MODULE 6 PROJECT – DATA ANALYSIS

ALY 6000 – Introduction to Analytics

Submitted To: Prof. Richard He

Submitted By: Aabhas Maru

NEU ID: 002955865

maru.a@northeastern.edu

January 19, 2022

INTRODUCTION

Why Data Analytics is Important in Sales.

A company's success can be determined by its customer base, growing sales figure, growing free cash flows or by its bottom line that is, its net profit. These are one of the key indicators to assess a company's success. The roles of data analytics is very vital in these situations. The information generated by the data analyst helps the company's to cater their customers better than competitor, understanding consumer purchasing habit, helps customer retention, provide customer churning info, helps in identifying weak and strong items, overall cost reduction and better market segmentation.

Dataset Description:
The given data set contains 1000 rows and 15 columns. Plus, it also contains 9 numerical value and 6 categorical variables. The given data set contains data of various products available in different continents. The name of the continents mentioned in the data set are; Africa, LATAM, Asia Pacific, Europe, and USCA.

Part I:

1. Summarize the data in a table.

```
> summary(sales2020)
   OrderID            OrderDate           ShipDate            ProductID
 Length:1000        Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


  CustomerID        CustomerName           City               State
 Length:1000        Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


   Country            Region             Market             Segment
 Length:1000        Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


  Department          Division          ProductName        OrderPriority
 Length:1000        Length:1000        Length:1000        Length:1000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


   ShipMode            Price             Quantity            Discount
 Length:1000        Min.   : 151.7     Min.   : 3.000     Min.   :0.0000
 Class :character   1st Qu.: 360.8     1st Qu.: 5.000     1st Qu.:0.0075
 Mode  :character   Median : 500.8     Median : 6.000     Median :0.1100
                    Mean   : 548.4     Mean   : 6.612     Mean   :0.1175
                    3rd Qu.: 683.6     3rd Qu.: 8.000     3rd Qu.:0.2100
                    Max.   :3558.0     Max.   :17.000     Max.   :0.3000


 ShippingCost_Product  Sales_Total           Returns          LossPerReturn
 Min.   : 1.09        Min.   :  349.3     Min.   :0.000     Min.   : 163.9
 1st Qu.:39.09        1st Qu.: 1455.8     1st Qu.:0.000     1st Qu.: 389.7
 Median :40.62        Median : 2481.9     Median :1.000     Median : 540.9
 Mean   :36.55        Mean   : 3107.4     Mean   :1.251     Mean   : 592.3
 3rd Qu.:42.27        3rd Qu.: 4006.7     3rd Qu.:2.000     3rd Qu.: 738.3
 Max.   :84.28        Max.   :29284.1     Max.   :6.000     Max.   :3842.6
   Total_loss           Net_Sale            Profits
 Min.   :    0.0      Min.   :  307.5     Min.   :-158.1
 1st Qu.:    0.0      1st Qu.: 1168.5     1st Qu.: 455.8
 Median :  578.3      Median : 1915.7     Median : 834.3
 Mean   :  823.9      Mean   : 2283.5     Mean   :1066.5
 3rd Qu.: 1103.9      3rd Qu.: 2914.1     3rd Qu.:1440.4
 Max.   :11527.9      Max.   :17756.2     Max.   :8641.6
> ^
```
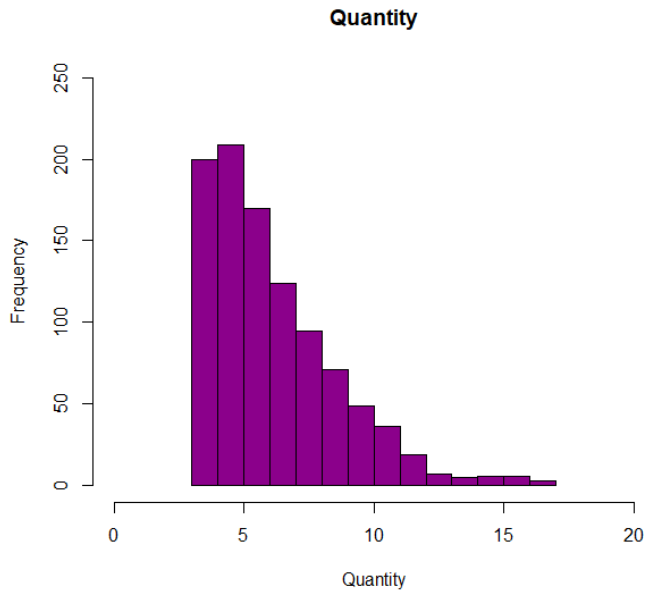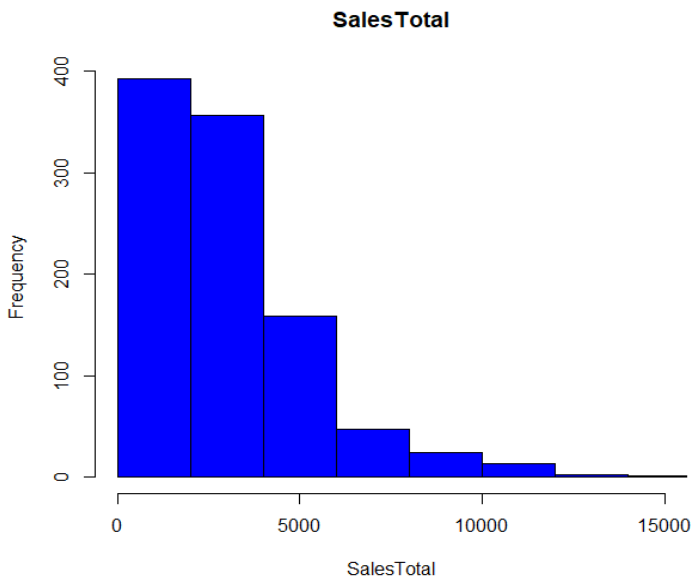
2. Visualizing the Data with the help of graph.

