Customer Personality Analysis

PROJECT REPORT SUBMITTED

IN FULFILMENT OF THE REQUIREMENTS FOR COURSE

STAT 467 – MULTIVARIATE ANALYSIS

DEPARTMENT OF STATISTICS OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

MELİKE ERDOĞAN- 2361269

YUSUF KAĞAN ÖZKAN- 2429215

January 2024

ABSTRACT

This report discusses Customer Personality Analysis, which is performed to understand a company's ideal customers and tailor its products to specific customer needs. Multivariate analyzes performed on customer data aim to reveal the complexities and relationships in customer behavior. The report begins by examining the effects of demographic characteristics on customer behavior with Exploratory Data Analysis. Principal Components Analysis and Factor Analysis provide a strategic framework for product modifications by identifying key factors affecting customer spending. In addition, customer behavior is understood more deeply through methods such as clustering, classification analysis and canonical correlation analysis and important suggestions for marketing strategies are offered. This report provides important findings that can help businesses develop customer-focused strategies and use their resources effectively.

1. Introduction

In today's competitive business world, companies are constantly striving to develop and maintain customer-focused strategies. In this context, in-depth analysis of customer data is critical for a business to optimize marketing strategies and increase customer satisfaction. This report contains comprehensive analysis performed on a customer data set. The focus of the analysis is the customer data set. This data set includes various factors such as customer demographics, shopping habits, and campaign reactions. The data set was collected between [date] and provides a rich and detailed view of the company's customers. This analysis was carried out for various purposes. The analyzes performed on the data set aim to segment customers according to certain characteristics. In this way, the company can understand its customers more effectively and develop target-oriented marketing strategies. Taking a detailed look at customer purchasing habits, campaign responses, and other important behaviors will help the company better align its products and services to customer expectations. The resulting analysis results will be used to create customized marketing strategies and increase customer loyalty. Various statistical and analytical methods were used during the analysis process. With Exploratory Data Analysis (EDA), the general characteristics of the data set were understood, and differences between customer groups were determined with Inferential Statistics. Principal Components Analysis (PCA) and Factor Analysis have been applied to explain and understand the relationships between variables in the data set. Customer segmentation was carried out with Discrimination and Classification methods, and it was aimed to identify similar customer groups with Cluster Analysis. Canonical Correlation Analysis was used to evaluate the relationships between customer behavior and the company's product and service features. This comprehensive analysis provides important information that will help company better understand its customer base, make strategic decisions, and gain a competitive advantage.

1.1 Data Description

With its 29 variables, the dataset enables Customer Personality Analysis, which helps businesses customize their products for their target audience. Product categories, retail locations, and descriptive customer statistics are all covered by the dataset, which offers numerical and categorical information to help with decision-making regarding targeted marketing and product customisation.

1.2. Research Questions

Is there any significant difference between Complaining of the customers?

Which Independent variables should be used while explaining the Income variable in multiple linear regression?

What types of factors explain which customer behaviors or/and attributes?

How to classify whether the customer accepted the offer in the last campaign based on the customer's information?

How many of the customers' acceptance of offers are predicted correctly and how many are predicted incorrectly?

How many groups can customers be divided into according to their various characteristics and which customers belong to which group?

1.3. The Aim of the Study

The main purpose of this study is to better understand the company's customer base, optimize marketing strategies and increase overall customer satisfaction through comprehensive analyzes performed on the customer data set. This research aims to provide a detailed insight into customer behavior, habits, and reactions to campaigns. The analyzes aim to provide information for the company to improve customer segmentation, create customized marketing strategies, make its products and services more suitable to customer expectations and gain competitive advantage. The aim of this study is to ensure that the findings support the company's strategic decisions and contribute to sustainable growth.

2. Methodology/Analysis

The methodology of this study includes a variety of analytical and statistical methods to support the comprehensive analysis performed on [company name]'s customer data set and to better understand the company's customer base.

In the first stage, Exploratory Data Analysis (EDA) was performed to understand the general structure of the data set. This step involves evaluating distributions, correlations, and missing data in the data set, as well as basic statistics. EDA is of critical importance in visualizing and understanding the general characteristics of the data set.

In the second stage, differential statistical methods were used to determine differences between customer groups. T tests, MANOVA, and similar statistical tests were applied to evaluate statistically significant differences between customer segments.

PCA and Factor Analysis were used to understand the relationships between variables in the data set. Identification of key components and factors was done to reduce the complexity of the data set and reveal key features.

Discrimination and Classification techniques were applied for customer segmentation. This step involves grouping and classifying according to specific customer characteristics.

Cluster Analysis was applied based on customer similarities. This step aims to identify homogeneous customer segments by dividing customers with similar characteristics into natural groups.

Canonical Correlation Analysis was used to evaluate the relationships between customer behavior and the company's product and service features. This method has been used to understand the interactions and connections between two different sets of variables. This comprehensive methodology aims to help the company understand its customer base and make strategic decisions by providing multifaceted analysis of the data set. Each analytical step provides valuable information for the company to strengthen customer relationships and increase competitive advantage.

3. Results and Findings

3.1 EXPLORATORY DATA ANALYSIS

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome
1	5524	1957	Graduation	Single	58138	0	0
2	2174	1954	Graduation	Single	46344	1	1
3	4141	1965	Graduation	Together	71613	0	0
4	6182	1984	Graduation	Together	26646	1	0
5	5324	1981	PhD	Married	58293	1	0
6	7446	1967	Master	Together	62513	0	1

	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts
1	04-09-2012	58	635	88	546
2	08-03-2014	38	11	1	6
3	21-08-2013	26	426	49	127
4	10-02-2014	26	11	4	20
5	19-01-2014	94	173	43	118
6	09-09-2013	16	520	42	98

	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases
1	172	88	88	3
2	2	1	6	2
3	111	21	42	1
4	10	3	5	2
5	46	27	15	5
6	0	42	14	2

	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
1	8	10	4
2	1	1	2
3	8	2	10
4	2	0	4
5	5	3	6
6	6	4	10

	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1
1	7	0	0	0	0
2	5	0	0	0	0
3	4	0	0	0	0
4	6	0	0	0	0
5	5	0	0	0	0
6	6	0	0	0	0

	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
1	0	0	3	11	1
2	0	0	3	11	0
3	0	0	3	11	0
4	0	0	3	11	0
5	0	0	3	11	0
6	0	0	3	11	0

Table 1: First 6 observations of the data

```
'data.frame':   2240 obs. of  29 variables:
 $ ID                 : int  5524 2174 4141 6182 5324 7446 965 6177 4855 5899 ...
 $ Year_Birth         : int  1957 1954 1965 1984 1981 1967 1971 1985 1974 1950 ...
 $ Education          : chr  "Graduation" "Graduation" "Graduation" "Graduation" ...
 $ Marital_Status     : chr  "Single" "Single" "Together" "Together" ...
 $ Income             : int  58138 46344 71613 26646 58293 62513 55635 33454 30351 5648 ...
 $ Kidhome            : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 1 1 2 2 2 ...
 $ Teenhome           : int  0 1 0 0 0 1 1 0 0 1 ...
 $ Dt_Customer        : chr  "04-09-2012" "08-03-2014" "21-08-2013" "10-02-2014" ...
 $ Recency            : int  58 38 26 26 94 16 34 32 19 68 ...
 $ MntWines           : int  635 11 426 11 173 520 235 76 14 28 ...
 $ MntFruits          : int  88 1 49 4 43 42 65 10 0 0 ...
 $ MntMeatProducts    : int  546 6 127 20 118 98 164 56 24 6 ...
 $ MntFishProducts    : int  172 2 111 10 46 0 50 3 3 1 ...
 $ MntSweetProducts   : int  88 1 21 3 27 42 49 1 3 1 ...
 $ MntGoldProds       : int  88 6 42 5 15 14 27 23 2 13 ...
 $ NumDealsPurchases  : int  3 2 1 2 5 2 4 2 1 1 ...
 $ NumWebPurchases    : int  8 1 8 2 5 6 7 4 3 1 ...
 $ NumCatalogPurchases: int  10 1 2 0 3 4 3 0 0 0 ...
 $ NumStorePurchases  : int  4 2 10 4 6 10 7 4 2 0 ...
 $ NumWebVisitsMonth  : int  7 5 4 6 5 6 6 8 9 20 ...
 $ AcceptedCmp3       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ AcceptedCmp4       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp5       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp1       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AcceptedCmp2       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Complain           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Z_CostContact      : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Z_Revenue          : int  11 11 11 11 11 11 11 11 11 11 ...
 $ Response           : int  1 0 0 0 0 0 0 0 1 0 ...
```

Table 2: Structure of the data

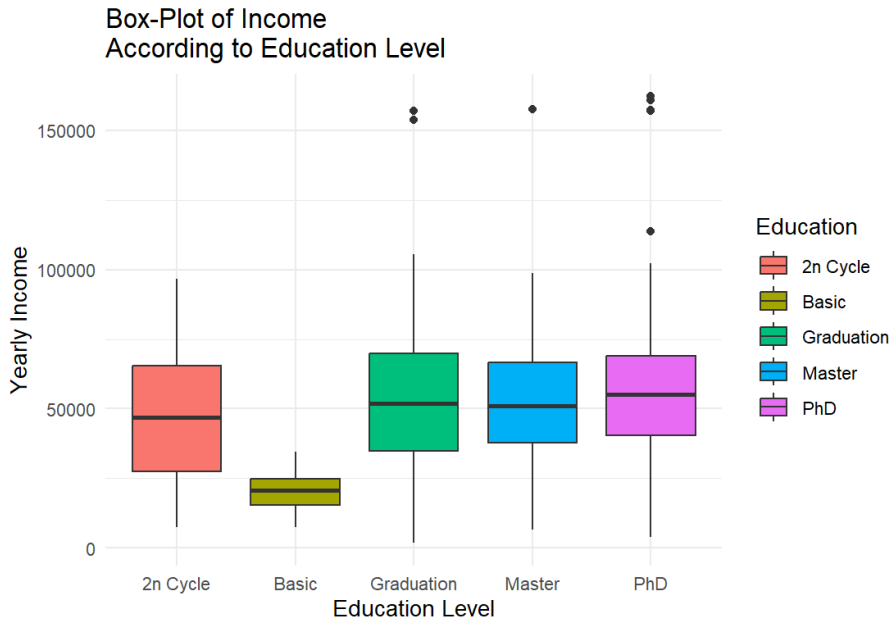
As we can see from the table 1 and table 2, we can observe that we have 29 different variables with 2240 observations. Despite it shows some of the variables are in the integer form, they are actually at factor class, so for further analysis we will convert their classes to factor to make the analysis more precise and correct.

```
       ID          Year_Birth    Education         Marital_Status
 Min.   :    0   Min.   :1893   Length:2240         Length:2240
 1st Qu.: 2828   1st Qu.:1959   Class :character    Class :character
 Median : 5458   Median :1970   Mode  :character    Mode  :character
 Mean   : 5592   Mean   :1969
 3rd Qu.: 8428   3rd Qu.:1977
 Max.   :11191   Max.   :1996

     Income        Kidhome     Teenhome      Dt_Customer
 Min.   :  1730   0:1293   Min.   :0.0000   Length:2240
 1st Qu.: 35303   1: 899   1st Qu.:0.0000   Class :character
 Median : 51382   2:  48   Median :0.0000   Mode  :character
 Mean   : 52247            Mean   :0.5062
 3rd Qu.: 68522            3rd Qu.:1.0000
 Max.   :666666            Max.   :2.0000
 NA's   :24
    Recency         MntWines         MntFruits     MntMeatProducts
 Min.   : 0.00   Min.   :   0.00   Min.   :  0.0   Min.   :   0.0
 1st Qu.:24.00   1st Qu.:  23.75   1st Qu.:  1.0   1st Qu.:  16.0
 Median :49.00   Median : 173.50   Median :  8.0   Median :  67.0
 Mean   :49.11   Mean   : 303.94   Mean   : 26.3   Mean   : 166.9
 3rd Qu.:74.00   3rd Qu.: 504.25   3rd Qu.: 33.0   3rd Qu.: 232.0
 Max.   :99.00   Max.   :1493.00   Max.   :199.0   Max.   :1725.0
```

```
 MntFishProducts  MntSweetProducts  MntGoldProds    NumDealsPurchases
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
 1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000
 Median : 12.00   Median :  8.00   Median : 24.00   Median : 2.000
 Mean   : 37.53   Mean   : 27.06   Mean   : 44.02   Mean   : 2.325
 3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000
 Max.   :259.00   Max.   :263.00   Max.   :362.00   Max.   :15.000

 NumWebPurchases  NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
 Min.   : 0.000   Min.   : 0.000      Min.   : 0.00     Min.   : 0.000
 1st Qu.: 2.000   1st Qu.: 0.000      1st Qu.: 3.00     1st Qu.: 3.000
 Median : 4.000   Median : 2.000      Median : 5.00     Median : 6.000
 Mean   : 4.085   Mean   : 2.662      Mean   : 5.79     Mean   : 5.317
 3rd Qu.: 6.000   3rd Qu.: 4.000      3rd Qu.: 8.00     3rd Qu.: 7.000
 Max.   :27.000   Max.   :28.000      Max.   :13.00     Max.   :20.000

  AcceptedCmp3      AcceptedCmp4      AcceptedCmp5      AcceptedCmp1
 Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
 Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
 Mean   :0.07277   Mean   :0.07455   Mean   :0.07277   Mean   :0.06429
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000

  AcceptedCmp2        Complain        Z_CostContact   Z_Revenue
 Min.   :0.00000   Min.   :0.000000   Min.   :3     Min.   :11
 1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:3     1st Qu.:11
 Median :0.00000   Median :0.000000   Median :3     Median :11
 Mean   :0.01339   Mean   :0.009375   Mean   :3     Mean   :11
 3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:3     3rd Qu.:11
 Max.   :1.00000   Max.   :1.000000   Max.   :3     Max.   :11

    Response
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1491
 3rd Qu.:0.0000
 Max.   :1.0000
```

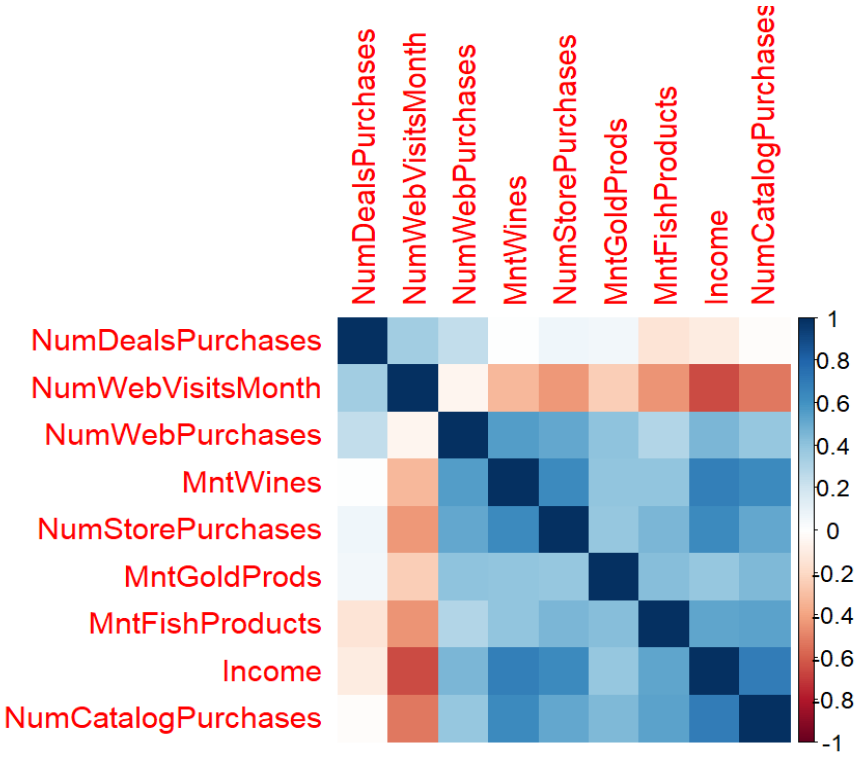
Table 3: Summary statistics of the data

From above table we can see the summary statistics of the variables in our dataset. We can see that our data contains 24 missing values for Income variable, so to make the analysis more robust we remove them from the data.



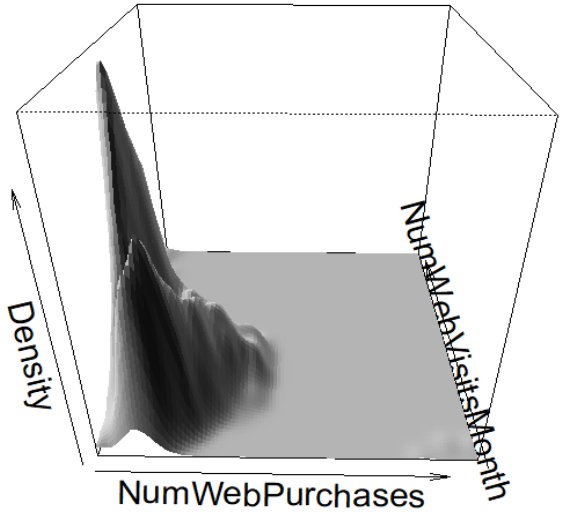
Plot 1: Boxplot of Income Among Education Levels

From the above Boxplot we can see the Income of the customers by their education levels, and we can easily observe that the highest education level which is PhD in this case has the maximum median value among other levels.

