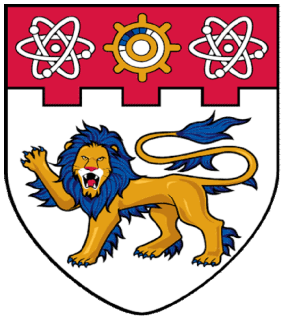


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH3511 Data Analysis with Computer

Group Project

Name	Matriculation Number
Bodipati Kiran	U1922759J
Cui Chenling	U1920157F
Iyengar Varun Srikant	U1922219K
Singh Ananya	U1923401D
Vaidyanathan Abhishek	U1923980D

Abstract:

In today's world, the success of any company depends on how well they market their products. Given that there are multiple alternatives to a consumer for buying anything, a poorly marketed product may not attract customers. To better market a product, it becomes imperative that a company understands its consumers. This includes but is not limited to the various factors that affect the income of a customer, the expenditure patterns of a consumer, other behavioural patterns of a consumer, the segmentation towards whom the marketing campaign was targeted, etc. To gain a better understanding of all these factors, we would like to perform a basic data analysis of the customer profiles and marketing data of a company.

1.Introduction

In today's world, the success of any company depends on how well they market their products. Given that there are multiple alternatives to a consumer for buying anything, a poorly marketed product may not attract the customers. Many multinational conglomerates spend billions of dollars annually just as a part of their marketing expenditures. To better market a product, it becomes imperative that a company understands its consumers.

In our project, we used a dataset containing the customer profiles and their spending patterns at a particular company. This included consumer details like income, education level, marital status, number of children, their expenditure on various products like fish, meat, wine, fruits, the etc. It also contained data of their response towards marketing campaigns. Based on this dataset, we aim to study the following:

1. How do factors like age, marital status, education level, number of children, country of residence, age etc affect the income of the consumers.
2. How does the income affect the expenditure of a person on various items like meat, fish, fruits, wine, gold etc.
3. Does the income level of a person affect the acceptance of the various marketing campaigns of a company?

In this report, we performed the various descriptions and analysis of the data using the R Language. We have formulated the appropriate hypotheses and conducted appropriate statistical analysis to draw appropriate conclusions for each of the research questions defined above with appropriate graphs and explanations.

2.Data Description

The dataset, titled '**Marketing Analytics**' is obtained from the largest data science competition community, Kaggle. The dataset contains 1 csv file named 'marketing_data.csv'. The dataset was provided to students for their final project in order to test their statistical analysis skills as part of a MSc. in Business Analytics, by Dr. Omar Romero-Hernandez. The dataset provides a comprehensive list of observations on customer profiles, product preferences, campaign success and failures and channel performances.
After all the preparation, 2240 observations with 19 categories are retained for statistical analysis:

1. ID: unique ID to identify the individual
2. Year_Birth: the year that the individual was born in
3. Education: the education level of the individual, categories include Graduation, Basic, PhD,master or 2nd Cycle
4. Marital_Status: the marital status of the individual, categories include single, divorced, widow, married, together, YOLO, alone and absurd
5. Income: yearly income of the individual
6. Kidhome: the number of kids that the individual lives with
7. Teenhome: the number of teenagers that the individual lives with
8. MntWines: the amount of money spent on wines in the past 2 years
9. MntFruits: the amount of money spent on fruits in the past 2 years
10. MntMeatProducts: the amount of money spent on meat products in the past 2 years
11. MntFishProducts: the amount of money spent on fish products in the past 2 years
12. MntSweetProducts: the amount of money spent on sweet products in the past 2 years
13. MntGoldProds: the amount of money spent on gold products in the past 2 years
14. AcceptedCmp1: 1 if the individual accepts the 1st campaign, 0 otherwise
15. AcceptedCmp2: 1 if the individual accepts the 2nd campaign, 0 otherwise
16. AcceptedCmp3: 1 if the individual accepts the 3rd campaign, 0 otherwise
17. AcceptedCmp4: 1 if the individual accepts the 4th campaign, 0 otherwise

18. AcceptedCmp5: 1 if the individual accepts the 5th campaign, 0 otherwise
19. Response: 1 if the individual accepts the 6th campaign, 0 otherwise
20. Country: the country where the individual lives in, categories include AUS, CA, GER, IND, ME, SA, SP and US.

3. Data Cleaning and Exploratory Analysis

In this section, we shall look into the data in more detail. Each variable is investigated individually to look for possible outliers, and/or to perform a transformation to avoid highly skewed data.

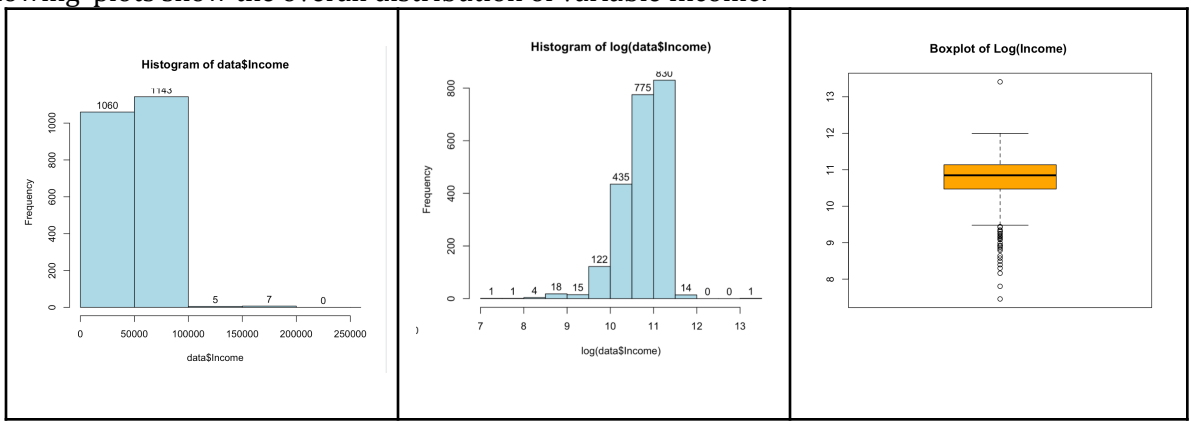
3.1 Summary statistics for the main variables of interest

As mentioned in the introduction, our objectives are to investigate the factors affecting income, how expenditures on different types of food are affected by income, and how campaign acceptance is affected by income. Hence, our main variables of interest in the dataset are *Income, various expenditures and various campaigns*.

3.1.1 Income

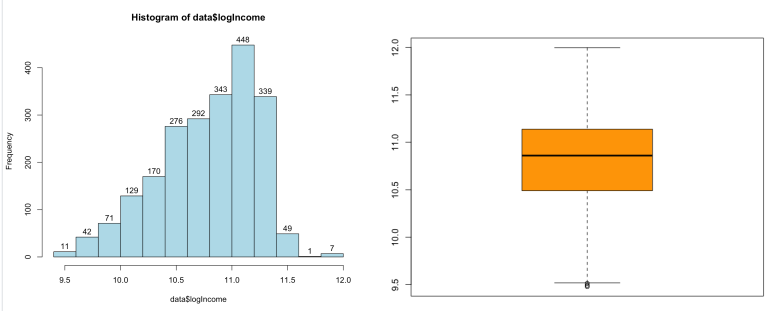
The data type of variable Income in the dataset is string with a dollar sign in front of the values. As individual income is a continuous variable, we removed the dollar sign and *converted* the data type to *numerical.* 24 rows with null value in Income were also removed from the dataset.