Quantity

Observation: As we can see that, as the quantity increases frequency starts to decrease. This because of the nature if the things been sold. This indicates that discount rates and profit margin need to be calculated on the quantity and not on the volume.

SalesTotal

Observation: We can observe that as soon as starts to increase frequency starts to decrease. This indicate that at pace at which the sales of a product increase the frequency at which product sold decreases.
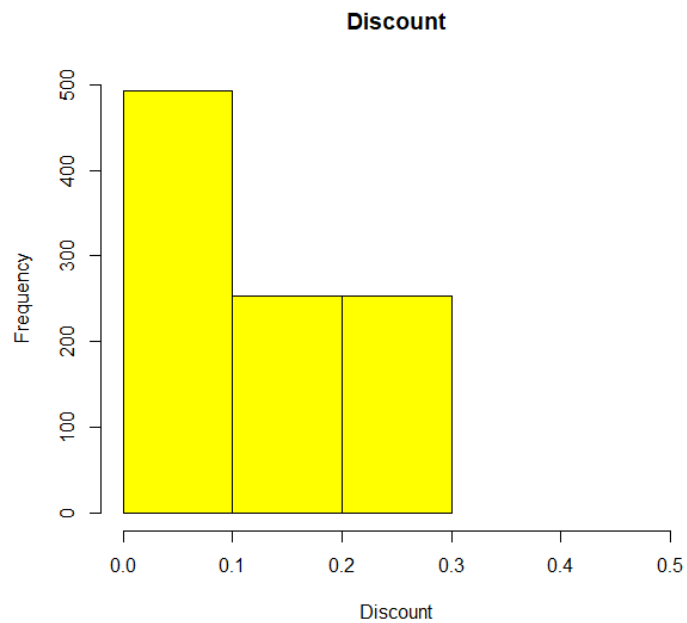
Discount

