R Notebook

Code

GRIP : The Sparks Foundation Data Science and Business Analytics Prepared by : Rutvi Shah TASK-3 : Exploratory Data Analysis - Retail —-

Step 1: Importing all necessary libraries

Hide

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(shiny)
library(plotly)
library(corrr)
library(treemap)


superstore <- read.csv('SampleSuperstore.csv')
head(superstore)
```

	Ship.Mode <fctr>	Segment <fctr>	Country <fctr>	City <fctr>	State <fctr>	Postal.Code <int>	Re <fc
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	Sc
2	Second Class	Consumer	United States	Henderson	Kentucky	42420	Sc
3	Second Class	Corporate	United States	Los Angeles	California	90036	W
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	Sc
5	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	Sc
6	Standard Class	Consumer	United States	Los Angeles	California	90032	W

6 rows | 1-8 of 13 columns

Hide

```
tail(superstore)
```

	Ship.Mode <fctr>	Segment <fctr>	Country <fctr>	City <fctr>	State <fctr>	Postal.Code <int>	Reg <fct
9989	Standard Class	Corporate	United States	Athens	Georgia	30605	Sou
9990	Second Class	Consumer	United States	Miami	Florida	33180	Sou
9991	Standard Class	Consumer	United States	Costa Mesa	California	92627	Wes
9992	Standard Class	Consumer	United States	Costa Mesa	California	92627	Wes
9993	Standard Class	Consumer	United States	Costa Mesa	California	92627	Wes

	Ship.Mode <fctr>	Segment <fctr>	Country <fctr>	City <fctr>	State <fctr>	Postal.Code <int>	Reg <fct
9994	Second Class	Consumer	United States	Westminster	California	92683	Wes

6 rows | 1-8 of 13 columns

Step 3: Finding more information about dataset

Hide

```
summary(superstore)
```

```
          Ship.Mode           Segment              Country                City
State       Postal.Code        Region
 First Class   :1538   Consumer   :5191   United States:9994   New York City: 915   Californi
a  :2001   Min.   : 1040   Central:2323
 Same Day      : 543   Corporate  :3020                        Los Angeles  : 747   New York
:1128   1st Qu.:23223   East   :2848
 Second Class  :1945   Home Office:1783                        Philadelphia : 537   Texas
: 985   Median :56431   South  :1620
 Standard Class:5968                                           San Francisco: 510   Pennsylva
nia: 587   Mean   :55190   West   :3203
                                                               Seattle      : 428   Washingto
n  : 506   3rd Qu.:90008
                                                               Houston      : 377   Illinois
: 492   Max.   :99301
                                                               (Other)      :6480   (Other)
:4295
            Category         Sub.Category      Sales              Quantity        Discount
Profit
 Furniture      :2121   Binders    :1523   Min.   :    0.444   Min.   : 1.00   Min.   :0.0000
Min.   :-6599.978
 Office Supplies:6026   Paper      :1370   1st Qu.:   17.280   1st Qu.: 2.00   1st Qu.:0.0000
1st Qu.:    1.729
 Technology     :1847   Furnishings: 957   Median :   54.490   Median : 3.00   Median :0.2000
Median :    8.666
                       Phones     : 889   Mean   :  229.858   Mean   : 3.79   Mean   :0.1562
Mean   :   28.657
                       Storage    : 846   3rd Qu.:  209.940   3rd Qu.: 5.00   3rd Qu.:0.2000
3rd Qu.:   29.364
                       Art        : 796   Max.   :22638.480   Max.   :14.00   Max.   :0.8000
Max.   : 8399.976
                       (Other)    :3613
```

#To check if there are any null values Step 4: Data Preparing and Cleaning

Hide

```
is.null(superstore)
```

```
[1] FALSE
```

#To check if there is any duplicacy and remove them too along with removing two columns (postal codes and country) #as I do not require them for further analysis.

Hide

```
data <- superstore %>%
         distinct() %>%
         select(-c(Country, Postal.Code))
data
```

Ship.Mode <fctr>	Segment <fctr>	City <fctr>	State <fctr>	Region <fctr>	Category <fctr>
Second Class	Consumer	Henderson	Kentucky	South	Furniture
Second Class	Consumer	Henderson	Kentucky	South	Furniture
Second Class	Corporate	Los Angeles	California	West	Office Supplies
Standard Class	Consumer	Fort Lauderdale	Florida	South	Furniture
Standard Class	Consumer	Fort Lauderdale	Florida	South	Office Supplies
Standard Class	Consumer	Los Angeles	California	West	Furniture
Standard Class	Consumer	Los Angeles	California	West	Office Supplies
Standard Class	Consumer	Los Angeles	California	West	Technology
Standard Class	Consumer	Los Angeles	California	West	Office Supplies
Standard Class	Consumer	Los Angeles	California	West	Office Supplies

1-10 of 9,977 rows | 1-7 of 11 columns Previous **1** 2 3 4 5 6 … 100 Next

Step 5: Checking Statistical Relationship between rows and columns.

1. Correlation between variables.

Hide

```
x <- data %>%select(Sales, Quantity, Discount, Profit)
corr_var <- correlate(x, method = 'pearson',use = "pairwise.complete.obs", diagonal = 1)
```

```
Correlation method: 'pearson'
Missing treated using: 'pairwise.complete.obs'
```

Hide

```
corr_var
```

term <chr>	Sales <dbl>	Quantity <dbl>	Discount <dbl>	Profit <dbl>
Sales	1.00000000	0.200722092	-0.028311117	0.47906731
Quantity	0.20072209	1.000000000	0.008678422	0.06621065
Discount	-0.02831112	0.008678422	1.000000000	-0.21966206

term <chr>	Sales <dbl>	Quantity <dbl>	Discount <dbl>	Profit <dbl>
Profit	0.47906731	0.066210646	-0.219662064	1.00000000

4 rows

2. Covariance between variables.

Hide

```
y<- data %>%select(Sales, Quantity, Discount, Profit)
cov_var <- cov(y)
cov_var
```

```
                Sales     Quantity      Discount      Profit
Sales    389028.396022 2.787656e+02  -3.645637429 70057.06713
Quantity    278.765576 4.958001e+00   0.003989513    34.56574
Discount     -3.645637 3.989513e-03   0.042623749   -10.63275
Profit    70057.067126 3.456574e+01 -10.632750986 54970.47882
```

3. Statistical Summary for sales¶

Hide

```
summary(data$Sales)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
    0.444    17.300    54.816   230.149   209.970 22638.480 
```

Step 6: Analysis and Visualization

1. Statewise Sales Analysis

Hide