The following plots show the overall distribution of variable Income.



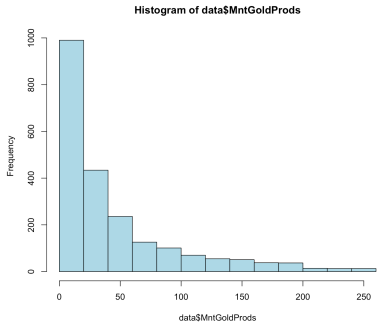
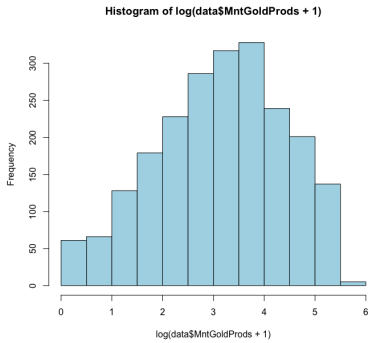
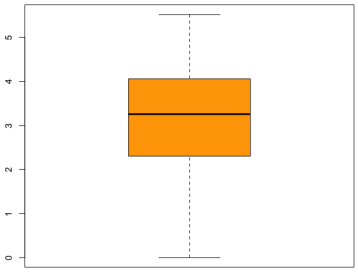
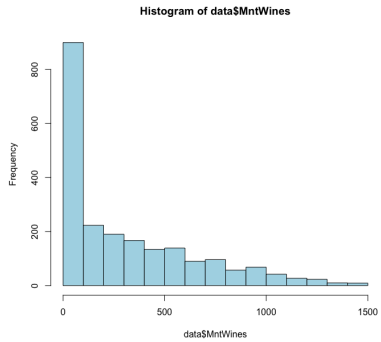
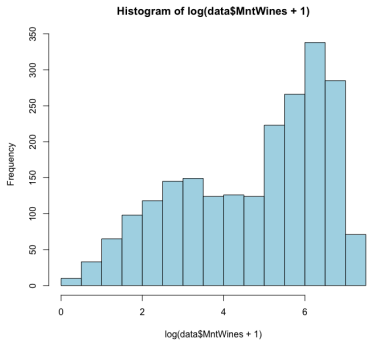
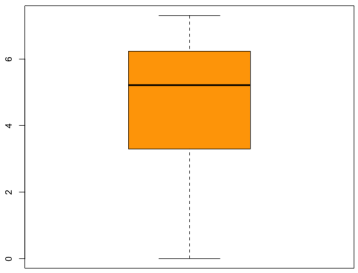
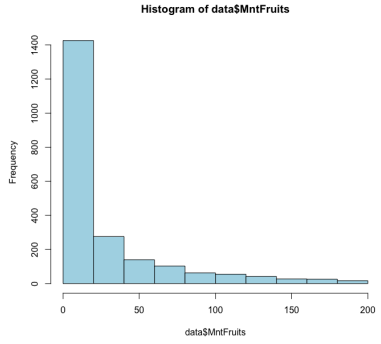
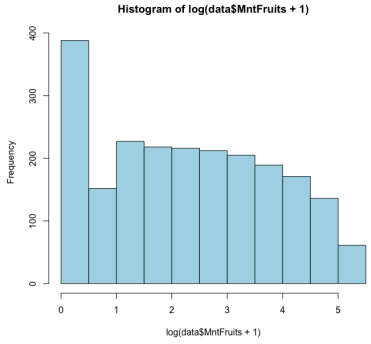
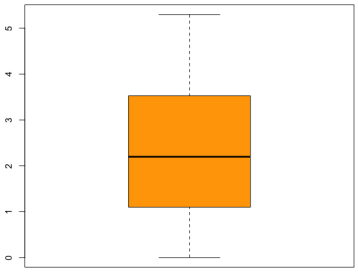
It appears that the variable *Income* is highly skewed, hence we apply a log transformation (base *e*) to the variable. The log-transformed data appears to have some outlying values at the left tail. Hence, we remove the individuals with log(Income) out of the whiskers of 1.5 IQR, that is, 62 rows in total.

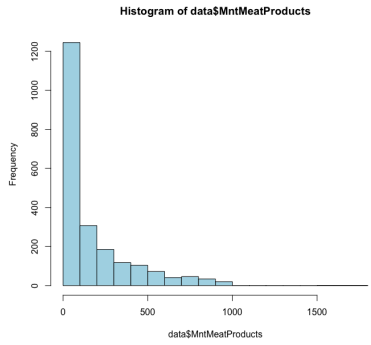
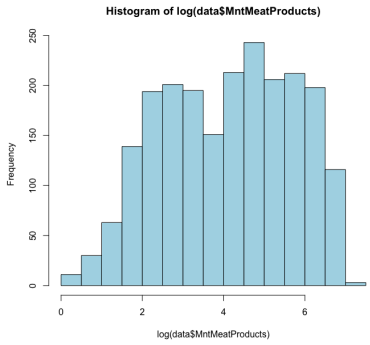
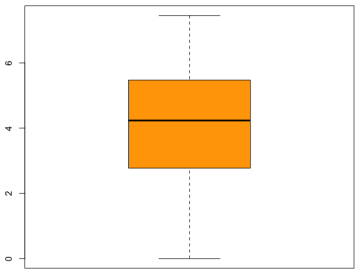
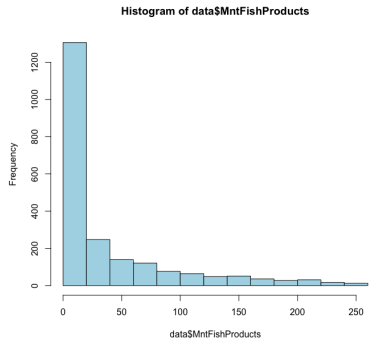
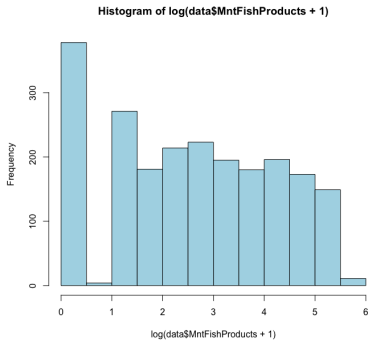
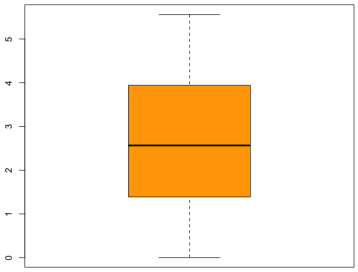
As shown below in the histogram and boxplot of the log-transformed variable, after removal of the outliers, along with summary statistics, the dataset is more symmetrically-distributed and with few outliers with respect to variable *Income*.

