



# MH3511 Data Analysis with Computer Group Project

## Marketing Analytics

Name	Matriculation Number
Zhou Wei	U2022264K
Huang Jingfang	U2020128A
Kelly Wong Jie Yin	U2020126H
Cao Qingtian	U2020646L
Lin Jacky	U2022138E

### ***Abstract:***

With the rapid improvement in the convenience and speed to access customer data, it has been easier for companies to analyse their business performance, and how various conditions and factors affect the sales of their products. The success of the company is also highly dependent on how well the marketing is done. Therefore, it is important for the companies to understand their customers, and generate better marketing strategies by harnessing the power of data. Hence, our project aims to analyse how the characteristics of customers would affect their acceptance level to the marketing campaigns as well as their consumption habits, by using the various data visualisation tools and statistical analysis methods in R.

# Table of Contents

<b>1 Introduction</b>	<b>1</b>
<b>2 Data Description</b>	<b>1</b>
<b>3 Cleaning of Dataset and Exploratory Data Analysis</b>	<b>2</b>
3.1 Pre-processing of Data	2
3.2 Summary Statistics for the Main Variables of Interest	2
3.2.1 Customer Profile	2
3.2.2 Customer Behaviours	3
3.3 Summary Statistics for other Variables Related to Expenditures	4
3.3.1 Amount Spent on Wine	4
3.3.2 Amount Spent on Fruit	5
3.3.3 Amount Spent on Meat	5
3.3.4 Amount Spent on Fish	5
3.3.5 Amount Spent on Sweets	5
3.3.6 Amount Spent on Gold	6
3.4 Final Dataset for Analysis	6
<b>4 Statistical Analysis</b>	<b>6</b>
4.1 Correlations between Continuous Variables	6
4.2 Income vs Total_Expenditure	7
4.3 Total_Cmp_acceptance vs Income	7
4.4 AcceptedCmp (number of acceptance for each campaign) vs Income	8
4.5 Total_Cmp_acceptance vs Total_Expenditure	9
4.6 Income vs Educationlvl	10
<b>5 Linear Regression</b>	<b>11</b>
5.1 Single Linear Regression	11
5.2 Multiple Linear Regression	14
<b>6 Conclusion and Discussion</b>	<b>15</b>
<b>7 Appendix</b>	<b>16</b>
7.1 R Code	16
7.2 Random Forest Model	24
<b>8 References</b>	<b>25</b>

## 1 Introduction

With the rapid improvement in the convenience and speed to access customer data, it has been easier for companies to analyse their business performance, and how various conditions and factors affect the sales of their products. The success of the company is also highly dependent on how well the marketing is done. Therefore, it is important for the companies to understand their customers, and generate better marketing strategies by harnessing the power of data.

For this project, a dataset that consists of customer profiles and customer behaviours of one company is chosen. It includes demographic of customers such as education level, marital status and income, their spendings on different items such as fish, meat, gold etc., as well as the customers' acceptance level towards different marketing campaigns. Based on the information available, we aim to investigate the following questions:

- Are campaigns in general effective in increasing the number of products sold and which campaign is the most successful?
- How do customers' income affect their acceptance level towards marketing campaigns?

The analysis will be performed using the R language. Appropriate hypotheses were formulated and statistical analysis were performed in order to draw the conclusions for the aforementioned objectives.

## 2 Data Description

The dataset, titled "Marketing Analytics", is obtained from Kaggle, an online community of data scientists and machine learning practitioners. The original data consists of one csv. file titled "marketing\_data.csv", and it was used for hiring data analytics for iFood. The dataset consists of 27 variables and 2,240 records in total, that provides detailed information on customer profiles, customer expenditures, campaign successes/failures.

The variables are described below:

1. **ID** = Customer's unique identifier
2. **Year\_Birth** = Customer's birth year
3. **Education** = Customer's education level, which includes *Graduation, Basic, PhD, master* and *2nd Cycle*
4. **Marital\_Status** = Customer's marital status, which includes *single, divorced, widow, married, together, YOLO, alone* and *absurd*
5. **Income** = Customer's yearly income
6. **Kidhome** = Number of children in customer's household
7. **Teenhome** = Number of teenagers in customer's household
8. **Dt\_Customer** = Date of customer's enrollment with the company
9. **Recency** = Number of days since customer's last purchase
10. **MntWines** = Amount spent on wine in the last 2 years
11. **MntFruits** = Amount spent on fruits in the last 2 years
12. **MntMeatProducts** = Amount spent on meat in the last 2 years
13. **MntFishProducts** = Amount spent on fish in the last 2 years
14. **MntSweetProducts** = Amount spent on sweets in the last 2 years
15. **MntGoldProds** = Amount spent on gold in the last 2 years
16. **NumDealsPurchases** = Number of purchases made with a discount
17. **NumWebPurchases** = Number of purchases made through the company's website
18. **NumCatalogPurchases** = Number of purchases made using a catalogue
19. **NumStorePurchases** = Number of purchases made directly in stores
20. **NumWebVisitsMonth** = Number of visits to company's web site in the last month
21. **AcceptedCmp1** = 1 if customer accepted the offer in the 1st campaign, 0 otherwise
22. **AcceptedCmp2** = 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
23. **AcceptedCmp3** = 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

24. **AcceptedCmp4** = 1 if customer accepted the offer in the 4th campaign, 0 otherwise
25. **AcceptedCmp5** = 1 if customer accepted the offer in the 5th campaign, 0 otherwise
26. **Complain** = 1 if customer complained in the last 2 years, 0 otherwise
27. **Country** = Customer's location

### 3 Cleaning of Dataset and Exploratory Data Analysis

In this section, each variable is investigated individually to look for possible outliers, where a transformation will be performed on data that are highly skewed.

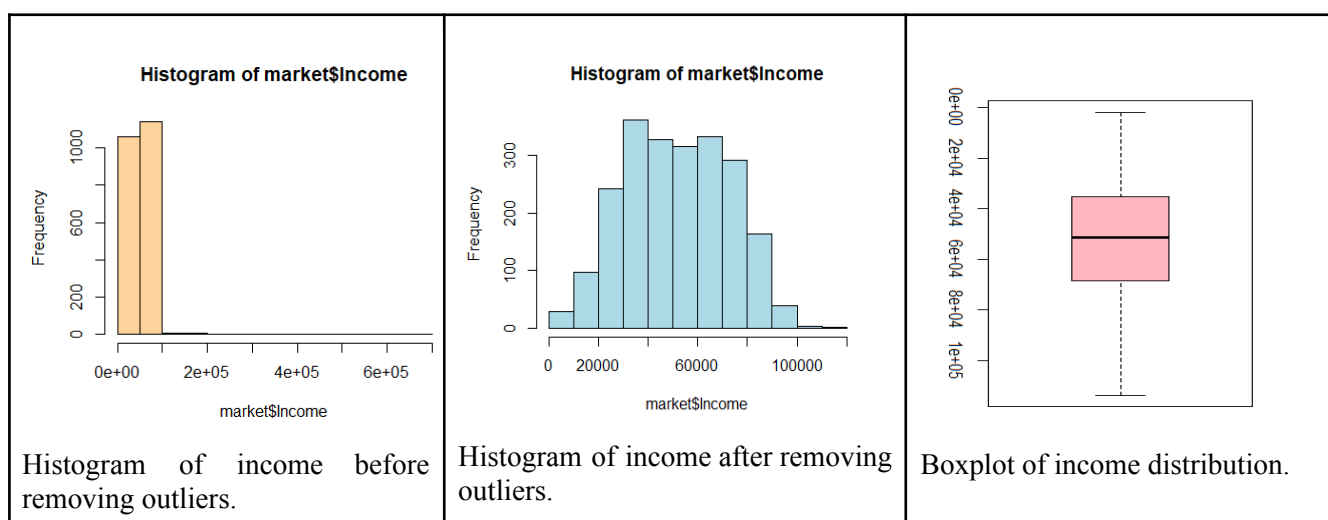
#### 3.1 Pre-processing of Data

To prepare the data for further analysis, a preliminary data cleaning was performed. Firstly, we select the main variables that we will be investigating later, such as income, age, and customers' expenses, and check if there is a need for data aggregation or transformation. These are the various data cleaning steps that we have performed:

