Task3 Spark Foundation

Sannidhya Das

2023-11-10

Domain : Data Science and Business Analytics

Batch : GRIPNOVEMBER23

Setting the working directory and loading Dataset

```
getwd()
```

```
## [1] "C:/Users/sanni/OneDrive/Documents/Internship projects/Spark Foundation/Task3"
```

```
setwd("C:/Users/sanni/OneDrive/Documents/Internship projects/Spark Foundation/Task3")
task3=read.csv("SampleSuperstore.csv")
```

Checking the dataset

```
head(task3)
```

```
##        Ship.Mode   Segment       Country            City      State Postal.Code
## 1   Second Class  Consumer United States       Henderson   Kentucky       42420
## 2   Second Class  Consumer United States       Henderson   Kentucky       42420
## 3   Second Class Corporate United States     Los Angeles California       90036
## 4 Standard Class  Consumer United States Fort Lauderdale    Florida       33311
## 5 Standard Class  Consumer United States Fort Lauderdale    Florida       33311
## 6 Standard Class  Consumer United States     Los Angeles California       90032
##   Region        Category Sub.Category    Sales Quantity Discount    Profit
## 1  South       Furniture    Bookcases 261.9600        2     0.00   41.9136
## 2  South       Furniture       Chairs 731.9400        3     0.00  219.5820
## 3   West Office Supplies       Labels  14.6200        2     0.00    6.8714
## 4  South       Furniture       Tables 957.5775        5     0.45 -383.0310
## 5  South Office Supplies      Storage  22.3680        2     0.20    2.5164
## 6   West       Furniture  Furnishings  48.8600        7     0.00   14.1694
```

```
tail(task3)
```

```
##           Ship.Mode   Segment       Country        City      State Postal.Code
## 9989 Standard Class Corporate United States      Athens    Georgia       30605
## 9990   Second Class  Consumer United States       Miami    Florida       33180
## 9991 Standard Class  Consumer United States  Costa Mesa California       92627
## 9992 Standard Class  Consumer United States  Costa Mesa California       92627
## 9993 Standard Class  Consumer United States  Costa Mesa California       92627
## 9994   Second Class  Consumer United States Westminster California       92683
##      Region   Category Sub.Category   Sales Quantity Discount  Profit
## 9989  South Technology       Phones 206.100        5      0.0 55.6470
## 9990  South  Furniture  Furnishings  25.248        3      0.2  4.1028
## 9991   West  Furniture  Furnishings  91.960        2      0.0 15.6332
## 9992   West Technology       Phones 258.576        2      0.2 19.3932
## 9993   West Office Supplies        Paper  29.600        4      0.0 13.3200
## 9994   West Office Supplies   Appliances 243.160        2      0.0 72.9480
```

Summary of the dataset

```r
summary(task3)
```

```
##   Ship.Mode           Segment            Country             City          
##  Length:9994        Length:9994        Length:9994        Length:9994       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##     State            Postal.Code       Region            Category        
##  Length:9994        Min.   : 1040   Length:9994        Length:9994       
##  Class :character   1st Qu.:23223   Class :character   Class :character  
##  Mode  :character   Median :56431   Mode  :character   Mode  :character  
##                     Mean   :55190                                        
##                     3rd Qu.:90008                                        
##                     Max.   :99301                                        
##  Sub.Category           Sales              Quantity        Discount     
##  Length:9994        Min.   :    0.444   Min.   : 1.00   Min.   :0.0000  
##  Class :character   1st Qu.:   17.280   1st Qu.: 2.00   1st Qu.:0.0000  
##  Mode  :character   Median :   54.490   Median : 3.00   Median :0.2000  
##                     Mean   :  229.858   Mean   : 3.79   Mean   :0.1562  
##                     3rd Qu.:  209.940   3rd Qu.: 5.00   3rd Qu.:0.2000  
##                     Max.   :22638.480   Max.   :14.00   Max.   :0.8000  
##      Profit         
##  Min.   :-6599.978  
##  1st Qu.:    1.729  
##  Median :    8.666  
##  Mean   :   28.657  
##  3rd Qu.:   29.364  
##  Max.   : 8399.976
```

```r
str(task3)
```

```
## 'data.frame':    9994 obs. of  13 variables:
##  $ Ship.Mode   : chr  "Second Class" "Second Class" "Second Class" "Standard Class" ...
##  $ Segment     : chr  "Consumer" "Consumer" "Corporate" "Consumer" ...
##  $ Country     : chr  "United States" "United States" "United States" "United States" ...
##  $ City        : chr  "Henderson" "Henderson" "Los Angeles" "Fort Lauderdale" ...
##  $ State       : chr  "Kentucky" "Kentucky" "California" "Florida" ...
##  $ Postal.Code : int  42420 42420 90036 33311 33311 90032 90032 90032 90032 90032 ...
##  $ Region      : chr  "South" "South" "West" "South" ...
##  $ Category    : chr  "Furniture" "Furniture" "Office Supplies" "Furniture" ...
##  $ Sub.Category: chr  "Bookcases" "Chairs" "Labels" "Tables" ...
##  $ Sales       : num  262 731.9 14.6 957.6 22.4 ...
##  $ Quantity    : int  2 3 2 5 2 7 4 6 3 5 ...
##  $ Discount    : num  0 0 0 0.45 0.2 0 0 0.2 0.2 0 ...
##  $ Profit      : num  41.91 219.58 6.87 -383.03 2.52 ...
```

```r
dim(task3)
```

```
## [1] 9994   13
```

```
colnames(task3)
```

```
##  [1] "Ship.Mode"    "Segment"      "Country"      "City"         "State"
##  [6] "Postal.Code"  "Region"       "Category"     "Sub.Category" "Sales"
## [11] "Quantity"     "Discount"     "Profit"
```

Checking is there is any null values in any columns

```
colSums(is.na(task3))
```

```
##    Ship.Mode      Segment      Country         City        State  Postal.Code
##            0            0            0            0            0            0
##       Region     Category Sub.Category        Sales     Quantity     Discount
##            0            0            0            0            0            0
##       Profit
##            0
```

Checking the dataset for duplicates and dropping the duplicate elements using unique()

```
sum(duplicated(task3))
```

```
## [1] 17
```

```
task3=unique(task3)
```

Finding the correlation and covariance of dataset using cor()and cov() method

```
cor(task3[,c("Sales","Quantity","Discount","Profit")])
```

```
##                Sales    Quantity     Discount      Profit
## 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
## Profit    0.47906731 0.066210646 -0.219662064  1.00000000
```

```
cov(task3[,c("Sales","Quantity","Discount","Profit")])
```

```
##                  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
```

Group the data by multiple columns and calculate the sum of Quantity, Discount, Sales, and Profit

```
grouped <- aggregate(cbind(Quantity, Discount, Sales, Profit) ~ Ship.Mode + Segment + Category +
                       Sub.Category+ State + Region,
                     data = task3, sum)
```

Print the grouped data

```
head(grouped)
```

```
##        Ship.Mode   Segment   Category Sub.Category    State  Region Quantity
## 1       Same Day  Consumer Technology  Accessories Illinois Central        3
## 2   Second Class  Consumer Technology  Accessories Illinois Central       18
## 3 Standard Class  Consumer Technology  Accessories Illinois Central       55
## 4    First Class Corporate Technology  Accessories Illinois Central        7
## 5   Second Class Corporate Technology  Accessories Illinois Central       17
## 6 Standard Class Corporate Technology  Accessories Illinois Central        6
##   Discount     Sales   Profit
## 1      0.2   39.264  -4.9080
## 2      0.8  983.728 231.2496
## 3      2.8 1603.768 240.9968
## 4      0.2  116.312  23.2624
## 5      0.6  490.184  94.7115
## 6      0.4  196.768  22.6196
```

Group the data by State and calculate the sum, mean, min, max, count, median, standard deviation, and variance of Profit

```
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
profit_summary <- task3%>%
  group_by(State) %>%
  summarise(sum = sum(Profit), mean = mean(Profit), min = min(Profit), max = max(Profit), count = n(),
            median = median(Profit), std = sd(Profit), var = var(Profit))
```

Print the summary statistics of Profit by State

```
profit_summary
```

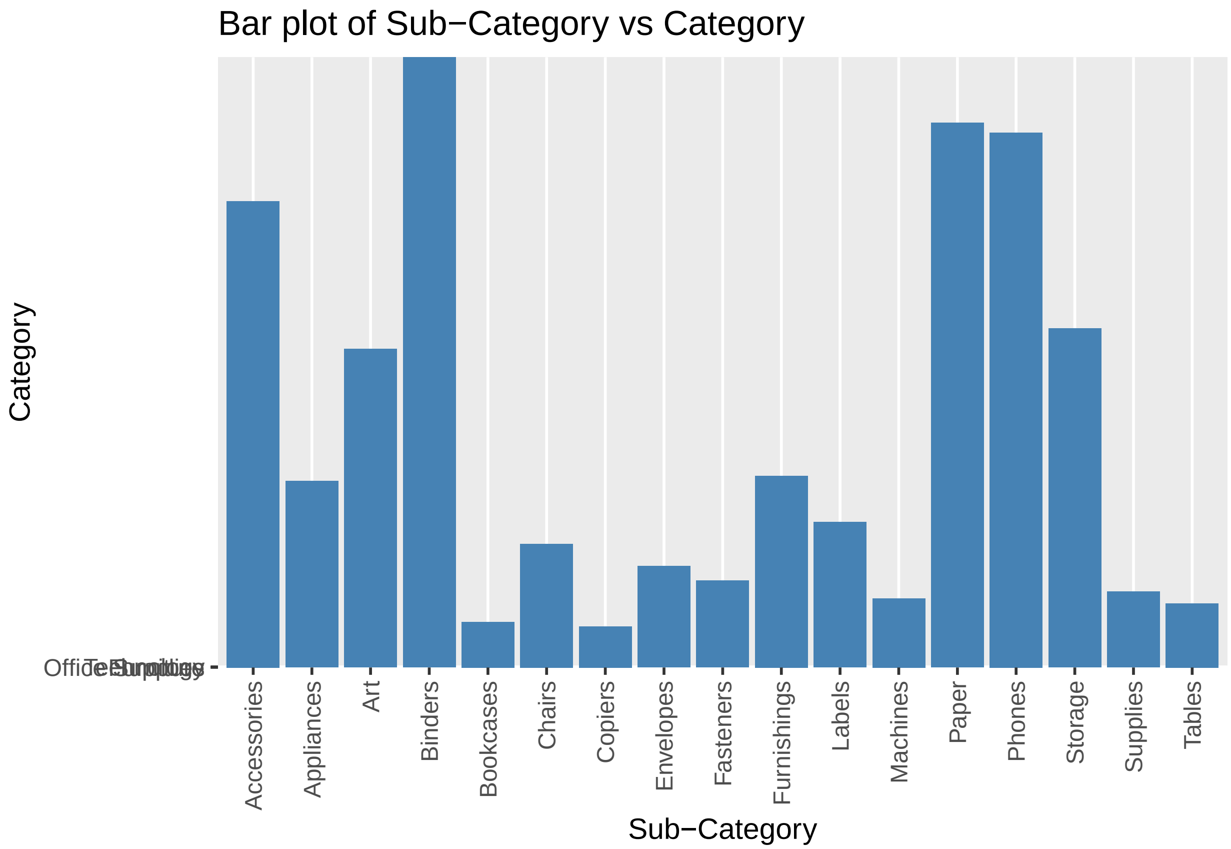
```
## # A tibble: 49 x 9
##    State                     sum   mean       min   max count median   std     var
##    <chr>                   <dbl>  <dbl>     <dbl> <dbl> <int>  <dbl> <dbl>   <dbl>
##  1 Alabama               5787.  94.9      0     1459.    61  16.9 211.   44480.
##  2 Arizona              -3428. -15.3   -814.     211.   224   2.53 109.   11939.
##  3 Arkansas              4009.  66.8      1.42   843.    60  18.3 123.   15191.
##  4 California           76331.  38.2   -326.    1906.  1996  13.3   97.8   9566.
##  5 Colorado             -6528. -35.9  -3400.     248.   182   3.12 276.   76410.
##  6 Connecticut           3511.  42.8    -15.6    295.    82  12.2   66.1   4374.
##  7 Delaware              9977. 104.     -48.8   5040.    96  19.2 519.  269313.
##  8 District of Columbia  1060. 106.       4.43   649.    10  14.5 213.   45566.
##  9 Florida              -3399.  -8.88 -1811.     328.   383   2.93 126.   15958.
## 10 Georgia              16250.  88.3      0.113 3177.   184  22.2 283.   80104.
## # i 39 more rows
```

