of customer heterogeneity (Allenby and Rossi, 1998). As a result, there are a lot of new research topics and methods being spawned by IT and the economic revolution to study (Rust and Huang, 2014). Behavioural attributes like brand, restaurant hygiene, innovation and location are the most important factors for customers when considering a restaurant. Demographic factors like gender and education also play an important role while determining targeting segments as they have a positive influence on revenue for restaurant chains. This can be confirmed as females are more likely to order at a high frequency and spend more than males. Moreover, customers with a higher level of education like Masters are more likely to spend in restaurants. Marketers can focus on these factors to form strategies to increase their chances of converting customers to enhance the restaurant chain's overall generated revenue and help increase the likelihood for the customer to order at a higher frequency and for them to convert to loyal customers.

## 5.0 Limitations

The limitation of this study is that the data being used is survey data. Although survey research techniques are a useful method for identifying trends, they are always predicated on a number of prior hypotheses regarding the consistency of consumers' perceptions of factors in relation to their choices and engagement with particular restaurants and restaurant categories (Clark and Wood, 1999).

## 6.0 Bibliography


Adhikari, A., Basu, A. and Raj, S.P. (2013) 'Pricing of experience products under consumer heterogeneity', International Journal of Hospitality Management, 33(1), pp. 6–18. Available at: https://doi.org/10.1016/j.ijhm.2013.01.002.

Allenby, G.M. and Rossi, P.E. (1998) 'Marketing models of consumer heterogeneity', Journal of Econometrics, 89(1), pp. 57–78. Available at: https://doi.org/https://doi.org/10.1016/S0304-4076(98)00055-4.

Ansari, A., Jedidi, K. and Jagpal, S. (2000) 'A hierarchical Bayesian methodology for treating heterogeneity in structural equation models, Marketing Science, 19(4), pp. 328–347.

Clark, M.A. and Wood, R.C. (1999) Loyalty in the restaurant industry Consumer loyalty in the restaurant industry A preliminary exploration of the issues, British Food Journal. # MCB University Press.

Cullen, F. (2004) 'Factors influencing restaurant selection in Dublin', Journal of Foodservice Business Research, 7(2), pp. 53–85. Available at: https://doi.org/10.1300/J369v07n02_05.

Davidson, I. (2002) 'Understanding K-means non-hierarchical clustering', Computer Science Department of State University of New York (SUNY), Albany [Preprint].

Hahn, C. et al. (2002) 'Capturing Customer Heterogeneity using a Finite Mixture PLS Approach', Schmalenbach Business Review, 54(3), pp. 243–269. Available at: https://doi.org/10.1007/BF03396655.

Hair, J.F. (2009) 'Multivariate data analysis'.

Kim, W.G., Lee, Y.-K. and Yoo, Y.-J. (2006) 'Predictors of relationship quality and relationship outcomes in luxury restaurants', Journal of Hospitality & Tourism Research, 30(2), pp. 143–169.

Koo, L.C., Tao, F.K.C. and Yeung, J.H.C. (1999) 'Preferential segmentation of restaurant attributes through conjoint analysis', international Journal of Contemporary Hospitality management [Preprint].

Murtagh, F. and Contreras, P. (2012) 'Algorithms for hierarchical clustering: An overview', Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp. 86–97. Available at: https://doi.org/10.1002/widm.53.

Nazari, Z. et al. (2015) 'A new hierarchical clustering algorithm', in 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), pp. 148–152. Available at: https://doi.org/10.1109/ICIIBMS.2015.7439517.

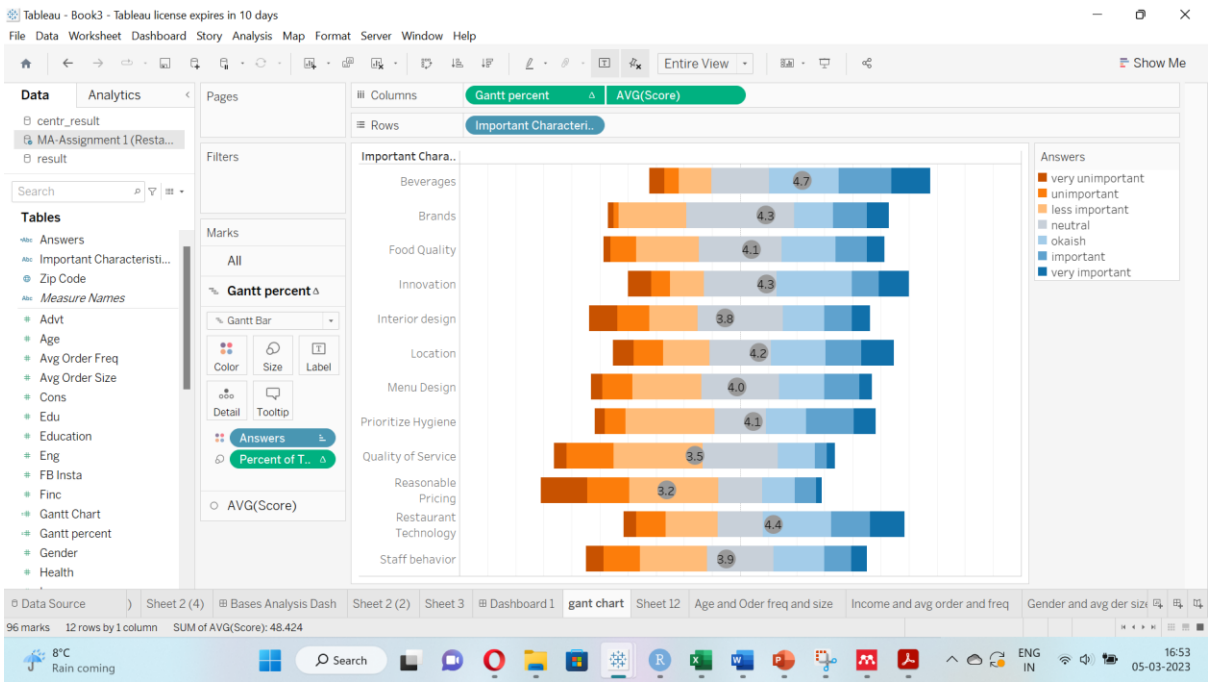
Nielsen, F. (2016) 'Hierarchical Clustering', in F. Nielsen (ed.) Introduction to HPC with MPI for Data Science. Cham: Springer International Publishing, pp. 195–211. Available at: https://doi.org/10.1007/978-3-319-21903-5_8.

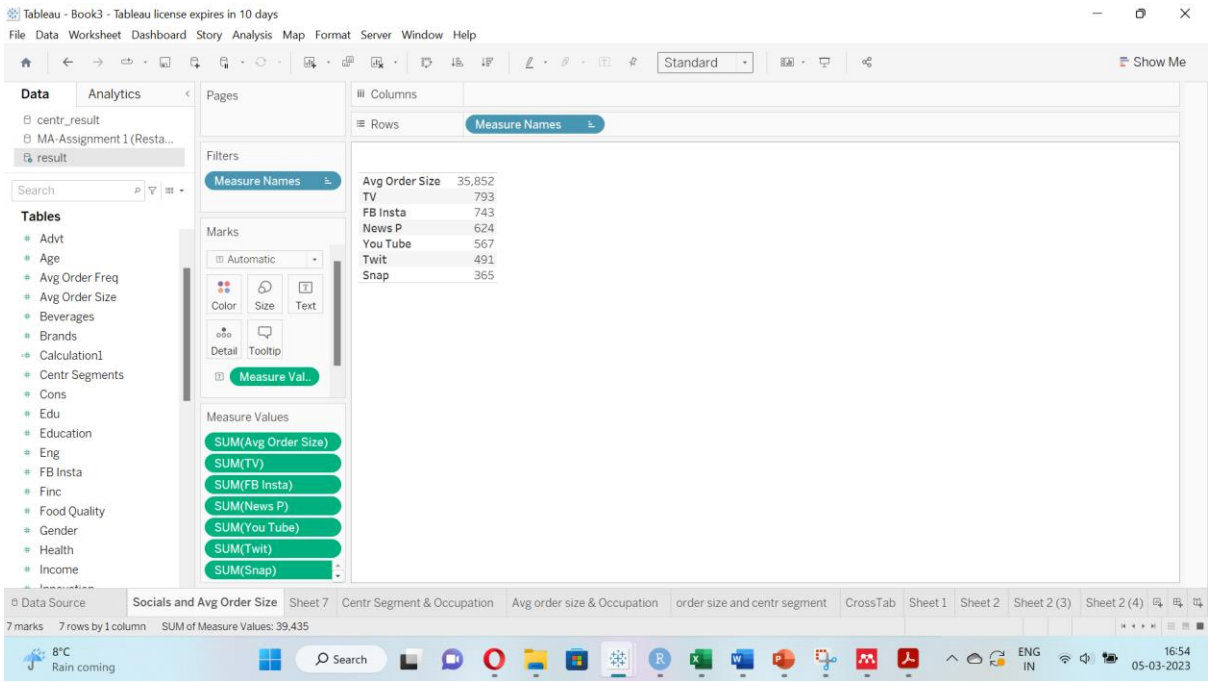
Rust, R.T. and Huang, M.-H. (2014) ‘The service revolution and the transformation of marketing science’, Marketing Science, 33(2), pp. 206–221.

Wu, J. et al. (2006) ‘A latent structure factor analytic approach for customer satisfaction measurement’, Marketing Letters, 17(3), pp. 221–238. Available at: https://doi.org/10.1007/s11002-006-7638-1.