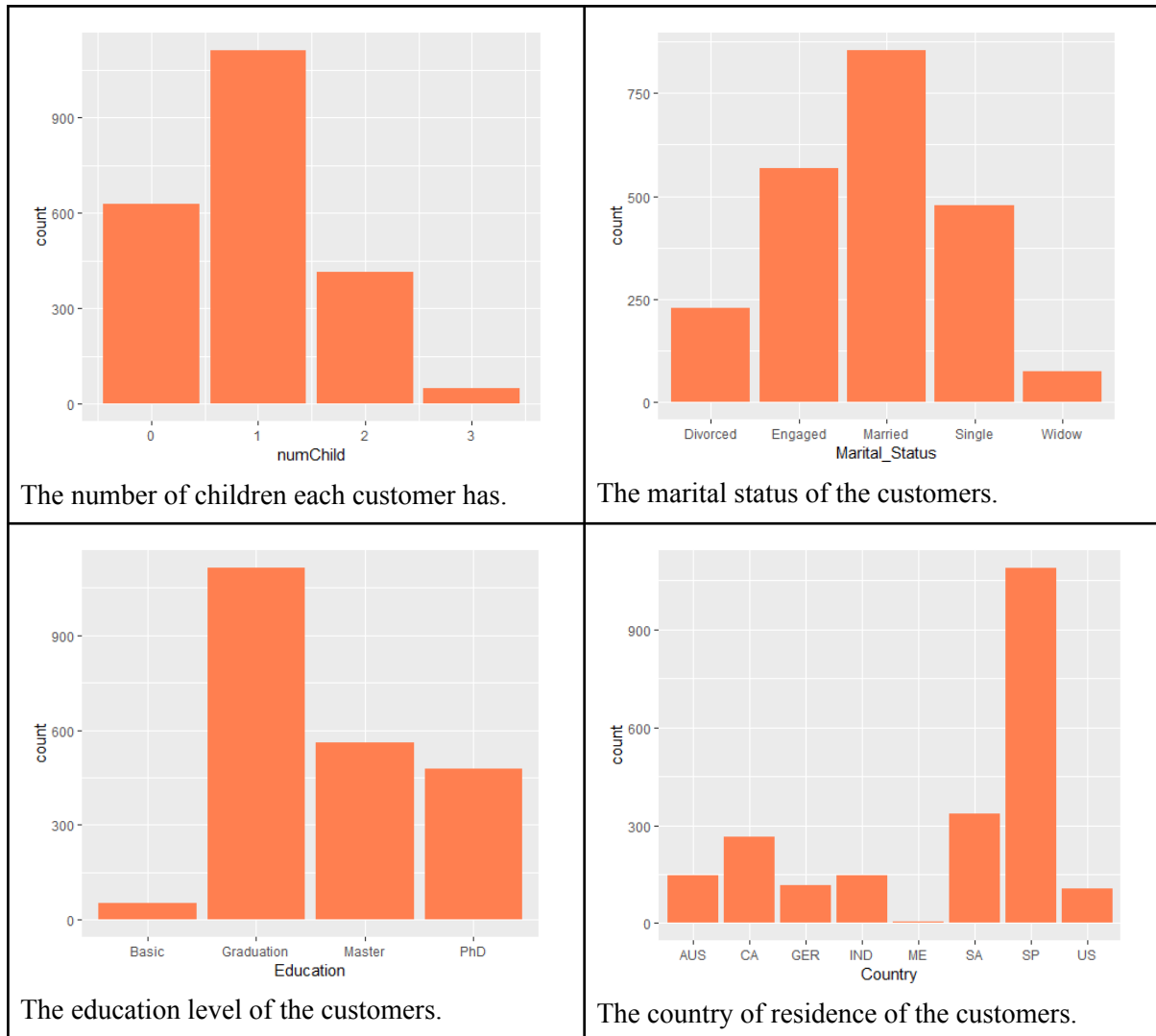
1. Transformed *Income* values to numerical and removed the rows with NULL values. A total of 24 rows with NULL values were removed from the dataset.
2. Removed outliers based on their income to avoid any bias for analysis.
3. Converted *Year\_Birth* values to *Age* by subtracting the value from the current year (2022).
4. Added a column *numChild* by summing up the total number of children in each household (summing the values from *Kidhome* and *Teenhome*).
5. For *education\_level*, we replaced *2nd cycle* with *Master* for better comprehension. Categories under *marital\_status* such as '*YOLO*', '*Alone*', and '*Absurd*' are renamed to '*Single*', and '*Together*' is renamed to '*Engaged*'.
6. *Education\_level*, *marital\_status* are encoded to discrete numerical values for analysis. Education\_level 1 to 4 are corresponding to '*Basic*', '*Graduation*', '*Master*', and '*PHD*' respectively. Marital\_status 1 to 5 are corresponding to '*Single*', '*Engaged*', '*Married*', '*Widow*', '*Divorced*' respectively.
7. Lastly, *Customer\_ID* was assigned to the remaining number of observations (starts from 1).

#### 3.2 Summary Statistics for the Main Variables of Interest

##### 3.2.1 Customer Profile



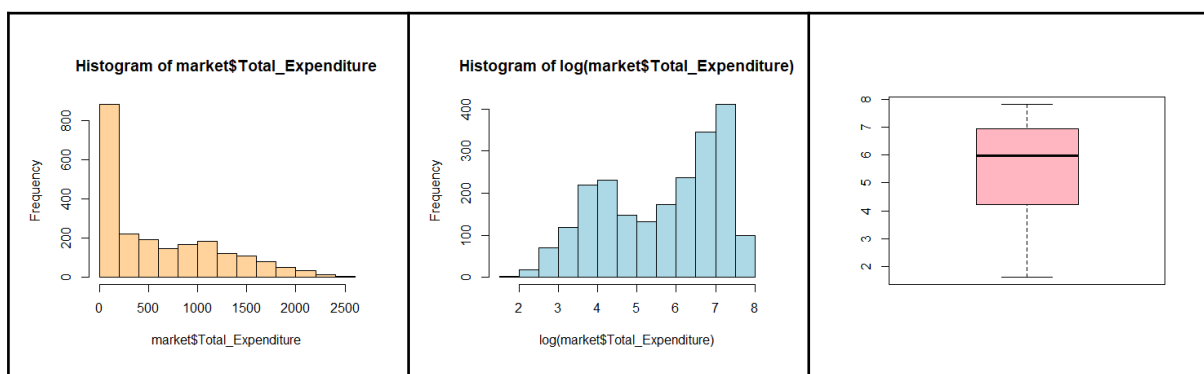
The income of customers after removing outliers appears to be normally distributed with mean = \$51,622.09 and standard deviation = \$20,713.06.

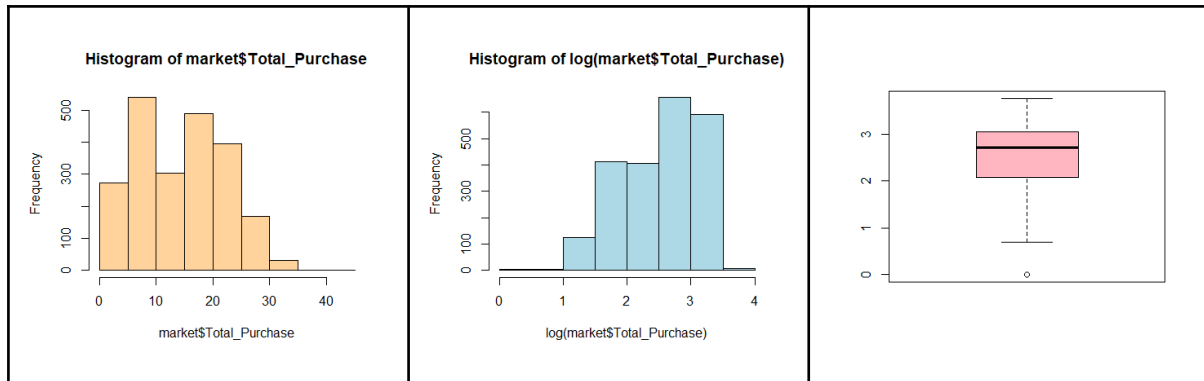


The majority of the customers are married and highly educated with either zero or one child. No outliers were removed from the above analysis.

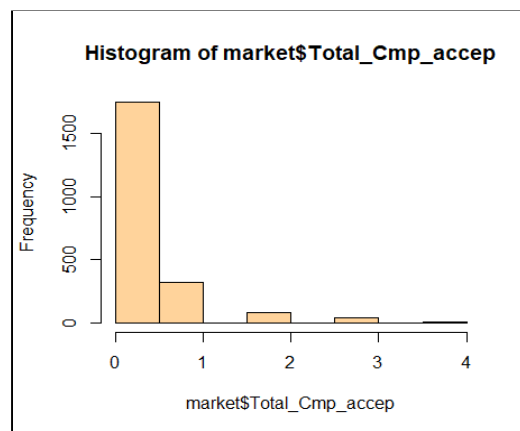
### 3.2.2 Customer Behaviours

To understand the overall customer behaviour, three columns “*Total\_Expenditure*”, “*Total\_Purchases*”, and “*Total\_Cmp\_acceptance*” are added to record the total expenditure on all the goods, total number of purchases made by each customer, and total number of campaigns each customer accepts respectively.





The distributions of total expenditure and total purchase are slightly right-skewed, so log transformation is applied. One outlier was observed and removed after transformation.

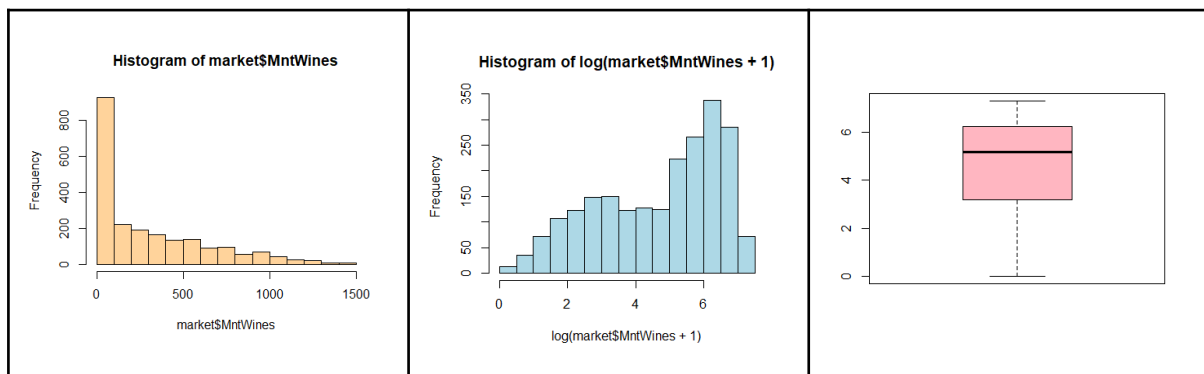


From the histogram, most of the customers accepted either zero or only one campaign. It shows that the marketing campaigns designed previously were largely unsuccessful.

