R Notebook

In this file I will analyze the data from a business that once a year creates a marketing campaign to try to seel 5 different offers to its clients.

The business wants to know how to maximize the number of offers that are accepted by its clients. Because of that, I will analyze how the different features of the dataframe increase or decrease the possibilities that the customers accept those offers.

```
data=read.csv("marketing_campaign.csv",sep=";")
length(data)
```

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

CLEANING THE DATA

This data frame has a total of 2240 registers and it has 24 NA's.

Considering that the lines containing NA's represent just 1% of the information, I will delete those lines. I want to poit out that if there were not that many registers or more NA's I would have searcher for an other way to manage the missing values.

In this case, since all the NA are in the "Income" feature, another good way to manage the NA's would be searching for the mean of incomes and substitute the NA's for the resulting number.

```
dim(data)
```

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

```
length(data[is.na(data)])
```

```
## [1] 24
```

```
apply(data, 2, function(x) any(is.na(x)))
```

```
##               ï..ID          Year_Birth           Education      Marital_Status
##               FALSE               FALSE               FALSE               FALSE
##              Income             Kidhome            Teenhome         Dt_Customer
##                TRUE               FALSE               FALSE               FALSE
##             Recency            MntWines           MntFruits     MntMeatProducts
##               FALSE               FALSE               FALSE               FALSE
##     MntFishProducts    MntSweetProducts        MntGoldProds   NumDealsPurchases
##               FALSE               FALSE               FALSE               FALSE
##     NumWebPurchases NumCatalogPurchases   NumStorePurchases   NumWebVisitsMonth
##               FALSE               FALSE               FALSE               FALSE
##        AcceptedCmp3        AcceptedCmp4        AcceptedCmp5        AcceptedCmp1
##               FALSE               FALSE               FALSE               FALSE
##        AcceptedCmp2            Complain       Z_CostContact           Z_Revenue
##               FALSE               FALSE               FALSE               FALSE
```

```
nrow(data[is.na(data$Income),])
```

```
## [1] 24
```

```
data=data[complete.cases(data), ]
```

ANALYZING THE DATA

The business wants to know how to increase the possibilities that a customer accepts an offer. In the dataframe, we can find a reference to 5 different offers in the following features:(data$AcceptedCmp1$, data$AcceptedCmp2, data$AcceptedCmp3$, data$AcceptedCmp4, data$AcceptedCmp5) There is a 0 in the columns were the offers were not accepted and 1 to those that were succesfully accepted.

I will now create a new feature to see how many offers has every client accepted. Since there is a 1 in the accepted offers and a 0 to the non accepted ones,we can just add all the columns. For example, if the new column has a 5 it would mean that the client has accepted all the offers, if it has a 3, that means that the client has accepted just 3 out of 5 offers.

```
library(ggplot2)
num.accepted=data.frame("num accepted"=data$AcceptedCmp1+data$AcceptedCmp2+data$AcceptedCmp3+data$Accep
```

We will start the analysis, by learning about the clients.

One of the most important factors when it comes to determine how a certain person will act is the age of the individual. That is why, we will firt analyze this variable.

```
table(data$Year_Birth)
```

```
##
## 1893 1899 1900 1940 1941 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953
##    1    1    1    1    1    6    7    8   16   16   21   30   29   42   52   35
## 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969
##   49   48   55   41   52   50   49   35   44   44   41   74   50   44   51   70
## 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985
##   75   86   78   72   69   83   89   52   76   53   39   38   44   41   38   32
## 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996
##   41   27   29   29   18   15   13    5    3    5    2
```

The first thing we can notice about this dataset is that a lot of the registers are from people born between 1970 and 1980. We will now examine the frequencies of the year of birth of the different participants, and we weill just select this 5 biggest groups.

```
sort(table(data$Year_Birth),decreasing=T)[1:5]
```

```
##
## 1976 1971 1975 1972 1978
##   89   86   83   78   76
```

Now, we will complete the information with a histogram to see the distribution of the ages. I asume that this business has that many data from people born in the 70's because this is their main customer target. However, I should also highlight the important number of people born in the 50's and 60's.

```
cuts=data.frame(cuts=(cut(data$Year_Birth,breaks = c(1850,1940,1950,1960,1970,1980,1990,2000),labels = 

library(dplyr)
```

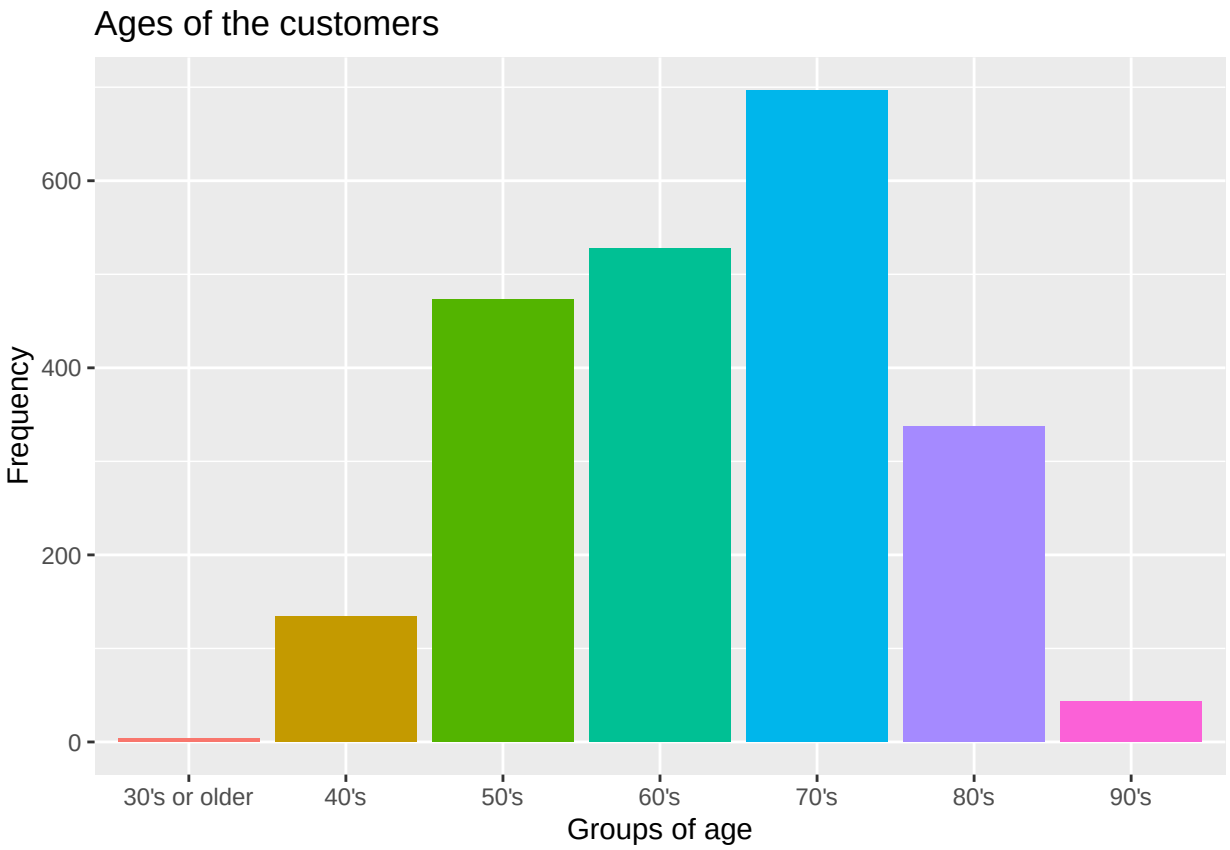
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
```