Visualization of dataset

```
library(ggplot2)
```

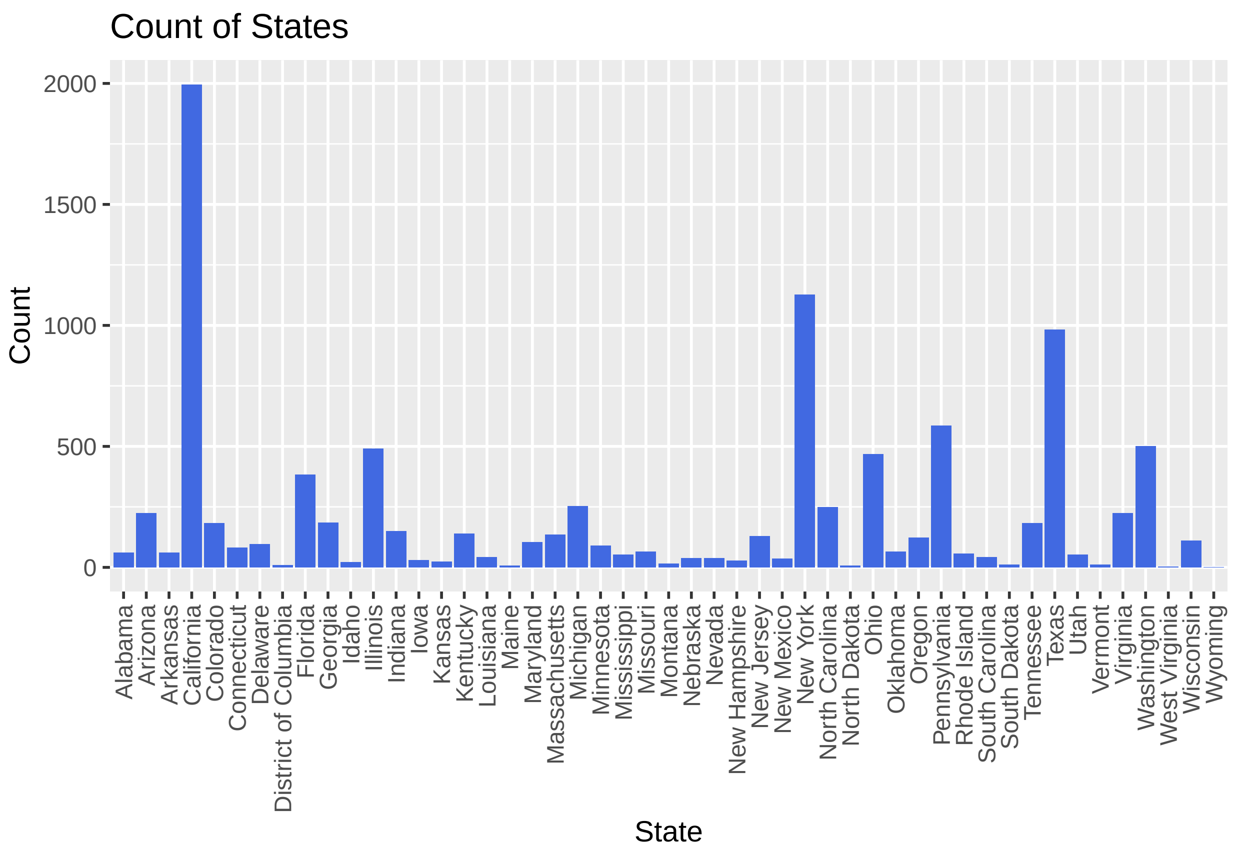
```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```r
ggplot(task3, aes(x = Sub.Category, y = Category)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ggtitle("Bar plot of Sub-Category vs Category") +
  xlab("Sub-Category") +
  ylab("Category")
```



# Note : Binders are purchased maximum times from the store followed by papers and phones .

```r
library(dplyr)

task3 %>%
  count(State) %>%
  ggplot(aes(x = State, y = n)) +
  geom_bar(stat = "identity", fill = "royalblue") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ggtitle("Count of States") +
  xlab("State") +
  ylab("Count")
```



Note :

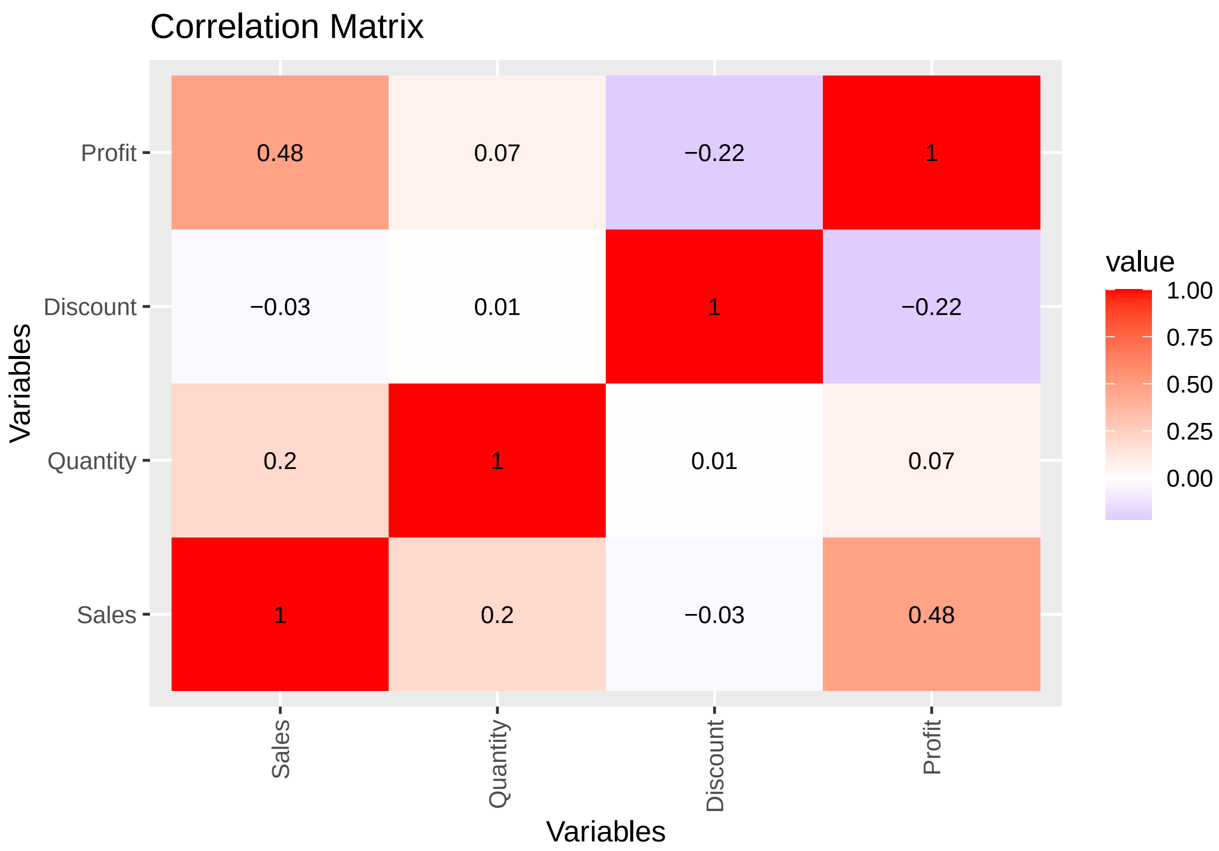
Products are very often ordered from California , New York and Texus .

Heatmap plot

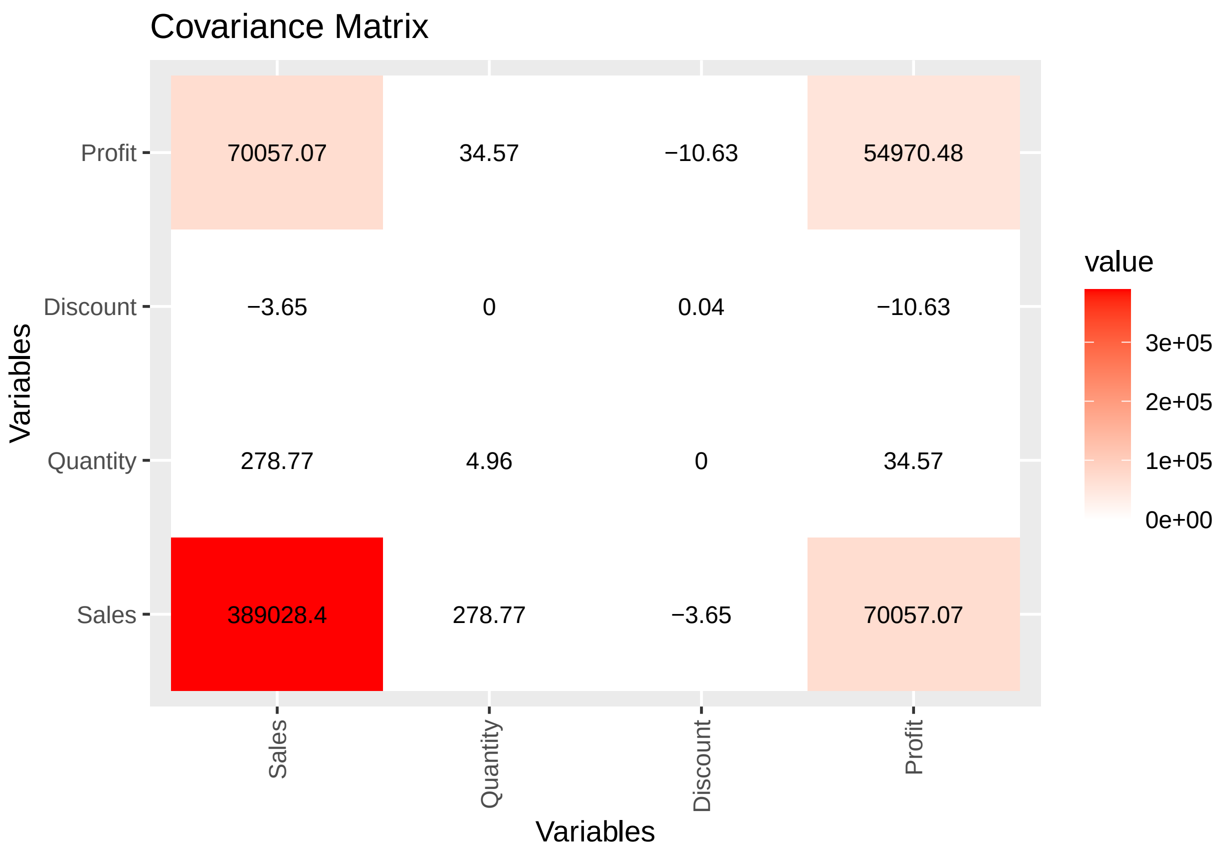
```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.1.3
```

```
# Creating a correlation matrix
corr <- cor(task3[,c("Sales","Quantity","Discount","Profit")])

# Creating a heatmap with annotations
ggplot(melt(corr), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Correlation Matrix") +
  xlab("Variables") +
  ylab("Variables")
```



```r
# Creating a covariance matrix
cov <- cov(task3[,c("Sales","Quantity","Discount","Profit")])

# Creating a heatmap with annotations
ggplot(melt(cov), aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Covariance Matrix") +
  xlab("Variables") +
  ylab("Variables")
```