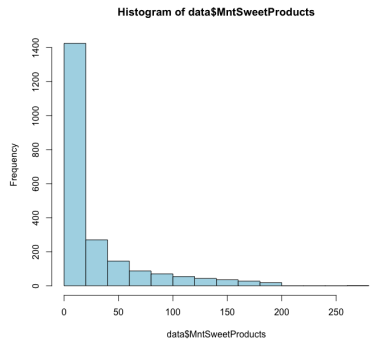
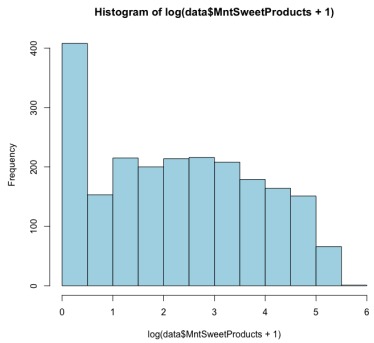
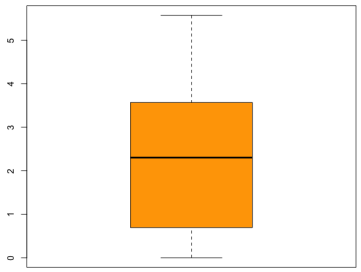


```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
13084   35966   52054   52721   68731  162397
```

3.1.2 Expenditures

		
Amount spent on gold product in the past 2 years, *MntGoldProds*	It appears that the variable *MntGoldProds* is highly skewed, hence we apply a log transformation (base *e*) to the variable.	The log-transformed data does not have any outliers and show a symmetrical distribution.
		
Amount spent on wine product in the past 2 years, *MntWines*	It appears that the variable *MntWines* is highly skewed, hence we apply a log transformation (base *e*) to the variable, after adding 1 to each value, to deal with 0 value.	The log-transformed data does not have any outliers and show a symmetrical distribution.
		

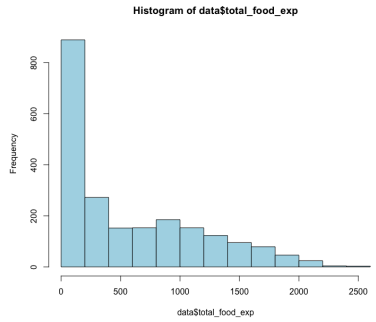
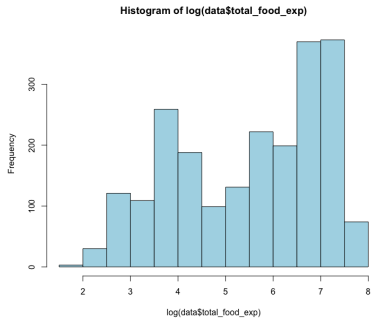
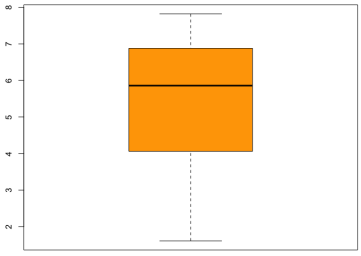
Amount spent on fruits in the past 2 years, *MntFruits*	It appears that the variable *MntFruits* is highly skewed, hence we apply a log transformation (base *e*) to the variable, after adding 1 to each value, to deal with 0 value.	The log-transformed data does not have any outliers and show a symmetrical distribution.
 Amount spent on meat products in the past 2 years, *MntMeatProducts*	 It appears that the variable *MntMeatProducts* is highly skewed, hence we apply a log transformation (base *e*) to the variable.	 The log-transformed data does not have any outliers and show a symmetrical distribution.
 Amount spent on fish products in the past 2 years, *MntFishProducts*	 It appears that the variable *MntFishProducts* is highly skewed, hence we apply a log transformation (base *e*) to the variable, after adding 1 to each value, to deal with 0 value.	 The log-transformed data does not have any outliers and show a symmetrical distribution.

		
Amount spent on sweet products in the past 2 years, *MntSweetProducts*	It appears that the variable *MntSweetProducts* is highly skewed, hence we apply a log transformation (base *e*) to the variable,after adding 1 to each value, to deal with 0 value.	The log-transformed data does not have any outliers and show a symmetrical distribution.

To further investigate the variation of expenditures, a new column named *total_food_exp* is constructed, denoting the total expenditure on all given types of food products through summing up the individual expenses on each type of food products, i.e. *MntMeatProds, MntWines, MntFruits, MntFishProducts,MntSweetProducts.*

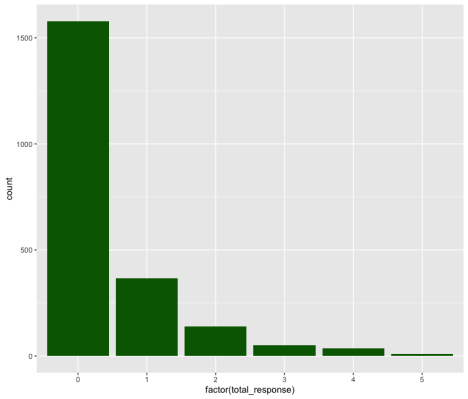
Total expenditure on food items, *total_food_exp :*

It appears that the variable *total_food_exp* is highly skewed, hence we apply a log transformation (base *e*) to the variable. The log-transformed data does not have any outliers and show a relatively symmetrical distribution.

3.1.3 Response to campaigns

To investigate the factors causing acceptance of campaigns, a new column named *total_response* is constructed through summing up all responses from all 6 campaigns. This denotes the total number of campaigns each individual accepts. **Ggplot** is used to see the distribution of the total number of responses to all campaigns.

	No outlying values of total_response are removed.

3.2 Summary statistics for other variables

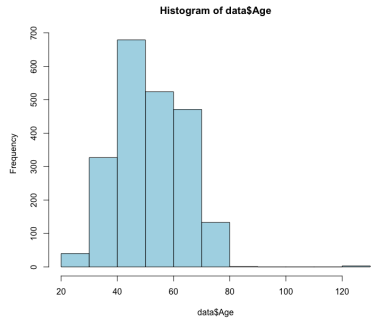
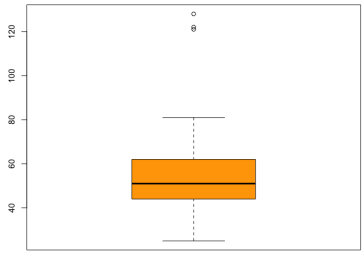
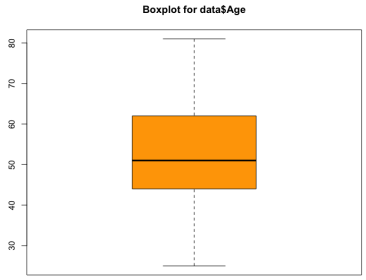
In the dataset, there is a combination of continuous and categorical variables. We will look at them separately.

3.2.1 Continuous variables

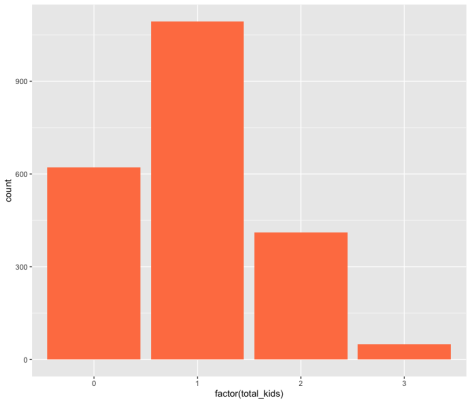
The histogram, the boxplot, the transformation applied and the outliers removed from the variable are shown in the below tables.