## 7.0 Appendix – Tableau

## 7.1 Summary Statistics Tableau



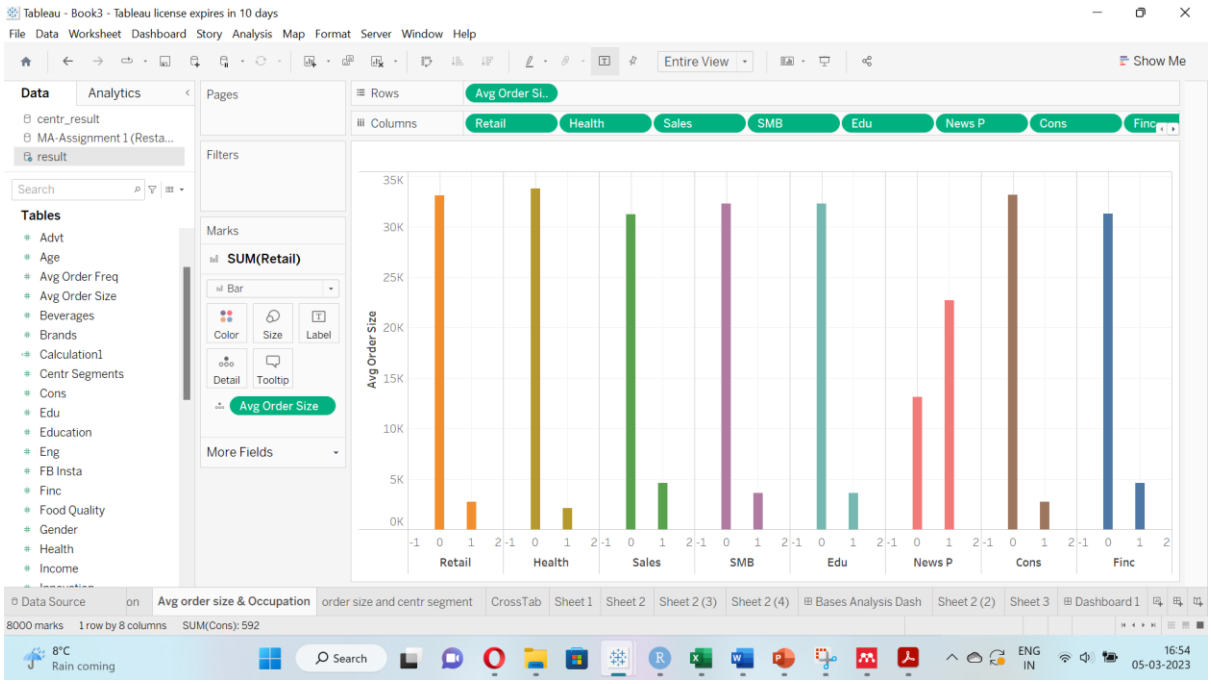
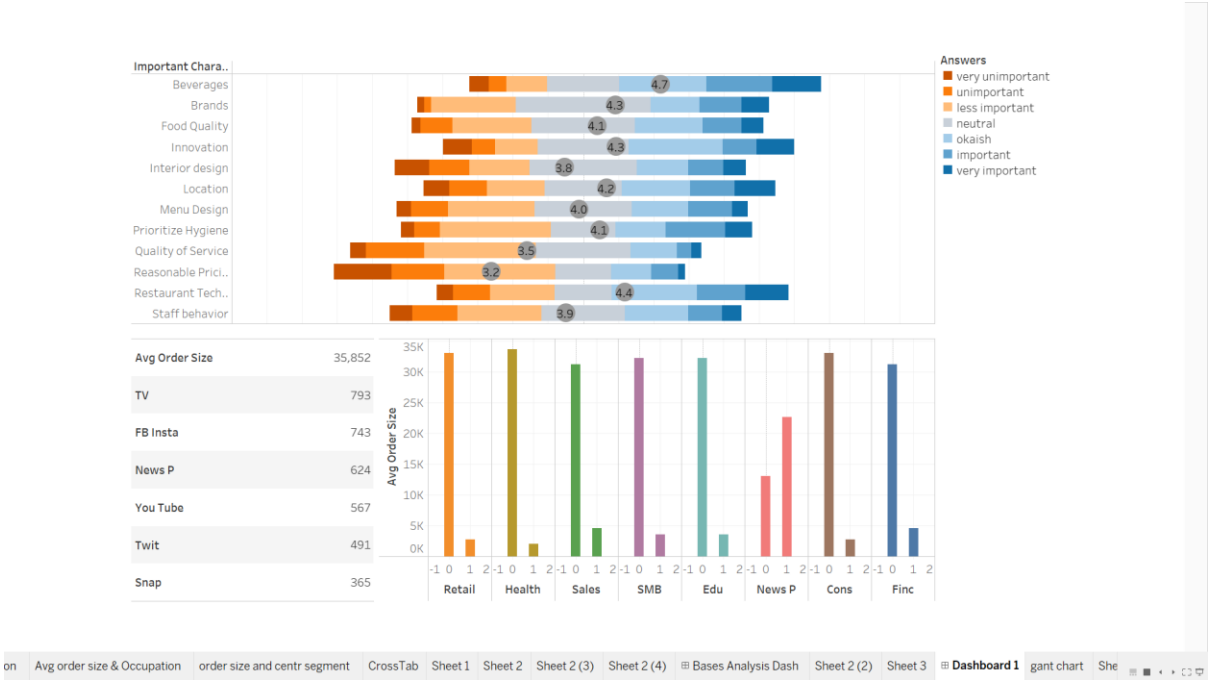
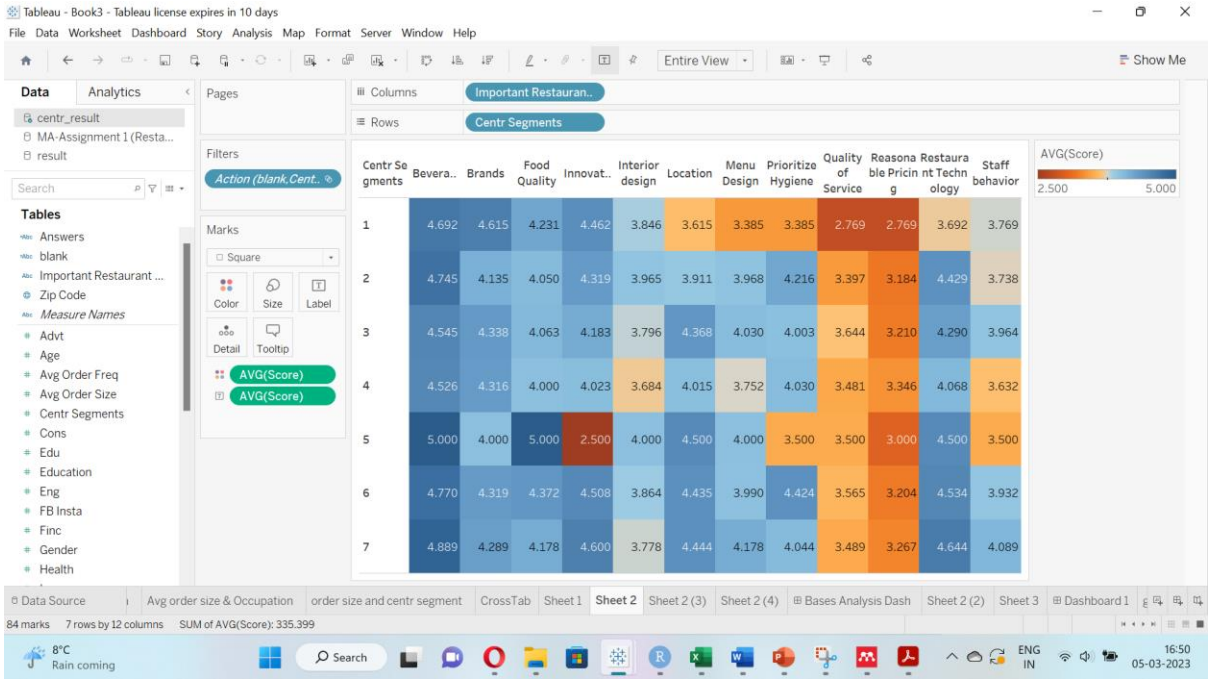
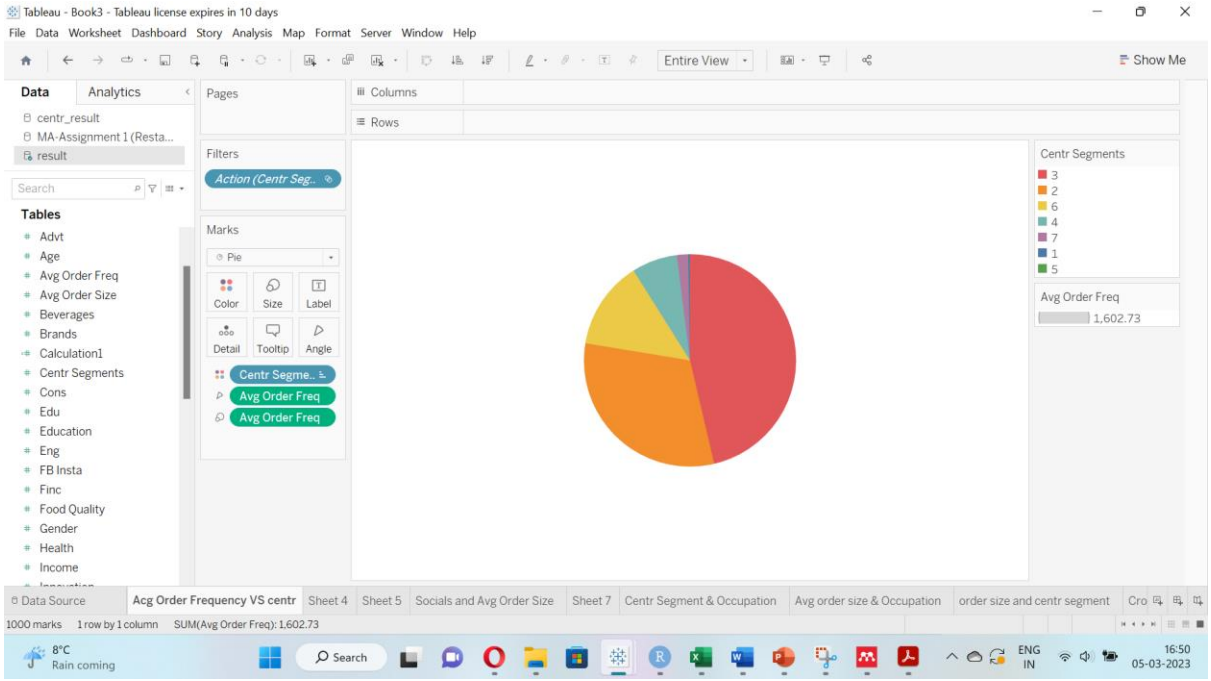


Tableau - Book3 - Tableau license expires in 10 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Entire View
Show Me
Data
Analytics
centr_result
MA-Assignment 1 (Resta...
result
Search
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Calculation1
Centr Segments
Cons
Edu
Education
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Pages
Filters
Marks
SUM(Retail)
Bar
Color
Size
Label
Detail
Tooltip
Avg Order Size
More Fields
Rows
Avg Order Si..
Columns
Retail
Health
Sales
SMB
Edu
News P
Cons
Finc
Avg Order Size
35K
30K
25K
20K
15K
10K
5K
0K
Retail
Health
Sales
SMB
Edu
News P
Cons
Finc
Data Source
on
Avg order size & Occupation
order size and centr segment
CrossTab
Sheet 1
Sheet 2
Sheet 2 (3)
Sheet 2 (4)
Bases Analysis Dash
Sheet 2 (2)
Sheet 3
Dashboard 1
8000 marks
1 row by 8 columns
SUM(Cons): 592
8°C
Rain coming
Search
ENG
IN
16:54
05-03-2023


Important Chara..
Beverages
Brands
Food Quality
Innovation
Interior design
Location
Menu Design
Prioritize Hygiene
Quality of Service
Reasonable Prici..
Restaurant Tech..
Staff behavior
4.7
4.3
4.1
4.3
3.8
4.2
4.0
4.1
3.5
3.2
4.4
3.9
Answers
very unimportant
unimportant
less important
neutral
okaish
important
very important
Avg Order Size 35,852
TV 793
FB Insta 743
News P 624
You Tube 567
Twit 491
Snap 365
Avg Order Size
35K
30K
25K
20K
15K
10K
5K
0K
Retail
Health
Sales
SMB
Edu
News P
Cons
Finc
on
Avg order size & Occupation
order size and centr segment
CrossTab
Sheet 1
Sheet 2
Sheet 2 (3)
Sheet 2 (4)
Bases Analysis Dash
Sheet 2 (2)
Sheet 3
Dashboard 1
gant chart
She