```
 statewise_sales <- data %>%
                 group_by(State) %>%
                 summarise(total_sales = sum(Sales)) %>%
                 arrange(desc(total_sales))
statewise_sales
```

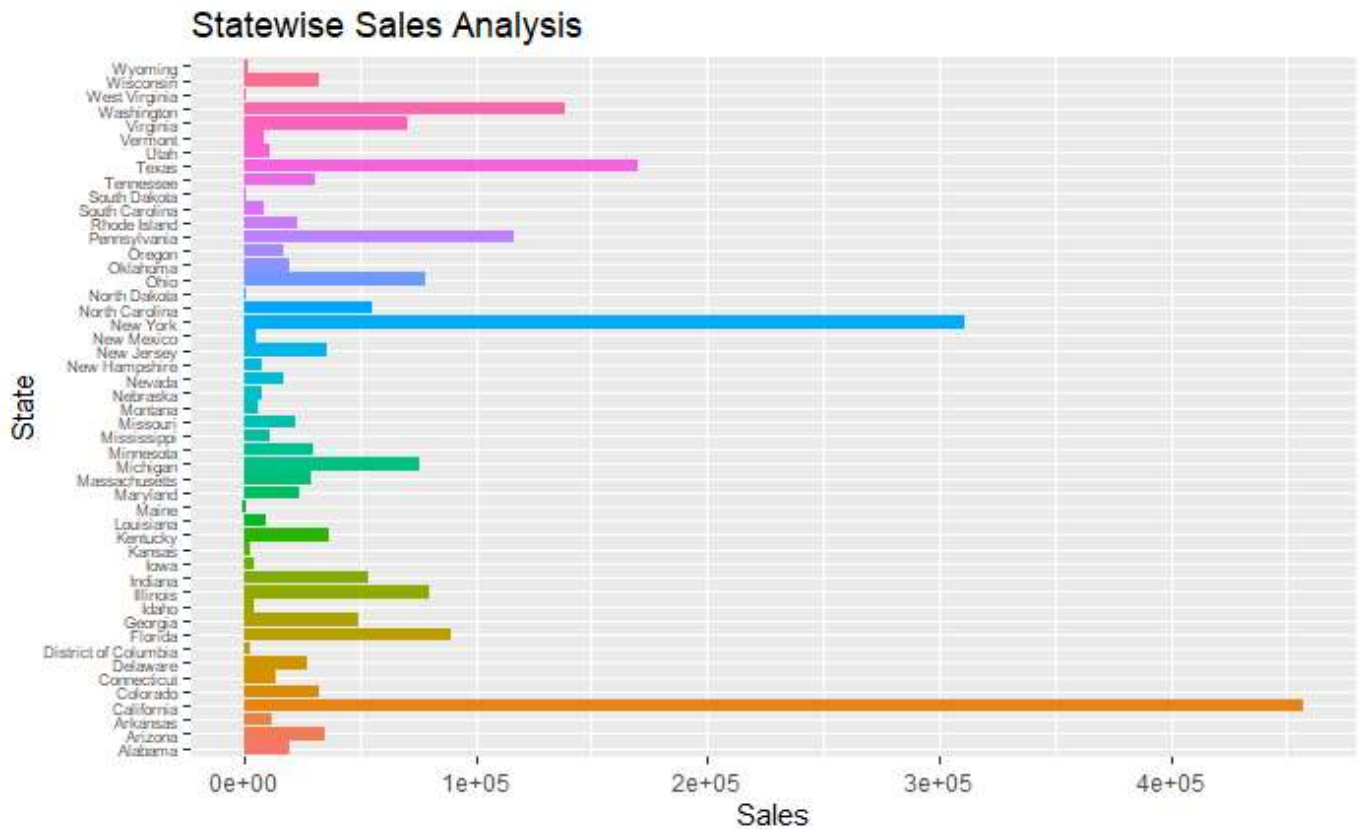
State <fctr>	total_sales <dbl>
California	457576.271
New York	310827.151
Texas	170124.542
Washington	138560.810
Pennsylvania	116496.362
Florida	89473.708

State <fctr>	total_sales <dbl>
Illinois	80162.537
Ohio	77976.764
Michigan	75879.644
Virginia	70636.720

1-10 of 49 rows Previous **1** 2 3 4 5 Next

Hide

```
ggplot(data, aes( x= State, y= Sales, fill= State),options(scipen=999)) +
          geom_col()+
          ggtitle("Statewise Sales Analysis") +
          coord_flip() +
        theme(legend.position = "None", axis.text.y = element_text(size=6))
```



Observation: State of California recorded the highest Sales of around 4,50,000 USD; followed by New York, Texas and Washington at second, third and fourth position respectively. On the other hand, North Dakota records the least sales among all the states with nearly 900 USD.

2. Regionwise Sales Analysis

Hide

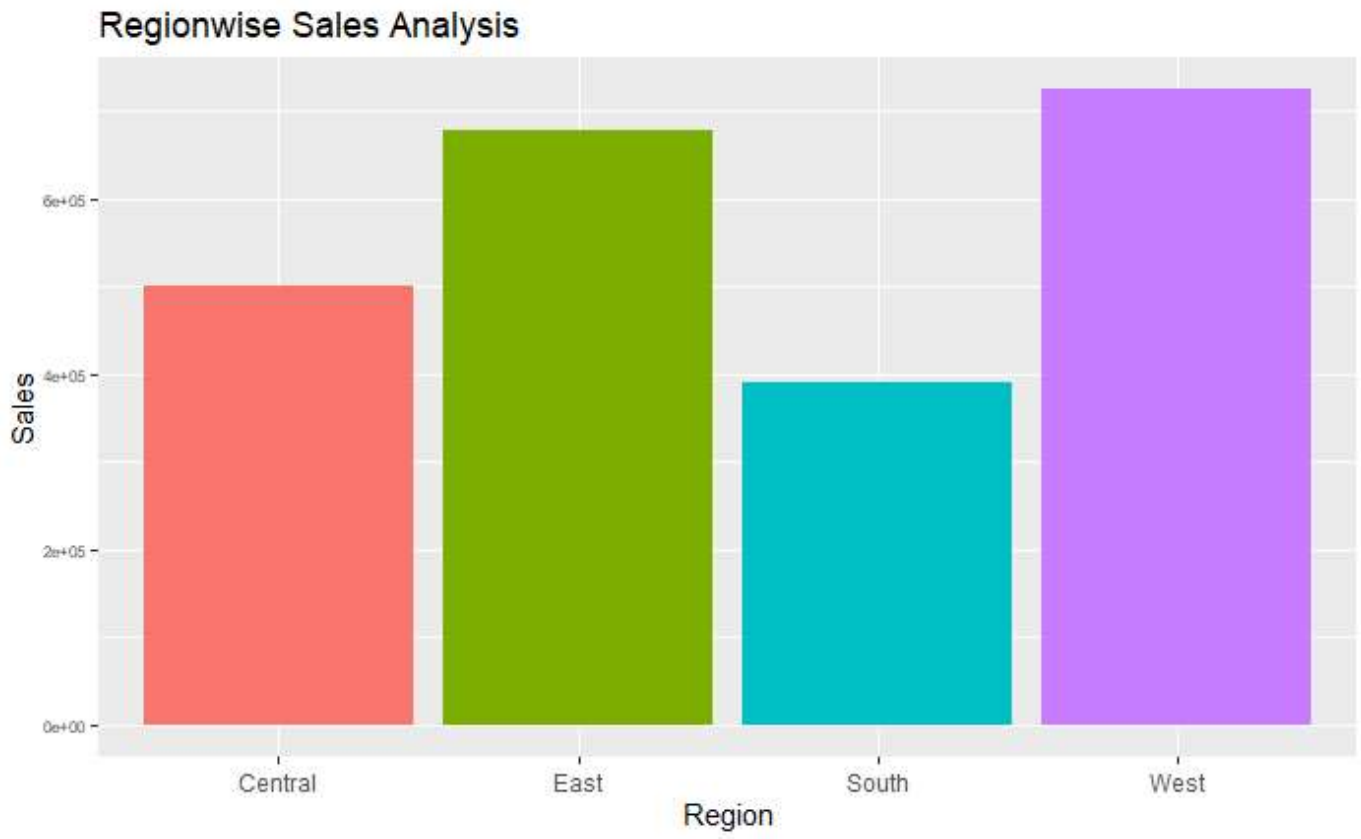
```
regionwise_sales = data %>%
        group_by(Region) %>%
       summarize(totalS= sum(Sales)) %>%
       arrange(desc(totalS))
regionwise_sales
```

Region <fctr>	totalS <dbl>
West	725255.6
East	678435.2
Central	500782.9
South	391721.9

4 rows

Hide

