Assignment on the data set bigstoredotcom.

Saba Akram

10/20/2020

library(readxl)

## Warning: package 'readxl' was built under R version 3.6.3

bigstore\_1<-read\_excel("C:/Users/akram/Downloads/bigstore.xlsx",1)  
bigstore\_2<-read\_excel("C:/Users/akram/Downloads/bigstore.xlsx",2)  
bigstore\_3<-read\_excel("C:/Users/akram/Downloads/bigstore.xlsx",3)  
head(bigstore\_1)

## # 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>

## Intrduction of the dataset.

The bigstoredotcom is a dataset seems to be of the online seller of various products. It mentions various variable factors in the dataset such as Row ID, Order ID, Order date, ship date, ship mode,customer ID, customer name, segment, city, state, country, postal code, Market, Region, Product ID, category, sub category, product name, Sales, Quantity, discount, profit, shipping Order and Order Priority in the first sheet of the excel file from which we imported the data.  
In the second sheet, we have Order ID of those who have returned their products.

## What we need to do

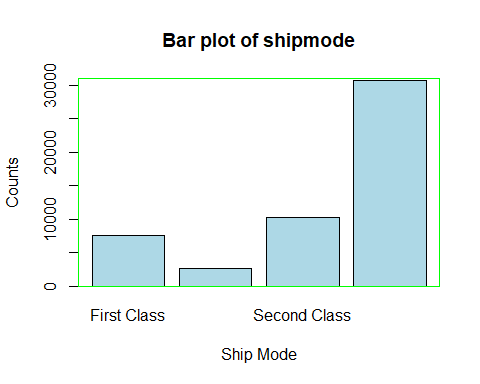
By observing the dataset we got to know that the variables which are identity variables need not to be analysed. However We need to summerise and visualise all the important variables of the bigstore such as shipmode, segment, market, category, sub\_category, sales, quantity, discount, profit, order priority of the "order sheet" and returned variable of the "return sheet". We will try to analyse the patterns of these variables individually, then we will find the target variable and try to show the effect of the variables (predictive variables) to the selected target variable or variables.

## Observing the signifant variables one by one.

table\_shipmode<-table(bigstore\_1$`Ship Mode`)  
table\_shipmode

##   
## First Class Same Day Second Class Standard Class   
## 7505 2701 10309 30775

barplot(table\_shipmode,ylim = c(0,31000),xlab = "Ship Mode",ylab = "Counts",main = "Bar plot of shipmode",col = "lightblue")  
box(which = "plot",lty="solid",col="green")

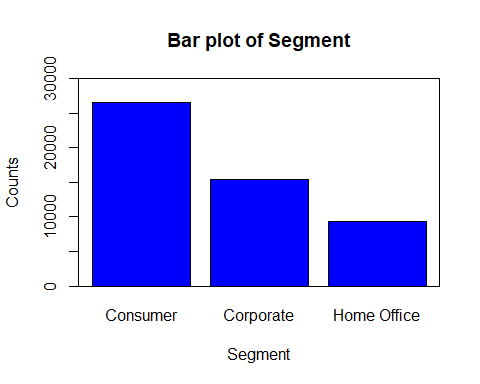


On plotting the bargraph of the 'shipmode' variable indiviadually, we get to know that the customers like the Standard\_Class shipping mode the most and they like Same\_day shipping day the least.  
   
Hense the company should motivate the use of Standard\_class and discaurage the use of same\_day shipping mode. The company should provide an incentive to the customer to use Standard\_Class, First\_Class and Second\_CLass shiping mode and thus can save the time, mnagement and the cost of maintaining the Same\_Day shipping mode system.

table\_segment<-table(bigstore\_1$Segment)  
table\_segment

##   
## Consumer Corporate Home Office   
## 26518 15429 9343

barplot(table\_segment,ylim = c(0,30000),xlab = "Segment",ylab = "Counts",main = "Bar plot of Segment",col = "blue")  
box(which = "plot",lty="solid",col="black")

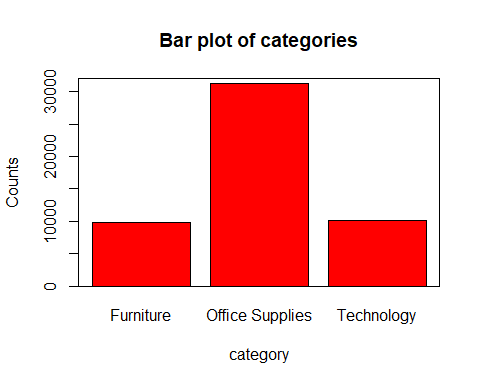


By observing the barplot of the 'segment' variable of the customers, we got to know that the customers from the consumer segment have the maximum number of purchases and the customers from the home office segment have the least number of purchases and the customers from the corporate segment have purchases in the middle of these two.  
  