## 7.2 Bases Dashboard Visualisations



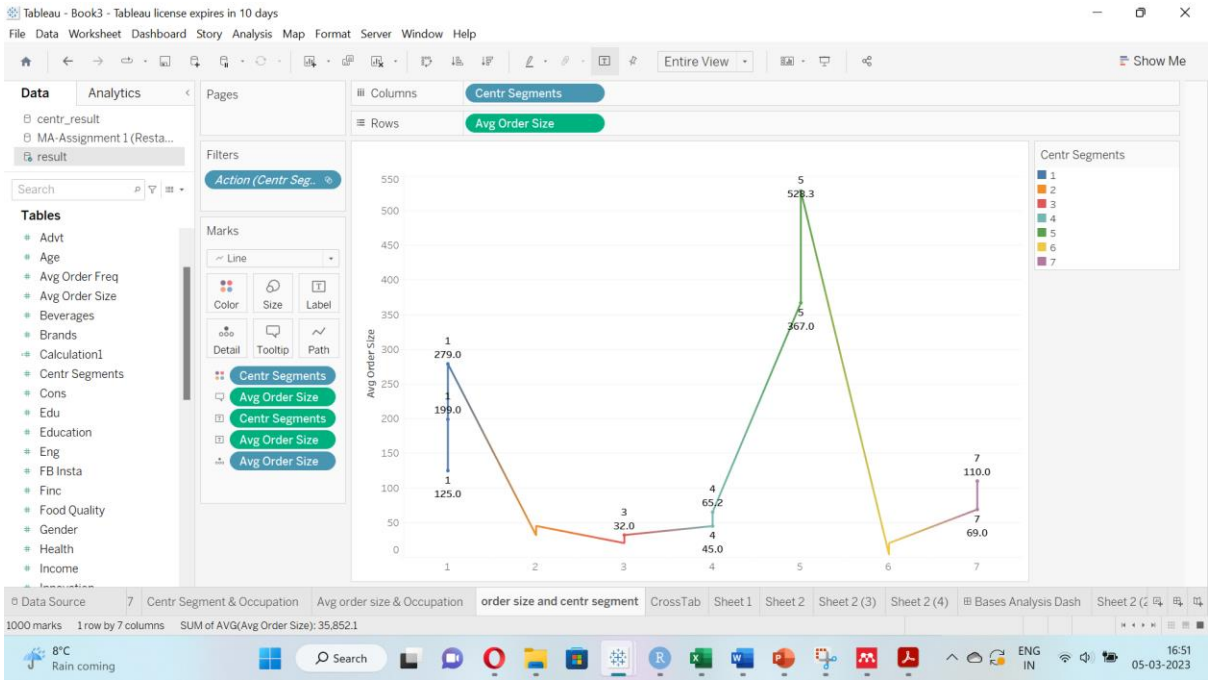
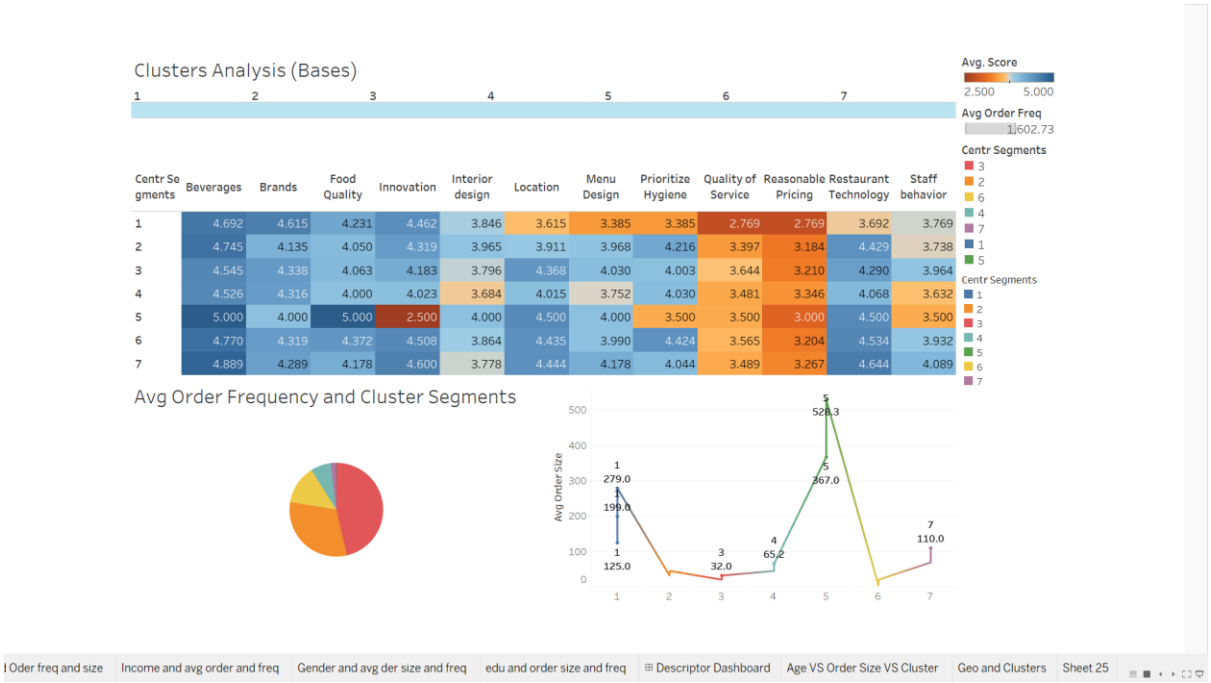
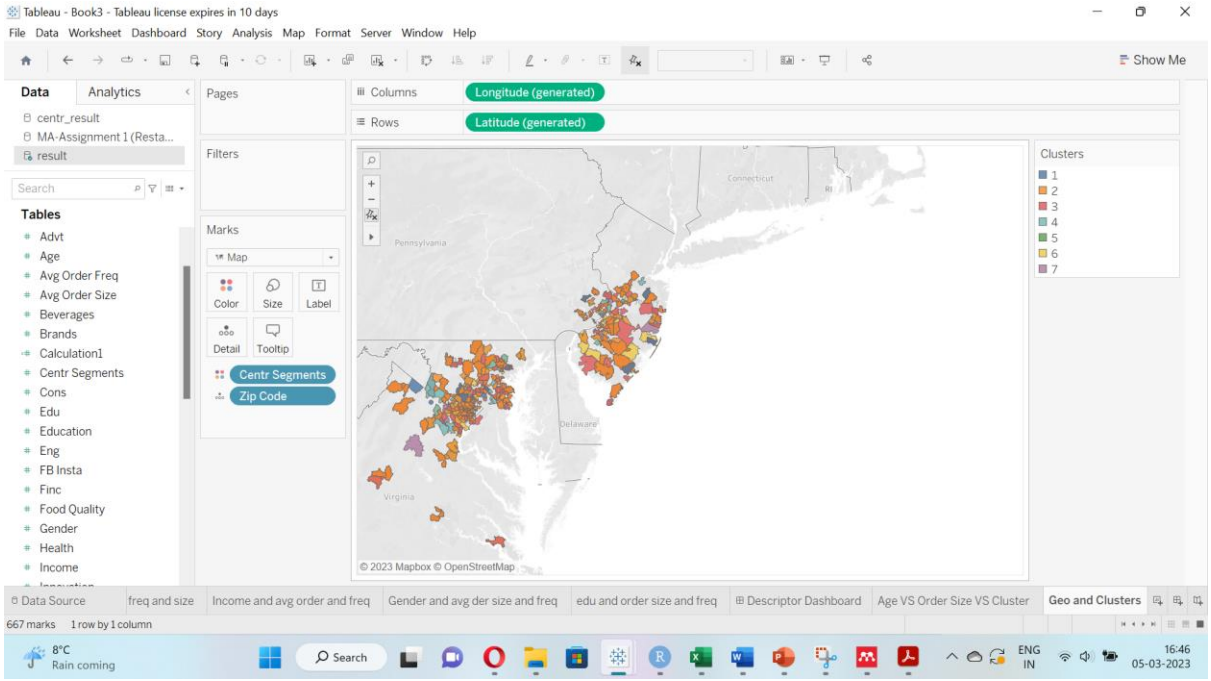
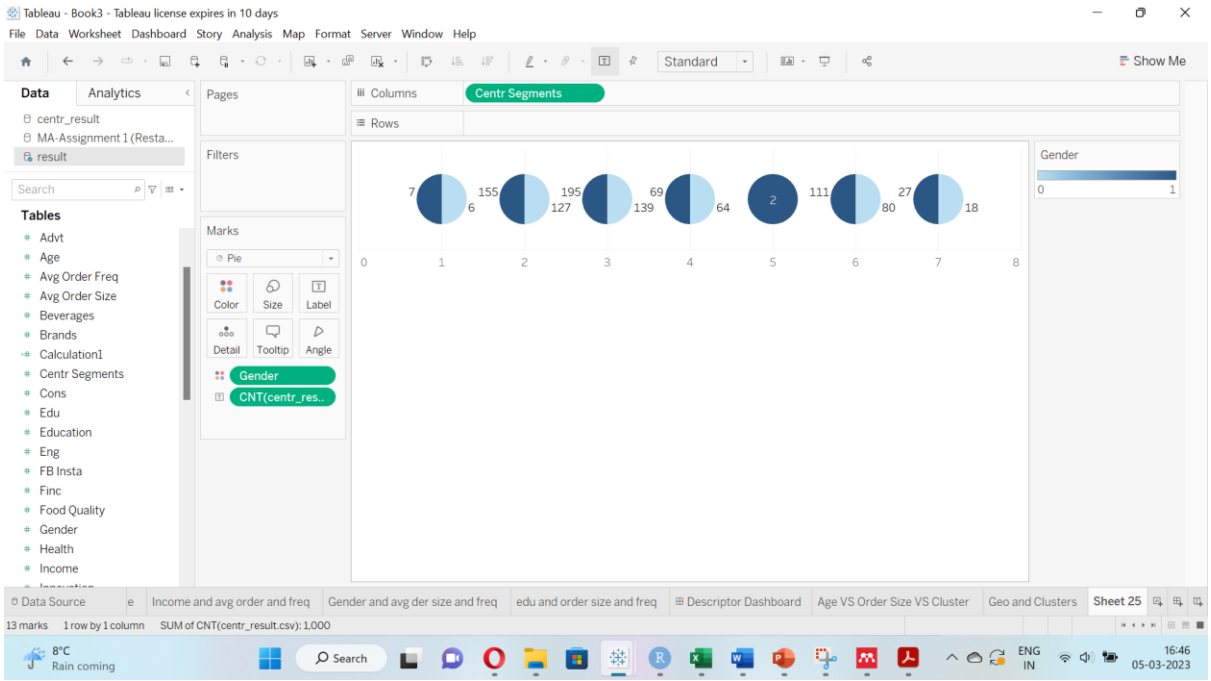


Tableau - Book3 - Tableau license expires in 10 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Entire View
Show Me
Data
Analytics
centr_result
MA-Assignment 1 (Resta...
result
Search
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Calculation1
Centr Segments
Cons
Edu
Education
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Pages
Filters
Action (Centr Seg..
Marks
Line
Color
Size
Label
Detail
Tooltip
Path
Centr Segments
Avg Order Size
Centr Segments
Avg Order Size
Avg Order Size
Columns
Centr Segments
Rows
Avg Order Size
Centr Segments
1
2
3
4
5
6
7
Avg Order Size
550
500
450
400
350
300
250
200
150
100
50
0
1
2
3
4
5
6
7
279.0
199.0
125.0
32.0
65.2
45.0
528.3
367.0
110.0
69.0
Data Source
7
Centr Segment & Occupation
Avg order size & Occupation
order size and centr segment
CrossTab
Sheet 1
Sheet 2
Sheet 2 (3)
Sheet 2 (4)
Bases Analysis Dash
1000 marks
1 row by 7 columns
SUM of AVG(Avg Order Size): 35,852.1
8°C
Rain coming
Search
ENG
IN
16:51
05-03-2023