```
ggplot(data, aes( x= Region, y= Sales, fill= Region),options(scipen=99)) +
          geom_col()+
          ggtitle("Regionwise Sales Analysis") +
        theme(legend.position = "None", axis.text.y = element_text(size=6))
```



Observation : From Regionwise Sales chart, we can see that company's sales are mostly concentrated on the Eastern and Western Region of America

3. Statewise Profit Analysis

Hide

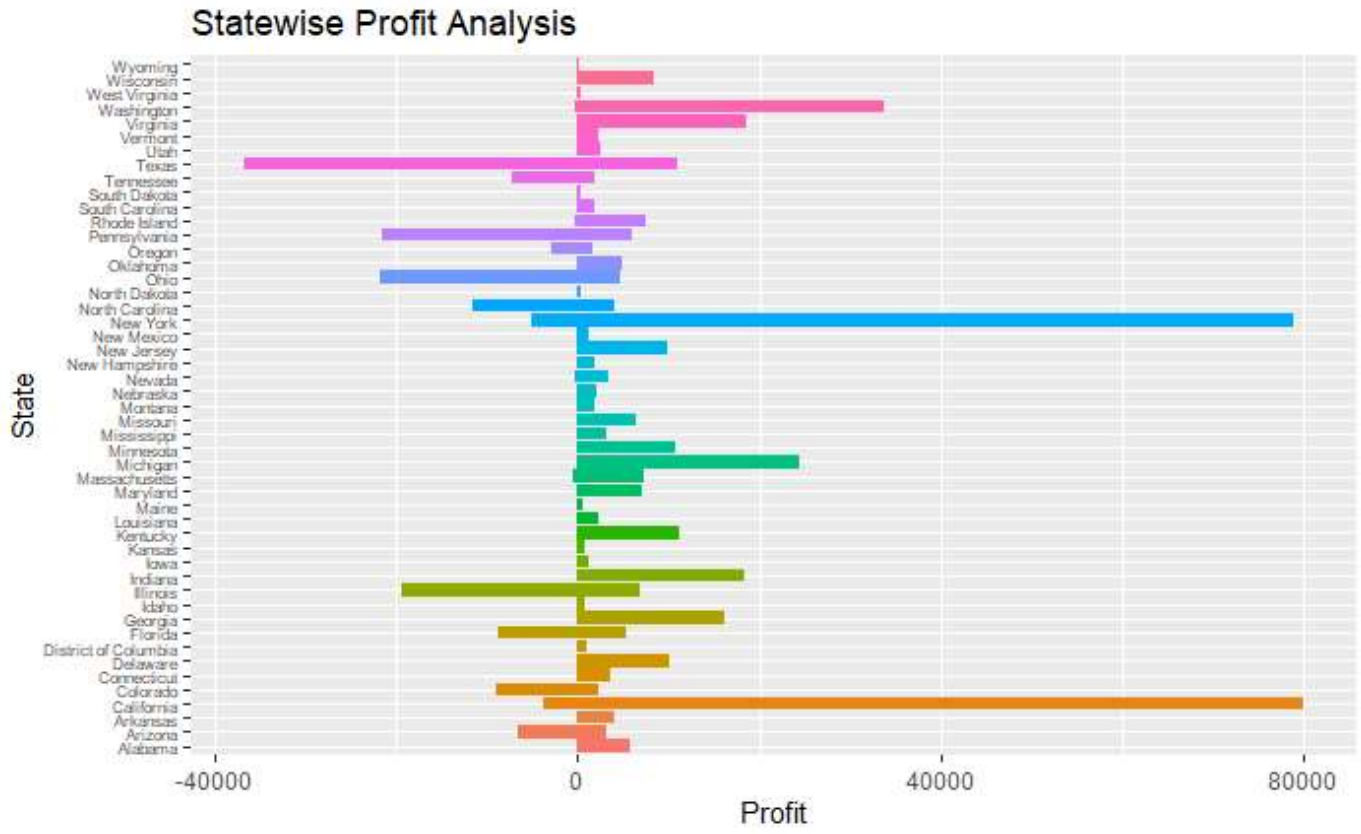
```
Statewise_profit = data%>%
          group_by(State)%>%
          summarise(totalP= sum(Profit))%>%
           arrange(desc(totalP))
Statewise_profit
```

State <fctr>	totalP <dbl>
California	76330.7891
New York	74015.4622
Washington	33368.2375
Michigan	24428.0903
Virginia	18597.9504
Indiana	18382.9363
Georgia	16250.0433
Kentucky	11199.6966
Minnesota	10823.1874
Delaware	9977.3748

1-10 of 49 rows Previous **1** 2 3 4 5 Next

Hide

