Capstone Project

2023-06-29

#This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com. When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(tidyverse)
library(skimr)

df =read.csv("C:/Users/vamsh/Downloads/superstore_dataset2011-2015.csv")
head(df)

##   Row.ID        Order.ID Order.Date Ship.Date      Ship.Mode Customer.ID
## 1  42433    AG-2011-2040   1/1/2011  6/1/2011 Standard Class    TB-11280
## 2  22253   IN-2011-47883   1/1/2011  8/1/2011 Standard Class    JH-15985
## 3  48883    HU-2011-1220   1/1/2011  5/1/2011   Second Class      AT-735
## 4  11731 IT-2011-3647632   1/1/2011  5/1/2011   Second Class    EM-14140
## 5  22255   IN-2011-47883   1/1/2011  8/1/2011 Standard Class    JH-15985
## 6  22254   IN-2011-47883   1/1/2011  8/1/2011 Standard Class    JH-15985
##     Customer.Name     Segment        City           State   Country
Postal.Code
## 1 Toby Braunhardt    Consumer Constantine     Constantine   Algeria
NA
## 2     Joseph Holt    Consumer Wagga Wagga New South Wales Australia
NA
## 3   Annie Thurman    Consumer    Budapest        Budapest   Hungary
NA
## 4    Eugene Moren Home Office   Stockholm       Stockholm    Sweden
NA
## 5     Joseph Holt    Consumer Wagga Wagga New South Wales Australia
NA
## 6     Joseph Holt    Consumer Wagga Wagga New South Wales Australia
NA
##   Market  Region       Product.ID        Category Sub.Category
## 1 Africa  Africa OFF-TEN-10000025 Office Supplies      Storage
## 2   APAC Oceania  OFF-SU-10000618 Office Supplies     Supplies
## 3   EMEA    EMEA OFF-TEN-10001585 Office Supplies      Storage
## 4     EU   North  OFF-PA-10001492 Office Supplies        Paper
## 5   APAC Oceania  FUR-FU-10003447       Furniture  Furnishings
## 6   APAC Oceania  OFF-PA-10001968 Office Supplies        Paper
##                               Product.Name   Sales Quantity Discount
Profit
## 1                     Tenex Lockers, Blue 408.300        2      0.0
106.140
```

```
## 2                   Acme Trimmer, High Speed 120.366        3      0.1
36.036
## 3                    Tenex Box, Single Width  66.120        4      0.0
29.640
## 4                 Enermax Note Cards, Premium  44.865        3      0.5 -
26.055
## 5                  Eldon Light Bulb, Duo Pack 113.670        5      0.1
37.770
## 6 Eaton Computer Printout Paper, 8.5 x 11  55.242        2      0.1
15.342
##   Shipping.Cost Order.Priority
## 1         35.46         Medium
## 2          9.72         Medium
## 3          8.17           High
## 4          4.82           High
## 5          4.70         Medium
## 6          1.80         Medium
```

shape of the data

```
dim(df)

## [1] 51290    24
```

Structure of the data

```
str(df)

## 'data.frame':    51290 obs. of  24 variables:
##  $ Row.ID        : int  42433 22253 48883 11731 22255 22254 21613 34662
44508 23688 ...
##  $ Order.ID      : chr  "AG-2011-2040" "IN-2011-47883" "HU-2011-1220" "IT-
2011-3647632" ...
##  $ Order.Date    : chr  "1/1/2011" "1/1/2011" "1/1/2011" "1/1/2011" ...
##  $ Ship.Date     : chr  "6/1/2011" "8/1/2011" "5/1/2011" "5/1/2011" ...
##  $ Ship.Mode     : chr  "Standard Class" "Standard Class" "Second Class"
"Second Class" ...
##  $ Customer.ID   : chr  "TB-11280" "JH-15985" "AT-735" "EM-14140" ...
##  $ Customer.Name : chr  "Toby Braunhardt" "Joseph Holt" "Annie Thurman"
"Eugene Moren" ...
##  $ Segment       : chr  "Consumer" "Consumer" "Consumer" "Home Office" ...
##  $ City          : chr  "Constantine" "Wagga Wagga" "Budapest" "Stockholm"
...
##  $ State         : chr  "Constantine" "New South Wales" "Budapest"
"Stockholm" ...
##  $ Country       : chr  "Algeria" "Australia" "Hungary" "Sweden" ...
##  $ Postal.Code   : int  NA NA NA NA NA NA NA 92691 NA NA ...
##  $ Market        : chr  "Africa" "APAC" "EMEA" "EU" ...
##  $ Region        : chr  "Africa" "Oceania" "EMEA" "North" ...
##  $ Product.ID    : chr  "OFF-TEN-10000025" "OFF-SU-10000618" "OFF-TEN-
10001585" "OFF-PA-10001492" ...
##  $ Category      : chr  "Office Supplies" "Office Supplies" "Office
```

