Retail Store

Oriol Monjo

27/7/2020

The business wants to identify which is the country with the highest levels of income. From that country, the organization wants to identify the most profitable segment of costumers. Finally, it wants to obtain as much information as possible about the costumers in that segment.

For this analysis, the enterprise has just given us the data, without providing any context nor other information about how the comporation works. However, the questions that it asks, require a basic understanding of the business.

That is why, I will focus on answering the questions of the corporation and at the same time I will also be searching for patters in the business behaviour, to determine the strategies that it is following.

In order to analyze the data, I will use a top-down approach, starting with a general vision of the whole world and then highlighting and focusing on specific details of the data.

During this case of study, I will provide an explanation of the most important things I see in the different plots.

```
library("readxl")
library("ggplot2")
data <- read_excel("Retail.xls")
head(data)
```

```
## # A tibble: 6 x 24
##   `Row ID` `Order ID` `Order Date`        `Ship Date`         `Ship Mode`
##      <dbl> <chr>      <dttm>              <dttm>              <chr>
## 1    32298 CA-2012-1~ 2012-07-31 00:00:00 2012-07-31 00:00:00 Same Day
## 2    26341 IN-2013-7~ 2013-02-05 00:00:00 2013-02-07 00:00:00 Second Cla~
## 3    25330 IN-2013-7~ 2013-10-17 00:00:00 2013-10-18 00:00:00 First Class
## 4    13524 ES-2013-1~ 2013-01-28 00:00:00 2013-01-30 00:00:00 First Class
## 5    47221 SG-2013-4~ 2013-11-05 00:00:00 2013-11-06 00:00:00 Same Day
## 6    22732 IN-2013-4~ 2013-06-28 00:00:00 2013-07-01 00:00:00 Second Cla~
## # ... with 19 more variables: `Customer ID` <chr>, `Customer Name` <chr>,
## #   Segment <chr>, City <chr>, State <chr>, Country <chr>, `Postal Code` <chr>,
## #   Market <chr>, Region <chr>, `Product ID` <chr>, Category <chr>,
## #   `Sub-Category` <chr>, `Product Name` <chr>, Sales <dbl>, Quantity <dbl>,
## #   Discount <dbl>, Profit <dbl>, `Shipping Cost` <dbl>, `Order Priority` <chr>
```

I will start the analysis by taking a look at the profit that the business obtains from the different countries. In the following plot, we can see that some countries are not profitable, for example Turkey, Nigeria, Africa and Latin America, seem to obtain mixed result.

Germany, France, United Kingdom, India and China are profitable markets. However, the most important market for this business is the USA.

```
Profit=aggregate(x = data$Profit, by = list(data$Country), FUN = "sum")
```

```r
Profit=Profit[order(Profit$x),]
head(Profit)#Less profitable Countries
```

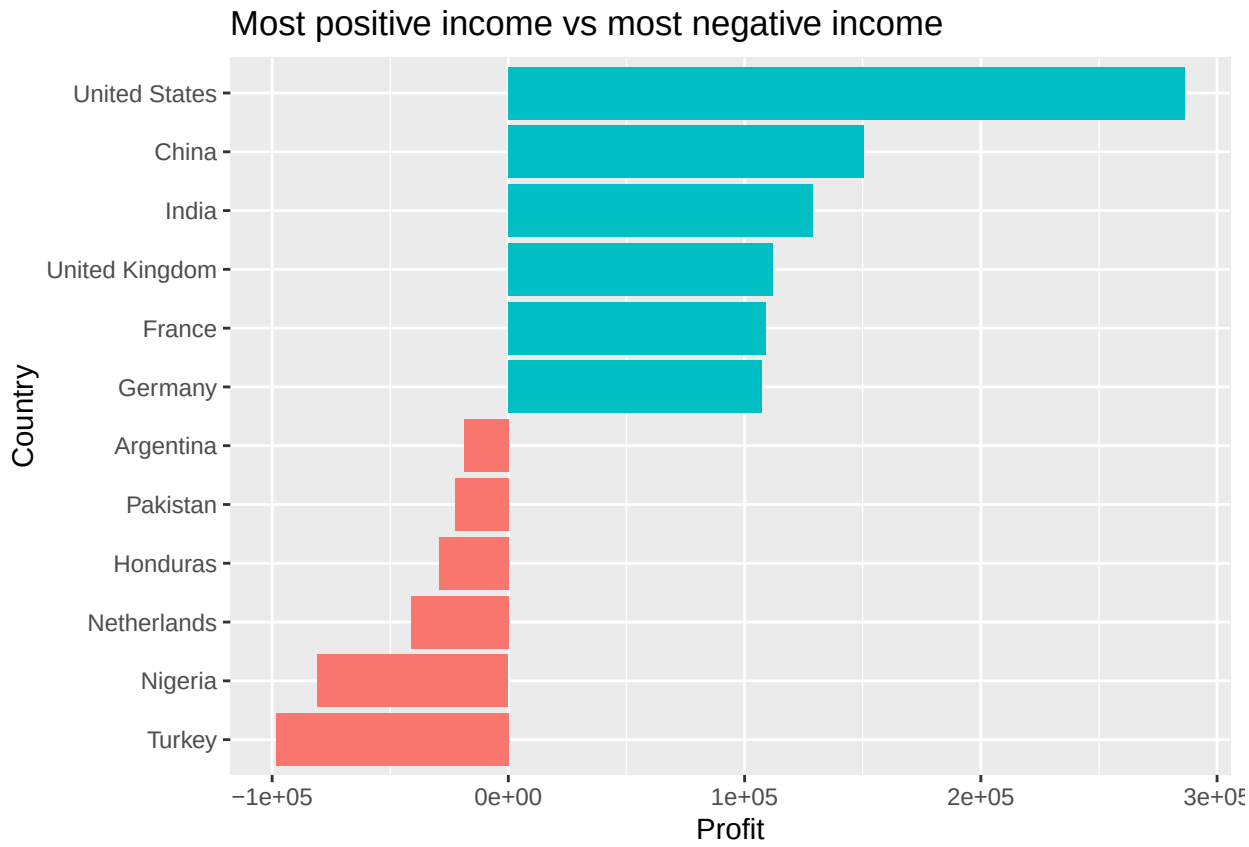
```
##         Group.1         x
## 134      Turkey -98447.23
## 95      Nigeria -80750.72
## 91  Netherlands -41070.07
## 55     Honduras -29482.37
## 97     Pakistan -22446.65
## 5     Argentina -18693.80
```

```r
tail(Profit)#Most profitable countries
```

```
##            Group.1        x
## 48         Germany 107322.8
## 45          France 109029.0
## 139 United Kingdom 111900.1
## 58           India 129071.8
## 27           China 150683.1
## 140  United States 286397.0
```

```r
Profit2=rbind(head(Profit),tail(Profit))
Profit2$Group.1=factor(Profit2$Group.1,levels =Profit2$Group.1[order(Profit2$x)])

ggplot(Profit2)+
  geom_col(aes(Group.1,x,fill=x>0),show.legend = FALSE)+
  coord_flip()+
  labs(x = "Country",y="Profit",title="Most positive income vs most negative income")
```



I will also search for the countries with the most sales to see if the USA is also the most relevant country in this field.

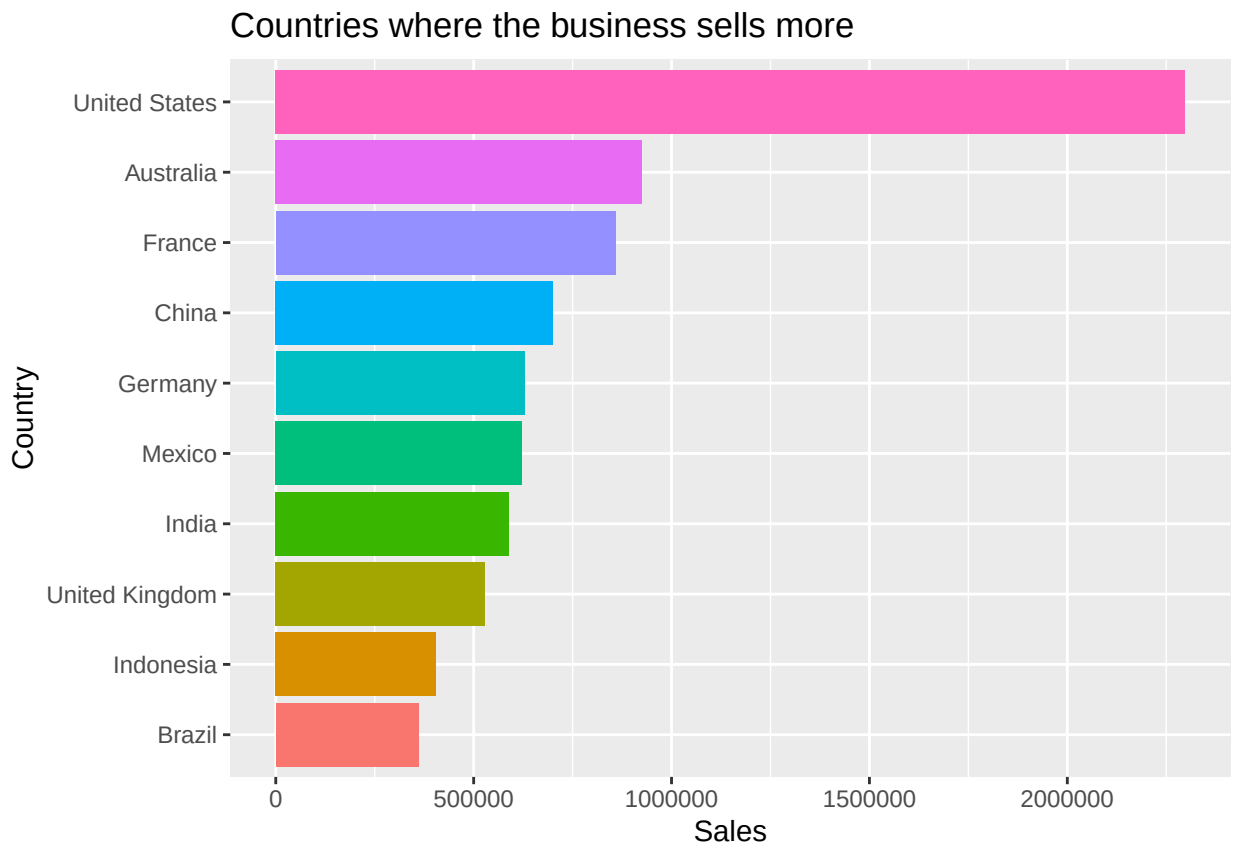
With the following plot we can see that the EUA is again in the first place.

```
Sales=aggregate(x = data$Sales, by = list(data$Country), FUN = "sum")
Sales=Sales[order(Sales$x),]
tail(Sales)
```

```
##           Group.1         x
## 82         Mexico  622590.6
## 48        Germany  628840.0
## 27          China  700562.0
## 45         France  858931.1
## 7       Australia  925235.9
## 140 United States 2297200.9
```

```
Sales2=tail(Sales,10)
Sales2$Group.1=factor(Sales2$Group.1,levels =Sales2$Group.1[order(Sales2$x)])

ggplot(Sales2)+
  geom_col(aes(Group.1,x,fill=Group.1),show.legend = FALSE)+
  coord_flip()+
  labs(x = "Country",y="Sales",title="Countries where the business sells more ")
```