Note:

1.There is a positive Correlation between Sales and profit.(Sales Increase Profit Increases)

2.There is a positive Correlation between Quantity and Profit.(Quantity Increase Profit Increases)

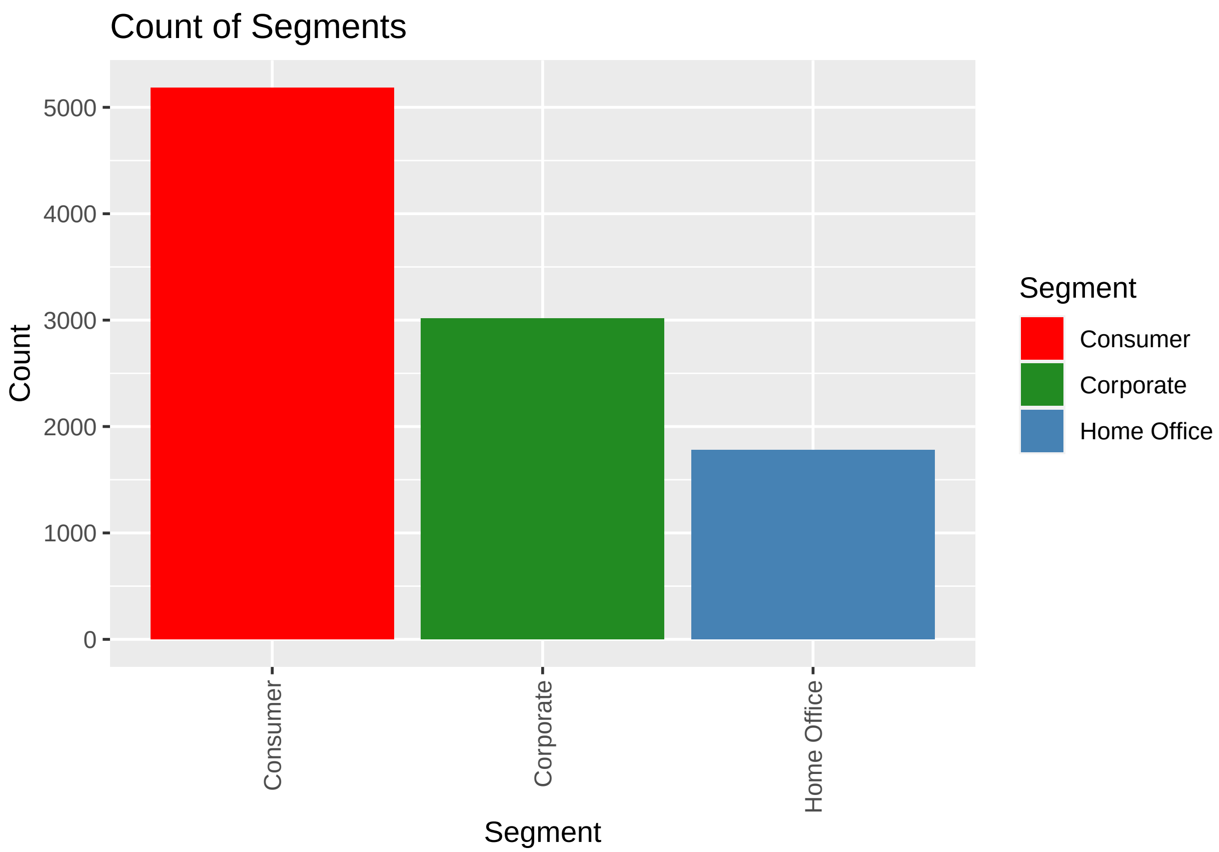
3.There is a Negetive Correlation between Profit and Discount.(Discount Increase Profit Dicreases)

4.There is Negative Correlation between Sales and Discount.(Sales Increase Discount Decreases)

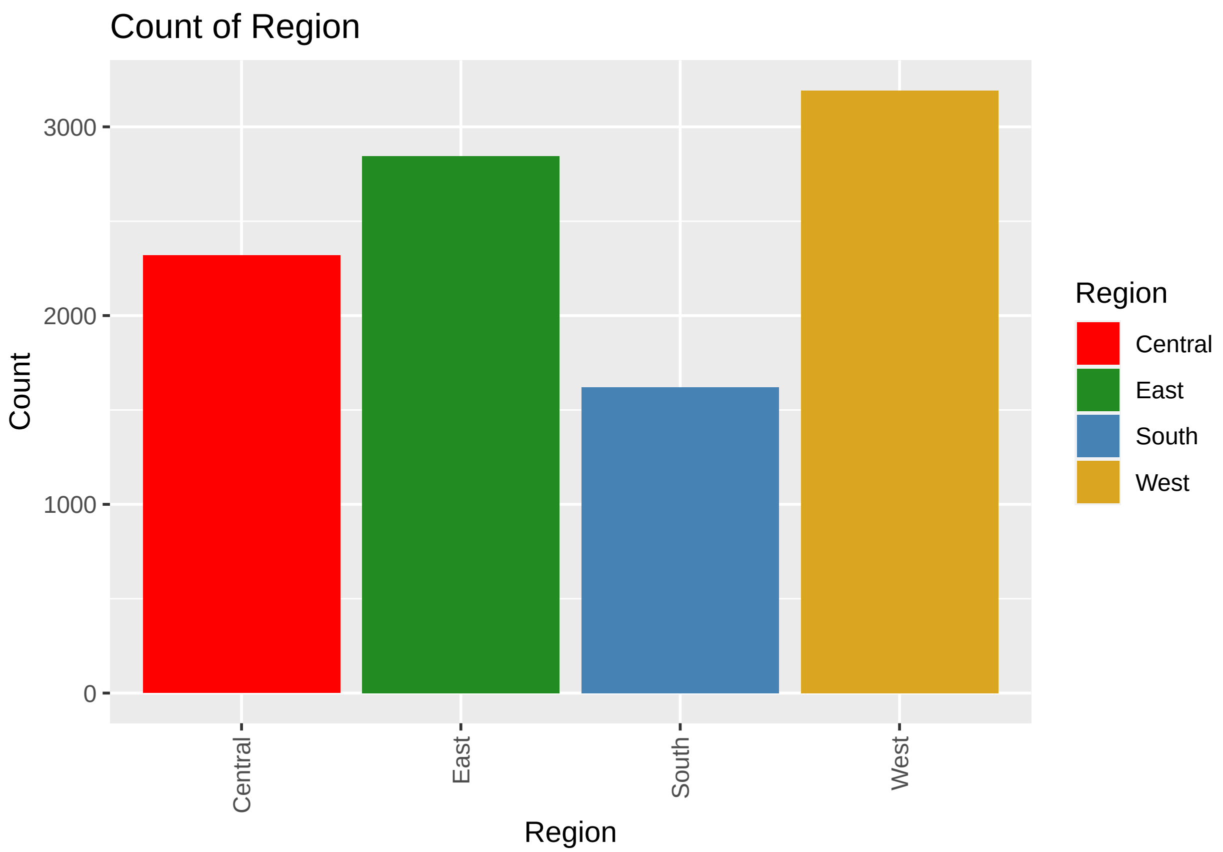
5.There is Nearly no Correlation between Quantity and Discount.(0 Correlation)

Creating a count plot

```r
ggplot(task3, aes(x = Segment, fill = Segment)) +
  geom_bar() +
  scale_fill_manual(values = c("red", "forestgreen", "steelblue")) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Count of Segments") +
  xlab("Segment") +
  ylab("Count")
```



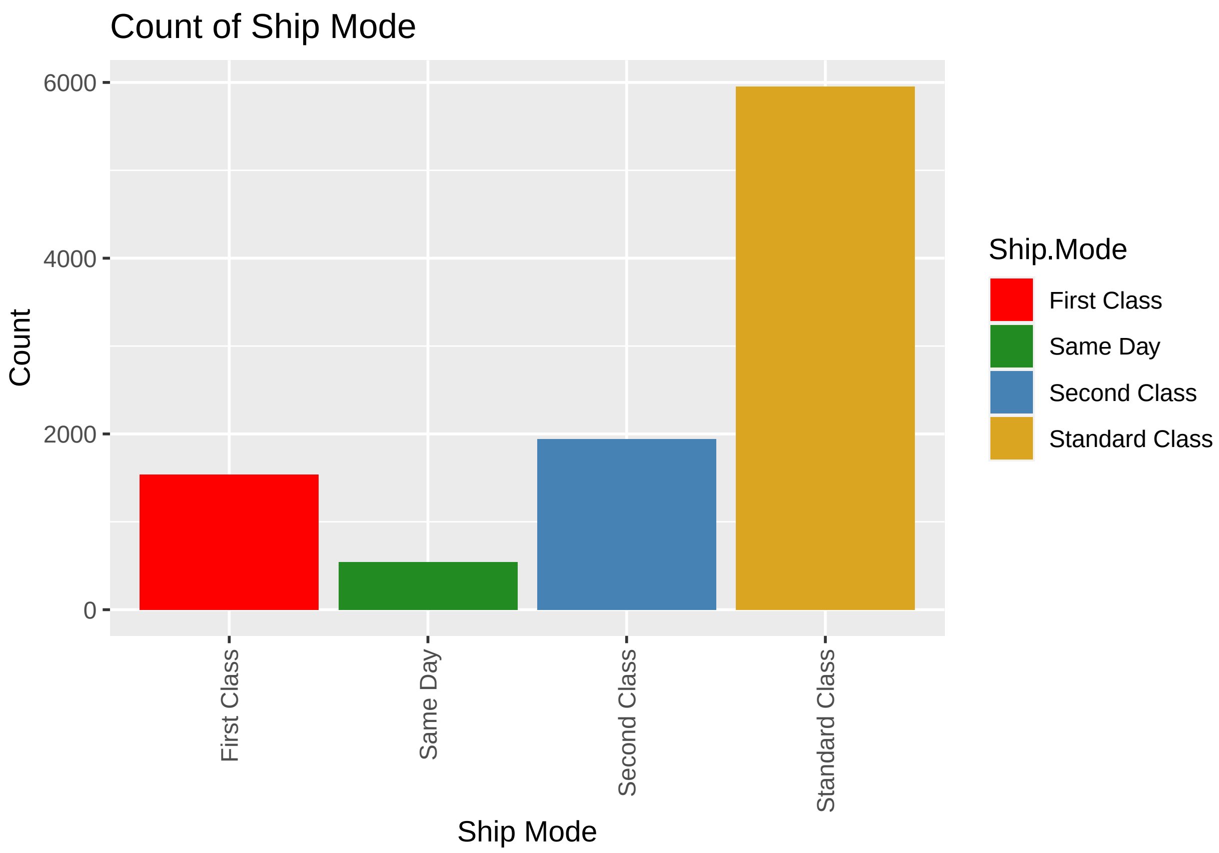
```
ggplot(task3, aes(x = Region, fill = Region)) +
  geom_bar() +
  scale_fill_manual(values = c("red", "forestgreen", "steelblue","goldenrod")) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Count of Region") +
  xlab("Region") +
  ylab("Count")
```



Note :

People from Western region orders more products from this store than East , Central and South .