3.2.1.1 Age of individuals as of 2021, *Age*

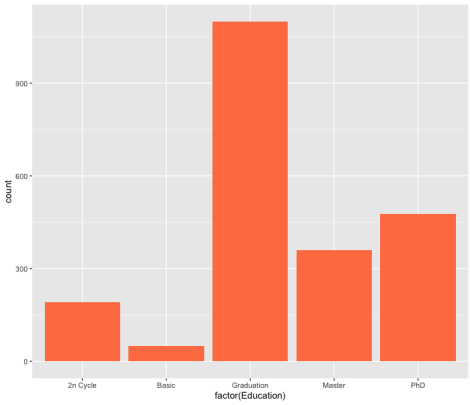
The variable **Age** is created from subtracting the year of birth from the Year_Birth column from 2021.

		
Data is not skewed, no log-transformation.	A few outliers are observed.	Removed outliers that are outside the whiskers of 1.5IQR.

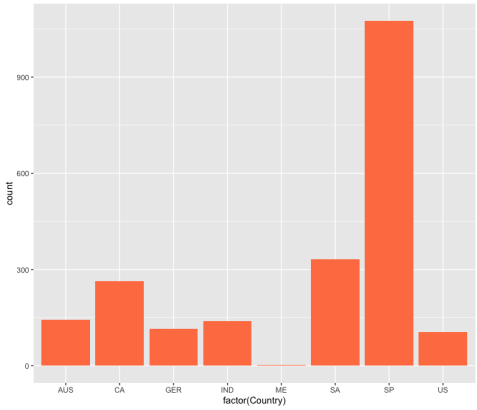
3.2.2 Categorical variables



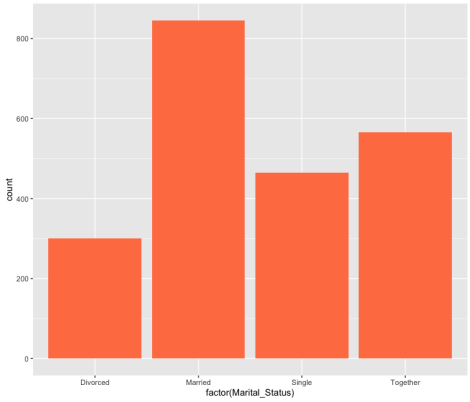
Total number of kids at home, *total_kids* (created from summing up Kidhome and Teenhome)



Education level, *Education*



Country of residence of the individuals, *Country*



Marital status of the individuals, *Marital_Status* (Marital_Status has been reduced to 4 categories: Divorced, Married, Single and Together, after combining Widow, YOLO, absurd and alone to Single)

No outlying data points are removed from any variables.

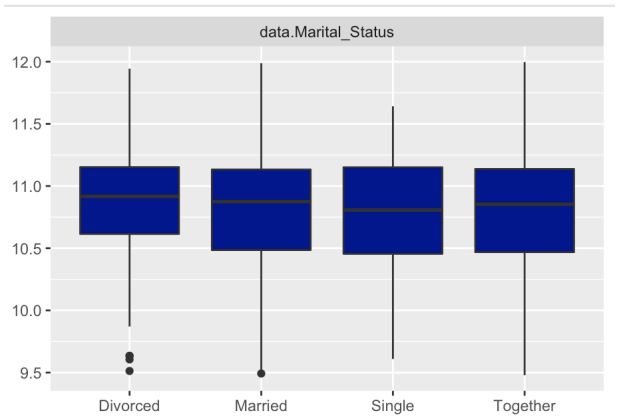
Based on the above analysis and removal of outliers, the dataset is reduced to **2175** observations.

4. Statistical Analysis

4.1 Factors affecting *logIncome*

In this section, we will investigate among *Marital_Status, Country, Education, total_kids and Age*, which factor affects *logIncome* the most. Here, t-test and ANOVA are performed for *Marital_Status, Country, Education* and *total_kids* since they are categorical data. Correlation test is performed for *Age* and *logIncome* since Age is a continuous variable.

4.1.1 Relation between logIncome and Marital_Status



Looking at the boxplot, we see that the spread of logIncome is similar for all 4 marital status. Hence, the ANOVA test is appropriate for testing the equality of the means($\boldsymbol{\mu}$). We test,

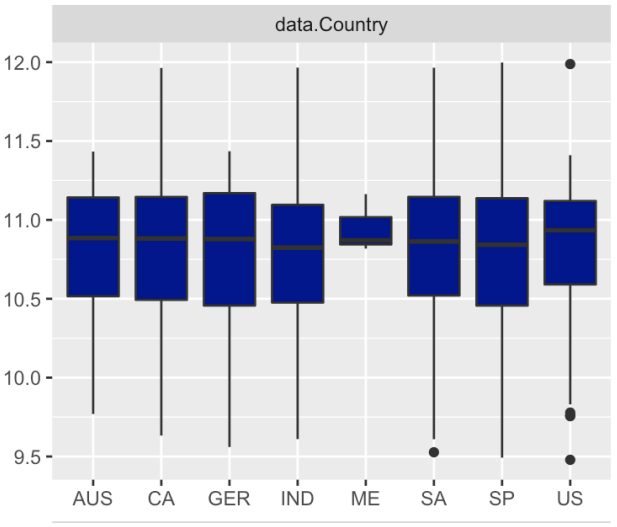
$H_0: \mu_{divorced} = \mu_{married} = \mu_{single} = \mu_{togethe}$

against H_1: *not all* μ_i are equal.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(data$Marital_Status)	3	1.2	0.4104	2.122	0.0954
Residuals	2171	419.9	0.1934		

The ANOVA test returns a p-value of **0.0954**, which shows that the means are not significantly different at a significance level of 0.05. Thus, we conclude that the income of individuals is independent of his/her marital status.

4.1.2 Relation between logIncome and Country



Looking at the boxplot, we see that the spread of *logIncome* is similar for all 8 countries. Hence, the ANOVA test is appropriate for testing the equality of the means($\boldsymbol{\mu}$). We test,

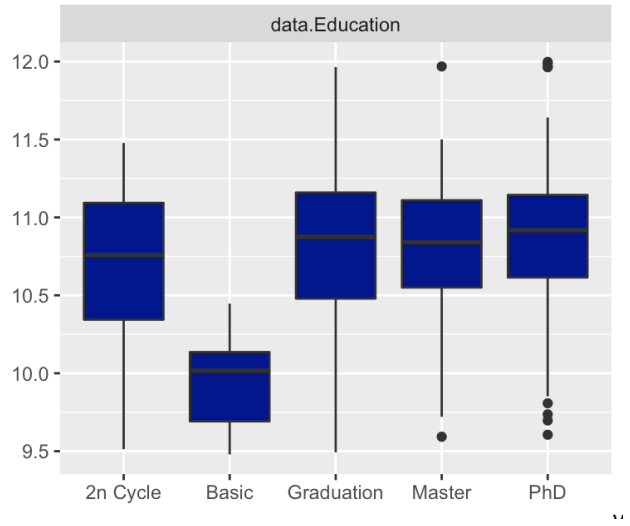
$H_0: \mu_{AUS} = \mu_{CA} = \mu_{GER} = \mu_{IND} = \mu_{ME} = \mu_{SA} = \mu_{SP} = \mu_{US}$

