A3: Team Project

Team 12 – Shimon Takahashi, Paolo Musone, Khizer Sultan, Yash Maradia, Pablo Ramírez

Data Science with R & Data Analytics: Hult International Business School, MBA

Prof. Omar Romero – Hernandez and Thomas Kurnicki

November 14, 2021

Our aim in this document is to present a better understanding of the company's online sales, determine what variables most impact Online sales performance and finally design a set of actions to improve and support them as much as possible to the best of their potential.

We ran multiple regressions to see if there are any statistically significant variables that would be capable of explaining and predict online sales. Next, based on the statistic insight, we have developed some tailored business strategies related to that specific sales channel.

Firstly, we cleaned and massaged the data set for us to start analyzing the data through code; we created a "train" and "test" data set from the original numerical data that allowed us to run multiple linear regressions.
We eliminated some categorical variables that weren't statistically significant (variables with a p-value higher than 0.05) such as "Recency", "Complain", "Response", "Age", etc…
Afterwards, we normalized the data set to compare only the remaining significant variables in a linear regression, realizing that the most significant variables that explain Web purchases "Newebpruchase" are the following (in descendent order):

- "Income" → 0.41
- "NumWebVisitsMonth" → 0.22
- "MntWines" → 0.18
- "MntFishProducts" → 0.07
- "Teenhome" → 0.05

The only variable which predicts negatively the outcome was "Kidhome" → -0.05, which probably suggests that families with kids are more prompt to buy products physically in the store.

Our Model is capable to explain only **38.96%** of the data set (Multiple R-squared: 0.3896) and has a Medium/Low Adjust R-squared (0.3875). The p-value is far lower than 0.05 (2.2e-16) meaning that the model is, overall, statistically significant.

The results suggest that customers with high income are more prompt to purchase through the Web site store, especially if they are looking for Wines and Fish products. The Number of visits obviously positively predicts the amount of web purchases, but the level of prediction is only 0.22 meaning that there should be a deeper analysis to understand why only 22% of the customers who visit the website finalize the purchase. Most likely, we think the problem is related to the website's general navigation path, which might not be as user-friendly to customers as intended.

In short, knowing the importance of the Online sales channel for the company and the continuously growing e-Commerce industry, we'd suggest the following actions to be done in order to improve Online purchases:

- Consolidate the current "Online Customers" offering a rewards program for returning customers which will include discounts, preferable sale prices and reward points sales.
- Improve the Website's general navigation path into a short, "easy to purchase" or "1-click" purchase experience.
- Improve the Website's SEO (Search Engine Optimization) to guarantee a flawless first online purchase to new customers.

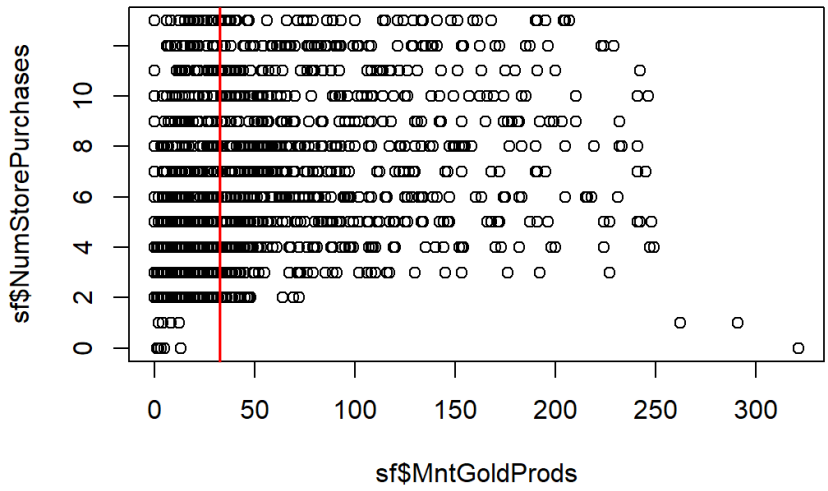
- Increase the Number of “Online Customers” by targeting customers with children (not Teenagers) with a Marketing campaign selling the benefits of the reward program and stating the “easiness” of online purchasing.