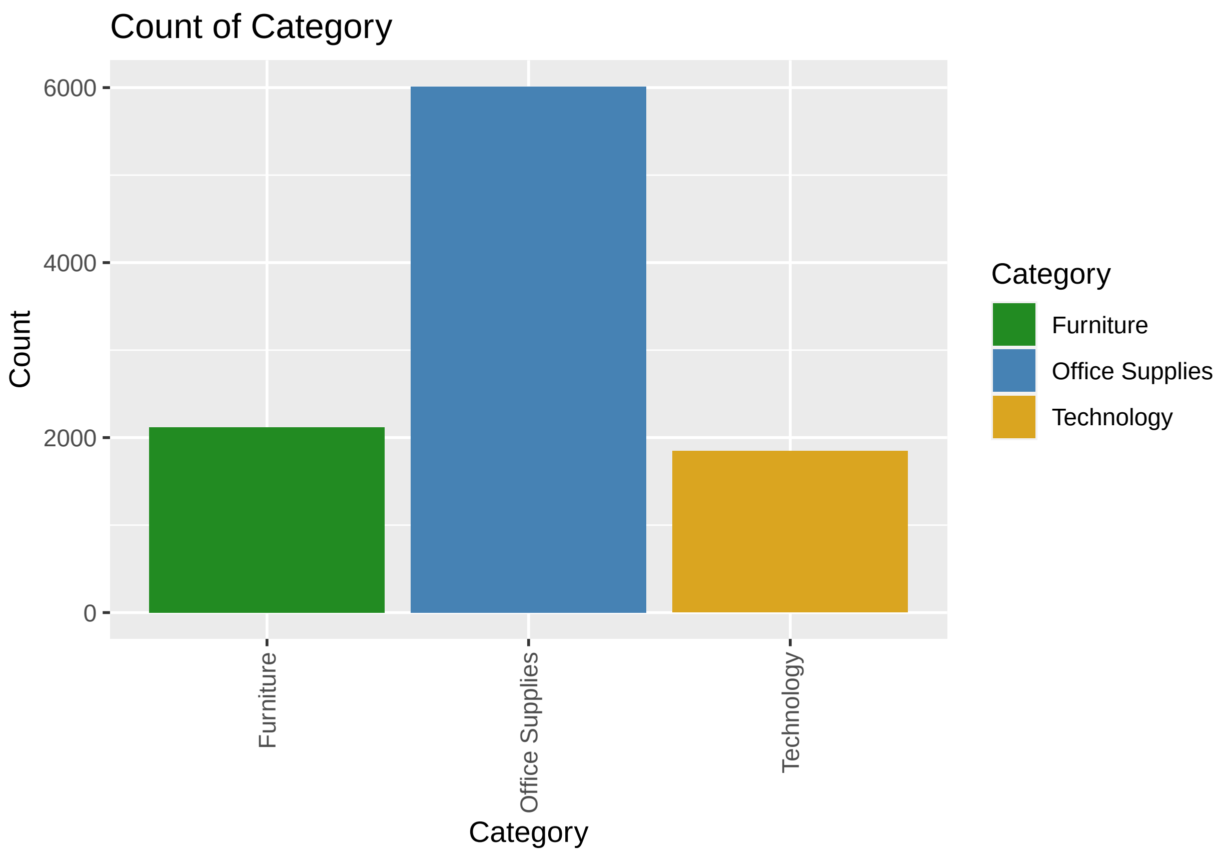
```
ggplot(task3, aes(x = Ship.Mode, fill = Ship.Mode)) +
  geom_bar() +
  scale_fill_manual(values = c("red", "forestgreen", "steelblue","goldenrod")) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Count of Ship Mode") +
  xlab("Ship Mode") +
  ylab("Count")
```



Note :

When purchasing goods from the store, most customers choose Standard class shipment.

```
ggplot(task3, aes(x = Category, fill = Category)) +
  geom_bar() +
  scale_fill_manual(values = c("forestgreen", "steelblue","goldenrod")) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Count of Category") +
  xlab("Category") +
  ylab("Count")
```

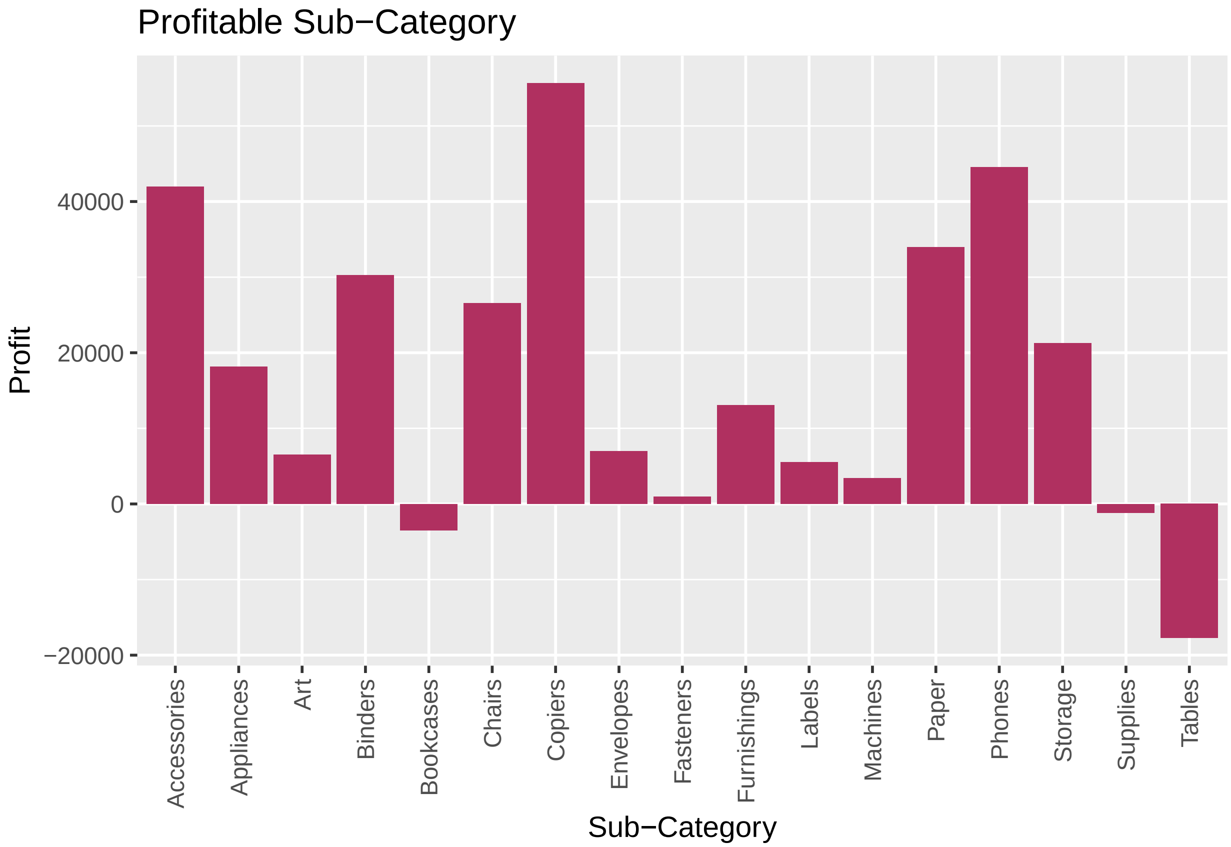


Note :

People prefers to purchase Office supplies than tech and furniture goods .

Profit Associated with Sub-Category

```
profit_SubCategory <- aggregate(Profit ~ Sub.Category, data = task3, FUN = sum)

ggplot(profit_SubCategory, aes(x = Sub.Category, y = Profit)) +
  geom_bar(stat = "identity",fill="Maroon") +
  labs(x = "Sub-Category", y = "Profit" ) +
  ggtitle("Profitable Sub-Category") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

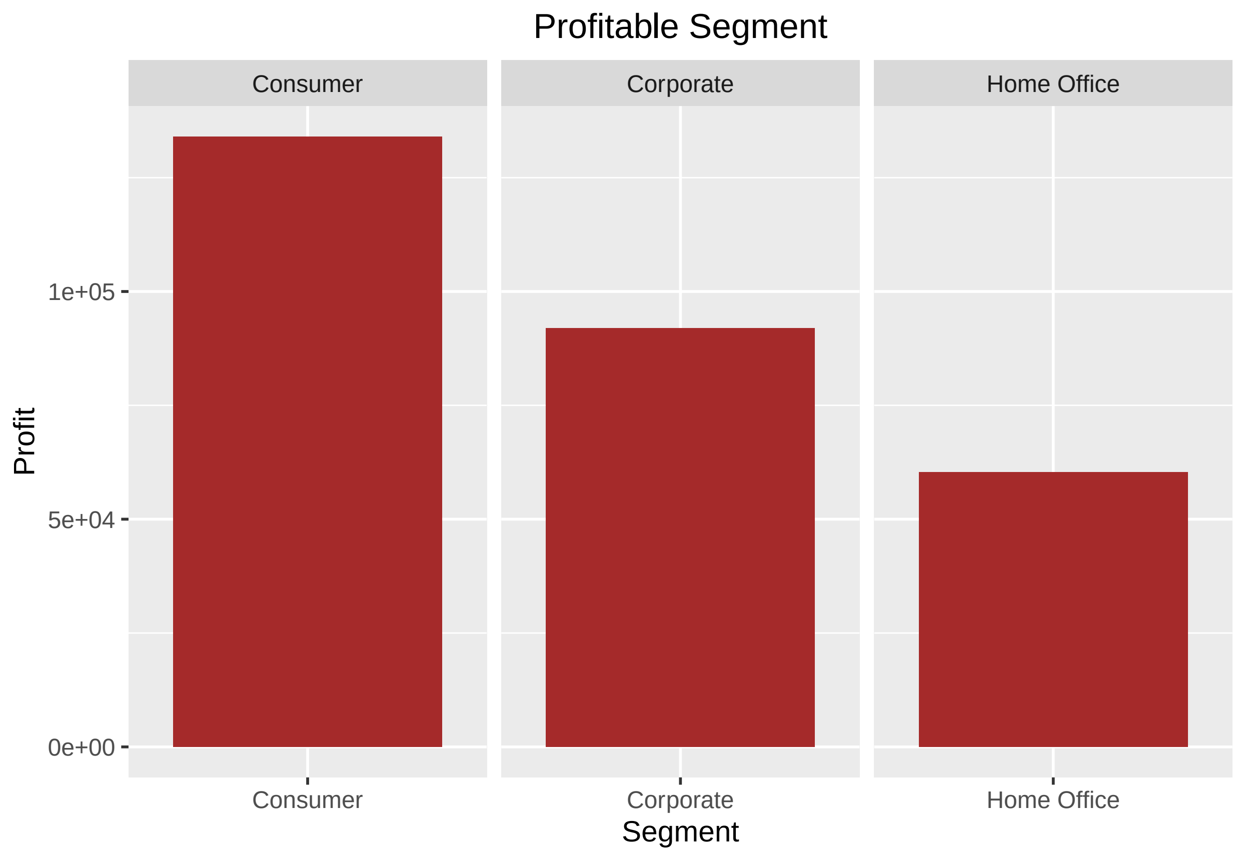


Note :

From the above it is clear that Copies gives the maximum profit to the store .

Profit Associated with Segment