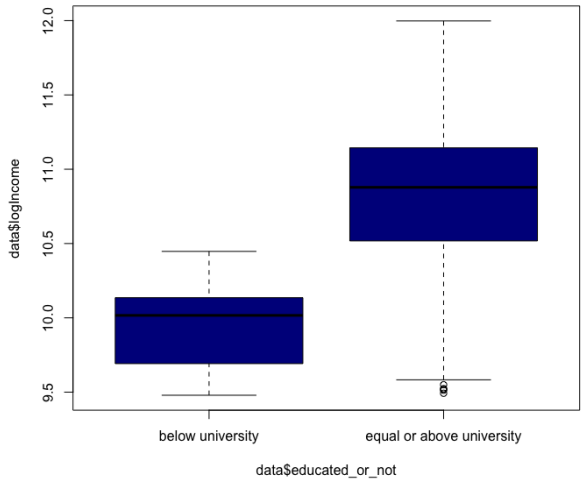
against H_1: *not all* μ_i are equal.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(data$Country)	7	1.0	0.1403	0.724	0.652
Residuals	2167	420.1	0.1939		

The ANOVA test returns a p-value of 0.652, which shows that the means are not significantly different at a significance level of 0.05. We conclude that the income of individuals is independent of their country.

4.1.3 Relation between logIncome and Education





As shown in the 1st boxplot, we see that the spread of logIncome is similar for individuals having education level above basic, but is obviously different between individuals having education level above basic and those having basic education.Hence, we constructed a new column named educated_or_not and assigned values == 'Basic' to 'below university', and the rest of the education levels to 'equal or above university'.

As shown in the 2nd boxplot, it is apparent that a *logIncome* gap exists between the individuals having education level below university and those equal or above university. We conducted further statistical tests(variance F-test and t-test) to support this observation.

Variance test

```
> var.test(data$logIncome[data$educated_or_not=='below university'],data$logIncome[data$educated_or_not=='equal or above university'])

        F test to compare two variances

data:  data$logIncome[data$educated_or_not == "below university"] and data$logIncome[data$educated_or_not == "equal or above university"]
F = 0.34702, num df = 48, denom df = 2125, p-value = 1.603e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2400259 0.5435397
sample estimates:
ratio of variances
         0.3470171
```

H_0: *Variance of below university and => university's logIncome are equal*

H_1: *Variance of below university and => university'slogIncome are not equal*

At a significance level of 0.05, we **reject** the null hypothesis and conclude that the variances of the 2 samples are not equal, since p-value = 1.603e-05< 0.05.

T-test

```
> t.test(data$logIncome[data$educated_or_not=='below university'],data$logIncome[data$educated_or_not=='equal or above university'],var.equal=FALSE,
 alternative='less')

        Welch Two Sample t-test

data:  data$logIncome[data$educated_or_not == "below university"] and data$logIncome[data$educated_or_not == "equal or above university"]
t = -23.253, df = 54.582, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.7960357
sample estimates:
mean of x mean of y
  9.94571  10.80347
```

H_0: *Mean of below university and => university'slogIncome are equal*

H_1: *Mean of below university's logIncome is less than that of equal or above university's logIn*

At a significance level of 0.05, we reject the null hypothesis and conclude that the mean of logIncome of individuals with below university degree is significantly less than that of those with equal or above university degree , since p-value = 2.2e-16< 0.05.

For the individuals who have equal or above university degree, we further perform ANOVA test to check the equality of mean(μ) for all 4 samples: 2nd cycle, master , graduate, PhD

ANOVA test

H_0: $\mu_{2nd\ cycle} = \mu_{master} = \mu_{graduate} = \mu_{PhD}$ against H_1: *not all* μ_i are equal.

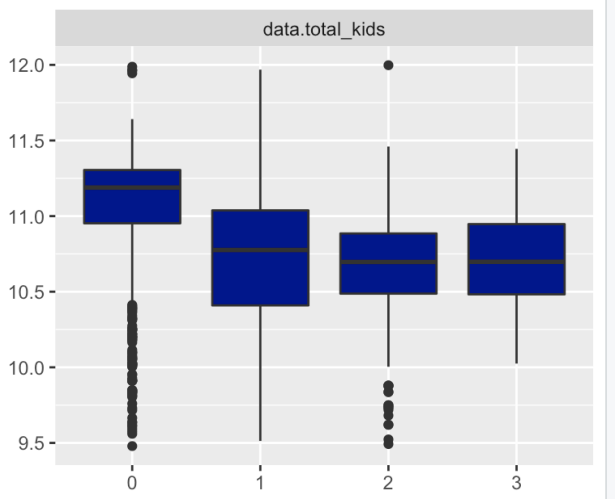
```
                                    Df Sum Sq Mean Sq F value   Pr(>F)
factor(educated_people$Education)    3    5.5  1.8175   10.22 1.11e-06 ***
Residuals                         2122  377.4  0.1779
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

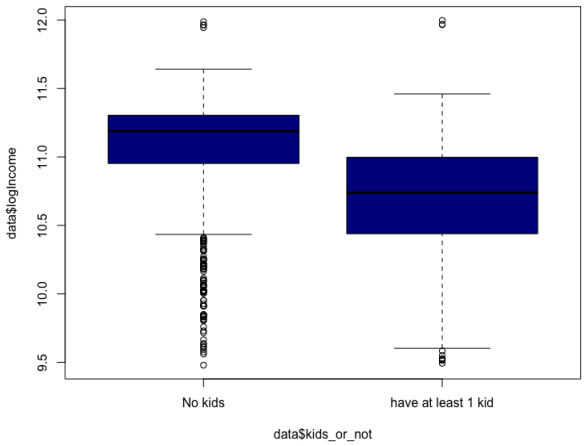
The ANOVA test returns a p-value of 1.11e-06, we reject the null hypothesis. A pairwise T-test is performed to see which sample deviates from the rest.