Going more into detail, in order to get the results for USA and Rest of the world total purchases, we ran a “for loop” with an “if” statement where we described our parameters, computing the sum of the USA purchases only VS the sum of the rest of the world (RoW) purchases. We then ran a linear regression to see if there is a linear relationship between USA and rest of the world.

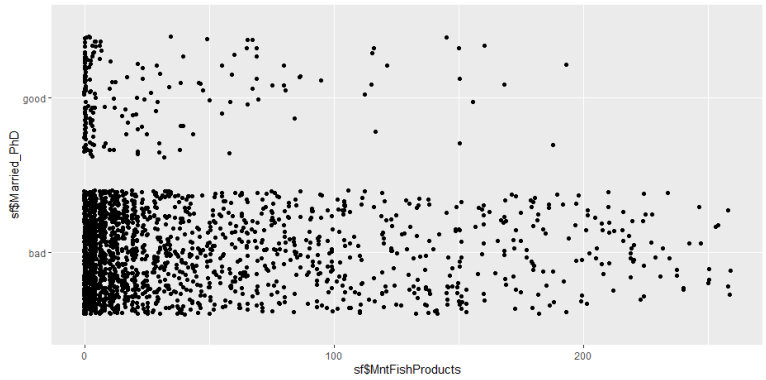
We got a **strong p-value of 2X10^-16** and an **R-squared of 0.1152**. Meaning that only 11.52% of the data from rest of the world can be explained because of the purchases that people make in USA, which is not very significant / impactful on our data. The strong P value just represents how strong does the people who purchase in the United States impacts that 11.52% of the number of store purchases in the rest of the world.

We do not agree with the supervisor’s claim. There is not enough evidence to confidently conclude that people who spend above the average amount on gold would have more store purchases. We ran a logistic and linear regression between “NumStorePurchases” VS a new *Dummy variable* that we created which assigned “1” to all observations above the mean spent in gold (USD $43.96, in the red line) and “0” for the observations below.

We got a **strong p-value of 2X10^-16** and an **R-squared of 0.1473**. Meaning that 14.73% of the number of store purchases can be explained because of the purchases that people who spend more than the mean in Gold, which is not very significant / impactful. The strong p-value just represents how strong does the people who spend more than the mean in Gold impacts that 14.74% of the number of store purchases.

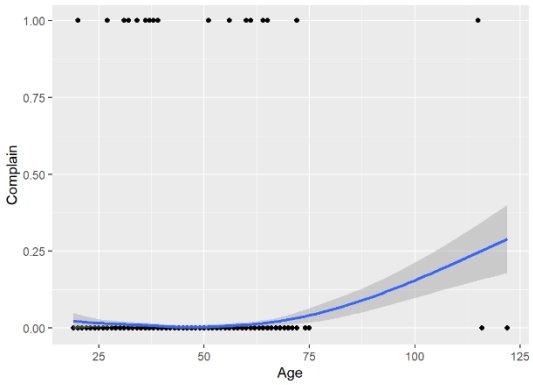


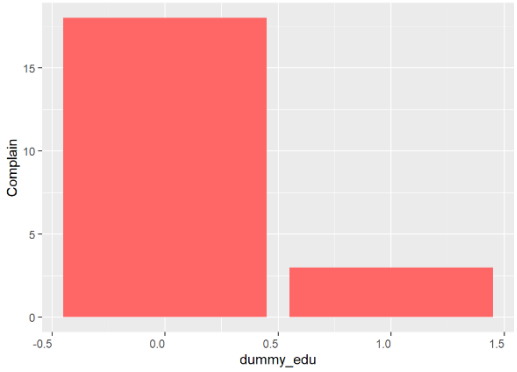
We can see a very significant relation between married and PhD candidate with amount spend on fish. As we can see from the graph the number of people who are married and has a PhD degree are “good” and the rest others are “bad”. The p-value we got was 0.005029 and R-squared of 0.003549, which means that only 0.3% of the data can be explained even though we have a strong p-value because there are less people who are “good” and spend amount on fish. The strong p-value shows how strongly do the people who are “good” and how much amount they spend on fish.