Observation: We can easily understand by looking at the graph, that the highest number of products offer discounts which lies between 0.0 to 0.1 (0% to 10%) and the other lies between 0.1 to 0.3 (10% to 30%).
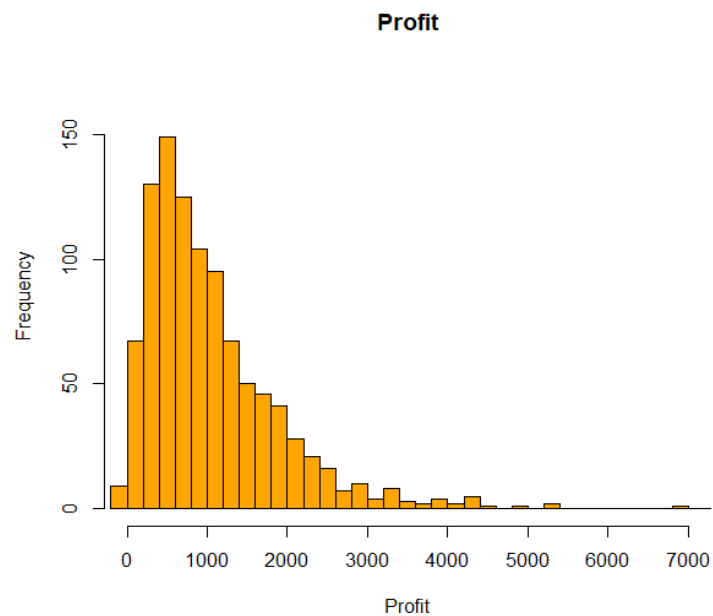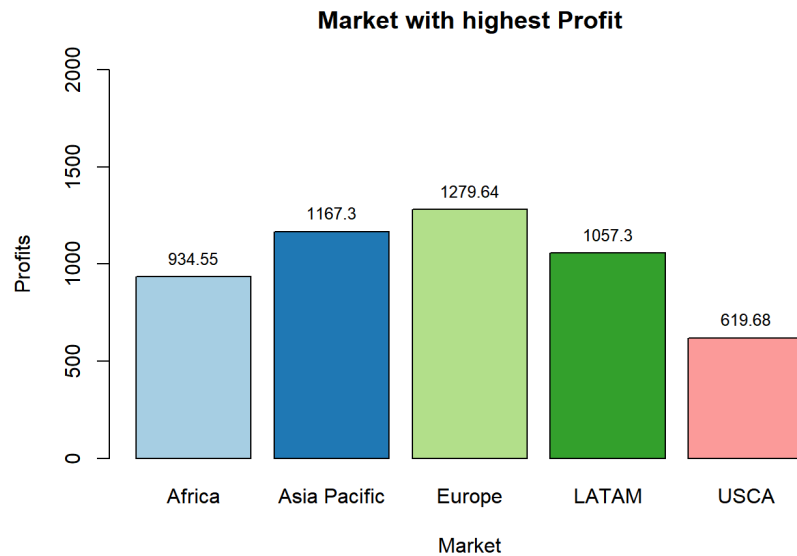
Profit

Observation: It indicates that product used to give good amount of profit between 0 to 1400 and after that it suddenly starts to decline and never covered up.
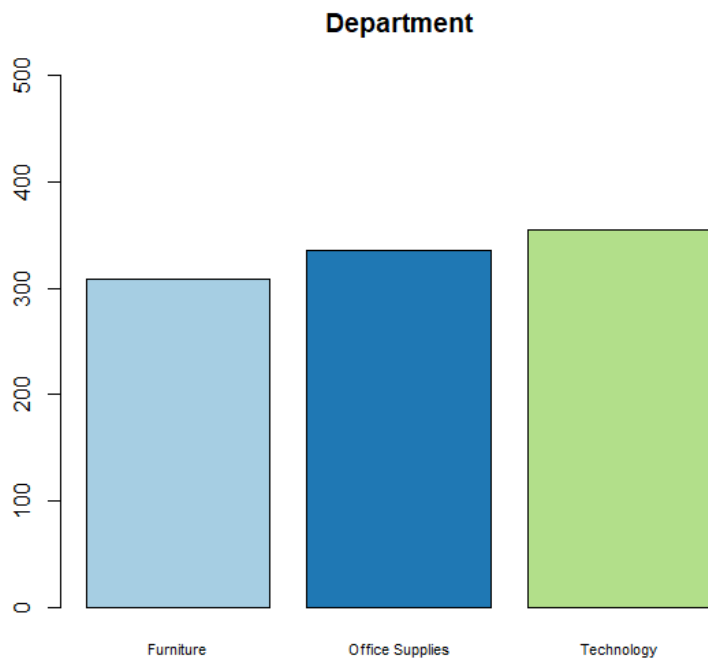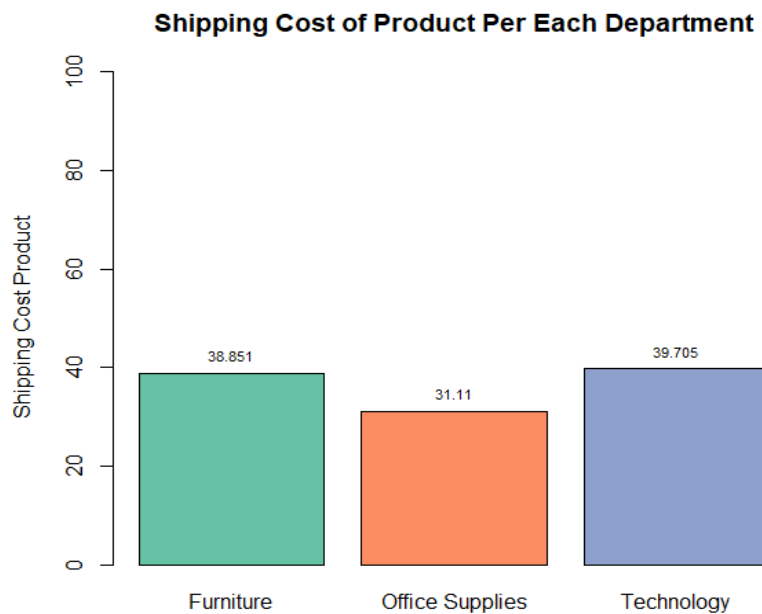
Department