Before proceeding to focus on the USA market, I would like to continue analyzing a little bit more the whole business. I do this because I think that this will be the best way to understand how the organization works and it may be helpful to extract conclusions from one especific market.

We will now take a look at the different Sub-categories that this business has.

The most important thing to see in the next graphic, is that I plotted the categories in the X-axis and the Profit in the Y-axis and thanks to that, we can see that this business obtains profits from a good part of the total operations. However, there is a great number of orders that make the business to lose money.

Specifically, 38078 orders are profitable and 12544 are income-negative.

With the fist "aggregate" function, it is possible to see that adding the profitable and non profitable operations, the three categories of products show a positive income.

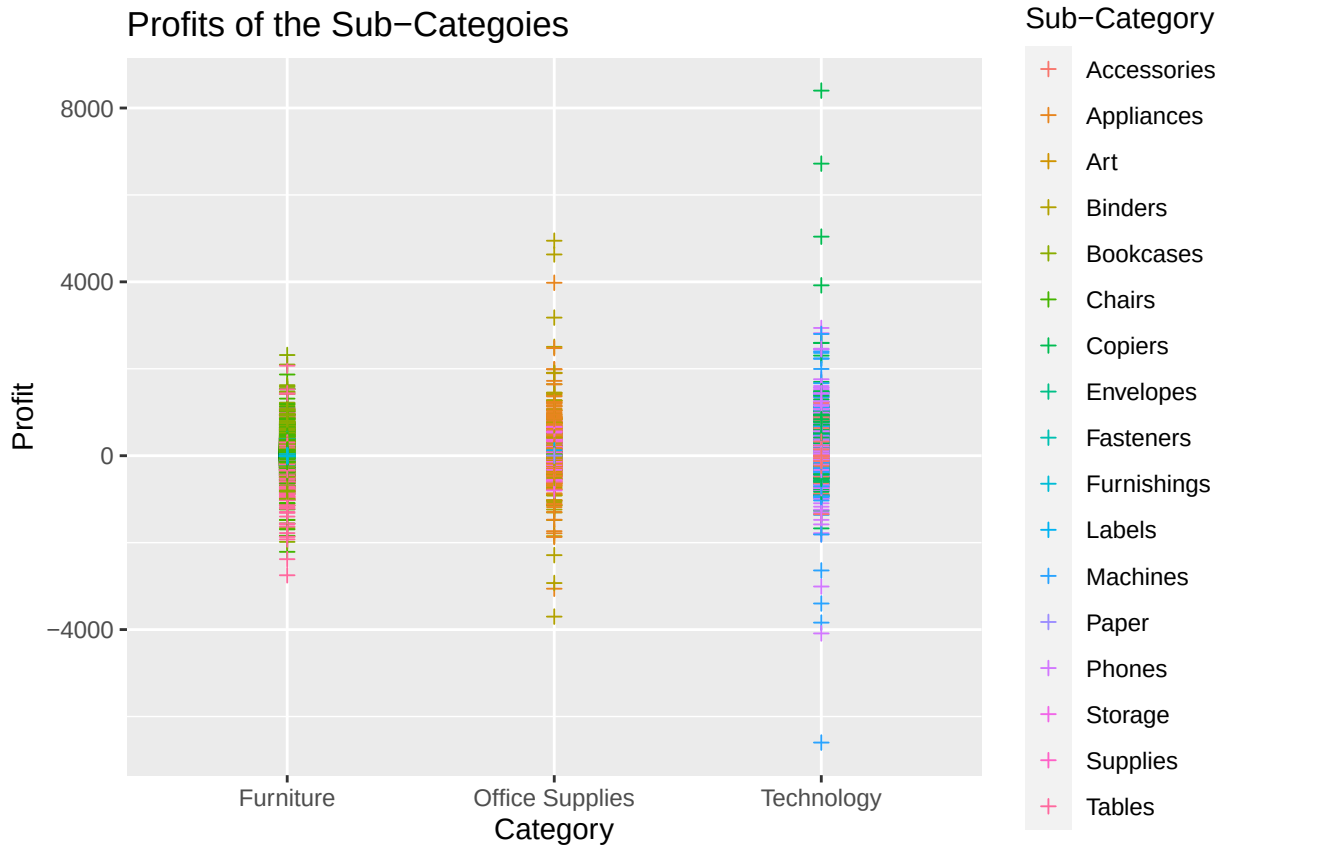
With the second "aggregate" function, we can see that the firsts quantile is negative in the "Furniture" Category and very close to 0 in the other 2 categories. That means that the business loses money or earns very little in at least 25% of the operations.

In the second aggregate, there is an other interesting factor to notice. The difference between the "regular products", those that produce a benefit that is in the Interquartile range (between the 1st and 3rd quantile) and the products that are near the maximum profit or the minimum.

The median and the mean are similar, and the Interquartile range, in all three cases, is less than 100$. This is a clear indicator that the business applies, in most of the products, a very reduced margin.

Nonetheless, as we can see in the "Profit.Max.", the business also sells very expensive products with high margins; or, as we can see in "Profit.Min.", expensive products that had to be sold losing a lot of money.

```
ggplot(data)+
  geom_point(aes(Category,Profit,color=`Sub-Category`),shape=3)+
  labs(x = "Category",y="Profit", colour = "Sub-Category",title="Profits of the Sub-Categoies")
```




```
dim(data[data$Profit<0,])[1] #The income is negative
```

```
## [1] 12544
```

```
dim(data[data$Profit>0,])[1] #The income is positive
```

```
## [1] 38078
```

```
aggregate(Profit~Category,data=data, FUN=sum)
```

```
##          Category   Profit
## 1       Furniture 285204.7
## 2 Office Supplies 518473.8
## 3      Technology 663778.7
```

```
aggregate(Profit~Category,data=data, FUN=summary)
```

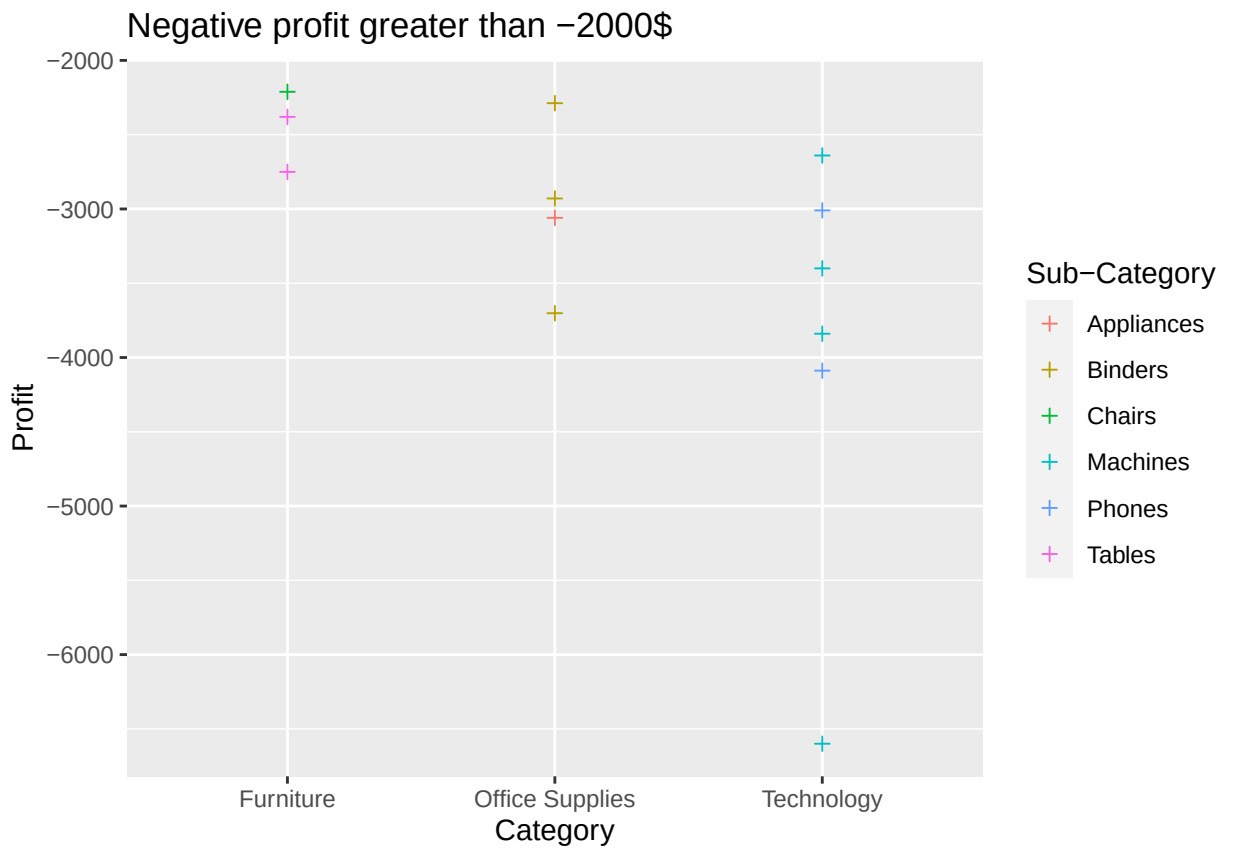
```
##          Category Profit.Min. Profit.1st Qu. Profit.Median Profit.Mean
## 1       Furniture -2750.28000      -12.17500       15.50220    28.87857
## 2 Office Supplies -3701.89280        0.45000        6.55380    16.57896
## 3      Technology -6599.97800        0.50000       29.94000    65.45496
##   Profit.3rd Qu. Profit.Max.
## 1       69.36000  2316.51000
## 2       20.58000  4946.37000
```

```
## 3       98.85000  8399.97600
```