```
ggplot(data, aes( x= State, y= Profit, fill= State),options(scipen=99)) +
          geom_col()+
          ggtitle("Statewise Profit Analysis") +
          coord_flip() +
          theme(legend.position = "None", axis.text.y = element_text(size=6))
```



Observation : From Statewise Profit Analysis , California and New York recorded the most profits. Texas was the most unprofitable among all, causing the company severe losses.

4)Regionwise Profit Analysis

Hide

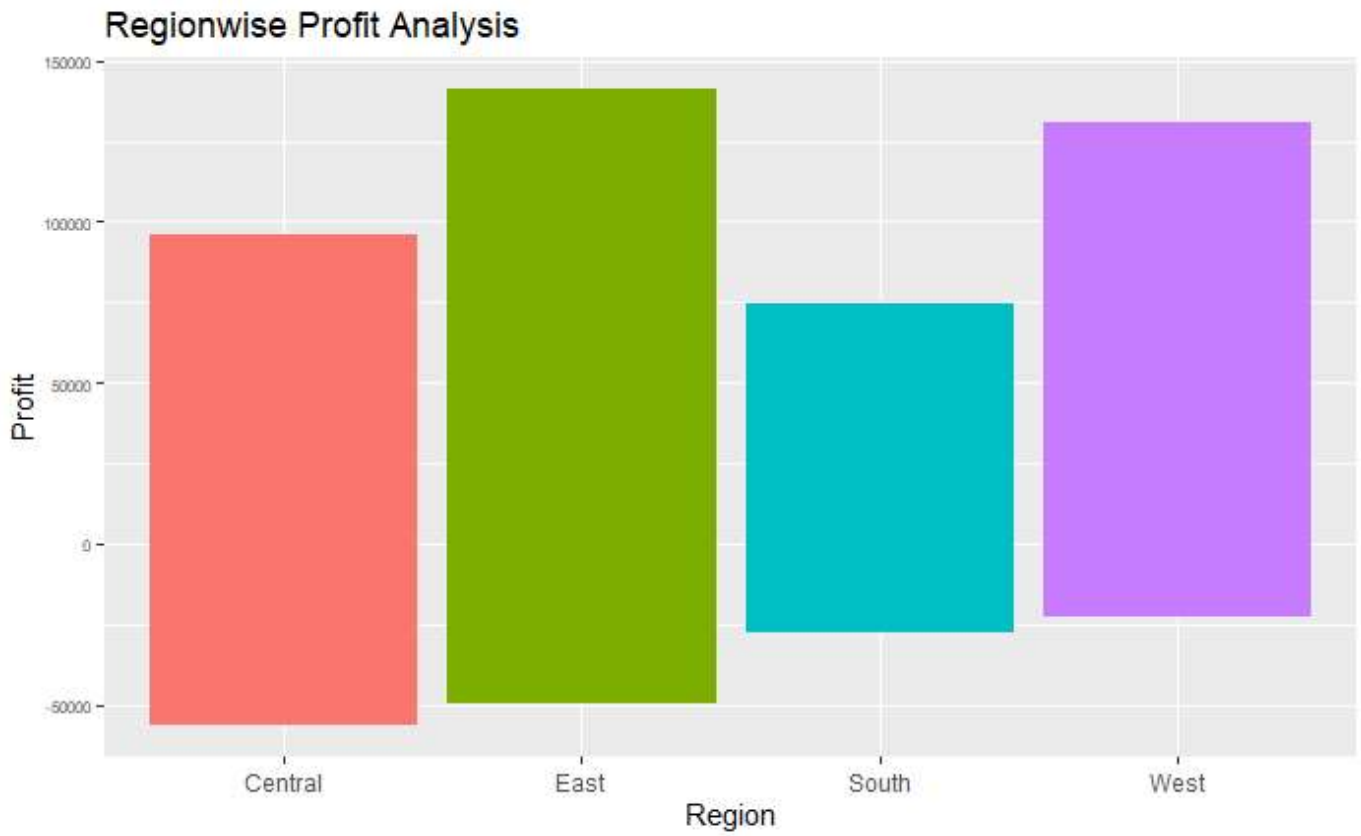
```
regionwise_profit = data %>%
        group_by(Region) %>%
       summarize(totalP= sum(Profit)) %>%
       arrange(desc(totalP))
regionwise_sales
```

Region <fctr>	totalS <dbl>
West	725255.6
East	678435.2
Central	500782.9
South	391721.9

4 rows

Hide

```
ggplot(data, aes( x= Region, y= Profit, fill= Region),options(scipen=99)) +
          geom_col()+
          ggtitle("Regionwise Profit Analysis") +
        theme(legend.position = "None", axis.text.y = element_text(size=6))
```



5. Statewise Profit-Sales Ratio Analysis (To measure how much profits are produced at a certain level of sales.)

Hide

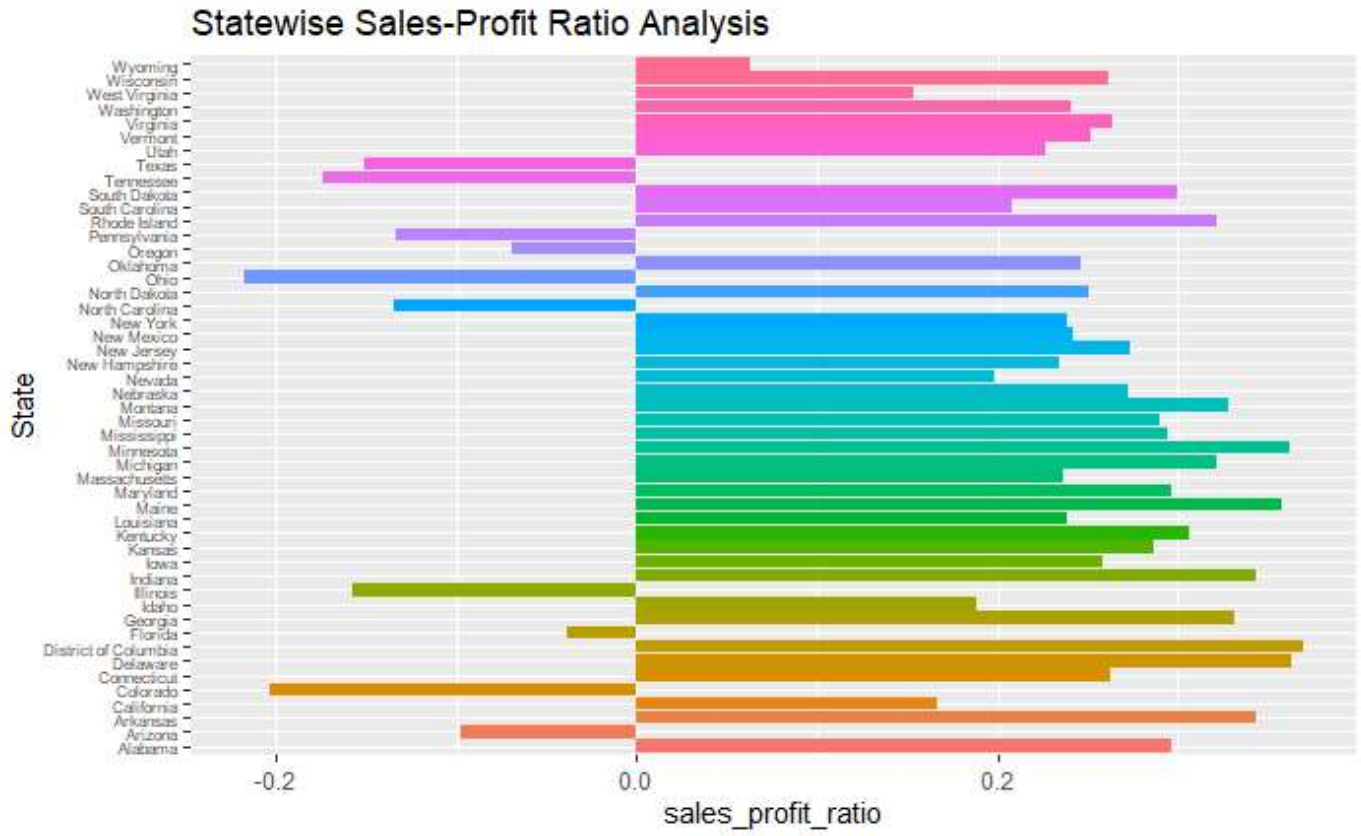
```
BarPlot = data %>%
        group_by(State) %>%
       summarize(sales_profit_ratio= sum(Profit)/sum(Sales)) %>%
       arrange(desc(sales_profit_ratio))
BarPlot
```

State <fctr>	sales_profit_ratio <dbl>
District of Columbia	0.36983662
Delaware	0.36346034
Minnesota	0.36242618
Maine	0.35771387
Arkansas	0.34326447
Indiana	0.34325110
Georgia	0.33098615
Montana	0.32800377
Rhode Island	0.32197470
Michigan	0.32193206

1-10 of 49 rows Previous **1** 2 3 4 5 Next

Hide