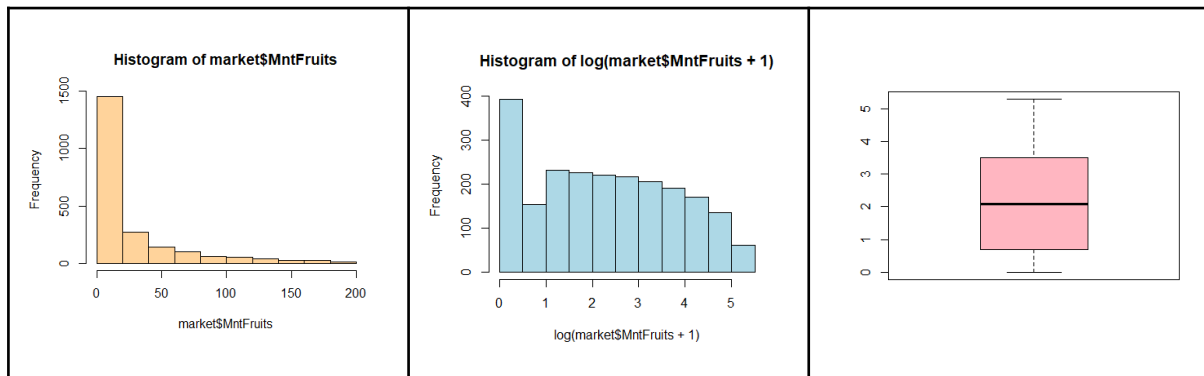
### 3.3 Summary Statistics for other Variables Related to Expenditures

After analysing the main variables, we went to explore each variable that contributes to the distribution of total expenditure (our main area of focus).

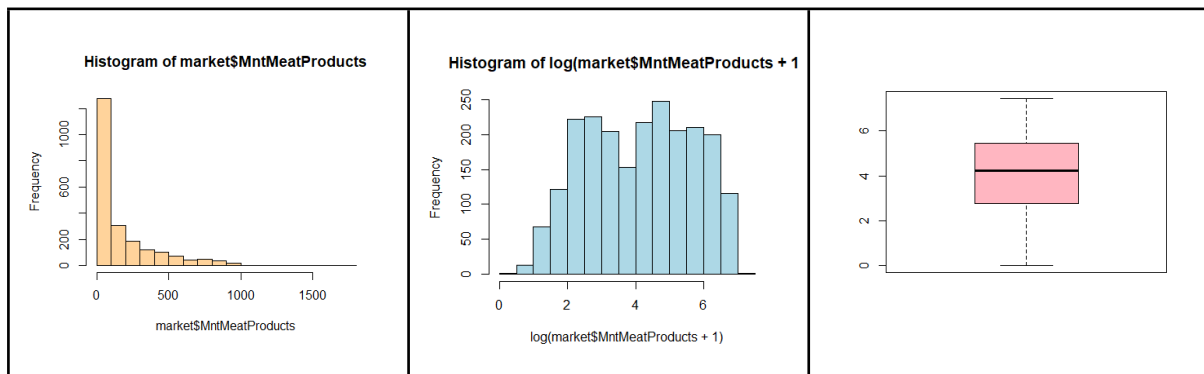
#### 3.3.1 Amount Spent on Wine



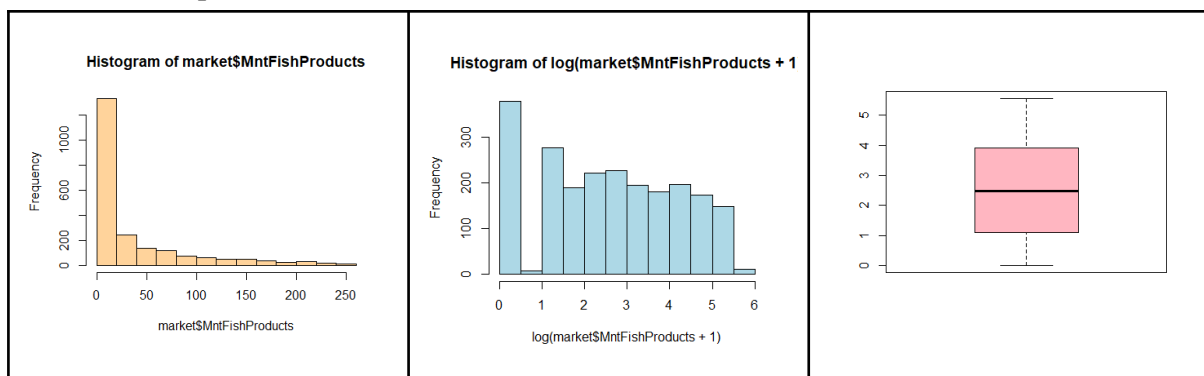
### 3.3.2 Amount Spent on Fruit



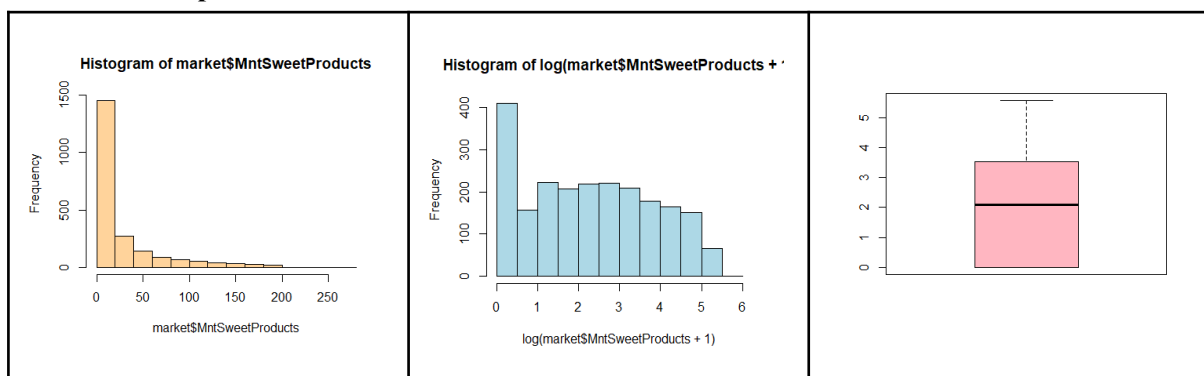
### 3.3.3 Amount Spent on Meat



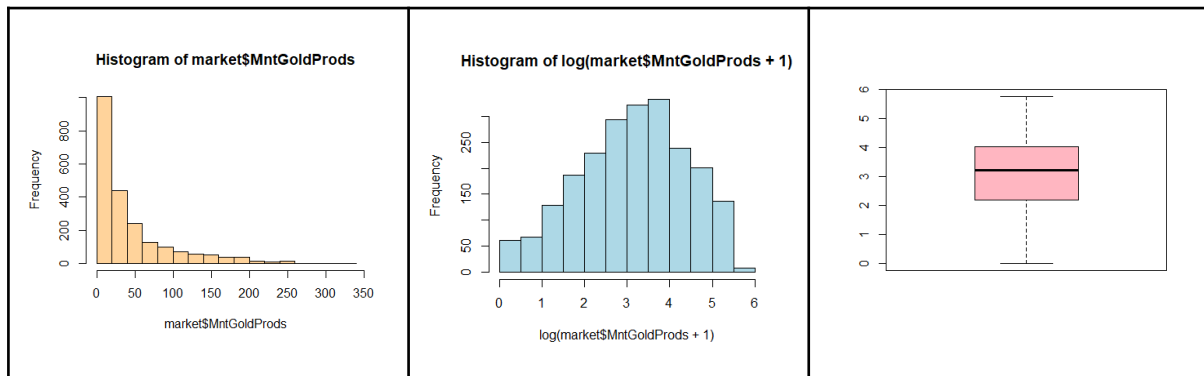
### 3.3.4 Amount Spent on Fish



### 3.3.5 Amount Spent on Sweets



### 3.3.6 Amount Spent on Gold



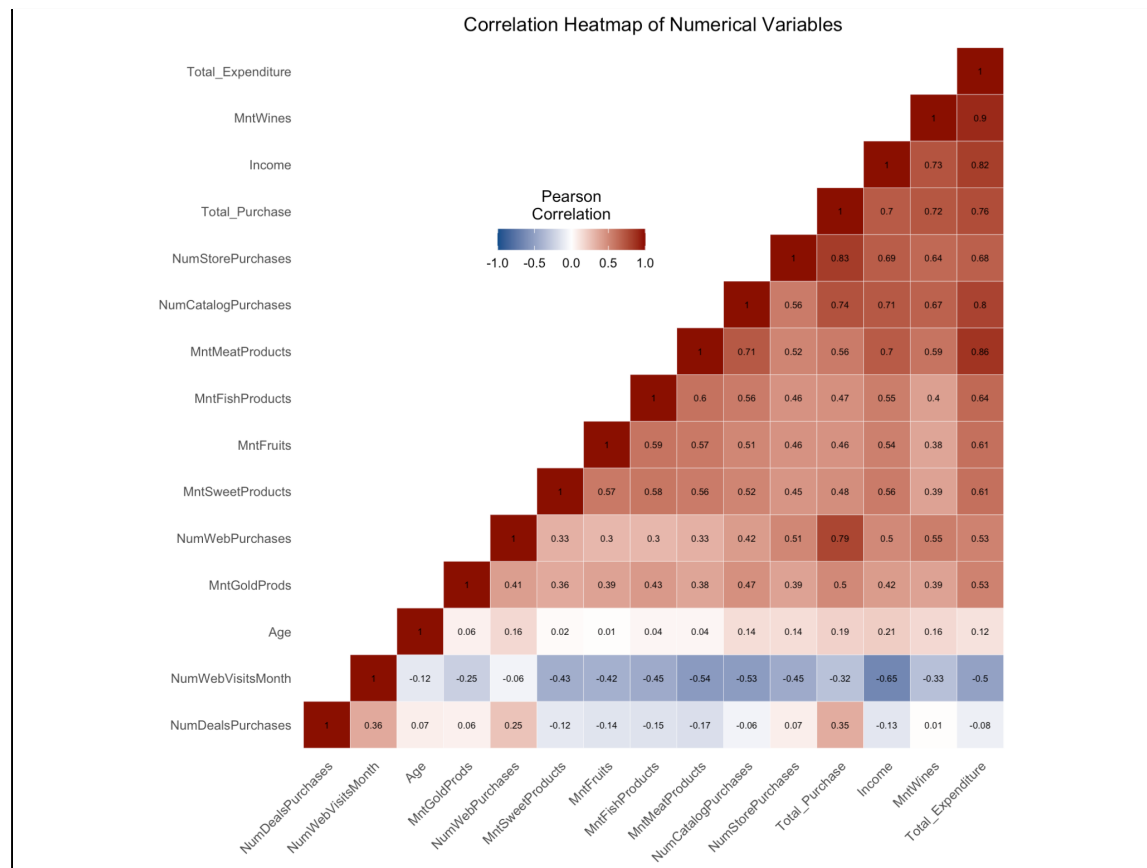
The histograms show that the variables are highly right-skewed, so we apply a log transformation to all the variables and generate the corresponding plots. Additionally, all the values are added by 1 before transformation to deal with 0 values. No outliers were observed from the boxplots.