By observing we got to knoe that the company or the bigstore needs to promote his sales in the Home Office segment and the corporate segment. There may be many reasons behind the low sales of these two segments. They need to take ups issues regarding these like providing offers on purchases etc. so that the customer feels satisfied, comfortable in purchasing fro mthe bigstore.

table\_category<-table(bigstore\_1$Category)  
table\_category

##   
## Furniture Office Supplies Technology   
## 9876 31273 10141

barplot(table\_category,ylim = c(0,32000),xlab = "category",ylab = "Counts",main = "Bar plot of categories",col = "red")  
box(which = "plot",lty="solid",col="black")

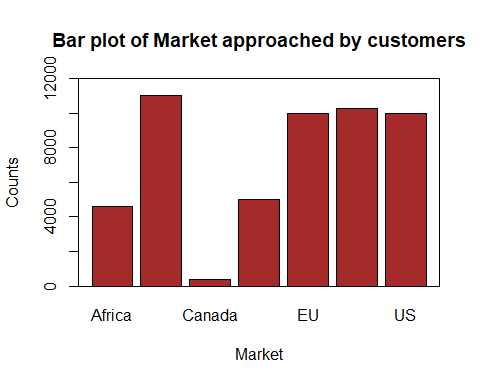


By observing the bar graph of the 'category' variable of the bigstore, we get that the customer purchased the 'Office Supplies' the most but the frequency of the furnitures and the technoligy products are low. Bigstore needs to look at these.

table\_market<-table(bigstore\_1$Market)  
table\_market

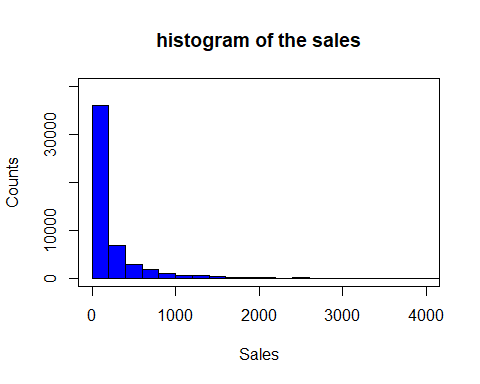
##   
## Africa APAC Canada EMEA EU LATAM US   
## 4587 11002 384 5029 10000 10294 9994

barplot(table\_market,ylim = c(0,12000),xlab = "Market",ylab = "Counts",main = "Bar plot of Market approached by customers",col = "brown")  
box(which = "plot",lty = "solid",col="black")



By observing the above graph of the 'Market Variable', the bigstore must observe that that they have very low reach in the markets of Canada the most, followed by Africa which in turn is followed by EMEA Market.  
The bigstore needs to focus on the cuatomers of these markets so that thaie products reach these markets.

hist(bigstore\_1$Sales,breaks = 100,xlim = c(0,4000),col = "blue",border = "black",ylim = c(0,40000),xlab = "Sales",ylab = "Counts",main = "histogram of the sales")  
box(which = "plot",lty = "solid",col="black")

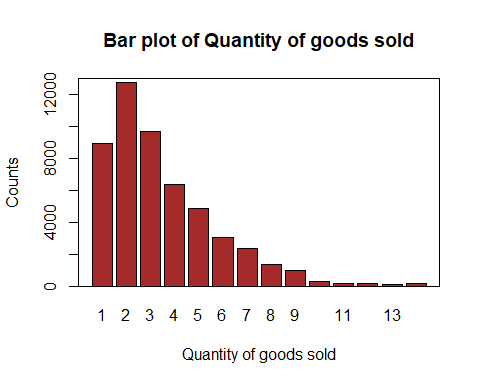


By Observing the above histogram of the 'sales' variable we got to know that maximum number of customers are purchasing products in small quantities i.e. shopping for less amount. As the sales amount is increasing the number of customers are decreasing and beyond a certain limit and there is not any customer who purchases goods worth of more than Rs. 3000.   
Bigstore must observe that it needs to improve its credibility and its brand value and must improve its internal loopholes.

