Final Project

Ankit Bhalekar

2023-02-18

ALY600 INTRODUCTION TO ANALYTICS
Professor Name: Dr. Dee Chiluiza, PhD
NORTHEASTERN UNIVERSITY
ANKIT VILAS BHALEKAR
Date: 18 February, 2023
Project Report

Library Data

```
library(readxl)
library(readr)
library(tidyverse)
library(dplyr)
library(knitr)
library(kableExtra)
library(DT)
library(magrittr)
library(FSA)
library(RColorBrewer)
library(ggplot2)

M6data <- read_excel("DataSets/M6_project_dataset(2023).xlsx")
```

ACKNOWLEDGMENT:

I have learned a ton from the course, and I will use the knowledge and abilities I have acquired to further my profession. I now know how to assess and analyze complicated data sets, draw meaningful conclusions from data, and present data in an engaging and clear way.

Moreover, I want to thank the entire class for their outstanding assistance, quick responses to questions, and facilitation of a positive learning atmosphere. Last but not least, I want to thank my friends for working together to make the learning process more meaningful and fun.

Once more, I would want to express my gratitude to Professor Dr. Dee Chiluiza for all of her tremendous efforts in making this course a success. I'm sure that whatever knowledge and skills I've gained about data analytics will be crucial for my professional and personal development.

Sincerely,

Ankit Vilas Bhalekar

INTRODUCTION:

The information in the provided data set includes information on all international marketplaces, broken down by area. This is a sizable data collection with a ton of information about the market for specific product categories like phones, copiers, table chairs, etc., as well as market segments, departments, shipping mode, shipping cost, losses due to returns, etc.

Because this data is so complex in tabular form and so challenging to grasp, we need to analyse it to make it simpler and put it into a more understandable format so that everyone can understand it and make predictions based on it.

I'll be analyzing the data using the R program. R is an extremely effective tool for data analysis. I'm going to perform several tasks for this report that will analyze the data and create some significant tables and graphs. These graphs will provide us a clear understanding of the most significant features of the data.

Descriptive statistics:-A statistical method which is known as descriptive statistics uses specific values like mean, mode, median,etc. to characterize data. As they contain no extension or assumption outside of what is directly available, descriptive statistics just describe the data that is currently available (a sample) and do not draw their conclusions from any probabilistic theories.

Inferential statistics :-A statistical method known as inferential statistics allows for the inference of features of a larger group can be inferred from a sample this small but representative. It enables the scientist to draw conclusions from a smaller sample of a larger group..

R Script versus R Markdown:

The R programming language uses both R script and R Markdown files, but they have different functions. R code is contained in plain text files called R script files. They are used to create and execute a collection of R commands that may be saved and used again at a later time. The R console or an integrated development environment (IDE) like RStudio are two places where R scripts can be run.

R Markdown files, on the other hand, are documents that combine text and code. In a single document, you can mix structured text, code, and the output of code execution, such as tables and charts. R Markdown files are perfect for creating reports, presentations, and even complete books because they can be transformed to a variety of output formats like HTML, PDF, and Word.

The similarities between R script and R Markdown files are that both may execute in the R terminal or an IDE like RStudio and both can contain R code.

The primary distinction between the two files is that R Markdown files incorporate text and code whereas R Script files exclusively contain code. Other advantages offered by R Markdown files include the capacity to incorporate automatically created tables, graphs, and text formatting.

R script files' primary benefit is their simplicity, as they offer a basic method for writing and executing R code. On the other hand, R Markdown files' key benefit is their capacity to generate dynamic, repeatable reports and documents that incorporate both code and text.

In conclusion, R Markdown files are useful for creating dynamic, reproducible documents that combine code and text, whereas R script files are ideal for writing and running code.

ANALYSIS:

TASK 1: Provide descriptive statistics of the entire data set.