### 3.4 Final Dataset for Analysis

Based on the above analysis, the dataset is further reduced to 2205 observations with the suggested transformations. Namely, log-transformation (base  $e$ ) to be applied to all the expenditures (columns that have “*Mnt*” as their sub-string).

## 4 Statistical Analysis

### 4.1 Correlations between Continuous Variables





The Correlation Heatmap is effective in determining the correlation between each pair of continuous variables. Possible linear relationships can also be identified from this plot.

A few interesting observations were drawn from the above plot:

- Total\_Purchases vs Total\_Expenditure ( $r=0.76$ )
- Income vs Total\_Purchases ( $r=0.7$ )
- Income vs Total\_Expenditure ( $r=0.82$ ) \*

Some other interesting observation explored:

- Total\_Cmp\_acceptance vs Income \*
- Total\_Cmp\_acceptance vs Total\_Expenditure \*
- Total\_Cmp\_acceptance vs Total\_Purchases
- Income vs Educationlvl

\*Only these observations will be explored in more details as they will provide the most insight to our analysis.

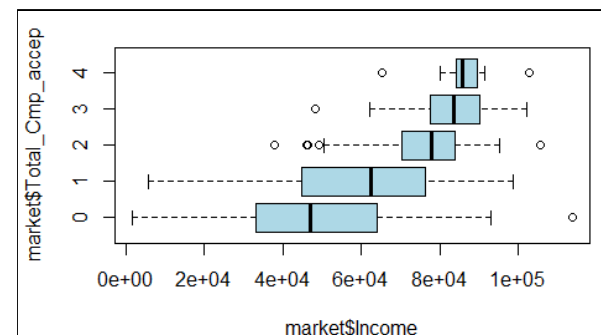
Alpha ( $\alpha$ ) = 0.05 will be used for all our statistical tests below.

## 4.2 Income vs Total\_Expenditure



There is a clear positive correlation between the income of the customers and their expenditure. Since the p-value ( $< 2e-16$ ) has a level of significance of less than 0.05, a statistically significant relationship between income and total expenditure can be highlighted. In addition,  $R^2 = 0.6784$  suggests a strong relationship between our model and the dependent variable, which supports the high correlation observed in the heat map ( $r = 0.82$ ). Hence, with a small p-value and relatively large  $R^2$ , it can be concluded that the two variables are highly correlated and are statistically significant.

## 4.3 Total\_Cmp\_acceptance vs Income



The sample sizes for these 5 categories are 1745, 322, 81, 41 and 11 respectively, and they are approximately normally distributed. Assuming normality, we apply t-test to determine if the total number of campaigns accepted increases with the average income.

Firstly, F-test was conducted to test if the variances are equal:

Total_Cmp_acceptance A	Total_Cmp_acceptance B	F-test p-value	T-test p-value	Result
0	1	0.3148	2.2e-16	Unequal
1	2	7.638e-05	3.502e-15	Unequal
2	3	0.07974	0.003361	Unequal
3	4	0.5839	0.3992	Equal

T-test:

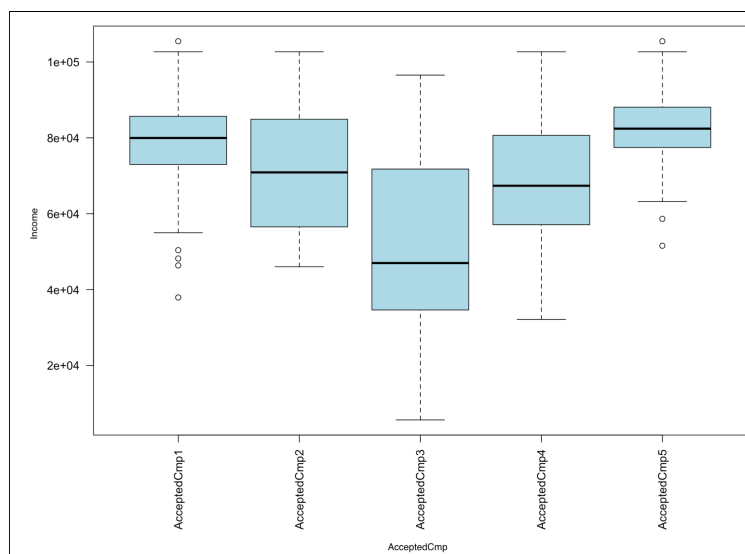
**Null hypothesis ( $H_0$ ):** The mean of Total\_Cmp\_acceptance B is smaller than or equal to that of Total\_Cmp\_acceptance A.

**Alternative hypothesis ( $H_1$ ):** The mean of Total\_Cmp\_acceptance B is greater than that of Total\_Cmp\_acceptance A.

Total_Cmp_acceptance A	Total_Cmp_acceptance B	p-value	Result
0	1	< 2.2e-16	Reject $H_0$ , infer it is greater
1	2	1.751e-15	Reject $H_0$ , infer it is greater
2	3	0.00168	Reject $H_0$ , infer it is greater

Hence, the number of campaigns accepted increases with the income of the customers.

#### 4.4 AcceptedCmp (number of acceptance for each campaign) vs Income



From the boxplot, we can observe that the data are approximately normally distributed and hence we can use t-test for our analysis. Starting from the leftmost boxplot, their average incomes are \$78,872.63, \$71,054.83, \$50,802.58, \$68,663.23 and \$82,345.50.

**Null hypothesis ( $H_0$ ):**

Campaign 5 has less than or equal average income as compared to Campaign X.

**Alternative hypothesis ( $H_1$ ):**

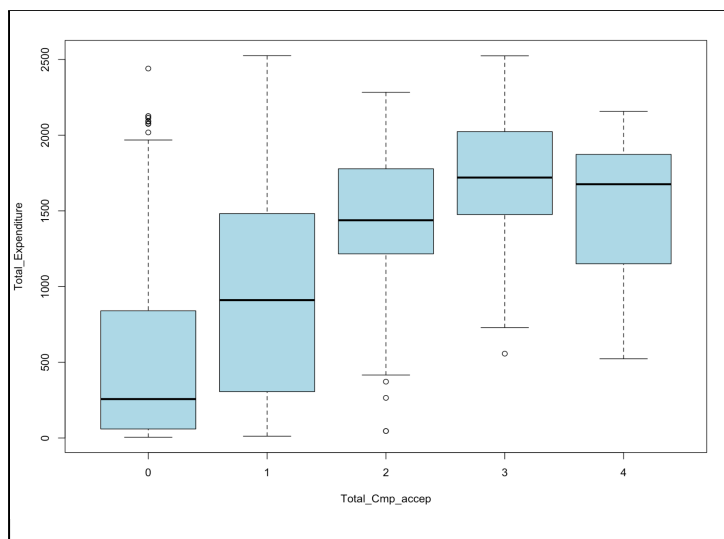
Campaign 5 has a greater average income than campaign X

Campaign X	F-test p-value	T-test p-value	Result
------------	----------------	----------------	--------

1	0.001941	0.001733	Reject $H_0$ , infer 5 is greater
2	1.316e-06	0.0003526	Reject $H_0$ , infer 5 is greater
3	< 2.2e-16	< 2.2e-16	Reject $H_0$ , infer 5 is greater
4	3.056e-12	< 2.2e-16	Reject $H_0$ , infer 5 is greater

From the table above, campaign 5 managed to attract the customers with the highest average income among all 5 campaigns at the significance level of 0.05.

#### 4.5 Total\_Cmp\_acceptance vs Total\_Expenditure



The plot shows that customers who have accepted 3 or 4 campaigns spent about the same amount on products on average.

##### Null Hypothesis ( $H_0$ ):

The average amount of money spent on products by customers who have accepted 3 or 4 campaigns is equal.

##### Alternative Hypothesis ( $H_1$ ):

The average amount of money spent on products by customers who have accepted 3 or 4 campaigns is not equal.

F-Test was conducted to check if the variances are equal:

Total_Cmp_acceptance A	Total_Cmp_acceptance B	F-test p-value	T-test p-value	Result
3	4	0.7103	0.2844	Do not reject $H_0$

There is insufficient evidence to reject  $H_0$  since  $p\text{-value} > 0.05$ . Further analysis will be done to investigate whether customers who accept one more campaign than the previous group spend more on average.

**Null hypothesis ( $H_0$ ):** Customers who accept B number of campaigns spend less than or equal to the customers who accept A number of campaigns on average.

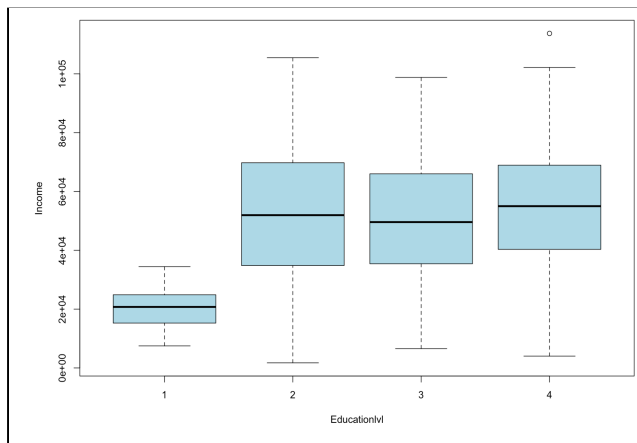
**Alternative hypothesis ( $H_1$ ):** Customers who accept B number of campaigns spend more than customers who accept A number of campaigns on average.

