ALY6000 Introduction to Analytics
Northeastern University

Prof. - Dr. Dee Chiluiza
Student's Name - Suprit Mestry

Date: 30 October,2021

Project Report

Library Data

```
# Libraries
library(ggplot2)
library(readxl)
library(knitr)
library(kableExtra)
library(RColorBrewer)
library(colorspace)
library(dplyr)
sales2020 <- read_excel("sales2020.xlsx")
xd = sales2020
```

Introduction

General Topic:

In today's market, there is much more demand for data analysts. Data analytics helps businesses grow with enhanced and optimized performances. It helps to derive and generate beautiful insights and dashboards. For example, In a energy Industry, data analysts are required to calculate and analyze the power usage delivery with the proper investment of the revenue used for the generation of power. Analysts in this industry, mainly focuses on having high productivity and efficiency in power delivery with using less number of resources. They also predict which source of energy like wind, hydro, nuclear or solar will be useful with respect to the need of power and revenue required. Thus this idea signifies the need of Data Analysts.

Data set:

In this project the data set file used is "sale2020.xlsx" which is a modern database. The data set contain data which represents user data of products sold to customers. The data is arranged in columns of date, order id, customer id, customer name to the respective city, state, country, region and many other cost and sale statistics respectively.

Problem Identification:

In a company, working as an analyst if were given this dataset of sales the one entity on which i would have focused will be "Loss per Return". According to me this is the important factor to look into as this factor will import the business of the goods on big scale.

Plan:

The plan in which Loss per Return can overcomes with few methods. Manufacturing only the products which have huge demand in the market and reducing the manufacturing of products while have less demand. Also icrease the use of goods which are cheap to produce and have have return value.

References:

Frankenfield, J. (2021, October 22). What is data analytics? Investopedia. Retrieved October 31, 2021, from https://www.investopedia.com/terms/d/data-analytics.asp (https://www.investopedia.com/terms/d/data-analytics.asp)

Analysis Section

Task 1

```
Mean = c(round(mean(xd$Price),2),round(mean(xd$Quantity),2),round(mean(xd$Sales_Total),2),round(mean(xd$Net_Sale)
,2),round(mean(xd$Profits),2))

Sd = c(round(sd(xd$Price),2),round(sd(xd$Quantity),2),round(sd(xd$Sales_Total),2),round(sd(xd$Net_Sale),2),round(
sd(xd$Profits),2))

Max = c(round(max(xd$Price),2),round(max(xd$Quantity),2),round(max(xd$Sales_Total),2),round(max(xd$Net_Sale),2),r
ound(max(xd$Profits),2))

Min = c(round(min(xd$Price),2),round(min(xd$Quantity),2),round(min(xd$Sales_Total),2),round(min(xd$Net_Sale),2),r
ound(min(xd$Profits),2))

Range = c(Max-Min)

Median = c(round(median(xd$Price),2),round(median(xd$Quantity),2),round(median(xd$Sales_Total),2),round(median(xd
$Net_Sale),2),round(median(xd$Profits),2))

Row = c(Mean, Sd, Range, Median)
Var = matrix(Row, nrow = 4, byrow=TRUE)

calc= c("Mean","Sd", "Range", "Median")
Variables = c("Profits", "Net_Sale", "Price", "Sales_Total", "Quantity")

colnames(Var) = Variables
rownames(Var) = calc
kable(Var, align = "ccccc") %>%
kable_styling(bootstrap_options = c("striped", "hover"), position = "center")%>%
  column_spec(1, bold = "T", background = "#E1C16E")%>%
row_spec(0,background = "#5F9EA0")
```

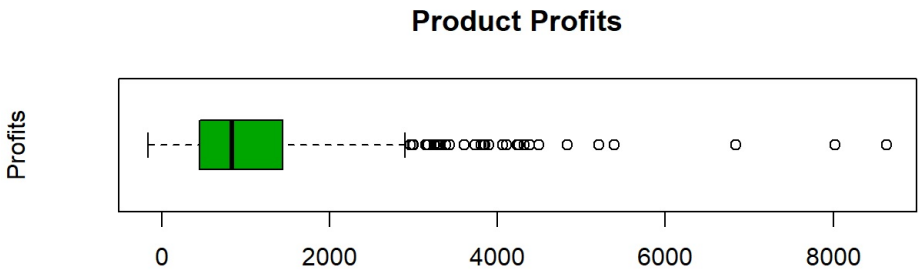
	Profits	Net_Sale	Price	Sales_Total	Quantity
Mean	548.41	6.61	3107.40	2283.54	1066.54
Sd	293.95	2.53	2574.06	1691.58	901.22
Range	3406.28	14.00	28934.83	17448.71	8799.69
Median	500.84	6.00	2481.89	1915.69	834.35