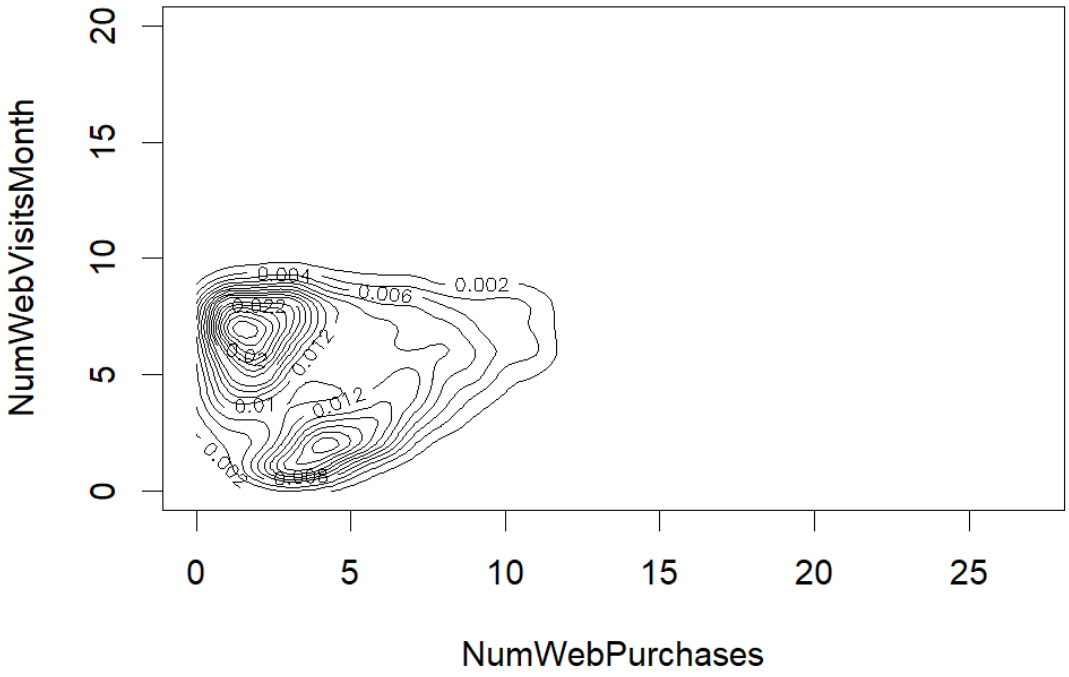


Plot 2: Correlation Plot of Some Variables

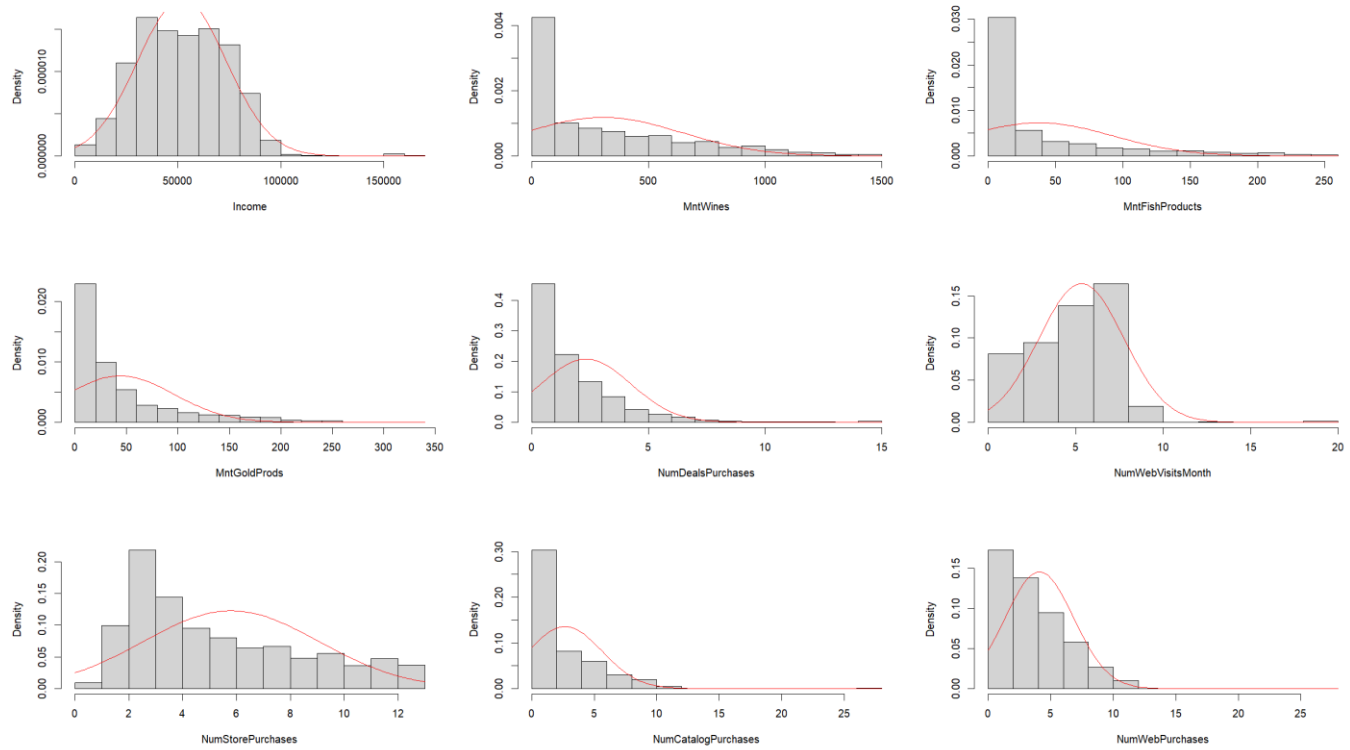
From above plot we can see the correlation values of some variables in the data.



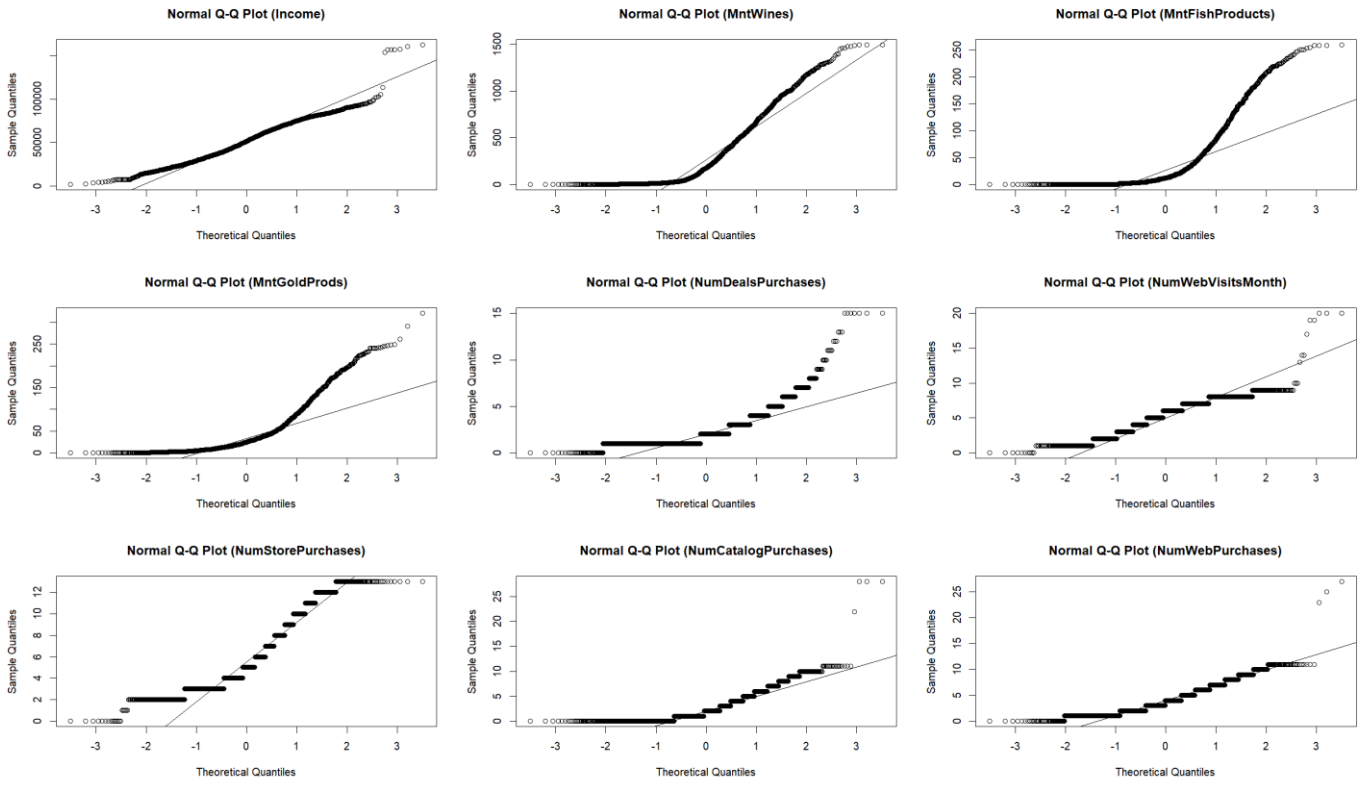
Plot 3: Perspective plot of Number of web purchase and Number of website visits per month



Plot 4: Contour plot of Number of web purchase and Number of website visits per month



Plot 5: Histograms of different variables



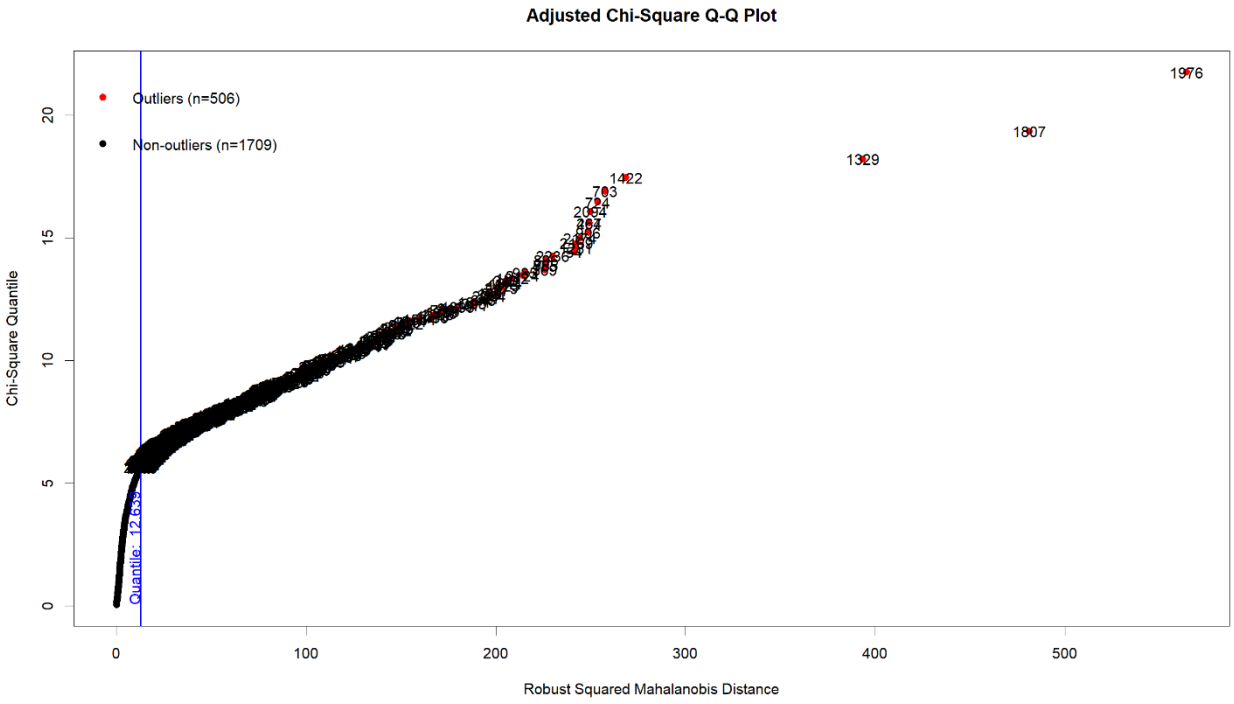
Plot 6: Normal QQ-plots of different variables

3.2 INFERENCES ABOUT A MEAN VECTOR

For this part we select our data with 4 variables as follows:

	Income	MntGoldProds	NumWebVisitsMonth	NumWebPurchases
1	58138	88	7	8
2	46344	6	5	1
3	71613	42	4	8
4	26646	5	6	2
5	58293	15	5	5
6	62513	14	6	6

Table 4: First 6 observations of new dataset



Plot 7: Adjusted Chi-Square QQ- plot

As seen in the plot we can observe that we have 506 outliers, since these outliers creates 25% of the data, we do not remove them from the dataset.

```
$multivariateNormality
           Test       HZ p value MVN
1 Henze-Zirkler 32.54736       0  NO

$univariateNormality
             Test          Variable Statistic   p value Normality
1 Anderson-Darling      Income          8.1234  <0.001      NO
2 Anderson-Darling   MntGoldProds     169.9323  <0.001      NO
3 Anderson-Darling NumWebVisitsMonth   38.1471  <0.001      NO
4 Anderson-Darling  NumWebPurchases    46.6300  <0.001      NO

$Descriptives
                    n         Mean      Std.Dev Median  Min    Max  25th
Income           2215 51969.861400 21526.320095  51373 1730 162397 35284
MntGoldProds     2215    43.979684    51.822660     25    0    321     9
NumWebVisitsMonth 2215     5.318736     2.425863      6    0     20     3
NumWebPurchases  2215     4.085779     2.741473      4    0     27     2
                  75th      Skew  Kurtosis
Income           68487 0.3468794 0.7058831
MntGoldProds        56 1.8360712 3.1381982
NumWebVisitsMonth    7 0.2180804 1.8395862
NumWebPurchases      6 1.1947659 4.0508248
```

Table 5: Multivariate Normality Test for data

As we can see the data does not follow multivariate normality. To obtain this we should do some transformations to go on with our analysis.

Normalizing the Data:

First, we try to apply log transformations to data but after we applied the transformations we could not reach the multivariate normality.

```
$multivariateNormality
           Test       HZ p value MVN
1 Henze-Zirkler 17.52559       0  NO
```

Table 6 : Log transformed data normality test

Then we applied square root transformation the data, but we again could not catch multivariate normality.

```
$multivariateNormality
           Test       HZ p value MVN
1 Henze-Zirkler 15.12294       0  NO
```

Table 7: Square root transformed data normality test

We also tried inverse transformation to data, but again we could not reach the multivariate normality.

```
$multivariateNormality
           Test       HZ p value MVN
1 Henze-Zirkler 196.4269       0  NO
```

Table 8: Inverse transformed data normality test

Then we applied Best Normalize function in R to our data, from there we applied order Norm function to make it normal. This time we catch univariate normality in one variable but not in the others. However, this was the best transformation that we applied for now, because it has the lowest test statistic value among the others.

```
$multivariateNormality
           Test       HZ p value MVN
1 Henze-Zirkler 8.169321       0  NO

$univariateNormality
             Test         Variable Statistic   p value Normality
1 Anderson-Darling    INCOME.x.t       0.0049     1        YES
2 Anderson-Darling     GOLD.x.t        0.8645  0.0265      NO
3 Anderson-Darling  Web_visit.x.t     21.5700  <0.001      NO
4 Anderson-Darling Web_Purchase.x.t   20.5792  <0.001      NO
```

Table 9: Order Norm transformation applied data normality test

Now, we will continue with this transformation and we will assume that our data is follows multivariate normal distribution for further analysis.

```
  v1    variable statistic        p
  <fct> <chr>        <dbl>    <dbl>
1 0     v2           1.00  1.00e+ 0
2 0     v3           0.996 6.41e- 6
3 0     v4           0.973 7.56e-20
4 0     v5           0.974 1.07e-19
5 1     v2           0.960 5.23e- 1
6 1     v3           0.948 3.05e- 1
7 1     v4           0.953 3.83e- 1
8 1     v5           0.952 3.79e- 1
```

Table 10: Shapiro test grouped by Complain

In the above table we can see that the observations grouped by the Complain variable which is a factor with levels 0 and 1. According to this we can see that except Amount of Gold, web visits, and web purchases with complain 0, all the other variables are significant. Which satisfy bivariate normality.

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  cbind(v2, v3, v4, v5)
Chi-Sq (approx.) = 5.9501, df = 10, p-value = 0.8194
```

Table 11: Box-M Test

Since the p-value from Box's M test is not statistically significant, we do not have enough evidence to reject the null hypothesis. Therefore, we can conclude that the variance-covariance

matrices are equal across all combinations of dependent variables formed by each group in the independent variable.