```
Supplies" "Office Supplies" ...
##  $ Sub.Category  : chr  "Storage" "Supplies" "Storage" "Paper" ...
##  $ Product.Name  : chr  "Tenex Lockers, Blue" "Acme Trimmer, High Speed"
"Tenex Box, Single Width" "Enermax Note Cards, Premium" ...
##  $ Sales         : num  408.3 120.4 66.1 44.9 113.7 ...
##  $ Quantity      : int  2 3 4 3 5 2 2 2 1 3 ...
##  $ Discount      : num  0 0.1 0 0.5 0.1 0.1 0 0.15 0 0 ...
##  $ Profit        : num  106.1 36 29.6 -26.1 37.8 ...
##  $ Shipping.Cost : num  35.46 9.72 8.17 4.82 4.7 ...
##  $ Order.Priority: chr  "Medium" "Medium" "High" "High" ...
```

Missing values

```
colMeans(is.na(df))
```

```
##         Row.ID       Order.ID     Order.Date      Ship.Date      Ship.Mode
##      0.0000000      0.0000000      0.0000000      0.0000000      0.0000000
##    Customer.ID  Customer.Name        Segment           City          State
##      0.0000000      0.0000000      0.0000000      0.0000000      0.0000000
##        Country    Postal.Code         Market         Region     Product.ID
##      0.0000000      0.8051472      0.0000000      0.0000000      0.0000000
##       Category   Sub.Category   Product.Name          Sales       Quantity
##      0.0000000      0.0000000      0.0000000      0.0000000      0.0000000
##       Discount         Profit  Shipping.Cost Order.Priority
##      0.0000000      0.0000000      0.0000000      0.0000000
```

###As the Postal.COde variable has over 80% of the missing values , we will be discarding it.

###Also removing columns like “Row.ID” , “Order.ID” , “Order.Date” “Ship.Date” ,“Customer.ID” , “Customer.Name”,“City” , “State” , “Country”,“Product.ID” and “Product.Name” that wont add any informative insights

```
df = df %>% dplyr::select(-c('Postal.Code',"Row.ID"  , "Order.ID"   ,
"Order.Date",     "Ship.Date" ,"Customer.ID"   ,  "Customer.Name","City"   ,
"State"  ,   "Country","Product.ID"  , "Product.Name"))
```

Descriptive statistics