```
ggplot(BarPlot, aes( x= sales_profit_ratio, y= State, fill= State),options(scipen=99)) +
          geom_col()+
          ggtitle("Statewise Sales-Profit Ratio Analysis ") +
        theme(legend.position = "None", axis.text.y = element_text(size=6))
```

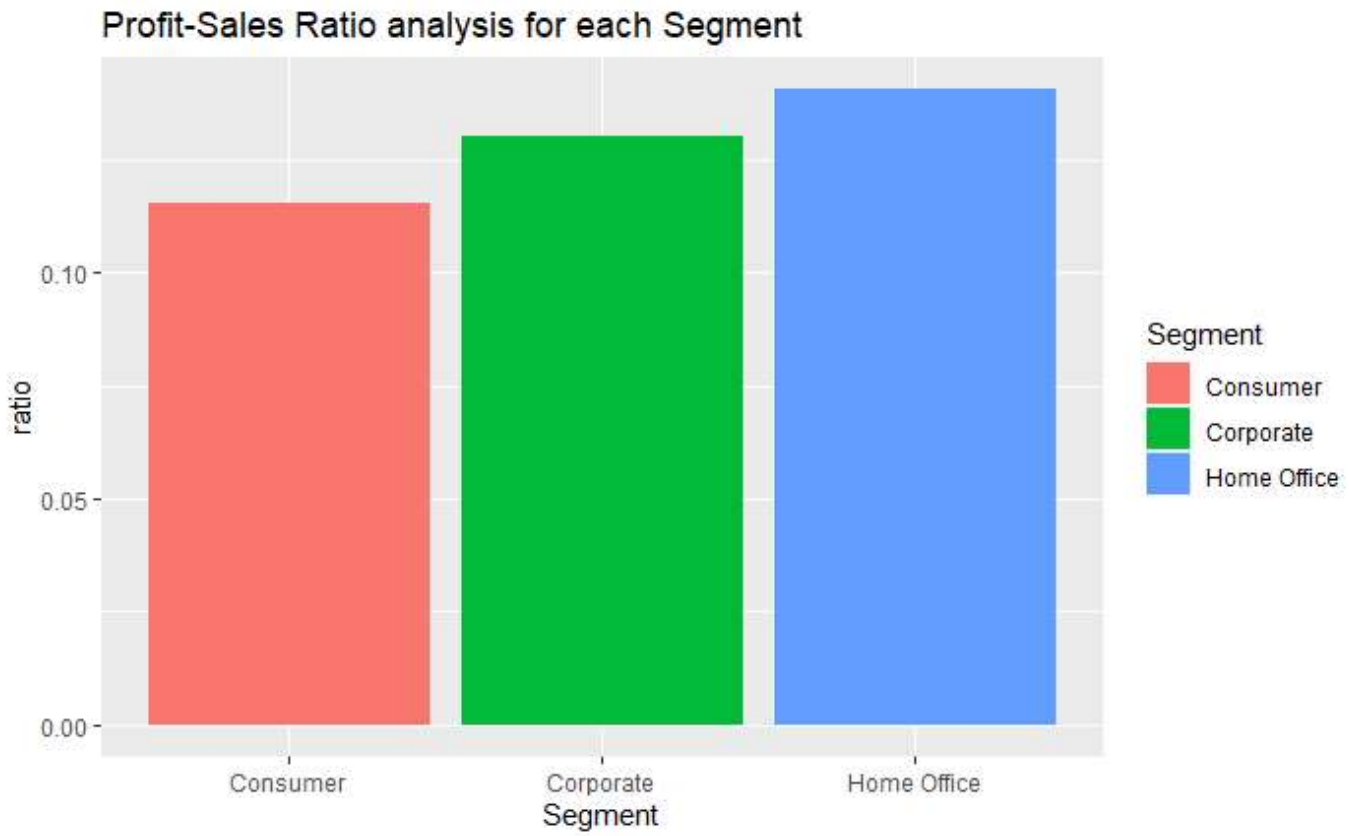


Observations : 1) Ohio has the worst profit-sales ratio(-0.2168). 2) The company should improve their Profit-Sales ratio in California, New York and Washington because even though these states have highest profits, the profit-sales ratio is not very impressive.

6. Segmentwise Sales and Profit Analysis

Hide

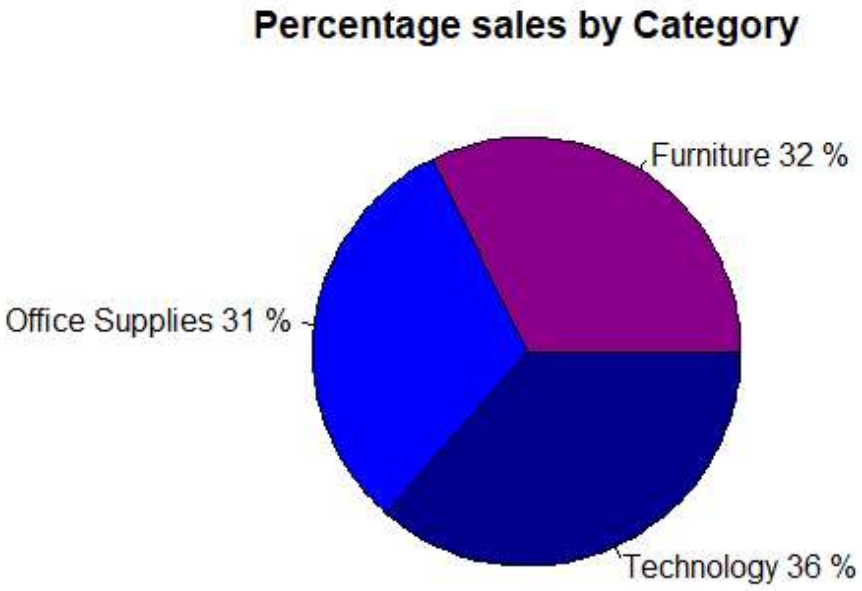
```
Segment_analysis = data %>%
        group_by(Segment) %>%
        summarize(ratio= sum(Profit)/sum(Sales)) %>%
        arrange(desc(ratio)) %>%
        ggplot( aes( x= Segment, y= ratio, fill= Segment),options(scipen=99)) +
        geom_col()+
        ggtitle("Profit-Sales Ratio analysis for each Segment ")
Segment_analysis
```



Observations: The profit-sales ratio is highest for Home Office segment. The company can improve its Profit-Sales ratio in the Consumer and Corporate Segment. 7) Percentage sales by Category.

Hide

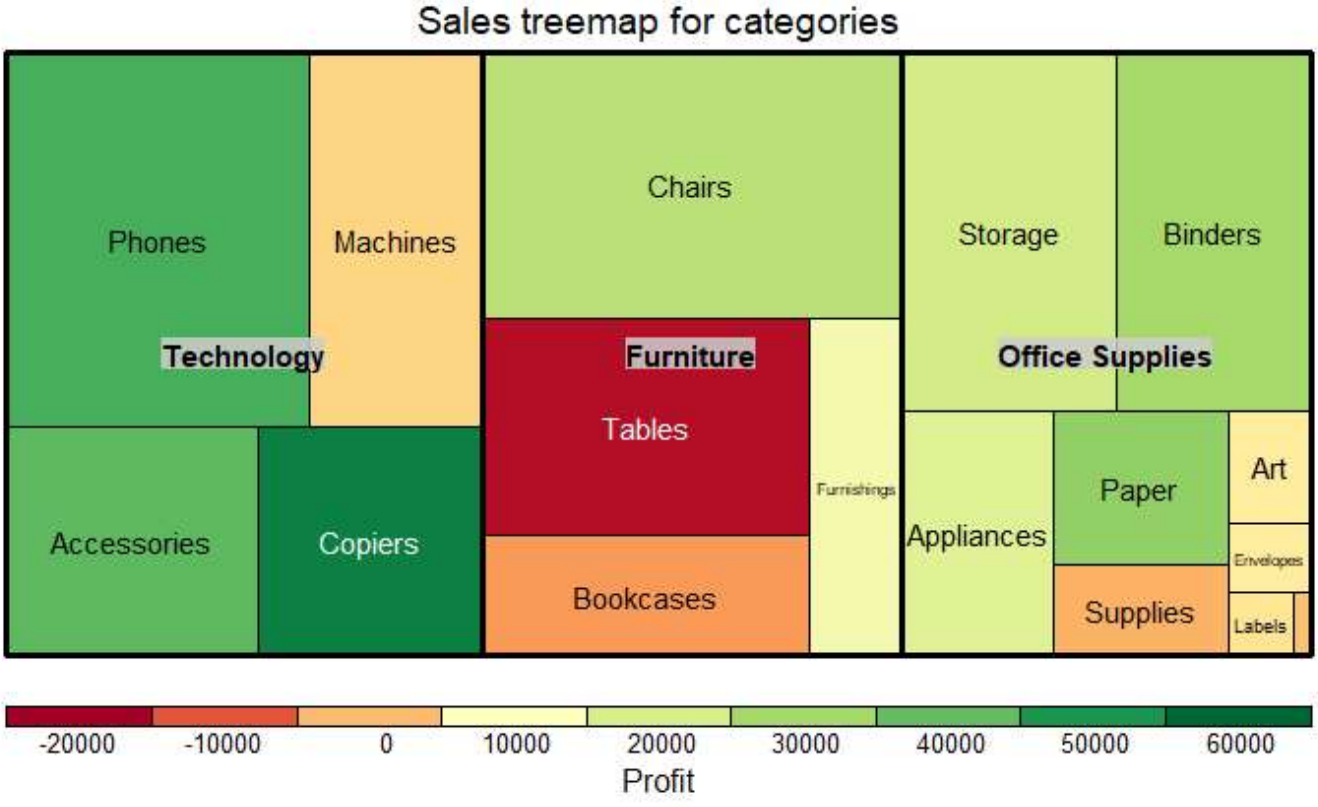
```
category_s = data %>%
                  group_by(Category) %>%
                  summarize(Sales=sum(Sales))
pct <- round((category_s$Sales/sum(category_s$Sales))*100)
lbls <- paste(category_s$Category , pct)
lbls <- paste(lbls, "%", sep = " ")
pie(category_s$Sales, labels = lbls, main =" Percentage sales by Category ", col= c('darkmage
nta','blue','blue4'))
```



8. Overall sales for Category ans Sub-Category

Hide