Observation: We can observe that Technology has the highest department (360), then come Office Supplies (320) and Furniture (300).
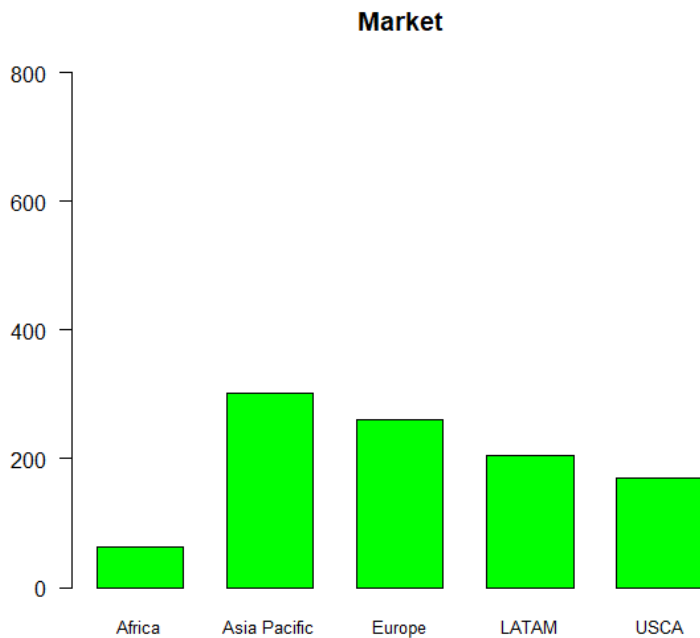
Market

Observation: We can see that Asia Pacific has the highest demand of product and Africa being the lowest.

Market with highest Profit

Observation: As the chart portray that Europe market is the most profitable market, then comes Asia, LATAM, Africa and at last USCA.

Shipping Cost of Product Per Each Department

Observation:

We can see that Technology department has the highest amount of shipping cost, then comes Furniture and then Office Supplies.