Clusters Analysis (Bases)
1
2
3
4
5
6
7
Avg. Score
2.500
5.000
Avg Order Freq
1,602.73
Centr Segments
3
2
6
4
7
1
5
Centr Segments
1
2
3
4
5
6
7
Centr Segments
Beverages
Brands
Food Quality
Innovation
Interior design
Location
Menu Design
Prioritize Hygiene
Quality of Service
Reasonable Pricing
Restaurant Technology
Staff behavior
1 4.692 4.615 4.231 4.462 3.846 3.615 3.385 3.385 2.769 2.769 3.692 3.769
2 4.745 4.135 4.050 4.319 3.965 3.911 3.968 4.216 3.397 3.184 4.429 3.738
3 4.545 4.338 4.063 4.183 3.796 4.368 4.030 4.003 3.644 3.210 4.290 3.964
4 4.526 4.316 4.000 4.023 3.684 4.015 3.752 4.030 3.481 3.346 4.068 3.632
5 5.000 4.000 5.000 2.500 4.000 4.500 4.000 3.500 3.500 3.000 4.500 3.500
6 4.770 4.319 4.372 4.508 3.864 4.435 3.990 4.424 3.565 3.204 4.534 3.932
7 4.889 4.289 4.178 4.600 3.778 4.444 4.178 4.044 3.489 3.267 4.644 4.089
Avg Order Frequency and Cluster Segments
Avg Order Size
500
400
300
200
100
0
1
2
3
4
5
6
7
279.0
199.0
125.0
32.0
65.2
528.3
367.0
110.0
Oder freq and size
Income and avg order and freq
Gender and avg der size and freq
edu and order size and freq
Descriptor Dashboard
Age VS Order Size VS Cluster
Geo and Clusters
Sheet 25

## 7.3 Descriptor Dashboard Visualisations



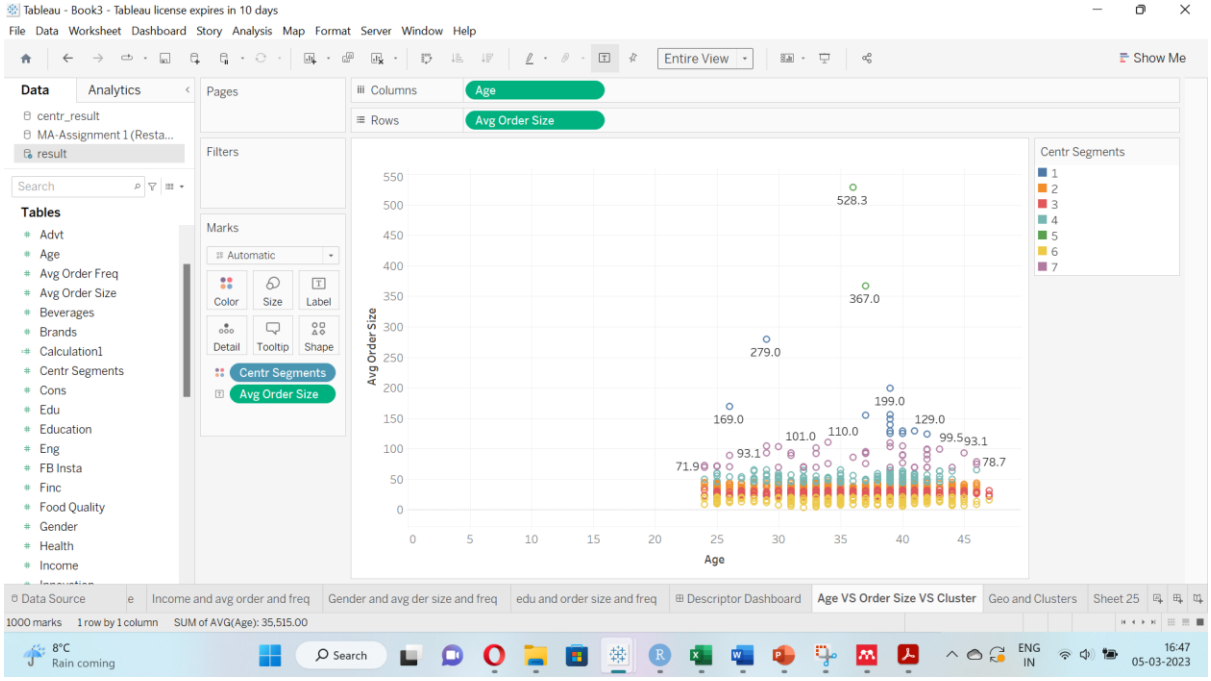


Tableau - Book3 - Tableau license expires in 10 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Entire View
Show Me
Data
Analytics
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Shape
Centr Segments
Avg Order Size
Columns
Age
Rows
Avg Order Size
Centr Segments
1
2
3
4
5
6
7
Avg Order Size
Age
528.3
367.0
279.0
199.0
169.0
129.0
110.0
101.0
99.5
93.1
93.1
78.7
71.9
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Calculation1
Centr Segments
Cons
Edu
Education
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Data Source
Income and avg order and freq
Gender and avg der size and freq
edu and order size and freq
Descriptor Dashboard
Age VS Order Size VS Cluster
Geo and Clusters
Sheet 25
1000 marks
1 row by 1 column
SUM of AVG(Age): 35,515.00
8°C
Rain coming
Search
ENG
IN
16:47
05-03-2023

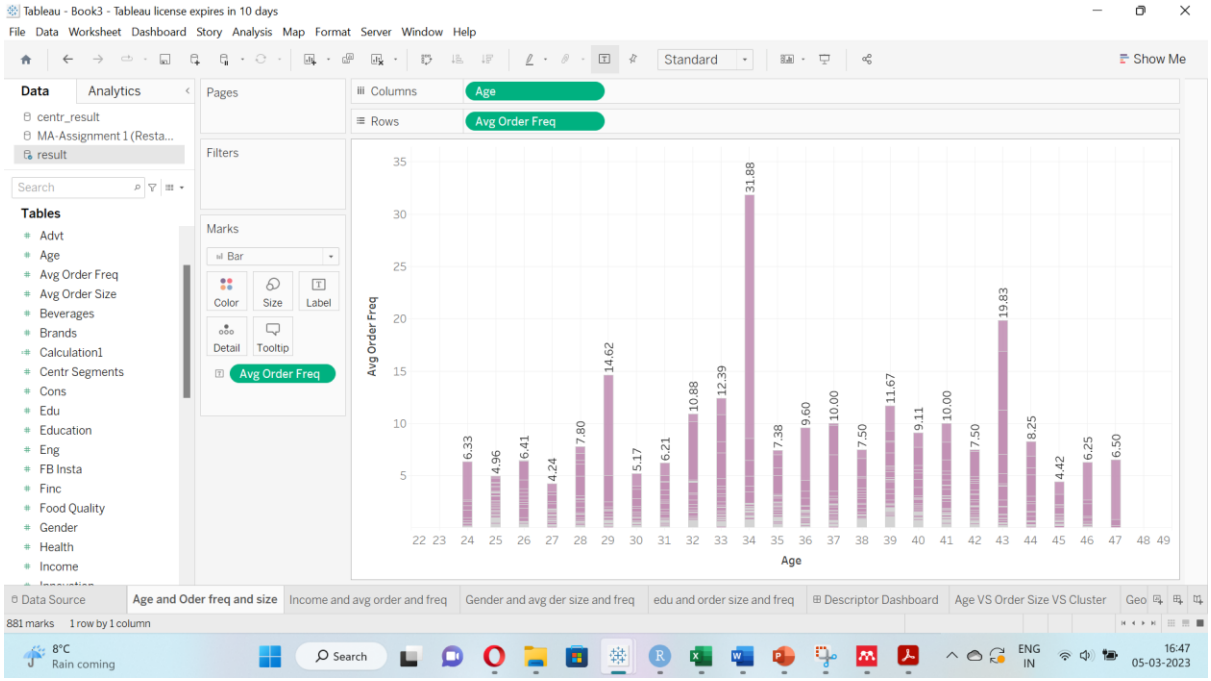


Tableau - Book3 - Tableau license expires in 10 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
Pages
Filters
Marks
Bar
Color
Size
Label
Detail
Tooltip
Avg Order Freq
Columns
Age
Rows
Avg Order Freq
Avg Order Freq
Age
6.33
4.96
6.41
4.24
7.80
14.62
5.17
6.21
10.88
12.39
31.88
7.38
9.60
10.00
7.50
11.67
9.11
10.00
7.50
19.83
8.25
4.42
6.25
6.50
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Calculation1
Centr Segments
Cons
Edu
Education
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Data Source
Age and Oder freq and size
Income and avg order and freq
Gender and avg der size and freq
edu and order size and freq
Descriptor Dashboard
Age VS Order Size VS Cluster
881 marks
1 row by 1 column
8°C
Rain coming
Search
ENG
IN
16:47
05-03-2023

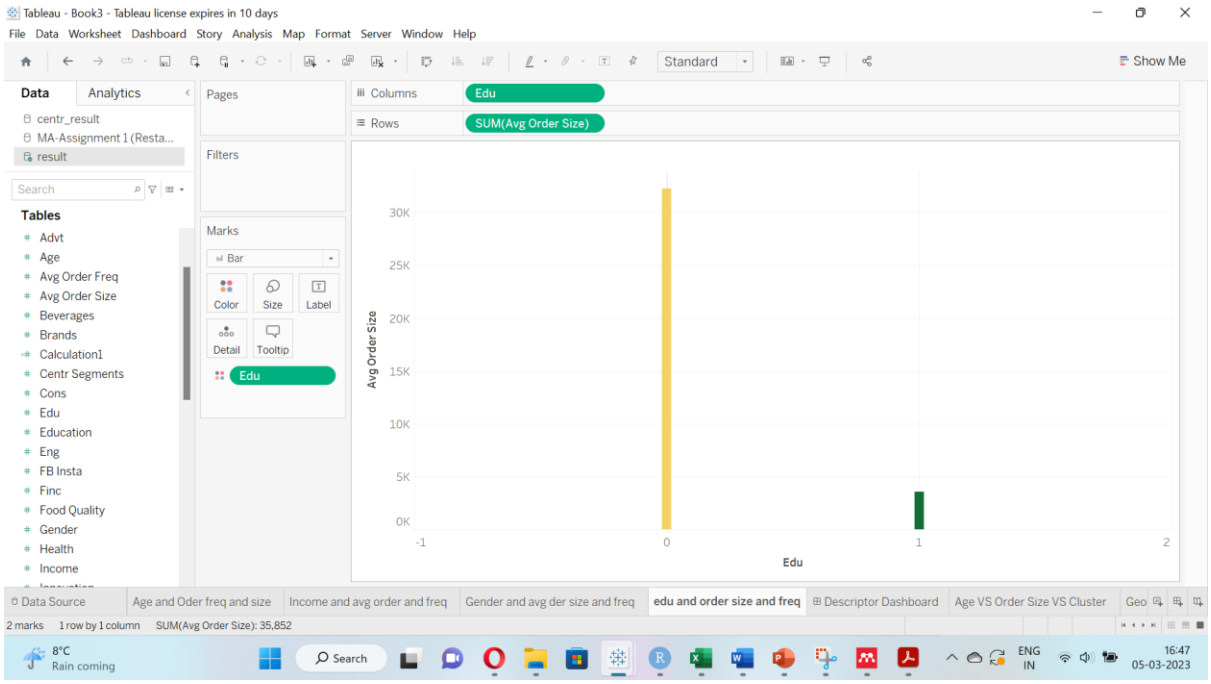


Tableau - Book3 - Tableau license expires in 10 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
centr_result
MA-Assignment 1 (Resta...
result
Search
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Calculation1
Centr Segments
Cons
Edu
Education
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Pages
Filters
Marks
Bar
Color
Size
Label
Detail
Tooltip
Edu
Columns
Edu
Rows
SUM(Avg Order Size)
Avg Order Size
30K
25K
20K
15K
10K
5K
0K
-1
0
1
2
Edu
Data Source
Age and Oder freq and size
Income and avg order and freq
Gender and avg der size and freq
edu and order size and freq
Descriptor Dashboard
Age VS Order Size VS Cluster
2 marks
1 row by 1 column
SUM(Avg Order Size): 35,852
8°C
Rain coming
Search
ENG
IN
16:47
05-03-2023