Task 2.1

```
par(mfcol=c(2,1), mar=c(4, 4, 4, 4))

boxplot(xd$Profits,
        col = terrain.colors(3),
        ylab = "Profits",main = "Product Profits",
        horizontal = T)

hist(xd$Profits,
     col = heat.colors(15),
     xlab = "Profits",
     xlim = c(0,3600),
     ylim = c(0,150),
     las=1,
     breaks = 60,
     main = "Product Profits")
```



Summary

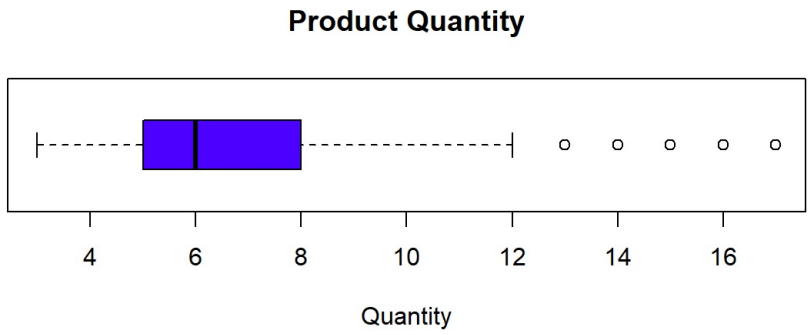
It can be inferred from the output graphs that the majority of the profit of products lies between the range of “0 to 2000”. Here profits can be increased with reducing the sale value of products and also giving more discount on the respective products.

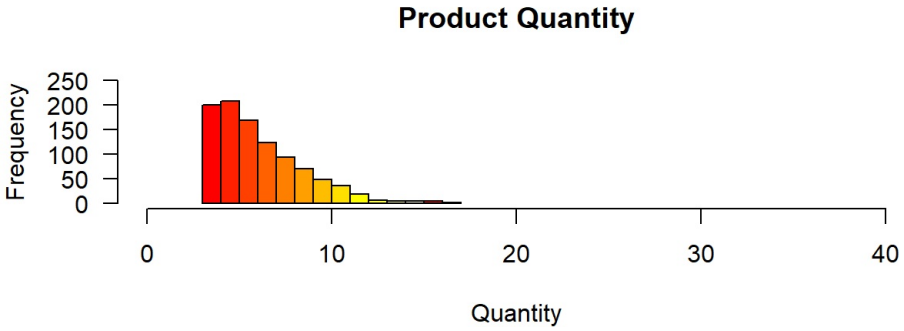
Task 2.2

```
par(mfcol=c(2,1), mar=c(4, 4, 4, 4))

boxplot(xd$Quantity,
        col = topo.colors(1),
        xlab = "Quantity",
        main = "Product Quantity",
        horizontal = T)

hist(xd$Quantity,
     col = heat.colors(12),
     xlab = "Quantity",
     xlim = c(0,40),
     ylim = c(0,250),
     las=1,
     breaks = 10,
     main = "Product Quantity")
```





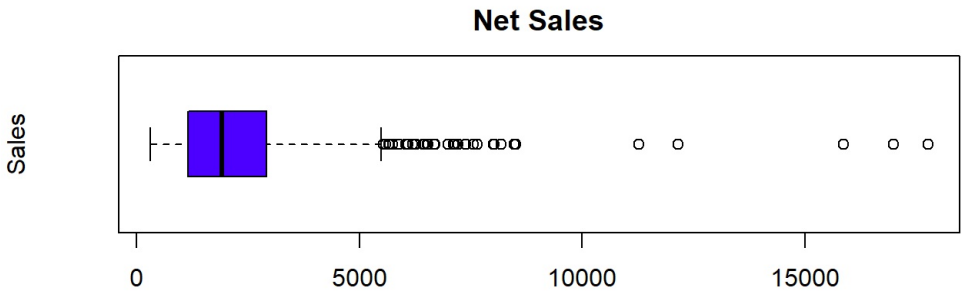
Summary
It can be inferred from the output that the frequency of the number of quantity of products purchase is decreasing gradually between 0 to 20. Product quantity can be increased with increasing manufacturing and also reducing selling price.

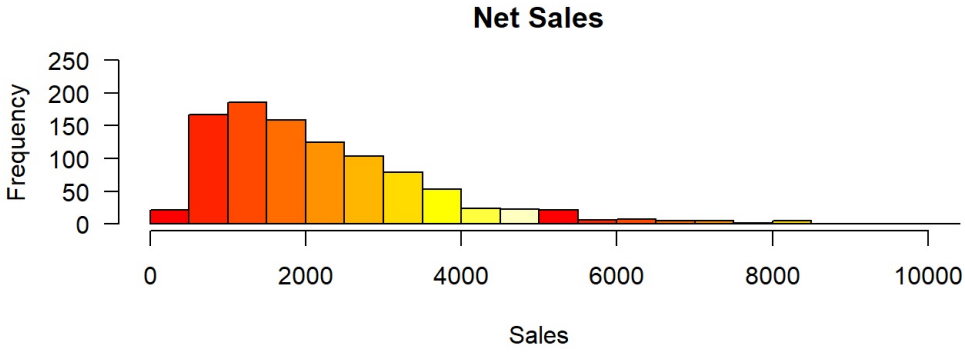
Task 2.3

```
par(mfcol=c(2,1), mar=c(4, 4, 2.5, 2.55))

boxplot(xd$Net_Sale,
        col = topo.colors(1),
        ylab = "Sales",main = "Net Sales",
        horizontal = T)

hist(xd$Net_Sale,
     col = heat.colors(10),
     xlab = "Sales",
     xlim = c(0,10000),
     ylim = c(0,250),
     las=1,
     breaks = 50,
     main = "Net Sales")
```