Analyzing the data, some interesting insights caught our attention that are definitely worth mentioning:

There are some indications within certain variables in the data regarding the **complains**. 0.9% of the total observations complained. More than half of those complaints were made by people aging between 25 to 38 years old. The rest of the complaints were from people aging from 50 to 75. Also, most of the people who complained were not highly educated (2n cycle, Basic and Graduation levels) the rest of the complaints were from people who had a Master's and Phd degrees.

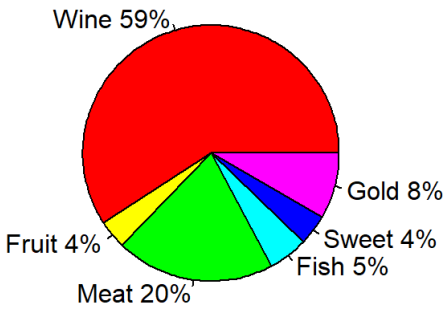




We tried to get some insights depending on people's marital status by splitting the categories; we assigned "1" to `Married` and `Together` and the rest categories were assigned to "0" … but the results in the regression returned that Marital status did not influence the `complain` variable to determine whether people complained or not.

We think that a household's total Income is distributed among family members, if there are any… Considering the relationship between Income and Marital status, we dug deeper into analyzing Teenagers. We chose this particular variable because `married` or `together` type of observations have the larger purchases, not to mention that the households that have Kids and teenagers make up 90% of our data. Furthermore, a Teenager's life stage is more expensive for a household's income than when kids; 41% of the total purchases in our data are from Teenagers, while 14% are from Kids. In short, a teenager's life stage is more expensive (significant) for a household's total income than the kids stage. So we analyzed customer's with Teenagers purchases and found out that Wine makes up for 59% of their purchases, followed by Meat with 20%.

Product consumption by families with Teenagers



Marketing is a very important pillar that thrives Revenues and psychologically pushes customers to purchase a service or product. Based on the data, we considered important to analyze the performance of the different campaigns implemented to determine which strategy is working or not. We found that campaign #4 is the most successful, followed by campaign #3 and #5 which had very similar performance and results.