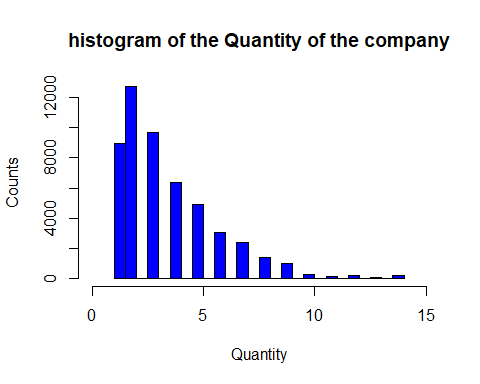
table\_quantity<-table(bigstore\_1$Quantity)  
table\_quantity

##   
## 1 2 3 4 5 6 7 8 9 10 11 12 13   
## 8963 12748 9682 6385 4882 3020 2385 1361 987 276 156 176 83   
## 14   
## 186

barplot(table\_quantity,ylim = c(0,13000),xlab = "Quantity of goods sold",ylab = "Counts",main = "Bar plot of Quantity of goods sold",col = "brown")  
box(which = "plot",lty = "solid",col="black")



hist(bigstore\_1$Quantity,breaks =30,xlim = c(0,15),col = "blue",border = "black",xlab = "Quantity",ylab = "Counts",main = "histogram of the Quantity of the company")

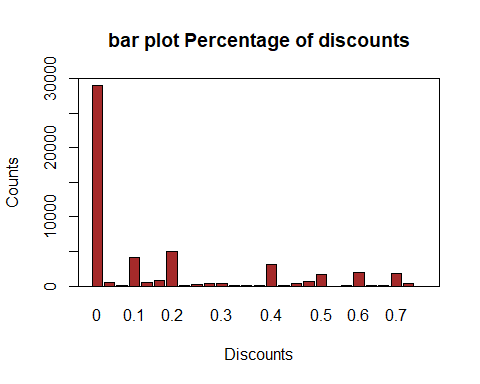


As we observed in the case of 'sales' variable that customers are not purchasing the products worth beyond Rs. 3000 and the graph was showing the decreasing trend from the very beginning.  
Now our observation becomes more powerful after observing this 'Quantity' variable. Here we observe that as the quantity of the products purchased is increasing there is decreasing number of customers. In other words the customers prefer to purchase products from bigstore in less quantities. The observation suggests that bigstore must either improve its credbility or improve its product value.

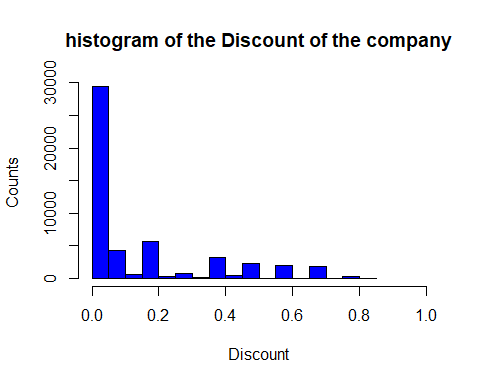
table\_discount<-table(bigstore\_1$Discount)  
table\_discount

##   
## 0 0.002 0.07 0.1 0.15 0.17 0.2 0.202 0.25 0.27 0.3 0.32 0.35   
## 29009 461 150 4068 541 735 4998 41 198 388 340 27 122   
## 0.37 0.4 0.402 0.45 0.47 0.5 0.55 0.57 0.6 0.602 0.65 0.7 0.8   
## 74 3177 104 327 725 1633 10 12 2006 23 17 1786 316   
## 0.85   
## 2

barplot(table\_discount,ylim = c(0,30000),xlab = "Discounts",ylab = "Counts",main = "bar plot Percentage of discounts",col = "brown")  
box(which = "plot",lty = "solid",col="black")

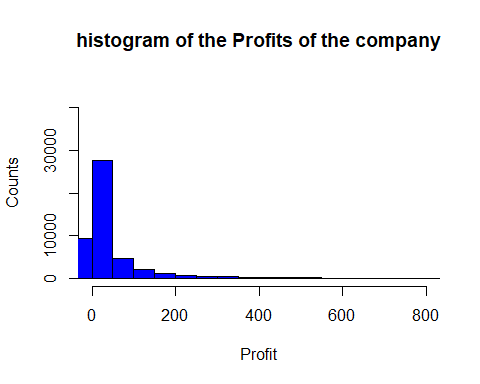


hist(bigstore\_1$Discount,breaks =30,xlim = c(0,1),col = "blue",border = "black",xlab = "Discount",ylab = "Counts",main = "histogram of the Discount of the company")