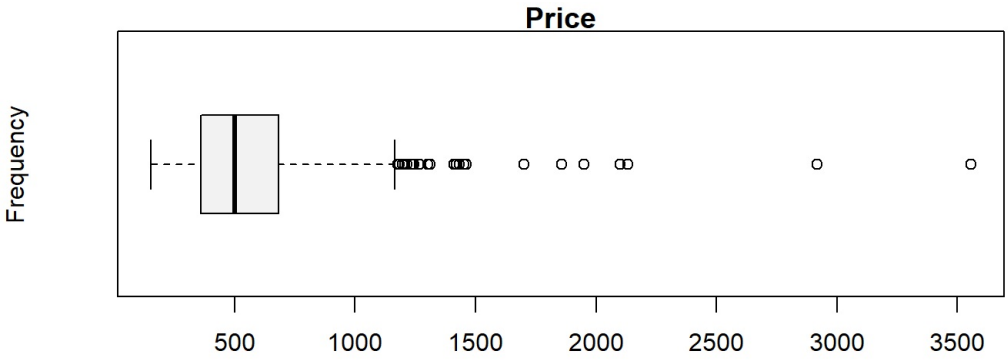
Summary
It can be inferred from the output that the Net Sales are maximum in the range of 1000 and then gradually decreases. Even Net Sales can be increased with high yeild and sale of products.

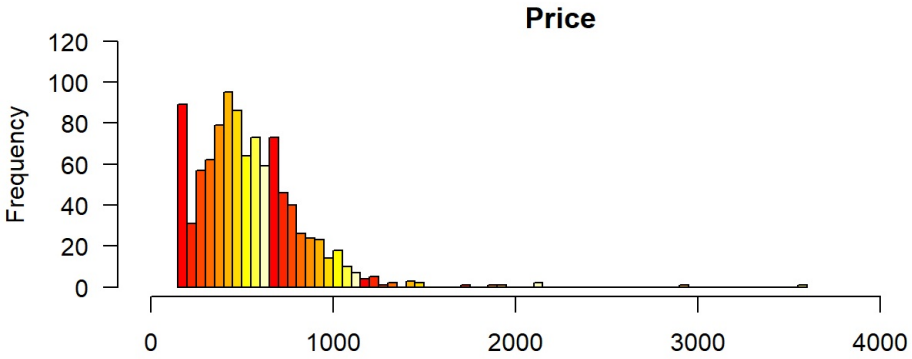
Task 2.4

```
par(mfcol=c(2,1), mar=c(2.5, 4, 1, 1))

boxplot(xd$Price,
        col = terrain.colors(1),
        ylab = "Frequency",main = "Price",
        horizontal = T)

hist(xd$Price,
     col = heat.colors(10),
     xlab = "Value",
     xlim = c(0,4500),
     ylim = c(0,120),
     las=1,
     breaks = 85,
     main = "Price")
```





Summary

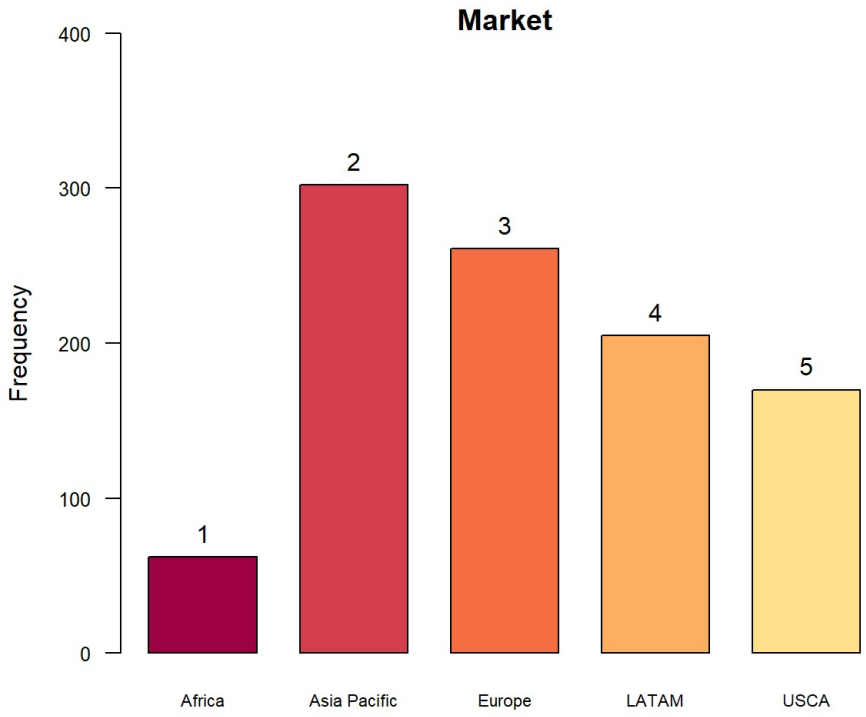
It can be inferred from the output that the maximum price has peak value at 500 and ranges between 200 to 2000. Price can be increased with high yeild of products and increasing number of sales.