```
treemap(data, index = c("Category","Sub.Category"),title='Sales treemap for categories', vSize = "Sales",vColor ="Profit", type= "value",palette = "RdYlGn", range=c(-20000,60000),mapping = c(-20000,10000,60000))
```

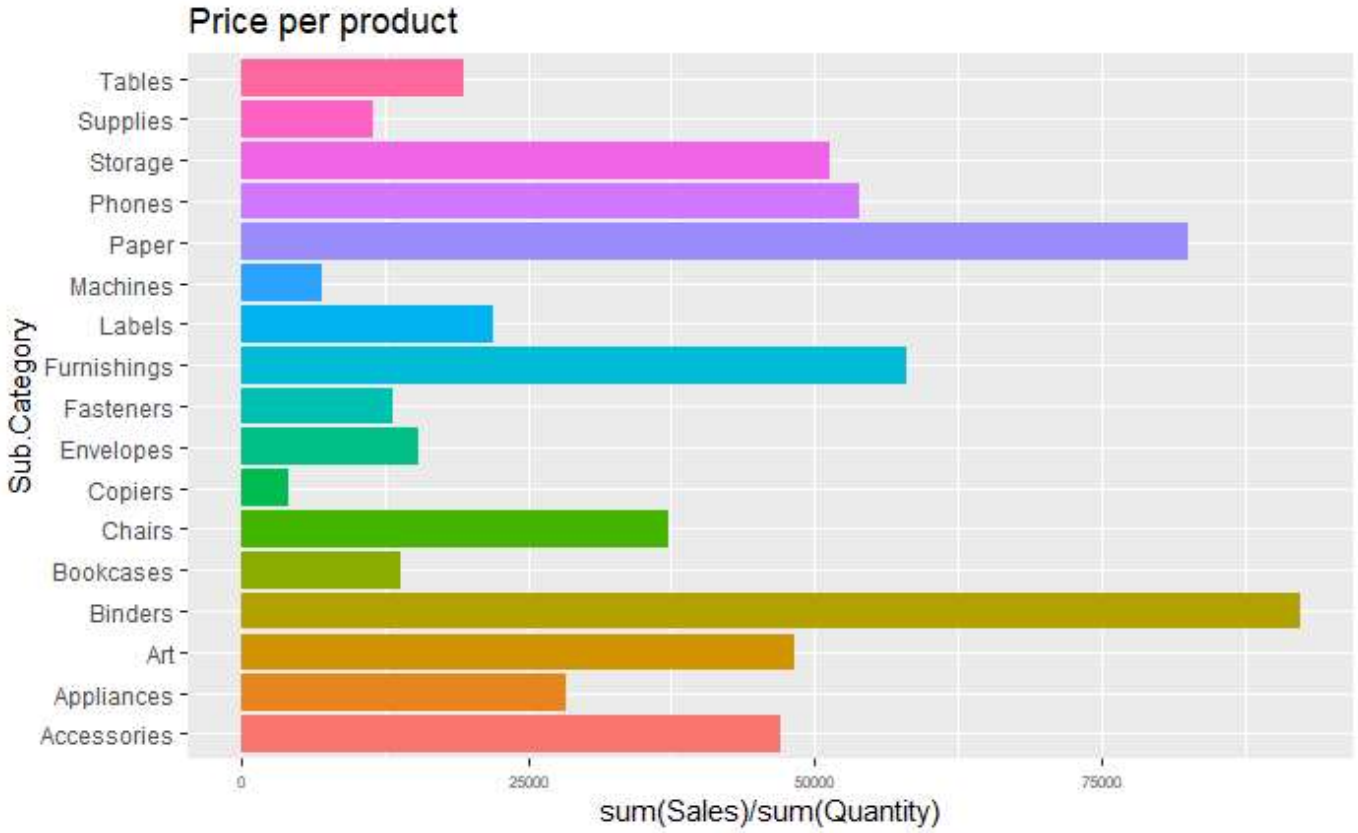


Observation : The above Treemap provides information about the sales and profit of various product category and sub-category. The cell size in the treemap is decided by the sales and the colour gradient describes profit. It can be concluded that “Phones” under “Technology” has the highest sale. “Tables” under “Furniture” incurred highest loss, while “Copiers” under “Technology” was most profitable.

9. Price per Product in different Sub-Categories

Hide