```
        Hotelling's two sample T2-test

data:  cbind(v2, v3, v4, v5) by v1
T.2 = 0.66844, df1 = 4, df2 = 2210, p-value = 0.6139
alternative hypothesis: true location difference is not equal to c(0,0,0,0)
```

Table 12: Hotelling' s Two sample T2- test

Since we do not reject the null hypothesis (H0), it implies that there is not enough evidence to support the claim that the mean responses significantly vary based on complain.

Confidence Intervals:

Simultaneous Confidence Intervals:

	Upper	Lower
Income	-0.3853699	-0.8758579
Gld	0.9671881	0.4296452
Visit	-0.3701101	-0.4717800
Purchase	0.9708613	0.8348648

Table 13: Simultaneous Confidence Intervals

Bonferroni Confidence Intervals

	Upper	Lower
Income	0.009882066	-0.49435652
Gld	0.571936080	0.04814384
Visit	0.021756012	-0.08994497
Purchase	0.578995229	0.45302985

Table 14: Bonferroni Confidence Intervals

Bonferroni confidence intervals gives narrower intervals, and simultaneous confidence intervals do not contain 0, so the rejection of Hotelling's T^2comes from here.

3.3 COMPARISONS OF SEVERAL MULTIVARIATE MEANS

One Way Manova

For one way anova we will use this variables :

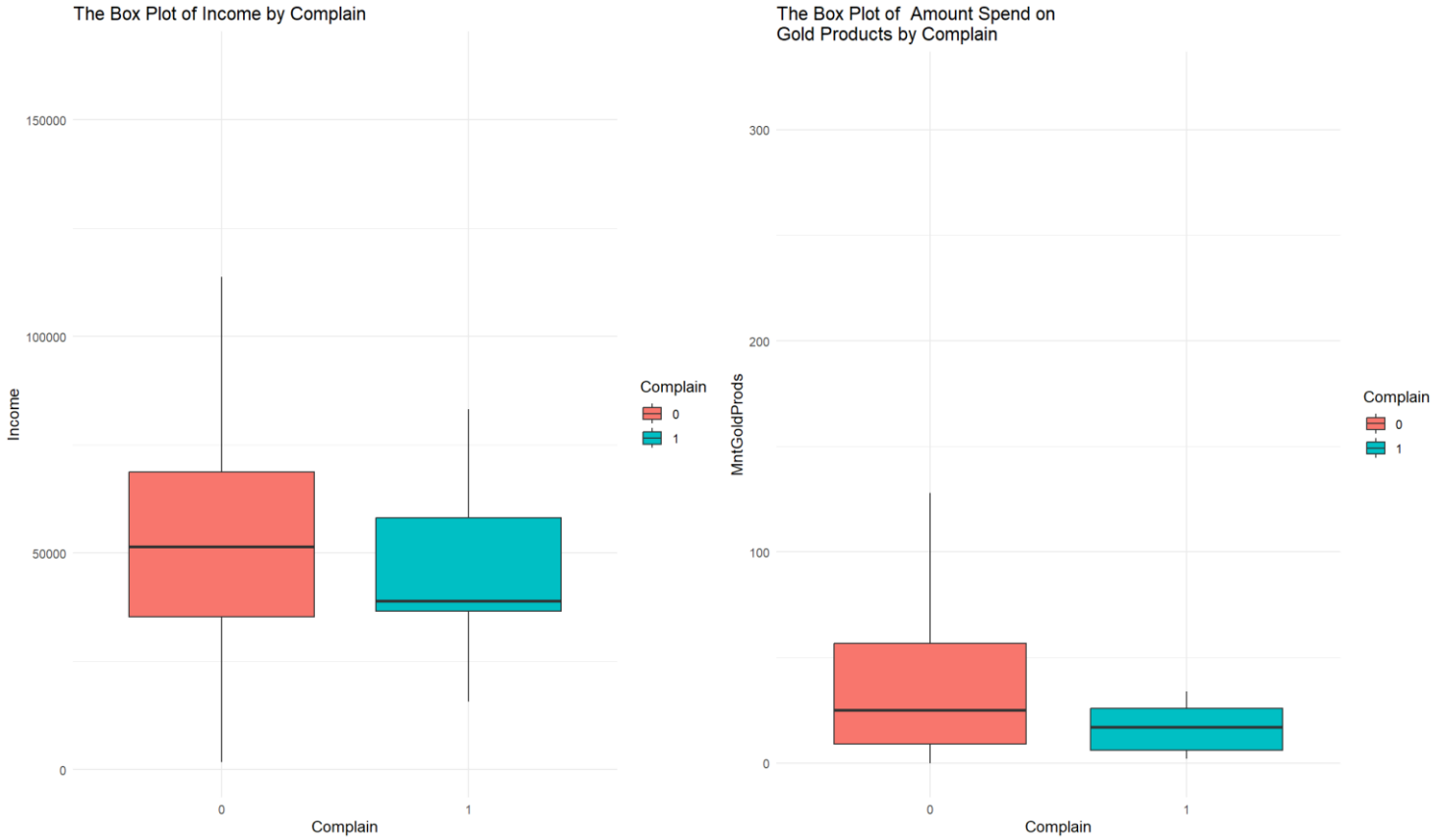
	Complain	Income	Gld	Visit	Purchase
1	0	0.2422611	0.9841202	0.5730787	1.2958306
2	0	-0.1837927	-0.8768154	-0.2132253	-1.2802664
3	0	0.8327847	0.4442181	-0.5140408	1.2958306
4	0	-1.1402646	-0.9795389	0.1356483	-0.6376492
5	0	0.2457586	-0.3211704	-0.2132253	0.4392276
6	0	0.4021350	-0.3698256	0.1356483	0.7210474

Table 15: First 6 observations of different variables

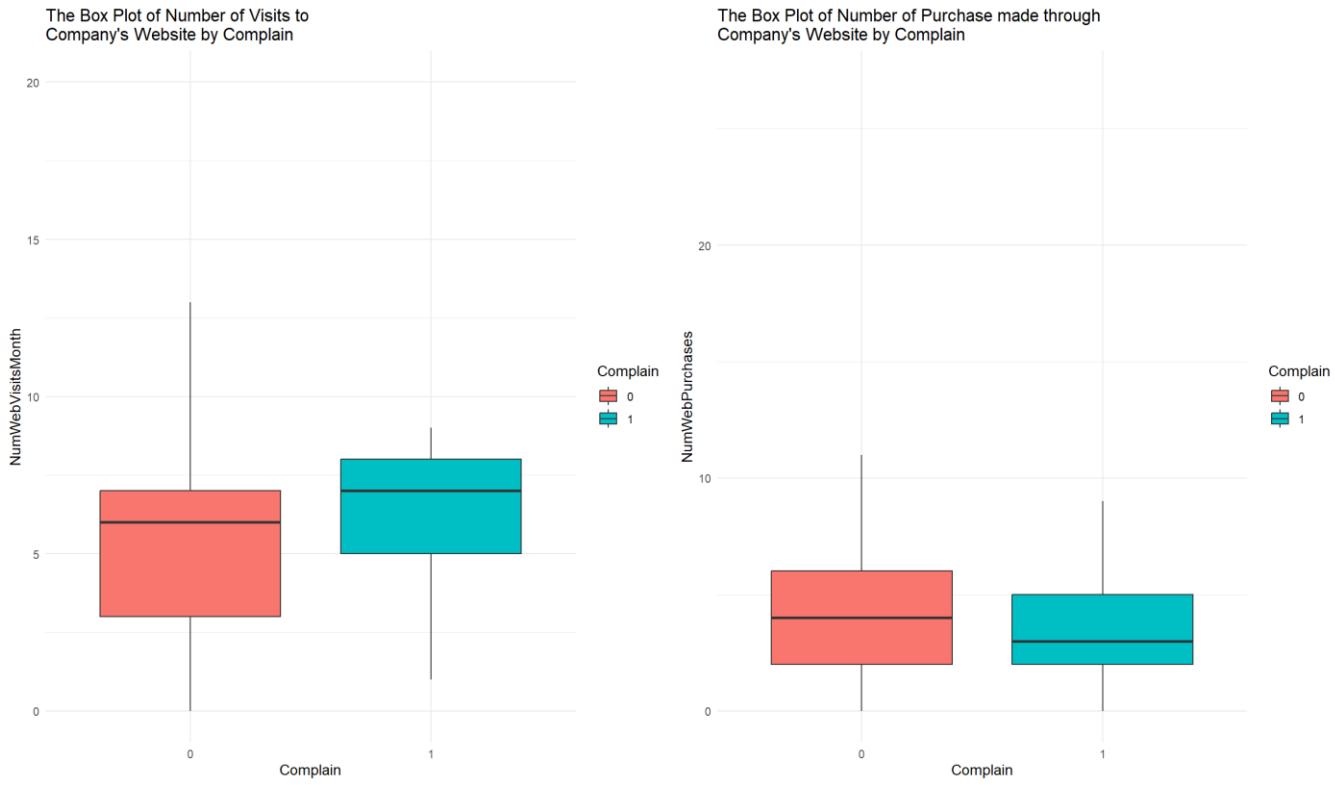
	Complain	n	mean_Income	sd_Income	mean_Gld	mean_Gld	sd_Gld	mean_Visit	sd_Visit	mean_Purchase	sd_Purchase
	<fct>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	2194	0.00279	1.00	0.00595	0.00595	0.993	-0.00378	0.965	0.0111	0.965
2	1	21	-0.288	0.908	-0.294	-0.294	0.803	0.219	0.973	-0.170	1.07

Table 16: Descriptive statistics

We can see the summary statistics or this data before we construct Manova analysis.



Plot 8: Boxplots of Income and Amount of Gold products by Complain



Plot 9: Boxplots of Number of web visits per month and Number of Web purchases of by Complain

	Complain	variable	statistic	p
	<fct>	<chr>	<dbl>	<dbl>
1	0	Gld	0.996	6.41e- 6
2	0	Income	1.00	1.00e+ 0
3	0	Purchase	0.974	1.07e-19
4	0	Visit	0.973	7.56e-20
5	1	Gld	0.948	3.05e- 1
6	1	Income	0.960	5.23e- 1
7	1	Purchase	0.952	3.79e- 1
8	1	Visit	0.953	3.83e- 1

Table 17: Normality Test for data

When we check the p values of variables we can see that except Purchase and Visit variables by there is no complain, other variables satisfies normality by Shapiro Test.

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  cbind(Income, Gld, Visit, Purchase)
Chi-Sq (approx.) = 5.9501, df = 10, p-value = 0.8194
```

Table 18: Box' s M-Test

Since p-value from Box's M test is non-significant, we do not reject the null hypothesis. Consequently, we can conclude that the variance-covariance matrices are equal for each combination of dependent variables formed by the groups in the independent variable.

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
Complain	1	0.0012084	0.66844	4	2210	0.6139
Residuals	2213					

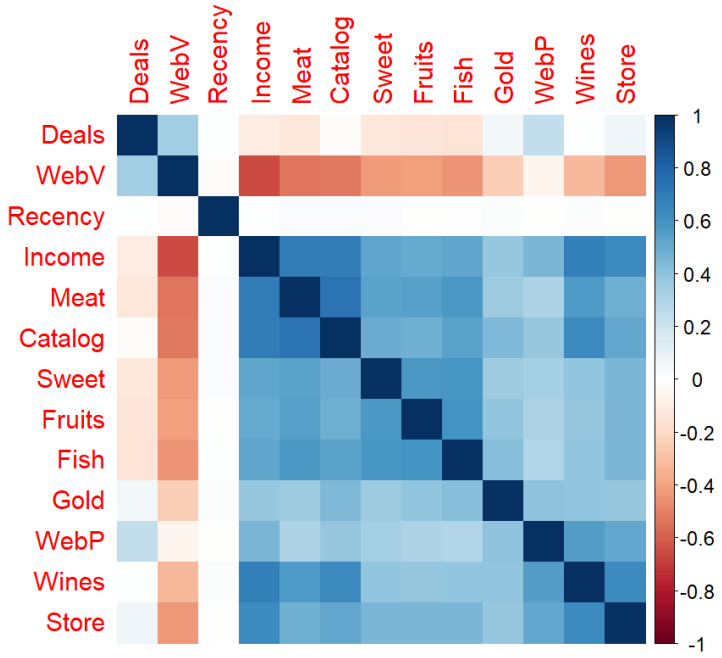
Table 19: Manova Table

From the table we can say that, since p value indicates that we fail to reject null hypothesis, which means that there is no significant difference between the complain levels at %95 confidence level.

3.4 PRINCIPAL COMPONENTS ANALAYSIS AND PRINCIPAL COMPONENTS REGRESSION

	Income	Recency	Wines	Fruits	Meat	Fish	Sweet	Gold	Deals	WebP	Catalog	Store	WebV
5524	58138	58	635	88	546	172	88	88	3	8	10	4	7
2174	46344	38	11	1	6	2	1	6	2	1	1	2	5
4141	71613	26	426	49	127	111	21	42	1	8	2	10	4
6182	26646	26	11	4	20	10	3	5	2	2	0	4	6
5324	58293	94	173	43	118	46	27	15	5	5	3	6	5
7446	62513	16	520	42	98	0	42	14	2	6	4	10	6

Table 20: First six observations of the 13 variables will be used in PCA Analysis Plot



Plot 10: Correlation plot of variables

From above table we can see that scales of the variables looks different, and form the correlation plot of variables some of the variables look highly correlated, so this might cause a problem in PCA. Therefore, we need to scale them.

	Income	Recency	Wines	Fruits	Meat	Fish	Sweet	Gold
5524	0.2865394	0.3100429	0.9775581	1.5486237	1.6893324	2.4533778	1.4840053225	0.84944146
2174	-0.2613480	-0.3808127	-0.8721782	-0.6371942	-0.7180335	-0.6510314	-0.6339376190	-0.73287793
4141	0.9125173	-0.7953261	0.3580150	0.5687743	-0.1786052	1.3394427	-0.1470541841	-0.03820113
6182	-1.1764139	-0.7953261	-0.8721782	-0.5618211	-0.6556203	-0.5049416	-0.5852492755	-0.75217451
5324	0.2937399	1.5535831	-0.3919582	0.4180282	-0.2187279	0.1524627	-0.0009891537	-0.55920873
7446	0.4897790	-1.1407539	0.6366612	0.3929039	-0.3078896	-0.6875539	0.3641734224	-0.57850531

	Deals	WebP	Catalog	Store	WebV
5524	0.3520085	1.4277804	2.5035082	-0.55437464	0.6930581
2174	-0.1677907	-1.1255915	-0.5711244	-1.16957495	-0.1313907
4141	-0.6875898	1.4277804	-0.2294985	1.29122630	-0.5436152
6182	-0.1677907	-0.7608241	-0.9127502	-0.55437464	0.2808337
5324	1.3916068	0.3334781	0.1121273	0.06082567	-0.1313907
7446	-0.1677907	0.6982455	0.4537532	1.29122630	0.2808337

Table 21: Scaled Variables

As we can see we scaled the data.

	Income	Recency	Wines	Fruits	Meat	Fish	Sweet	Gold	Deals	WebP	Catalog	Store	WebV
Income	1.0000	0.0069	0.6883	0.5080	0.6925	0.5204	0.5237	0.3892	-0.1085	0.4588	0.6965	0.6304	-0.6506
Recency	0.0069	1.0000	0.0154	-0.0060	0.0223	0.0003	0.0249	0.0174	0.0025	-0.0058	0.0239	-0.0008	-0.0185
Wines	0.6883	0.0154	1.0000	0.3870	0.5688	0.3976	0.3902	0.3926	0.0092	0.5537	0.6347	0.6399	-0.3219
Fruits	0.5080	-0.0060	0.3870	1.0000	0.5478	0.5934	0.5716	0.3964	-0.1344	0.3020	0.4862	0.4585	-0.4187
Meat	0.6925	0.0223	0.5688	0.5478	1.0000	0.5735	0.5350	0.3593	-0.1211	0.3070	0.7341	0.4859	-0.5395
Fish	0.5204	0.0003	0.3976	0.5934	0.5735	1.0000	0.5838	0.4271	-0.1431	0.2996	0.5327	0.4576	-0.4464
Sweet	0.5237	0.0249	0.3902	0.5716	0.5350	0.5838	1.0000	0.3573	-0.1212	0.3339	0.4951	0.4551	-0.4223
Gold	0.3892	0.0174	0.3926	0.3964	0.3593	0.4271	0.3573	1.0000	0.0522	0.4070	0.4423	0.3890	-0.2476
Deals	-0.1085	0.0025	0.0092	-0.1344	-0.1211	-0.1431	-0.1212	0.0522	1.0000	0.2416	-0.0119	0.0665	0.3460
WebP	0.4588	-0.0058	0.5537	0.3020	0.3070	0.2996	0.3339	0.4070	0.2416	1.0000	0.3868	0.5162	-0.0512
Catalog	0.6965	0.0239	0.6347	0.4862	0.7341	0.5327	0.4951	0.4423	-0.0119	0.3868	1.0000	0.5177	-0.5220
Store	0.6304	-0.0008	0.6399	0.4585	0.4859	0.4576	0.4551	0.3890	0.0665	0.5162	0.5177	1.0000	-0.4324
WebV	-0.6506	-0.0185	-0.3219	-0.4187	-0.5395	-0.4464	-0.4223	-0.2476	0.3460	-0.0512	-0.5220	-0.4324	1.0000

Table 22: Covariance matrix of Scaled Variables

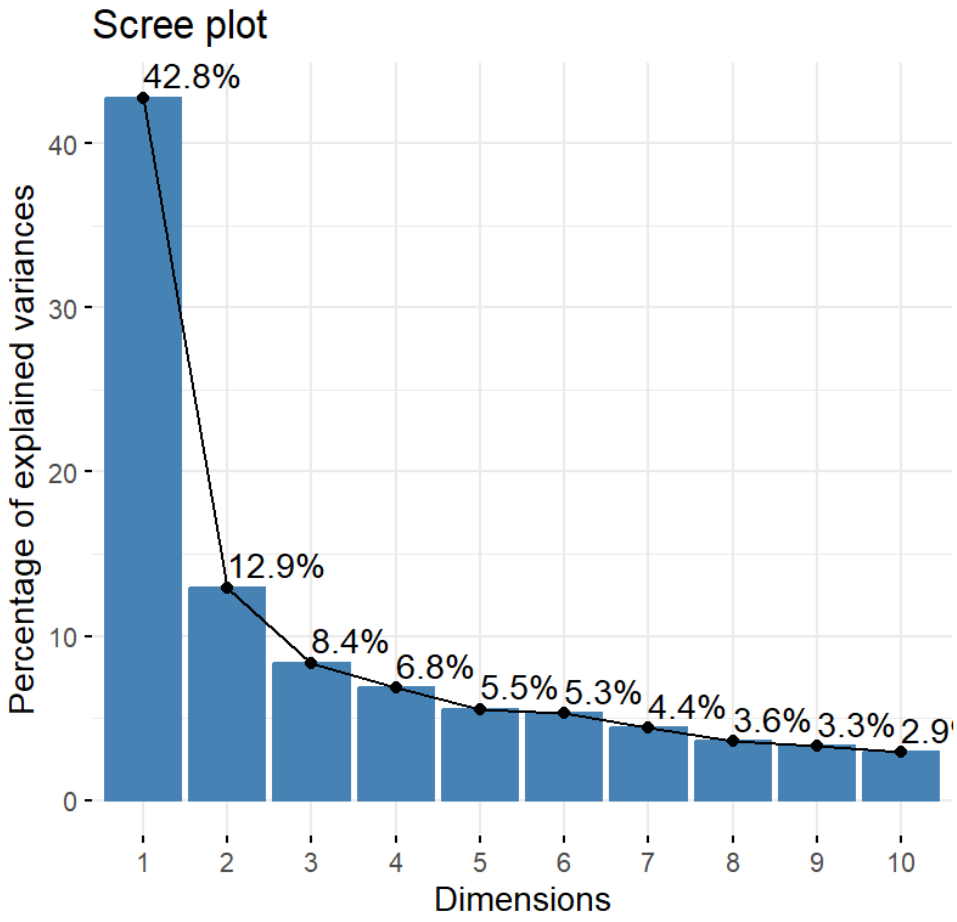
Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.2652	1.2464	1.00132	0.90655	0.81574	0.79766
Proportion of Variance	0.4276	0.1295	0.08355	0.06849	0.05545	0.05302
Cumulative Proportion	0.4276	0.5570	0.64060	0.70908	0.76454	0.81756

	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.72530	0.65648	0.62920	0.59324	0.50829	0.47548
Proportion of Variance	0.04384	0.03591	0.03299	0.02933	0.02153	0.01884
Cumulative Proportion	0.86140	0.89731	0.93030	0.95963	0.98116	1.00000

Table 23: PCA Result

As we can see from above after we conduct PCA, the proportion of variance explained by each principal component, along with the cumulative proportion of variance explained, provides valuable insights. For instance, by examining the data, we observe that the first six components collectively account for a substantial 81.75% of the variability. This is considered highly favorable in terms of capturing and summarizing the essential information in the dataset.