Task 3.1

```
par(mar=c(3, 5, 1, 4))
Market=table(xd$Market)
S = barplot(Market,
                main="Market",
                ylab="Frequency",

                col = brewer.pal(11,"Spectral"),
                las=1,
                ylim = c(0,400),
                cex.names=0.7,
                cex.axis=0.8,
                space=0.4)

text(y=Market,
     S,
     pos = 3)
```



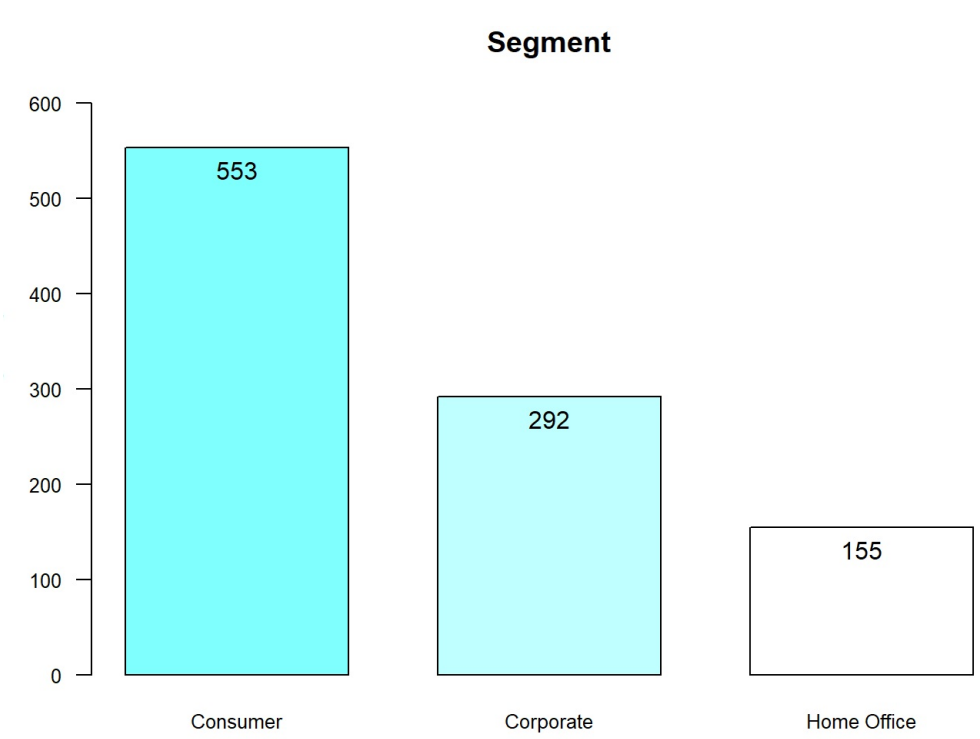
Summary
From the above graph we understand that which Market has high number of yield of the products.

Task 3.2

```
par(mar=c(3, 3, 1, 1))
Segment=table(xd$Segment)
T = barplot(Segment,
                main="Segment",
                ylab="Frequency",

                col = cm.colors(5),
                las=1,
                ylim = c(0,650),
                cex.names=0.8,
                cex.axis=0.8,
                space=0.4)

text(y=Segment,
     T,
     Segment,
     pos = 1)
```