Profit in Each Department
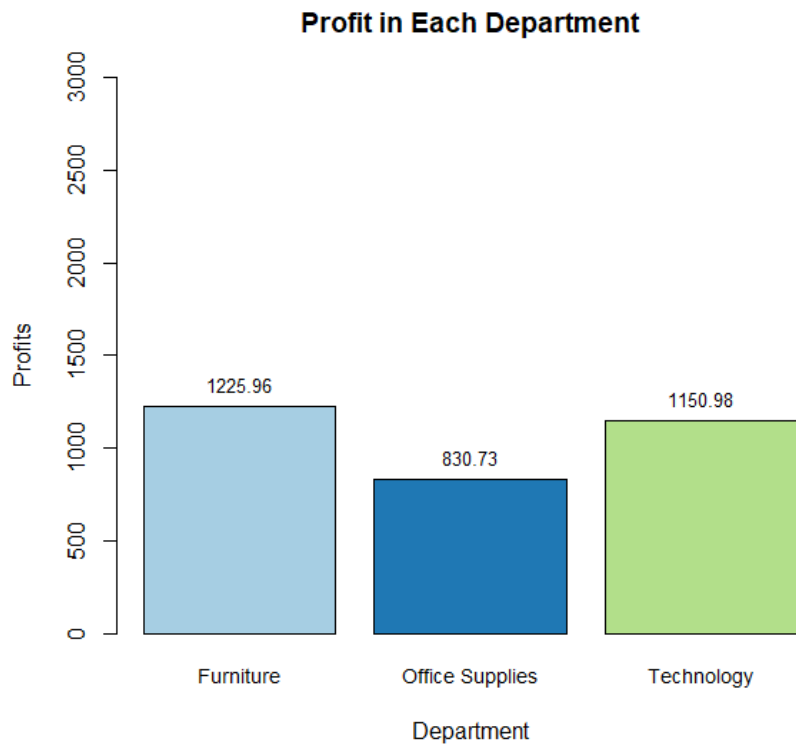
Observation: Furniture is the most profitable department, then Technology and then Office Supplies.

3. What data is telling?

It has provided with the information that which department is generating good profit for us and which department needs to be focused.

It has also provided the information regarding Technology department has the highest shipment cost product.

It also provided with the market which has highest customer base and generating most profit. Plus, also provided the data about the least profit generating market and indicates that it needs to be focused on.

Part II

```
> descriptive_statistics
                       Mean        Median
SalesTotal             6.61200     2.5345374
Quantity               3107.40286  2574.0623301
Discount               0.11746     0.1004402
Profit                 1066.54488  901.2249241
ShippingCostProduct    36.55309    13.1965054
> ^
```

It shows the mean, median, standard deviation, range and variance of the variables.

The mean provides the knowledge about the average of the numerical values. In the given analysis mean has allowed me to analyze the average of sales and let me know how different or a particular market is affecting the mean. Same in case of discount, profit and shipping cost.

It provides information about how your overall profit is in term of all markets and help you to draw attention towards the market which is affecting overall mean to drop. Same in case of Shipping cost.

Part III:

The data is telling that there is a issue and if resolved will provide an competitive edge over our competitor. The problem lies when we analyzed the market capture of different market, there we saw that Asia Pacific has highest market share. But, European market is the one which is most profitable though have less market share then Asia Pacific, as compared to Asia Pacific market.

This problem can be caused because of multiple reasons, consumer purchasing power, economic policy etc. If the household of the people is high, then they will be willing to spend more, and you don't have to provide discounts. We can further understand the Asian Market and tries to find different option to boost up the sales. Another thing that we can do to increase the profit, is to try to cut down the shipping expenses and which increase out profit. Another thing which we can do to increase the profit is to start leveraging the resources in the Asia Pacific Market, through this we will be able to reduce both fixed cost and operational cost and this can increase our profit in that market.

Conclusion:

The analysis provides the insight about how the shipping cost is affecting the companies profit and giving the management indication of reducing it by improving their logistics and supply chain method.

To overcome this problem, what we can do is, we can setup a fixed percentage on the profit generating from the product price, this will help us to select the best mode of shipment for the product.

The given analysis also provides the insight about the different markets. As Asia Pacific has the highest share but the most profit generating was European Market. This could be possible because of the purchasing power of the people is good in that market or economy. Which provides a good opportunity for the company to target this market more and try to increase their market share here. The other reason for this is discounts also. With the overall information we can strategize our market according to the variables and the purchasing power of the targeted market.

Learnings:

The above analysis has allowed me to learn how to present the data, which variables to choose either numerical or categorical, also depending upon the type of data analysis to perform. Plus, I learned how to plot different graphs to analyze the data in R. Learned in detail the use of vector in R.

Overall, I got a brief overview of how to perform and present the data, analyze new techniques, identifying and understanding the patterns and try to understand what data is trying to communicate.