```
##
##     intersect, setdiff, setequal, union
```

```
table(cuts) %>%
  as.data.frame() %>%
  ggplot(aes(x = cuts, y = Freq, fill=cuts)) +
  geom_col()+
  labs(x="Groups of age", y="Frequency",  title="Ages of the customers"    )+
  theme(legend.position = "none")
```



We will continue including the number of offers accepted to the dataframe "cuts" and watching how many offers did the customers from different age groups accepted.

The resulting table shows which percentage of customers of every decade accepted 0,1,2,3 or 4 offers.

For example, the 68.65% of people from the 40's did not accept any offer, 20.89 % accepted 1, 8.95% accepted 2 and 1.49 % accepted 3.

If, as I said, the main target of the business is the people from the 70's, then, I would recommend to change it, because this is the decade that shows a smaller percentage of acceptance. 83.78% of them didn't accept a single offer and less than 5% accepted 2 or more offers (2.43%+2.29%+0.14%).

```
cuts=cbind(cuts,num.accepted)
prop.table(table(cuts$cuts,cuts$num.accepted),1)
```

```
##
##                          0           1           2           3           4
##   30's or older 0.750000000 0.250000000 0.000000000 0.000000000 0.000000000
##   40's          0.686567164 0.208955224 0.089552239 0.014925373 0.000000000
```

```
##   50's          0.792811839 0.152219873 0.027484144 0.016913319 0.010570825
##   60's          0.787878788 0.164772727 0.030303030 0.015151515 0.001893939
##   70's          0.837876614 0.113342898 0.024390244 0.022955524 0.001434720
##   80's          0.765578635 0.142433234 0.056379822 0.026706231 0.008902077
##   90's          0.674418605 0.186046512 0.093023256 0.023255814 0.023255814
```

The groups that show the best percentages for the business are the customers born in the 90's and the ones born in the 40's.

If we make a comparison of the mean incomes by groups, it is possible to see that the wealthiest groups are the 40's and the 90's, and thus, this are again the ones with the most potential.

Considering that the mean income for the persons who were born during the 70's is the second lowest, maybe, the high percentage of rejection of the offers is due to a reduced income. In order to learn more about that, we will examine if the income has an important effect on whether the person accepts the offers or not. I assume that the costumers with less money will be more conservative when it comes to decide whether to accept or decline an offer that is not expected.

```
cuts$Income=data$Income
aggregate(Income~cuts, data = cuts, FUN = mean)
```

```
##            cuts   Income
## 1 30's or older 57873.75
## 2          40's 61129.13
## 3          50's 56777.65
## 4          60's 53359.80
## 5          70's 49689.69
## 6          80's 45117.66
## 7          90's 57882.56
```

We will now create a table with the number of offers that the customers have accepted and whether their income is more than 50000$ or not. In the following table, FALSE will mean that they earn less than 50000$ and TRUE will mean that they earn more than that.

As we can see, 90% of those who earn less than 50000$ have rejected all the offers and if they have accepted any, they mostly accepted just one. The costumers who earn more than 50000$ show a different behavior. More than 30% of the clients who have an income greater than 50000$ have accepted at least one offer, almost 1% of these customers accepted up to 4 offers.

```
prop.table(table(cuts$Income>50000,cuts$num.accepted),1)
```

```
##
##                    0           1           2           3           4
##   FALSE 0.9000000000 0.0943396226 0.0047169811 0.0009433962 0.0000000000
##   TRUE  0.6946366782 0.1929065744 0.0657439446 0.0371972318 0.0095155709
```

Since the income is such an important factor, and it is normally related to the level of studies, maybe the education will be an other important factor. We will now create a boxplot to see the distribution of the incomes based on the education.

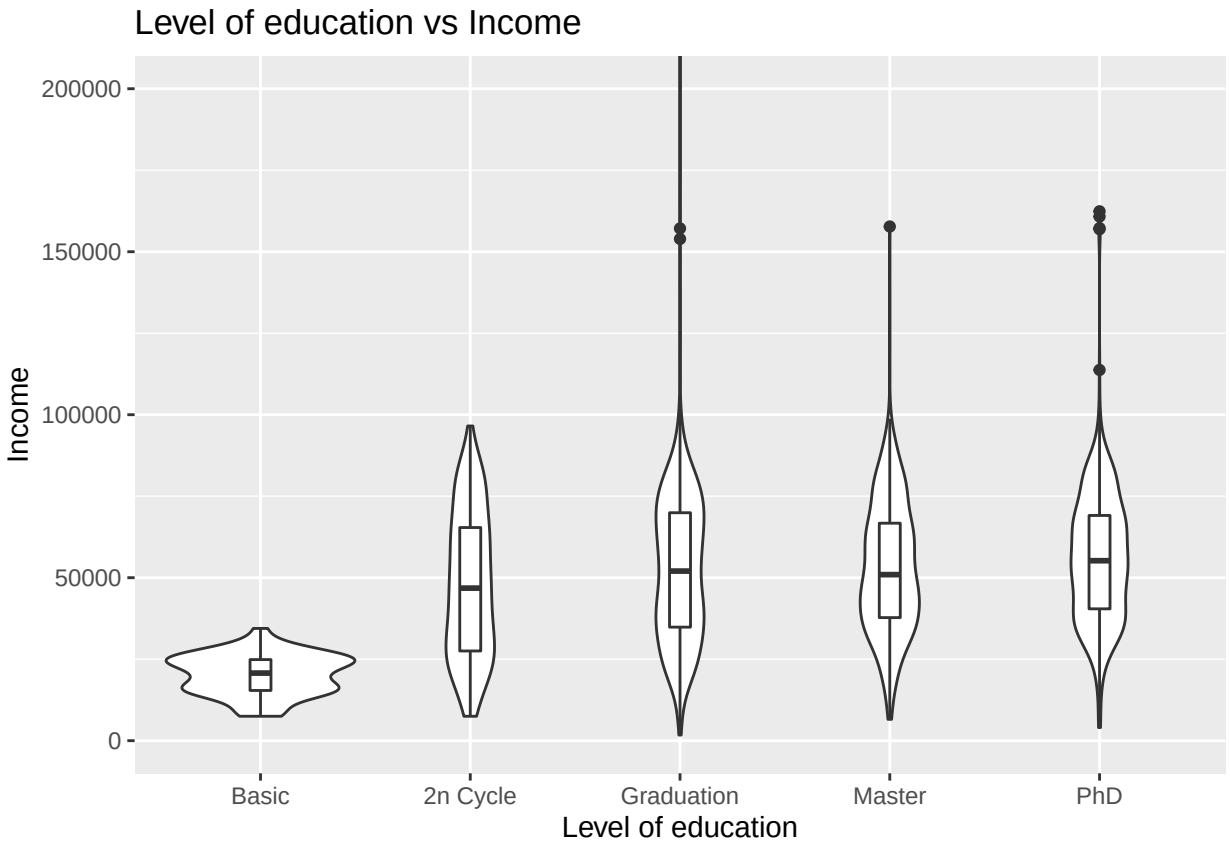
I presume that it will show that the people with lower education levels earn less and if this is the case, I will conclude that most of the customers with less studies, show less inclination to accept the offer.

As we can see in the plot, the people with Basic education earns less than the clients with more studies.