Summary

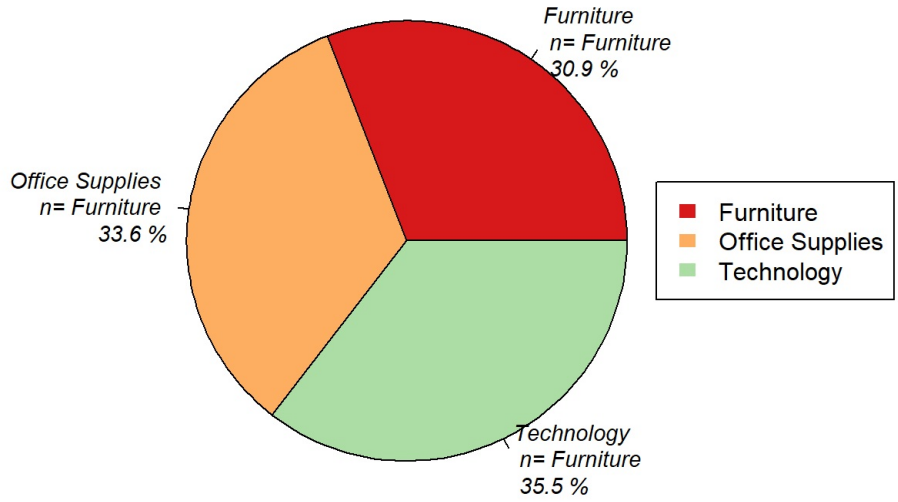
From above graph we understand the number of frequency at which the entity is repeated in the segment.

Task 3.3

```
par(mar=c(1, 1, 1, 1))
Depart=table(xd$Department)
perc = (Depart/sum(Depart))*100
pieLabels = paste(unique(sort(xd$Department)),
                  "\n",
                  "n=",
                  sort(xd$Department),
                  "\n",
                  round(perc,1),
                  "%")

pie(Depart,
    labels = pieLabels,
    radius = 0.7,
    col = brewer.pal(4,"Spectral"),
    lty = 1,
    cex=0.9,
    font = 3
    )

par(mar=c(1, 1, 1, 1))
legend("right",
       legend = paste(unique(sort(xd$Department))),
       fill = brewer.pal(4,"Spectral"),
       border = "white",
       cex = 1,
       )
```

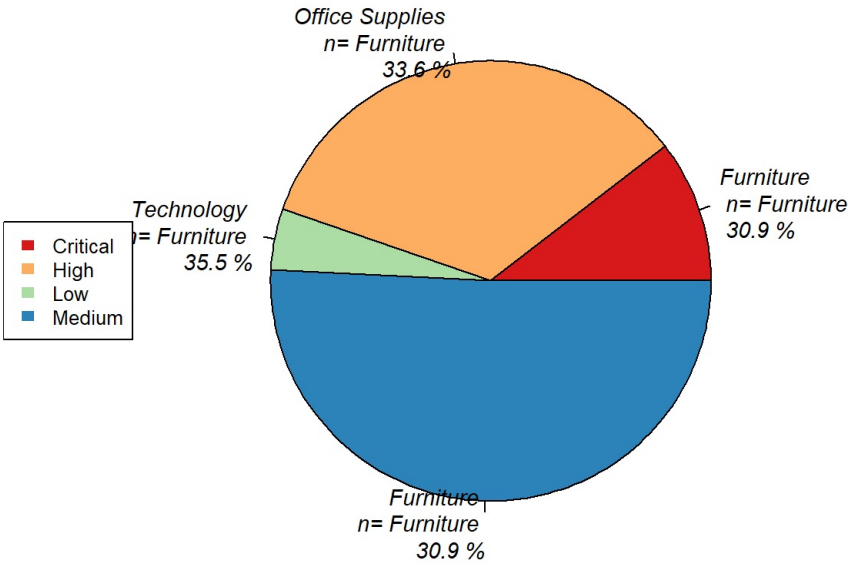


Summary

From above graph we understand in which department the products are been used in terms of pie chart and the percentage respectively.

Task 3.4

```
par(mar=c(1, 1, 1, 1))
OrderPriority=table(xd$OrderPriority)
D_percent = (OrderPriority/sum(OrderPriority))*100
pie(OrderPriority,
    labels = pieLabels,
    radius = 0.7,
    col = brewer.pal(4,"Spectral"),
    lty = 1,
    cex=0.9,
    font = 3
    )
pieLabels = paste(unique(sort(xd$OrderPriority)),
                  "\n",
                  "n=",
                  sort(xd$OrderPriority),
                  "\n",
                  round(D_percent,1),
                  "%")
par(mar=c(1, 1, 1, 1))
legend("left",
       legend = paste(unique(sort(xd$OrderPriority))),
       fill = brewer.pal(4,"Spectral"),
       border = "white",
       cex = 0.8
       )
```