```
profit_Segment <- aggregate(Profit ~ Segment, data = task3, FUN = sum)

ggplot(profit_Segment, aes(x = Segment, y = Profit)) +
  geom_bar(stat = "identity",fill= "brown") +
  facet_wrap(~ Segment, scales = "free_x") +
  labs(x = "Segment", y = "Profit") +
  ggtitle("Profitable Segment") +
  theme(plot.title = element_text(hjust = 0.5))
```

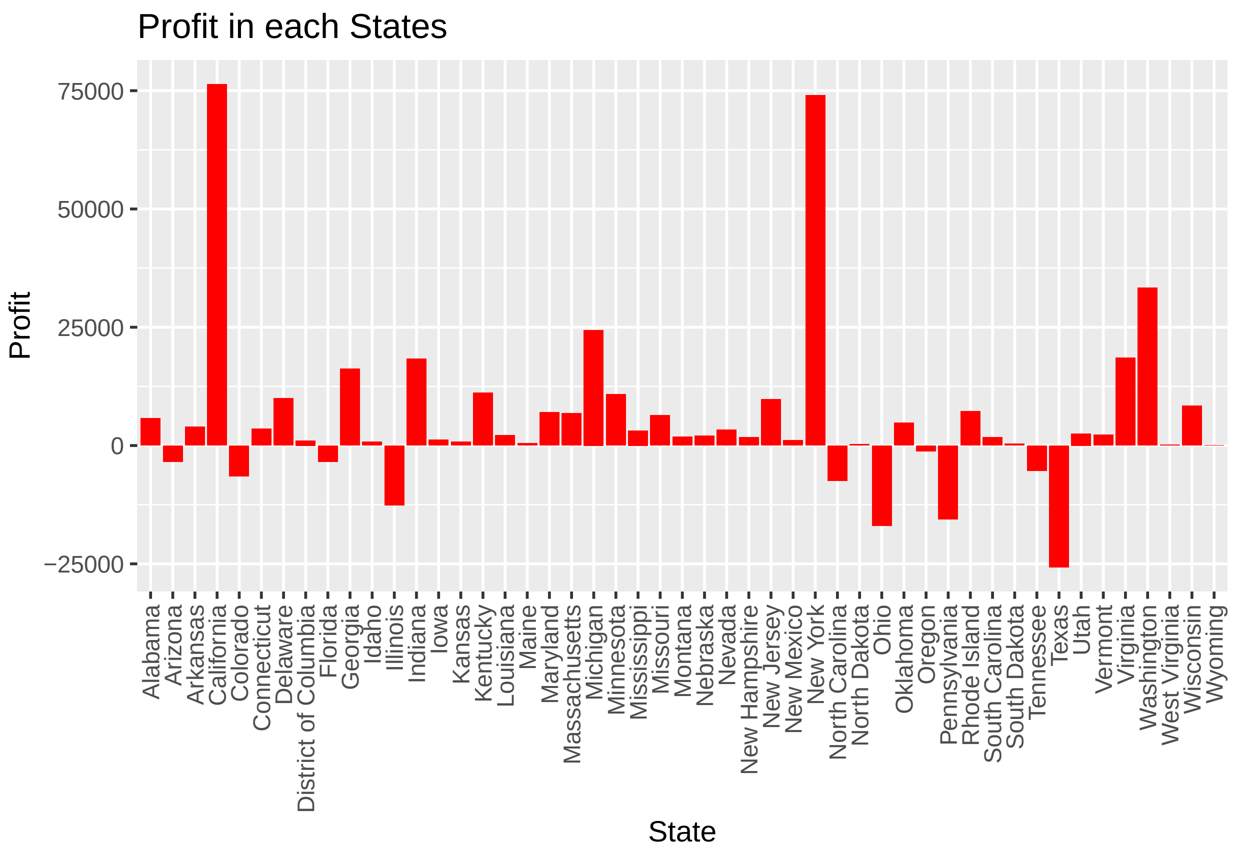


Note:

Consumer give the most profit

Profit in Association with states

```
profit_states <- aggregate(Profit ~ State, data = task3, FUN = sum)

ggplot(profit_states, aes(x = State, y = Profit)) +
  geom_bar(stat = "identity",fill="red") +
  labs(x = "State", y = "Profit") +
  ggtitle("Profit in each States")  +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

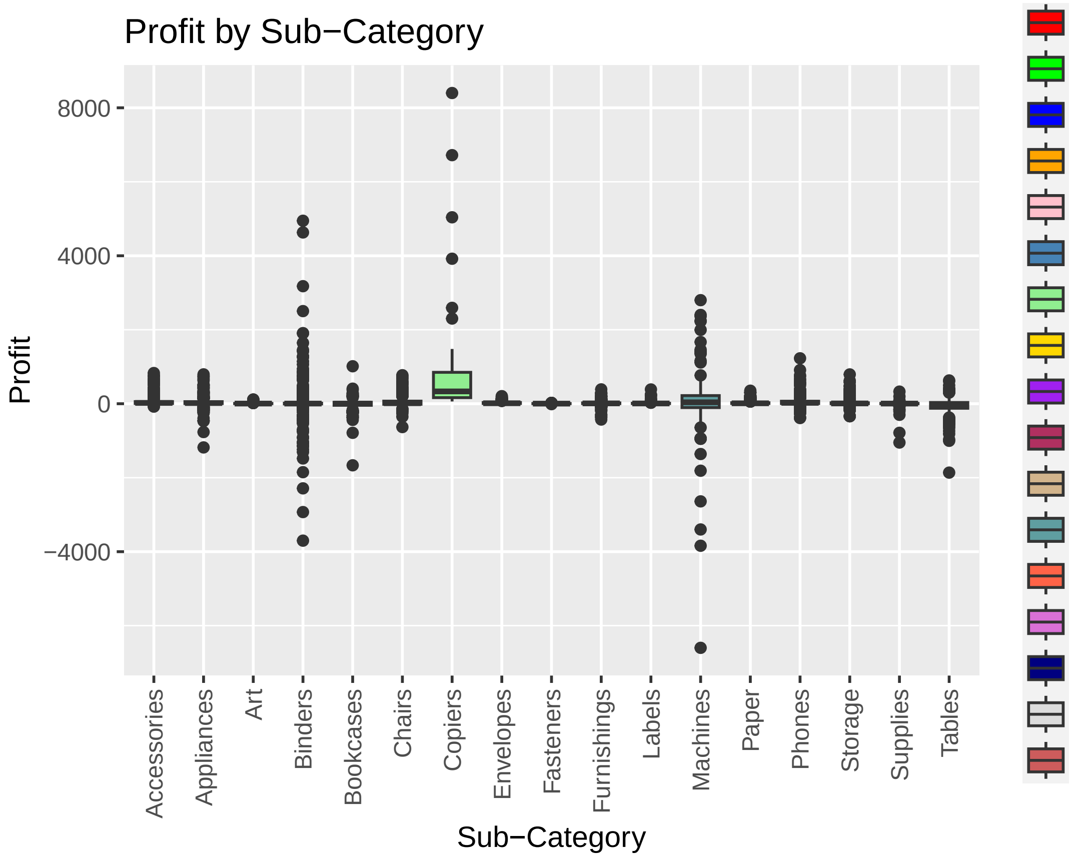


Note:

California gives maximum profit followed by New York.

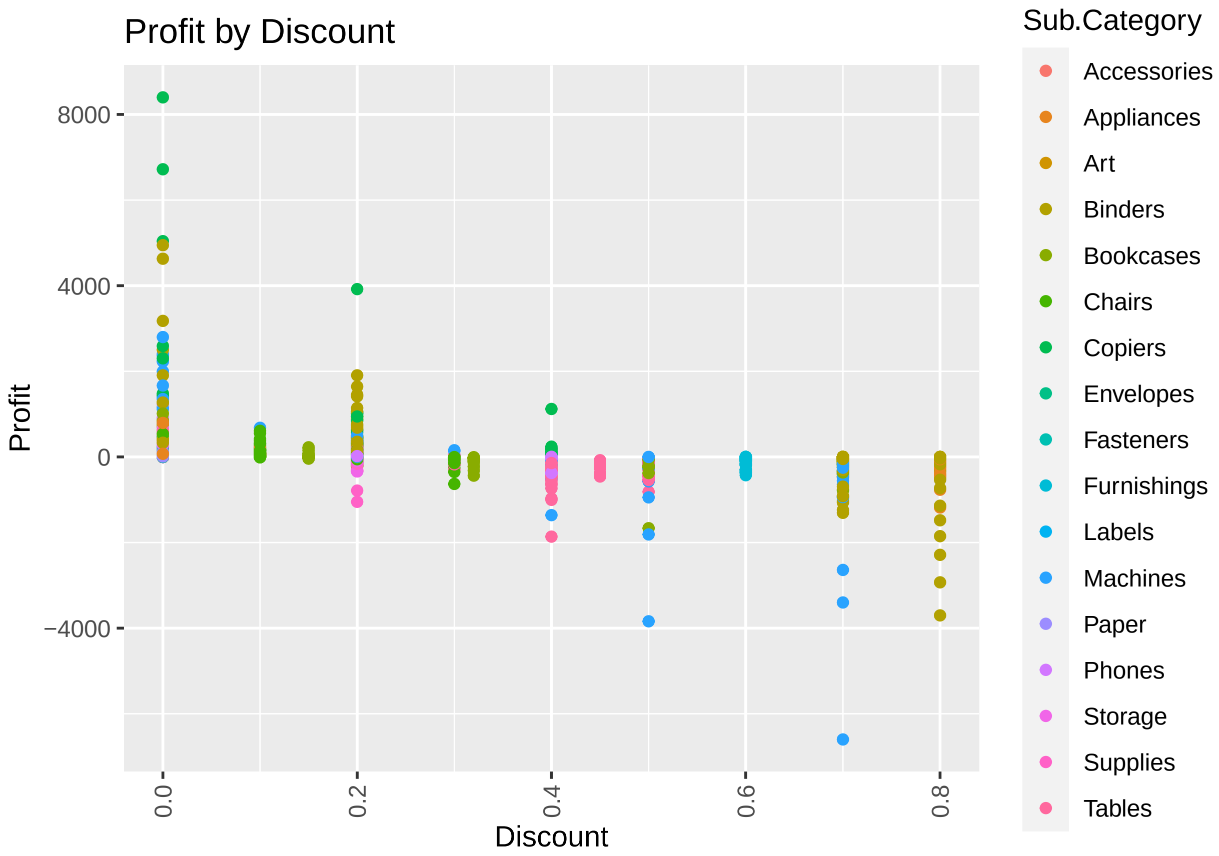
Creating a box plot with different colors for different bars

```
ggplot(task3, aes(x = Sub.Category, y = Profit, fill = Sub.Category)) +
  geom_boxplot() +
  scale_fill_manual(values = c("red", "green", "blue", "orange","pink","steelblue","lightgreen","gold",
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Profit by Sub-Category") +
  xlab("Sub-Category") +
  ylab("Profit")
```