```
data$Education=factor(data$Education, levels=c("Basic","2n Cycle","Graduation","Master","PhD"))

ggplot()+
  geom_violin(aes(data$Education,data$Income))+
  geom_boxplot(aes(data$Education,data$Income),width =0.1)+
```

```
  coord_cartesian(ylim = c(0, 200000))+
  labs(x="Level of education", y="Income", title="Level of education vs Income")
```



Now that we know the relation between the studies and the earnings, we will now proceed to see if it is also true that most of the people with basic education refused the offer.

As we expected, the clients that studied the most, are more likely to accept the offer of the supermarket.

We can conclude then, that the level of studies and the income share great similarities, and that they are a great factor to determine how the client will act.

```
prop.table(table(data$Education,cuts$num.accepted),1)
```

```
##
##                          0           1           2           3           4
##   Basic      0.888888889 0.111111111 0.000000000 0.000000000 0.000000000
##   2n Cycle   0.815000000 0.135000000 0.035000000 0.015000000 0.000000000
##   Graduation 0.795698925 0.142473118 0.031362007 0.023297491 0.007168459
##   Master     0.797260274 0.147945205 0.035616438 0.016438356 0.002739726
##   PhD        0.762993763 0.160083160 0.054054054 0.018711019 0.004158004
```

Another factor that can also be important is whether the clients have Kids or not, and if they have Teenagers at home.

We will start by introducing the information in the dataset “cuts”, and making a bar plot.

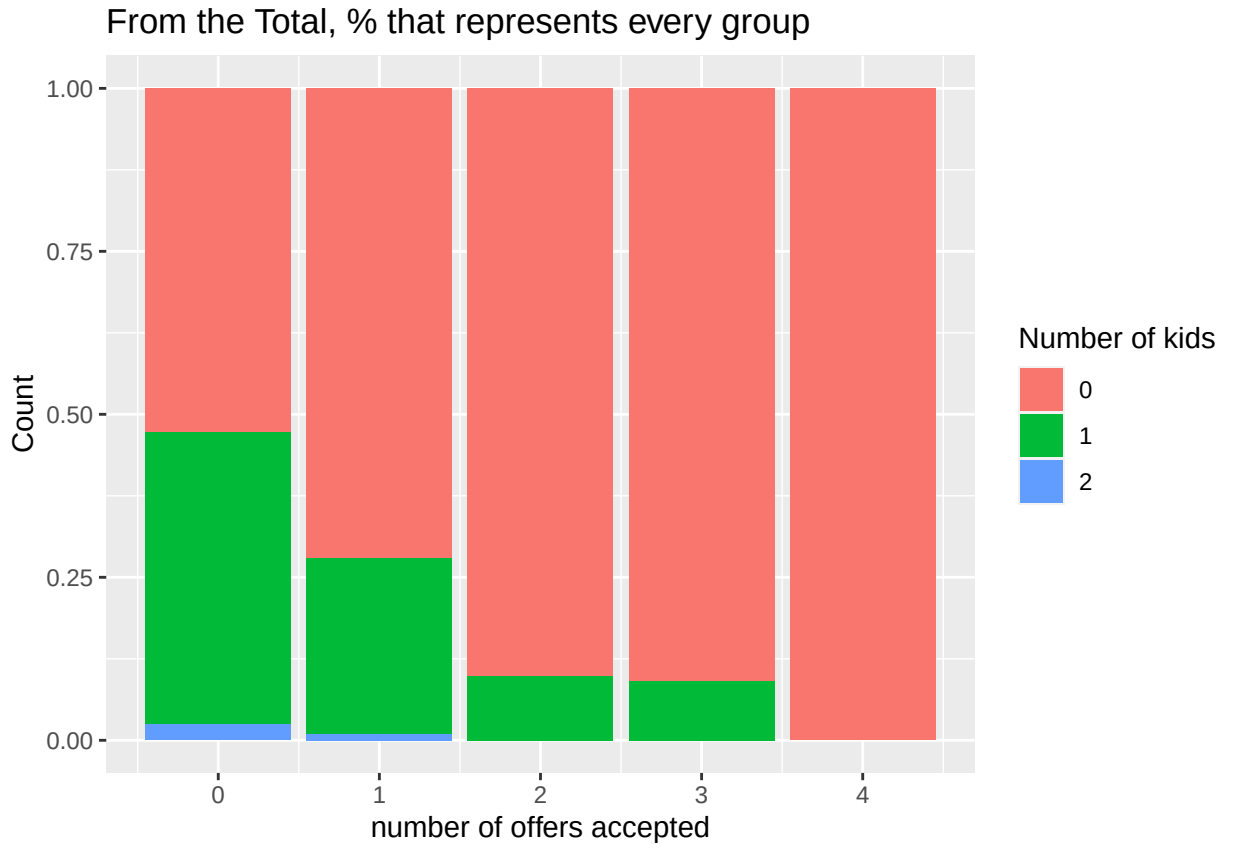
The plot that we have made, has the position=“fill”. That means that it shows the percentage of clients that did a certain thing.

For example, as we can see, the whole bar of 4 offers accepted, is filled with the color of “0 kids”.

That means that 100% of the clients who took 4 offers had 0 kids. The table that we made after the plot, explains the same information, but showing the exact numbers.