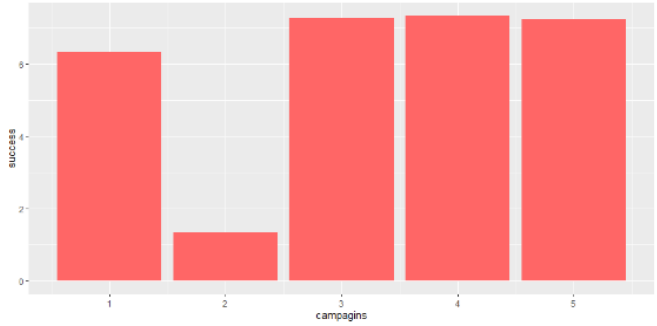
Campaign #1 – 6.33%

Campaign #2 – 1.33%

Campaign #3 – 7.27%

Campaign #4 – 7.32%

Campaign #5 – 7.23%



Code:

```
###################################

## Created by Team 12 - Shimon Takahashi, Paolo Musone, Khizer Sultan, Yash Maradia, Pablo Ramirez##

## MBAN HULT 2021 ##

## Date: 11.14.2021 ##

## Version 0.1 ##

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


#Downloading the data#

library(readxl)

datasets_marketing_campaign_SF_1_ <- read_excel("C:/Users/ramir/Desktop/MBA/HULT/Courses/MBAN/Data Science with R/datasets_marketing_campaign_SF (1).xlsx")

View(datasets_marketing_campaign_SF_1_)

sf <- datasets_marketing_campaign_SF_1_



install.packages("ggplot2")

install.packages("plotly")


library(ggplot2)

library(plotly)


 ######### Massaging the data ###########


#Exploring if there are any "NA" elements in the data set... 34 NA elements in the data set= 1,5% of the total data#

is.na(sf)

which(is.na(sf))
```

```
#Cancel the 34 NA elements from the dataset.
sf<-na.omit(sf)

#Replace each variable from Education column to a numeric variable#
sf$Education <- gsub("PhD","0", sf$Education)
sf$Education <- gsub("Master","1", sf$Education)
sf$Education <- gsub("Graduation", "2", sf$Education)
sf$Education <- gsub("2n Cycle", "3", sf$Education)
sf$Education <- gsub("Basic", "4", sf$Education)
sf$Education <- as.numeric(sf$Education)
print(sf$Education)

#Replace each variable from Marital_Status  column to a numeric variable#
sf$Marital_Status <- gsub("Widow", "0", sf$Marital_Status)
sf$Marital_Status <- gsub("Single", "1", sf$Marital_Status)
sf$Marital_Status <- gsub("Divorced", "2", sf$Marital_Status)
sf$Marital_Status <- gsub("Married", "3", sf$Marital_Status)
sf$Marital_Status <- gsub("Together", "4", sf$Marital_Status)
sf$Marital_Status <- gsub("Alone", "5", sf$Marital_Status)
sf$Marital_Status <- gsub("Absurd", "6", sf$Marital_Status)
sf$Marital_Status <- gsub("YOLO", "7", sf$Marital_Status)
sf$Marital_Status <- as.numeric(sf$Marital_Status)
print(sf$Marital_Status)

#Replace each variable from Country column to a numeric variable#
sf$Country <- gsub("AUS", "0", sf$Country)
sf$Country <- gsub("US", "1", sf$Country)
sf$Country <- gsub("CA", "2", sf$Country)
```

```
sf$Country <- gsub("GER", "3", sf$Country)
sf$Country <- gsub("IND", "4", sf$Country)
sf$Country <- gsub("ME", "5", sf$Country)
sf$Country <- gsub("SA", "6", sf$Country)
sf$Country <- gsub("SP", "7", sf$Country)
sf$Country<- as.numeric(sf$Country)
print(sf$Country)

#Modifying dates#

#Converting Dt_customer in as.date format
sf$Dt_Customer<- as.Date(sf$Dt_Customer)

#creating a "Today Variable" and converting in as.date format
sf$Year_2014 <- c()
for(i in 1:nrow(sf)){
 sf$Year_2014[1] <-("2014-12-31")
}
sf$Year_2014<-as.Date(sf$Year_2014)

#Exploring the Age variable#
summary(sf$Age)
hist(sf$Age)

# Creating new variables #

# Total purchases #
sf$Total_Purchases<-
sf$NumDealsPurchases+sf$NumWebPurchases+sf$NumCatalogPurchases+sf$NumStorePurchases
```

```
# Customers Age #
sf$Age<-2015 - sf$Year_Birth

# How many days are the customers enrolled in the Retailer subscription #
sf$days_of_Subscription<- (sf$Year_2014 - sf$Dt_Customer)

############ Part 1 - a: Variables that impact most "NumWebPurchases" and CMO recommendations ############
# Q1: What variables drive Web purchases #

# Creating new variables for regression purposes #
sf$binary_Education <- c()
for(i in 1: nrow(sf)){
  if(sf$NumWebPurchases[i]>0 ){
    sf$binary_Education[i]<-"1"
  }else{
    sf$binary_Education[i]<-"0"
  }#closing my if statement