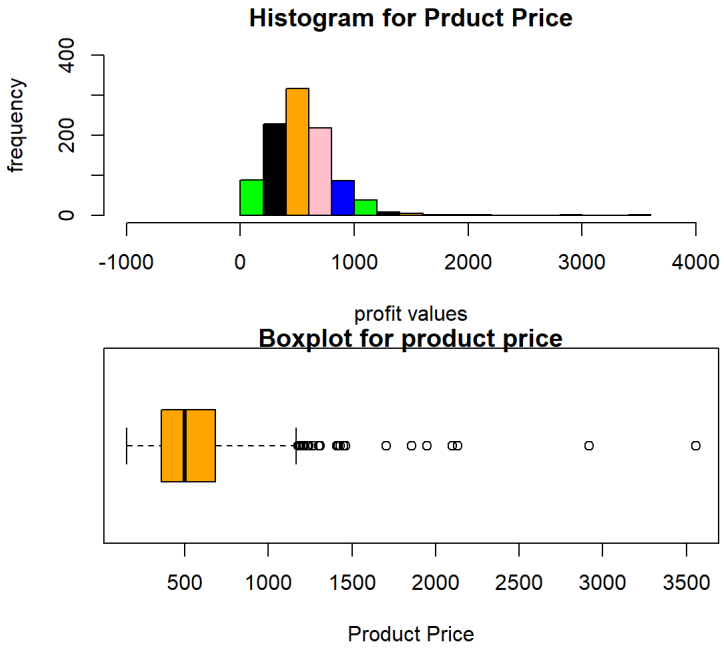
Description:
First task is about presenting the descriptive statistics of the entire data set. The basic descriptive statistics would be mean, median, mode, max values, min values etc. I had to chose one table and produce two graphs for that table. i chose the variable "product price and produced Histogram and bar plot for the same.

```
t1a=summary(M6data)


t1a %>%
  kbl(caption = "Summary Table") %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

Summary Table

OrderDate	ProductID	City	State	Country	Region	Market	Segment	Department	Division	OrderPriority	ShipM
Min. :2020-01-01 00:00:00.0	Length:1000	Length:1000	Length:1000	Length:1000	Length:1000	Length:1000	Length:1000	Length:1000	Length:1000	Length:1000	Lengtl
1st Qu.:2020-05-09 00:00:00.0	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :chara
Median :2020-08-06 00:00:00.0	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :chara
Mean :2020-07-24 05:12:28.73	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3rd Qu.:2020-10-19 00:00:00.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Max. :2020-12-25 00:00:00.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

```
par(mfcol=c(2,1),
    mai=c(1,1,4,1),
    mar=c(4,8,0.9,3))

tab1=M6data$Product_Price
hist(tab1,
      main="Histogram for Prduct Price",
      breaks=20,
     xlab="profit values",
     ylab="frequency",
     xlim =c(-1000,4000),
     ylim = c(0,450),
    col = c("green","black","orange","pink","blue"))


boxplot(tab1,
           horizontal = T,
           main= "Boxplot for product price",
           xlab="Product Price",
           col="orange")
```




```
mean_price=mean(tab1)
med_price=median(tab1)
min_price=min(tab1)
max_price=max(tab1)


vec1=c(mean_price,med_price,min_price,max_price)

mat1= round(matrix(c(vec1), ncol =4, byrow = TRUE),2)
col_names= c("Mean","Median","Min price","max price")
colnames(mat1)=col_names