```
cuts$Kid=as.factor(data$Kidhome)

ggplot()+
  geom_bar(aes(cuts$num.accepted,fill=cuts$Kid),position="fill")+
  labs(x="number of offers accepted", y="Count", title="From the Total, % that represents every group")+
  scale_fill_discrete(name = "Number of kids")
```




```
prop.table(table(cuts$Kid,cuts$num.accepted),2)
```

```
##
##                0           1           2           3           4
##   0 0.527034718 0.721362229 0.901234568 0.909090909 1.000000000
##   1 0.448491747 0.269349845 0.098765432 0.090909091 0.000000000
##   2 0.024473534 0.009287926 0.000000000 0.000000000 0.000000000
```

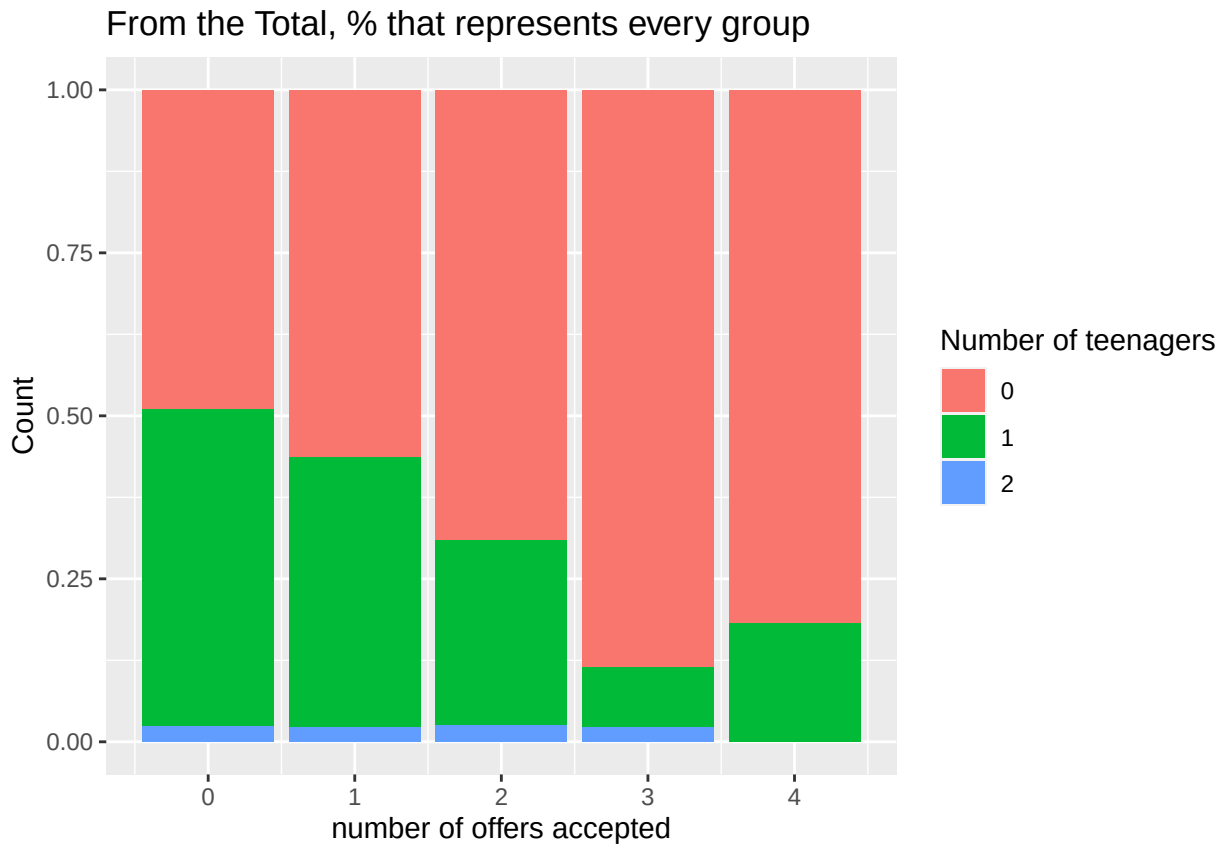
If we apply the same procedure with the number of teenagers that the customer has we will be able to see a similar trend, but in a more moderated way.

We can conclude the analysis of kids and teenagers, saying that this seems to be a very influential factor, especially if the consumers has kids.

```
cuts$teen=as.factor(data$Teenhome)

ggplot()+
  geom_bar(aes(cuts$num.accepted,fill=cuts$teen),position="fill")+
  labs(x="number of offers accepted", y="Count", title="From the Total, % that represents every group")+
```

```
  scale_fill_discrete(name = "Number of teenagers")
```




```
prop.table(table(cuts$teen,cuts$num.accepted),2)
```

```
##
##              0          1          2          3          4
##   0 0.49003984 0.56346749 0.69135802 0.88636364 0.81818182
##   1 0.48662493 0.41486068 0.28395062 0.09090909 0.18181818
##   2 0.02333523 0.02167183 0.02469136 0.02272727 0.00000000
```

```
0.098765432 +0.090909091
```

```
## [1] 0.1896745
```

Now we will take a closer look at the kind of food that the clients buy and if these habits make a difference when it comes to take decisions regarding the offers.

We will start creating a new dataframe named “food” with the data that makes reference to food and drinks.

We will also include the number of offers that the clients have accepted, we will use a boxplot to examine it better.

This next plot is a very interesting one, there are a couple of important ideas we can obtain from it:

1-Generally, the persons who buy more, tend to be more open to the offers. Even if it is more difficult to detect a client willing to accept 4 offers, to focus on the clients whose consume is higher than the 95th percentile of its respective line of products, increases greatly the possibilities to contact consumers who will accept 2 or 3 offers.

2-The consumption of wine is a very good indication of how interested a certain customer could be in the

offers.

```
food=data.frame("Wine"=data$MntWines,"Fish"=data$MntFishProducts,"Fruits"=data$MntFruits,"Meat"=data$Mn
food$offer=factor(cuts$num.accepted)