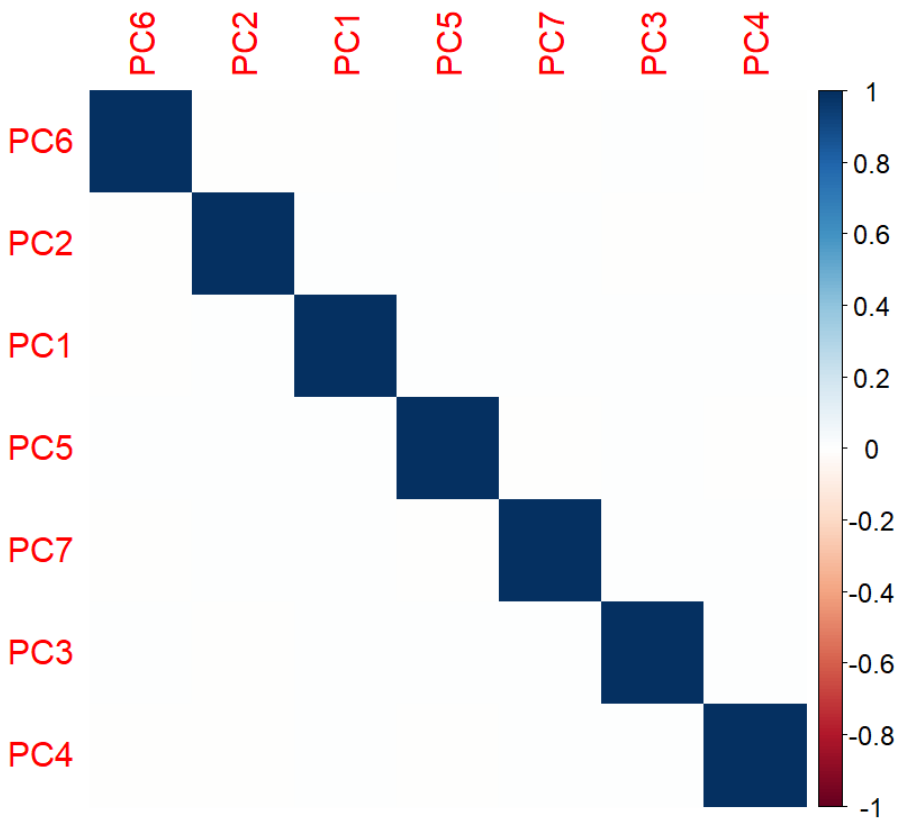


Plot 11: Scree Plot (line plot of the eigenvalues of factors)

As observed, the utilization of three components appears to be satisfactory. Further support for this choice is found in the summary output, indicating that the first seven components account for nearly 86.14% of the variability. This high percentage underscores the effectiveness of these components in capturing and summarizing the essential patterns within the data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
5524	-3.80731257	-0.5666256	0.2699971	-1.35717500	0.9661105	-0.2503492	2.37193912
2174	2.19032418	0.9306120	-0.3357695	0.13313381	0.3112445	-0.3632488	-0.09320716
4141	-1.49136294	-0.1457490	-0.9484400	0.01139915	-0.9600031	1.0688255	-0.59914242
6182	2.03272253	0.5127521	-0.7955286	0.01539069	0.1006477	0.1260047	-0.22100538
5324	0.03641763	-0.7474273	1.5383382	-0.38822738	1.1136274	0.3411494	-0.52672828
7446	-0.64126458	-0.6969099	-1.1678298	0.82750487	-0.2420594	1.1246073	-0.23136417

Table 24: First seven variables extracted Table



Plot 12: Correlation plot of seven variables

As we can observe all the components are linearly independent.

Interpretation of Components:

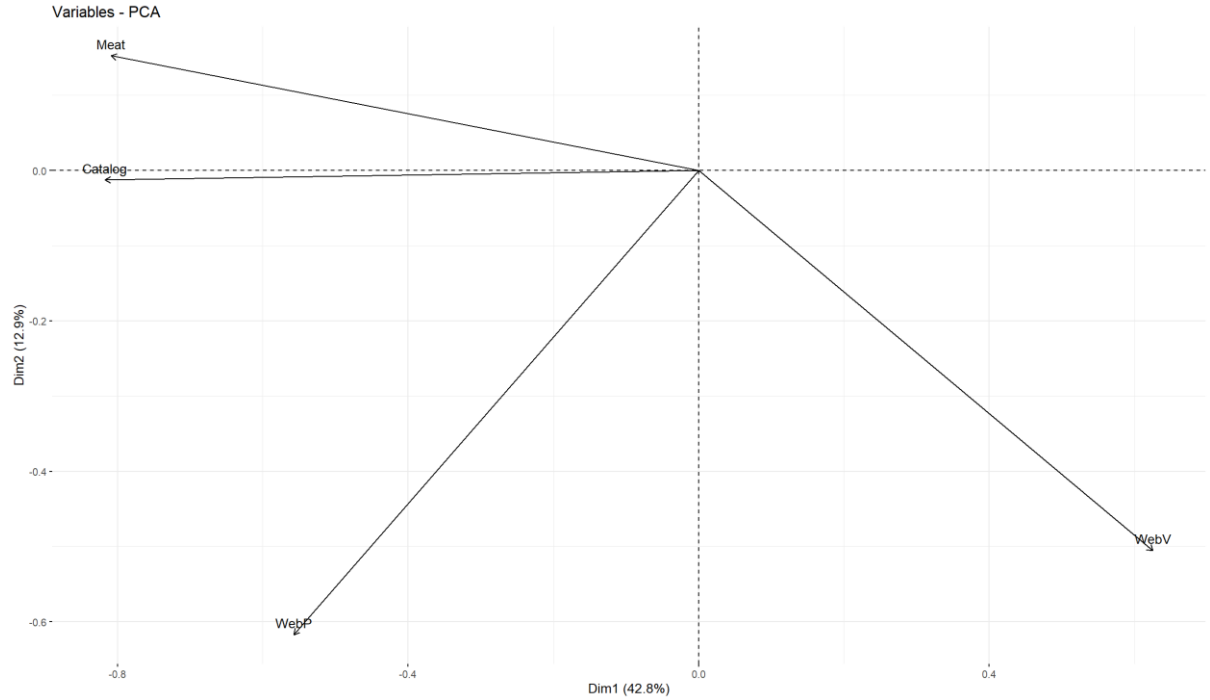
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Recency	-0.01956273	0.01554768	0.9971383731	-0.0423407806	-0.02215687	0.04293615	-0.008176735
Wines	-0.74250158	-0.28154041	0.0176368010	0.3947319264	-0.15453576	0.07322930	0.176913254
Fruits	-0.72947834	0.16469966	-0.0525925223	-0.3562880651	0.08161631	0.16369159	-0.006759164
Meat	-0.80895047	0.15258792	0.0237192964	0.1678103861	0.25415253	-0.12878790	0.253734397
Fish	-0.75281816	0.18015760	-0.0369989814	-0.3269132996	0.08385387	0.03088785	0.041145245
Sweet	-0.72637753	0.14444288	0.0005401344	-0.3172315050	0.12945241	0.28837959	0.018074960
Gold	-0.60076454	-0.22789236	0.0104443813	-0.3512458775	-0.39770650	-0.52855309	-0.081783741
Deals	0.09568769	-0.79125191	0.0286524369	-0.1186824073	0.52528057	-0.15124295	-0.196729601
WebP	-0.55748437	-0.61737514	-0.0233867842	0.0002361941	-0.25404424	0.24816229	0.109361416
Catalog	-0.81739326	-0.01262666	0.0355621550	0.2293349792	0.19602587	-0.26111043	0.191038638
Store	-0.74713239	-0.23538810	-0.0198081839	0.2148732819	-0.08456183	0.22092966	-0.419979290
WebV	0.62554100	-0.50544049	-0.0148229155	-0.2557174586	-0.04260233	0.14400524	0.397516893

Table 25: Corelation plot of scaled variables and components

From above table we can say for PC1 it is positively correlated with Deals and Web visits, also it is negatively correlated with all the other variables. The component is most correlated with Catalog variable. After this let us check some plots.

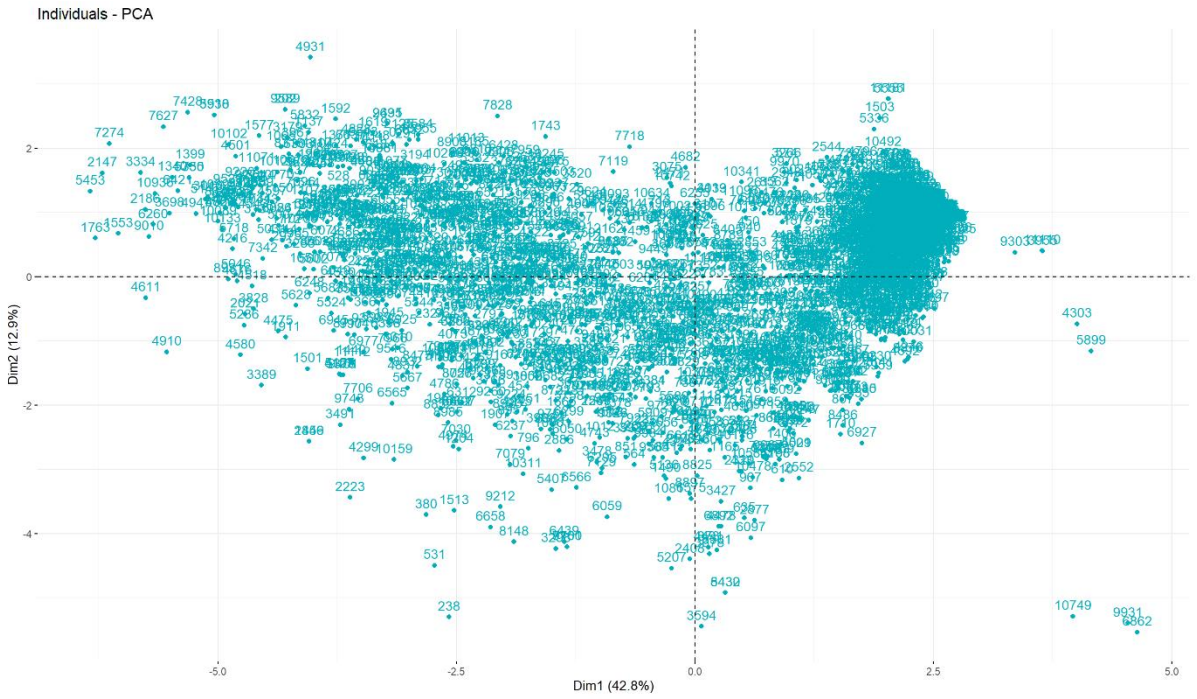
Plot 13: Biplot for PC1 and PC2

From this plot we can see that closer the vectors more correlated with each other, aslo we say that customer with ID 4611 has the most associaation with Catalog variable.



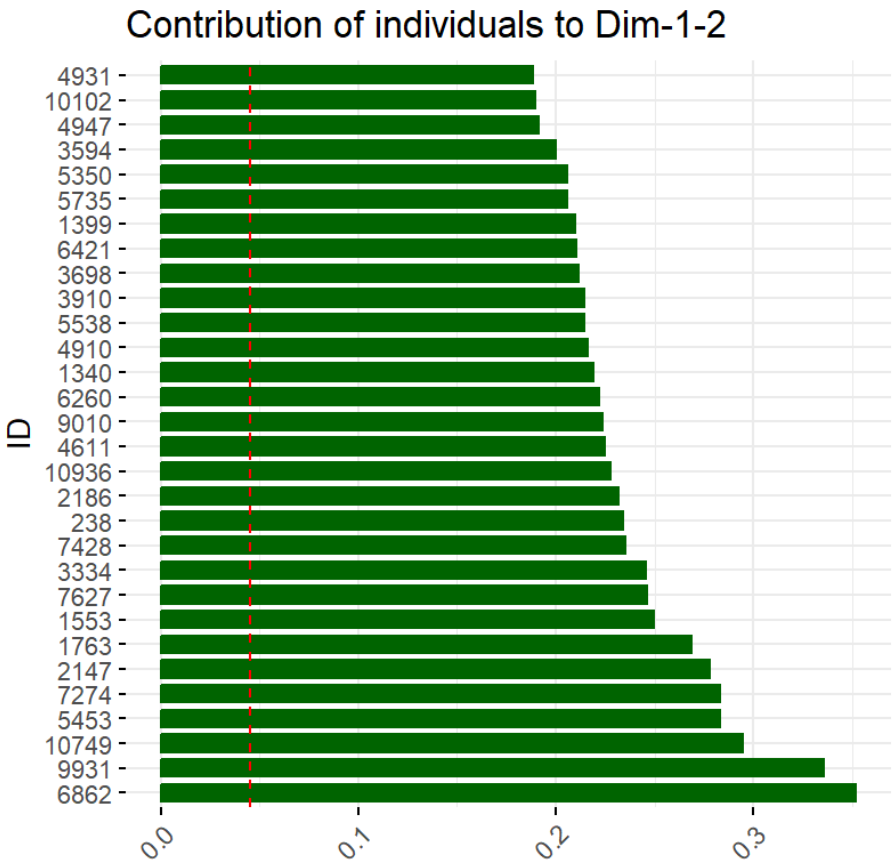
Plot 14: First 4 variables having the highest contribution

As we can see from the plot Meat, Catalog, Web purchase, and Web visit variables having the highest contribution to the first two components.



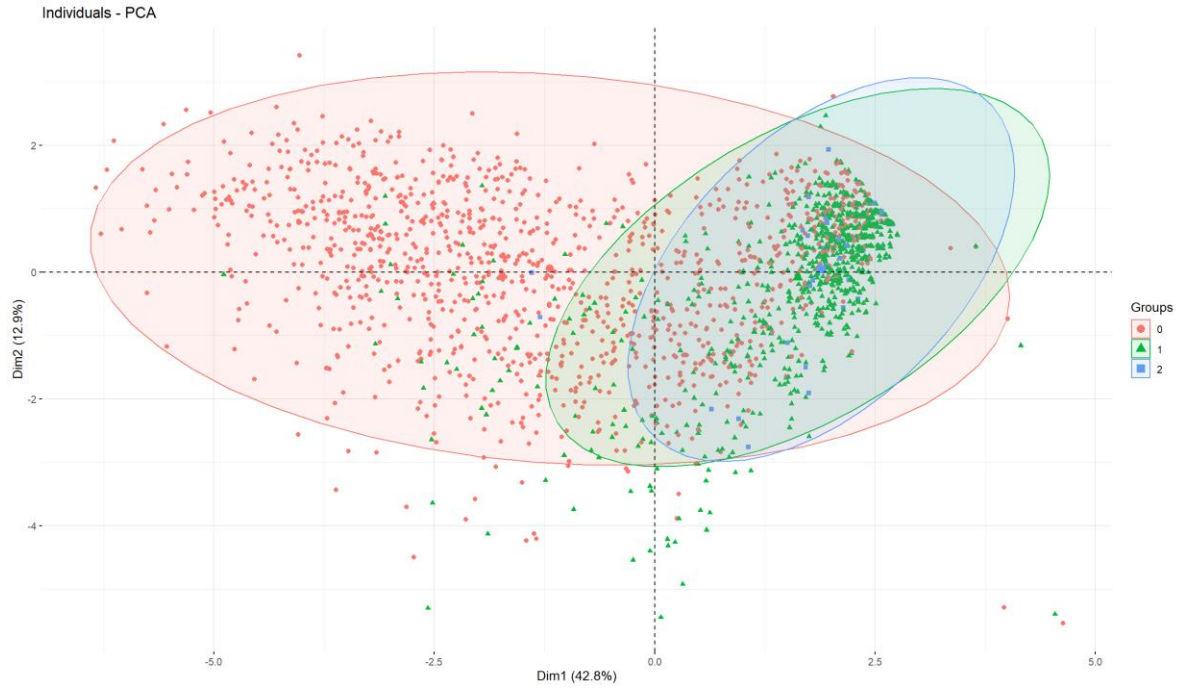
Plot 15: Individuals – PCA

From this plot we can see that which component is good in explanation of variables.



Plot 16: Contribution of Individuals

From the above plot we can see that customer with ID 6862 provides the highest contribution to the first two component.



Plot 17: Observatons by Kidhome with factor 0,1, and 2

```
Call:
lm(formula = Income ~ ., data = ols.data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4385 -0.2894 -0.0019  0.2841  5.5213

Coefficients:
                         Estimate             Std. Error t value             Pr(>|t|)
(Intercept) -0.00000000000000002179  0.01115170489882117907   0.000              1.00000
PC1         -0.35802823629727154042  0.00492418505397152623 -72.708 < 0.0000000000000002 ***
PC2          0.02610864063909990734  0.00894936234671424129   2.917              0.00357 **
PC3          0.00195075894703244370  0.01113948677660865985   0.175              0.86100
PC4          0.27943043553573904125  0.01230404644106897899  22.710 < 0.0000000000000002 ***
PC5          0.00920591515170010068  0.01367369614076540361   0.673              0.50085
PC6          0.00007880937638983541  0.01398372417043868172   0.006              0.99550
PC7         -0.06752038790817622982  0.01537876263868326147  -4.390            0.0000118 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5248 on 2207 degrees of freedom
Multiple R-squared:  0.7254,    Adjusted R-squared:  0.7245
F-statistic: 832.9 on 7 and 2207 DF,  p-value: < 0.00000000000000022
```

Table 26: Regression output with reduced model

```
Call:
lm(formula = Income ~ ., data = pca)

Residuals:
   Min     1Q Median     3Q    Max
-93586  -6104   -129   5797 111981

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 51719.7347  1082.3444  47.785 < 0.0000000000000002 ***
Recency        -7.1074     8.0360  -0.884              0.37655
Wines          17.2209     1.0906  15.790 < 0.0000000000000002 ***
Fruits         10.7533     8.1173   1.325              0.18539
Meat           15.7076     1.7191   9.137 < 0.0000000000000002 ***
Fish           -0.6152     6.1062  -0.101              0.91976
Sweet          21.9755     7.7765   2.826              0.00476 **
Gold           -9.2537     5.4155  -1.709              0.08764 .
Deals          45.8170   139.2248   0.329              0.74212
WebP         1220.1062   115.4954  10.564 < 0.0000000000000002 ***
Catalog       716.3626   135.8052   5.275          0.000000146 ***
Store         457.7062   107.5284   4.257          0.000021626 ***
WebV        -3269.6949   135.5459 -24.122 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10930 on 2202 degrees of freedom
Multiple R-squared:  0.7436,    Adjusted R-squared:  0.7422
F-statistic: 532.2 on 12 and 2202 DF,  p-value: < 0.00000000000000022
```

Table 27: Regression output with full model

When we look at the above tables, if we want to compare them MSE with reduced model is highly lower than the full model. Also Adjusted R-squared value is bigger at reduced model,

and the difference between maximum and minimum point is significantly decreased after the PCA with the reduced model. To sum up, using PCA gives us better regression coefficients and lower standard error with less parameters.

3.5 - FACTOR ANALYSIS AND FACTOR ROTATION

We continue the report with Factor Analysis. In this analysis, we considered 18 variables, 3 of which were categorical. Their structure and first 10 lines are as follows.