mat1
```

```
##       Mean Median Min price max price
## [1,] 548.9 501.35    151.93   3558.41
```

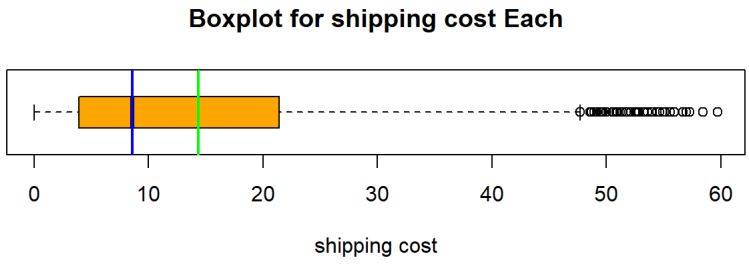
Observation:

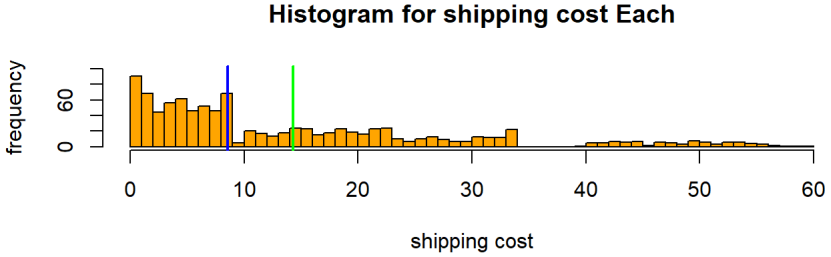
1)I got basic descriptive statistic of the data set.
2)Mean Product price is 548.9
3)Maximum shipping cost is 59.71
4)Maximum loss per returns is 1352.20
5)From the barplot and histogram mean and median product price is 548.9 and 501.35 respectively
6)From the barplot and histogram min and max product price is 151.93 and 3558.41 respectively

Task2: Displaying the distribution of variable “Shipping Cost Each”.

Description:
The distribution of the variable Shipping Cost Each by using a horizontal box plot and a histogram is displayed. also the code par(mfcol=c(2,1)) are used at the beginning of your R chunk. Median and the mean on each graph is displayed. To make comparisons easier, the graph’s are set to the same limit magnitude. The variable distribution is described.

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

tab2=M6data$Shipping_Cost_Each

mean_ship=mean(tab2)
med_ship=median(tab2)

boxplot(tab2,
           horizontal = T,
           main= "Boxplot for shipping cost Each",
           xlab="shipping cost",
           col="orange")
         #abline for mean

        abline (v =mean_ship,col = "green",lwd=2)

        #abline for median

        abline (v =med_ship,col = "blue",lwd=2)



  text(x=mean(tab2),
     y = 390,
     paste("Median:", round(median(tab2),1),"K"),
     col = "red",
     cex = 0.8,
     pos = 2)

hist(tab2,
      main="Histogram for shipping cost Each",
      breaks=50,
     xlab="shipping cost",
     ylab="frequency",
     xlim =c(0,60),
     ylim = c(0,100),
     col="orange")

         #abline for mean

        abline (v =mean_ship,col = "green",lwd=2)

        #abline for median

        abline (v =med_ship,col = "blue",lwd=2)
```





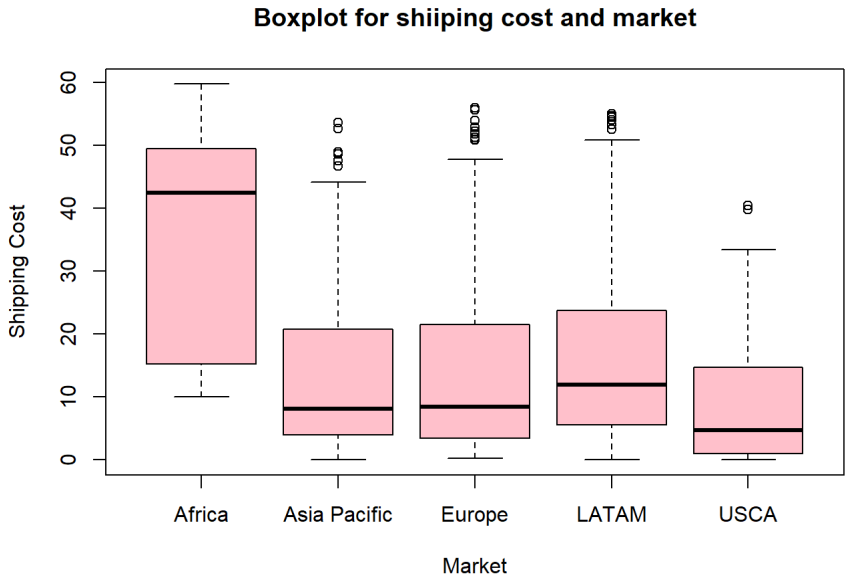
Observations:

1)Boxplot and Histogram for shipping cost each is plotted.
2)Mean shipping cost each is 14.32 which is seen from the graphs
3)Median shipping cost each is 8.585 which is seen from the graphs

Task3: Ploting a box plot for Shipping Cost versus Markets.