library(reshape)
```

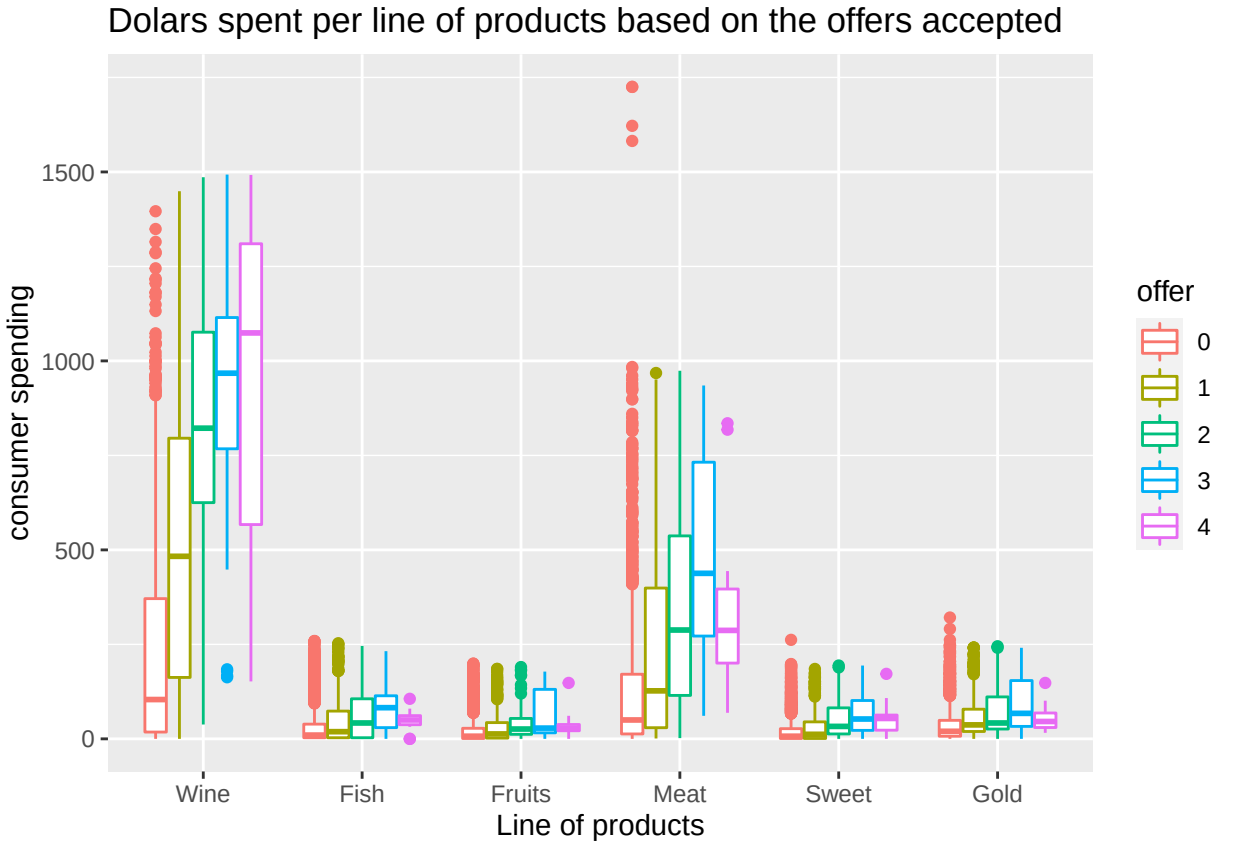
```
##
## Attaching package: 'reshape'
```

```
## The following object is masked from 'package:dplyr':
##
##     rename
```

```
meltData <- melt(food)
```

```
## Using offer as id variables
```

```
ggplot(meltData)+
  geom_boxplot(aes(factor(variable), value, color=offer))+
  labs(x="Line of products", y="consumer spending", title="Dolars spent per line of products based on tl
  scale_fill_discrete(name = "Offers accepted")
```



We have already determined that in order to know if a client will be a good target for the offers, we need to focus on those that have a consume of products higher than a certain level. We will now search for the minimum level of consume that a client must have, to be considered as a good target.

The following function will determine the different quartiles. The business should target any person who buys products from ANY line with a total cost that overpasses the top 5 %.

```
meltData=data.frame(meltData)
library(dplyr)


summary(food)
```

```
##       Wine              Fish            Fruits            Meat
##  Min.   :   0.0   Min.   :  0.00   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:  24.0   1st Qu.:  3.00   1st Qu.:  2.00   1st Qu.:  16.0
##  Median : 174.5   Median : 12.00   Median :  8.00   Median :  68.0
##  Mean   : 305.1   Mean   : 37.64   Mean   : 26.36   Mean   : 167.0
##  3rd Qu.: 505.0   3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 232.2
##  Max.   :1493.0   Max.   :259.00   Max.   :199.00   Max.   :1725.0
##      Sweet             Gold        offer
##  Min.   :  0.00   Min.   :  0.00   0:1757
##  1st Qu.:  1.00   1st Qu.:  9.00   1: 323
##  Median :  8.00   Median : 24.50   2:  81
##  Mean   : 27.03   Mean   : 43.97   3:  44
##  3rd Qu.: 33.00   3rd Qu.: 56.00   4:  11
##  Max.   :262.00   Max.   :321.00
```

```
meltData%>%
  group_by(variable)%>%
  summarise("10%"=quantile(value, 0.95, na.rm=TRUE)) #these are the values that need to overpass the cl
```

```
## # A tibble: 6 x 2
##   variable `10%`
##   <fct>    <dbl>
## 1 Wine     1000.
## 2 Fish      169
## 3 Fruits    122.
## 4 Meat      688.
## 5 Sweet     125.
## 6 Gold      165.
```

Wine 1000 Fish 169 Fruits 122 Meat 688 Sweet 125 Gold 165

```
meltData=data.frame(meltData)

A=meltData[(meltData$variable=="Wine" & meltData$value>1000)|(meltData$variable=="Fish" & meltData$valu

table(A$offer!=0)
```

```
##
## FALSE  TRUE
##   348   317
```

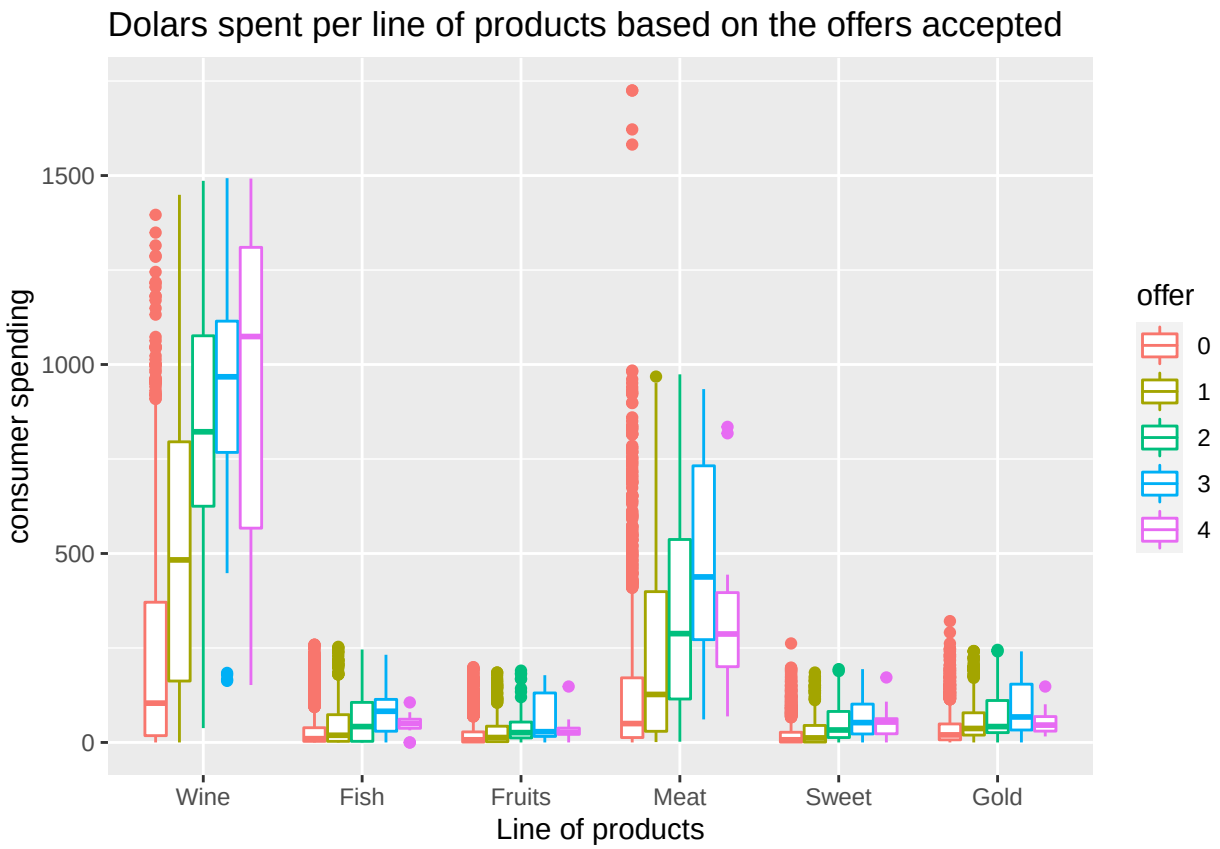
```
317/(348+317) #47.66% would accept at least 1 offer.
```

```
## [1] 0.4766917
```

As we can see in the last table, if the business targets just the people in the upper 5% of consumption, the possibilities of reaching a client that accepts at least one offer is almost 50%.

Regarding the wine consumption distinction of the clients that I said in the second point, as we can see in the following plot.

```
ggplot(meltData)+
  geom_boxplot(aes(factor(variable), value, color=offer))+
  labs(x="Line of products", y="consumer spending", title="Dolars spent per line of products based on tl
  scale_fill_discrete(name = "Offers accepted")
```



If from the wine line of products, we take the consumers that have a consumed a wine with a total cost that overpasses the median cost of the group of consumers that accepted 2 offers, the possibilities to find a client who REFUSES the offer decreases greatly

As we can see from all the consumers who have payed more than 822$ for wine during the last 2 years, 70 would say no to the offer and 153 would accept at least one offer.

```
food%>%
  select(Wine,offer)%>%
  group_by(offer)%>%
  summarise("summary"=median(Wine))
```

```
## # A tibble: 5 x 2
##   offer summary
##   <fct>   <dbl>
## 1 0         104
## 2 1         483
## 3 2         822
## 4 3         968.
## 5 4        1074
```

```
table(food[food$Wine>822,"offer"]!=0)
```

```
##
## FALSE  TRUE
##    70   153
```

```
153/(70+153)#68.60% would accept at least one offer.
```

```
## [1] 0.6860987
```

We now know what the customers buy, now is time to understand how they buy. According to many marketing studies this is one of the most important aspects to consider when choosing how to address the consumer.

In order to be able to visualize all the information regarding how the consumer buys, I will put together diferent information in the same plot.

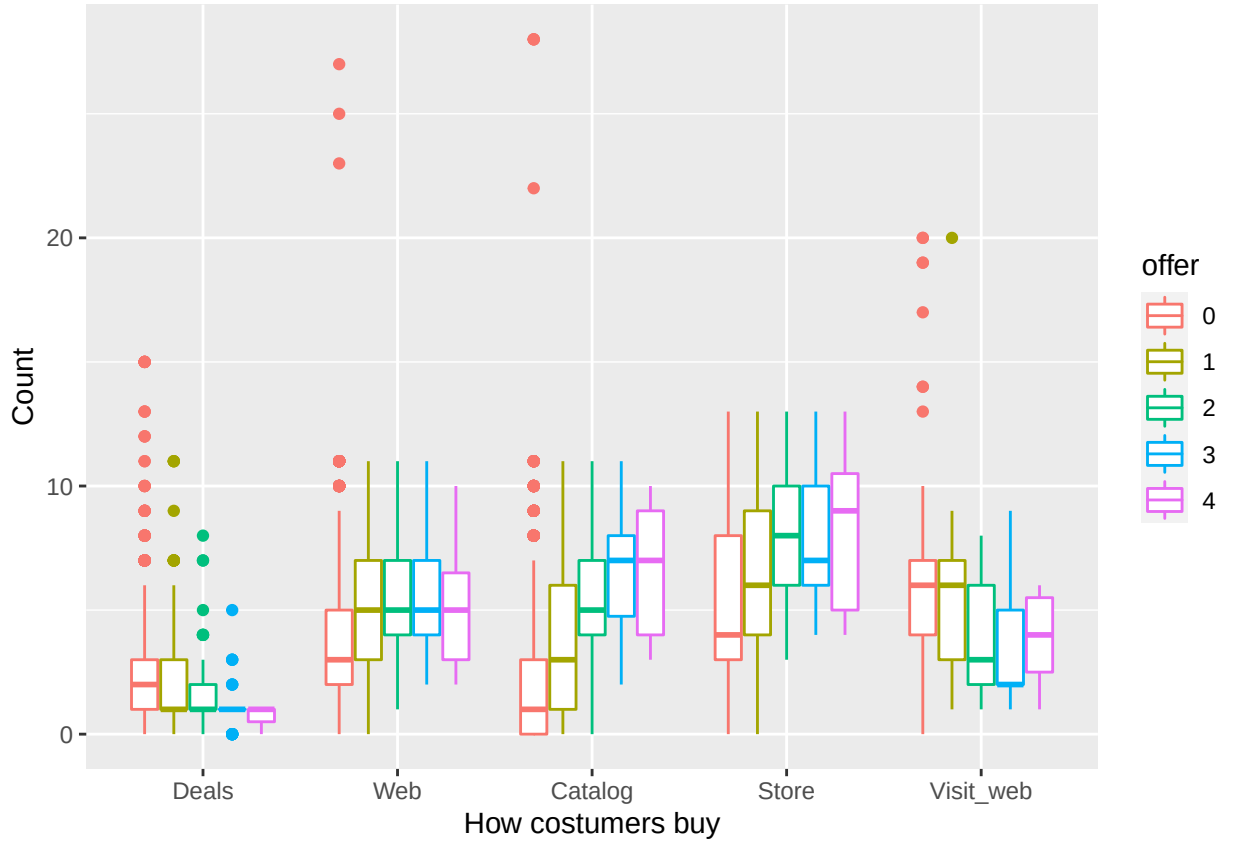
In the following graph, startin from the left side, we have, the number of "Deals" that the consumer has bought during the last month. Then we can find the number of times the client has bought something trough the web. The one that follows is the number of times the client boutgh something trough the catalog. "Store" means the number of times that the client bought something in the phisical shops. Finally the "Visit_web" show how many times a client has visited the webpage of the supermarket.

```
Buy=data.frame("Deals"=data$NumDealsPurchases,"Web"=data$NumWebPurchases,"Catalog"=data$NumCatalogPurcha
Buy$offer=factor(cuts$num.accepted)

meltData <- melt(Buy)
```

```
## Using offer as id variables
```

```
ggplot(meltData)+
  geom_boxplot(aes(factor(variable), value,color=offer))+
  labs(x="How costumers buy", y="Count")+
  scale_fill_discrete(name = "Offers accepted")
```



From this last plot it is possible to obtain 3 interesting informations:

1-Eventhoug the behavior that the clients have depend on how they buy (trough web, using catalog or in the physical shop) if what we want to know is how do we maximize the possibilities to contact a costumer who at least accepts one offer, then the output that we get from this plot is the same as the last one. The more times a client buys, the more likely he/she is to accept the offers, no matter how they buy.

2- Another important conclusion that we can take out of this chart is that according to the data, the consumers who constantly visit the webpage tend to refuse the offers according to the data, the customers who visited the webpage more than 9 times during the last month will reject the offer in almost every case.

```
prop.table(table(Buy$offer,Buy$Visit_web),2)
```

```
##
##                0           1           2           3           4           5
##   0 1.000000000 0.653333333 0.666666667 0.738916256 0.797235023 0.845878136
##   1 0.000000000 0.193333333 0.189054726 0.157635468 0.129032258 0.114695341
##   2 0.000000000 0.100000000 0.049751244 0.088669951 0.041474654 0.010752688
##   3 0.000000000 0.040000000 0.089552239 0.009852217 0.023041475 0.021505376
##   4 0.000000000 0.013333333 0.004975124 0.004926108 0.009216590 0.007168459
##
##                6           7           8           9          10          13
##   0 0.791044776 0.842377261 0.850000000 0.792682927 1.000000000 1.000000000
##   1 0.164179104 0.136950904 0.114705882 0.195121951 0.000000000 0.000000000
##   2 0.026865672 0.018087855 0.029411765 0.000000000 0.000000000 0.000000000
##   3 0.008955224 0.002583979 0.005882353 0.012195122 0.000000000 0.000000000
##   4 0.008955224 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000
##
```

```
##              14          17          19          20
##   0 1.000000000 1.000000000 1.000000000 0.666666667
##   1 0.000000000 0.000000000 0.000000000 0.333333333
##   2 0.000000000 0.000000000 0.000000000 0.000000000
##   3 0.000000000 0.000000000 0.000000000 0.000000000
##   4 0.000000000 0.000000000 0.000000000 0.000000000
```

3- The last conclusion we can obtain from this plot is that the more deals a customer buys, the less tendency he/she will have to accept an offer. The data reveals that the customers who are more likely to accept offers have bought 0 or 1 discounted products in the last month.

As we can see in the following table only 15.67% (190/(1022+190)) of those who bought more than one discounted item accepted any offer. That's why I would recommend to focus on those who bought 1 or less discounted items.

```
Buy%>%
   select(offer,Deals)%>%
   group_by(offer)%>%
   summarise("Less than"=quantile(Deals, 0.5, na.rm=TRUE))
```

```
## # A tibble: 5 x 2
##   offer `Less than`
##   <fct>       <dbl>
## 1 0               2
## 2 1               1
## 3 2               1
## 4 3               1
## 5 4               1
```

```
  Discout=Buy[Buy$Deals>1,"offer"]
  table(Discout!=0)
```

```
##
## FALSE  TRUE
##  1022   190
```

```
  190/(1022+190) #15.67% accepted at least 1 offer
```

```
## [1] 0.1567657
```

As we all know the client satisfaction is a very important factor to determine if the customer trust a business enought to buy its products or not.

Thats why I presume that how happy s consumer is with the corporation will be an inportant factor. In the data frame, we have the feature "complain". This variable explains if the client complained to the costumer service in the last 2 years.

I asume that complaining is the equivalent of a non-happy customer.

With the following table, we can see that, as we expected, the clients who complained show a lower percentage of acceptance, more than 90% of then refused all the offers.

```
prop.table(table(data$Complain,cuts$num.accepted),1)
```

```
##
##              0          1          2          3          4
##   0 0.79179954 0.14669704 0.03644647 0.02004556 0.00501139
##   1 0.90476190 0.04761905 0.04761905 0.00000000 0.00000000
```

```
table(data$Complain,cuts$num.accepted!=0)
```

```
##
##     FALSE TRUE
##   0  1738  457
##   1    19    2
```

```
2/(19+2)#9.52% accepted at least one offer.
```

```
## [1] 0.0952381
```

We will finish this analysis by extracting some ideas from the variables that refer to time: Dt_Customer : date of customer's enrolment with the company Recency : number of days since the last purchase

We will examine first the Dt_Customer, this may seem an irrelevant data, however, it has been proven by many studies that the day of the week in which people buy, show a certain pattern of behavior.

For example, the person who goes to the supermarket on Saturday tend to buy larger quantities of food than those who go during the week, since these last ones, are more likelly to go at least 2 time per week to buy groceries.

Before applying the next function, I should clarify that, it transforms a date into de day of the week. It is a Spanish function and thus the days of the week will be in Spanish, here you have the translation:

#lu is lunes means Monday. #ma is martes means Tuesday. #mi is miercoles means Wednesday. #ju is jueves means Thursday. #vi is viernes means Friday. #sa is sabado means Saturday. #do is domingo means Sunday.

Even though most of the percentual differences that we can see in the following table may seem irrelevant, there is a fact that is worth mentioning.

The customers who enrolled to the company on Thursday or Friday are less likely to accept the offer.

The difference between the costumers who enrolled on Monday and those who enrolled on Thursday is more than 5.55%, I think is something that needs to be considered.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:reshape':
##
##     stamp
```

```
## The following objects are masked from 'package:dplyr':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(anytime)

time=data.frame("Recency"=data$Recency,"Date_customer"=data$Dt_Customer)
time$Date_customer=anytime(time$Date_customer)

daily=time%>%
  mutate(wday=wday(Date_customer,label=TRUE))
```

```
day=factor(daily$wday,levels=c("lu","ma","mi","ju","vi","sá","do"))

prop.table(table(day,cuts$num.accepted),1)
```

```
##
## day            0           1           2           3           4
##   lu 0.777777778 0.157407407 0.046296296 0.009259259 0.009259259
##   ma 0.758865248 0.156028369 0.049645390 0.035460993 0.000000000
##   mi 0.800000000 0.135714286 0.007142857 0.057142857 0.000000000
##   ju 0.833333333 0.142857143 0.015873016 0.000000000 0.007936508
##   vi 0.775862069 0.163793103 0.051724138 0.008620690 0.000000000
##   sá 0.790322581 0.129032258 0.024193548 0.048387097 0.008064516
##   do 0.813333333 0.120000000 0.053333333 0.013333333 0.000000000
```

```
0.833333333-0.777777778
```

```
## [1] 0.05555555
```

Regarding the Recency feature, I will start by plotting the data.

```
table(time$Recency)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## 28 24 28 29 26 15 21 12 25 24 26 25 24 27 19 16 21 19 19 24 21 21 13 27 28 24
## 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
## 21 27 23 29 32 18 21 17 23 17 18 24 25 18 25 15 20 20 11 16 31 19 28 29 22 28
## 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77
## 21 19 32 20 37 23 15 14 22 21 19 23 23 30 16 17 25 15 18 29 27 20 26 16 19 25
## 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
## 26 17 27 27 20 21 23 18 26 25 16 20 20 18 30 21 26 18 23 20 22 17
```

```
ggplot()+
  geom_bar(aes(time$Recency,fill=factor(cuts$num.accepted)))+
labs(x="Days since the last purchase", y="count", title="Days since the last purchase vs offers accepte
  scale_fill_discrete(name = "Offers accepted")
```

There is no sign of any pattern in the last plot nor in the table. I will try to cut the data in different groups grouping the days in weeks. For example, the first group will group all the customers who went to buy less than 7 days ago, the next group will have the customers who went between 7 and 14 days ago...

In my opinion, more data would be required in order to get a good conclusion out of this next table. However, the numbers reveal a certain patter that is very interesting. The percentage of customers who refused all the offers, reach the lower points during the following periodes:

#From 7 days to 14: 77% didn't accept any offer #From 21 days to 28: 73% didn't accept any offer #From 28 days to 35: 74% didn't accept any offer #From 56 days to 63: 75% didn't accept any offer

The rest of percentages of clients that rejected all the offers are closer or above 80%.

This data is interesting because it is possible to see that:

1-The customers are more likely to accept at least one offer after a week (from 7 days to 14), after a month (21 days to 35), after two months (from 56 days to 63)

2-Most of the "optimal time" to make the offer to the costumer are during the first month.

These two conclusions could be explained because according to psychology principles: people is more likely to spend their money in a business that they have already bought in, because they feel linked to it, this link disappears with time.

```
days=cut(time$Recency,c(-0.1,7,14,21,28,35,42,49,56,63,70,77,84,91,98,105),levels=c("7","14","21","28",
prop.table(table(days,cuts$num.accepted),1 )
```

```
##
## days                  0           1           2           3           4
##   (-0.1,7] 0.808743169 0.125683060 0.038251366 0.010928962 0.016393443
```

```
##   (7,14]   0.770588235 0.176470588 0.035294118 0.011764706 0.005882353
##   (14,21]  0.801418440 0.148936170 0.035460993 0.014184397 0.000000000
##   (21,28]  0.730061350 0.159509202 0.079754601 0.024539877 0.006134969
##   (28,35]  0.745222930 0.197452229 0.025477707 0.025477707 0.006369427
##   (35,42]  0.827586207 0.110344828 0.041379310 0.020689655 0.000000000
##   (42,49]  0.824675325 0.129870130 0.025974026 0.019480519 0.000000000
##   (49,56]  0.815642458 0.145251397 0.022346369 0.016759777 0.000000000
##   (56,63]  0.751824818 0.160583942 0.043795620 0.029197080 0.014598540
##   (63,70]  0.784722222 0.159722222 0.041666667 0.013888889 0.000000000
##   (70,77]  0.833333333 0.104938272 0.024691358 0.037037037 0.000000000
##   (77,84]  0.819875776 0.118012422 0.024844720 0.024844720 0.012422360
##   (84,91]  0.783216783 0.160839161 0.041958042 0.013986014 0.000000000
##   (91,98]  0.787500000 0.156250000 0.031250000 0.018750000 0.006250000
##   (98,105] 0.882352941 0.058823529 0.058823529 0.000000000 0.000000000
```

CONCLUSIONS OF THE ANALYSIS

I will divide the conclusions in three different parts, family and finances, habits, and time. The first part will be more focused on which group of consumers should be targeted for the offers that the business makes. In the second one I will be explaining how to increase the success of the marketing campaigns. The last one will be more anecdotical, some data to consider but not to as important as the data explained in the habits part.

Family and finances: The customer target of the business, according to the amount of data that the business has, seems to be people born in the decade of the 70's. However, this is the customer that tends to reject the most offers. According to the data, 83.78% of the consumers born in the 70's refused all the marketing campaigns that the business made, and just 5% accepted 2 or more of them. In contrast, just the 67.44% of those born in the decade of the 90's rejected all the offers and 13.95% of them accepted 2 of more offers.

Another important factor to decide who should the business marketing campaigns focus to, is the personal finances. Even though people born in the decade of the 70's may seem to have one of the biggest budgets to spend, the data reveals that in fact they are the second demographic group with less income, the first one is the group born in the decade of the 80's.

To determine if the Income was a decisive factor to decide if the customer would accept or decline the offer we created a table of percentages. There it is possible to see that those consumers who earn less than 50000 $ per year, show a 90% provability of refusing all the offers and those who earn more than that have a provability of just 69%. As we have seen with the data, the best groups, both in percentage of offer acceptance and the highest budget are the customers born in the decade of the 40's and the 90's.

Another very important point to consider is if the consumer has kids of teenagers. The data shows that the customers who have kids or teenagers tends to prefer to refuse all or most of the offers, in fact, the consumers who have kids represent just the 18.96% of all the consumers who accepted 2 or more offers. The number of clients having teenegers do not refuse that many offers, but the percentage is still high.

Habits: In this part of the conclusion, I will explain 2 very interesting facts that seem to be crucial when explaining why some costumers refuse the offers and others accept them, which products do they consume and how they buy their groceries.

From the plot we did that explains which lines of products do the consumers buy, there are 2 important facts to highlight.

The fist one is that in all the of the product lines, but the wine one, a pattern has appeared. The level of spending shows a positive correlation with the number of offers accepted from 1 offer until 3 and then the level of consume reduces, the consumers who accepted 4 offers, show a low spending.

If the business just contacts the consumers who's spending is in the top 5% of the different lines of products, the possibility of finding a client who will accept at least 1 offer is 47.6%

The second important fact that it is possible to see in the plot is that the consume of wine is one of the

best indicators. The more a consumer spends in wine, the more possible is for him to accept an offer. If the business contacts those consumers who have a spent more than 822$ on wine in the last two years, the possibilities to fins a client who will accept at least 1 offer if 68.60%.

Obviously these two facts, have a certain relation with the income of the consumers that I highlighted in the "Family and finances" part. However, to know the influence that the level of spending has, is very important for the business. Since the supermarket is the organization that sells the products, it will always be able to know who are the clients who consume the most and thus, to whom make the offers.

Another important fact is that according to the data, the consumers who visit more the business' webpage are less likely to accept the offers. From those consumers who visited the webpage more than 9 times during the last month, just one accepted one offer.

A similar pattern appears with the number of discounted products that the consumer has bought during the last month. The customers who bought more than 1 discounted product in the last month have a probability of just 15.67% to accept one offer or more.

Finally, regarding to the happiness of the costumers with the corporation, if a client has complained about the business in the last 2 years, they have a probability of 90% to reject all the offers, I would recommend not to contact them for these offers.

Time: Before finishing the conclusion, I would like to highlight two facts that, in my opinion, are important to consider. According to the data, the costumers who enrolled to the company on Thursday or Friday are less likely to accept the offer. In fact, those who enroller on Monday have a probability to accept 5.55% higher than the costumers enrolled on Thursday.

The data also reveals that the customers are more likely to accept an offer if they are contacted during the following periods after the last time they shop from the business: #-From 7 days to 14 #-From 21 days to 28 #-From 28 days to 35 #-From 56 days to 63