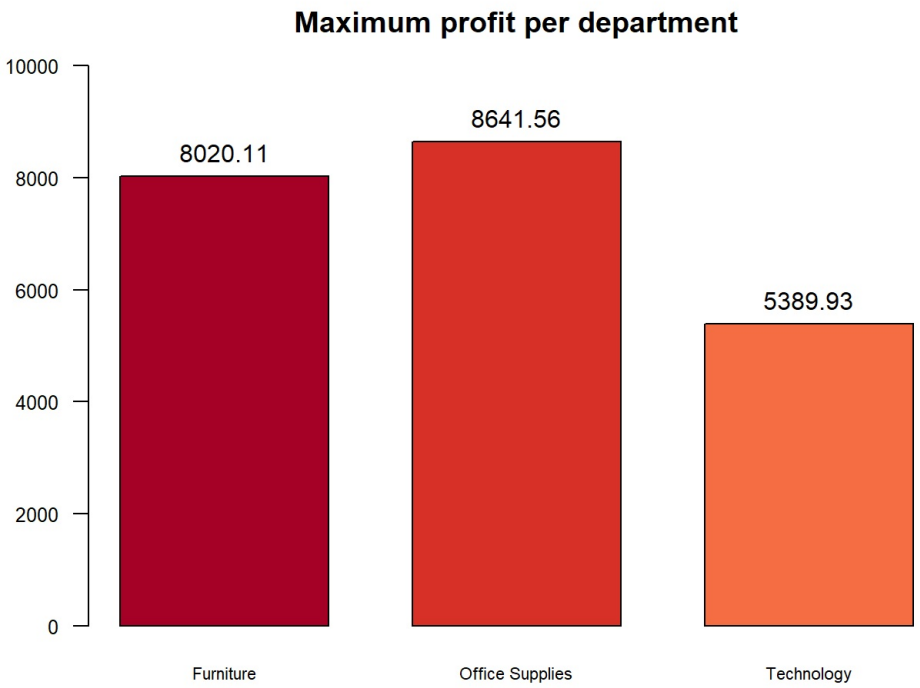


Summary
From above graph we understand the order priority at which the products are been used in terms of pie chart and the percentage respectively.

Task 4.1

Question - What is the Maximum profit per Department?
Logic - It states the the maximum profit that each department would make after the sale of products.

```
par(mar=c(3, 3, 3, 3))
M=round(tapply(xd$Profits, INDEX = xd$Department, FUN = max),2)
A = barplot(M,
                main="Maximum profit per department",
                xlab = "Profit",
                col = brewer.pal(11,"RdYlBu"),
                las=1,
                ylim = c(0,10000),
                cex.names=0.7,
                cex.axis=0.8,
                space=0.4)
par(mar=c(1, 1, 1, 1))
text(y=M,
     T,
     M,
     pos = 3)
```



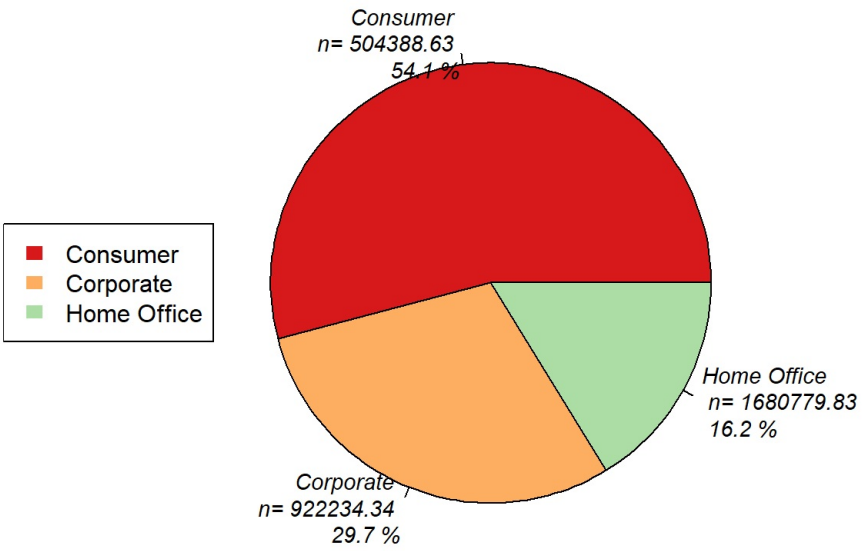
Observation - We can conclude the way it states maximum value of profit made in each department.

Task 4.2

Question - What is the Sum of Total Sale per Segment?
Logic - It states which segment has the total number of sale value through pie chart.

```
par(mar=c(1, 1, 1, 1))
N=round(tapply(xd$Sales_Total, INDEX = xd$Segment, FUN = sum),2)
perC = (N/sum(N))*100
pieLabels = paste(unique(sort(xd$Segment)),
                  "\n",
                  "n=",
                  sort(N),
                  "\n",
                  round(perC,1),
                  "%")

pie(N,
    labels = pieLabels,
    radius = 0.7,
    col = brewer.pal(4,"Spectral"),
    lty = 1,
    cex=0.9,
    font = 3
    )
par(mar=c(1, 1, 1, 1))
legend("left",
       legend = paste(unique(sort(xd$Segment))),
       fill = brewer.pal(4,"Spectral"),
       border = "white",
       cex = 1
       )
```