Description:
The task is about displaying the Shipping Cost versus Markets, plotting a boxplot.The purpose of this task is to fine out in which market the shipping cost is higher and in which it is lower.

```
boxplot(M6data$Shipping_Cost_Each~M6data$Market,
        main="Boxplot for shiiping cost and market",
        xlab = "Market",
        ylab = "Shipping Cost",
        col="pink")
```



Observations:

1)Box plots Shipping Cost versus Markets for each market are plotted.
2)African market has the highest shipping cost.
3)USCA market has the lowest shipping cost
4)No outliers are shown for the box plot of African market.

Task4: Bar plot for mean shipping cost per market.

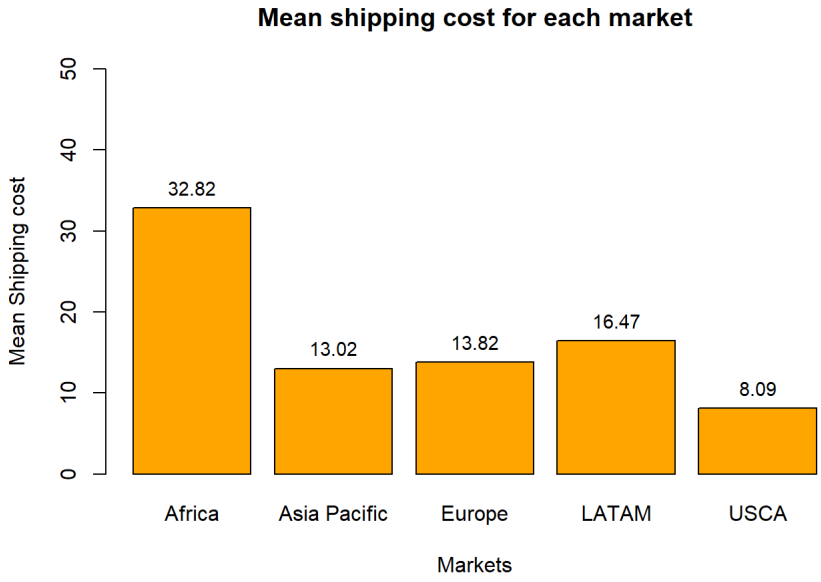
Description:
The task is about applying the tapply() code and tracking the average delivery cost per market. By using the aforementioned code, i have to make an object that I may use to prepare and present a bar plot. This graph is then to be compared with the box plot from the preceding task.

```
t4=round(tapply(M6data$Shipping_Cost_Each,M6data$Market,mean),2)

kbl(t4,caption = "**************************************************************Sum of Sales*****************************
**************************************************") %>%
  kable_material_dark()
```

Sum of Sales

	x
Africa	32.82
Asia Pacific	13.02
Europe	13.82
LATAM	16.47
USCA	8.09

```
bar=barplot(t4,
        main = "Mean shipping cost for each market",
        ylim = c(0,50),
        xlab = "Markets",
        ylab = "Mean Shipping cost",
        col = "orange")

text(y=t4,bar,labels =t4, cex=0.9,pos=3)
```



Observations:

1)Box plots for mean Shipping Cost versus Markets for each market are plotted.
2)African market shows the highest shipping cost which is 32.82.
3)USCA market shows the lowest shipping cost which is 8.09
4)When I compared this graph to the previous box plots it shows that Mean shipping cost of african market is also highest.
5)similarly mean shiping cost of USCA market is lowest as lowest shiiping cost in previous task.

Task4 (Extra Points): Bar plot for median loss per returns for each markets.