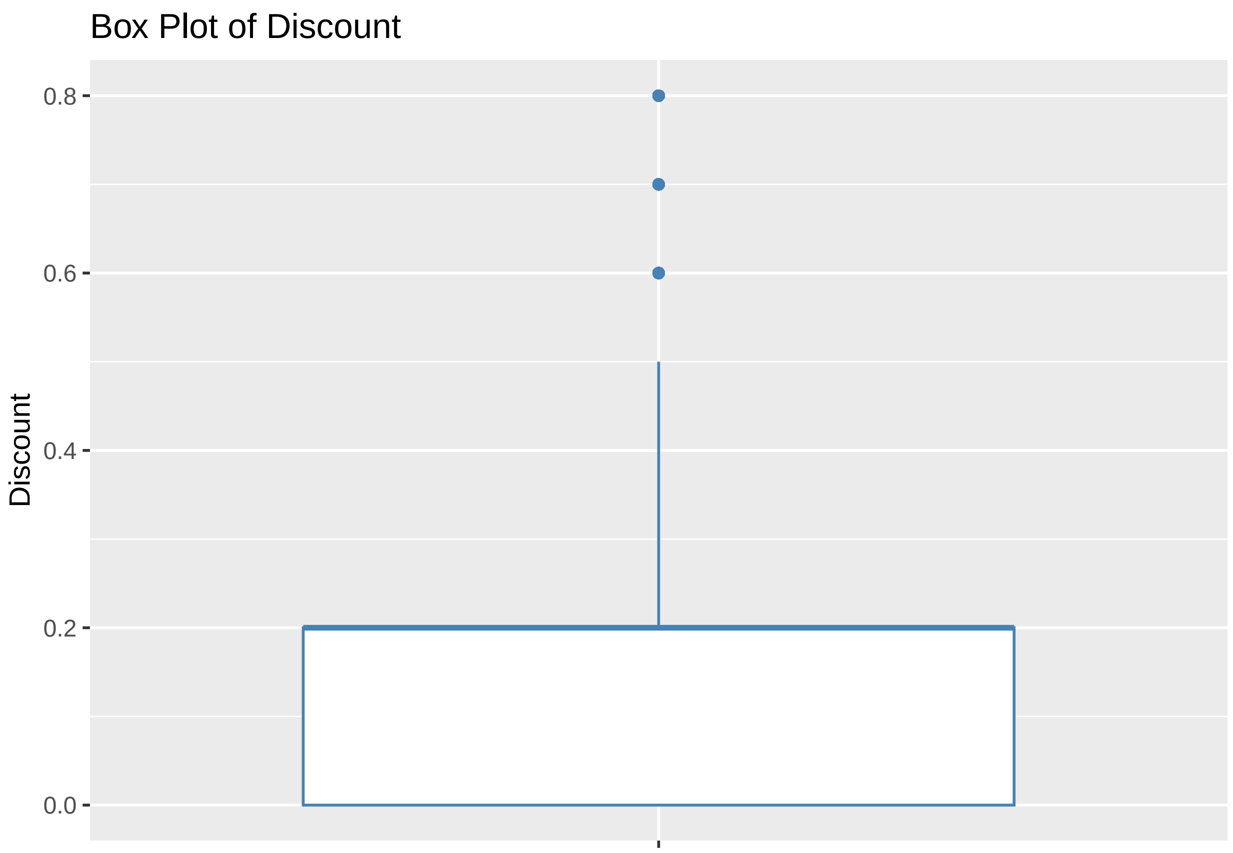
Creating a point plot with different colors for different lines

```
ggplot(task3, aes(x = Discount, y = Profit, color = Sub.Category)) +
  geom_point() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  ggtitle("Profit by Discount") +
  xlab("Discount") +
  ylab("Profit")
```



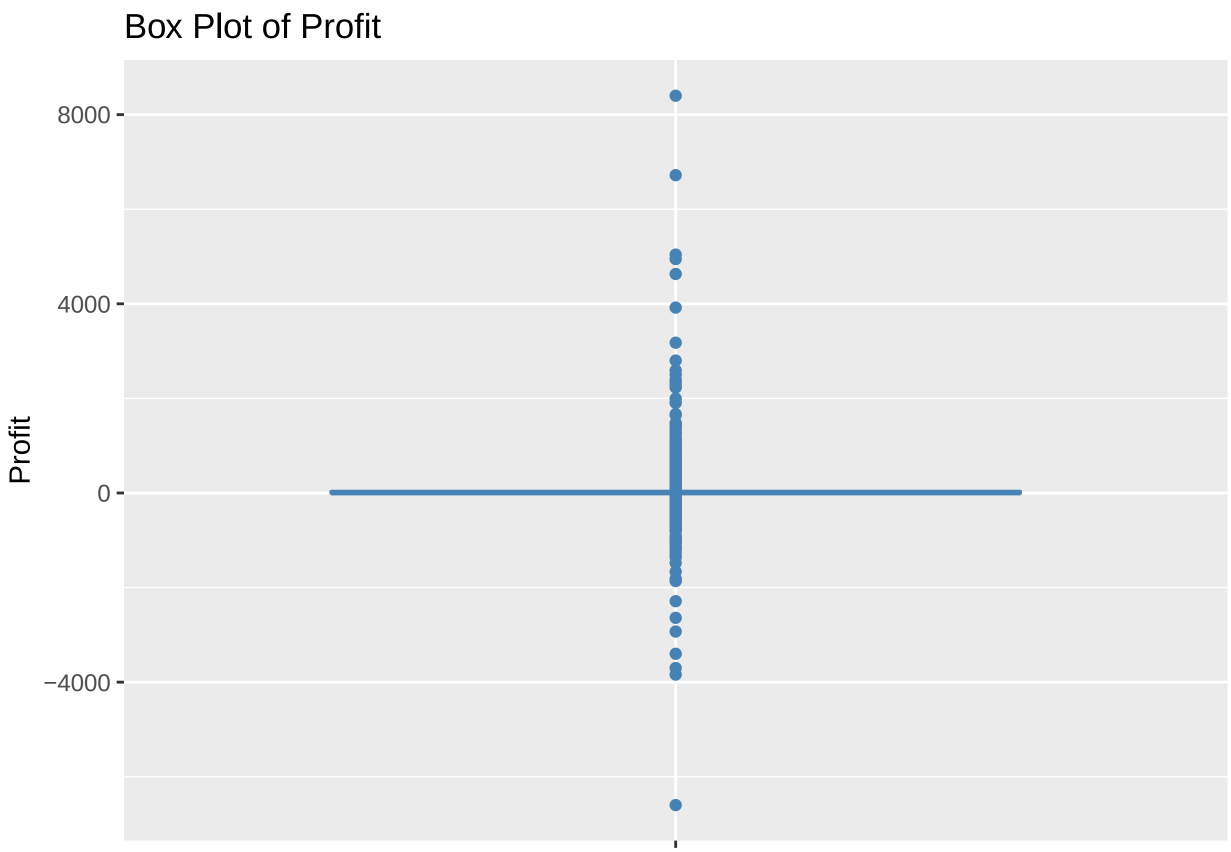
Boxplot of Discount

```
ggplot(task3, aes(x = "", y = Discount)) +
  geom_boxplot(color="steelblue") +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank()) +
  ggtitle("Box Plot of Discount") +
  ylab("Discount")
```



Boxplot of profit

```
ggplot(task3, aes(x = "", y = Profit)) +
  geom_boxplot(color="steelblue") +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank()) +
  ggtitle("Box Plot of Profit") +
  ylab("Profit")
```