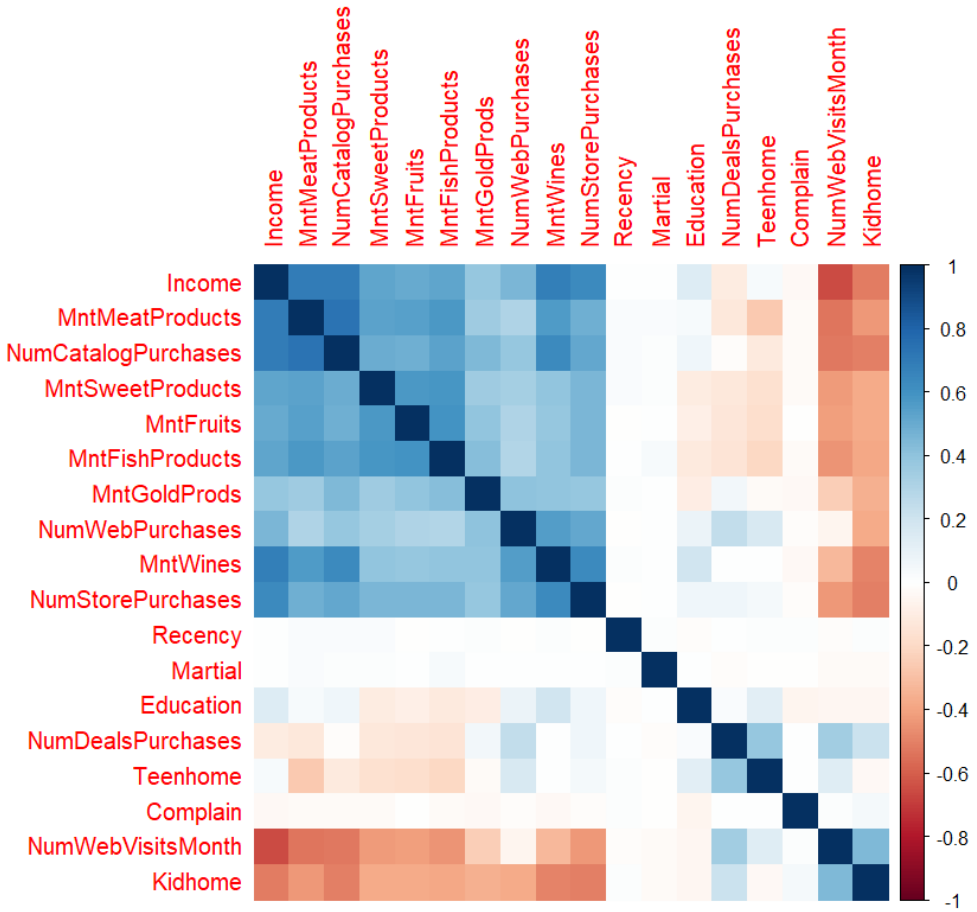
```
> str(fa)
'data.frame':   2215 obs. of  18 variables:
 $ Income             : int  58138 46344 71613 26646 58293 62513 55635 33454 30351 5648 ...
 $ Recency            : int  58 38 26 26 94 16 34 32 19 68 ...
 $ MntWines           : int  635 11 426 11 173 520 235 76 14 28 ...
 $ MntFruits          : int  88 1 49 4 43 42 65 10 0 0 ...
 $ MntMeatProducts    : int  546 6 127 20 118 98 164 56 24 6 ...
 $ MntFishProducts    : int  172 2 111 10 46 0 50 3 3 1 ...
 $ MntSweetProducts   : int  88 1 21 3 27 42 49 1 3 1 ...
 $ MntGoldProds       : int  88 6 42 5 15 14 27 23 2 13 ...
 $ NumDealsPurchases  : int  3 2 1 2 5 2 4 2 1 1 ...
 $ NumWebPurchases    : int  8 1 8 2 5 6 7 4 3 1 ...
 $ NumCatalogPurchases: int  10 1 2 0 3 4 3 0 0 0 ...
 $ NumStorePurchases  : int  4 2 10 4 6 10 7 4 2 0 ...
 $ NumWebVisitsMonth  : int  7 5 4 6 5 6 6 8 9 20 ...
 $ Kidhome            : int  1 2 1 2 2 1 1 2 2 2 ...
 $ Teenhome           : int  0 1 0 0 0 1 1 0 0 1 ...
 $ Complain           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Education          : int  3 3 3 3 5 4 3 5 5 5 ...
 $ Martial            : int  5 5 6 6 4 6 3 4 6 6 ...
```

Table 28: Structure of the data for factor analysis

```
> head(fa)
  Income Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds
1  58138      58      635        88             546             172              88           88
2  46344      38       11         1               6               2               1            6
3  71613      26      426        49             127             111              21           42
4  26646      26       11         4              20              10               3            5
5  58293      94      173        43             118              46              27           15
6  62513      16      520        42              98               0              42           14
  NumDealsPurchases NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth Kidhome
1                 3                8                  10                 4                 7       1
2                 2                1                   1                 2                 5       2
3                 1                8                   2                10                 4       1
4                 2                2                   0                 4                 6       2
5                 5                5                   3                 6                 5       2
6                 2                6                   4                10                 6       1
  Teenhome Complain Education Martial
1        0        0         3       5
2        1        0         3       5
3        0        0         3       6
4        0        0         3       6
5        0        0         5       4
6        1        0         4       6
```

Table 29: The first 6 observations of the data

After removing the NA values in our data, we can examine our FA analysis. For this, we first visualized the correlation in our data. As we can see from below, we have many variables that are correlated. For example, “Income” and “NumCatalogPurchase” are variables that show high correlation with “MntMeatProducts”.



Plot 18: Correlation plot of variables

Then, it is checked whether Factor Analysis is appropriate or not. There are 2 methods for this. The first one, the Kaiser-Meyer-Olkin (KMO) test, was performed by removing the dependent variable from the data. Looking at the result of this, it can be seen that the Overall MSA value is 0.88 and is greater than 0.5.

```
> KMO(r=cm)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cm)
Overall MSA =  0.88
MSA for each item =
              Income             Recency            MntWines           MntFruits     MntMeatProducts
                0.88                0.32                0.88                0.94                0.91
   MntFishProducts    MntSweetProducts        MntGoldProds   NumDealsPurchases     NumWebPurchases
                0.94                0.94                0.94                0.51                0.86
NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth            Kidhome            Teenhome
                0.92                0.91                0.82                0.89                0.52
           Complain           Education             Martial
                0.62                0.63                0.52
```

Table 30: Kaiser-Meyer-Olkin (KMO) test

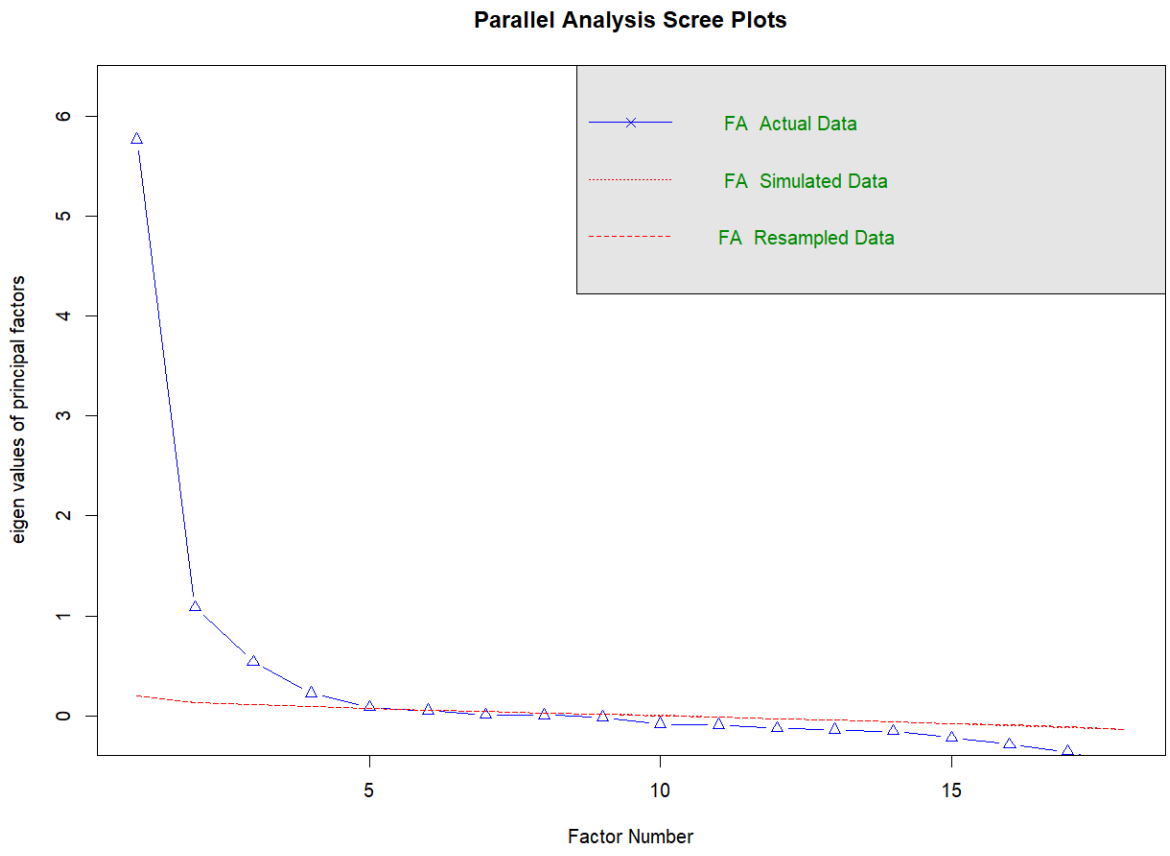
The second method we can use to prove that Factor Analysis is a suitable analysis for our data is Bartlett's test of sphericity.

```
> print(cortest.bartlett(cm,nrow(fa)))
$chisq
[1] 17157.21

$p.value
[1] 0

$df
[1] 153
```

Table 31: Bartlett's test of sphericity

The Chi-square value is 17157.21 with 153 degrees of freedom. Therefore, test is significant at 0.05 significance level. So, it can be stated that Factor Analysis will be an appropriate technique for Customer Personality data.

In the continuation of the factor analysis, it is decided how many factors there will be. There are multiple methods for this. First, we can look at the scree plot.



Plot 19: Scree Plot

Looking at this graph shows that after factor 9, the total variance explains much smaller amounts.

It is necessary to test to understand whether the result obtained from the graph is really correct. For this, the factanal function was used and by testing how many factors there should be, it was observed that when there were 9 factors, the p-value was greater than 0.05. As a result, it was

decided that 9 factors were sufficient and Varimax solution or rotation was performed with them.

```
> factanal(fa, factors = 7)$PVAL
        objective
0.000000006375027
> factanal(fa, factors = 8)$PVAL
  objective
0.007651545
> factanal(fa, factors = 9)$PVAL
objective
 0.186481
```

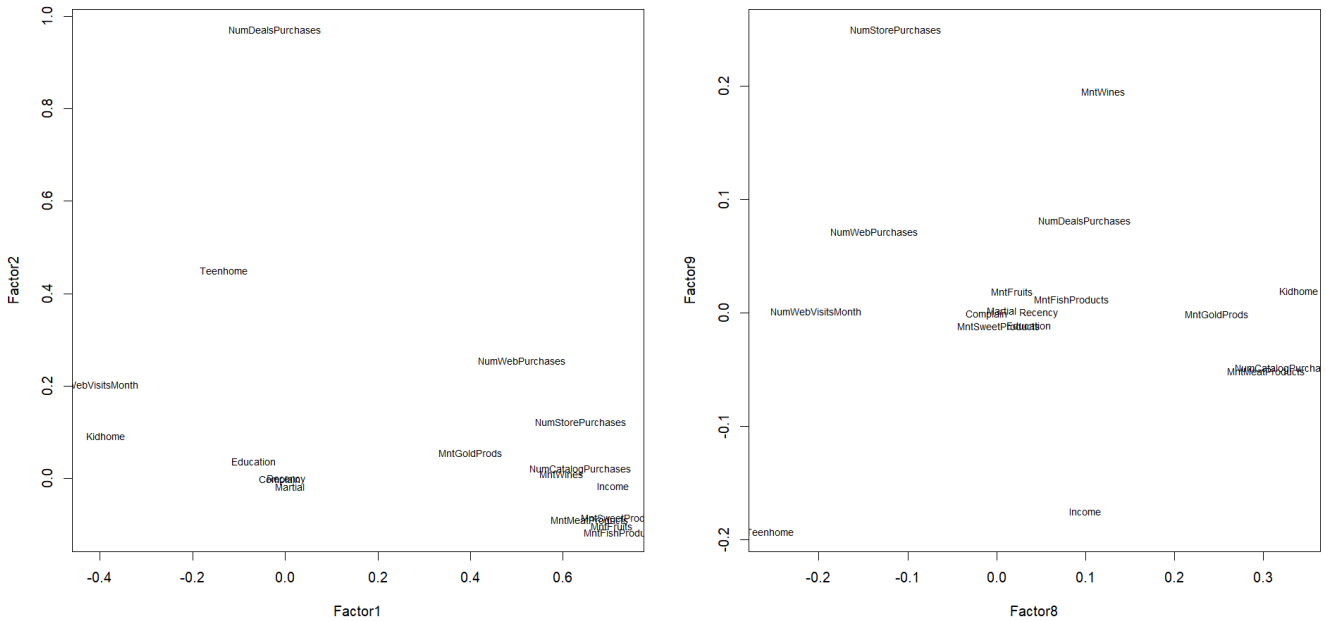
Table 32: Factanal test

```
> fa_test

Call:
factanal(x = fa, factors = 9)

Uniquenesses:
            Income             Recency           MntWines          MntFruits     MntMeatProducts     MntFishProducts    MntSweetProducts
             0.124               0.998               0.172               0.435               0.005               0.384               0.422
       MntGoldProds   NumDealsPurchases     NumWebPurchases NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth            Kidhome
             0.005               0.005               0.449               0.265               0.322               0.005               0.005
          Teenhome            Complain           Education             Martial
             0.565               0.997               0.819               0.998

Loadings:
                  Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8 Factor9
Income             0.709           0.191   0.431  -0.300   0.111           0.100  -0.175
Recency
MntWines           0.597           0.347   0.529           0.117           0.120   0.195
MntFruits          0.706  -0.101           -0.134  -0.103           0.111
MntMeatProducts    0.657           0.143   0.170                   0.635   0.303
MntFishProducts    0.725  -0.115   0.115  -0.171  -0.104
MntSweetProducts   0.729                   -0.124  -0.109
MntGoldProds       0.401           0.314  -0.148           0.802           0.248
NumDealsPurchases          0.972                   0.172
NumWebPurchases    0.511   0.255   0.200   0.268   0.158   0.244          -0.138
NumCatalogPurchases 0.638          0.368   0.184  -0.106           0.193   0.325
NumStorePurchases  0.639   0.124   0.193   0.278  -0.183   0.169          -0.113   0.250
NumWebVisitsMonth -0.414   0.205  -0.172           0.825          -0.150  -0.202
Kidhome           -0.386          -0.790           0.197          -0.214   0.340
Teenhome          -0.132   0.451           0.233                  -0.199  -0.255  -0.193
Complain
Education                                  0.406
Martial

                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8 Factor9
SS loadings       4.425   1.323   1.179   0.995   0.958   0.801   0.599   0.558   0.186
Proportion Var    0.246   0.073   0.066   0.055   0.053   0.044   0.033   0.031   0.010
Cumulative Var    0.246   0.319   0.385   0.440   0.493   0.538   0.571   0.602   0.612

Test of the hypothesis that 9 factors are sufficient.
The chi square statistic is 33.33 on 27 degrees of freedom.
The p-value is 0.186
```

Table 33: 9-factor solution for factanal test

Various information about factor loadings can be obtained by looking at this solution. For instance, Factor 1 reflects “MntFishProducts” and “MntSweetProducts” while Factor 2 reflects only “NumDealsPurchases”. Also, it can be said that Factor 5 is dominated by the “NumWebVisitsMonth” and Factor 6 is dominated by the “MntGoldProds” as an example interpretation of test result.

Moreover, according to the test results, The first 9 factors explain almost 61% of the variance. We can effectively reduce dimensionality from 18 to 9 while losing about 39% of the variance. The first factor explains 24.6% of the variance, the second factor explains 7.3%, and the ninth factor explains 1% of the variance.



Plot 20: Visualization of the factor model

In addition, we can reach the same test results above, for example, the first factor is dominated by “MntFishProducts” and “MntSweetProducts”, by looking at this plot.

Then, Cronbach’s alpha should be run. Because of this, consistency of each factor is determined.

```
> names(fa_test$loadings[,1])[abs(fa_test$loadings[,1])>0.4]
 [1] "Income"             "MntWines"           "MntFruits"          "MntMeatProducts"
 [5] "MntFishProducts"    "MntSweetProducts"   "MntGoldProds"       "NumWebPurchases"
 [9] "NumCatalogPurchases" "NumStorePurchases"  "NumWebVisitsMonth"
```

Table 33: Consistency of the first factor

```
> names(fa_test$loadings[,2])[abs(fa_test$loadings[,2])>0.4]
[1] "NumDealsPurchases" "Teenhome"
> names(fa_test$loadings[,3])[abs(fa_test$loadings[,3])>0.4]
[1] "Kidhome"
> names(fa_test$loadings[,4])[abs(fa_test$loadings[,4])>0.4]
[1] "Income"    "MntWines"  "Education"
> names(fa_test$loadings[,5])[abs(fa_test$loadings[,5])>0.4]
[1] "NumWebVisitsMonth"
> names(fa_test$loadings[,6])[abs(fa_test$loadings[,6])>0.4]
[1] "MntGoldProds"
> names(fa_test$loadings[,7])[abs(fa_test$loadings[,7])>0.4]
[1] "MntMeatProducts"
> names(fa_test$loadings[,8])[abs(fa_test$loadings[,8])>0.4]
character(0)
> names(fa_test$loadings[,9])[abs(fa_test$loadings[,9])>0.4]
character(0)
```

Table 34: Consistency of the other factors

After that, Reliability analysis is conducted for each actor which can show the consistency of variables within the factor.

```
Reliability analysis
 raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
      0.34      0.56    0.39      0.39 1.3 0.014  1.4 1.1     0.39
```

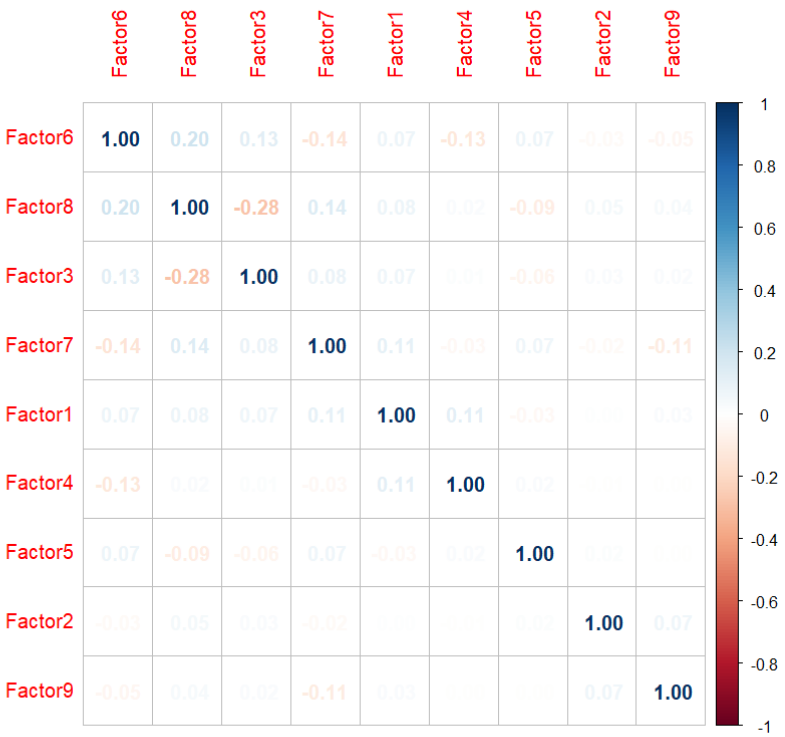
Table 35: Reliability analysis

For example, raw alpha is 0.34 for the first factor. So, it can be said that it is not very consistant. Then, estimated factor scores for each individual are calculated and the first 6 are shown below.

```
> head(factor_scores)
     Factor1     Factor2     Factor3     Factor4    Factor5     Factor6     Factor7    Factor8     Factor9
1  1.7821857  0.03439395  0.98650247 -1.04551309  2.1295177 -0.5610586  0.73639976  0.7908102 -0.4286473
2 -0.7777491 -0.10835937 -0.81861510  0.09335976 -0.6696502 -0.3414170 -0.44704773  0.3716546 -0.7174014
3  0.8951545 -0.53386045  0.09861736  0.22285113 -0.4742977 -0.1420164 -0.76091919 -1.2301580  0.3925446
4 -0.7470346 -0.30308809 -0.78723568 -0.53154784 -0.1210048 -0.4133867 -0.02265031  0.2568259  0.6485175
5  0.3664295  1.35298102 -0.92763040 -0.03434035 -0.4569876 -0.8625252 -0.55123148  0.6305138  0.1622148
6  0.3987658 -0.07770346  0.46811851  0.98005072  0.3293109 -0.5500861 -0.54348923 -1.3097633  0.1007780
```

Table 36: Estimated factor scores for individuals

At the end of the Factor Analysis, it can be observed that there is almost no correlation.