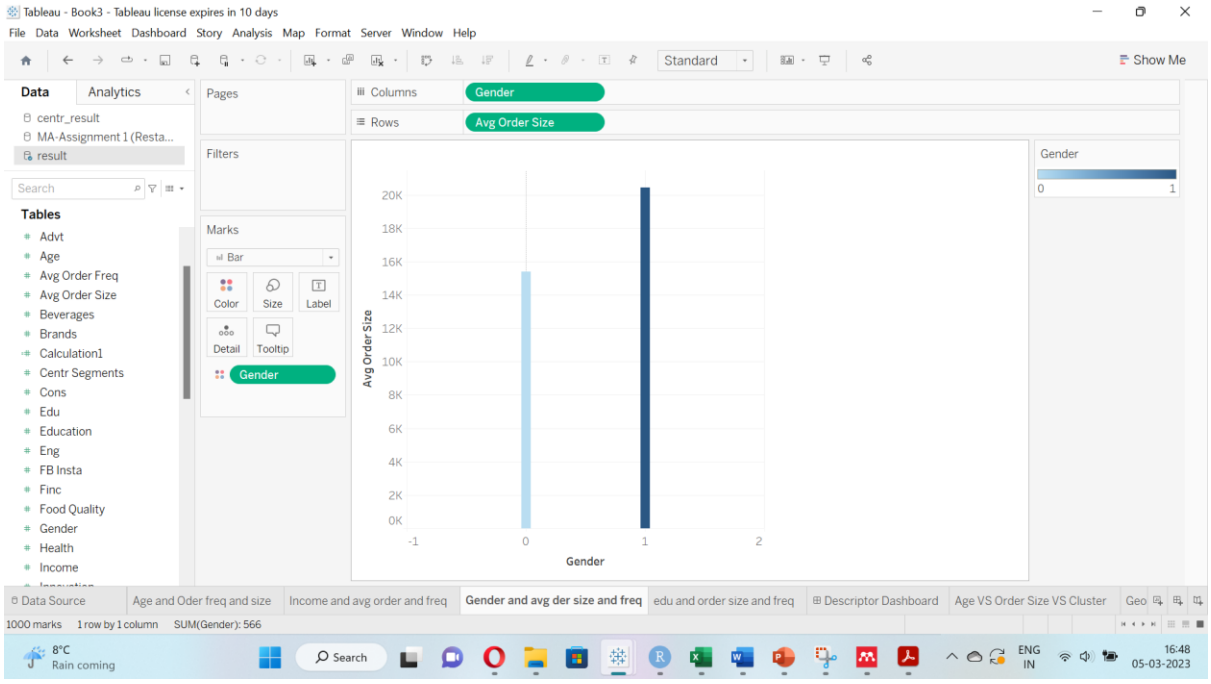
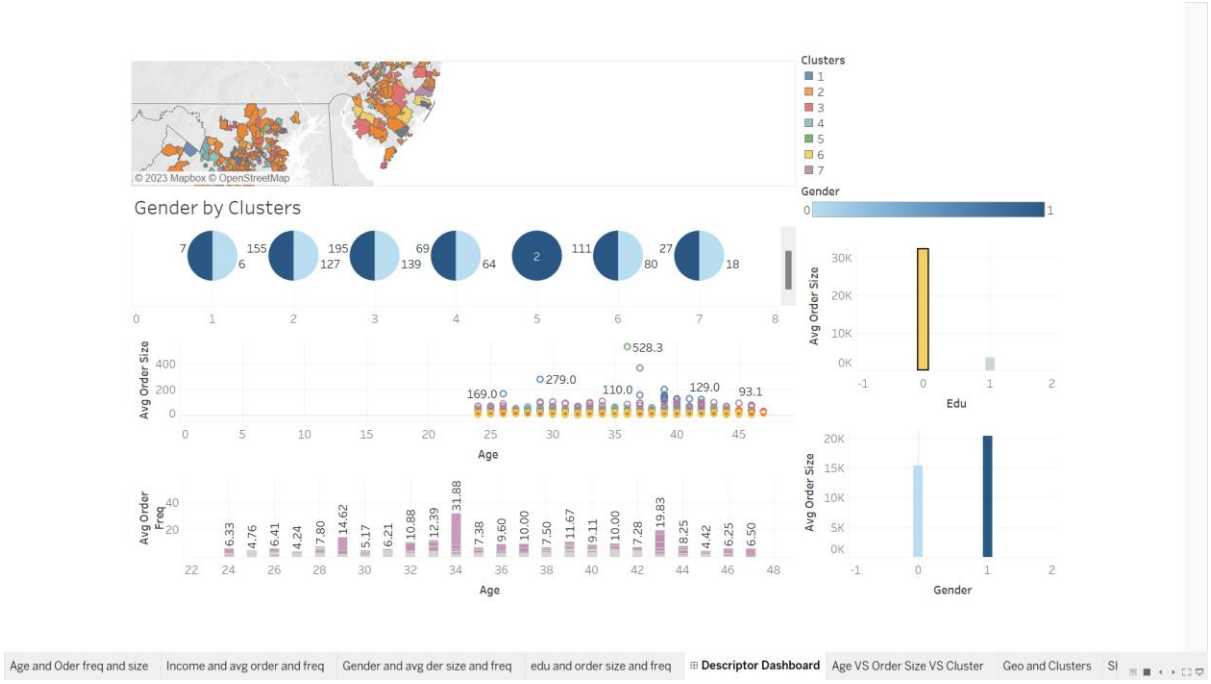


Tableau - Book3 - Tableau license expires in 10 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
centr_result
MA-Assignment 1 (Resta...
result
Search
Tables
Advt
Age
Avg Order Freq
Avg Order Size
Beverages
Brands
Calculation1
Centr Segments
Cons
Edu
Education
Eng
FB Insta
Finc
Food Quality
Gender
Health
Income
Pages
Filters
Marks
Bar
Color
Size
Label
Detail
Tooltip
Gender
Columns
Gender
Rows
Avg Order Size
Gender
0
1
Avg Order Size
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
-1
0
1
2
Gender
Data Source
Age and Oder freq and size
Income and avg order and freq
Gender and avg der size and freq
edu and order size and freq
Descriptor Dashboard
Age VS Order Size VS Cluster
1000 marks
1 row by 1 column
SUM(Gender): 566
8°C
Rain coming
Search
ENG
IN
16:48
05-03-2023


Clusters
1
2
3
4
5
6
7
© 2023 Mapbox © OpenStreetMap
Gender by Clusters
Gender
0
1
Avg Order Size
Edu
Age
Avg Order Freq
Gender
Age and Oder freq and size
Income and avg order and freq
Gender and avg der size and freq
edu and order size and freq
Descriptor Dashboard
Age VS Order Size VS Cluster
Geo and Clusters

## 8.0 Appendix - R Code

```
##installing libraries

install.packages("readxl")

## reading libraries
library("readxl")

### checking working directory
getwd()


##setting working directory
setwd("C:/Users/kanik/Desktop/Marketing Analytics")


##reading in dataset
cus_data <- read_xlsx("Restaurant Data.xlsx")


## removing observation column from dataset
cus_data <- cus_data[-1]

##setting up seed so that data doesn't change once code is run again
set.seed(14)

###Checking for null values
is.na(cus_data)
summary(cus_data)
```

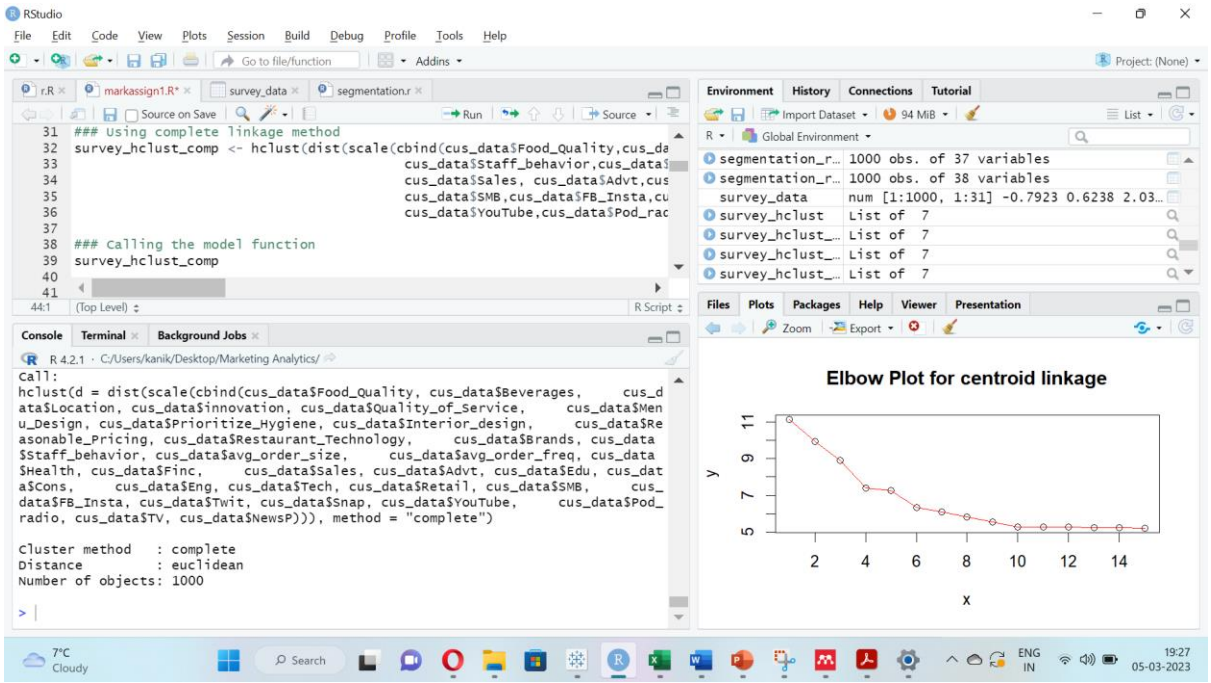
##Scaling the dataset and calculating euclidean distance

### Using complete linkage method

```
survey_hclust_comp <-
hclust(dist(scale(cbind(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_
data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,


cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Hea
lth,cus_data$Finc,

                    cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail
,

                    cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,


cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP))),
method="complete")
```

### Calling the model function

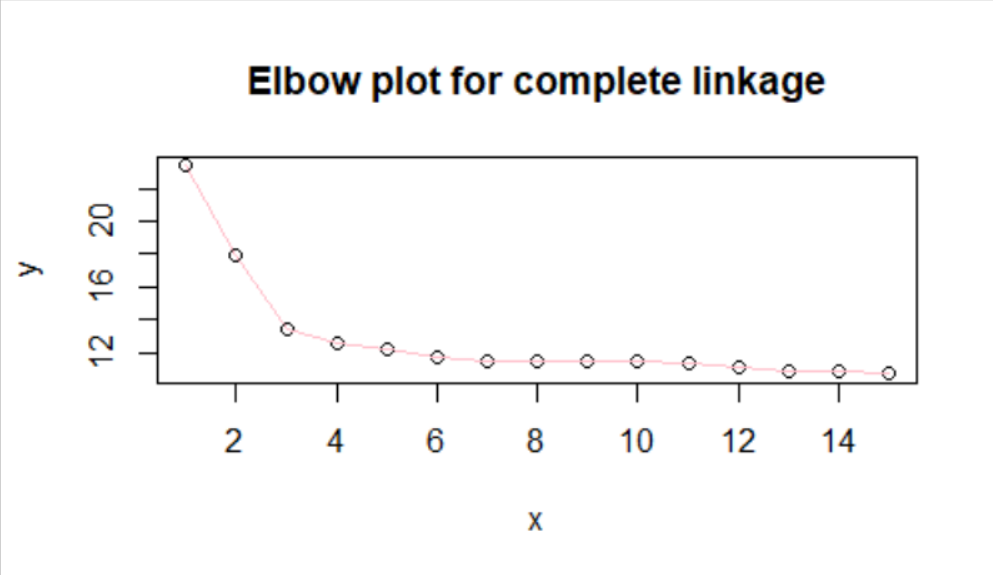
```
survey_hclust_comp
```



### creating elbow plots that determines the no. of clusters;

##Usually the no. present at the kink of the chart is used to determine no. of clusters

```
x <- c(1:15)

y <- sort(survey_hclust_comp$height, decreasing = TRUE)[1:15] ## average distance between clusters

plot(x,y, main = "Elbow plot for complete linkage");lines(x,y,col = "PINK")
```



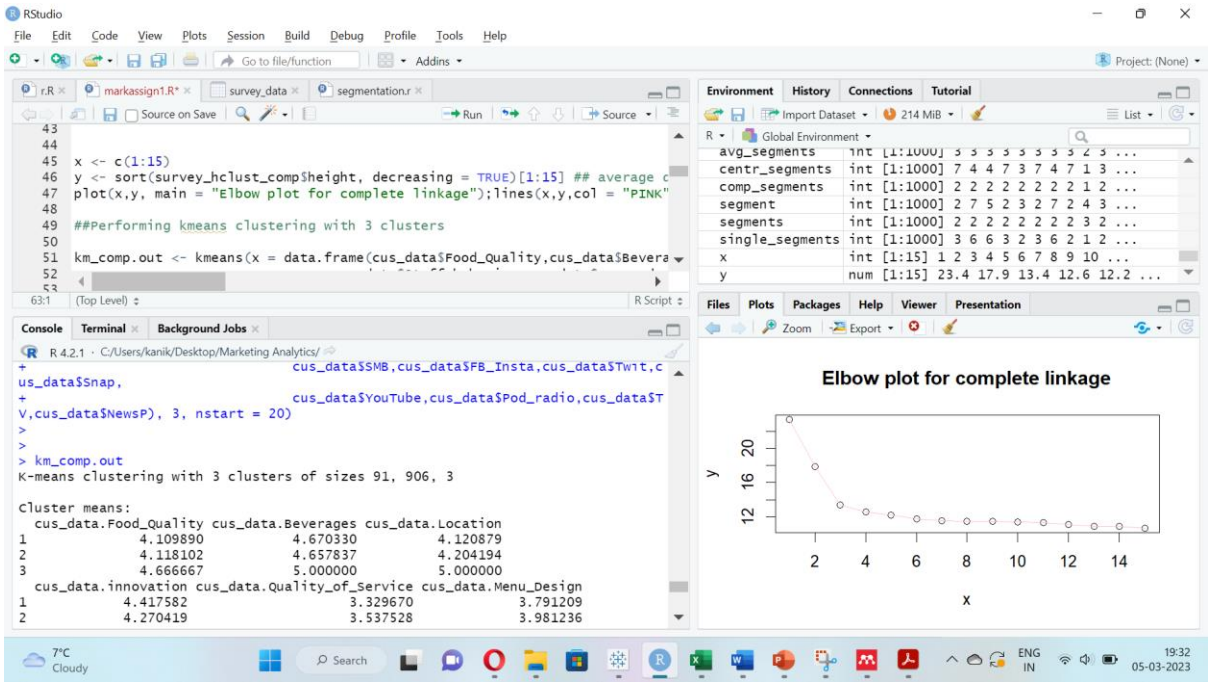

##Performing kmeans clustering with 3 clusters

```
km_comp.out <- kmeans(x =
data.frame(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,

cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Health,cus_data$Finc,

                    cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail
,
```

cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,

cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP),
3, nstart = 20)

km_comp.out




comp_segments = km_comp.out$cluster

segment_comp_result <- cbind(cus_data, segments) #add new column to original dataset

##shows within cluster variation for every cluster

km_comp.out$withinss

##shows the total within cluster variations for different clusters

km_comp.out$tot.withinss

```
Within cluster sum of squares by cluster:
[1]  78368.38 145687.84  31995.74
 (between_SS / total_SS =  70.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
>
> comp_segments = km_comp.out$cluster
> segment_comp_result <- cbind(cus_data, segments) #add new column to original d
ataset
>
> ##shows within cluster variation for every cluster
>
> km_comp.out$withinss
[1]  78368.38 145687.84  31995.74
>
> ##shows the total within cluster variations for different clusters
>
> km_comp.out$tot.withinss
[1] 256052
>
```

##########################################################################

### Using single linkage method

survey_hclust_single <-
hclust(dist(scale(cbind(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,

cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Health,cus_data$Finc,

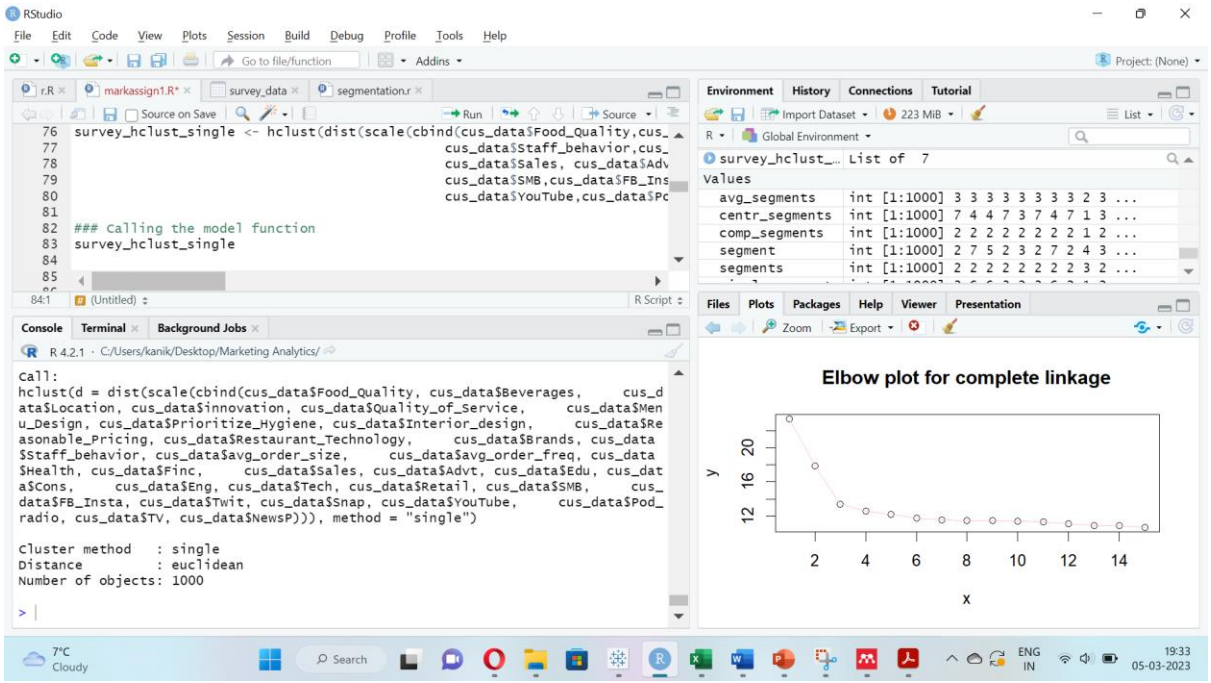
cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail,

cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,

cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP))), method="single")

### Calling the model function

survey_hclust_single



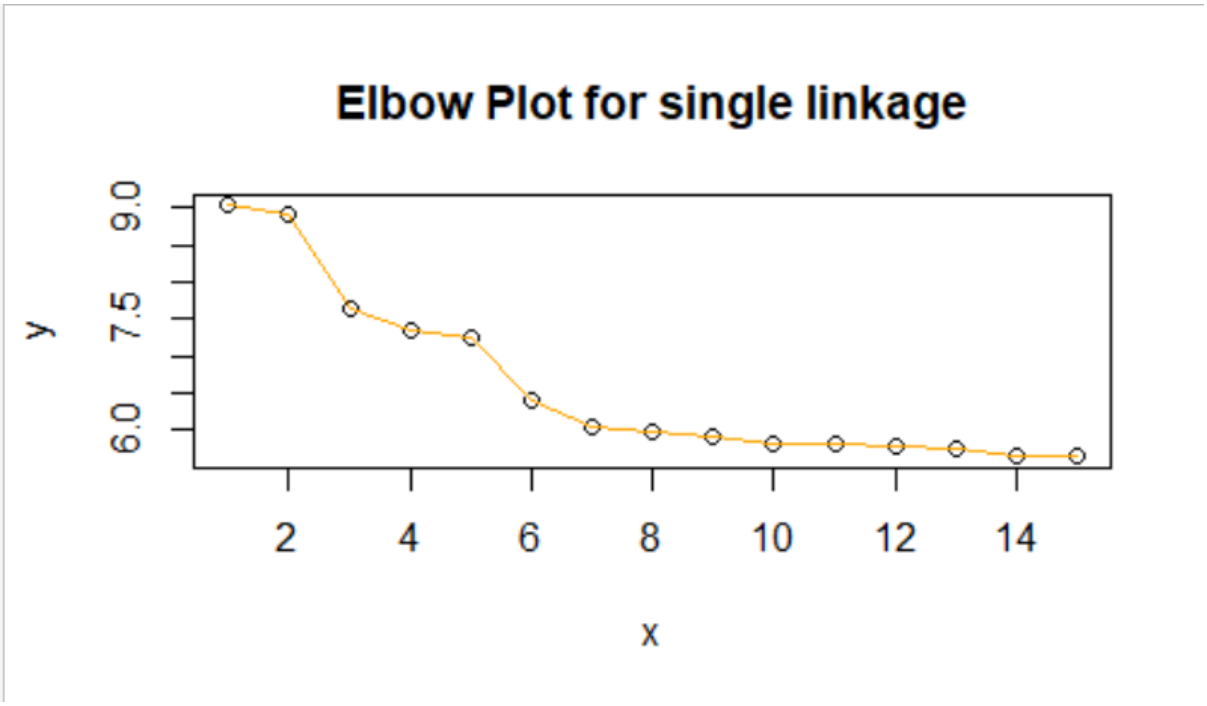

### creating elbow plots that determines the no. of clusters;

##Usually the no. present at the kink of the chart is used to determine no. of clusters

x <- c(1:15)

y <- sort(survey_hclust_single$height, decreasing = TRUE)[1:15] ## average distance between clusters

plot(x,y, main = "Elbow Plot for single linkage");lines(x,y,col = "ORANGE")




km_single.out <- kmeans(x =
data.frame(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,

cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Health,cus_data$Finc,

cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail,

cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,

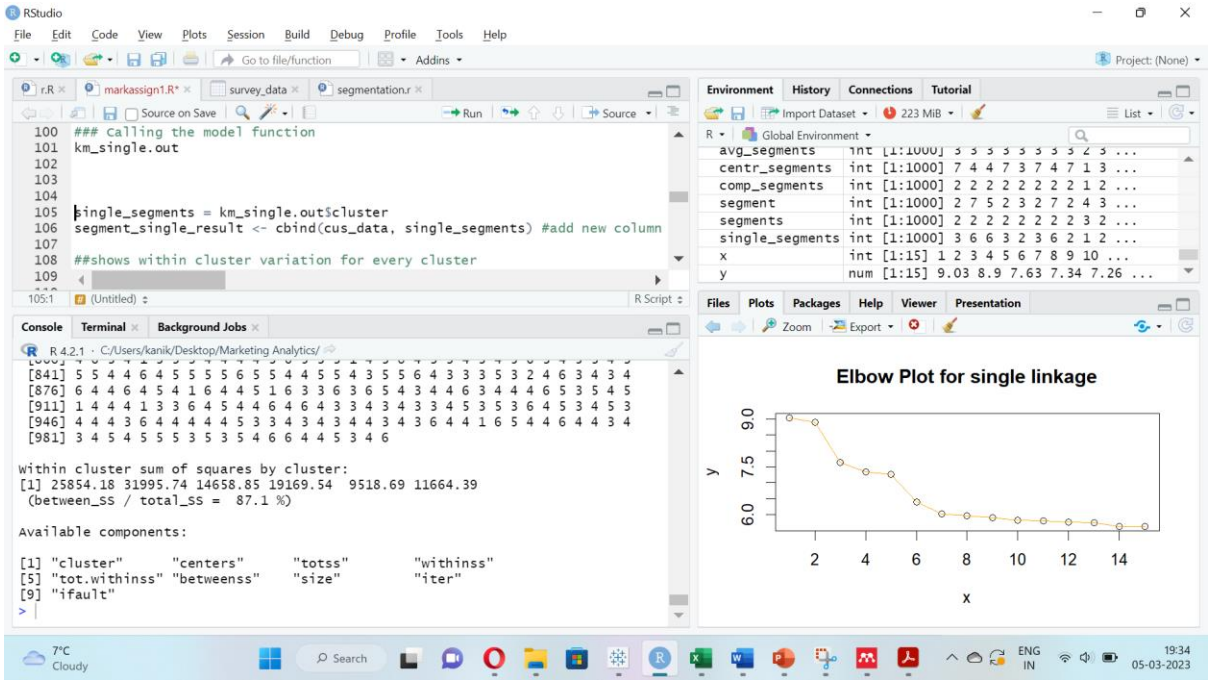
cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP), 6, nstart = 20)

### Calling the model function

km_single.out

single_segments = km_single.out$cluster

segment_single_result <- cbind(cus_data, single_segments) #add new column to original dataset




##shows within cluster variation for every cluster

km_single.out$withinss

##shows the total within cluster variations for different clusters

km_single.out$tot.withinss

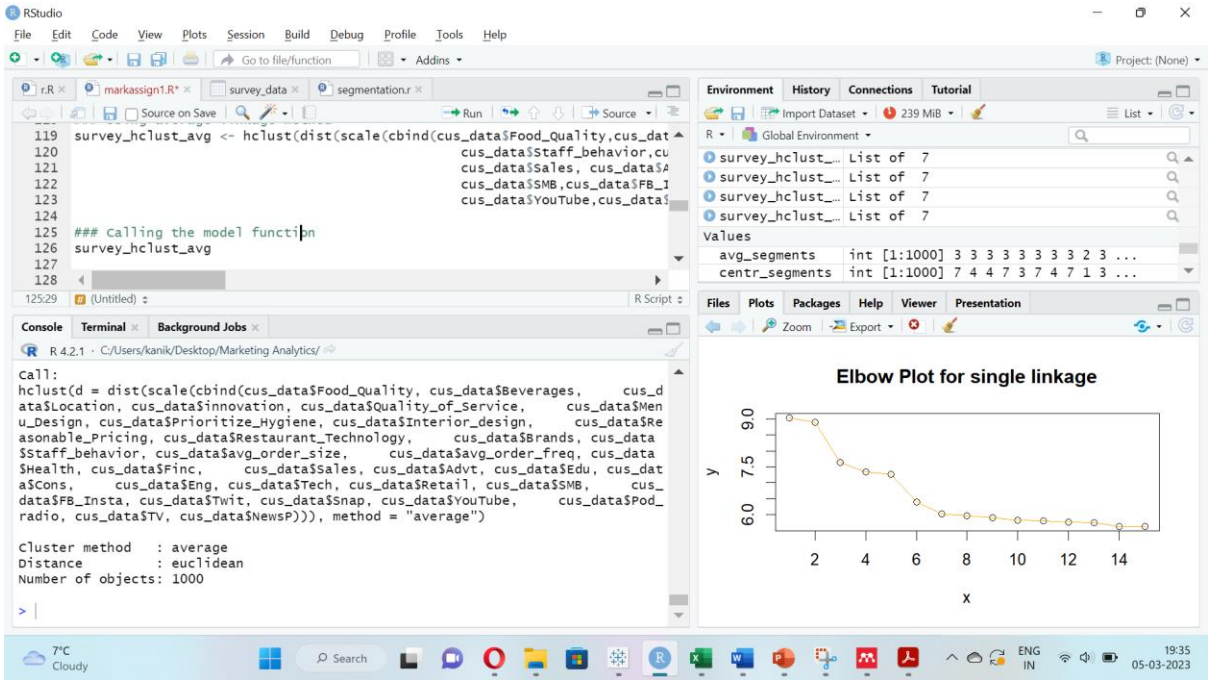
```
Within cluster sum of squares by cluster:
[1] 11664.39  9518.69 19169.54 31995.74 25854.18 14658.85
 (between_SS / total_SS =  87.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
>
>
> single_segments = km_single.out$cluster
> segment_single_result <- cbind(cus_data, single_segments) #add new column to o
riginal dataset
>
> ##shows within cluster variation for every cluster
>
> km_single.out$withinss
[1] 11664.39  9518.69 19169.54 31995.74 25854.18 14658.85
> ##shows the total within cluster variations for different clusters
>
> km_single.out$tot.withinss
[1] 112861.4
```

### Using average linkage method

```
survey_hclust_avg <-
hclust(dist(scale(cbind(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_
data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,


cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Hea
lth,cus_data$Finc,

                        cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail
,

                        cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,


cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP))),
method="average")
```

### Calling the model function

```
survey_hclust_avg
```



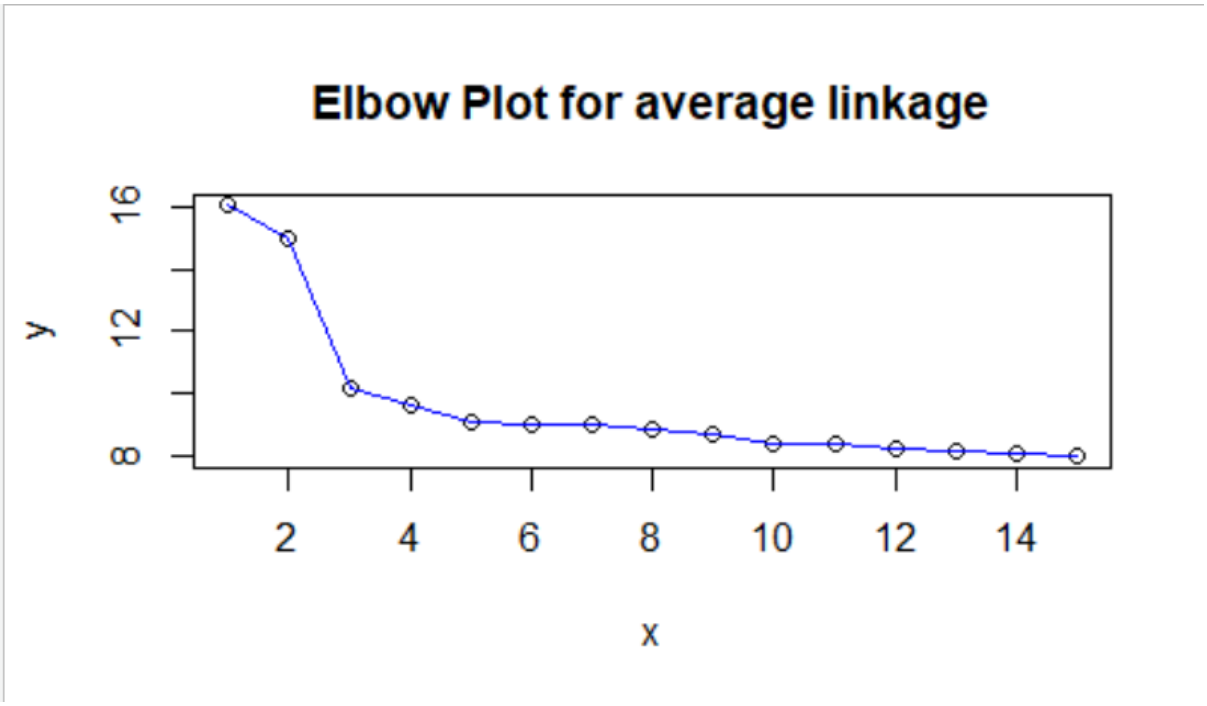

### creating elbow plots that determines the no. of clusters;

##Usually the no. present at the kink of the chart is used to determine no. of clusters

x <- c(1:15)

y <- sort(survey_hclust_avg$height, decreasing = TRUE)[1:15] ## average distance between clusters

plot(x,y,main = "Elbow Plot for average linkage");lines(x,y,col = "BLUE")




km_avg.out <- kmeans(x =
data.frame(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,

cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Health,cus_data$Finc,

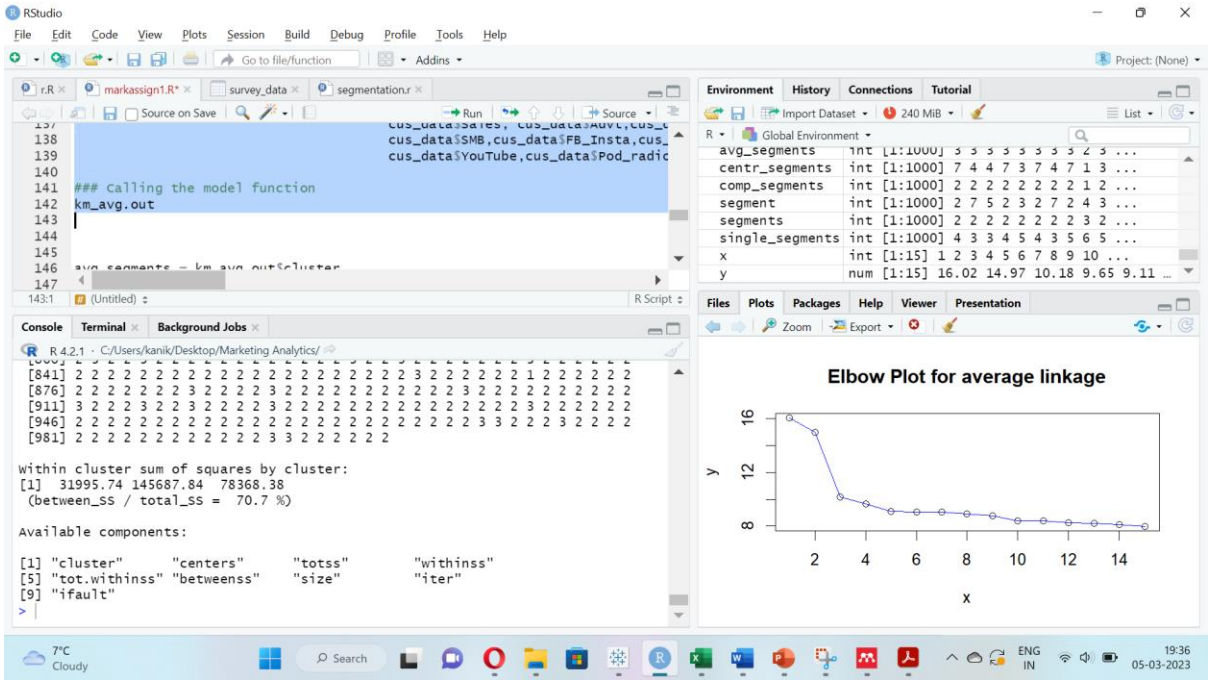
cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail,

cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,

cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP), 3, nstart = 20)