Total_Cmp_acceptance A	Total_Cmp_acceptance B	F-test p-value	T-test p-value	Result
0	1	1.668e-10	< 2.2e-16	Reject $H_0$ , infer 1 is greater

1	2	0.004875	3.346e-12	Reject $H_0$ , infer 2 is greater
2	3	0.7466	0.0008421	Reject $H_0$ , infer 3 is greater

Here we see that as the total number of campaign acceptance increases from 0 to 3, the average amount that customers spent increases.

#### 4.6 Income vs Educationlvl



The sample sizes for these 4 categories are 54, 1111, 562, 476 respectively, and they are approximately normally distributed. Hence, t-test can be applied to determine if there are customers with different education levels with the same average income.

**Null hypothesis ( $H_0$ ):** Customers with education level A and B have the same average income.

**Alternative hypothesis ( $H_1$ ):** Customers with education level A and B have different average income.

Education Level A	Education Level B	F-test p-value	T-test p-value	Result
2	3	0.6978	0.2562	Mean is roughly equal.
3	4	0.01258	0.0002611	Mean is not equal. Education Level 3: 50857.81 and education level 4: 55279.94

At this significance level, customers who have education level 2 appear to have the same mean income as customers who have education level 3. Further tests will be done to investigate whether an increase in education level increases the average income for the rest of the pairs:

**Null hypothesis ( $H_0$ ):** Customers with education level A and B have less or the same average income.

**Alternative hypothesis ( $H_1$ ):** Customers with education level B have a higher average income than customers with education level A.

Education Level A	Education Level B	F-test p-value	T-test p-value	Result
1	2	< 2.2e-16	< 2.2e-16	Reject $H_0$ , infer that Education level 2 has a greater average income than 1.
3	4	0.01258	0.000165	Reject $H_0$ , infer that Education Level 4 has an average income of 55279.94 is greater than the average income of education level 3 (50857.81).

At 95% significance level, it can be suggested that the higher the education level, the higher the average income. However there is an exception for education level 2 and 3.

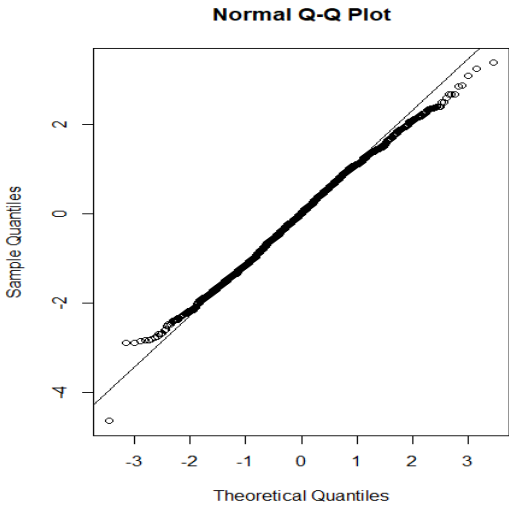
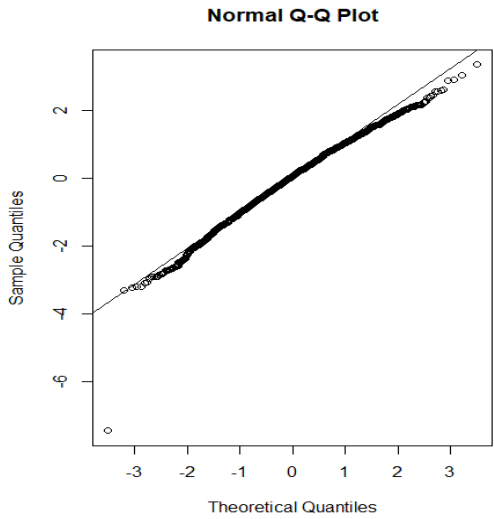
## 5 Linear Regression

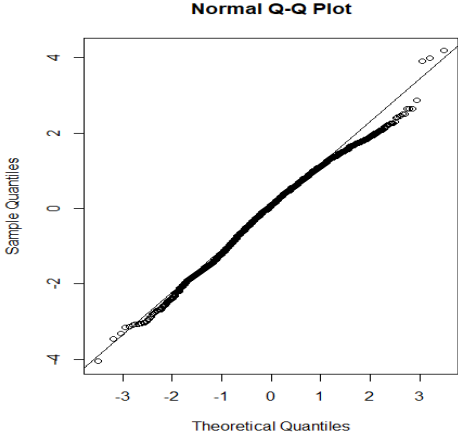
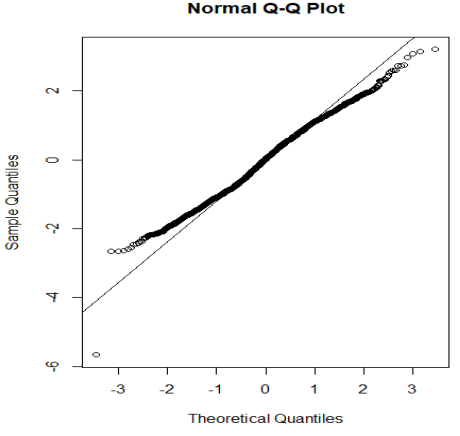
### 5.1 Single Linear Regression

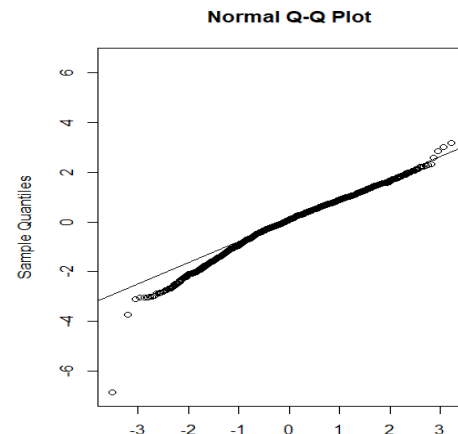
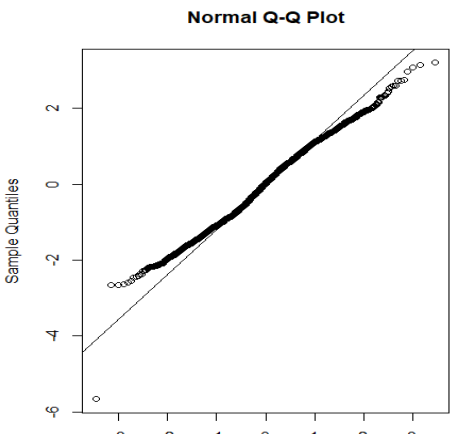
The relationship between Income and MntFruits, MntWine, MntGoldProds, MntFishProducts, MntMeatProducts, MntSweetProducts will be studied respectively. The correlations between these continuous variables are as follows: 0.539, 0.730, 0.702, 0.702, 0.552, 0.419, 0.556.

It can be observed that income is linearly related to the number of purchases of wines and meat products. However, after observing the plots of income vs amount spent on meat, it can be observed that there are many '0' values on the amount spent on meat.

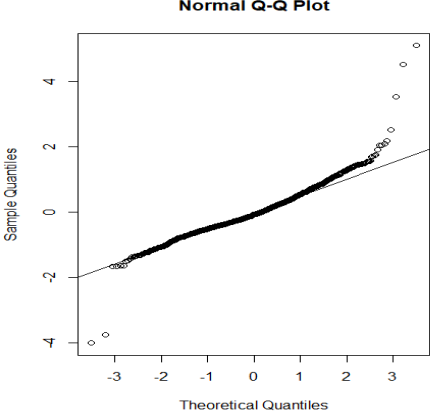
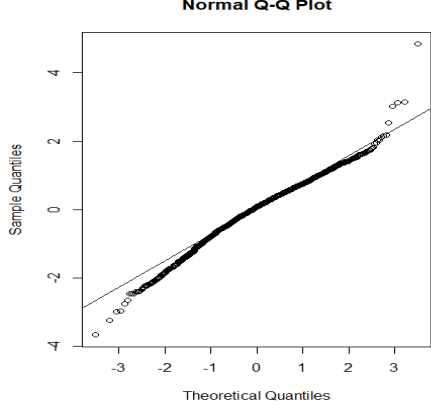
Hence, it can be hypothesised that income and other variables may have a linear relationship only above a certain threshold. Therefore, purchases with values '0' were removed. Linear regression was performed again, and the following results were obtained:  
(Y is log(variable), and X is Income)

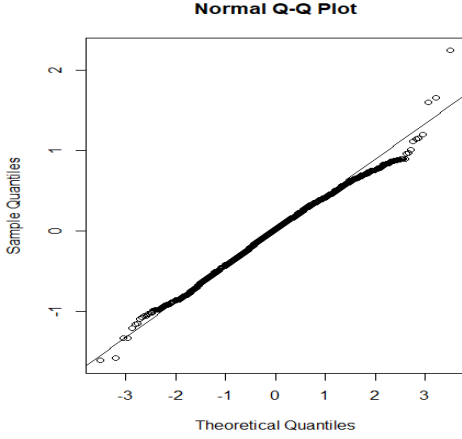
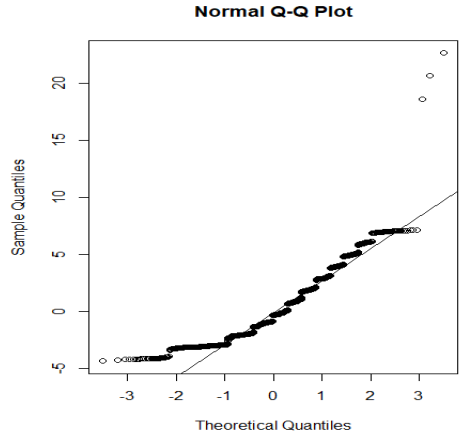
QQ-plot of residuals		QQ-plot of residuals	
			