}# closing the i loop

sf$binary_Education<-as.numeric(sf$binary_Education)

sf$binary_Marital_Status <- c()
for(i in 1: nrow(sf)){
  if(sf$Marital_Status[i]==3 & 4){
    sf$binary_Marital_Status[i] <- "1"
  }else{
    sf$binary_Marital_Status[i]<- "0"
```

```
  }#closing my if statement

}# closing the i loop

sf$binary_Marital_Status<-as.numeric(sf$binary_Marital_Status)


# Data Frame with all the numeric data#

Numeric_data<- as.data.frame(sf[c(
"binary_Education","binary_Marital_Status","Income","Kidhome","Teenhome","Recency","MntWines",

                    "MntFruits",
"MntMeatProducts","MntFishProducts","MntSweetProducts","MntGoldProds", "NumDealsPurchases",

                    "NumCatalogPurchases","NumStorePurchases"
,"NumWebVisitsMonth","NumWebPurchases","AcceptedCmp3", "AcceptedCmp4" , "AcceptedCmp5",

                    "AcceptedCmp1", "AcceptedCmp2","Complain" , "Response" ,"Total_Purchases",

                    "Age","Z_CostContact","Z_Revenue")])



#Train 80% of "Numeric_data" and test them with the outstanding 20% of the data#

train_index <- sample(1:nrow(Numeric_data),size= 0.8*nrow(Numeric_data))

sf_train <- Numeric_data[train_index,]

sf_test <- Numeric_data[-train_index,]


#Linear regression using the trained data#


#The regression is clearly over fitted with unsignificant variables, which will be removed later to perform our linear regression analysis

LR<-lm(NumWebPurchases~Income+Kidhome+Teenhome+Recency+MntWines+

      MntFruits+
MntMeatProducts+MntFishProducts+MntSweetProducts+MntGoldProds+NumDealsPurchases+

      NumCatalogPurchases+NumStorePurchases+NumWebVisitsMonth+AcceptedCmp3+
AcceptedCmp4+ AcceptedCmp5+

      AcceptedCmp1+ AcceptedCmp2+Complain+Response+Total_Purchases+

      Age+binary_Education+binary_Marital_Status, data=sf_train)
```

```
summary(LR)

LR_significant <- lm(NumWebPurchases~Income+Kidhome+Teenhome+MntWines+
                     MntFishProducts+NumWebVisitsMonth, data=sf_train)
summary(LR_significant)

#Data Normalization
min_max_norm <- function(Numeric_data) {
  ( Numeric_data- min(Numeric_data)) / (max(Numeric_data) - min(Numeric_data))
}

#apply Min-Max normalization to the data set
Normalized_data <- as.data.frame(lapply(Numeric_data, min_max_norm))

train_index1 <- sample(1:nrow(Normalized_data),size= 0.8*nrow(Normalized_data))
sf_train_norm <-Normalized_data[train_index,]
sf_test_norm <- Normalized_data[-train_index,]

#Linear Regression of Normalized data for comparing the estimated slopes among the significant
variables#
LR_significant_norm<-lm(NumWebPurchases~Income+Kidhome+Teenhome+MntWines+
                          MntFishProducts+NumWebVisitsMonth, data=sf_train_norm)
summary(LR_significant_norm)

############ Part 1 - b: The US VS RoW purchases

#for usa
sf$Usa_total_purchase <- c()
for(i in 1: nrow(sf)){
```

```
  if(sf$Country[i]==1 ){
    sf$Usa_total_purchase[i]<-sf$Total_Purchases[i]
  }else{
    sf$Usa_total_purchase[i]<- 0
  }#closing my if statement
}# closing the i loop
sum(sf$Usa_total_purchase)



#for rest of the world
sf$RoW_total_purchase <- c()
for(i in 1: nrow(sf)){

  if(sf$Country[i] !=1 ){
    sf$RoW_total_purchase[i]<-sf$Total_Purchases[i]
  }else{
    sf$RoW_total_purchase[i]<- 0
  }#closing my if statement
}# closing the i loop
sum(sf$RoW_total_purchase)

train_index <- sample(1:nrow(sf), size=0.8*nrow(sf))

SF_train_3 <- sf[train_index,]
SF_test_3 <- sf[-train_index,]

linear_model <- lm(Usa_total_purchase ~ RoW_total_purchase
                   ,data=SF_train_3)
```

```
summary(linear_model)