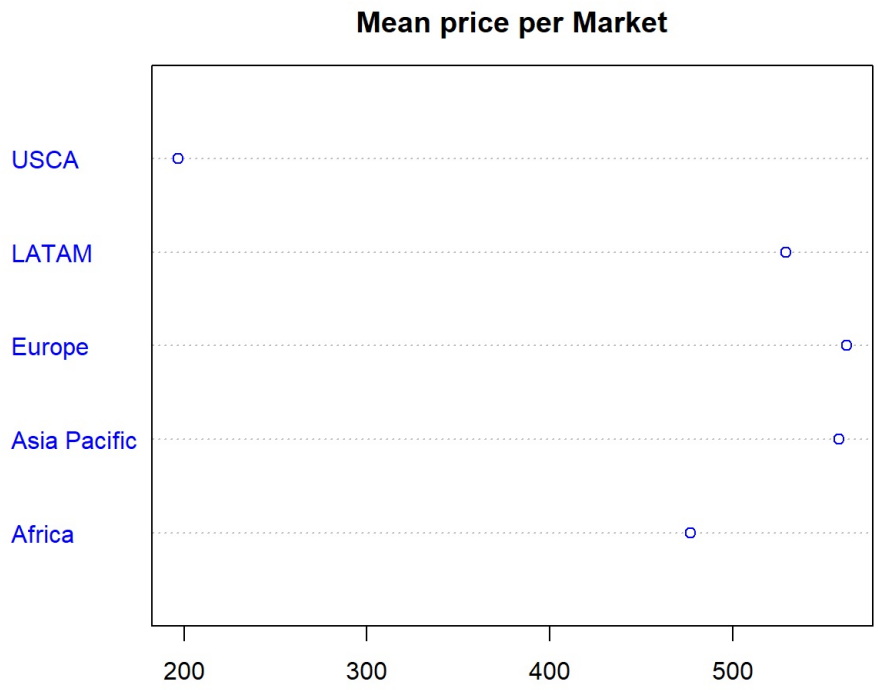


Observation - We can conclude the way it states maximum value in each segment.

Task 4.3

Question - What is the Mean price of each in the Market?
Logic - It represents wthe mean price of the products of each Market.

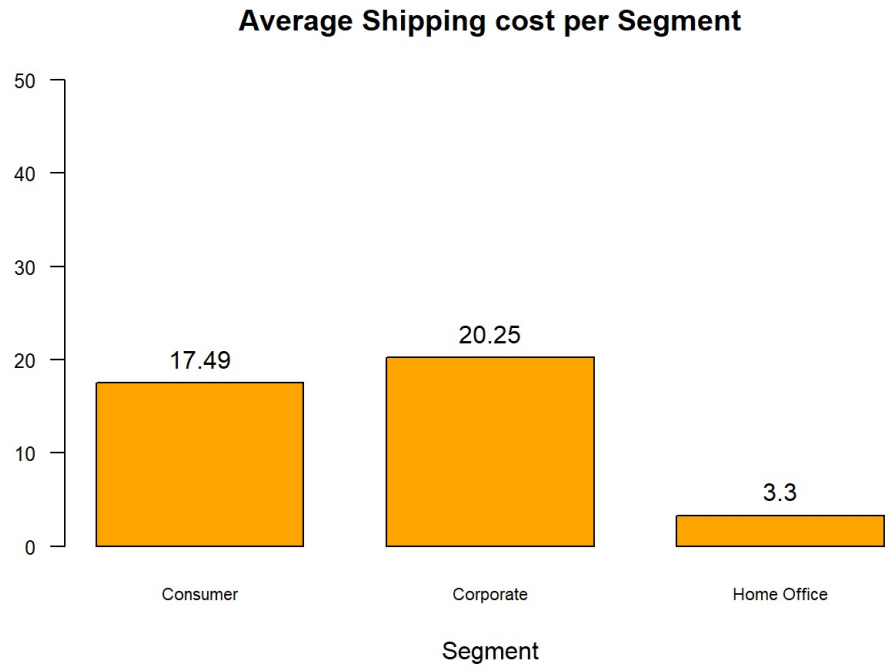
```
par(mar=c(3, 3, 3, 3))
P=round(tapply(xd$Price, INDEX = xd$Market, FUN = mean,2))
E = dotchart(P,
                main = "Mean price per Market",
                xlab = "Mean",
                col = "blue",
                las = 1,
                ylim = c(0,10000),
                cex.names = 0.7,
                cex.axis = 0.8,
                space = 0.4)
par(mar=c(1, 1, 1, 1))
text(y=P,
     T,
     P,
     pos = 4)
```



Observation - We can conclude the way it represents market price.