```
> pairwise.t.test(educated_people$logIncome,educated_people$Education,p.adjust.method = "none")

        Pairwise comparisons using t tests with pooled SD

data:  educated_people$logIncome and educated_people$Education

          2n Cycle Graduation Master
Graduation 0.0047   -          -
Master     0.0011   0.2388     -
PhD        2.7e-07  6.2e-05    0.0335

P value adjustment method: none
```

From the result, we can deduce that individuals with 2nd Cycle education level or PhD education level differ from other people with respect to *logIncome*.

4.1.4 Relation between logIncome and total_kids





As shown in the 1st boxplot, we see that the spread of *logIncome* is similar for individuals having 1 to 3 kids, but is obviously different between individuals having no kids and individuals having at least 1 kid. Hence, we constructed a new column named kids_or_not and assigned value 1 for individuals who have at least 1 kid, 0 for individuals who do not have any kids.
As shown in the 2nd boxplot, it is apparent that a *logIncome* gap exists between the individuals with no kids and those with at least 1 kid. We conducted further statistical tests(variance F-test and t-test) to support this observation.

Variance test

```
> var.test(data$logIncome[data$kids_or_not==0],data$logIncome[data$kids_or_not==1])

        F test to compare two variances

data:  data$logIncome[data$kids_or_not == 0] and data$logIncome[data$kids_or_not == 1]
F = 1.3146, num df = 620, denom df = 1553, p-value = 3.238e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.154504 1.502631
sample estimates:
ratio of variances
          1.31463
```

H_0: *Variance of no kids and have >= 1 kid's logIncome are equal*

H_1: *Variance of no kids and have >= 1 kid'slogIncome are not equal*

At a significance level of 0.05, we reject the null hypothesis and conclude that the variances of the 2 samples are not equal, since p-value = 3.238e-05< 0.05.

T-test

```
> t.test(data$logIncome[data$kids_or_not==0],data$logIncome[data$kids_or_not==1],alternative= 'greater',var.equal = FALSE)

        Welch Two Sample t-test

data:  data$logIncome[data$kids_or_not == 0] and data$logIncome[data$kids_or_not == 1]
t = 16.249, df = 1016.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3034753       Inf
sample estimates:
mean of x mean of y
 11.02542  10.68773
```

H_0: *Mean of no kids and have >= 1 kid'slogIncome are equal*

H_1: *Mean of no kids'logIncome is larger than the mean of have >= 1 kid's logIncome*

At a significance level of 0.05, we reject the null hypothesis and conclude that the mean of *logIncome* of individuals with no kids is significantly larger than those with kids, since p-value = 2.2e-05< 0.05.

For the individuals who have at least 1 kid, we further perform ANOVA test to check the equality of mean(μ) for all 3 samples: 1 kid, 2 kids and 3 kids.

ANOVA test

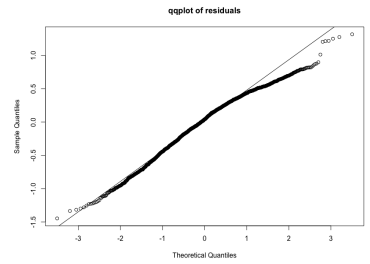
$H_0: \mu_1 = \mu_2 = \mu_3$ against H_1: *not all* μ_i are equal.