Variable	MntFruits	Variable	MntWines
Formula	$\hat{Y}=4.551e-05x+1.562e-01$	Formula	$\hat{Y}=7.409e-05x+8.244e-01$
P-value	<2.2e-16	P-value	< 2.2e-16
R <sup>2</sup>	0.444	R <sup>2</sup>	0.6798

QQ-plot of residuals		QQ-plot of residuals	
			
Variable	MntGoldProds	Variable	MntFishProducts
Formula	$\hat{Y}=3.257e-05x+1.446e+00$	Formula	$\hat{Y}=4.451e-05x+6.042e-01$
P-value	$< 2.2e-16$	P-value	$< 2.2e-16$
R <sup>2</sup>	0.2711	R <sup>2</sup>	0.4557

QQ-plot of residuals		QQ-plot of residuals	
			
Variable	MntMeatProducts	Variable	MntSweetProducts
Formula	$\hat{Y}=6.280e-05x+8.386e-01$	Formula	$\hat{Y}=4.719e-05x+1.065e-01$
P-value	$< 2.2e-16$	P-value	$< 2.2e-16$
R <sup>2</sup>	0.6459	R <sup>2</sup>	0.4745

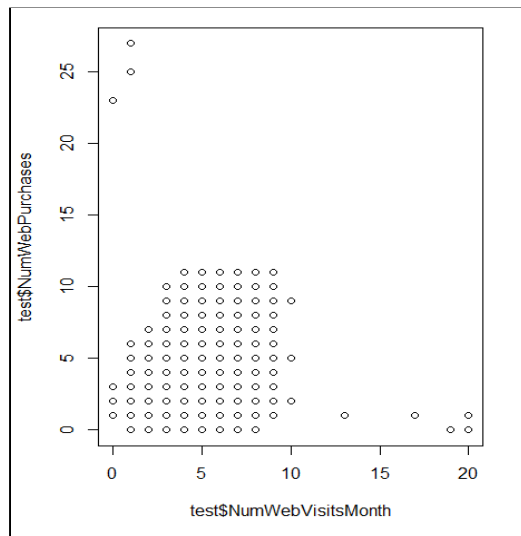
From the plots above, only log(meat) and log(wine) seem to be strongly linearly related to income. The remaining variables were then tested:

QQ-plot of residuals		QQ-plot of residuals	
			
Y	log(Total_Expenditure)	Y	log(Total_Expenditure)
X	log(TotalPurchases)	X	Income
Formula	$\hat{Y}=2.26745x-0.19622$	Formula	$\hat{Y}=5.993e-05x+2.469e+00$
P-value	$< 2.2e-16$	P-value	$< 2.2e-16$
R <sup>2</sup>	0.842	R <sup>2</sup>	0.6958

QQ-plot of residuals		QQ-plot of residuals	
			
Y	log(TotalPurchases)	Y	NumWebPurchases
X	Income	X	NumWebVisitsMonth
Formula	$\hat{Y}=2.070e-05x+1.472e+00$	Formula	$\hat{Y}=-0.06440x+4.44500$
P-value	$< 2.2e-16$	P-value	0.00792
R <sup>2</sup>	0.5067	R <sup>2</sup>	0.003195

The analysis shows that log(Total\_Expenditure) and log(Total\_Purchases), log(Total\_Expenditure) and Income have strong linear relationships, while surprisingly, the amount of web visits is not

linearly related to the number of web purchases. After plotting the graph, it is clear that they are indeed not linearly correlated.



Since the sample of having 0 web visits contains 6 records while the group with more than 0 web visits has 2200 points, there is a large difference between the two sample sizes. Hence, the result of comparing their average number of web purchases probably should not be taken into consideration.

## 5.2 Multiple Linear Regression

In this section, we attempt to build a multiple linear model for `Total_Cmp_accep` based on the 4 given performance measures, namely `Income`, `Age`, `NumWebVisitsMonth` and `numChild`. To avoid multicollinearity, `Total_Purchase` and `logTotal_Expenditure` were excluded due to having high correlation with `Income`. A backward elimination method was used to select the most appropriate model. The result is shown in the R output below.

We conclude that `Income`, `Age`, `NumWebVisitsMonth` and `numChild` are the most significant measures that could be used to model `Total_Cmp_accep`. The fitted model is:

$$Total\_Cmp\_accep = -4.748e^{-1} + 1.553e^{-5} - 3.516e^{-3} + 5.505e^{-2} - 1.422e^{-1}$$

```
> modelA = lm(Total_Cmp_accep~Income+Age+NumWebVisitsMonth+numChild, data = market)
> step(modelA, direction="backward")
Start: AIC=-2151.19
Total_Cmp_accep ~ Income + Age + NumWebVisitsMonth + numChild

      Df Sum of Sq  RSS   AIC
<none>             827.45 -2151.2
- Age              1    3.443  830.89 -2144.0
- numChild         1   19.715  847.17 -2101.3
- NumWebVisitsMonth 1   20.964  848.42 -2098.0
- Income           1  125.319  952.77 -1842.2

Call:
lm(formula = Total_Cmp_accep ~ Income + Age + NumWebVisitsMonth +
    numChild, data = market)

Coefficients:
(Intercept)      Income           Age  NumWebVisitsMonth
-4.748e-01    1.553e-05   -3.516e-03    5.505e-02
numChild
-1.422e-01
```



## 6 Conclusion and Discussion

Ifood is a Brazilian company that specialises in food delivery, and it has a rather high presence in over a thousand cities. In this report, we attempt to find out if campaigns are effective in increasing the products sold, which campaign is more effective and if customers' income affects their acceptance level towards each marketing campaign.

We conclude that:

- In general, customers with a higher average income tend to accept more campaigns, but the average income for customers who accept 3 and 4 campaigns is the same.
- Campaign 5 is the most successful in attracting high-income customers, as the average income for those who accepted it is the highest among all.
- In addition, customers who accepted more campaigns will spend more on products, but the average amount of money spent for customers who accepted 3 and 4 campaigns is the same.
- The higher the educational level, the higher the average income, except for customers who have a master's degree and graduation degree, who generally have the same average income.
- The single linear regression suggests that customers with higher income tend to buy more meat and wine. The log of the amount they buy is strongly linearly related to income.
- The multiple linear regression suggests that Income, Age, NumWebVisitsMonth and numChild are the significant measures that could be used to model Total\_Cmp\_accep.

Our analysis provided various insightful results, but it must be noted that it is only for a company based in Brazil, so some of the trends derived from the data may not be extended to the rest of the countries. Additionally, the data is dated back to around 2020, which is not recent and may not be representative of the current trends. Some key entries missing for more in-depth analysis as well. For instance, the customer's total expenditures and ifood's total profits for categories such as meat and fish are not provided. If these data are included, further investigation could be done to find out which category yields the most profits, and the company can focus their marketing strategies accordingly.

## 7 Appendix

### 7.1 R Code

```
#import data
market <- read.csv("marketing_data.csv",header = TRUE)
#view the data information
head(market)
length(market)

#Data Cleaning
#transform Income column to numerical
market$Income <- gsub('[/,$,]',",",market$Income)
#remove any null value
market$Income <- as.numeric(market$Income)
market <- market[!is.na(market$Income),]
#remove outliers based on the customer income
lowerl = quantile(market$Income, 0.25)-1.5*IQR(market$Income)
upperl = quantile(market$Income, 0.75)+1.5*IQR(market$Income)
market = market[!(market$Income > upperl|market$Income<lowerl),]
#convert the Year_Birth to Age
typeof(market$Year_Birth)
market$Year_Birth <- 2022 - market$Year_Birth
names(market)[2] <- "Age"
#remove those who are older than 120 and younger than 5
market <- market[(market$Age<120)&(market$Age>5),]
#Calculate the total number of children in each household
market$numChild <- market$Kidhome+market$Teenhome
#education - replace 2n cycle with Master as they are the same
market$Education[market$Education == '2n Cycle'] <- 'Master'
#encode education level to numerical values
market$Educationlvl <- market$Education
market$Educationlvl = factor(market$Education,
                             levels = c('Basic', 'Graduation', 'Master','PhD'),
                             labels = c(1,2,3,4))
market$Educationlvl <- as.numeric(market$Educationlvl)
#Marital_Status - re-categorize
market$Marital_Status[market$Marital_Status == 'Together'] <- 'Engaged'
market$Marital_Status[market$Marital_Status == 'YOLO'] <- 'Single'
market$Marital_Status[market$Marital_Status == 'Alone'] <- 'Single'
market$Marital_Status[market$Marital_Status == 'Absurd'] <- 'Single'
#encode Marital_Status to numerical values
market$Marital_encoded = factor(market$Marital_Status,
                                levels = c('Single', 'Engaged', 'Married', 'Widow','Divorced'),
                                labels = c(1,2,3,4,5))
market$Marital_encoded <- as.numeric(market$Marital_encoded)
#rename the first column name to Customer_ID and assign IDs to all the customers
names(market)[1] <- "Customer_ID"
market$Customer_ID <- c(1:nrow(market))

#EDA
#A. Summary Statistics for customer profile
#Check the distribution of the main variable - income
hist(market$Income,col='burlywood1')
boxplot(market$Income,col='lightpink',horizontal = TRUE)
```

```

mean(market$Income)
sd(market$Income)
#Check the distribution of the main variable - Age
hist(market$Age,col='burlywood1')
boxplot(market$Age,col='lightpink')
#Count plot for other categorical variables
ggplot(market,aes(numChild))+geom_bar(fill='coral')
ggplot(market,aes(Education))+geom_bar(fill='coral')
ggplot(market,aes(Country))+geom_bar(fill='coral')
ggplot(market,aes(Marital_Status))+geom_bar(fill='coral')
#B.Summary statistics for customer behaviours
expense <- market[,grepl("Mnt", colnames(market))]
#Distribution of amount spent on Wines
hist(market$MntWines,col='burlywood1')
hist(log(market$MntWines+1),col='lightblue')
boxplot(log(market$MntWines+1),col = 'lightpink')
#Distribution of amount spent on Fruits
hist(market$MntFruits,col='burlywood1')
hist(log(market$MntFruits+1),col='lightblue')
boxplot(log(market$MntFruits+1),col = 'lightpink')
#Distribution of amount spent on Meat
hist(market$MntMeatProducts,col='burlywood1')
hist(log(market$MntMeatProducts+1),col='lightblue')
boxplot(log(market$MntMeatProducts+1),col = 'lightpink')
#Distribution of amount spent on Fish
hist(market$MntFishProducts,col='burlywood1')
hist(log(market$MntFishProducts+1),col='lightblue')
boxplot(log(market$MntFishProducts+1),col = 'lightpink')
#Distribution of amount spent on Sweet
hist(market$MntSweetProducts,col='burlywood1')
hist(log(market$MntSweetProducts+1),col='lightblue')
boxplot(log(market$MntSweetProducts+1),col = 'lightpink')
#Distribution of amount spent on Gold
hist(market$MntGoldProds,col='burlywood1')
hist(log(market$MntGoldProds+1),col='lightblue')
boxplot(log(market$MntGoldProds+1),col = 'lightpink')
#Distribution of Total Expenditure
market$Total_Expenditure <- rowSums(market[,c(10:15)])
hist(market$Total_Expenditure,col='burlywood1')
hist(log(market$Total_Expenditure),col='lightblue')
boxplot(log(market$Total_Expenditure),col = 'lightpink')
#Distribution of Total number of purchases
purchase <- market[,grepl("Purchases", colnames(market))]
market$Total_Purchase <- rowSums(market[,c(16:19)])
hist(market$Total_Purchase,col='burlywood1')
hist(log(market$Total_Purchase),col='lightblue')
boxplot(log(market$Total_Purchase),col = 'lightpink')
#Distribution of Total Campaign acceptance
market$Total_Cmp_accep <- rowSums(market[,c(21:25)])
hist(market$Total_Cmp_accep,col='burlywood1')

#Heatmap
install.packages("corrplot")
install.packages("PerformanceAnalytics")
install.packages("tidyverse")