Task 5.1

```
Q = dplyr::filter(xd, Region=="Western US")
G = round(tapply(Q$ShippingCost_Product, INDEX = Q$Segment, FUN = mean),2)
Z = kable(G,"pipe")
K = barplot(G,
                main="Average Shipping cost per Segment",
                xlab = "Segment",
                las=1,
                col = "orange",
                ylim = c(0,50),
                cex.names=0.7,
                cex.axis=0.8,
                space=0.4)
text(y=G,
     K,
     G,
     pos = 3)
```

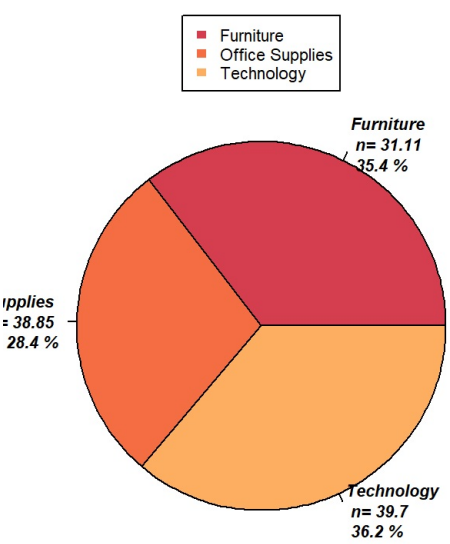


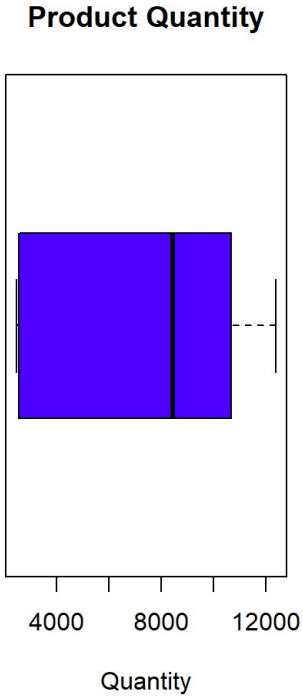
Observation - The above graph represents average shipping cost for each of the segment. If shipping cost is reduced, the sale price can be increased thus will help generate high revenue.

Task 6

Question - What is the Mean price of each in the Market?

```
par(mfcol= c(1,2),mar= c(2, 2, 2, 2))
pl=round(tapply(xd$ShippingCost_Product, INDEX = xd$Department, FUN = mean),2)
kl=tapply(xd$ShippingCost_Product, INDEX = xd$Market, FUN = sum)
per = (pl/sum(pl))*100
pieLabels = paste(unique(sort(xd$Department)),
                  "\n",
                  "n=",
                  sort(pl),
                  "\n", round(per,1), "%")
pie(pl,
    labels = pieLabels,
    radius = 1,
    col = brewer.pal(9,"Spectral"),
    lty = 1,
    cex = 0.7,
    font = 4
    )
par(mar=c(1, 1, 1, 1))
legend("top",
       legend = paste(unique(sort(xd$Department))),
       fill = brewer.pal(9,"Spectral"),
       border = "white",
       cex = 0.7,
       )
par(mar=c(4, 4, 4, 4))
boxplot(kl,
        col = topo.colors(1),
        xlab = "Quantity",
        main = "Product Quantity",
        horizontal = T)
```





Observation - The above graph represents the similarities from different graphs of same variable which is department. Quantity and Shipping Cost are two numerical variables which are related to each other with the department variable.Thus this task helps to analyse two different graphs with different approach to analyse with same categorical variable.

CONCLUSION

From this final project report we concluded the complete descriptive statistics learned and many other aspects of R programming which helped to analyse multiple task analysis and generating visualizations and understanding them deeply. I got familiar with many function of libraries dplyr(), ggplot2(), RColorBrewer() and many more. We created function within the analysis and generated barplots, boxplots, piecharts, histograms and dotcharts. We created table with knitr:kable() function and learned to import datasets from excel and csv files with readxl() & readcsv() functions respectively.

For example, From Task 5.1, we understood in a given segment, the Corporate segment has the highest average shipping cost among all. Also in task 4.1 we focused on the barplot which represented relation between the department and the maximum profit made in that field which is in “Office supplies”. Thus in the same way we concluded all the Task Analysis in the similar way of preentng via visualizations and insights.

REFERENCES

1 - Frankenfield, J. (2021, October 22). What is data analytics? Investopedia. Retrieved October 31, 2021, from https://www.investopedia.com/terms/d/data-analytics.asp (https://www.investopedia.com/terms/d/data-analytics.asp)

2 - National Center for Ecological Analysis and Synthesis. (n.d.). Retrieved October 31, 2021, from https://www.nceas.ucsb.edu/sites/default/files/2020-02/Product-Communications-Strategy-Template.docx (https://www.nceas.ucsb.edu/sites/default/files/2020-02/Product-Communications-Strategy-Template.docx)

3 - Functions. R. (n.d.). Retrieved October 31, 2021, from https://www.tutorialspoint.com/r/r_functions.htm (https://www.tutorialspoint.com/r/r_functions.htm)

APPENDIX

The following “M6_Project.Rmd” file in included and attached in the submission report.