```
                                    Df Sum Sq Mean Sq F value Pr(>F)
factor(have_kids_data$total_kids)    2   0.37  0.1852   1.184  0.306
Residuals                         1551 242.62  0.1564
```

The ANOVA test returns a p-value of 0.306, which shows that the means are not significantly different at a significance level of 0.05.We conclude that the income of individuals who have kids is independent of the number of kids that he has.

4.1.5 Relation between logIncome and Age

Since Age is a continuous variable, we perform single linear regression to investigate the relationship between logIncome and Age.

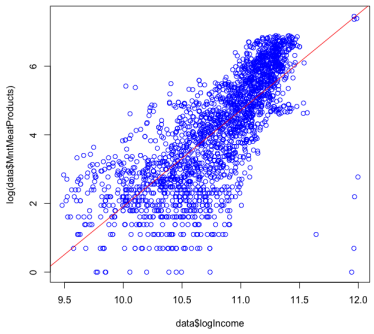
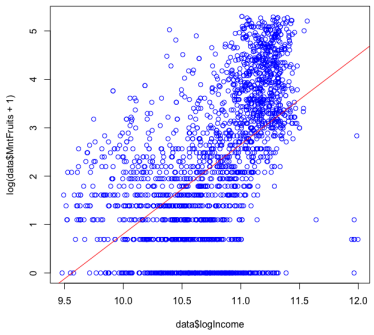
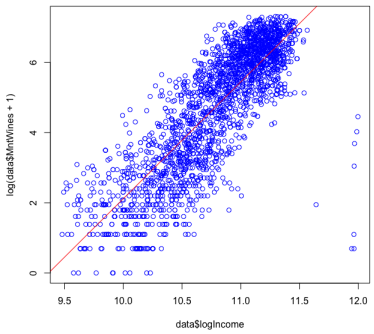
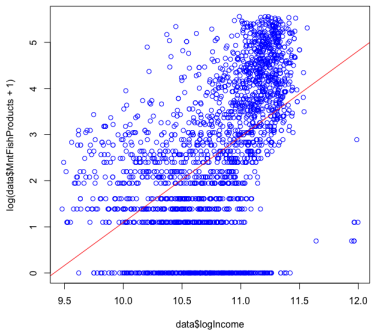
Variable	Fitted model, with Y being logIncome	p-value	R-squared	qq-plot of residuals
Age	Y =1.034e+01 +8.482e-03(Age)	< 2.2e-16	0.05058	

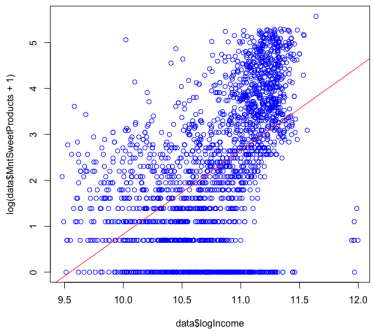
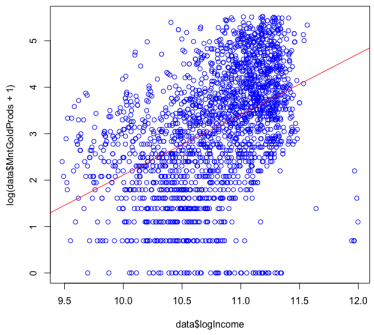
By observing the R^2 value of Age vs LogIncome, we find that Income is independent of the age of the person. This implies that companies can focus on segmenting their target audience in a way that does not involve

4.2 How does logIncome affect various expenditures

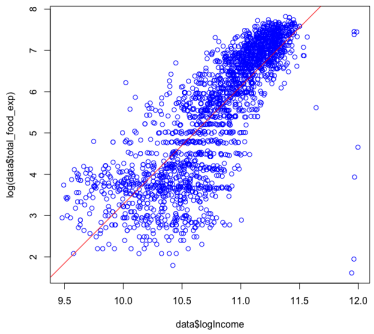
4.2.1 Income vs. Amount spent on individual product types

Y	Fitted model, with X being logIncome	p-value	R-squared	Scatter plot with regression line

log(MntMeatProds)	*Y* =-26.10609 +2.80211*X*	< 2.2e-16	0.5858	
log(MntFruits)	*Y* =-17.69221 +1.84974 *X*	< 2.2e-16	0.2675	
log(MntWines)	*Y* =-31.1744 +3.3291*X*	<2.2e-16	0.6799	
log(MntFishProducts)	*Y*=-17.51458+ 1.86091 *X*	<2.2e-16	0.2419	

log(MntSweetProducts)	*Y* =-17.40269 +1.82295*X*	<2.2e-16	0.2519	
log(MntGoldProds)	*Y* = -10.93785+ 1.30461*X*	< 2.2e-16	0.1998	
By comparing the R-squared values,we found that *log(MntWines) and log(MntMeatProds)* have a high positive correlation to *logIncome*. This means that higher the income, higher is the expenditure on **Meat Products** and **Wine**. We also observe that while the expenditure on other plots is also positively related to Income, their correlation is really very high, which shows that the expenditure on these products is not that strongly related to Income.				

4.2.2 Income vs. Total Food Expenditure on all products total_food_exp

Y	Fitted model, with X being logIncome	p-value	R-squared	Scatter plot with regression line
log(total_food_exp)	*Y* =-25.2427+ 2.8517 *X*	< 2.2e-16	0.6703	
We conduct a regression on the log of total expenditure on food products against *logIncome* and we found that again as expected there is a high positive correlation implying that people with higher				

income splurge on food. This means that the company could market food products more to those with higher income, as they are more likely to purchase these products.

4.3 How does campaign acceptance vary with income?

4.3.1 logIncome vs. Individual campaign acceptance

Tests performed for each campaign:

Variance test

sample 1: Individuals who accepted 1st(or 2nd, 3rd, 4th, 5th, 6th) campaign
sample 2: Individuals who did not accept the campaign

H_0: $Variance\ of\ \ sample\ 1\ and\ 2\ are\ equal$

H_1: $Variance\ of\ \ sample\ 1\ and\ 2\ are\ not\ equal$

T-test

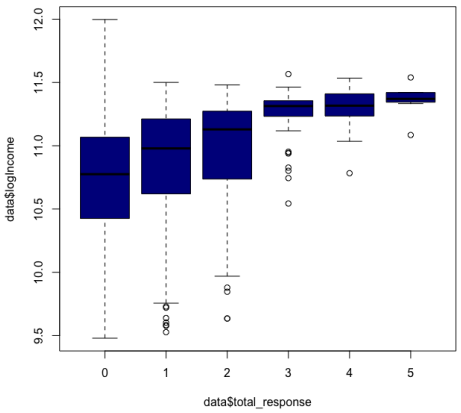
H_0: $Mean\ of\ sample\ 1\ and\ \ 2\ are\ equal.$

H_1: $Mean\ of\ sample\ 1\ is\ less\ than\ sample\ 2.$

Campaign	Variance test p-value	T-test p-value	Analysis
1	2.2e-16<0.05	2.2e-16<0.05	**Variance test** Reject the null hypothesis and conclude that the variances of the 2 samples are not equal. **T-test** Reject the null hypothesis and conclude that the mean of logIncome of individuals who did not accept the 1st campaign is significantly less than those who accepted it.
2	0.0001097<0.05	7.661e-10<0.05	**Variance test** Reject the null hypothesis and conclude that the variances of the 2 samples are not equal. **T-test** Reject the null hypothesis and conclude that the mean of logIncome of individuals who did not accept the 2nd campaign is significantly less than those who accepted it.
3	0.3998>0.05	0.8078>0.05	**Variance test** Do not reject the null hypothesis and conclude that the variances of the 2 samples are equal. **T-test** Do not reject the null hypothesis as there is no sufficient data to prove that the means are not equal.
4	2.2e-16<0.05	2.2e-16<0.05	**Variance test** Reject the null hypothesis and conclude that the variances of the 2 samples are not equal. **T-test** Reject the null hypothesis and conclude that the

			mean of logIncome of individuals who did not accept the 4th campaign is significantly less than those who accepted it.
5	2.2e-16<0.05	2.2e-16<0.05	**Variance test** Reject the null hypothesis and conclude that the variances of the 2 samples are not equal. **T-test** Reject the null hypothesis and conclude that the mean of logIncome of individuals who did not accept the 5th campaign is significantly less than those who accepted it.
6	0.2562>0.05	1.436e-10	**Variance test** Do not reject the null hypothesis and conclude that the variances of the 2 samples are equal. **T-test** Reject the null hypothesis and conclude that the mean of logIncome of individuals who did not accept the 6th campaign is significantly less than those who accepted it.
From the above data, it is evident that for all campaigns except campaign 3, there is a significant correlation between campaign acceptance and income. Although no information regarding the campaigns are available, it is likely that these campaigns featured premium or non-essential products. Hence, the company could target future campaigns at higher income brackets, focusing a larger portion of their resources on advertising to them in order to drive higher sales.			

4.3.2 logIncome vs. Total campaign acceptance *total_response*



Looking at the boxplot, we see that the spread of logIncome is different for different numbers of total response. Hence, the ANOVA test is appropriate for testing the equality of the means($\boldsymbol{\mu}$). We test,

$H_0: \mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

against H_1: *not all* μ_i are equal.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(data$total_response)	5	39.4	7.873	44.73	<2e-16 ***
Residuals	2169	381.8	0.176		

The ANOVA test returns a p-value of 2e-16, we reject the null hypothesis in favour of the alternative hypothesis. This substantiates that the income does affect the attitudes of the people towards accepting marketing campaigns. From the graph, we see that people with a higher income are more accepting of marketing campaigns.

5. Conclusion

Today, each company is trying to improve its business model by trying to find the right combination of their marketing strategies. There is fierce competition between companies to push their products over the ones of other companies. We attempted to provide a marketing analysis of consumer data of a company that included studying the various factors that affected the income of a family, the expenditure patterns that varied with income as well as the response of the various segments towards the marketing campaigns launched by the company.

We were able to draw the following interesting conclusions:

- The income of the customers is independent of their age, marital status and country of residence
- The income of the customers depends on their education level and whether they have kids.
- Among the various products that people spend on, we see that an increase in income increases the amount spent on wines and meat. The money spent on products like fish, sweets and gold does not depend as heavily on income.
- The acceptance of marketing campaigns among consumers depends on income. A person with higher income is more likely to accept more marketing campaigns.

Additionally, if we combine our results from the tests that analysed the effects of income on expenditure and the marketing campaigns, we can extrapolate that certain marketing campaigns featured certain products. While this was not statistically tested in our study, it is an interesting observation.

Given the very large number of columns in our dataset, we had to restrict our analysis to cover only certain observations related to the analysis of customer profiles and marketing behaviour. This study provides a basic analysis of the marketing data; however, a deeper and extensive study can be conducted to find various other factors that can affect the expenditure as well as the acceptance of various marketing campaigns to gain a deeper understanding of the data.

Appendix - R code

```
#import data
data = read.csv('/Users/chenlingcui/Desktop/MH3511/marketing_data.csv',sep=',',header= TRUE)
#remove dollar sign for Income, and remove null values
data$Income = as.numeric(gsub("[\\$,]", "", data$Income))
data = data[!is.na(data$Income), ]
#show distribution of Income
hist(data$Income,col='lightblue')
#log transformation of Income and store as new variable logIncome
hist(log(data$Income),col='lightblue')
boxplot(log(data$Income),col='orange')
data$logIncome = log(data$Income)
#remove outliers for logIncome
data = data[!(data$logIncome>quantile(data$logIncome, 0.75)+1.5*IQR(data$logIncome) |
data$logIncome<quantile(data$logIncome, 0.25)-1.5*IQR(data$logIncome)), ]
hist(data$logIncome,col='lightblue')
boxplot(data$logIncome,col='orange')
length(data$logIncome)
#create new variable named Age
data$Age = 2021 - data$Year_Birth
#reduce to 4 categories in Marital_Status
a<-gsub("Widow","Divorced",data$Marital_Status)
a<-gsub("YOLO","Single",a)
a<-gsub("Alone","Single",a)
a<-gsub("Absurd","Single",a)
data$Marital_Status<-a
#show distirbution of amount of gold products purchased
hist(data$MntGoldProds,col='lightblue')
hist(log(data$MntGoldProds+1),col='lightblue')
boxplot(log(data$MntGoldProds+1),col='orange')

#show distirbution of amount of wine products purchased
hist(data$MntWines,col='lightblue')
hist(log(data$MntWines+1),col='lightblue')
boxplot(log(data$MntWines+1))
#show distirbution of amount of fruit products purchased
hist(data$MntFruits,col='lightblue')
hist(log(data$MntFruits+1),col=)
boxplot(log(data$MntFruits+1),col='orange')
#show distirbution of amount of meat products purchased
hist(data$MntMeatProducts,col='lightblue')
hist(log(data$MntMeatProducts),col='lightblue')
boxplot(log(data$MntMeatProducts),col='orange')
#show distirbution of amount of fish products purchased
hist(data$MntFishProducts,col='lightblue')
hist(log(data$MntFishProducts+1),col='lightblue')
boxplot(log(data$MntFishProducts+1),col='orange')
#show distirbution of amount of sweet products purchased
```

```
hist(data$MntSweetProducts,col='lightblue')
hist(log(data$MntSweetProducts+1),col='lightblue')
boxplot(log(data$MntSweetProducts+1),col='orange')
#create new variable named total food expenditure
data$total_food_exp =
data$MntWines+data$MntFruits+data$MntMeatProducts+data$MntFishProducts+data$MntSweetProducts
hist(data$total_food_exp,col='lightblue')
hist(log(data$total_food_exp+1),col='lightblue')
boxplot(log(data$total_food_exp+1),col='orange')
#remove outlier for Age
data = data[!(data$Age>quantile(data$Age, 0.75)+1.5*IQR(data$Age) | data$Age<quantile(data$Age,
0.25)-1.5*IQR(data$Age)), ]

#create new variable named total_response
data$total_response =
data$AcceptedCmp1+data$AcceptedCmp2+data$AcceptedCmp3+data$AcceptedCmp4+data$AcceptedCmp5+data$R
esponse

#show distirbution of Age
hist(data$Age,col='lightblue')
boxplot(data$Age,col='orange')
#create new variable named total_kids
data$total_kids= data$Kidhome+data$Teenhome
#count plot for all categorical variables
ggplot(data,aes(data$total_kids))+geom_bar(fill='coral')
ggplot(data,aes(data$Education))+geom_bar(fill='coral')
ggplot(data,aes(data$Country))+geom_bar(fill='coral')
ggplot(data,aes(data$Marital_Status))+geom_bar(fill='coral')

#show what affects logIncome
data1 =  data.frame(data$Marital_Status,data$Country,data$Education,data$Age,data$total_kids,data$logIncome)
data1.m <- melt(data1, "data.logIncome")
ggplot(data1.m, aes(value, data.logIncome)) +
  +     geom_boxplot() +
  +     facet_wrap(~variable, scales = "free")
#anova test
summary(aov(data$logIncome~factor(data$Marital_Status)))

summary(aov(data$logIncome~factor(data$Country)))

summary(aov(data$logIncome~factor(data$Education)))

data$educated_or_not = data$Education
data$educated_or_not[data$educated_or_not!='Basic']='equal or above university'
data$educated_or_not[data$educated_or_not=='Basic']='below university'
boxplot(data$logIncome~data$educated_or_not,col='darkblue')
var.test(data$logIncome[data$educated_or_not=='below university'],data$logIncome[data$educated_or_not=='equal or
above university'])
t.test(data$logIncome[data$educated_or_not=='below university'],data$logIncome[data$educated_or_not=='equal or
above university'],alternative = 'less',var.equal = FALSE)

educated_people = data[data$Education!='Basic',]
summary(aov(educated_people$logIncome~factor(educated_people$Education)))

data$kids_or_not = data$total_kids
data$kids_or_not[data$kids_or_not!=0]='have at least 1 kid'
```

```
data$kids_or_not[data$kids_or_not==0]='no kids'
boxplot(data$logIncome~data$kids_or_not)
var.test(data$logIncome[data$kids_or_not=='no kids'],data$logIncome[data$kids_or_not=='have at least 1 kid'])

t.test(data$logIncome[data$kids_or_not=='no kids'],data$logIncome[data$kids_or_not=='have at least 1 kid'],var.equal =
FAL, alternative = 'greater')
summary(aov(have_kids_data$logIncome~factor(have_kids_data$total_kids)))
#linear regression model for all logIncome vs all variables
model1= lm(data$logIncome~data$Age)
summary(model1)
model2= lm(log(data$MntMeatProducts)~data$logIncome)
summary(model2)
plot(data$logIncome, log(data$MntMeatProducts))
abline( model2)
model2= lm(log(data$MntFruits+1)~data$logIncome)
summary(model2)
plot(data$logIncome, log(data$MntFruits+1))
abline( model2)
model2= lm(log(data$MntWines+1)~data$logIncome)
summary(model2)
plot(data$logIncome, log(data$MntWines+1))
abline( model2)
model2= lm(log(data$MntFishProducts+1)~data$logIncome)
summary(model2)
plot(data$logIncome, log(data$MntFishProducts+1))
abline( model2)
model2= lm(log(data$MntSweetProducts+1)~data$logIncome)
summary(model2)
plot(data$logIncome, log(data$MntSweetProducts+1))
abline( model2)
model2= lm(log(data$MntGoldProds+1)~data$logIncome)
summary(model2)
plot(data$logIncome, log(data$MntGoldProds+1))
abline( model2)
model2= lm(log(data$total_food_exp)~data$logIncome)
summary(model2)
plot(data$logIncome, log(data$total_food_exp))
abline( model2)
#variance and t-test for all different campaigns and acceptance
var.test(data$AcceptedCmp1,data$logIncome)
t.test(data$AcceptedCmp1,data$logIncome,var.equal = FALSE,alternative='less')
var.test(data$AcceptedCmp2,data$logIncome)
t.test(data$AcceptedCmp2,data$logIncome,var.equal = FALSE,alternative='less')
var.test(data$AcceptedCmp3,data$logIncome)
t.test(data$AcceptedCmp3,data$logIncome,var.equal = TRUE,alternative='less')
var.test(data$AcceptedCmp4,data$logIncome)
t.test(data$AcceptedCmp4,data$logIncome,var.equal = FALSE,alternative='less')
var.test(data$AcceptedCmp5,data$logIncome)
t.test(data$AcceptedCmp5,data$logIncome,var.equal = FALSE,alternative='less')
var.test(data$Response,data$logIncome)
t.test(data$Response,data$logIncome,var.equal = TRUE,alternative='less')
boxplot(data$logIncome~data$total_response,col='darkblue')
#anova test to test if the mean of logIncome of people who have different total number of response is the same
summary(aov(data$logIncome~data$total_response))
```

References

1. https://www.kaggle.com/jackdaoud/marketing-data