```
Price_per_product = ggplot(data, aes( x= Sub.Category, y=sum(Sales)/sum(Quantity), fill= Sub.
Category),options(scipen=99)) +
          geom_col()+
          ggtitle("Price per product") +
          coord_flip() +
        theme(legend.position = "None", axis.text.x = element_text(size=6))
Price_per_product
```

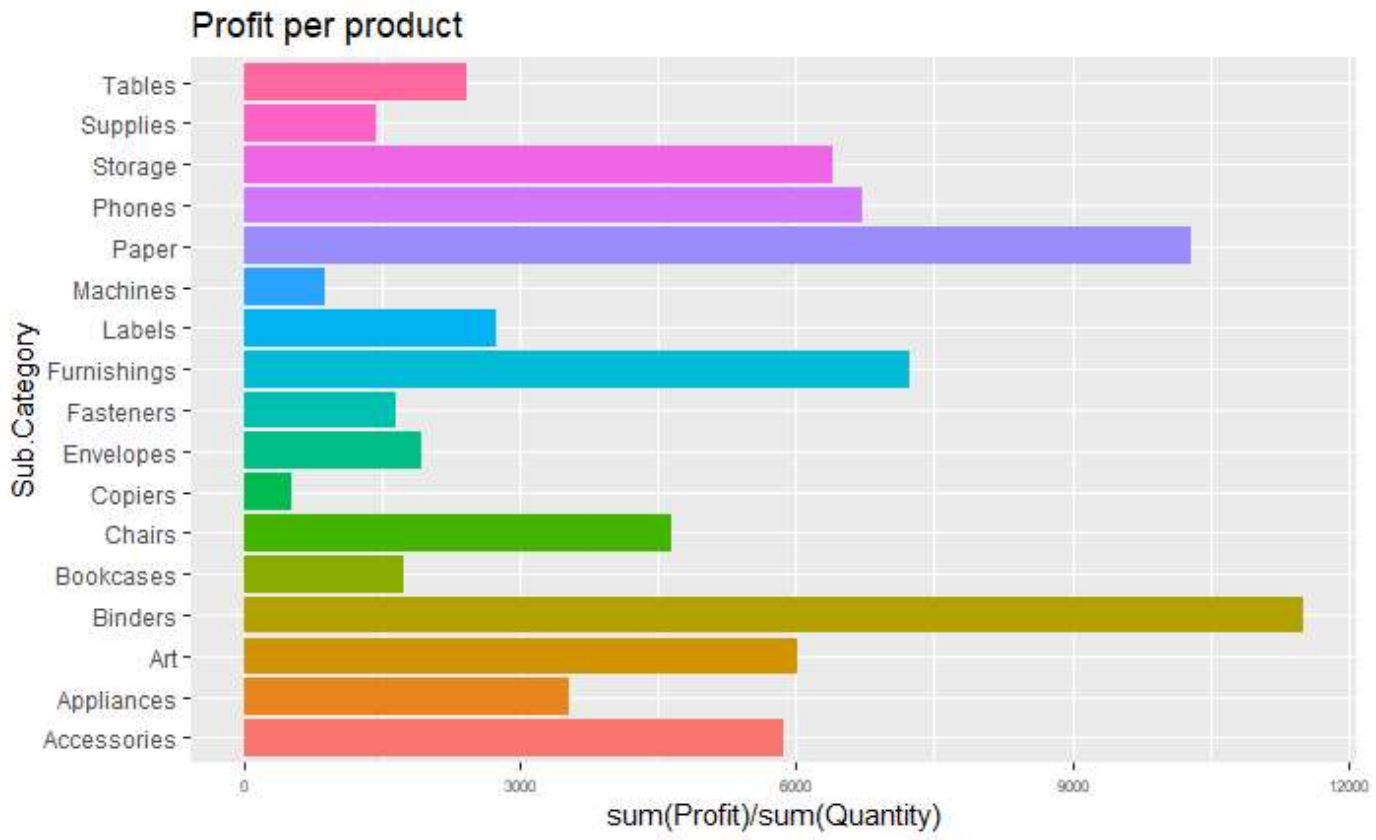


Observation : Binders have the highest price, followed by paper. Copiers had the least price among all.

10. Profit per product in different Sub-Category

Hide

```
profit_per_product = ggplot(data, aes( x= Sub.Category, y=sum(Profit)/sum(Quantity), fill= Su
b.Category),options(scipen=99)) +
          geom_col()+
          ggtitle("Profit per product") +
          coord_flip()+
        theme(legend.position = "None", axis.text.x = element_text(size=6))
profit_per_product
```

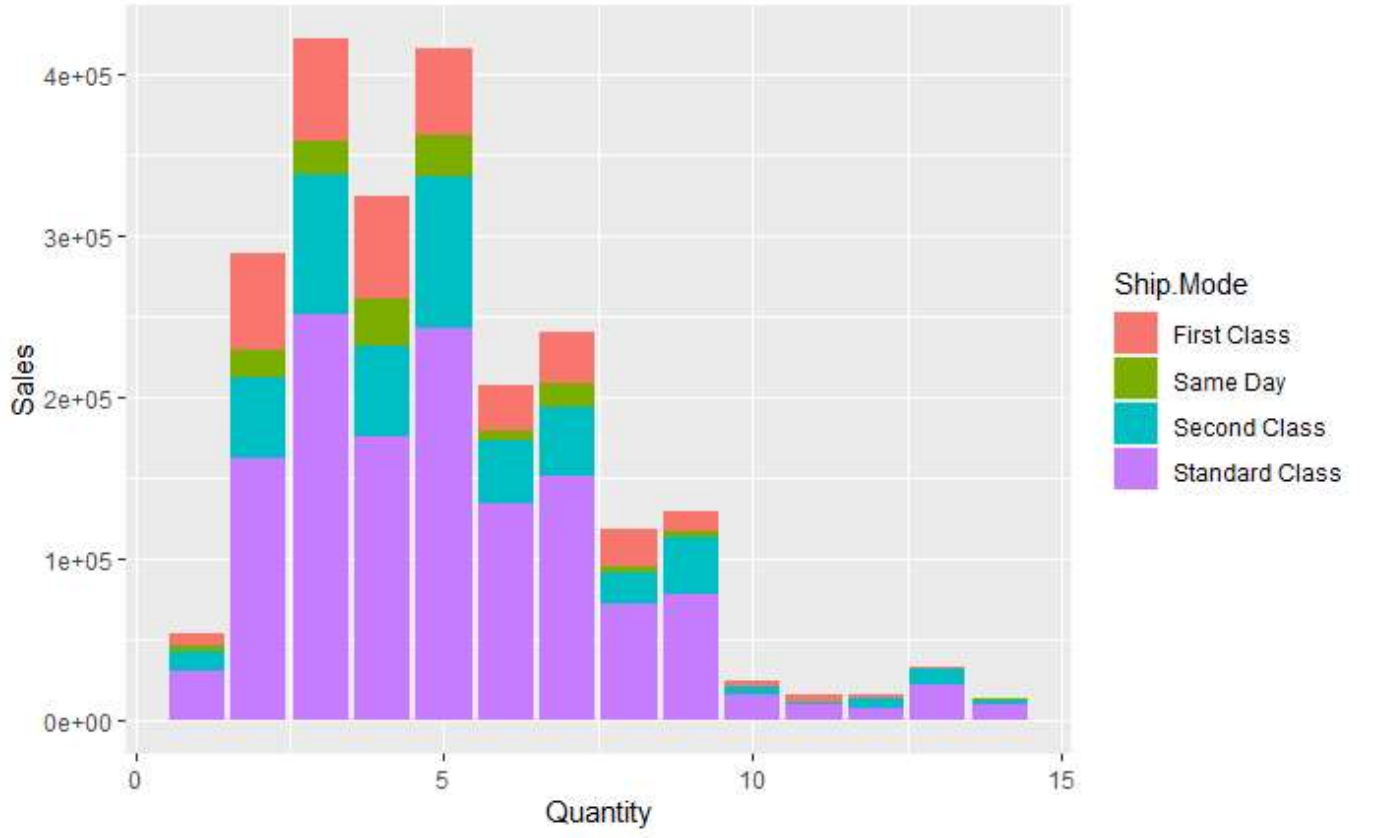


Observation : Binders make highest profit per product.

11. Sales vs Quantity

Hide