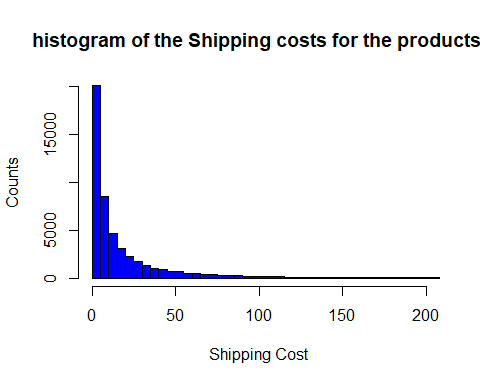
By observing the 'discount' variable we didnot reach to a solid conclusion but only few. We observe that the discount near the 0 is maximum and after that there is a drastic decrease in the frequency as the discouny rate increases. However after that thare is a gradual decrease in the number of frequency as the discount rate decreases.  
We will investigate this variable later in thi assigment with the target variale.

hist(bigstore\_1$Profit,breaks = 300,xlim = c(0,800),col = "blue",border = "black",ylim = c(0,45000),xlab = "Profit",ylab = "Counts",main = "histogram of the Profits of the company")



By observing the 'profit' variable', we observe that there is not a regular pattern in this histigram but we only observe that firstly the counts of the customer is increasing first and then decreases further as the amount of profit for the company increases.  
We wil investigate it further by investigating it by setting a target variable.

hist(bigstore\_1$`Shipping Cost`,breaks = 300,xlim = c(0,200),col = "blue",border = "black",ylim = c(0,20000),xlab = "Shipping Cost",ylab = "Counts",main = "histogram of the Shipping costs for the products.")

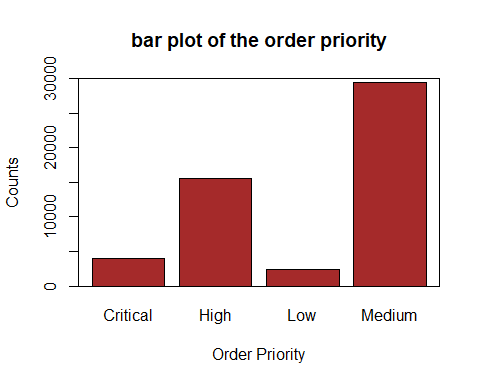


Here we observe that as the shipping cost increases the counts of the customers is continuously decreasing. However we will investigate it further.

table\_orderpriority<-table(bigstore\_1$`Order Priority`)  
table\_orderpriority

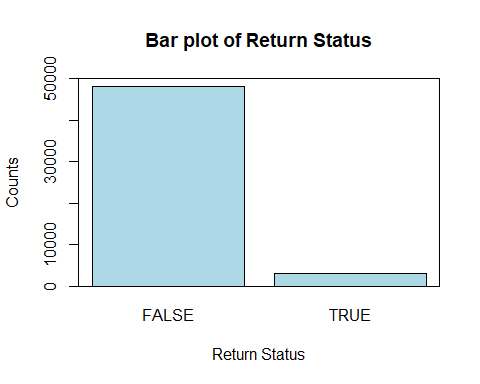
##   
## Critical High Low Medium   
## 3932 15501 2424 29433

barplot(table\_orderpriority,ylim = c(0,30000),xlab = "Order Priority",ylab = "Counts",main = "bar plot of the order priority",col = "brown")  
box(which = "plot",lty = "solid",col="black")



By observing the above bargrph we observe that 'medium order priority counts maximum followed by high followed by critical and then low which is minimum.

x<-c(bigstore\_1$`Order ID`)  
y<-c(bigstore\_2$`Order ID`)  
Returned<-x %in% y  
table\_returned<-table(Returned)  
barplot(table\_returned,ylim = c(0,50000),xlab = "Return Status",ylab = "Counts",main = "Bar plot of Return Status",col = "lightblue")  
box(which = "plot",lty = "solid",col="black")



round(prop.table(table\_returned),4)\*100

## Returned  
## FALSE TRUE   
## 94.05 5.95

Now we had the 'return' variable in the second sheet but we didnot have the the complete information because it had the customer id's of the customers who had returned the products, We thus created a new variable named 'returned' and filled it with status 'TRUE' and 'FALSE' and thus we can observe how much customers have returned the products and how many didnot return. From the graph and the table we observe that 3050 customers have returned their product and 48240 customers didnot return their products. However we will investigate it in details later.