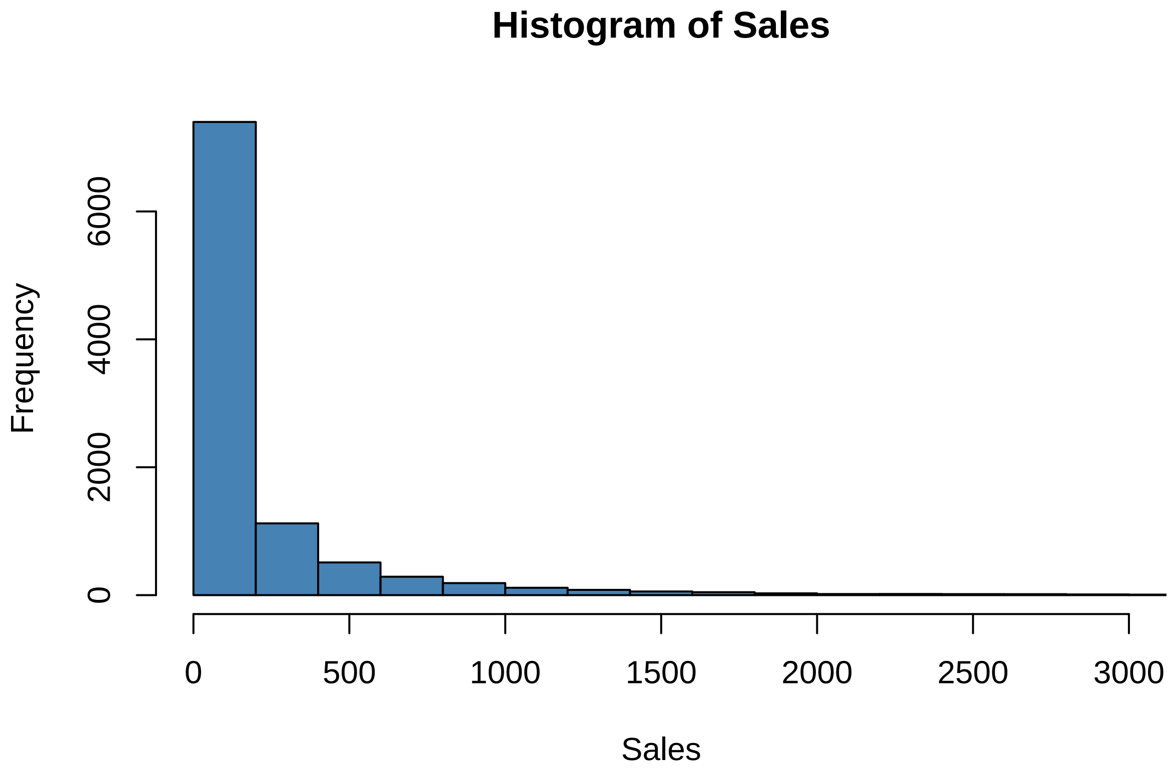


Distribution Plot

```
summary(task3$Sales)
```

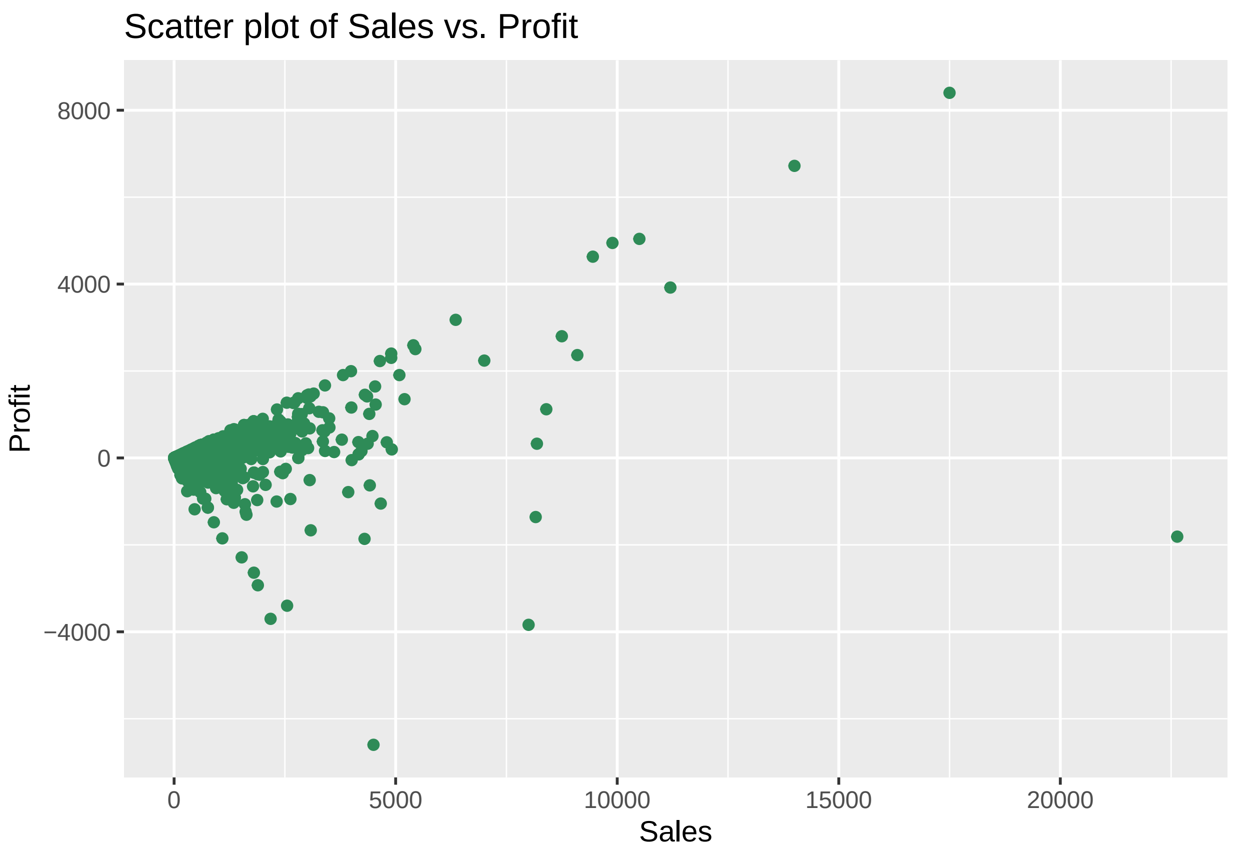
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     0.444    17.300    54.816   230.149   209.970 22638.480
```

```
# Creating a histogram of Sales
hist(task3$Sales, col = "steelblue", breaks = 100, main = "Histogram of Sales", xlab = "Sales",
     ylab = "Frequency",xlim = c(0,3000))
```



Scatter plot of sales

```
ggplot(task3, aes(x = Sales, y = Profit)) +
  geom_point(color="seagreen") +
  labs(x = "Sales", y = "Profit") +
  ggtitle("Scatter plot of Sales vs. Profit")
```

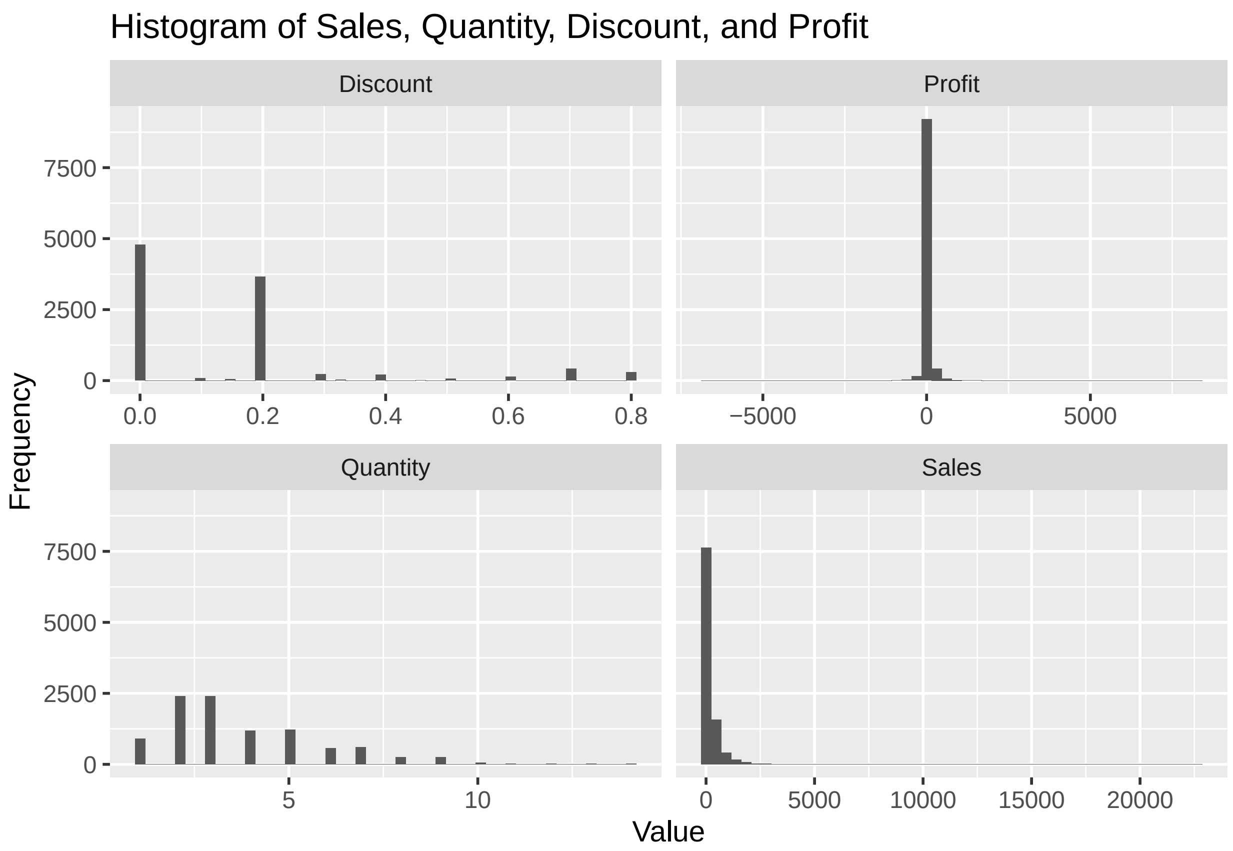


Histogram of all quantitative variables

```
library(tidyr)
```

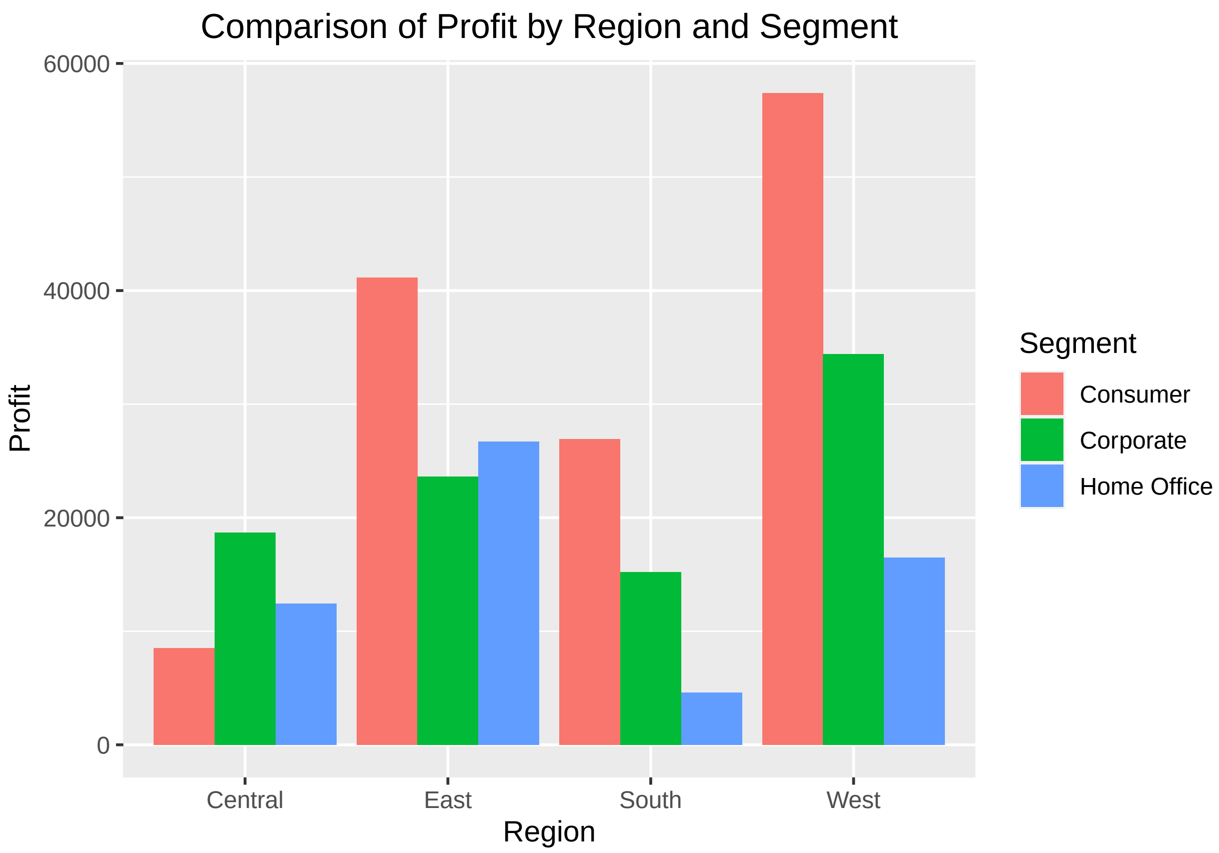
```
## Warning: package 'tidyr' was built under R version 4.1.3

##
## Attaching package: 'tidyr'

## The following object is masked from 'package:reshape2':
##
##     smiths

## The following object is masked from 'package:magrittr':
##
##     extract
```

```
h <- gather(task3, key = "variable", value = "value", Sales:Profit)

ggplot(h, aes(x = value)) +
  geom_histogram(bins = 50) +
  facet_wrap(~ variable, scales = "free_x") +
  labs(x = "Value", y = "Frequency") +
  ggtitle("Histogram of Sales, Quantity, Discount, and Profit")
```



Segment*Region wise profit