```

```

install.packages("reshape2")
library("PerformanceAnalytics")
library(corrplot)
library(tidyverse)
library(reshape2)
numeric_marketing_data <- market %>%
  select_if(.,is.numeric) %>%
  select(-c("AcceptedCmp1",
            "AcceptedCmp2",
            "AcceptedCmp3",
            "AcceptedCmp4",
            "AcceptedCmp5",
            "Complain",
            "Recency",
            "Customer_ID",
            "Educationlvl",
            "Marital_encoded",
            "numChild",
            "Teenhome",
            "Kidhome",
            "Response",
            "Total_Cmp_accep"),)
# Create correlation matrix
cormat <- round(cor(numeric_marketing_data),2)
# Function to get lower triangle of the correlation matrix
get_lower_tri<-function(cormat){
  cormat[upper.tri(cormat)] <- NA
  return(cormat)
}
# Function to get upper triangle of the correlation matrix
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}
# Function to reorder correlation matrix
reorder_cormat <- function(cormat){
  # Use correlation between variables as distance
  dd <- as.dist((1-cormat)/2)
  hc <- hclust(dd)
  cormat <-cormat[hc$order, hc$order]
}
# Reorder the correlation matrix
cormat <- reorder_cormat(cormat)
upper_tri <- get_upper_tri(cormat)
# Melt the correlation matrix
melted_cormat <- melt(upper_tri, na.rm = TRUE)
# Create a heatmap
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "dodgerblue4", high = "red4", mid = "white",
                      midpoint = 0, limit = c(-1,1), space = "Lab",
                      name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                    size = 9, hjust = 1))+

```

```

coord_fixed()
# Plot gg heatmap with values on graph as text
ggheatmap +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 2.20) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, -0.5),
    legend.position = c(0.5, 0.6),
    legend.direction = "horizontal") +
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
    title.position = "top", title.hjust = 0.5)) +
  labs(title = "Correlation Heatmap of Numerical Variables") +
  theme(plot.title = element_text(hjust = 0.5))

```

### **#Test for Income vs Total Expenditure**

```

attach(market)
plot(Income, Total_Expenditure, col="darkseagreen3")
abline(lm(Total_Expenditure ~ Income), col="blue") # regression line (y~x)
detach(market)
summary(lm(market$Total_Expenditure ~ market$Income))

```

### **#Test for total acceptance vs income**

```

market$Total_Cmp_acceptance=market$AcceptedCmp1+market$AcceptedCmp2+market$AcceptedCmp3+market$AcceptedCmp4+market$AcceptedCmp5

```

#### **#Boxplot**

```

boxplot(market$Income~market$Total_Cmp_accep, horizontal=TRUE, col='lightblue')

```

#### **#Test total acceptance = 1 vs 0**

```

var.test(market$Income[market$Total_Cmp_acceptance==1], market$Income[market$Total_Cmp_acceptance==0])

```

```

t.test(market$Income[market$Total_Cmp_acceptance==1], market$Income[market$Total_Cmp_acceptance==0], var.equal=T)

```

#### **#Test total acceptance = 2 vs 1**

```

var.test(market$Income[market$Total_Cmp_acceptance==2], market$Income[market$Total_Cmp_acceptance==1])

```

```

t.test(market$Income[market$Total_Cmp_acceptance==2], market$Income[market$Total_Cmp_acceptance==1], var.equal=F)

```

#### **#Test total acceptance = 3 vs 2**

```

var.test(market$Income[market$Total_Cmp_acceptance==4], market$Income[market$Total_Cmp_acceptance==2])

```

```

t.test(market$Income[market$Total_Cmp_acceptance==3], market$Income[market$Total_Cmp_acceptance==2], var.equal=T)

```

#### **#Test total acceptance = 4 vs 3**

```

var.test(market$Income[market$Total_Cmp_acceptance==4], market$Income[market$Total_Cmp_acceptance==3])

```

```

t.test(market$Income[market$Total_Cmp_acceptance==4], market$Income[market$Total_Cmp_acceptance==3], var.equal=T)

```

#### **#Test total acceptance = 1 vs 0**

```

t.test(market$Income[market$Total_Cmp_acceptance==1], market$Income[market$Total_Cmp_acceptance==0], var.equal=T, alternative='greater')

```

**#Test total acceptance = 2 vs 1**

```
t.test(market$Income[market$Total_Cmp_acceptance==2],market$Income[market$Total_Cmp_acceptance==1],var.equal=F,alternative='greater')
```

**#Test total acceptance = 3 vs 2**

```
t.test(market$Income[market$Total_Cmp_acceptance==3],market$Income[market$Total_Cmp_acceptance==2],var.equal=T,alternative='greater')
```

**#Test which campaign has the greatest average income**

**#Boxplot**

```
AcceptedCmp1 <- market$Income[market$AcceptedCmp1==1]
AcceptedCmp2 <- market$Income[market$AcceptedCmp2==1]
AcceptedCmp3 <- market$Income[market$AcceptedCmp3==1]
AcceptedCmp4 <- market$Income[market$AcceptedCmp4==1]
AcceptedCmp5 <- market$Income[market$AcceptedCmp5==1]
par(mar = c(12,6,1,1))
boxplot(AcceptedCmp1,AcceptedCmp2,AcceptedCmp3,AcceptedCmp4,AcceptedCmp5,beside=TRUE,
        names=c('AcceptedCmp1','AcceptedCmp2','AcceptedCmp3','AcceptedCmp4','AcceptedCmp5'),
        las=2,col="lightblue")
```

**#Test campaign 1 vs 5**

```
var.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp1==1])
t.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp1==1],var.equal=F,alternative='greater')
```

**#Test campaign 2 vs 5**

```
var.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp2==1])
t.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp2==1],var.equal=F,alternative='greater')
```

**#Test campaign 3 vs 5**

```
var.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp3==1])
t.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp3==1],var.equal=F,alternative='greater')
```

**#Test campaign 4 vs 5**

```
var.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp4==1])
t.test(market$Income[market$AcceptedCmp5==1],market$Income[market$AcceptedCmp4==1],var.equal=F,alternative='greater')
```

**#Test total acceptance vs total expenditure**

**#Boxplot**

```
attach(market)
boxplot(Total_Expenditure~Total_Cmp_accept,col='lightblue')
detach(market)
```

**#Test total acceptance 0 vs 1**

```
var.test(market$MntSum[market$Total_Cmp_acceptance==1],market$MntSum[market$Total_Cmp_acceptance==0])
t.test(market$Total_Expenditure[market$Total_Cmp_acceptance==1],market$Total_Expenditure[market$Total_Cmp_acceptance==0],var.equal=F,alternative='greater')
```

**#Test total acceptance 1 vs 2**

```
var.test(market$MntSum[market$Total_Cmp_acceptance==1],market$MntSum[market$Total_Cmp_acceptance==2])
t.test(market$Total_Expenditure[market$Total_Cmp_acceptance==2],market$Total_Expenditure[market$Total_Cmp_acceptance==1],var.equal=F,alternative='greater')
```

**#Test total acceptance 3 vs 2**

```
var.test(market$MntSum[market$Total_Cmp_acceptance==3],market$MntSum[market$Total_Cmp_acceptance==2])
```

```
t.test(market$Total_Expenditure[market$Total_Cmp_acceptance==3],market$Total_Expenditure[mar
ket$Total_Cmp_acceptance==2],var.equal=T,alternative='greater')
#Test total acceptance 4 vs 3
var.test(market$MntSum[market$Total_Cmp_acceptance==3],market$MntSum[market$Total_Cmp_a
cceptance==4])
t.test(market$Total_Expenditure[market$Total_Cmp_acceptance==3],market$Total_Expenditure[mar
ket$Total_Cmp_acceptance==4],var.equal=T)
```