Now, I will take a closer look to the orders that produce the greatest negative profit, -2000 $ or less.

As we can see in the plot, among these products, we can find furniture like “tables” and “chairs” or Office supplies such as “Binders” and “Appliances”. However, the most important negative income, comes from the technology category were we can find sub-categories like Machines and Phones.

```
data_redu=data[data$Profit<=-2000,]
ggplot(data_redu)+
  geom_point(aes(Category,Profit,color=`Sub-Category`),shape=3)+
  labs(x = "Category",y="Profit", colour = "Sub-Category",title="Negative profit greater than -2000$")
```



I will now create a plot that will contain all the sub-categories of the business, with the total profit that every sub-category has.

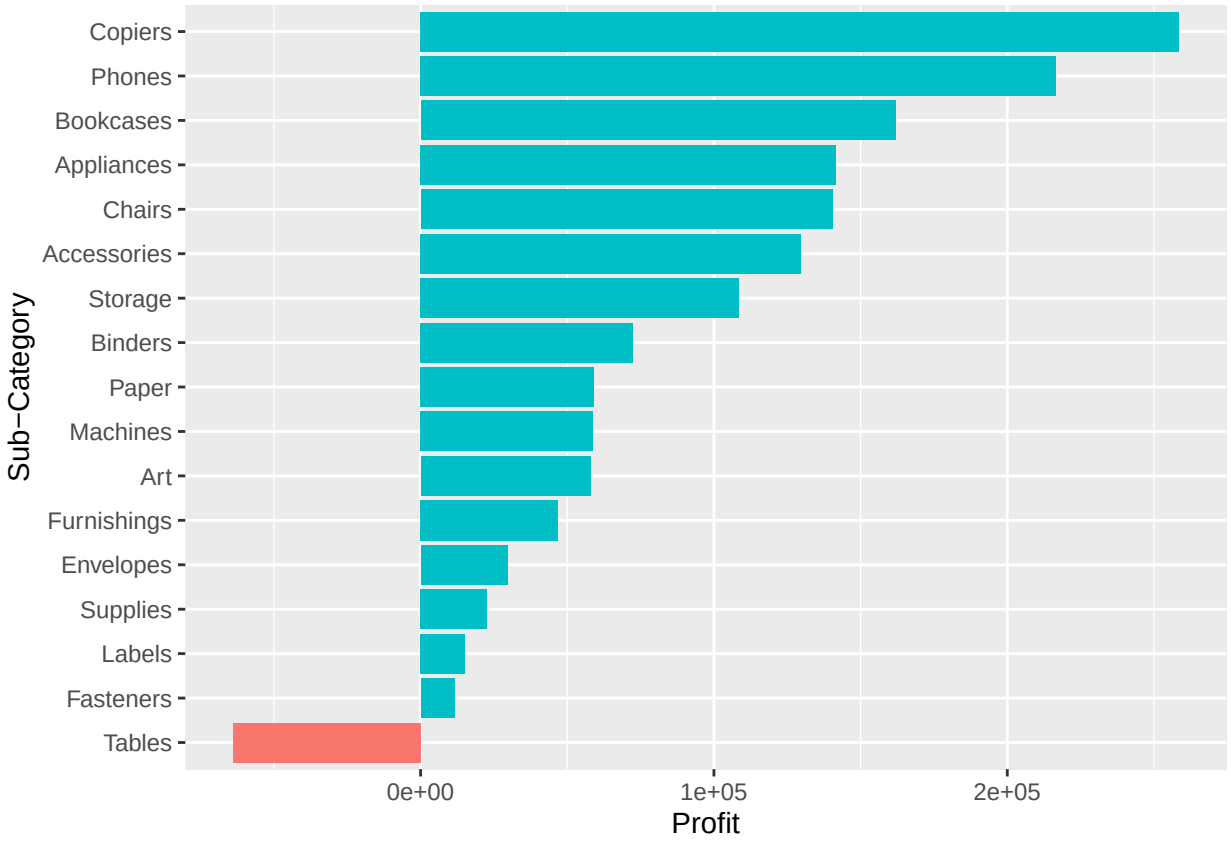
As we can see in the following plot, the sub-category “tables” is generating a great negative profit. Considering that this sub-category contains products with an high negative-margin, the organization should consider if it is worth to continue selling “tables”.

Another conclusion that we can take out of this chart, is that the product “Machines” is profitable even if the business suffers losses selling certain “Machines”. Instead of deleting the “Machines”, we can search for which “Machines” are making the organization to lose money and to think whether to keep them or delete them.

```
data_sub=aggregate(x = data$Profit, by = list(data$`Sub-Category`), FUN = "sum")


data_sub$Group.1=factor(data_sub$Group.1,levels =data_sub$Group.1[order(data_sub$x)])
ggplot(data_sub)+
```

```
  geom_col(aes(Group.1,x,fill=x>0),show.legend = FALSE)+
  coord_flip()+
  labs(x = "Sub-Category",y="Profit")
```



As we can see in the following data frame, these non-profitable machines are: Cubify CubeX 3D Printer Triple Head Print, Cubify CubeX 3D Printer Double Head Print, Cubify CubeX 3D Printer Double Head Print, Lexmark MX611dhe Monochrome Laser Printer.

Considering that out of these 4 machines, 3 are "Cubify CubeX 3D", we will continue our analysis, examining the profit that we obtained from this line of products.

```
data[data$`Sub-Category`=="Machines" & data$Profit<=-2000,18]
```

```
## # A tibble: 4 x 1
##   `Product Name`
##   <chr>
## 1 Cubify CubeX 3D Printer Triple Head Print
## 2 Cubify CubeX 3D Printer Double Head Print
## 3 Cubify CubeX 3D Printer Double Head Print
## 4 Lexmark MX611dhe Monochrome Laser Printer
```

Now we will search how many products of the "Cubify CubeX 3D" line the business has and which profit obtains from them.

As we can see, in the following data frame, there are just 4 products of this line, 3 of them are non-profitable, and just one is, however, its margin is low.

The business should seriously consider if is worth it to keep selling this line of products, I personally recommend to delete it.

```
data[grepl("Cubify CubeX 3D", data$`Product Name`),c(18,22)]
```

```
## # A tibble: 4 x 2
##   `Product Name`                                Profit
##   <chr>                                          <dbl>
## 1 Cubify CubeX 3D Printer Triple Head Print -3840.
## 2 Cubify CubeX 3D Printer Double Head Print -6600.
## 3 Cubify CubeX 3D Printer Double Head Print -2640.
## 4 Cubify CubeX 3D Printer Double Head Print   360.
```

If we check the information that we have of these products we find that all of them have been sold in the USA and that the only one that is profitable is from the west part of the country.

```
data_Cubify=data.frame(data[grepl("Cubify CubeX 3D", data$`Product Name`),])
data_Cubify["Country"]
```

```
##         Country
## 1 United States
## 2 United States
## 3 United States
## 4 United States
```

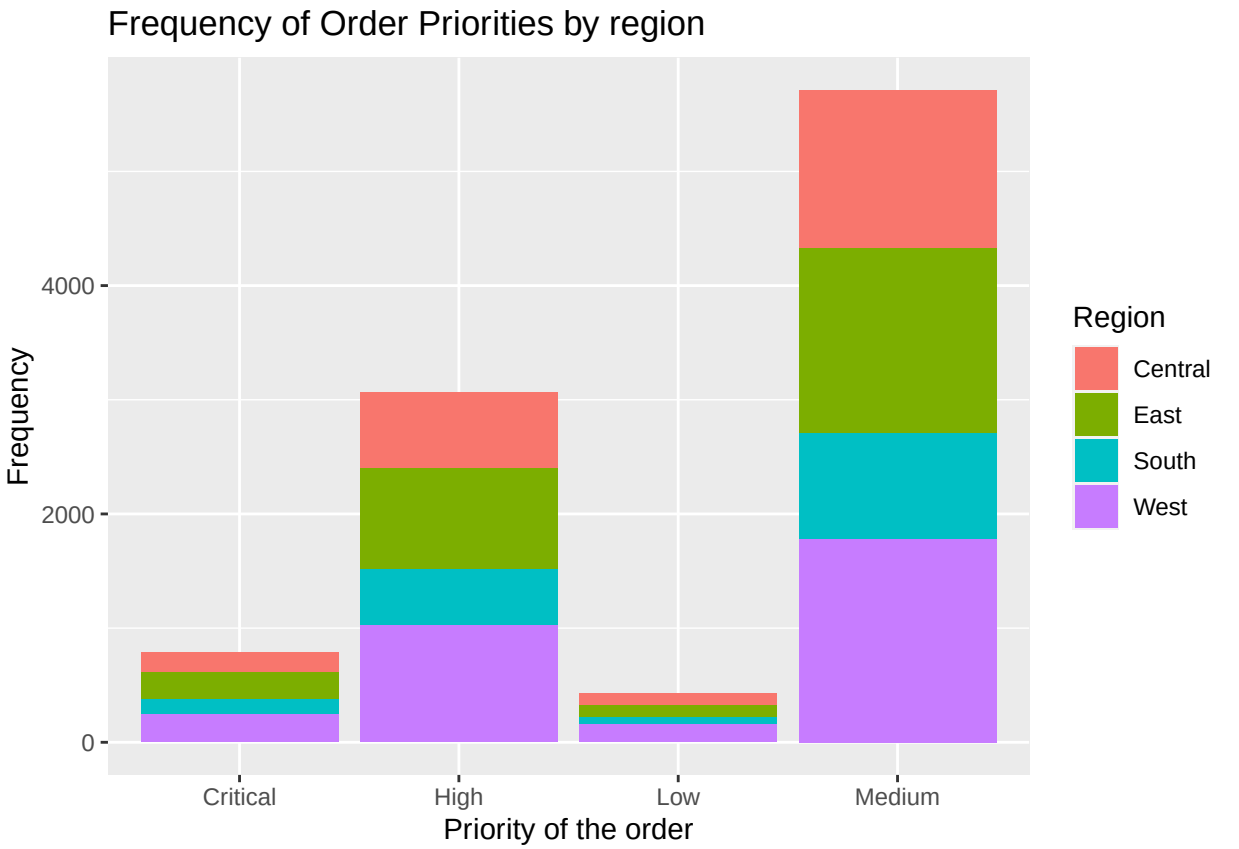
Since the 4 machines are for the USA, and as we said at the very beginning of this case, we need to focus on this country, we will now focus on it.

I will start the USA analysis, understanding how the business manages the costumer's orders.

Fist, we will take a look at the priorities of the orders. In the next plot we can see that the business uses mainly "Medium" and "High" priorities for its shipments.

```
data_USA=data[data$Market=="US",]

ggplot()+
  geom_bar(aes(data_USA$`Order Priority`,fill=data_USA$Region))+
  labs(x = "Priority of the order",y="Frequency", fill = "Region",title="Frequency of Order Priorities
```



It is possible to see in the last plot, that in all the columns, but specially on the "Medium" and "High" priority ones, the "East" and "West" Regions have a greater proportion of shipments than the "Central" and "South" part.

I assume that this is caused because of a higher demand on the "East" and "West" regions, but to be sure, it will be better to check it.

As we can see in the following table, the "East" and "West" Regions are the ones with the highest demands.

```
table(data_USA$Region)
```

```
##
## Central    East   South    West
##    2323    2848    1620    3203
```

Now I will analyze how the "Order Priority" is affected by the "Category" of the sent products.

In the following table, we will find how the percentage of the different "order priorities" is distributed between the different categories of products.

For example, it is possible to see that the 20% of the "Critical order priorities" that the business sends in the USA, are used for "Furniture" products, 60% for "Office Supplies" and 19% for "Technology".

It is interesting to notice that in every case, the 60% of every order priority is used to send Office supplies.

```
prop.table(table(data_USA$`Order Priority`,data_USA$Category),margin=1)
```

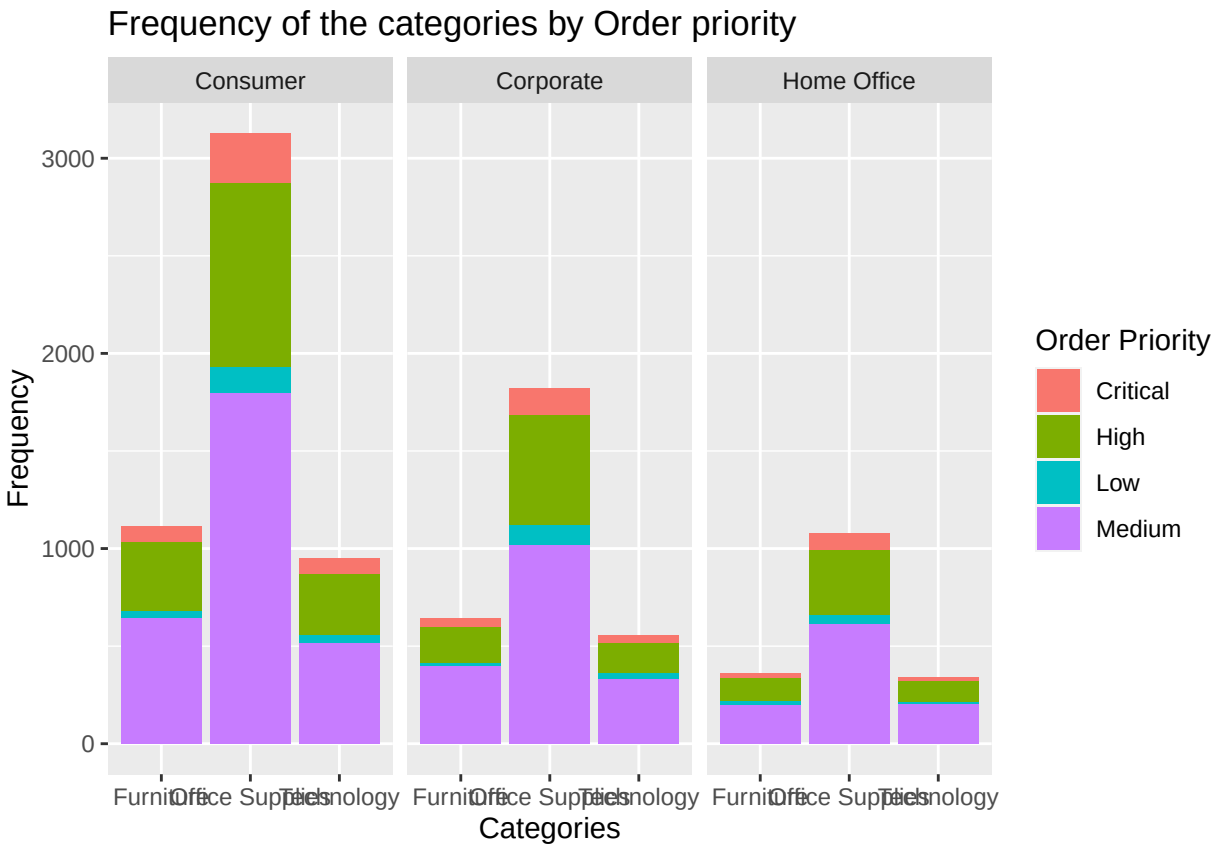
```
##
##             Furniture Office Supplies Technology
##   Critical 0.2030651       0.6104725  0.1864623
##   High     0.2121212       0.6001955  0.1876833
```

```
##    Low       0.1689815        0.6527778  0.1782407
##    Medium    0.2168126        0.5996497  0.1835377
```

In order to understand better this data, we will use a bar plot to see if this percentage can be explained by the quantity of Office supplies products.

As we suspected, the Office supplies is the most sold category in the USA and, as the following plot shows, almost half of the Office Supplies aim the Consumer Market.

```
ggplot(data_USA)+
  geom_bar(aes(Category,fill=`Order Priority`))+
  facet_wrap(~Segment)+
  labs(x = "Categories",y="Frequency", fill = "Order Priority",title="Frequency of the categories by Or
```



The business wants to know more about the segment that produces more money in the USA. That's why I will now focus on the Consumer one.

```
USA_consumer=data_USA[data_USA$Segment=="Consumer",]
```

I will proceed to take a closer look to the column "Order Date", just because it can give us a good understanding of the consumer market.

I want find out if the consumers show a different behavior depending on how close it is the end of the month.

In other words, do they buy more during the first days of the month? Do they buy more when the month is about to end? or there is no significant difference?

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.3.0 -
```

```
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## v purrr   0.3.4
```

```
## -- Conflicts ---------------------------------------------------------------- tidyverse_conflicts() -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(modelr)
library(lubridate)
```

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

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

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

```
library(stringi)

USA_consumer$Order_date_str=as.character(USA_consumer$`Order Date`)

USA_consumer$Order_date_of_month=as.integer(stri_sub(USA_consumer$Order_date_str, -2))

USA_consumer$Order_date_of_month=cut(USA_consumer$Order_date_of_month, breaks=4,labels=c("1st_week","2n
```

To answer the questions, I will divide the whole month in 4 equal groups. I do that because I want to know if the clients show a different behavior during any of the quarters of the months.

Just for simplicity, I will call these quarters "weeks", it is easier to understand. However, not all these "weeks" will have 7 days, some of them will have 8.

That is because, the maximum number of days in a month is 31. If I divide 31/4 the result is 7.75 and thus, since the program cannot cut "3/4" of a day, it puts 8 days in some groups.

Another thing to consider is that not all the months have 31 days. Because of that the 4th week some months will have the full 8 days, some months if will have 7, and on february even less days. When we examine the results, we should take both things into consideration.

The first thing we will do with this "weeks" is to use them in a proportion table that will tell us when the USA Consumers buy more.

As we can see in the first table, the percentages of orders placed, does not vary much during the months. However, considering that this is the data from a big company, 1% of sells represent a lot a money. That is why we need to investigate more these findings.

We can see that the "Central" Region, buys much more during the first week and that the "West" Region buys more during the second. Moreover, the "East" and "South" regions show a consume of 26% during the 2 fist weeks, and during the 3rd and 4rth week the percentage decreases.

During the first 2 weeks, the proportion of consume is 52.93% (0.5293778), and during the 2 last weeks is 47.06% (0.4706222).

We can then affirm that during the two first weeks of the month the consumers buy more than during the 2 last weeks.

There is a difference of almost 6% between these two periods. (0.5293778-0.4706222=0.0587556)

```
prop.table(table(USA_consumer$Region,USA_consumer$Order_date_of_month),1)
```

```
##
##            1st_week  2nd_week  3rd_week  4th_week
##   Central 0.2986799 0.2318482 0.2541254 0.2153465
##   East    0.2668482 0.2634445 0.2457454 0.2239619
##   South   0.2684964 0.2613365 0.2446301 0.2255370
##   West    0.2338517 0.2936603 0.2362440 0.2362440
```

```
Con_region=(prop.table(table(USA_consumer$Region,USA_consumer$Order_date_of_month),1))

#First part of the month
sum(prop.table(table(USA_consumer$Order_date_of_month))[c(1,2)])
```

```
## [1] 0.5293778
```

```
#Second part of the month
sum(prop.table(table(USA_consumer$Order_date_of_month))[c(3,4)])
```

```
## [1] 0.4706222
```

Now, I would like to check if in all the sub-categories is true that during the fist part of the month people tend to buy more.

In the next table we can see that it seems that most of the sub-categories don't show any different behavior between the weeks. Nonetheless, "Envelopes", "Copiers" and "accessories" have more sells during the first part of the month and "Suppliers" during the second half. Furthermore, the customers buy more "Fasteners" during the 4rth week and "Machines" during the 1st and 4rt week.

```
prop.table(table(USA_consumer$`Sub-Category`
                 ,USA_consumer$Order_date_of_month),1)
```

```
##
##                1st_week  2nd_week  3rd_week  4th_week
##   Accessories 0.2720588 0.3137255 0.2352941 0.1789216
##   Appliances  0.2745902 0.2663934 0.2336066 0.2254098
##   Art         0.2803738 0.2453271 0.2453271 0.2289720
##   Binders     0.2679487 0.2628205 0.2653846 0.2038462
##   Bookcases   0.2748092 0.2213740 0.2595420 0.2442748
##   Chairs      0.2583587 0.2401216 0.2644377 0.2370821
##   Copiers     0.3714286 0.3428571 0.1142857 0.1714286
##   Envelopes   0.2093023 0.3178295 0.2635659 0.2093023
##   Fasteners   0.2456140 0.2192982 0.2017544 0.3333333
##   Furnishings 0.2692308 0.2631579 0.2226721 0.2449393
##   Labels      0.2804233 0.2645503 0.2116402 0.2433862
##   Machines    0.2982456 0.1929825 0.2280702 0.2807018
##   Paper       0.2424242 0.2987013 0.2366522 0.2222222
##   Phones      0.2505543 0.2483370 0.2616408 0.2394678
##   Storage     0.2666667 0.2422222 0.2600000 0.2311111
##   Supplies    0.2200000 0.2200000 0.3200000 0.2400000
##   Tables      0.3018868 0.3018868 0.1761006 0.2201258
```

I will now examine these sub-categories that we have highlighted to see if there is any pattern.

In order to do that, I will first find the price of the products. In the data we can find information about the total Sales for a specific product and the quantity of products that contained the order.

So, “price” is equal to (“Total Sale”/“Total quantity bought”).

In the following boxplot, it is quite clear that “Copiers” and “Machines” are the two sub-categories with the highest price. However, we should point out that that “Machines” sub-category, cover the whole range of prices. It goes from very cheap to very expensive.

The median in the “Machines” sub-category is in the low part of the box, that means that most of the products sold has a ‘reduced’ price.

Another interesting point that we can see, is that “Fasteners” and “Supplies” are cheap compared with the other Sub-Categories. In other words, the two categories that the customers buy the most during the last part of the month are the cheapest ones.

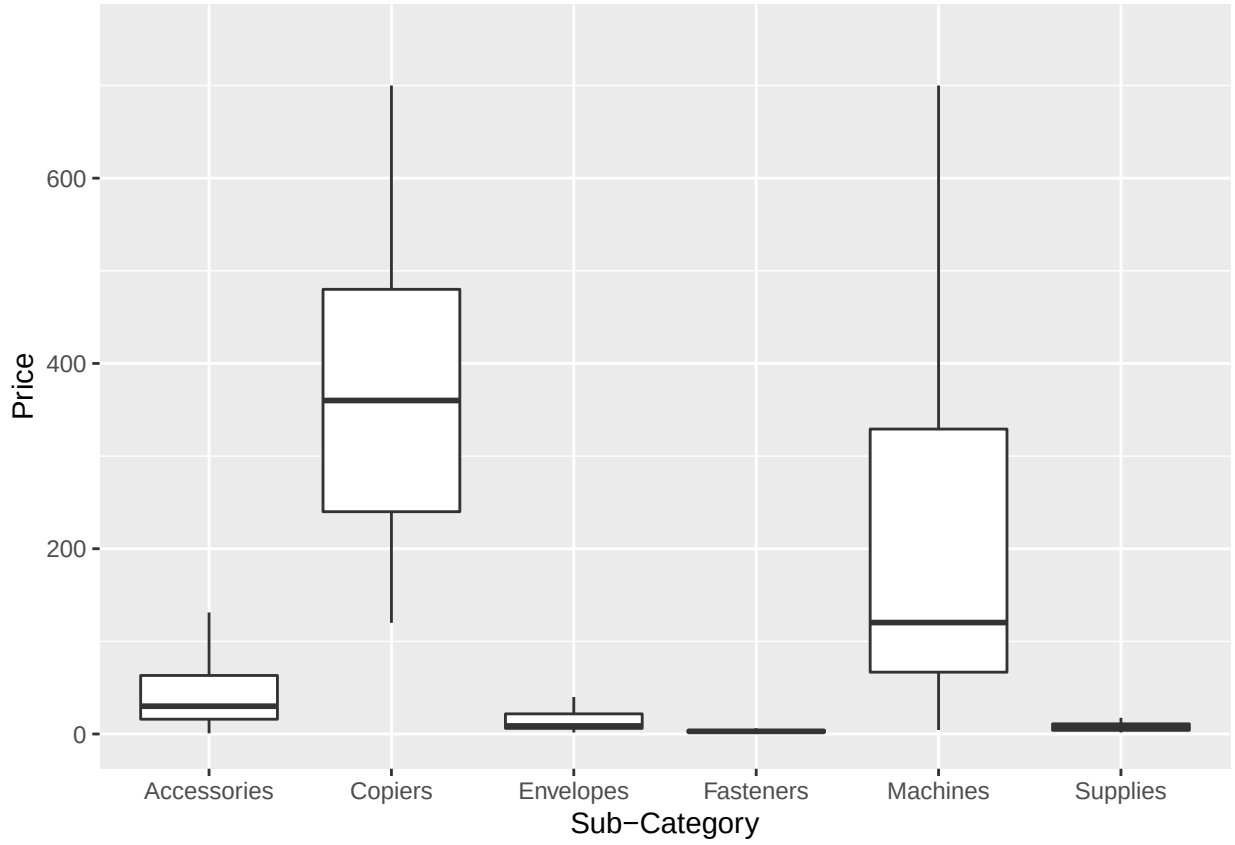
It makes sense, since people get paid at the beginning of the month, is during this first days when they can spend the most.

```
compras=USA_consumer[USA_consumer$`Sub-Category`=="Envelopes"| USA_consumer$`Sub-Category`=="Copiers"|US

compras=mutate(compras, price=Sales/Quantity)

ggplot(compras)+
  geom_boxplot(aes(`Sub-Category`,price), outlier.shape=NA)+
  ylim(c(0,750))+
  labs(x = "Sub-Category",y="Price")
```

```
## Warning: Removed 15 rows containing non-finite values (stat_boxplot).
```



We know that the orders of products of the sub-category “Machines”, are more pronounced during the first and last week of the month and that this sub-category have a great range of prices. We have just found out

that the products that people buy during the first half of the month tend to be more expensive than those that the customers buy in the second half.

It would be logical to expect that the products from the Sub-Category "Machines" bought during the first part of the month, would be more expensive than the ones ordered during the second part.

To obtain this information I will create a boxplot with the 2 weeks. In order to make it simpler to visualize it, I will hide the outliers.

As we can see, the price seems to be slightly higher in the first week than in the 4th week. However, both range of prices are similar and thus, we can't conclude that the prices are different.

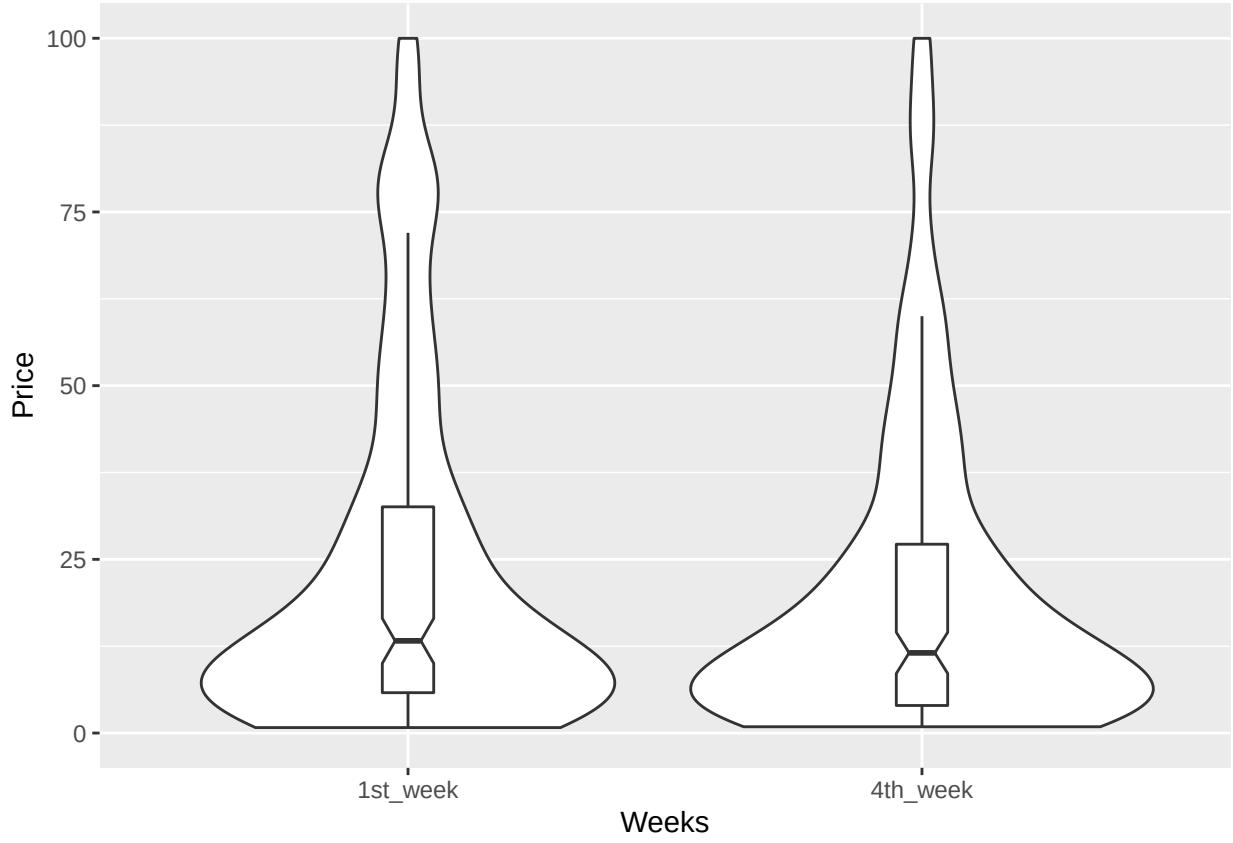
One thing that we need to consider is that there are a lot of outliers in the plot that I have just taken out of it.

```
compras2=compras[compras$Order_date_of_month=="1st_week"|compras$Order_date_of_month=="4th_week",]

ggplot(compras2)+
  geom_violin(aes(Order_date_of_month,price))+
  geom_boxplot(aes(Order_date_of_month,price),width=0.1,outlier.shape=NA, notch=TRUE)+
  ylim(c(0,100))+
  labs(x = "Weeks",y="Price")
```

```
## Warning: Removed 78 rows containing non-finite values (stat_ydensity).
```

```
## Warning: Removed 78 rows containing non-finite values (stat_boxplot).
```



The outliers are specially interesting to examine in this case, because, as we said before, a huge part of the products in the "Machines" Sub-category, have a reduced price. If that happens, the 3 quantiles will be in the lower part of the plot, leaving a lot of data as "outlier".

Now we will proceed to examine better the outliers to explain if there is a difference of prices between the weeks.

To do that I will select just those values above 75$.

According to the next table, both groups have very similar quantiles, and thus this won't be a interesting indicator. There are two important information to take out of this table, the "mean" and the "number".

The number, because there are 61 outliers from the first week and 37 from the forth one. Even though the number of registers is higher, that does not confirm that the price of the registers is also higher.

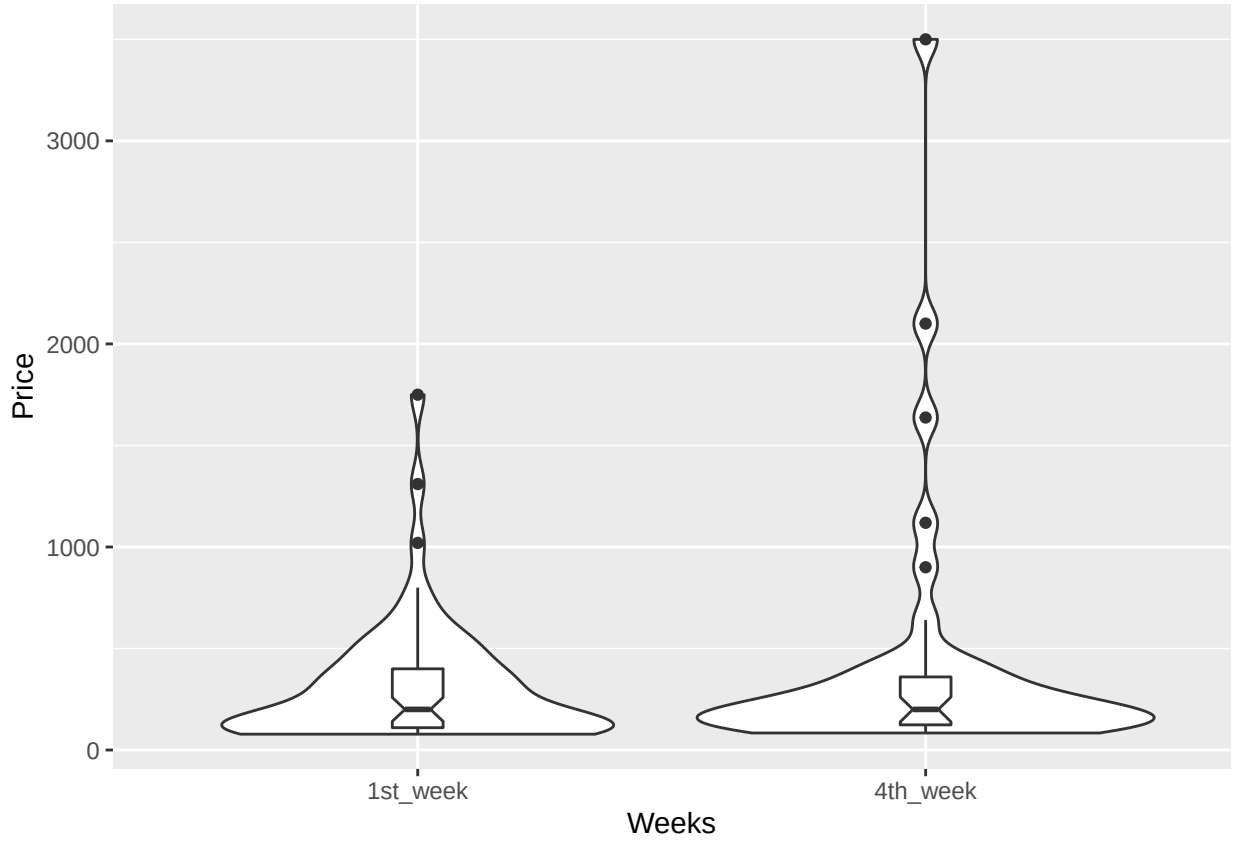
In this case, both weeks show a similar range of prices. It is important to stress that the 4rth week has some really high values and this makes the mean to increase its value.

We can conclude then, that we can't prove that the two groups have different range of prices.

```
filter(compras2,price>75)%>%
  select(Order_date_of_month,price)%>%
  group_by(Order_date_of_month)%>%
  dplyr::summarise(number=n(),
                   quant1=quantile(price,0.25),
                   quant2=median(price),
                   quant3=quantile(price,0.75),
                   mean=mean(price))
```

```
## # A tibble: 2 x 6
##   Order_date_of_month number quant1 quant2 quant3  mean
##   <fct>                <int>  <dbl>  <dbl>  <dbl> <dbl>
## 1 1st_week                61   110.   200.   400.  314.
## 2 4th_week                37   124.   200.   360.  441.
```

```
com3=filter(compras2,price>75)
ggplot(com3)+
  geom_violin(aes(Order_date_of_month,price))+
  geom_boxplot(aes(Order_date_of_month,price),width=0.1,notch=TRUE)+
  labs(x = "Weeks",y="Price")
```



Another interesting information that we can obtain from the "Order date" is which day of the week are the clients buying.

Before applying the next function, I should clarify that, it transforms a date into the day of the week that the date refers to.

The function has been programed in Spanish and thus, the days of the weeks will be in Spanish. Nonetheless I will offer the following translation to identify which are the days.

#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.

From the data in the following table, we can determine that most of the transactions take place during the workdays. This points out that the business is provably operating online. If the shop were 100% physic, then, most of the sells would tend to group during the weekends.

```
USA_consumer2=USA_consumer
class(USA_consumer$`Order Date`)
```

```
## [1] "POSIXct" "POSIXt"
```

```
Day_con=USA_consumer2%>%
  mutate(wday=wday(`Order Date`,label=TRUE))%>%
  select(wday,Category,Region)

table(Day_con$wday)
```

```
##
##  do  lu  ma  mi  ju  vi  sá
```

```
## 110 938 965 821 909 932 516
```

In the following proportion table, I asked for the percentage of operations of every category that takes place in every day.

For example, as we can see, on "do"(domingo -> Sunday), the shop has sold 2.06% (0.02066487) of the total furniture sold, 2.14% (0.02142629) of the Office Supplies and 2.10% (0.02103049) of the Technology. On "lu"(lunes -> Monday) the business sells 17.78% of the total furniture, 18.19% of office Supplies and 17.98% of the "Technology...

"lu"(lunes -> Monday) and "ma"(martes -> Tuesday) show a very regular distribution of the sells across the different categories. Both days have a high demand. I would recommend the business to make sure there are enough employees to manage all the operations.

"mi"(miercoles -> Wednesday) seems to be a calmer day, especially in the Furniture department which won't need that many people working to attend the demand.

"ju"(jueves -> Thursday) show a high activity in the "Office Supplies" department, and during "vi"(viernes -> Friday) the department that shows a higher level of sells is "furniture". Both days seem to be busy in all the departments, however the two departments that we have highlighted, should be the priority of the business.

```
prop.table(table(Day_con$wday,Day_con$Category),2)
```

```
##
##       Furniture Office Supplies Technology
##   do 0.02066487      0.02142629 0.02103049
##   lu 0.17789757      0.18196354 0.17981073
##   ma 0.19047619      0.18356252 0.18822292
##   mi 0.14645103      0.16021746 0.16508938
##   ju 0.16891285      0.18164375 0.16088328
##   vi 0.20125786      0.17173009 0.17981073
##   sá 0.09433962      0.09945635 0.10515247
```

If we divide the orders of the regions by day, we can complete the information that we saw in the last table.

We now know which are the most active departments in the whole USA, but do all the departments show a similar behavior across the country? This is what we will figure out now.

According to the information we have in the next table, the "Central" and "East" Regions show a greatest activity during Monday, Thursday and Friday and the "South" and "West" Regions, on Tuesday, Thursday and Friday.

I would recommend taking this information into consideration when deciding the schedules of the employees.

```
prop.table(table(Day_con$wday,Day_con$Region),2)
```

```
##
##         Central       East      South       West
##   do 0.00990099 0.02450647 0.03460621 0.01973684
##   lu 0.20462046 0.18243703 0.16229117 0.17105263
##   ma 0.14356436 0.18924438 0.21957041 0.19677033
##   mi 0.16831683 0.15724983 0.13007160 0.16566986
##   ju 0.18151815 0.15724983 0.18257757 0.18241627
##   vi 0.19719472 0.18175630 0.17541766 0.16686603
##   sá 0.09488449 0.10755616 0.09546539 0.09748804
```

We already know the number of operations that take place. However, as we said, this data frame also shows the number of units that a client acquires from a certain product in a single operation.

That is why now I would like to know if there is a substantial difference between the number of units that people buy in every operation across the country.

To find that information I will create another proportion table with the different quantities bought in the 4 regions.

In the following table, we will see that, for example, the "Central" region is responsible for the 22.92%(0.22925764) of the orders that contained just one item. 29.25% of the operations with 1 item are from the East, 17.46% from the South and 30.34% from the West.

As we can see in the table, the "West" Region is the part of the country that consumes the most in almost every case. The level of sells for every number of items, represent normally from a 30% to 40% of the total, but for example the "West" Region holds the 63% of the sells that had 12 elements.

This could have different explanations, maybe the business was created in that region and has more presence or maybe the company policies applied in this territory have been well received by the consumers.

I don't know which is the case, however, if this high consume is due to the policies applied, then, they need to be reproduced in other Regions as well.

```
prop.table(table(USA_consumer$Region,USA_consumer$Quantity),2)
```

```
##
##                    1          2          3          4          5          6
##   Central 0.22925764 0.23923445 0.22411128 0.24188312 0.25856698 0.23183391
##   East    0.29257642 0.31578947 0.28438949 0.25649351 0.24454829 0.32179931
##   South   0.17467249 0.14194577 0.16537867 0.16396104 0.18068536 0.14878893
##   West    0.30349345 0.30303030 0.32612056 0.33766234 0.31619938 0.29757785
##
##                    7          8          9         10         11         12
##   Central 0.21070234 0.22222222 0.19852941 0.20000000 0.27777778 0.18181818
##   East    0.24749164 0.25641026 0.28676471 0.36000000 0.11111111 0.09090909
##   South   0.18394649 0.16239316 0.16911765 0.00000000 0.11111111 0.09090909
##   West    0.35785953 0.35897436 0.34558824 0.44000000 0.50000000 0.63636364
##
##                   13         14
##   Central 0.18750000 0.25000000
##   East    0.18750000 0.31250000
##   South   0.25000000 0.12500000
##   West    0.37500000 0.31250000
```

Now that we know that every region consumes in its particular way, I would like to see if the regions show a different behavior when they are offered different discounts.

I will now create a bar plot coloring the bars based on the region and dividing the plot by the discounts offered.

In this way I will be able to see which are the territories that makes the most use of the discounts and if the territories increase their consume because of the reduced prices.

In the plot we can see that there is one region that shows a strong response to the discounts, the "Central" one. According to the next table of data, most of the orders placed in this Region had some discount applied.

Other regions also make use of discounts, but in the following plot, we can see that the "Central" region show a clear majority of orders placed with discounts of 30% , 60% and 80%.

There are 2 possible lectures of this scenario. 1-Discounts are used to motivate people to buy more. 2-The consumers are not buying and the business uses the discounts to sell the stock and free storage space.

In this case, I assume that the business uses both. It seems to use from 10% to 20% discount to motivate costumers to buy, and from 20% to 80% to delete stocks.

I affirm that, because as we can see the 20% discount is offered to all the regions, and this shows that this can be a regular marketing campaign. Furthermore, is quite normal for shops to offer this kind of discounts.

I think that from 30% to 80% is more aimed to sell the stocks, because it is not so common for a shop to offer this kind of discount. Moreover, most of the operations in this range of discounts, are sold in the "Central" Region. That may mean that this is not a regular policy of the business, but something that was applied because they were forced to.

There still one last information that we can extract from this plot, the number of products bought.

Normally if a business applies a great discount such as 60% or 80%, it is expected to see a different distribution of the quantities for every order.

That happens because the consumer that wanted one chair, considering the new price of chairs, may decide to buy two or three. However, as we can see in the 60% or 80% discount plots, the number of units bought, have a similar distribution as if there were no discounts applied. That shows that they just buy what they need.

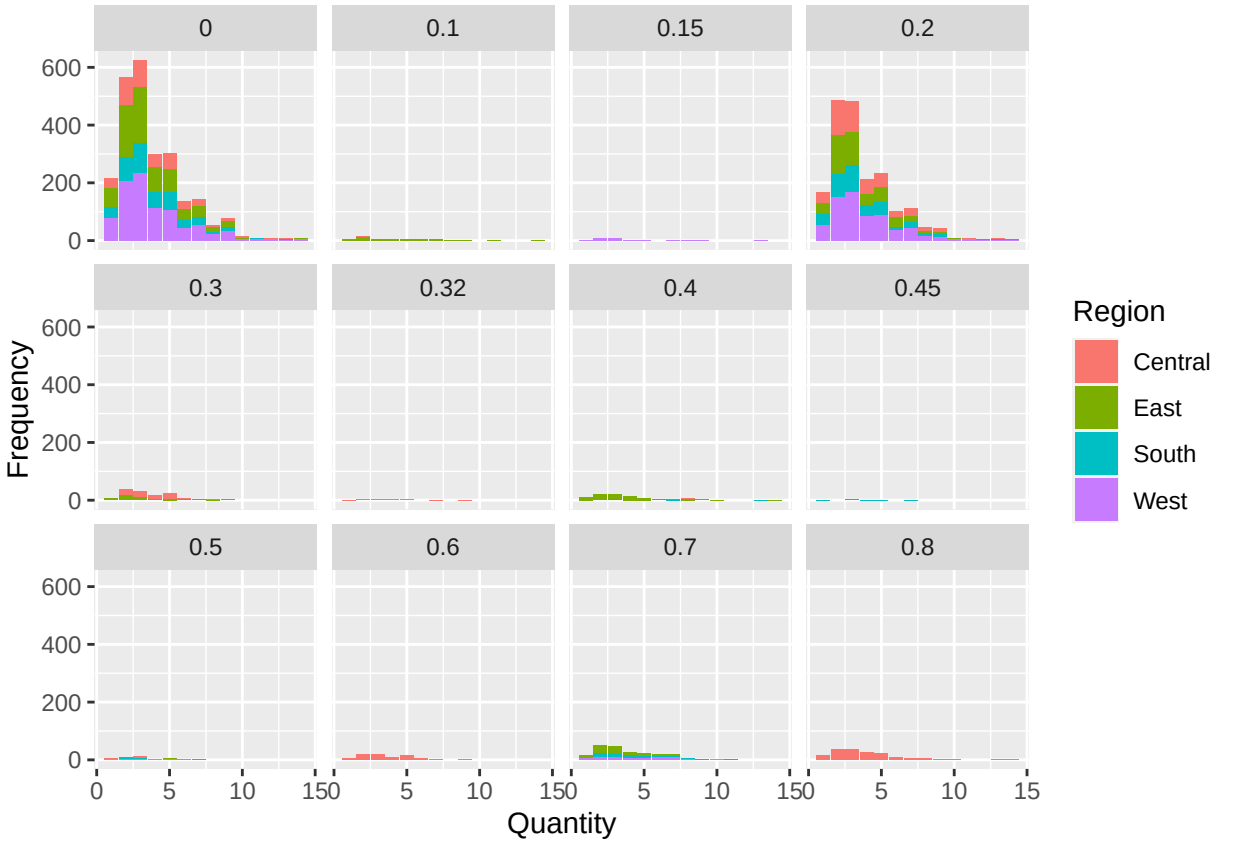
I would require more information about the consumers to make a confident guess. Nonetheless, based on the data I have, I assume that the central region is poorer than the other regions. If this was not the case, I suppose that they would buy more without discount or use the discount to buy more quantity.

One interesting detail of this chart, is that, as we can see, some discounts, like the ones used by the "Central" region, are offered to a specific a territory.

If all the consume was made trough an online portal, the discounts would be used by different regions. The fact that just one specific territory is taking advantage of all the discounts, means that there are some phisical shops offering this deals in the "Central" Region.

This fact is important because we didn't know if the shop used only the online channel or also the phisical shop one.

```
ggplot(USA_consumer)+
  geom_bar(aes(Quantity,fill=Region))+
  facet_wrap(~Discount)+
  labs(x = "Quantity",y="Frequency")
```




```
#Orders placed in the Central Region with the discount offered.
table(USA_consumer[USA_consumer$Region=="Central","Discount"])
```

```
##
##    0  0.1  0.2  0.3 0.32  0.4  0.5  0.6  0.8
##  406    6  443   84   13    6   13   74  167
```

Now, I want to confirm if, as we said, the discounts from 10% until 20% are aiming to motivate the sells, and those that are more than that, have the objective of selling the stocks that the consumers have not bought in the first place.

To do that I will start finding the orders that had a negative profit and looking at their distribution.

If what we said is correct, the number of orders with negative income that have a discount of 20% or less, should be reduced compared with the total of orders offered with these discounts. At the same time, the orders with more than a 20% discount, should be, for the most part, a negative-income product.

As we can see in the following table, most of the orders that are not income positive, have a discount higher than 20%. However, there are 283 orders with losses that have just a 20% discount applied.

```
money_loser=filter(USA_consumer,Profit<0)%>%
  select(Profit,Region,`Product Name`,Discount)

table(money_loser$Discount)
```

```
##
##  0.1 0.15  0.2  0.3 0.32  0.4 0.45  0.5  0.6  0.7  0.8
##    1   12  283  121   13   79    6   37   74  210  167
```

For the next table, I will create a data frame that includes all the orders offered with a discount.

I will divide the number of orders with losses by the total of products offered with discount.

The data show that just a 2% of the products offered with a 10% discount, 41.37% of the products with a 15% discount and 14.77% of the products with a 20% discount, were income negative operations.

Even though this percentages are higher than I anticipated, specially the 41.37%, I must highlight that the products that had more than a 20% discount are, in almost every case, products with losses.

As we can see in the table, all the products offered with 32%,45%,50%,60%,70% and 80% discounts are income negative, and those offered with 30% and 40% are income negative in 91.66% and 82.22% of the cases.

```
discount=filter(USA_consumer,Discount>0)%>%
  select(Profit,Region,`Product Name`,Discount,Category)

table(money_loser$Discount)/table(discount$Discount)
```

```
##
##       0.1      0.15       0.2       0.3      0.32       0.4      0.45       0.5
## 0.0200000 0.4137931 0.1477035 0.9166667 1.0000000 0.8229167 1.0000000 1.0000000
##       0.6       0.7       0.8
## 1.0000000 1.0000000 1.0000000
```

One last detail that I would like to see is how the discounts are applied to the different Categories.

As we can see in the following table, most of the discounts offered are not higher than 20%. This case is specially true in the Technology Category, were just a few items are offered with a 40% discount and a very small quantity of products have even higher discounts than that.

The technology margins tend to be reduced. That would explain why the business can't offer discounts like the ones it is offering in the "Office Supplies" Category.

As we can see in the following data frame that contains only the 70% and 80% discounted products that are part of the "Office Supplies", the business offers a large line of cheap products like for example "Hanging Post Binders". This kind of products tend to have a higher margin, and, to experiment losses in cheap products, is not as financial damaging as it is to sell a machine for a low price.

There is one interesting information that we can take out of this data. It is regarding to the policy of rotation of asses that the business uses.

Products such as machines, food, clothes and in some cases even furniture, have a short life circle, meaning that they need to be sold in a certain period of time or they will be outdated.

This is not the case for Office products, but according to the data that we have, the business is applying the same philosophy to these products as well. As we have seen, when the corporation sees that a product is not selling well, they offer it with great discounts.

That is showing that the business cares to ensure a rotation of its assets, even if that means to concur in extra losses.

```
table(discount$Category,discount$Discount)
```

```
##
##                   0.1 0.15  0.2  0.3 0.32  0.4 0.45  0.5  0.6  0.7  0.8
##   Furniture        44   29  315  131   13   28    6   31   74   11    0
##   Office Supplies   6    0 1153    0    0    0    0    0    0  188  167
##   Technology        0    0  448    1    0   68    0    6    0   11    0
```

```
Ofice_Supplies=filter(discount,Category=="Office Supplies",Discount>0.6)%>%
  select(`Product Name`, Discount)

tail(Ofice_Supplies)
```

```
## # A tibble: 6 x 2
##   `Product Name`                                                Discount
##   <chr>                                                            <dbl>
## 1 Insertable Tab Indexes For Data Binders                            0.7
## 2 Wilson Jones Easy Flow II Sheet Lifters                            0.7
## 3 Insertable Tab Indexes For Data Binders                            0.8
## 4 Storex Dura Pro Binders                                            0.8
## 5 GBC White Gloss Covers, Plain Front                                0.8
## 6 Hoover Replacement Belt for Commercial Guardsman Heavy-Duty Upright ~      0.8
```

I consider that we have already answered the questions that the business asked us at the beginning of this case. However, before finishing the analysis, I would like to have a clear view of which region has reached the greatest number of losses and which region obtains a better total profit.

To do that, I will aggregate the profits by Region and then applying the function "sum".

As we expected, the "Central" region is the one with a greatest negative profit, but it is close to the negative income from the "East" region.

Finally, if we take a look at the total profit, adding both positive and negative incomes, we can see that all the regions are profitable, but the "Central" region obtains a very low result.

```
aggregate(Profit~Region, data=money_loser, FUN=sum)
```

```
##    Region     Profit
## 1 Central -34630.456
## 2    East -27951.649
## 3   South -12468.253
## 4    West  -9895.353
```

```
aggregate(cbind(Profit)~Region, data=USA_consumer, FUN=sum)
```

```
##    Region    Profit
## 1 Central  8564.048
## 2    East 41190.984
## 3   South 26913.573
## 4    West 57450.604
```

To conclude this case of study, I will write two different texts. The first one will explain the information that we now have about the business. The second one, will be more focused on answering the questions of the business and to offer some recommendations.

#About the business:

The enterprise that we have analyzed, is a big company. Based on our findings, it operates online and using physical shops.

This organization seems to be in an expansive phase, because even though it has losses in many of the markets that it operates, it maintains its activity on them. This is a clear behavior of businesses that are trying to increase their market share in countries where they are not yet consolidated.

This business is offering high discounts to eliminate its unwanted stocks. This kind of management is common in industries such as cloth, technology, food, or any other industry that operates with assets that need to be sold in a certain period.

However, this corporation also applies this philosophy to products that do not need to rotate that much, such as "Office Supplies", showing that it is a deliberate action. Furthermore, in this organization, most of the products are sold with a reduced margin.

It is possible to determine that the organization is applying a business strategy. The high rotation and low margins are key parts of a low-cost businesses' philosophy. This behavior is specially indicated for enterprises

that are expanding.

The used strategy gives to the business a high negotiation power with their suppliers. The high volume of products the company is buying, allows it to get better prices and several days of credit. These days of credit, enables the business to sell the products before paying for them, generating a high level of liquidity.

Many of the businesses that apply this strategy use this liquidity to open new shops, and it seems to be that this corporation is also doing it. That may be the reason why it is operating in that many countries even if they are not profitable.

#The business wants to identify which is the country with the greatest income. From that country, the organization wants to identify the most profitable segment of costumers. Finally, it wants to obtain as much information as possible about the costumers in that segment.

The country with the greatest income is the USA. From that country, the most profitable segment is the "consumer segment". The information that we have found about them is the following one:

-They buy more during the working days. The "Central" and "East" has the highest volume of consumption on "Mondays", "Thursdays" and "Fridays" and the "South" and "West" on "Tuesday", "Thursday" and "Friday".

-The costumers buy more during the first half of the month. Furthermore, what they buy, tend to be more expensive. The price of the products of the "Machines" category, seems to be an exception, it is not influenced by when the consumer buys the product.

-The region that buys more is the "West". We have focused on which is the region that consumes the most volume of products in a single operation. The "West" is the responsible for at least 30% to 40% of orders containing from 1 product until 14.

#Recommendations for the business:

-The company need to be careful with the "Central" region. It seems to be that the products are not always well received. That is why in many cases, the business needs to use great discounts to delete the unwanted stocks. The products offered with such discounts, are always income negative. I would recommend reducing the quantity of products that are sold in the shops of the "Central" region to avoid this situation.

-The sub-Category "tables" is the only one that is not performing. I would recommend to delete this sub-Category or to think about how to cover this demand with a substitutive product.

-During our exploration of the data, we have found a line of products where 3 out of the 4 offered references were income negative. With a more exhaustive analysis, the company could identify more of those lines of products and consider if it is worth to keep selling them. I would recommend deleting the specific line that we have found ("Cubify CubeX 3D").