linear_model_2 <- lm(Usa_total_purchase ~ RoW_total_purchase
                     ,data=SF_test_3)

summary(linear_model_2)

cor(SF_train_3$Usa_total_purchase, SF_train_3$RoW_total_purchase)

Usa_perCapita <- sum(sf$Usa_total_purchase)/sum(sf$Country==1)
RoW_perCapita <- sum(sf$RoW_total_purchase)/sum(sf$Country==0)
Usa_perCapita
RoW_perCapita

############ Part 1 - c: Gold consumption VS In-store purcahsing #################

#Exploring the MntGoldProds data#
hist(sf$MntGoldProds)
#People spend an average of USD $43.96 in gold#
Gold_Mean <- mean(sf$MntGoldProds)
summary(sf$MntGoldProds)

#Creating a Dummy variable for people who spend in Gold above USD $43.96 = "1" and below = "0"#
sf$DummyGold_Above_Mean <- c()
for(i in 1: nrow(sf)){
  if(sf$MntGoldProds[i] > 43.96){
    sf$DummyGold_Above_Mean[i] <- 1
  }else{
```

```
   sf$DummyGold_Above_Mean[i] <- 0
  }#closing my if statement
}# closing the i loop
sum(sf$DummyGold_Above_Mean)

####### Logistic Regression: DummyGold_Above_Mean VS NumStorePurchases##########

DummyGold_Above_Mean_VS_StorePurchases_logreg <-
glm(DummyGold_Above_Mean~NumStorePurchases, data = sf, family = "binomial")
summary(DummyGold_Above_Mean_VS_StorePurchases_logreg)

####### Linear Regression: MntGoldProds VS NumStorePurchases ##########

Gold_VS_StorePurchases_linear <- lm(MntGoldProds~NumStorePurchases, data=sf)
summary(Gold_VS_StorePurchases_linear)

plot(x = sf$MntGoldProds, y = sf$NumStorePurchases)

######### Part 1 - d: Does Married and PhD customer's have any relationship with Fish consumption?
###########

sf$Married_PhD<- c()
for(i in 1: nrow(sf)){
  if(sf$Marital_Status[i] == 3 & sf$Education[i] == 0){
    sf$Married_PhD[i] <- "good"
  }else{
    sf$Married_PhD[i] <- "bad"
  }#closing my if statement
}# closing the i loop
```

```
linearMod <- lm(`MntFishProducts`~ .,
                data = sf)
summary(linearMod)


linearMod2<-  lm(`MntFishProducts`~ sf$Married_PhD,
                 data = sf)
summary(linearMod2)


library(ggplot2)


ggplot(data=sf, aes(x=sf$`MntFishProducts`, y=sf$Married_PhD)) + geom_jitter()


######### Part 1 - e: Analyzing complaints, Teenagers purchases and Campaign performance
############


# Q7 - Teenagers spend most on Wines... but Meat is the most consumed food #


#Assigning "1" and "0" to the "Teenhome" variable to distinguish observations that actually have 1 or 2
teens (doesn't matter how many) from those observations that don't have any teens#

sf$Teenhome <- gsub(2, 1, sf$Teenhome)

#Convert "HaveTeenhome" data from chr to num#

sf$Teenhome <- as.numeric(sf$Teenhome)

#Creating a new vector "Have_Teenhome" for observations that actually have teens#

Have_Teenhome <- which(sf$Teenhome == 1) #This vector is the one we use to plot it VS Mnt of Food#


#Filtering the addition of each Food Amount category by "Have_Teenhome == 1"

sf$Wine_HaveTeenhome <- c()

for(i in 1: nrow(sf)){
```

```
  if(sf$HaveTeenhome[i] == 1){
    sf$Wine_HaveTeenhome[i] <- sf$MntWines[i]
  }else{
    sf$Wine_HaveTeenhome[i] <- 0
  }#closing my if statement
}# closing the i loop
sum(sf$Wine_HaveTeenhome)

sf$Fruits_HaveTeenhome <- c()
for(i in 1: nrow(sf)){
  if(sf$HaveTeenhome[i] == 1){
    sf$Fruits_HaveTeenhome[i] <- sf$MntFruits[i]
  }else{
    sf$Fruits_HaveTeenhome[i] <- 0
  }#closing my if statement
}# closing the i loop
sum(sf$Fruits_HaveTeenhome)

sf$Meat_HaveTeenhome <- c()
for(i in 1: nrow(sf)){
  if(sf$HaveTeenhome[i] == 1){
    sf$Meat_HaveTeenhome[i] <- sf$MntMeatProducts[i]
  }else{
    sf$Meat_HaveTeenhome[i] <- 0
  }#closing my if statement
}# closing the i loop
sum(sf$Meat_HaveTeenhome)