Plot 21: Correlation plot of variables

3.6 - DISCRIMINATION AND CLASSIFICATION

In this analysis, our objective is classifying the customer according to whether or not he/she accepts the offer in the last campaign, based on information such as the customer's age, income, the money he/she has spent on wine or fruit in the last two years, and how much he/she has purchased from the company's website or catalogue.

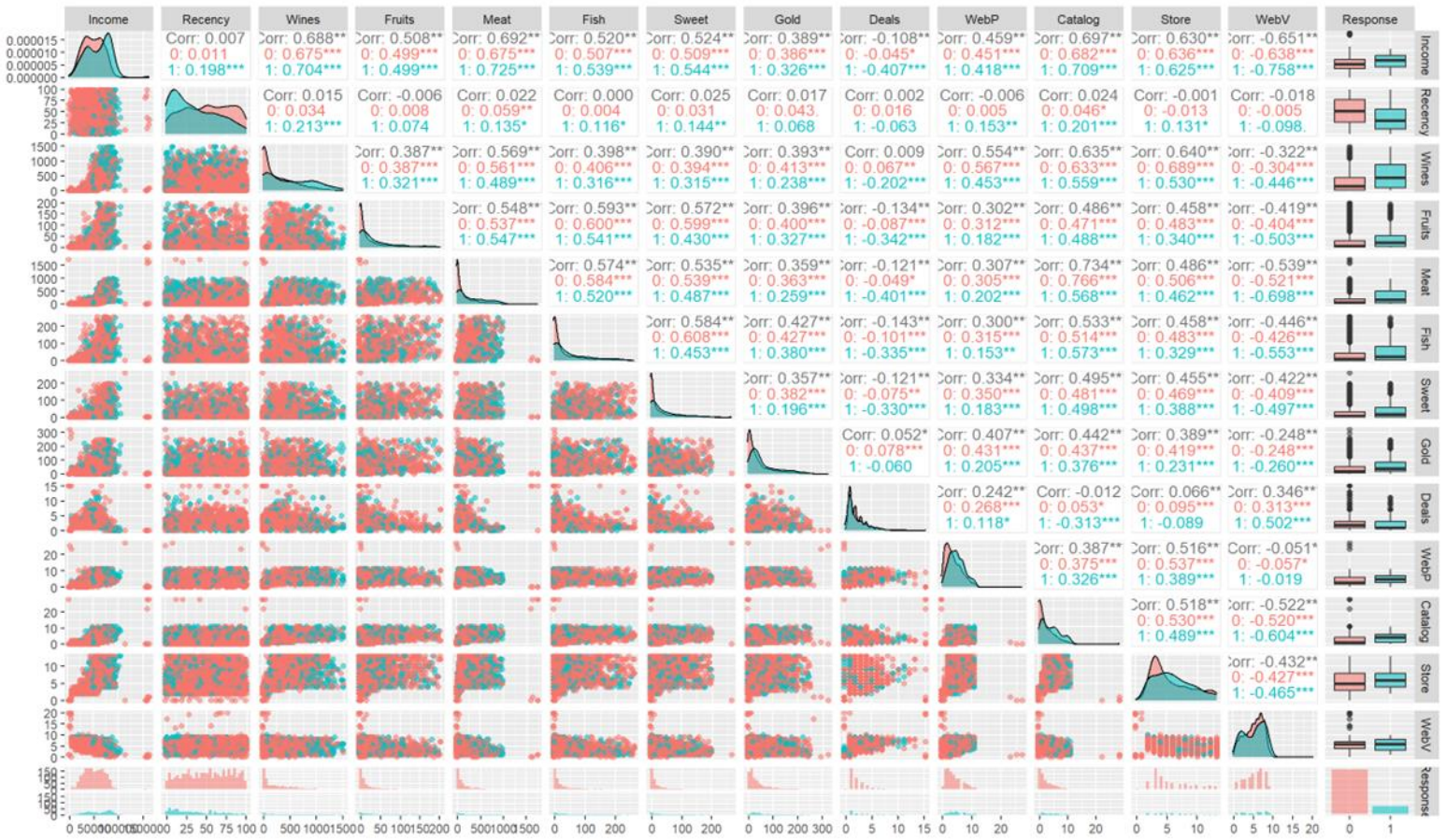
Below is the summary and structure of the data created for Fisher Discriminant Analysis.

```
> str(LDA)
'data.frame':   2215 obs. of  15 variables:
 $ Income  : int  58138 46344 71613 26646 58293 62513 55635 33454 30351 5648 ...
 $ Recency : int  58 38 26 26 94 16 34 32 19 68 ...
 $ Wines   : int  635 11 426 11 173 520 235 76 14 28 ...
 $ Fruits  : int  88 1 49 4 43 42 65 10 0 0 ...
 $ Meat    : int  546 6 127 20 118 98 164 56 24 6 ...
 $ Fish    : int  172 2 111 10 46 0 50 3 3 1 ...
 $ Sweet   : int  88 1 21 3 27 42 49 1 3 1 ...
 $ Gold    : int  88 6 42 5 15 14 27 23 2 13 ...
 $ Deals   : int  3 2 1 2 5 2 4 2 1 1 ...
 $ WebP    : int  8 1 8 2 5 6 7 4 3 1 ...
 $ Catalog : int  10 1 2 0 3 4 3 0 0 0 ...
 $ Store   : int  4 2 10 4 6 10 7 4 2 0 ...
 $ WebV    : int  7 5 4 6 5 6 6 8 9 20 ...
 $ Response: Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
 $ Age     : num  66 69 58 39 42 56 52 38 49 73 ...
```

Table 37: Structure of the data

```
> summary(LDA)
     Income          Recency          Wines           Fruits           Meat            Fish      
 Min.   :  1730   Min.   : 0.00   Min.   :   0.0   Min.   :  0.00   Min.   :   0.0   Min.   :  0.00  
 1st Qu.: 35284   1st Qu.:24.00   1st Qu.:  24.0   1st Qu.:  2.00   1st Qu.:  16.0   1st Qu.:  3.00  
 Median : 51373   Median :49.00   Median : 175.0   Median :  8.00   Median :  68.0   Median : 12.00  
 Mean   : 51970   Mean   :49.02   Mean   : 305.2   Mean   : 26.36   Mean   : 167.1   Mean   : 37.65  
 3rd Qu.: 68487   3rd Qu.:74.00   3rd Qu.: 505.0   3rd Qu.: 33.00   3rd Qu.: 232.5   3rd Qu.: 50.00  
 Max.   :162397   Max.   :99.00   Max.   :1493.0   Max.   :199.00   Max.   :1725.0   Max.   :259.00  
     Sweet            Gold            Deals            WebP          Catalog          Store       
 Min.   :  0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 0.000   1st Qu.: 3.000  
 Median :  8.00   Median : 25.00   Median : 2.000   Median : 4.000   Median : 2.000   Median : 5.000  
 Mean   : 27.04   Mean   : 43.98   Mean   : 2.323   Mean   : 4.086   Mean   : 2.672   Mean   : 5.802  
 3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000   3rd Qu.: 6.000   3rd Qu.: 4.000   3rd Qu.: 8.000  
 Max.   :262.00   Max.   :321.00   Max.   :15.000   Max.   :27.000   Max.   :28.000   Max.   :13.000  
      WebV          Response       Age       
 Min.   : 0.000   0:1882   Min.   : 27.00  
 1st Qu.: 3.000   1: 333   1st Qu.: 46.00  
 Median : 6.000            Median : 53.00  
 Mean   : 5.319            Mean   : 54.18  
 3rd Qu.: 7.000            3rd Qu.: 64.00  
 Max.   :20.000            Max.   :130.00  
```

Table 38: Summary of the data



Plot 21: Multiple plots for variables

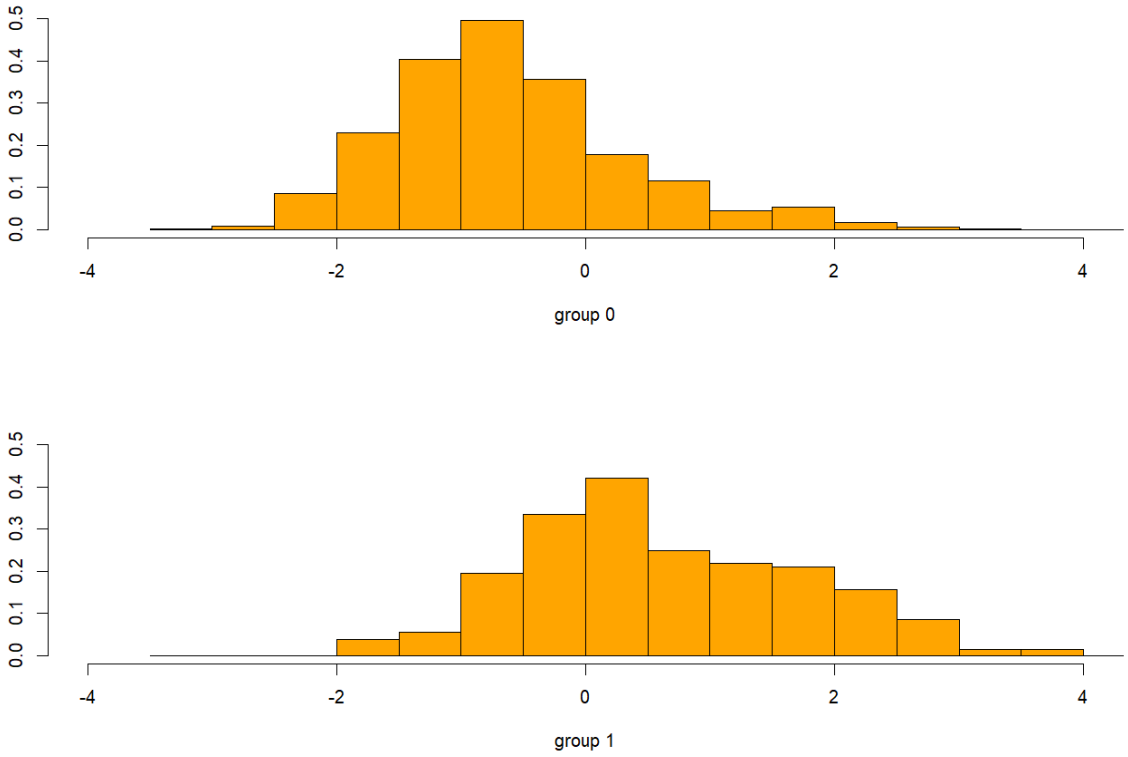
Then the data is divided into two: 80% for train and 20% for test and model is conducted to get the values of the loadings of the discriminant functions.

```
> model1
Call:
lda(Response ~ ., data = train)

Prior probabilities of groups:
       0        1
0.853211 0.146789

Group means:
    Income  Recency    Wines   Fruits     Meat     Fish    Sweet     Gold    Deals     WebP  Catalog   Store     WebV      Age
0 50089.59 51.33871 271.1532 24.32527 140.5114 35.05914 24.97245 40.74194 2.326613 3.913978 2.391129 5.77957 5.352151 54.43011
1 59514.79 36.04297 493.5195 37.82031 285.3828 53.29688 37.73828 59.32422 2.320312 5.117188 4.261719 6.12500 5.320312 53.51953

Coefficients of linear discriminants:
                  LD1
Income  -0.000001551038
Recency -0.017780115677
Wines    0.001507670044
Fruits   0.000926344988
Meat     0.002715977304
Fish    -0.001252428067
Sweet   -0.000190027418
Gold     0.001199416745
Deals    0.007697412988
WebP     0.046739256964
Catalog  0.137910662835
Store   -0.196457255391
WebV     0.134069641219
Age     -0.008895000557
```

Table 39: Model for all predictors

When it is examined the output, it can be seen that π for not accepting the offer in the last campaign which is prior probability of group 0 is 0.85 and π for accepting the offer in the last campaign which is prior probability of group 1 is 0.14. Furthermore, independent variables' averages for each of them within each class which are group means in the output can be seen above.

The model is -0.000001551038*Income - 0.017780115677*Recency +…… - 0.008895000557*Age from the output.



Plot 22: Hostogram of linear discriminants

Then, the prediction of the model is conducted.

```
> names(pred_values)
[1] "class"     "posterior" "x"
```

Table 40: Prediction of the model

To test performance of the model, confusion matrix can be used. Also, accuracy of the model is obtained.

```
> table_train_1
         Actual
Predicted    0    1
        0 1426  189
        1   62   67
> sum(diag(table_train_1))/sum(table_train_1)
[1] 0.856078
```

Table 41: Accuracy of the model for train

```
> table_test_2
         Actual
Predicted   0   1
        0 380  52
        1  14  25
> sum(diag(table_test_2))/sum(table_test_2)
[1] 0.8598726
```

Table 42: Accuracy of the model for test

According to the results, the model correctly classifies the customers with 0.856 probability for the training data which is good and the misclassification rate is 0.144. Moreover, 189 customers who really accept the offer in the last campaign and 62 customers who really do not accept the offer in the last campaign is peredicted wrong.

Fort he test performance, it can be said that the model correctly classifies the customers with 0.86 probability and the classification error rate is 1.14. Similarly, 25 customers who really accept the offer in the last campaign and 380 customers who really do not accept the offer in the last campaign is peredicted correct.