Description:
The task is for extra point and it is about about applying the tapply() code and tracking the median loss per return for each segments.
Observations:

```
te= tapply(M6data$Loss_Per_Return, M6data$Segment, median)

kbl(te,caption = "**********************************************************************Median Loss Per returns for Each segment***
**********************************************************************************") %>%
  kable_material_dark()
```

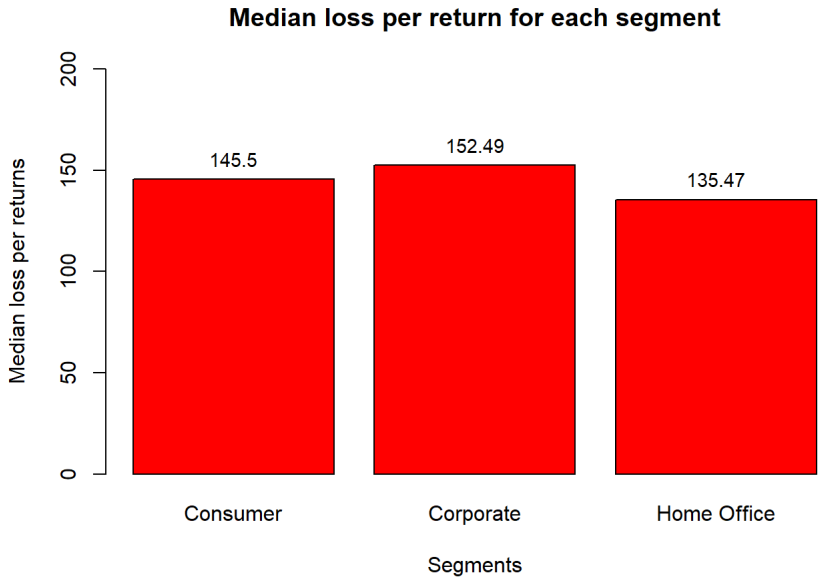
**********Median Loss Per returns for Each segment**********

Consumer

Corporate

Home Office

```
bar2=barplot(te,
        main = "Median loss per return for each segment",
        ylim = c(0,200),
        xlab = "Segments",
        ylab = "Median loss per returns",
        col = "red")

text(y=te,bar,labels =te, cex=0.9,pos=3)
```



1)Bar plots for median loss per return for each segment is plotted.
2)Corporate segment has the highest median loss per return which is 152.49.
3)Home office segment has the lowest median loss per return which is 135.47.

Task5: Using a box plot, display Shipping Cost versus Shipping Mode.

Description:
The purpose of this task is to proved the produce the boxplots which will display Shipping Cost versus Shipping Mode . there will be four box plots, one for each class. From these box plot i can clearly see which class will haev the hhighest and lowest shipping cost.

```
boxplot(M6data$Shipping_Cost_Each~M6data$ShipMode,
        col = c('green','pink','red','violet','gray'),
        ylab = "shipping Cost",
        xlab = "Shipping Mode")
```

Observations:

1)Corporate segment has the highest median loss per return which is 152.49.
2)Home office segment has the lowest median loss per return which is 135.47.
3)No outliers seen for the “same Day” class.

Task6: Bar plot for mean shipping cost per market.

Description:
This task involves adding a new column to dataset and using the pipes and mutate() code to calculate the total sales as the product of price and quantity. Also, I must display the codes; however, it is not necessary to display the results. To present all the variables, including the new column Total sales, use a glimpse of dataset.

```
M6data1= M6data%>%
  mutate(total_sale=M6data$Product_Price*M6data$Quantity)

glimpse(M6data1[1,])
```

```
## Rows: 1
## Columns: 18
## $ OrderDate          <dttm> 2020-05-08
## $ ProductID          <chr> "TEC75553"
## $ City               <chr> "Yaounde"
## $ State              <chr> "Centre"
## $ Country            <chr> "Cameroon"
## $ Region             <chr> "Central Africa"
## $ Market             <chr> "Africa"
## $ Segment            <chr> "Home Office"
## $ Department         <chr> "Technology"
## $ Division           <chr> "Accessories"
## $ OrderPriority      <chr> "High"
## $ ShipMode           <chr> "First Class"
## $ Product_Price      <dbl> 360.04
## $ Quantity           <dbl> 154
## $ Shipping_Cost_Each <dbl> 50.84
## $ Returns            <dbl> 46
## $ Loss_Per_Return    <dbl> 126.01
## $ total_sale         <dbl> 55446.16
```

Task7: Bar plot for mean shipping cost per market.