### #Income and educational level

#### #Boxplot

```
attach(market)
boxplot(Income~Educationlvl,col='lightblue')
detach(market)
```

#### #Test educational level 2 and 3

```
var.test(market$Income[market$Educationlvl==2],market$Income[market$Educationlvl==3])
t.test(market$Income[market$Educationlvl==2],market$Income[market$Educationlvl==3],var.equal=
T)
```

#### #Test educational level 4 and 3

```
var.test(market$Income[market$Educationlvl==4],market$Income[market$Educationlvl==3])
t.test(market$Income[market$Educationlvl==4],market$Income[market$Educationlvl==3],var.equal=
F)
```

#### #Test educational level 1 vs 2

```
var.test(market$Income[market$Educationlvl==2],market$Income[market$Educationlvl==1])
t.test(market$Income[market$Educationlvl==2],market$Income[market$Educationlvl==1],var.equal=
F,alternative='greater')
```

#### #Test educational level 3 vs 4

```
t.test(market$Income[market$Educationlvl==4],market$Income[market$Educationlvl==3],var.equal=
F,alternative='greater')
```

### #Single Linear Regression

#### #create a new data frame consisting of these 6 continuous variables

```
market_lr=data.frame(market$MntWines,market$MntFishProducts,market$MntFruits,market$MntSw
eetProducts,market$MntGoldProds,market$Income)
colnames(market_lr)=c("MntWines","MntFishProducts","MntFruits","MntSweetProducts","MntGold
Prods","Income")
```

#### #test for wine

```
market_lr_cp=market_lr[market_lr$MntWines!=0,]
market_lr_cp$logMntWines=log(market_lr_cp$MntWines)
model=lm(market_lr_cp$logMntWines ~ market_lr_cp$Income)
res=resid(model)
qqnorm(res)
qqline(res)
```

#### #test for fish

```
market_lr_cp=market_lr[market_lr$MntFishProducts!=0,]
market_lr_cp$logMntFishProducts=log(market_lr_cp$MntFishProducts)
model=lm(market_lr_cp$logMntFishProducts ~ market_lr_cp$Income)
res=resid(model)
qqnorm(res)
qqline(res)
```

#### #test for fruits

```
market_lr_cp=market_lr[market_lr$MntFruits!=0,]
market_lr_cp$logMntFruits=log(market_lr_cp$MntFruits)
model=lm(market_lr_cp$logMntFruits ~ market_lr_cp$Income)
res=resid(model)
qqnorm(res)
```

```

qqline(res)
#test for sweets
market_lr_cp=market_lr[market_lr$MntSweetProducts!=0,]
market_lr_cp$logMntSweetProducts=log(market_lr_cp$MntSweetProducts)
model=lm(market_lr_cp$logMntSweetProducts ~ market_lr_cp$Income)
res=resid(model)
qqnorm(res)
qqline(res)
#test for gold
market_lr_cp=market_lr[market_lr$MntGoldProds!=0,]
market_lr_cp$logMntGoldProds=log(market_lr_cp$MntGoldProds)
model=lm(market_lr_cp$logMntGoldProds ~ market_lr_cp$Income)
res=resid(model)
qqnorm(res)
qqline(res)
#test for log(TotalPurchases) vs Income
market$logTotalPurchases=log(market$TotalPurchases)
model=lm(market$logTotalPurchases ~ market$Income)
res=resid(model)
qqnorm(res)
qqline(res)
#test for log(Total_Expenditure) vs Income
market$logTotal_Expenditure=log(market$Total_Expenditure)
model=lm(market$logTotal_Expenditure ~ market$Income)
res=resid(model)
qqnorm(res)
qqline(res)
#test for log(TotalPurchases) vs log(Total_Expenditure)
model=lm(market$logTotal_Expenditure ~ market$logTotalPurchases)
res=resid(model)
qqnorm(res)
qqline(res)
#test for web purchases vs web visits
model=lm(market$NumWebPurchases ~ market$NumWebVisitsMonth)
res=resid(model)
qqnorm(res)
qqline(res)

```

## **#Multiple Linear Regression**

### **#Grouping of data**

```

market$Total_Expenditure<-market$MntFishProducts+market$MntWines+market$MntFruits+marke
t$MntGoldProds+market$MntMeatProducts+market$MntSweetProducts
market$Total_Purchase<-market$NumCatalogPurchases+market$NumDealsPurchases+market$Num
StorePurchases+market$NumWebPurchases
market$Total_Cmp_accep<-market$AcceptedCmp1+market$AcceptedCmp2+market$AcceptedCmp
3+market$AcceptedCmp4+market$AcceptedCmp5
dim(market)

```

### **#Data Visualisation + Further Cleaning**

```

hist(market$Total_Expenditure)
hist(market$Total_Purchase)
hist(market$Total_Cmp_accep)
#hist(market$Total_Cmp_acceptanceR)
market$logTotal_Expenditure <- log(market$Total_Expenditure+1)
hist(market$logTotal_Expenditure)
marketA <- market[market$Total_Cmp_accep>0,]

```



```

dim(marketA)
#Checking of correlations
attach(market)
cor(Income, Total_Purchase)
cor(Income, NumWebVisitsMonth)
cor(Income, Age)
cor(Income, numChild)
cor(Income, logTotal_Expenditure)
cor(Age, NumWebVisitsMonth)
cor(Age, Total_Purchase)
cor(Age, logTotal_Expenditure)
cor(Age, numChild)
cor(NumWebVisitsMonth, Total_Purchase)
cor(NumWebVisitsMonth, logTotal_Expenditure)
cor(NumWebVisitsMonth, numChild)
cor(Total_Purchase, numChild)
cor(Total_Purchase, logTotal_Expenditure)
#Multiple linear regression model
modelA = lm(Total_Cmp_accep~Income+Age+NumWebVisitsMonth+numChild, data = market)
summary(modelA)
step(modelA, direction="backward")

#Random Forest Model
#import library
library(randomForest)
library(caret)
require(caTools)
#clean data
market2 <- market
head(market2)
market2$Total_Cmp_accep[market2$Total_Cmp_accep>1] <- 1
market2$numChild[market2$numChild>1] <- 1
summary(market2)
dim(market2)
#remove irrelevant variables
market2 <- market2[, -c(1,3,4,6:19,21:26,32:43, 45)]
summary(market2)
sapply(market2, class)
#transform to categorical variables
market2 <- transform(
  market2,
  Country=as.factor(Country),
  Total_Cmp_accep=as.factor(Total_Cmp_accep),
  numChild=as.factor(numChild),
  Complain=as.factor(Complain)
)
sapply(market2, class)
summary(market2)
#train random forest model
sample = sample.split(market2$Total_Cmp_accep, SplitRatio = .75)
train = subset(market2, sample == TRUE)
test = subset(market2, sample == FALSE)
dim(train)
dim(test)
rf <- randomForest(

```

```

Total_Cmp_accep ~ .,
data=train
)
rf
#predict using trained random forest model
pred = predict(rf, newdata=test[-9])
confusionMatrix(pred, test$Total_Cmp_accep)
varImpPlot(rf, main = 'Variable Importance')

```

## 7.2 Random Forest Model

In this section, we attempt to build a random forest model for Total Campaign Acceptance (Total\_Cmp\_accep) based on 8 given performance measures, namely Income, Age, NumWebVisitsMonth, Country, Marital\_encoded, Educationlvl, numChild, Complain. We first split the data into the train and test dataset. Then we train the random forest model using the train dataset and use it to predict the test dataset. Lastly, we use a confusion matrix to evaluate the performance of our model. The result is shown in the R output below.

We conclude that Income, Age, NumWebVisitsMonth are the top 3 most significant measures that could be used to .

```

> sample = sample.split(market3$Total_Cmp_accep, SplitRatio = .75)
> train = subset(market3, sample == TRUE)
> test = subset(market3, sample == FALSE)
> dim(train)
[1] 1654 9
> dim(test)
[1] 551 9
> rf <- randomForest(
+   Total_Cmp_accep ~ .,
+   data=train
+ )
> rf

Call:
randomForest(formula = Total_Cmp_accep ~ ., data = train)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 2

OOB estimate of error rate: 17.65%
Confusion matrix:
  0 1 class.error
0 1263 47 0.03587786
1 245 99 0.71220930

```

```

> pred = predict(rf, newdata=test[-9])
> confusionMatrix(pred, test$Total_Cmp_accep)
Confusion Matrix and Statistics

          Reference
Prediction 0 1
0  426  88
1   11  26

      Accuracy : 0.8203
      95% CI   : (0.7857, 0.8515)
No Information Rate : 0.7931
P-Value [Acc > NIR] : 0.06182

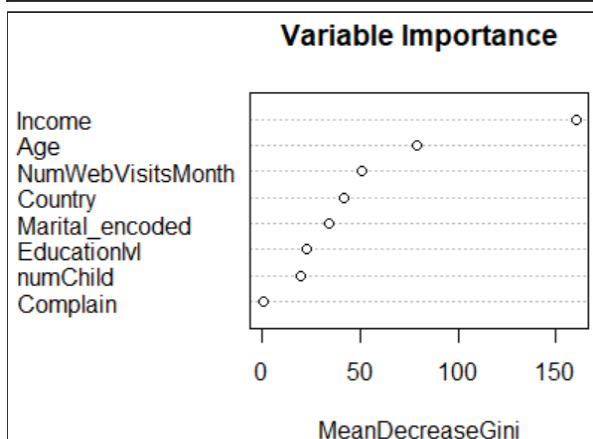
      Kappa : 0.2704

McNemar's Test P-Value : 2.201e-14

      Sensitivity : 0.9748
      Specificity : 0.2281
      Pos Pred Value : 0.8288
      Neg Pred Value : 0.7027
      Prevalence : 0.7931
      Detection Rate : 0.7731
      Detection Prevalence : 0.9328
      Balanced Accuracy : 0.6014

'Positive' Class : 0

```



## 8 References

1. Jack Daoud. (2021, April). Marketing Analytics, Version 1. Retrieved 16 March, 2022 from <https://www.kaggle.com/datasets/jackdaoud/marketing-data/version/1>