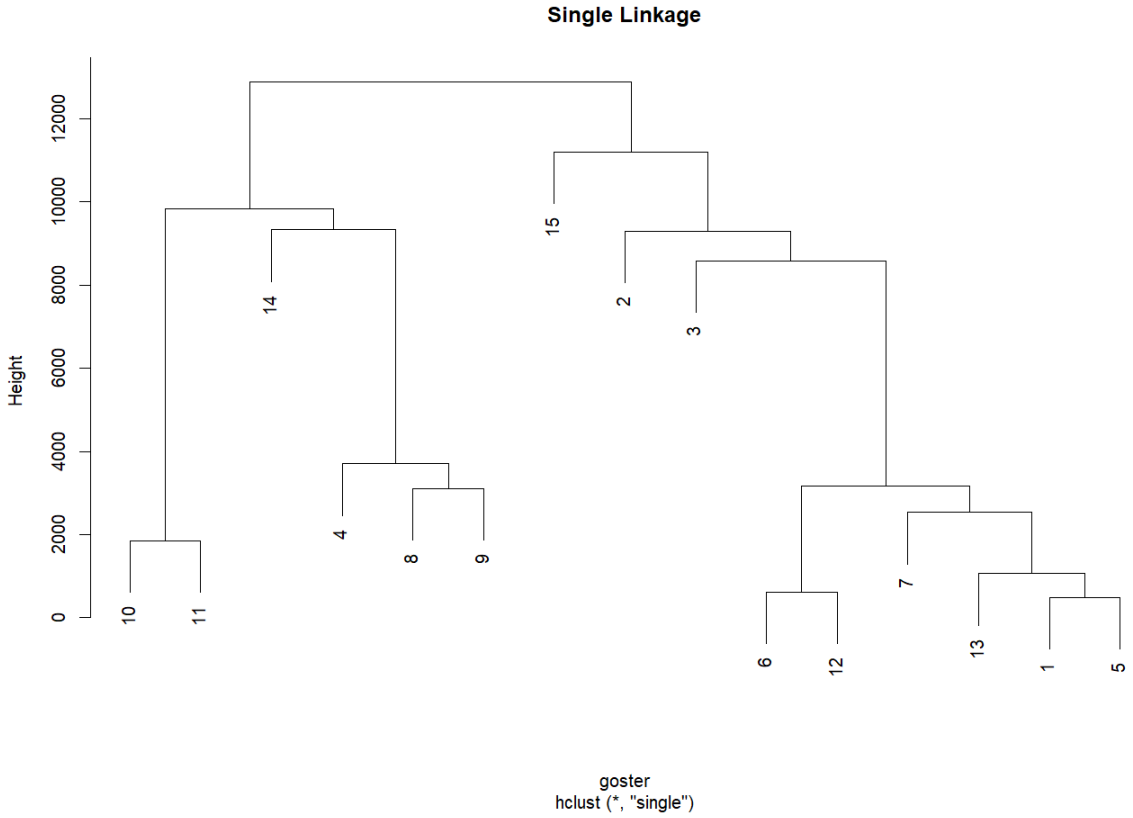
3.7 - CLUSTERING

Agglomerative hierarchical clustering

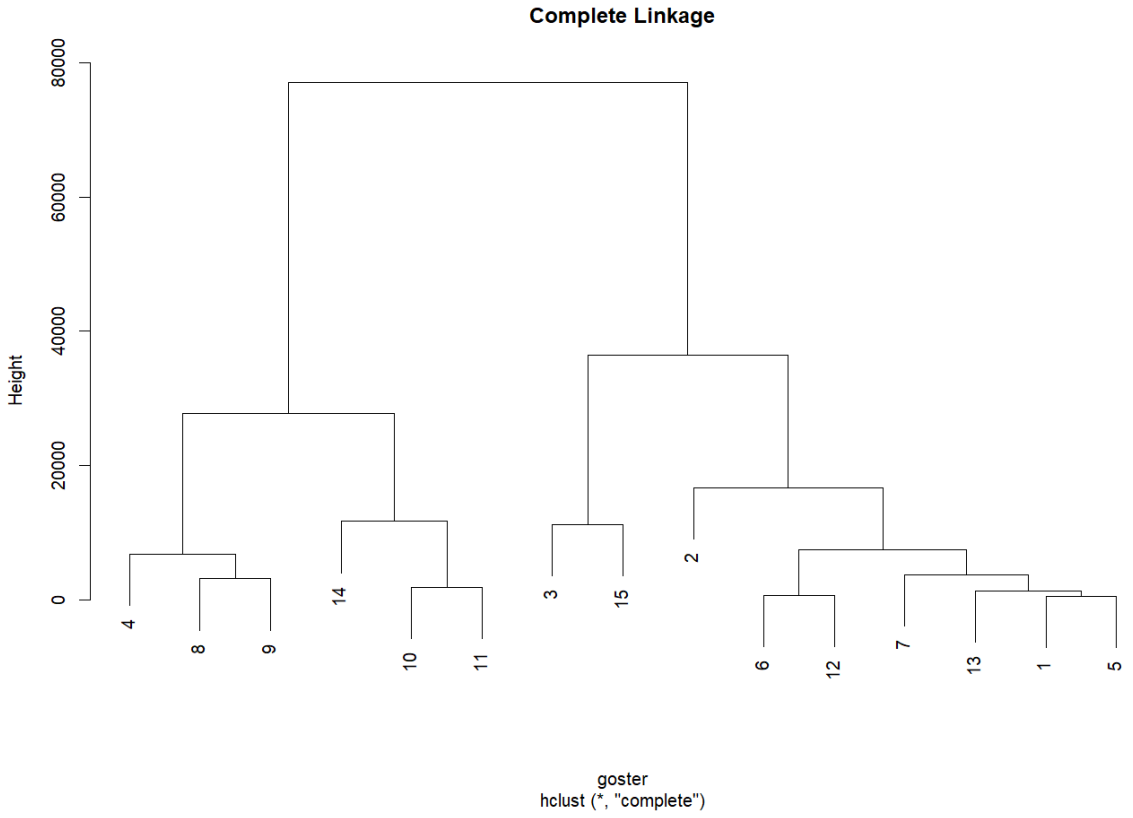
The variables Income, Age and MntWines were selected for this analysis. Then, Euclidean distances and dendrogram were created. Below you can see the first 15 lines of these distances and the created dendrograms.

```
> Distances_first_15
          1         2         3         4         5         6         7         8         9        10        11        12        13        14
2  11810.4962
3  13476.6231 25272.4100
4  31498.1931 19698.0228 44968.9190
5    487.8986 11950.1286 13322.4121 31647.4148
6   4376.5226 16177.0149  9100.4857 35870.6155  4234.2656
7   2534.7988  9293.7154 15979.1427 28989.8683  2658.7418  6883.9033
8  24690.3447 12890.2012 38160.6103  6808.3104 24839.1897 29062.3974 22181.5743
9  27793.9436 15993.0128 41264.0578  3705.0147 27942.4533 32165.9809 25284.9660  3103.6388
10 52493.5101 40696.0037 65966.2024 20998.0344 52645.2088 56867.1309 49987.4330 27806.0635 24703.0156
11 50641.9100 38844.0066 64114.3766 19146.0023 50793.2748 55015.4019 48135.5450 25954.0960 22851.0015  1852.3131
12  4914.8255 16690.0040  8583.1381 36387.4688  4740.0976   613.7915  7398.1233 29579.2468 32682.4991 57385.2408 55533.3208
13  1280.7361 13011.8941 12260.5260 32708.7690  1063.0908  3172.0459  3719.0491 25900.4969 29003.8352 53706.3913 51854.5024  3679.2134
14 40819.9038 29021.0199 54291.6523  9323.0039 40970.3531 45192.9617 38312.7058 16131.1653 13028.0111 11675.0854  9823.0066 45710.4076 42031.6439
15 24664.7928 36469.5767 11202.0413 56162.8274 24521.1778 20292.8314 27175.9506 49354.7783 52458.3878 77158.1986 75306.6458 19783.6751 23458.7400 65484.6945
```

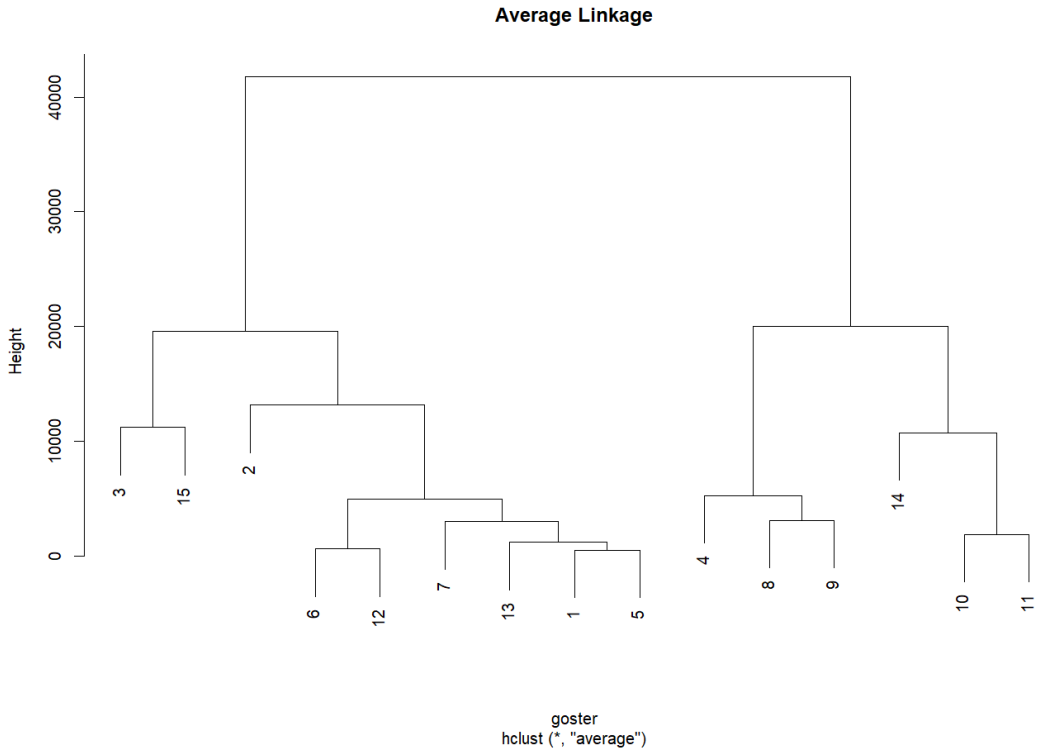
Table 43: Euclidean distances



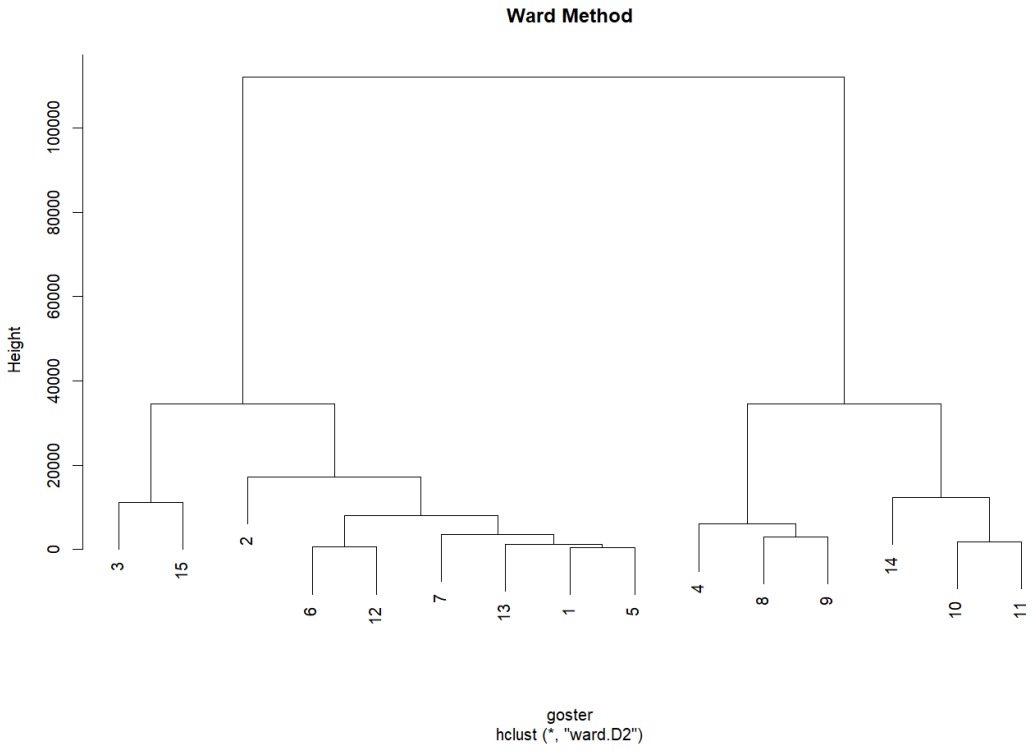
Plot 23: Dendrogram with Single method



Plot 24: Dendrogram with Complete method



Plot 25: Dendrogram with Average method

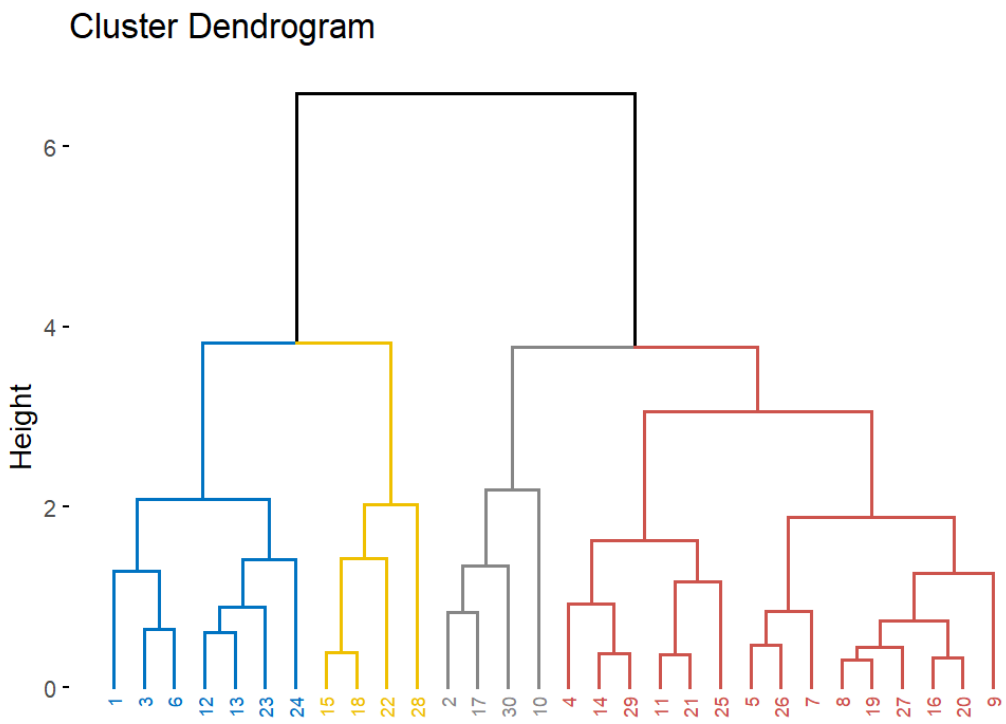


Plot 26: Dendrogram with Ward method

```
> cs

Call:
hclust(d = goster, method = "single")

Cluster method   : single
Distance         : euclidean
Number of objects: 15

> cc

Call:
hclust(d = goster, method = "complete")

Cluster method   : complete
Distance         : euclidean
Number of objects: 15

> ca

Call:
hclust(d = goster, method = "average")

Cluster method   : average
Distance         : euclidean
Number of objects: 15

> cw

Call:
hclust(d = goster, method = "ward.D2")

Cluster method   : ward.D2
Distance         : euclidean
Number of objects: 15
```

Table 44: Outputs of the dendrograms

Divisive Hierarchical Clustering



Plot 27: Cluster Dendrogram

K-means Clustering

For this analysis, some variables are selected and their first 6 rows can be seen below.

```
> head(cluster_data)
     Income Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds
5524  58138      58      635        88             546             172              88           88
2174  46344      38       11         1               6               2               1            6
4141  71613      26      426        49             127             111              21           42
6182  26646      26       11         4              20              10               3            5
5324  58293      94      173        43             118              46              27           15
7446  62513      16      520        42              98               0              42           14
     NumDealsPurchases NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth Age
5524                 3                8                  10                 4                 7  66
2174                 2                1                   1                 2                 5  69
4141                 1                8                   2                10                 4  58
6182                 2                2                   0                 4                 6  39
5324                 5                5                   3                 6                 5  42
7446                 2                6                   4                10                 6  56
```

Table 45: First 6 rows of the data

Then, variances of each predictor is computed.

```
> sapply(cluster_data, var)
            Income            Recency           MntWines           MntFruits      MntMeatProducts
   463382456.828568          838.079803       113801.905413         1584.202050        50315.675317
    MntFishProducts    MntSweetProducts        MntGoldProds   NumDealsPurchases     NumWebPurchases
       2998.747622         1687.368719         2685.588115            3.701083            7.515674
NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth                 Age
          8.568377           10.568830            5.884811          143.688158
```

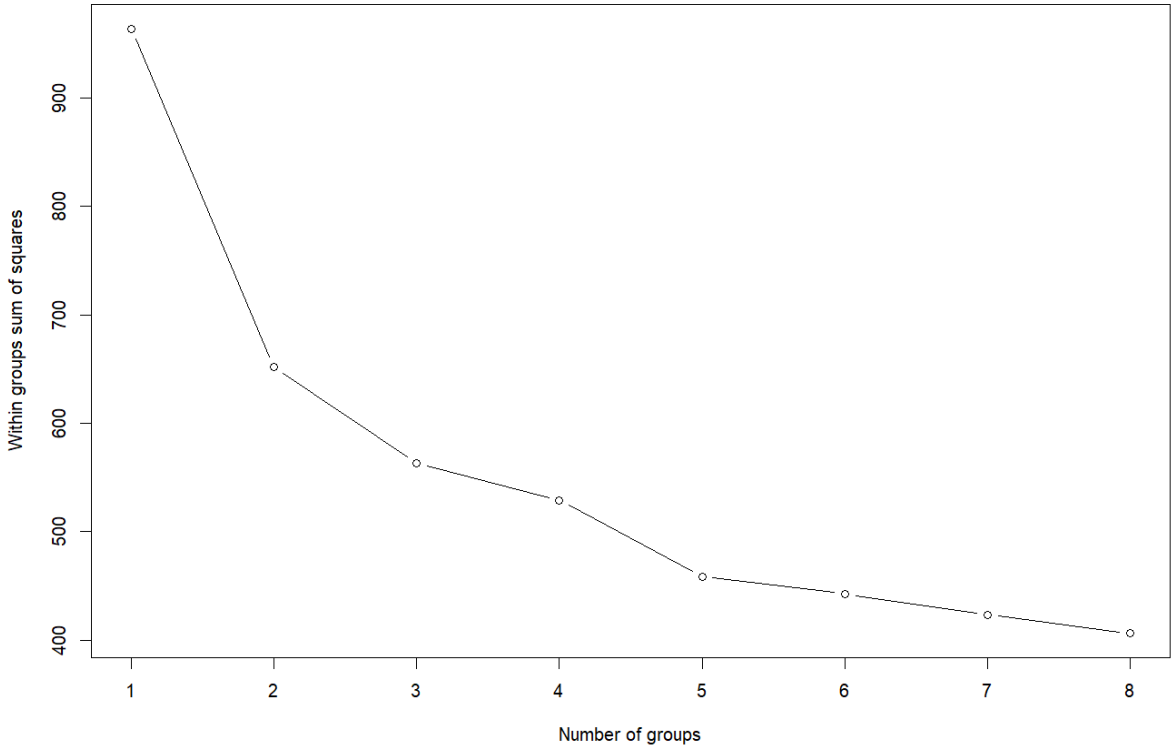
Table 46: Variances of the predictors

As it can be seen above, the variance values are very far from each other. Because of this, standardization on them must be done first.

```
> sapply(crime_s, var)
            Income            Recency           MntWines           MntFruits      MntMeatProducts
         0.01795090          0.08550962          0.05105402          0.04000409          0.01690928
    MntFishProducts    MntSweetProducts        MntGoldProds   NumDealsPurchases     NumWebPurchases
         0.04470338          0.02458145          0.02606330          0.01644926          0.01030957
NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth                 Age
         0.01092905          0.06253745          0.01471203          0.01354399
```

Table 47: Standardization of variables

After that, clustering can be applied.



Plot 28: Elbow plot

From the plot we can see that, there are 2 elbow which means 3 groups can be needed for K-means Clustering. The group means for 3 groups and the cluster number table can be shown below output.

```
> kmeans(crime_s, centers = 3)$centers * rge
      Income   Recency   MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds
1 36674.88682  97.27089  11.356171 0.7759108       0.3628672        3.332327         39.41484    20.387669
2    41.02554 916.08855 130.708705 3.6869243      14.9903790      201.619289         29.43175     2.883155
3   689.10603 118.59933   5.424843 5.7100678   40563.9863855       97.619288         84.77612     7.211675
  NumDealsPurchases NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth      Age
1          4.274099        9.981669           40.30632          68.78846          4.827492 6.617472
2          2.514573    35807.842843           31.62622         170.68289          6.286407 11.143902
3          2.019900       19.765340          340.83156         211.11309          4.287065 54.325871
```

Table 48: The group means for three groups

```
> head(final_clust,15)
5524 2174 4141 6182 5324 7446  965 6177 4855 5899  387 2125 8180 2569 2114
   1    3    1    3    2    2    2    3    3    3    3    1    3    3    2
```

Table 49: The cluster number for each customer

```
> table(final_clust)
final_clust
   1    2    3
 402  629 1184
```

Table 50: Table of clusters

Also, it can be obtained the cluster number for each customer and it can be seen that there are 402 customers in first cluster, 629 costumers in the second cluster and 1184 costumers in the third cluster.