APPENDIX:

```r
library("plyr")

library("FSA")

library("FSAdata")

library("magrittr")

library("dplyr")

library("plotrix")

library("ggplot2")

library("moments")

library(knitr)

library(readr)

library(DT)

sales2020 <- read_csv("C:/Users/hp/Downloads/salesdata.csv")

View(salesdata)

summary(salesdata)

df = data.frame("Quantity"=salesdata$Quantity,
"SalesTotal"=salesdata$Sales_Total,"Discount"= salesdata$Discount,
"Profit"=salesdata$Profits, "ShippingCost"=salesdata$ShippingCost_Product)

hist(salesdata$Quantity, main="Quantity", xlab="Quantity", xlim=c(0,20), ylim=c(0,250),
col="darkmagenta")

hist(salesdata$Sales_Total, main="SalesTotal", xlab="SalesTotal", xlim=c(0,15000),
breaks=20, col="blue")

hist(salesdata$Discount, main="Discount", xlab="Discount", xlim=c(0,0.5), ylim=c(0,500),
breaks=3, col="yellow")

hist(salesdata$Profits, main="Profit", xlab="Profit", xlim=c(0,7000), ylim=c(0,170), breaks=50,
col="orange")

library(RColorBrewer)

department<-table(salesdata$Department)

barplot(Department, main="Department", ylim  = c(0,500), cex.names = 0.7, col=
Brewer.pal(5,name = "Paired"))
```

```r
barplot(market, width = 0.8, space = 0.4, ylim = c(0,800), cex.names = 0.8, las = 1,
main="Market", col="green")

Profit_market = tapply(salesdata$Profits, INDEX = salesdata$Market, FUN = mean)

barplot(Profit_market, main="Market with highest Profit", xlab="Market", ylab="Profits",
ylim=c(0,2000), col=brewer.pal(5,name="Paired"))

text(y=Profit_table, np_mar, round(Profit_table, digits=2), cex=0.8, pos=3)

Furniture = dplyr :: filter(salesdata,Department = "Furniture")

Office Supplies = dplyr :: filter(salesdata,Department = "Office Supplies")

Technology = dplyr :: filter(salesdata,Department = "Technology")

mean(Furniture$ShippingCost_Product)

mean(OfficeSupplies$ShippingCost_Product)

mean(Technology$ShippingCost_Product)

vector_names=c("Furniture","Office Supplies","Technology")

vector_statis = c(mean_Furniture, mean_OfficeSupplies, mean_Technology)

names(vector_statis) = vector_names

c_sc = knitr :: kable(vector_statis, col.names = "Shipping Cost of Product")

c_plot = barplot(vector_statis, main = "Shipping Cost of Product Per Each Department", ylab =
"Shipping Cost Product", ylim = c(0,100), col = brewer.pal(3,name = "Set2"))

text(y = vector_statis, c_plot, round(vector_statis, digits = 3), cex = 0.71, pos = 3.1)

profit_dprt = tapply(salesdata$Profits, INDEX = salesdata$Department, FUN = mean)

profitplot = barplot(profit_deprt, main = "Profit in Each Department", xlab = "Department",
ylab = "Profits", ylim = c(0,3000), cex.name = 0.9, col = brewer.pal(5,name = "Paired"))

text(y = profit_deprt,proplot, round(profit_deprt, digits = 2), cex = 0.8, pos = 3)

vector_mm = c(mean_qty,mean_SalesTotal,mean_discount,mean_profit,mean_cost,
,median_qty,median_SalesTotal,median_discount,median_profit,median_cost)

row1 = c("Mean", "Median")

col1 = c("SalesTotal","Quantity","Discount","Profit","ShippingCostProduct")

descriptive_statistics = matrix(vector_mm, nrow = 5, ncol = 2, dimnames = list(col1,row1))

descriptive_statistics
```

References:

1. https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter
2. R Mean, Median and Mode
   https://www.tutorialspoint.com/r/r_mean_median_mode.htm
3. Kaggle your Machine Learning and Data Science Community
   https://www.kaggle.com
4. How to Select a Range of Date in R?
   https://stats.stackexchange.com/questions/220865/how-to-select-a-range-of-dates-in-r/220870
5. How to Assign Colors by Factor in ggplot2?, By Zach (November 16, 2020)
   https://www.statology.org/color-by-factor-ggplot2/
6. The Complete ggplot2 Tutorial – Part 1 | Introduction To ggplot2.
   http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html
7. R Vector
   https://www.datamentor.io/r-programming/vector/
8. R Vector – How to Create, Combine and Index Vectors in R?
   https://techvidvan.com/tutorials/r-vector/
9. LinkedIn Learning
   https://www.linkedin.com/learning/