## Selecting the target Variable.

First we select the target variable and use all the significant predictive variables so that we can reach to a certin conclusion.  
 Based on my observation, 'Returned variable in the second sheet will be the most significant target variable and we will do the Exploratory Data Analysis taking 'Return' as the target variable. We will match all the order ID's and their return status. I have done it in the above section and observed that out of 51290 customers,3050 i.e.5.95% have returned their products whereas 48240 i.e 94.05% didnot returned their products. We will add the return variable in the order sheet which will make our task easy. We will use both methods general method and also we will use the ggplot for the Exploratory Data Analysis.

library(ggplot2)

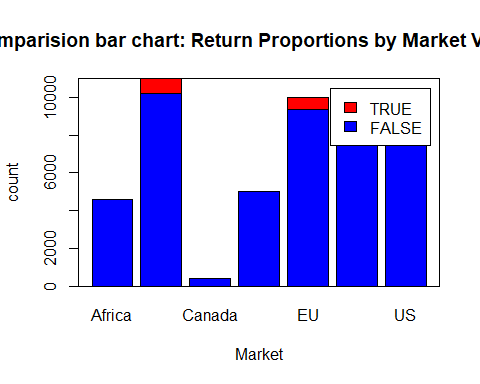
## Warning: package 'ggplot2' was built under R version 3.6.3

bigstore\_reor<-cbind(bigstore\_1,Returned) # Combining in one sheet

return\_market<-table(Returned,bigstore\_1$Market,dnn = c("Returned","Market"))  
return\_market

## Market  
## Returned Africa APAC Canada EMEA EU LATAM US  
## FALSE 4587 10216 384 5029 9363 9469 9192  
## TRUE 0 786 0 0 637 825 802

barplot(return\_market,legend=rownames(return\_market),col = c("blue","red"),ylim = c(0,11000),ylab = "count",xlab = "Market",main = "Comparision bar chart: Return Proportions by Market Variable")  
box(which = "plot",lty="solid",col="black")



addmargins(return\_market,FUN = sum)

## Margins computed over dimensions  
## in the following order:  
## 1: Returned  
## 2: Market

## Market  
## Returned Africa APAC Canada EMEA EU LATAM US sum  
## FALSE 4587 10216 384 5029 9363 9469 9192 48240  
## TRUE 0 786 0 0 637 825 802 3050  
## sum 4587 11002 384 5029 10000 10294 9994 51290

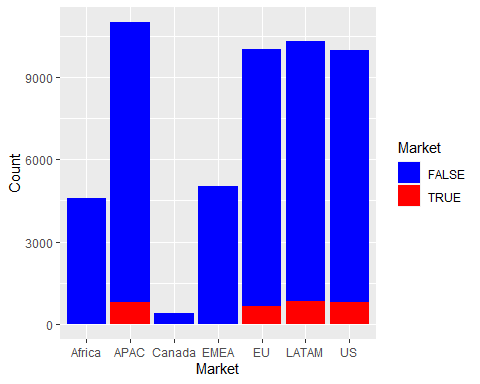
round(prop.table(return\_market,margin = 1),4)\*100

## Market  
## Returned Africa APAC Canada EMEA EU LATAM US  
## FALSE 9.51 21.18 0.80 10.42 19.41 19.63 19.05  
## TRUE 0.00 25.77 0.00 0.00 20.89 27.05 26.30

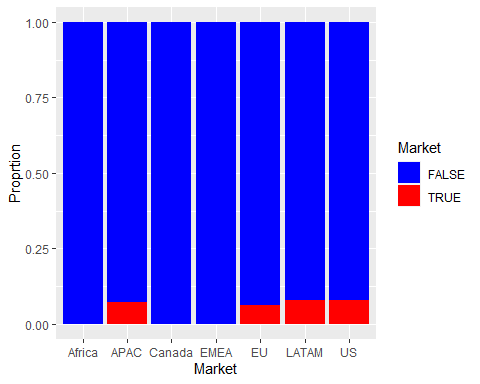
round(prop.table(return\_market,margin = 1),4)\*100

## Market  
## Returned Africa APAC Canada EMEA EU LATAM US  
## FALSE 9.51 21.18 0.80 10.42 19.41 19.63 19.05  
## TRUE 0.00 25.77 0.00 0.00 20.89 27.05 26.30