3.8 CANONICAL CORRELATION ANALYSIS

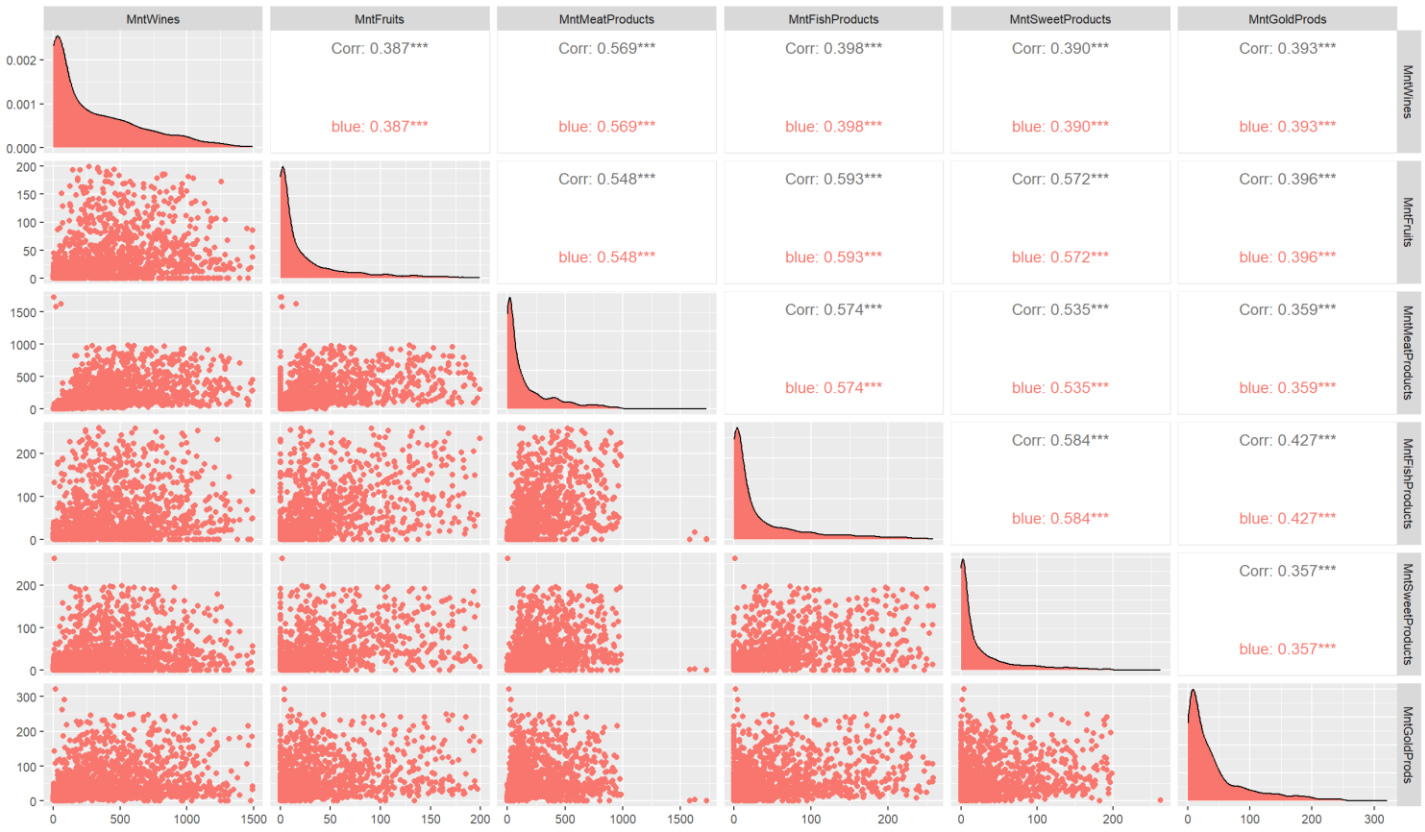
	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
1	635	88	546	172	88	88
2	11	1	6	2	1	6
3	426	49	127	111	21	42
4	11	4	20	10	3	5
5	173	43	118	46	27	15
6	520	42	98	0	42	14

	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
1	3	8	10	4	7
2	2	1	1	2	5
3	1	8	2	10	4
4	2	2	0	4	6
5	5	5	3	6	5
6	2	6	4	10	6

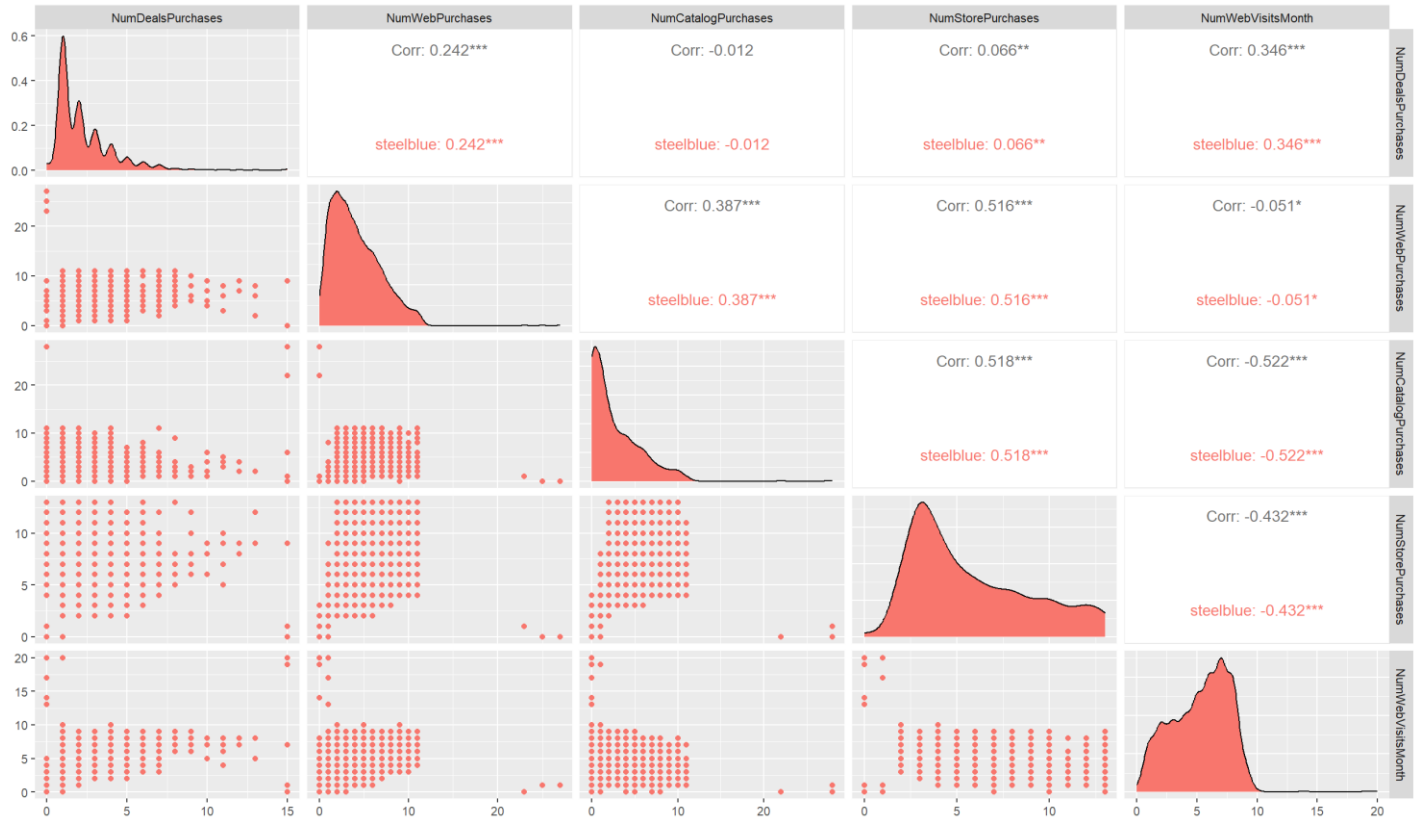
Table 51 : First 6 observations for CCN

From above table our data can be seen, so we can divide it into two sets which are Products and Place.

The observations starts with Mnt are into the Products set and with Num ones into the Place set.



Plot 29 : Pairs of plots for products set



Plot 30 : Pairs of plots for place set

	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts
MntWines	1.000000000	0.3869765	0.5687523	0.3976019	0.3901778
MntFruits	0.386976523	1.0000000	0.5477955	0.5934073	0.5715815
MntMeatProducts	0.568752276	0.5477955	1.0000000	0.5735067	0.5350478
MntFishProducts	0.397601940	0.5934073	0.5735067	1.0000000	0.5838036
MntSweetProducts	0.390177833	0.5715815	0.5350478	0.5838036	1.0000000
MntGoldProds	0.392588469	0.3964431	0.3593279	0.4270562	0.3573363
NumDealsPurchases	0.009234616	-0.1344159	-0.1210791	-0.1430620	-0.1212143
NumWebPurchases	0.553744921	0.3020006	0.3070130	0.2996211	0.3338660
NumCatalogPurchases	0.634683543	0.4862294	0.7340825	0.5326917	0.4950537
NumStorePurchases	0.639888931	0.4584570	0.4858769	0.4576414	0.4550961
NumWebVisitsMonth	-0.321928339	-0.4187061	-0.5394635	-0.4463921	-0.4223363

	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases
MntWines	0.39258847	0.009234616	0.55374492	0.63468354
MntFruits	0.39644311	-0.134415864	0.30200059	0.48622938
MntMeatProducts	0.35932787	-0.121079062	0.30701303	0.73408249
MntFishProducts	0.42705617	-0.143062024	0.29962114	0.53269166
MntSweetProducts	0.35733629	-0.121214264	0.33386598	0.49505374
MntGoldProds	1.00000000	0.052161082	0.40700473	0.44233975
NumDealsPurchases	0.05216108	1.000000000	0.24164617	-0.01189661
NumWebPurchases	0.40700473	0.241646174	1.00000000	0.38680770
NumCatalogPurchases	0.44233975	-0.011896609	0.38680770	1.00000000
NumStorePurchases	0.38903875	0.066468290	0.51619091	0.51774317
NumWebVisitsMonth	-0.24763802	0.346003374	-0.05117877	-0.52197909

	NumStorePurchases	NumWebVisitsMonth
MntWines	0.63988893	-0.32192834
MntFruits	0.45845705	-0.41870613
MntMeatProducts	0.48587690	-0.53946354
MntFishProducts	0.45764140	-0.44639213
MntSweetProducts	0.45509611	-0.42233626
MntGoldProds	0.38903875	-0.24763802
NumDealsPurchases	0.06646829	0.34600337
NumWebPurchases	0.51619091	-0.05117877
NumCatalogPurchases	0.51774317	-0.52197909
NumStorePurchases	1.00000000	-0.43236918
NumWebVisitsMonth	-0.43236918	1.00000000

Table 52 : correlations within and between the two sets of variables

$xcoef

	[,1]	[,2]	[,3]	[,4]	[,5]
MntWines	0.001439367	-0.002763766	-0.0002021915	-0.0019671832	-0.0004106794
MntFruits	0.001981194	-0.002052043	0.0162735227	-0.0019839408	0.0302355185
MntMeatProducts	0.001500597	0.005149613	-0.0031464845	-0.0009367272	0.0004099911
MntFishProducts	0.002130973	0.002136807	0.0090471246	-0.0032442659	-0.0120569523
MntSweetProducts	0.003151143	-0.002606799	0.0098646730	0.0156904632	-0.0153840676
MntGoldProds	0.002721474	-0.004261268	-0.0125022390	0.0164273701	0.0043342759

$ycoef

	[,1]	[,2]	[,3]	[,4]	[,5]
NumDealsPurchases	-0.06563624	0.01572002	-0.30098522	0.1927697	0.4562511
NumWebPurchases	0.08931099	-0.20757675	-0.01988135	0.3350855	-0.2171470
NumCatalogPurchases	0.20407159	0.17366476	-0.27613250	-0.2041229	-0.1216511
NumStorePurchases	0.09772063	-0.17364810	0.18968434	-0.2505006	0.1917927
NumWebVisitsMonth	-0.03564334	-0.23285985	-0.16226833	-0.4483348	-0.1895047

Table 53: raw canonical coefficients

For the variable NumDealsPurchases, a one unit increase in purchase leads to a .0656 decrease in the first canonical variate of set 2 when all of the other variables are held constant.

[1] 0.86609937 0.44116092 0.23019323 0.12756426 0.01385556

Table 54 : Cannonical correlation values

```
$corr.X.xscores
                      [,1]       [,2]       [,3]        [,4]        [,5]
MntWines         0.8597928 -0.3889258 -0.1182454 -0.29862624 -0.04143846
MntFruits        0.6502911  0.1109820  0.5034441  0.14959902  0.53624630
MntMeatProducts  0.8428228  0.5105675 -0.1216353 -0.08194738  0.03634135
MntFishProducts  0.6853892  0.2034913  0.4077704  0.13095589 -0.22146894
MntSweetProducts 0.6625905  0.0899002  0.4289098  0.42852435 -0.25412159
MntGoldProds     0.5799616 -0.1924641 -0.3151299  0.63841625  0.17258512

$corr.Y.xscores
                        [,1]       [,2]       [,3]        [,4]          [,5]
NumDealsPurchases   -0.07189891 -0.1527711 -0.15602704  0.02162249  0.0085975202
NumWebPurchases      0.53161530 -0.2768805 -0.03881521  0.05260944 -0.0024324801
NumCatalogPurchases  0.78224142  0.1281345 -0.06853777 -0.01282213 -0.0004714312
NumStorePurchases    0.67758533 -0.1538897  0.06946039 -0.01972694  0.0053893301
NumWebVisitsMonth   -0.51260247 -0.2411216 -0.10034485 -0.04366558 -0.0028993675

$corr.X.yscores
                      [,1]       [,2]       [,3]        [,4]          [,5]
MntWines         0.7446660 -0.17157886 -0.02721930 -0.03809403 -0.0005741530
MntFruits        0.5632167  0.04896093  0.11588942  0.01908349  0.0074299921
MntMeatProducts  0.7299683  0.22524244 -0.02799962 -0.01045356  0.0005035297
MntFishProducts  0.5936151  0.08977241  0.09386599  0.01670529 -0.0030685759
MntSweetProducts 0.5738692  0.03966045  0.09873213  0.05466439 -0.0035209966
MntGoldProds     0.5023043 -0.08490762 -0.07254078  0.08143909  0.0023912633

$corr.Y.yscores
                        [,1]       [,2]       [,3]       [,4]       [,5]
NumDealsPurchases   -0.08301462 -0.3462933 -0.6778090  0.1695028  0.62051053
NumWebPurchases      0.61380405 -0.6276179 -0.1686201  0.4124152 -0.17555987
NumCatalogPurchases  0.90317745  0.2904485 -0.2977402 -0.1005151 -0.03402469
NumStorePurchases    0.78234132 -0.3488290  0.3017482 -0.1546431  0.38896519
NumWebVisitsMonth   -0.59185179 -0.5465616 -0.4359157 -0.3423026 -0.20925663
```

Table 55 : Loadings of the variables on the canonical dimensions

From above table we can see the correlations between variables and the canonical variates.

```
Wilks' Lambda, using F-approximation (Rao's F):
              stat      approx df1      df2       p.value
1 to 5:  0.1874403 152.7842791  30 8818.000 0.00000000000
2 to 5:  0.7501458  33.1140841  20 7314.108 0.00000000000
3 to 5:  0.9314219  13.2376509  12 5836.819 0.00000000000
4 to 5:  0.9835385   6.1308828   6 4414.000 0.00000206078
5 to 5:  0.9998080   0.2119828   2 2208.000 0.80899510181
```

Table 56: test of canonical dimensions

As we can see from the table except dimension 5 all of the dimensions are statistically significant at 95% confidence level.

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	0.48556364	-0.9323437	-0.06820838	-0.66362018	-0.13854080
[2,]	0.07885554	-0.0816755	0.64771933	-0.07896488	1.20343519
[3,]	0.33660126	1.1551176	-0.70579284	-0.21011874	0.09196575
[4,]	0.11669382	0.1170133	0.49542798	-0.17765867	-0.66024862
[5,]	0.12944136	-0.1070810	0.40521703	0.64452647	-0.63194048
[6,]	0.14103401	-0.2208302	-0.64789928	0.85131002	0.22461371

Table 57 : Standardized producs canonical coefficients diagonal matrix of roducts sd'

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	-0.12627231	0.03024249	-0.5790414	0.3708543	0.8777450
[2,]	0.24484365	-0.56906606	-0.0545042	0.9186280	-0.5953026
[3,]	0.59735408	0.50834785	-0.8082893	-0.5975044	-0.3560947
[4,]	0.31768719	-0.56452540	0.6166588	-0.8143710	0.6235129
[5,]	-0.08646586	-0.56488610	-0.3936407	-1.0875989	-0.4597125

Table 58 : Standardized place canonical coefficients diagonal matrix of place sd's

The interpretation of standardized canonical coefficients is akin to the interpretation of standardized regression coefficients. For instance, in the case of the variable " NumDealsPurchases," a one standard deviation increase in reading corresponds to a 0.12 standard deviation decrease in the score on the first canonical variate for set 2, assuming all other variables in the model are held constant.

4. Discussion/Conclusion

This report presents comprehensive analyzes performed on the customer data set and highlights how the findings can contribute to the company's strategic decisions. Our analysis focuses on key topics such as customer segmentation, purchasing habits and marketing strategies. First, with Exploratory Data Analysis (EDA), the general characteristics of the data set were understood and potential differences between customer groups were determined. These differences have been evaluated statistically using Inferential Statistics, which increases the company's ability to identify target customer segments more effectively. Principal Components Analysis (PCA) and Factor Analysis have played an important role in identifying the key factors in customer behavior. This reveals the company's potential to better align its products and services to customer expectations. Customer segmentation has been achieved with Discrimination and Classification techniques, which makes it possible to create customized marketing strategies. The results provide valuable insights that allow the company to better understand its customer base and optimize marketing strategies. However, more data and long-term follow-up analyzes can be performed to determine the level of reliability of these results. Comparing the results obtained throughout the report with other companies in similar sectors and connecting them with relevant research in the literature can add more depth to future analyses. If more time and resources are allocated, more detailed analyzes can be made, especially in areas such as campaign interactions, customer satisfaction and brand loyalty. As a result, this analysis provides important insights that will help better understand the customer base and make more informed strategic decisions. This information will allow the company to increase its competitive advantage and achieve sustainable growth.

References

Kibar, Z. S. (2023). Multivariate Analysis. *Stat467 Recitation 10*.

Kibar, Z. S. (2023). Multivariate Analysis. *Stat467 Recitation 8*.

Kabacoff, R. (2023) *Quick-R: Cluster analysis*.

https://www.statmethods.net/advstats/cluster.html

Kabacoff, R. (2023) *Principal Components and Factor Analysis*

https://www.statmethods.net/advstats/factor.html

Data Source: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis