Marketing Campaign Analysis

Marshal Multani

March 2021

Introduction

In this project, a thorough analysis of a retail food company's marketing campaign is presented. It aims to understand the campaign's interactions with it's target audience, find business opportunities and insights, and to propose any data-driven actions to maximize the optimal results of the campaign and generate value to the company.
Products from 5 major categories are sold: wines, rare meat products, exotic fruits, specially prepared fish, and sweet products. These can further be divided into "gold" and "regular" products. The customers can order and acquire products through 3 sales channels: physical stores, catalogs, and the company's website.

Objective(s)

The key objectives are:
1. *EDA*: explore the data to understand the characteristic features of the respondents to the previous marketing campaigns by the company, to make better execution of the forthcoming one.
2. *Regression analysis*: build a regression model to identify significant factors that influence the number of store purchases by the respondents. Also, compare the performance of the previous campaigns by their respective geographical regions.
3. *Visualization*: plot and visualize the performances of the campaigns and individual products.

Dataset

The dataset contains socio-demographic and firmographic features of 2,240 customers. Additionally, it contains binary flags for those customers that responded to the campaign by buying a product.

```
df <- read.csv("marketing_data.csv")
dim(df)
```

```
## [1] 2240   28
```

```
head(df)
```

	ID <int>	Year_Birth <int>	Education <chr>	Marital_Status <chr>	Income <chr>	Kidhome <int>	Teenhome <int>	Dt_Customer <chr>	Recency <int>
1	1826	1970	Graduation	Divorced	$84,835.00	0	0	6/16/14	0
2	1	1961	Graduation	Single	$57,091.00	0	0	6/15/14	0
3	10476	1958	Graduation	Married	$67,267.00	0	1	5/13/14	0
4	1386	1967	Graduation	Together	$32,474.00	1	1	5/11/14	0
5	5371	1989	Graduation	Single	$21,474.00	1	0	4/8/14	0
6	7348	1958	PhD	Single	$71,691.00	0	0	3/17/14	0

6 rows | 1-10 of 29 columns

The "Income" column in the data frame is of "chr" data type containing commas and the Dollar ($) sign. To apply any arithmetic operation on it for the analysis, it needs to be *coerced* to a numeric data type by performing string replacement.
The "Dt_Customer" column is also of "chr" data type. This needs to be coerced to "Date" type.

```
df$Income <- str_replace_all(df$Income,"([$,])", "")
df$Income <- as.numeric(df$Income)
df$Dt_Customer <- as.Date(df$Dt_Customer,format = '%m/%d/%Y')
head(df)
```

	ID <int>	Year_Birth <int>	Education <chr>	Marital_Status <chr>	Income <dbl>	Kidhome <int>	Teenhome <int>	Dt_Customer <date>	Recency <int>
1	1826	1970	Graduation	Divorced	84835	0	0	0014-06-16	0
2	1	1961	Graduation	Single	57091	0	0	0014-06-15	0
3	10476	1958	Graduation	Married	67267	0	1	0014-05-13	0
4	1386	1967	Graduation	Together	32474	1	1	0014-05-11	0
5	5371	1989	Graduation	Single	21474	1	0	0014-04-08	0
6	7348	1958	PhD	Single	71691	0	0	0014-03-17	0

6 rows | 1-10 of 29 columns

Exploratory Data Anaalysis (EDA)

Duplicates, Outliers and Null Values

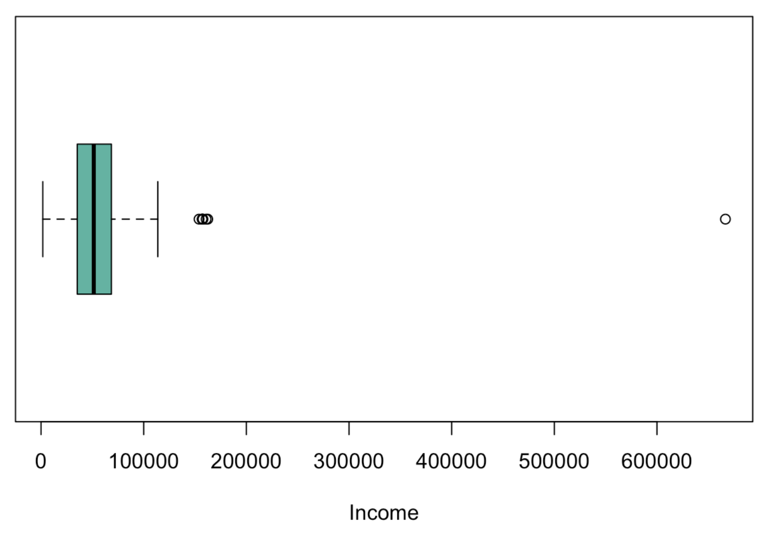
Count the number of duplicate values that may be in the data frame. Also, Identify features that contain NULL values. Then, using the distribution of any such feature can help to replace the NULL value with the *median* value to avoid the effects of outliers on the imputation value.

```
sapply(df, function(df) sum(is.na(df)))
```

```
##                ID          Year_Birth           Education      Marital_Status
##                 0                   0                   0                   0
##            Income             Kidhome            Teenhome         Dt_Customer
##                24                   0                   0                   0
##           Recency            MntWines           MntFruits     MntMeatProducts
##                 0                   0                   0                   0
##   MntFishProducts    MntSweetProducts        MntGoldProds   NumDealsPurchases
##                 0                   0                   0                   0
##   NumWebPurchases NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth
##                 0                   0                   0                   0
##      AcceptedCmp3        AcceptedCmp4        AcceptedCmp5        AcceptedCmp1
##                 0                   0                   0                   0
##      AcceptedCmp2            Response            Complain             Country
##                 0                   0                   0                   0
```

The 'Income' column contains 24 NULL values. It can be replaced by the median Income.

```
boxplot(df$Income , col="#69b3a2" , xlab="Income",horizontal = TRUE)
```




```
# Set the outlier, which may be a data entry error, in the Income column to N/A.
df <- df %>%
  mutate_at(vars(Income), na_if, 666666)
max(df$Income,na.rm = TRUE)
```

```
## [1] 162397
```

```
# Set the n/a entries in Income to the median income.
df$Income[is.na(df$Income)]<-median(df$Income,na.rm = TRUE)
# check for duplicate values
sum(anyDuplicated(df))
```

```
## [1] 0
```

There are no duplicate values in the data frame.

Feature Engineering

Review a list of variables in the data frame which can be combined to create new useful variables for the analysis.

```
str(df)
```

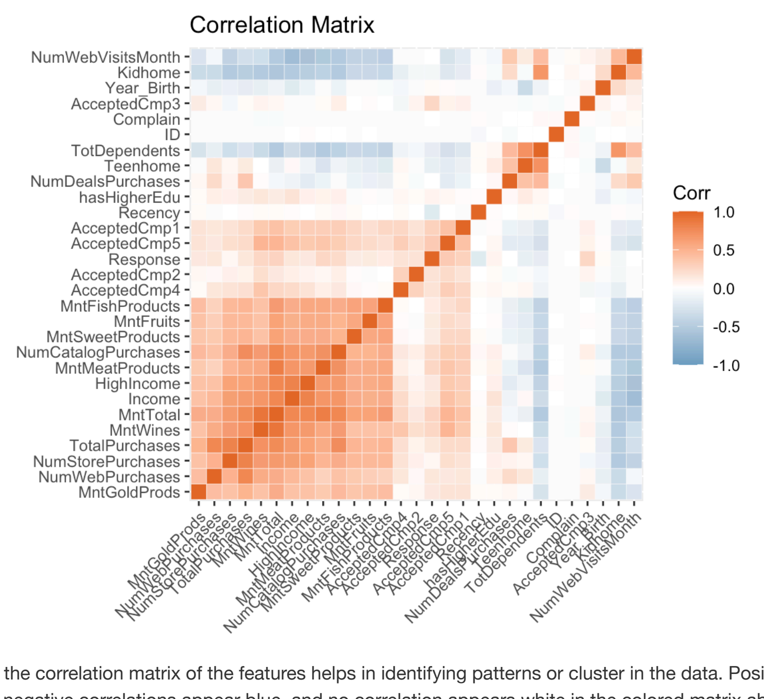
```
## 'data.frame':    2240 obs. of  28 variables:
##  $ ID                 : int  1826 1 10476 1386 5371 7348 4073 1991 4047 9477 ...
##  $ Year_Birth         : int  1970 1961 1958 1967 1989 1958 1954 1967 1954 1954 ...
##  $ Education          : chr  "Graduation" "Graduation" "Graduation" "Graduation" ...
##  $ Marital_Status     : chr  "Divorced" "Single" "Married" "Together" ...
##  $ Income             : num  84835 57091 67267 32474 21474 ...
##  $ Kidhome            : int  0 0 0 1 1 0 0 0 0 0 ...
##  $ Teenhome           : int  0 0 1 1 0 0 0 1 1 1 ...
##  $ Dt_Customer        : Date, format: "0014-06-16" "0014-06-15" ...
##  $ Recency            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ MntWines           : int  189 464 134 10 6 336 769 78 384 384 ...
##  $ MntFruits          : int  104 5 11 0 16 130 80 0 0 0 ...
##  $ MntMeatProducts    : int  379 64 59 1 24 411 252 11 102 102 ...
##  $ MntFishProducts    : int  111 7 15 0 11 240 15 0 21 21 ...
##  $ MntSweetProducts   : int  189 0 2 0 0 32 34 0 32 32 ...
##  $ MntGoldProds       : int  218 37 30 0 34 43 65 7 5 5 ...
##  $ NumDealsPurchases  : int  1 1 1 1 2 1 1 1 3 3 ...
##  $ NumWebPurchases    : int  4 7 3 1 3 4 10 2 6 6 ...
##  $ NumCatalogPurchases: int  4 3 2 0 1 7 10 1 2 2 ...
##  $ NumStorePurchases  : int  6 7 5 2 2 5 7 3 9 9 ...
##  $ NumWebVisitsMonth  : int  1 5 2 7 7 2 6 5 4 4 ...
##  $ AcceptedCmp3       : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ AcceptedCmp4       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp5       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp1       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ AcceptedCmp2       : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Response           : int  1 1 0 0 1 1 1 0 0 0 ...
##  $ Complain           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Country            : chr  "SP" "CA" "US" "AUS" ...
```

'Mnt…' variables can be summed to create a 'MntTotal', representing the total amount spent by a customer in all the years as a customer. 'Num…Purchases' variables can be summed to create a 'TotalPurchases' Variable. A 'TotDependents' variable can be created by adding together 'Kidhome' and 'Teenhome'. Customers with higher education and income of more than $60,000 can also be used to create two new variables.

```
# Total amount spent by far
df <- mutate(df, MntTotal = MntWines + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGold
Prods)

# Total number of purchases by far
df <- mutate(df, TotalPurchases = NumDealsPurchases + NumWebPurchases + NumCatalogPurchases + NumStorePurchases)

# Total number of dependents
df <- mutate(df, TotDependents = Kidhome + Teenhome)

# High income individuals
df <- mutate(df, HighIncome = Income > 60000)
df$HighIncome <- as.numeric(df$HighIncome)
# Customers with Higher Education
df <- mutate(df, hasHigherEdu = Education %in% c('Graduation', 'PhD', 'Master'))
df$hasHigherEdu <- as.numeric(df$hasHigherEdu)
```

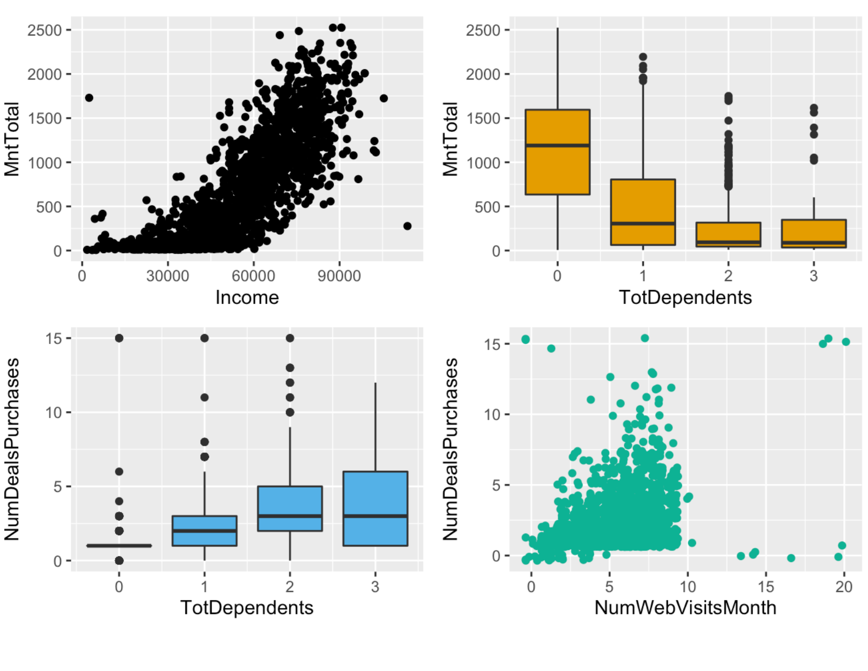
Plots and Patterns

```
# Subset of the data frame with only numeric variables
corrdf <- df[,sapply(df,is.numeric)]
# correlation matrix of the variables
corrMatrix <- cor(corrdf)
ggcorrplot(corrMatrix, title = "Correlation Matrix",
  hc.order = TRUE,
  outline.color = "white",
  ggtheme = ggplot2::theme_gray,
  colors = c("#6D9EC1", "white", "#E46726"))+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 9, hjust = 1),axis.text.y = element_text(size = 9))
```



Plotting the correlation matrix of the features helps in identifying patterns or cluster in the data. Positive correlations between features appear orange, negative correlations appear blue, and no correlation appears white in the colored matrix above.
Findings:
• Total Amount and Total Purchases:
- Total amount spent (MntTotal) and other 'Mnt' features, along with total purchases and other 'Purchases' features, are positively correlated with Income.
- Total number of purchases in all three categories of ways to purchase - store, web and catalog - are also positively correlated with Income and *negatively* correlated with the 'TotDependents'.
• NumDealsPurchases correlation
- 'NumDealsPurchases' is positively correlated with 'NumWebVisitsMonth', 'NumWebPurchases', and 'TotDependents'. *This suggests that customers with dependents prefer buying online with deals on products.*
• Anomalies:
- 'Income' seems to suggest a positive, but week, correlation with 'Response' to previous advertising campaigns.

```
plot1 <- ggplot(subset(df,Income<150000),aes(x=Income,y=MntTotal)) + geom_point()

plot2 <- ggplot(df,aes(x=TotDependents,y=MntTotal,group=TotDependents)) + geom_boxplot(fill = "#E69F00"
,show.legend = FALSE)

plot3 <- ggplot(df,aes(x=TotDependents,y=NumDealsPurchases,group=TotDependents)) + geom_boxplot(fill="#56B4E9",sh
ow.legend = FALSE)

plot4 <- ggplot(df,aes(x=NumWebVisitsMonth,y=NumDealsPurchases,group=NumWebPurchases)) + geom_point(position = 'j
itter',color='#16B295')

wrap_plots(plot1, plot2, plot3, plot4)
```



Regreesion Analysis

In order to gain further insight into what features explain the "response" variable in the dataset, regression analysis is to be performed for causal inference.
Consider the following linear probability model that includes the super set of variables in the dataset plus the error term 'u':

$$Response = \beta_0 + \beta_1 YearBirth + \beta_2 Education + \beta_3 MaritalStatus$$
$$+ \beta_4 Income + \ldots + \beta_{n-1} TotalPurchases + \beta_n TotalDependents + u$$

Conclusion