#Or  
  
ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Market),fill=factor(Returned)),position = "stack")+scale\_x\_discrete("Market")+scale\_y\_continuous("Count")+guides(fill=guide\_legend(title = "Market"))+scale\_fill\_manual(values = c("blue","red"))



ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Market),fill=factor(Returned)),position = "fill")+scale\_x\_discrete("Market")+scale\_y\_continuous("Proprtion")+guides(fill=guide\_legend(title = "Market"))+scale\_fill\_manual(values = c("blue","red"))

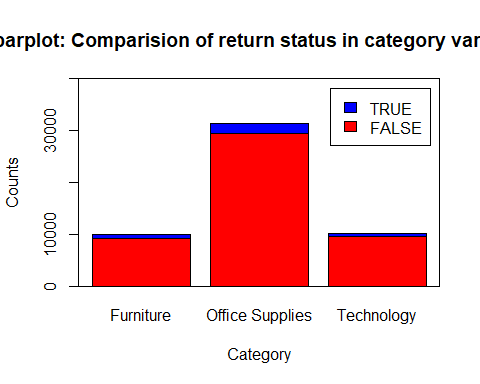


By plotting the bargraph, sumtable and row.margin above we observe that there are no returns in the markets of Africa, Canada, EMEA. However certain markets such as APAc, EU, LATAM and US have returned the products in certain proportions approx 20% to 27%. So the strategy should be to focus on these markets and observe the reason of returning the products.

return\_category<-table(Returned,bigstore\_1$Category,dnn = c("Returned","Category"))  
return\_category

## Category  
## Returned Furniture Office Supplies Technology  
## FALSE 9238 29485 9517  
## TRUE 638 1788 624

barplot(return\_category,ylim = c(0,40000),xlab = "Category",ylab = "Counts",col = c("Red","blue"),legend=rownames(return\_category),main = "barplot: Comparision of return status in category variable")  
box(which = "plot",lty="solid",col="black")



addmargins(return\_category,FUN = sum)

## Margins computed over dimensions  
## in the following order:  
## 1: Returned  
## 2: Category

## Category  
## Returned Furniture Office Supplies Technology sum  
## FALSE 9238 29485 9517 48240  
## TRUE 638 1788 624 3050  
## sum 9876 31273 10141 51290

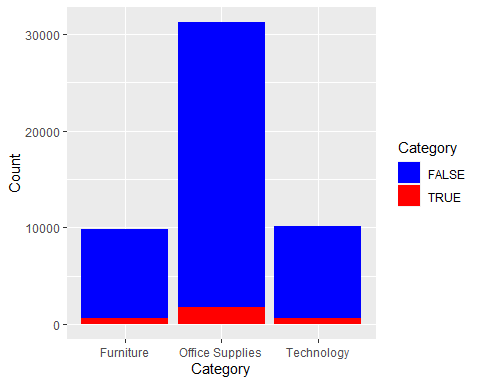
round(prop.table(return\_category,margin = 1),4)\*100

## Category  
## Returned Furniture Office Supplies Technology  
## FALSE 19.15 61.12 19.73  
## TRUE 20.92 58.62 20.46

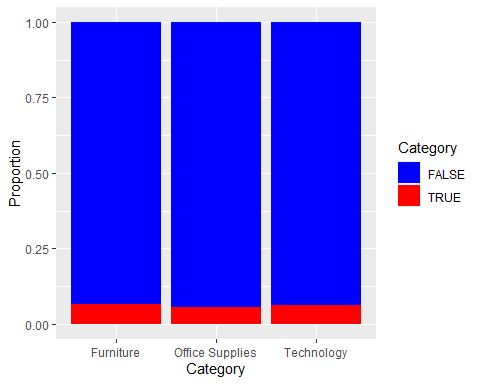
round(prop.table(return\_category,margin = 2),4)\*100

## Category  
## Returned Furniture Office Supplies Technology  
## FALSE 93.54 94.28 93.85  
## TRUE 6.46 5.72 6.15

#Or  
  
  
ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Category),fill=factor(Returned)),position = "stack")+scale\_x\_discrete("Category")+scale\_y\_continuous("Count")+guides(fill=guide\_legend(title = "Category"))+scale\_fill\_manual(values = c("blue","red"))



ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Category),fill=factor(Returned)),position = "fill")+scale\_x\_discrete("Category")+scale\_y\_continuous("Proportion")+guides(fill=guide\_legend(title = "Category"))+scale\_fill\_manual(values = c("blue","red"))

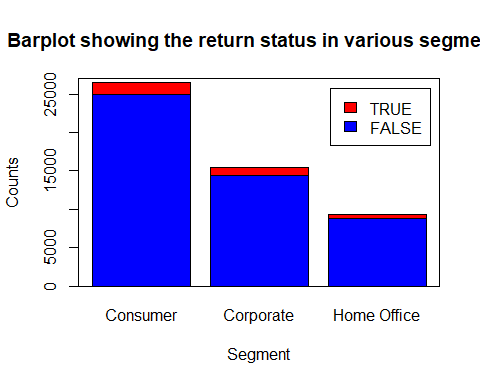


By the above observations we get that in each category approx 6% of the products returned.

return\_segment<-table(Returned,bigstore\_1$Segment,dnn=c("Returned","Segment"))  
return\_segment

## Segment  
## Returned Consumer Corporate Home Office  
## FALSE 24936 14458 8846  
## TRUE 1582 971 497

barplot(return\_segment,xlab = "Segment",ylab = "Counts",main = "Barplot showing the return status in various segments",col = c("blue","red"),ylim = c(0,27000),legend=rownames(return\_segment))  
box(which = "plot",lty="solid",col="black")



addmargins(return\_segment,FUN = sum)

## Margins computed over dimensions  
## in the following order:  
## 1: Returned  
## 2: Segment

## Segment  
## Returned Consumer Corporate Home Office sum  
## FALSE 24936 14458 8846 48240  
## TRUE 1582 971 497 3050  
## sum 26518 15429 9343 51290

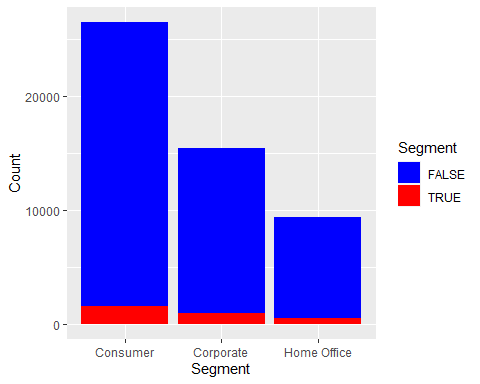
round(prop.table(return\_segment,margin = 1),4)\*100

## Segment  
## Returned Consumer Corporate Home Office  
## FALSE 51.69 29.97 18.34  
## TRUE 51.87 31.84 16.30

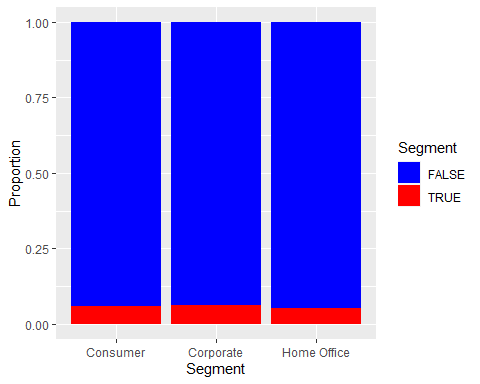
round(prop.table(return\_segment,margin = 2),4)\*100

## Segment  
## Returned Consumer Corporate Home Office  
## FALSE 94.03 93.71 94.68  
## TRUE 5.97 6.29 5.32

#Or  
  
ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Segment),fill=factor(Returned)),position = "stack")+scale\_x\_discrete("Segment")+scale\_y\_continuous("Count")+guides(fill=guide\_legend(title = "Segment"))+scale\_fill\_manual(values = c("blue","red"))

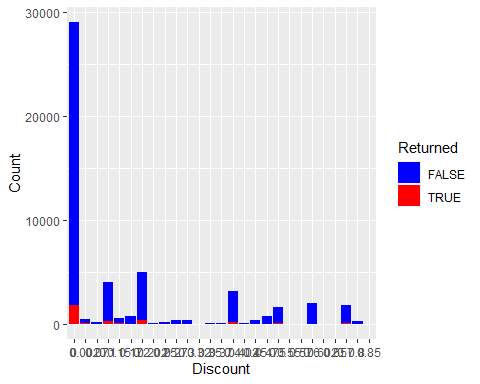


ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Segment),fill=factor(Returned)),position = "fill")+scale\_x\_discrete("Segment")+scale\_y\_continuous("Proportion")+guides(fill=guide\_legend(title = "Segment"))+scale\_fill\_manual(values = c("blue","red"))