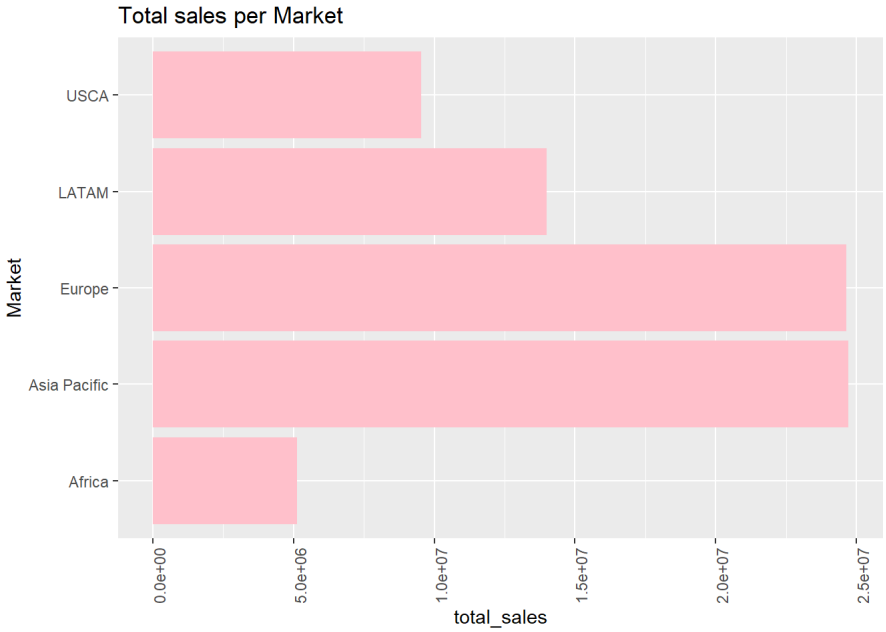
Description:
I have created the variable (column) total sales. In this task, I have to observe in which market or segment or department, the company had the highest sales. Choice of the variable is up to me.

```
t7= M6data1 %>%
       group_by(Market) %>%
       summarise(total_sales = sum(total_sale))


b=t7 %>%
  kbl(caption = "Total sales per market") %>%
  kable_classic(full_width = F, html_font = "Cambria")
b
```

Total sales per market

Market	total_sales
Africa	5127296
Asia Pacific	24704625
Europe	24627127
LATAM	14005415
USCA	9533931

```
plot_Market = ggplot(t7,
       aes(x = total_sales,
       y = Market, fill = total_sales))+ geom_bar(stat = "identity", fill = "pink")  +
       ggtitle("Total sales per Market")

plot_Market + theme(axis.text.x = element_text(angle=90))
```



Observations:
1)boxplot is plotted for total sales per market. 2)Asia pacific market has the highest total sales 24704625.
3)USCA market has the lowest total sales which is 9533931.

Task8: Bar plot for mean shipping cost per market.

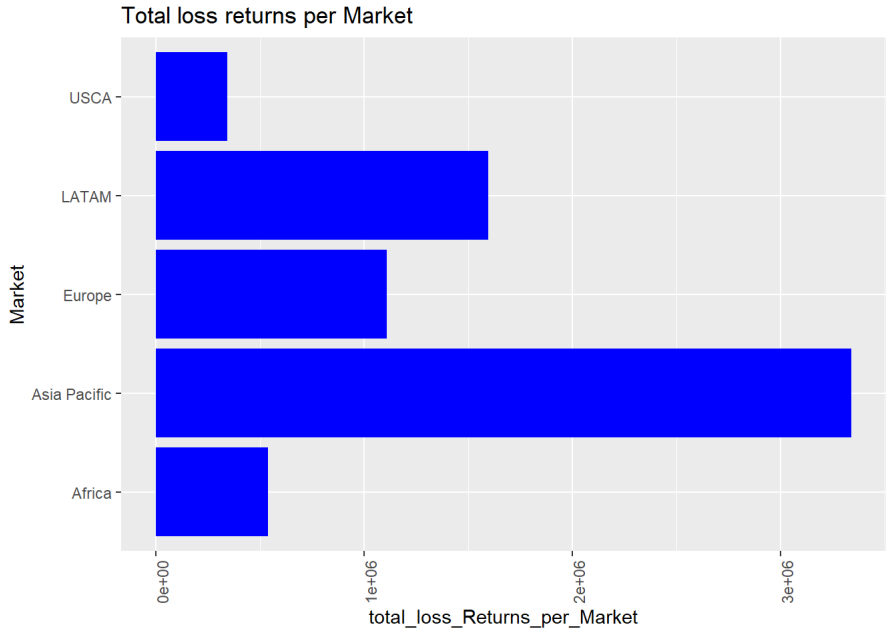
Description:
I have to create a question that uses the three codes group by(), filter(), and mutate() in this task by drawing on my understanding of those three codes. I am therefore using this algorithm and graphics to determine total loss returns for each market.

```
M6data2 = M6data %>%
mutate(totalloss_per_returns = M6data$Returns * M6data$Loss_Per_Return)
t8= M6data2 %>%
       group_by(Market)%>%
       summarise(total_loss_Returns_per_Market = sum(totalloss_per_returns))

knitr::kable(t8,
             digits = 2,
             caption = "Table : Total sales per Market",
             format = "html",table.attr = "style='width:40%:'", align = 'c')%>%
kable_classic(bootstrap_options = "striped",
full_width = TRUE, position = "center", font_size = 12)
```