### Calling the model function

km_avg.out




avg_segments = km_avg.out$cluster

segment_avg_result <- cbind(cus_data, avg_segments) #add new column to original dataset

##shows within cluster variation for every cluster

km_avg.out$withinss

##shows the total within cluster variations for different clusters

km_avg.out$tot.withinss

```
Within cluster sum of squares by cluster:
[1]  31995.74  78368.38 145687.84
 (between_SS / total_SS =  70.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
>
>
> avg_segments = km_avg.out$cluster
> segment_avg_result <- cbind(cus_data, avg_segments) #add new column to origina
l dataset
>
> ##shows within cluster variation for every cluster
>
> km_avg.out$withinss
[1]  31995.74  78368.38 145687.84
>
> ##shows the total within cluster variations for different clusters
>
> km_avg.out$tot.withinss
[1] 256052
>
```

### Using centroid linkage method

survey_hclust_centr <-
hclust(dist(scale(cbind(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,

cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Health,cus_data$Finc,

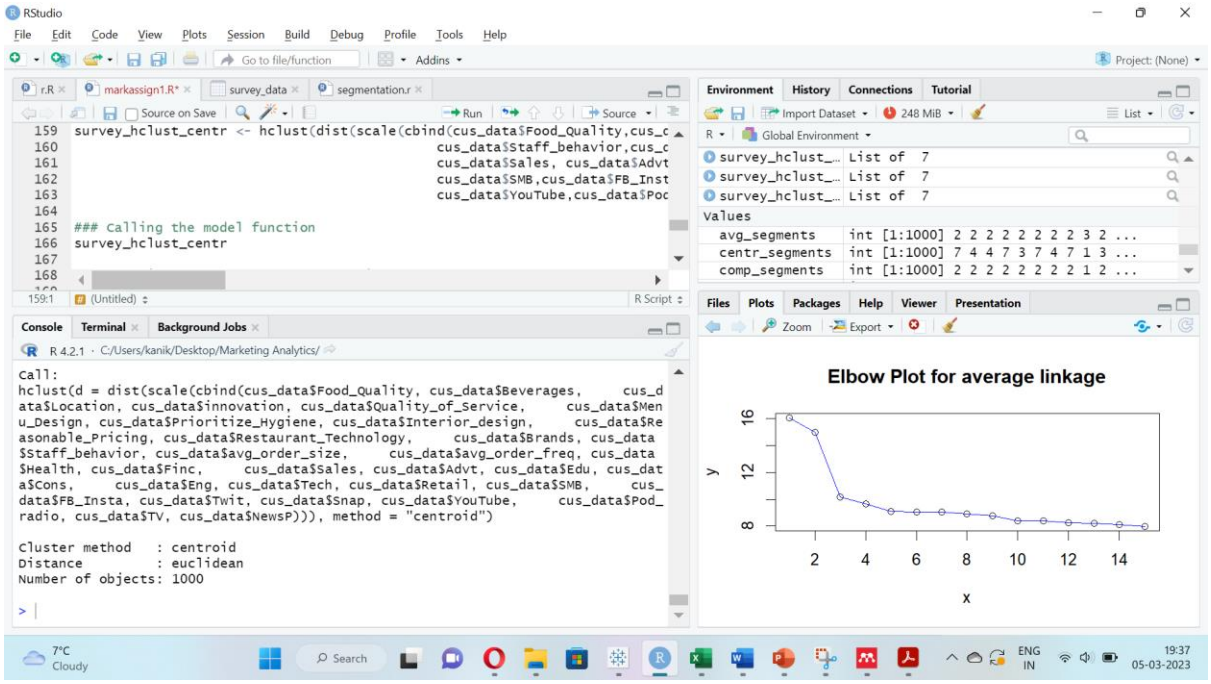
cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail,

cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,

cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP))),
method="centroid")

### Calling the model function

survey_hclust_centr



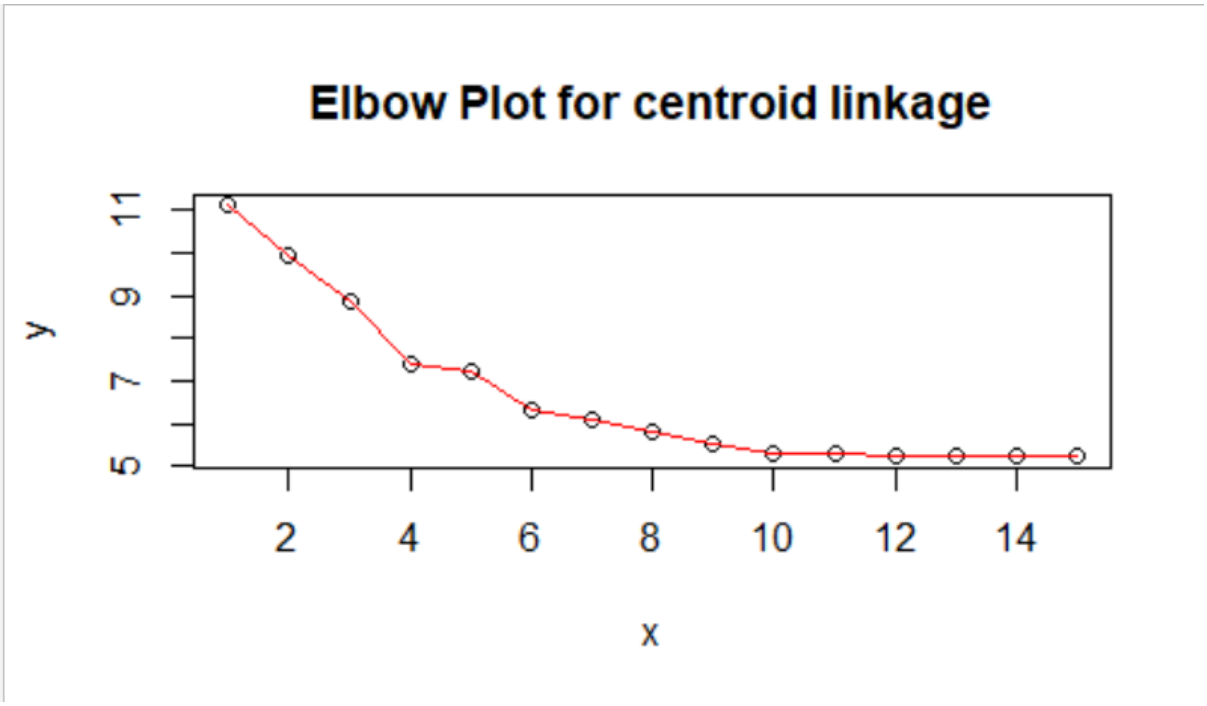

### creating elbow plots that determines the no. of clusters;

##Usually the no. present at the kink of the chart is used to determine no. of clusters

x <- c(1:15)

y <- sort(survey_hclust_centr$height, decreasing = TRUE)[1:15] ## average distance between clusters

plot(x,y,main = "Elbow Plot for centroid linkage");lines(x,y,col = "RED")




##Performing kmeans clustering

km_centr.out <- kmeans(x =
data.frame(cus_data$Food_Quality,cus_data$Beverages,cus_data$Location,cus_data$innovation, cus_data$Quality_of_Service,
cus_data$Menu_Design,cus_data$Prioritize_Hygiene,cus_data$Interior_design,
cus_data$Reasonable_Pricing,cus_data$Restaurant_Technology,cus_data$Brands,

cus_data$Staff_behavior,cus_data$avg_order_size,cus_data$avg_order_freq,cus_data$Health,cus_data$Finc,

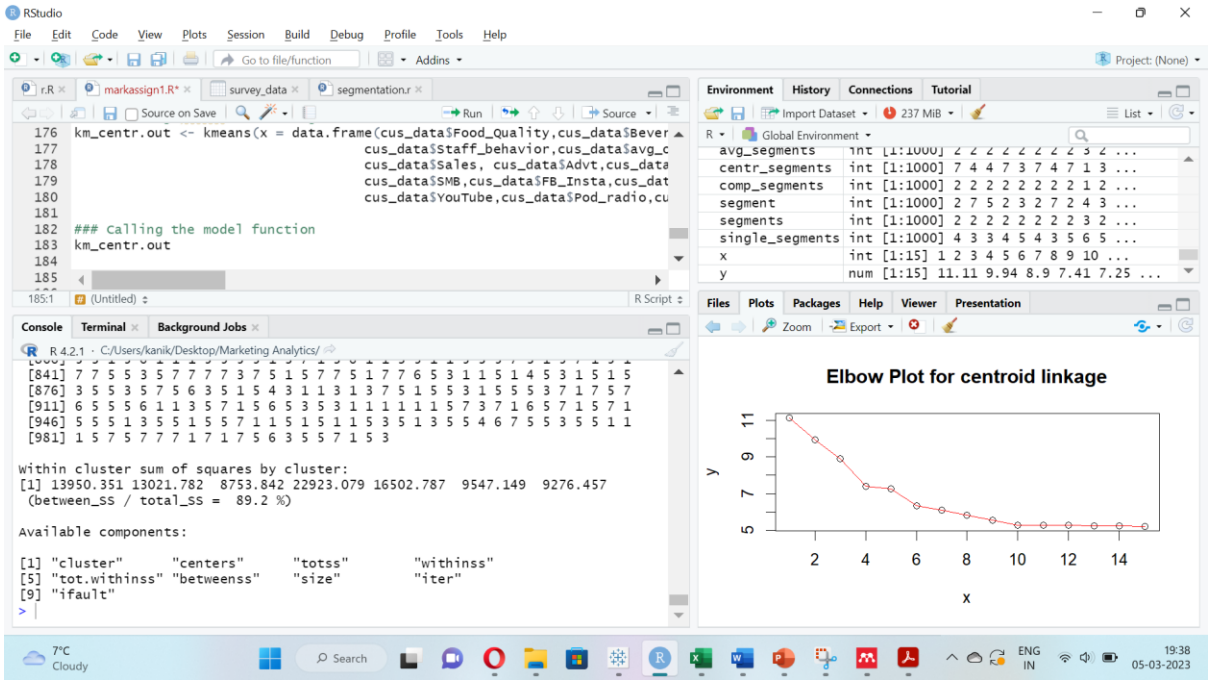
cus_data$Sales,
cus_data$Advt,cus_data$Edu,cus_data$Cons,cus_data$Eng,cus_data$Tech,cus_data$Retail,

cus_data$SMB,cus_data$FB_Insta,cus_data$Twit,cus_data$Snap,

cus_data$YouTube,cus_data$Pod_radio,cus_data$TV,cus_data$NewsP), 7, nstart = 20)

### Calling the model function

km_centr.out




centr_segments = km_centr.out$cluster

segment_centr_result <- cbind(cus_data, centr_segments)

##shows within cluster variation for every cluster

km_centr.out$withinss

##shows the total within cluster variations for different clusters

km_centr.out$tot.withinss

```
Within cluster sum of squares by cluster:
[1]  9547.149  8753.842  9276.457 13950.351 13021.782 22923.079 16502.787
 (between_SS / total_SS =  89.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
>
>
> centr_segments = km_centr.out$cluster
> segment_centr_result <- cbind(cus_data, centr_segments)
>
> ##shows within cluster variation for every cluster
>
> km_centr.out$withinss
[1]  9547.149  8753.842  9276.457 13950.351 13021.782 22923.079 16502.787
>
> ##shows the total within cluster variations for different clusters
>
> km_centr.out$tot.withinss
[1] 93975.45
>
```

##export excel file

write.csv(segment_centr_result, file = "C:/Users/kanik/Desktop/Marketing Analytics/centr_result.csv", row.names = FALSE)