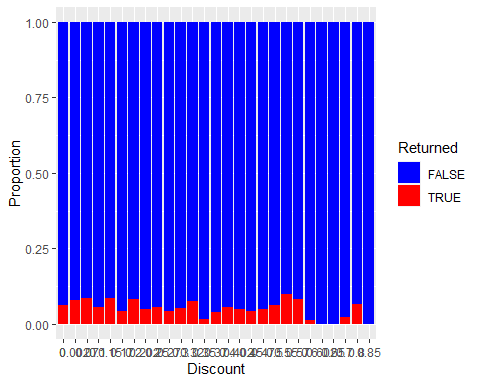


By observing the above table and graph we observe that in the segment also approx 6% of the product from each segment.

ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Discount),fill=factor(Returned)),position = "stack")+scale\_x\_discrete("Discount")+scale\_y\_continuous("Count")+guides(fill=guide\_legend(title = "Returned"))+scale\_fill\_manual(values = c("blue","red"))



ggplot()+geom\_bar(data = bigstore\_reor,aes(x=factor(Discount),fill=factor(Returned)),position = "fill")+scale\_x\_discrete("Discount")+scale\_y\_continuous("Proportion")+guides(fill=guide\_legend(title = "Returned"))+scale\_fill\_manual(values = c("blue","red"))

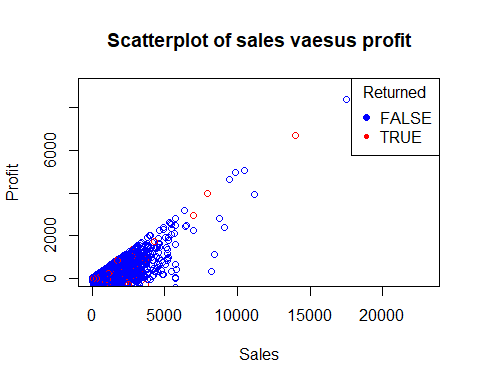


From the above graphwe observe that discount doesnot affect the return, because we donot see any regular pattern in the discount.

## Multvariate Analysis

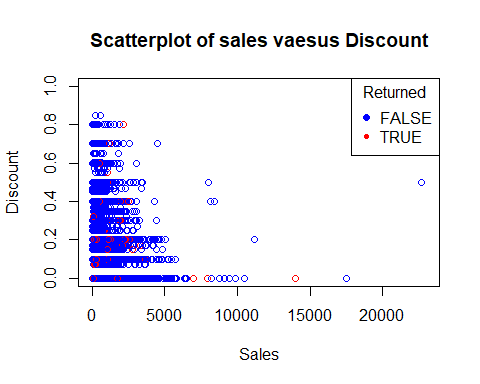
We will take 'Return' as the target variable against two variables namely sales and profit below.

plot(bigstore\_reor$Sales,bigstore\_reor$Profit,xlim = c(0,23000),ylim = c(0,9000),xlab = "Sales",ylab = "Profit",main = "Scatterplot of sales vaesus profit",col=ifelse(bigstore\_reor$Returned==TRUE,"red","blue"))  
  
legend("topright",c(rownames(table(bigstore\_reor$Returned))),col = c("blue","red"),pch = c(16,20),title = "Returned")



By observing the above graph which is the scatterplot of the profit and sales against the target variable 'Return'. We pbserve that the proportion of increase in sales is greater than the proportional increase in profit also the frequency of the 'return' is the greatest in the lower bottom of the graph.

plot(bigstore\_reor$Sales,bigstore\_reor$Discount,xlim = c(0,23000),ylim = c(0,1),xlab = "Sales",ylab = "Discount",main = "Scatterplot of sales vaesus Discount",col=ifelse(bigstore\_reor$Returned==TRUE,"red","blue"))  
  
legend("topright",c(rownames(table(bigstore\_reor$Returned))),col = c("blue","red"),pch = c(16,20),title = "Returned")



By observing the above graph which is the scatterplot of the sales and discount proportion against the target variable 'Returned'. We observe that just above the first line from below we get the red plot which mentions the frequency of the returns.

## Conclusion

We have investigated various variables individually like segment,category,market, quantity,sales etc. and reached to some conclusions. we have also taken a target variable 'return' and tied to to Exploratory Data Analysis. Also we did the multivariate analysis to the best of my knowledge.

## The End