Table : Total sales per Market

Market	total_loss_Returns_per_Market
Africa	538580.1
Asia Pacific	3338326.5
Europe	1109570.2
LATAM	1596964.4
USCA	343502.6

```
plot_Market = ggplot(t8,
       aes(x = total_loss_Returns_per_Market,
       y = Market, fill = total_loss_Returns_per_Market))+
       geom_bar(stat = "identity", fill = "blue")  + ggtitle("Total loss returns per Market")

plot_Market + theme(axis.text.x = element_text(angle=90))
```



Observations:
1)boxplot is plotted for total loss return per market.
2)Asia pacific market has the highest total loss returns which is 3338326.5
3)USCA market has the lowest total loss returns which is sales which is 343502.6

CONCLUSION:

Technical information regarding the global market that was broken down by region was given. After evaluating the data, we were able to derive some distilled information about the table's summary, the highest and lowest allowed product prices, the mean and median shipping costs for each market, and a number of other items.

I created boxplots, histograms, and bar graphs based on the provided data in order to meet the requirements and extract some useful information from the chart. Under each task, observations for each individual task are recorded. According to the aforementioned data, the USCA market has the lowest shipping cost and mean shipping coast, while the African market has the highest. I also came to the conclusion that corporate has the largest median loss per return and home office has the lowest. I came to the conclusion that "same day" class has the highest median shipping cost while "regular class" has the lowest shipping cost after plotting the boxplots for shipping cost versus shipping mode.

While completing this report I learned many skill about data analytics. I learned data visualization techniques using R. I Learned statistics and probabilities. I learned constructing the right graphs based on the information and drawing conclusions from the graphs. While completing the project I used a simple strategy to complete the project. I understood the data set first and the understood the instructions given by the professor. Then started solving the tasks one by one. I solve all the task and the started writing observations for each task. I tried my best to complete the project and construct the report.

REFERENCES:

[1]Satyapriya Chaudhari, "What Is Descriptive Statistics?", https://builtin.com/data-science/descriptive-statistics (https://builtin.com/data-science/descriptive-statistics)

[2]My accounting course, "What is Inferential Statistics?,"https://www.myaccountingcourse.com/accounting-dictionary/inferential-statistics (https://www.myaccountingcourse.com/accounting-dictionary/inferential-statistics)
[3]Gc digital fellows, "Intro to R markdown", https://digitalfellows.commons.gc.cuny.edu/2022/05/18/introduction-to-r-markdown/ (https://digitalfellows.commons.gc.cuny.edu/2022/05/18/introduction-to-r-markdown/)
[4]Dee Chiluiza, "Introduction to data analysis using R, R Studio and R Markdown", https://rpubs.com/Dee_Chiluiza/816756 (https://rpubs.com/Dee_Chiluiza/816756)

APPENDIX:

To this report, R markdown file is attached. Name of the file is "Bhalekar_ALY600Project6.Rmd"