```
ggplot(data, aes(x = Quantity, y = Sales, fill = Ship.Mode),options(scipen=99) )+ geom_bar(st
at = "identity")
```

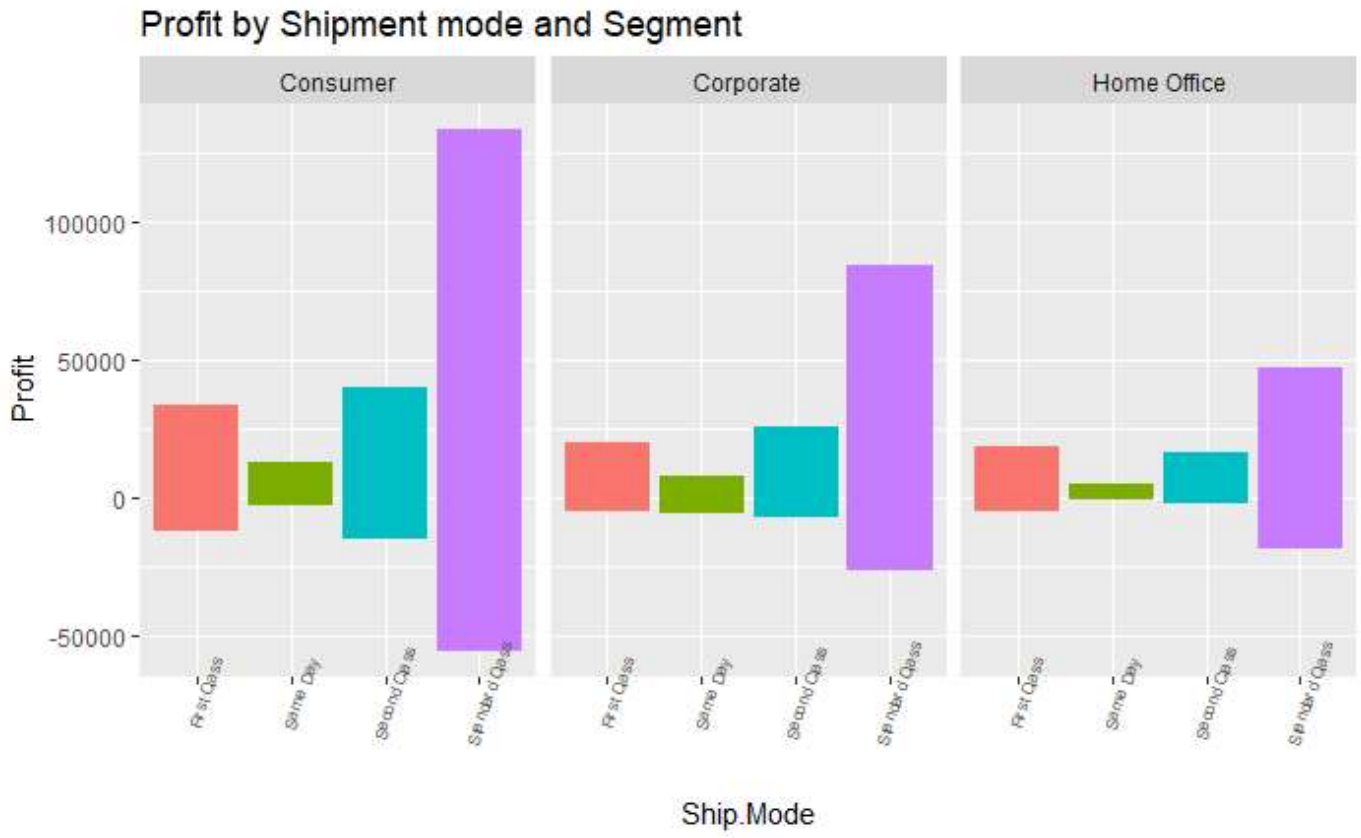


Observation : We see the following pattern that most of the sales have been triggered by the standard class of shipment mode.

12. Profit by Shipment mode and Segment

Hide

```
ggplot(data, aes( x=Ship.Mode, y= Profit, fill= Ship.Mode),options(scipen=99)) +
          geom_col()+
          ggtitle("Profit by Shipment mode and Segment") +
        theme(legend.position = "None", axis.text.x = element_text(angle = 70 ,size=6)) +
         facet_wrap(~Segment)
```



Observations : Standard Class is the most used shipping mode and Consumer segment is the largest segment among the three.

13. Sales with and without Discount

Hide

```
Sales_with_discount = data %>%
                      filter(Discount != 0) %>%
                      summarize(totals=sum(Sales))
Sales_with_discount
```

totals
<dbl>
1208918

1 row

Hide

```
Sales_without_discount = data %>%
                      filter(Discount == 0) %>%
                      summarize(totals=sum(Sales))
Sales_without_discount
```

totals <dbl>
1087278

1 row

Observation : Sales are high when discount is offered. 14) Profit with and without discount

Hide

```
profit_with_discount = data %>%
                      filter(Discount != 0) %>%
                      summarize(totalp=sum(Profit))
profit_with_discount
```

totalp <dbl>
-34602.98

1 row

Hide

```
profit_without_discount = data %>%
                      filter(Discount == 0) %>%
                      summarize(totalp=sum(Profit))
profit_without_discount
```

The company incurres loss when discount is given. So that area should be monitored.