sf$Fish_HaveTeenhome <- c()
```

```
for(i in 1: nrow(sf)){
  if(sf$HaveTeenhome[i] == 1){
    sf$Fish_HaveTeenhome[i] <- sf$MntFishProducts[i]
  }else{
    sf$Fish_HaveTeenhome[i] <- 0
  }#closing my if statement
}# closing the i loop
sum(sf$Fish_HaveTeenhome)

sf$Sweet_HaveTeenhome <- c()
for(i in 1: nrow(sf)){
  if(sf$HaveTeenhome[i] == 1){
    sf$Sweet_HaveTeenhome[i] <- sf$MntSweetProducts[i]
  }else{
    sf$Sweet_HaveTeenhome[i] <- 0
  }#closing my if statement
}# closing the i loop
sum(sf$Sweet_HaveTeenhome)

sf$Gold_HaveTeenhome <- c()
for(i in 1: nrow(sf)){
  if(sf$HaveTeenhome[i] == 1){
    sf$Gold_HaveTeenhome[i] <- sf$MntGoldProds[i]
  }else{
    sf$Gold_HaveTeenhome[i] <- 0
  }#closing my if statement
}# closing the i loop
sum(sf$Gold_HaveTeenhome)
```

```
#Pie Chart with Percentages for families with teens product consumption#
slices <- c(325143, 20159, 110127, 27418, 21414, 45682)
lbls <- c("Wine", "Fruit", "Meat", "Fish", "Sweet", "Gold")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
   main="Product consumption by families with Teenagers")

# Q9: People without high education and ranging from 25 to 38 years old file the most complaints.#

# Creating dummy variable for education level... Phd & Masters = 1, else 0.
for (i in 1:nrow(sf)){
  if (sf$Education[i] <= 1){
    sf$dummy_edu[i] <- "1"
  }else{
    sf$dummy_edu[i] <- "0"
  }
}
View(sf$dummy_edu)
sf$dummy_edu <- as.numeric(sf$dummy_edu)

### Creating a dummy variable whether they are married & together = 1, else 0.
for (i in 1:nrow(sf)){
  if (sf$Marital_Status[i] == 3 & 4){
    sf$dummy_marital[i] <- "1"
  }else{
    sf$dummy_marital[i] <- "0"
  }
```

```
}

View(sf$dummy_marital)

sf$dummy_marital <- as.numeric(sf$dummy_marital)


# Logistic Regression: Complain VS Age, dummy_edu adn marital_edu #

sf_logit <- glm(Complain ~  Age + dummy_edu + dummy_marital, data = sf, family = "binomial")

summary(sf_logit)


Complain_VS_dummy_edu_linear <- lm(Complain~Age, data = sf_train)

summary(Complain_VS_dummy_edu_linear)


plot(x = sf$Age, y = sf$Complain)


Scatter_Complain_VS_Age <- ggplot(data=sf, aes(x = Age, y = Complain)) + geom_point() +
scale_color_manual(values = c("#FF5733","#FF33E9")) + geom_smooth()

Scatter_Complain_VS_Age


ggplot(data = sf, aes(x = `dummy_edu`, y= `Complain`))+

  geom_bar(stat = "identity", fill = "#FF6666")+

  theme(axis.text.x = element_text(angle = 0, hjust = 1))


#Q8 - Campaign performance#


#Most succesfull Campaign is determined by the sum of successes in each Campaign... "AcceptedCmp4" is the most succesfull with 164 successes (purchases)#

Campaign1_successes <- sum(sf$AcceptedCmp1)

Campaign2_successes <- sum(sf$AcceptedCmp2)

Campaign3_successes <- sum(sf$AcceptedCmp3)

Campaign4_successes <- sum(sf$AcceptedCmp4)
```

```
Campaign5_successes <- sum(sf$AcceptedCmp5)


#Most efficient Campaign is determined by dividing the amount of successes by total observations or attempts#

sum(sf$AcceptedCmp1)/2240*100

sum(sf$AcceptedCmp2)/2240*100

sum(sf$AcceptedCmp3)/2240*100

sum(sf$AcceptedCmp4)/2240*100

sum(sf$AcceptedCmp5)/2240*100


success <- c(sum(sf$AcceptedCmp1)/2240*100,
             sum(sf$AcceptedCmp2)/2240*100,
             sum(sf$AcceptedCmp3)/2240*100,
             sum(sf$AcceptedCmp4)/2240*100,
             sum(sf$AcceptedCmp5)/2240*100)


campagins <- c(1,2,3,4,5)


as.data.frame(success)

as.data.frame(campagins)


new_df <- cbind.data.frame(success,campagins)


ggplot(data= new_df, aes(x= campagins,y = success))+
  geom_bar(stat = "identity", fill = "#FF6666")
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```