```r
plot1 <- aggregate(Profit ~ Region + Segment, data = task3, FUN = sum)

ggplot(plot1, aes(x = Region, y = Profit, fill = Segment)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Region", y = "Profit", title = "Comparison of Profit by Region and Segment")  +
  theme(plot.title = element_text(hjust = 0.5))
```

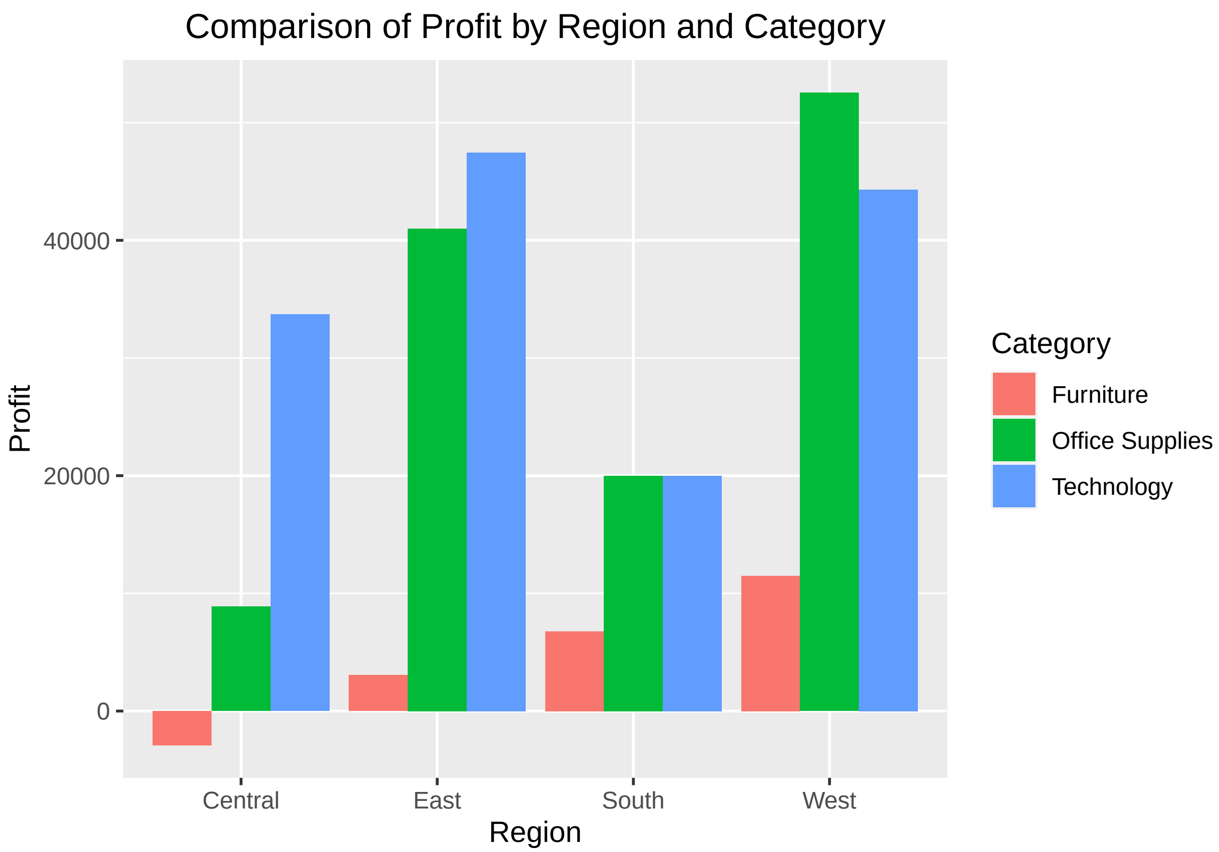


Note :

Central Region has Less number of consumers

Profit associated with different categories

```
plot2 <- aggregate(Profit ~ Region + Category, data = task3, FUN = sum)

ggplot(plot2, aes(x = Region, y = Profit, fill = Category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Region", y = "Profit", title = "Comparison of Profit by Region and Category") +
  theme(plot.title = element_text(hjust = 0.5))
```



Note:

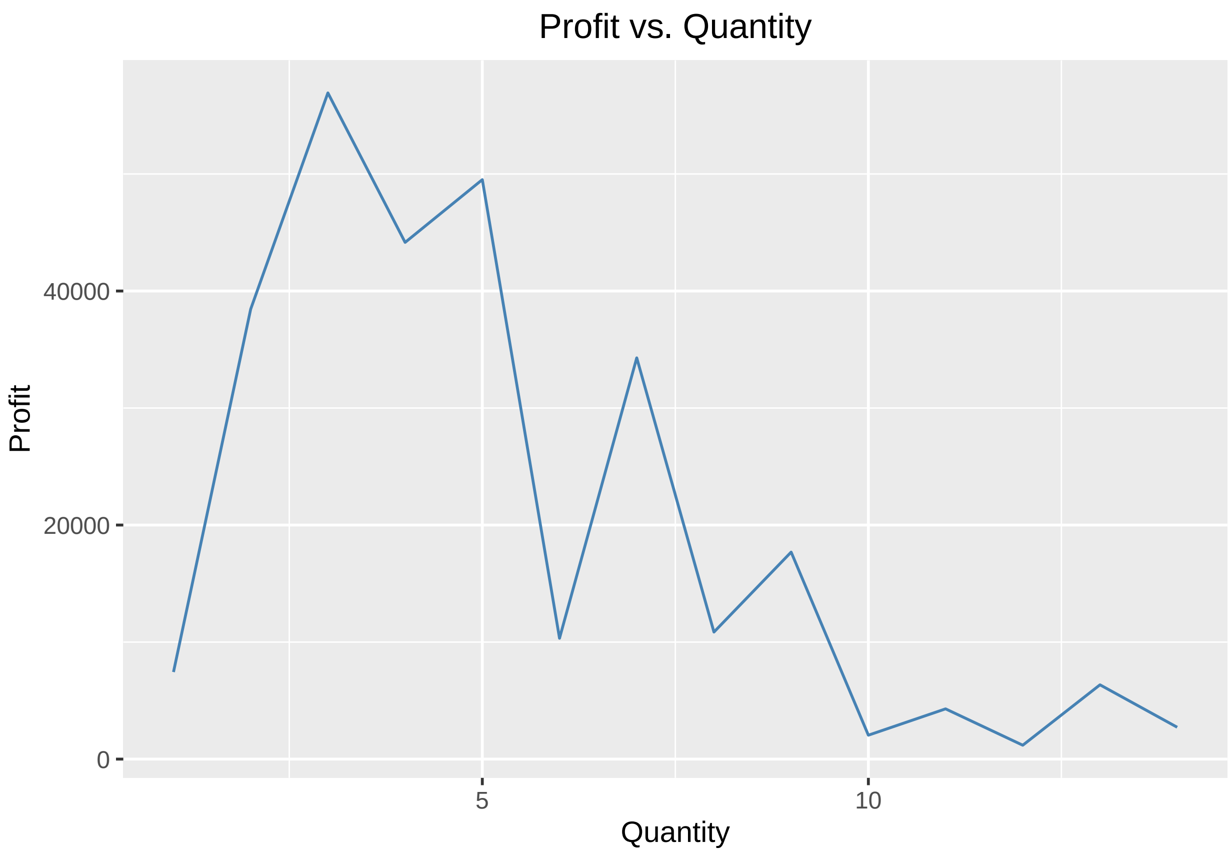
1.Sale of furniture is significantly low in Central and Eastern Regions.

2.There is very low office supply in Central Region.

Lineplots

1.Profit vs Quantity

```
plot3 <- aggregate(Profit ~ Quantity, data = task3, FUN = sum)

ggplot(plot3, aes(x = Quantity, y = Profit)) +
  geom_line(color = "steelblue") +
  labs(x = "Quantity", y = "Profit", title = "Profit vs. Quantity") +
  theme(plot.title = element_text(hjust = 0.5))
```

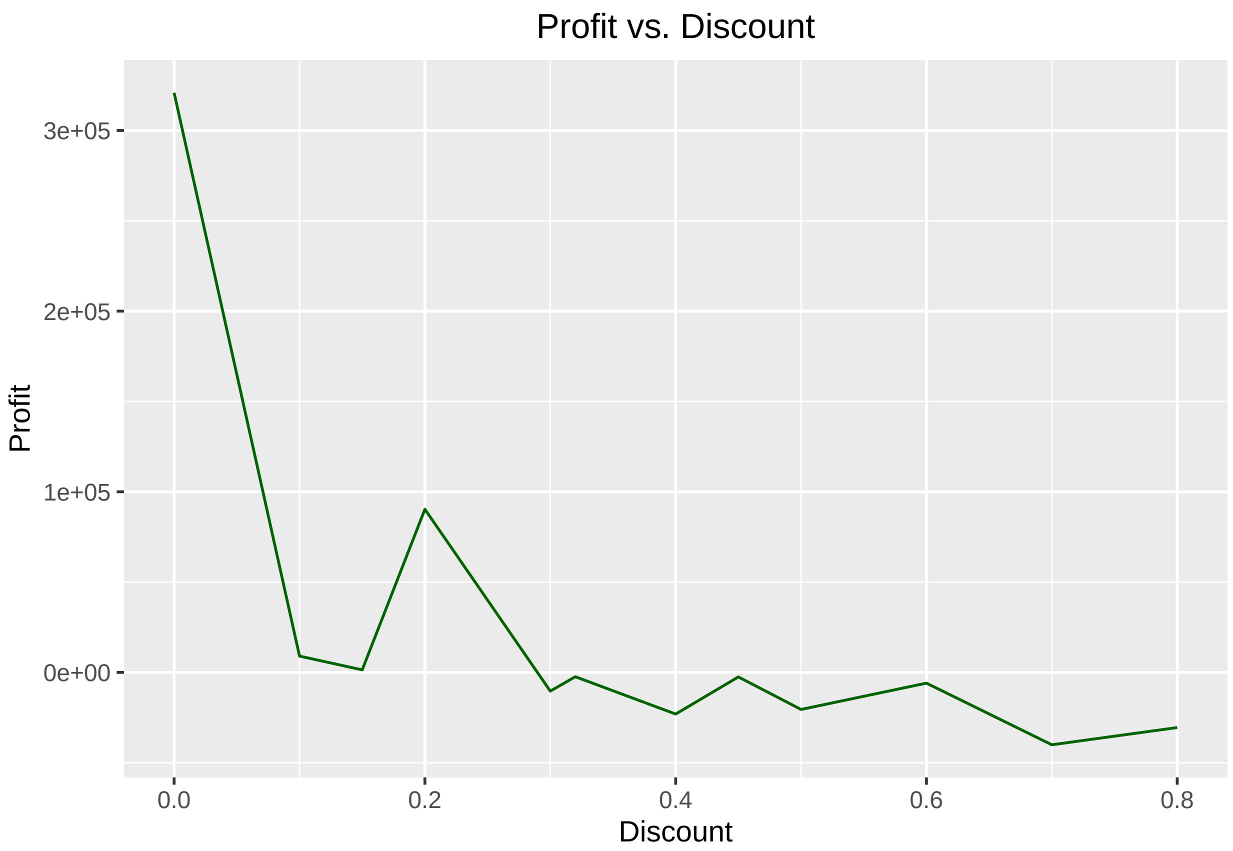


Note:

There is a Constant increment in Profit with the increase in Quantity

2. Profit vs Discount

```r
plot4 <- aggregate(Profit ~ Discount, data = task3, FUN = sum)

ggplot(plot4, aes(x = Discount, y = Profit)) +
  geom_line(color = "darkgreen") +
  labs(x = "Discount", y = "Profit", title = "Profit vs. Discount") +
  theme(plot.title = element_text(hjust = 0.5))
```



Conclusion

1. The superstore loses money when it offers discounts.
2. However, they will lose out on sales and be unable to draw in new, loyal clients if they cease offering discounts.
3. The shop offers discounts around holidays, end-of-season sales, and clearance sales in order to clear up room in their warehouses for new inventory.
4. The business benefits in the long run by gaining more devoted clients by taking on little losses.
5. A crucial aspect of the operation of the corporation is the little losses from discounts.