```
skim(df)
```

Data summary

Name	df
Number of rows	51290
Number of columns	12
_______________________	
Column type frequency:	
character	7
numeric	5

Group variables	None

Variable type: character

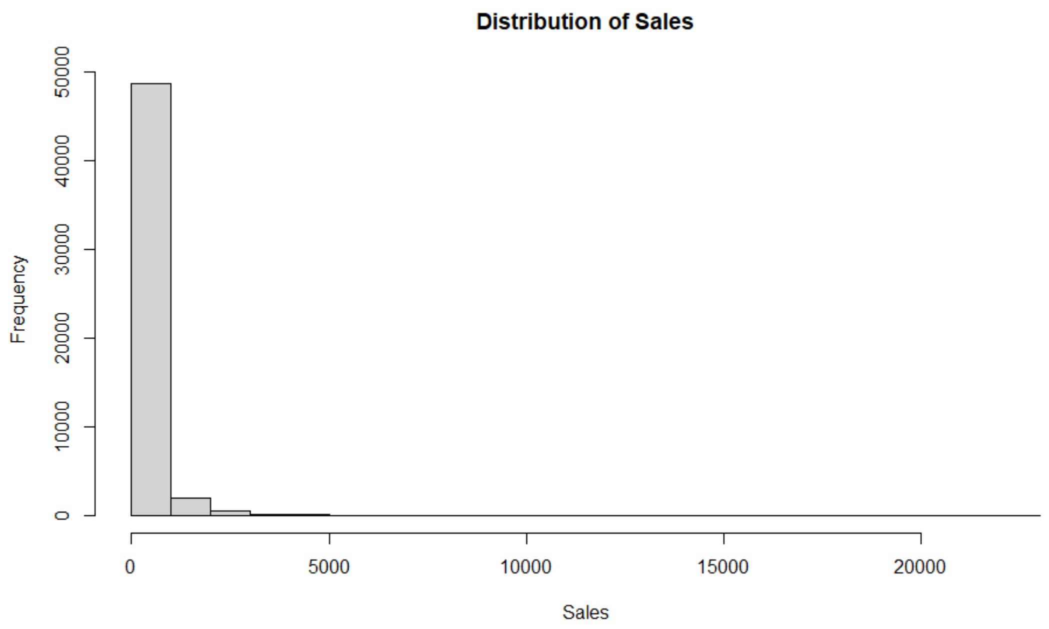
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Ship.Mode	0	1	8	14	0	4	0
Segment	0	1	8	11	0	3	0
Market	0	1	2	6	0	7	0
Region	0	1	4	14	0	13	0
Category	0	1	9	15	0	3	0
Sub.Category	0	1	3	11	0	17	0
Order.Priority	0	1	3	8	0	4	0

Variable type: numeric

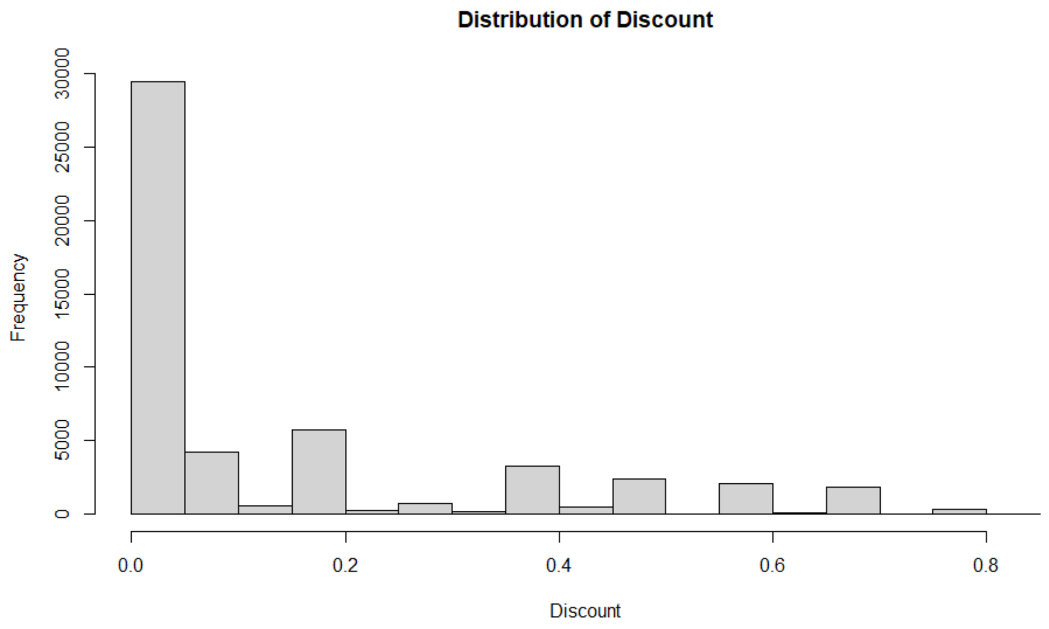
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Sales	0	1	246.49	487.57	0.44	30.76	85.05	251.05	22638.48	▇▁▁▁▁
Quantity	0	1	3.48	2.28	1.00	2.00	3.00	5.00	14.00	▇▃▁▁▁
Discount	0	1	0.14	0.21	0.00	0.00	0.00	0.20	0.85	▇▃▃▁▁
Profit	0	1	28.61	174.34	-6599.98	0.00	9.24	36.81	8399.98	▁▁▇▁▁
Shipping.Cost	0	1	26.38	57.30	0.00	2.61	7.79	24.45	933.57	▇▁▁▁▁

histograms of numeric variables

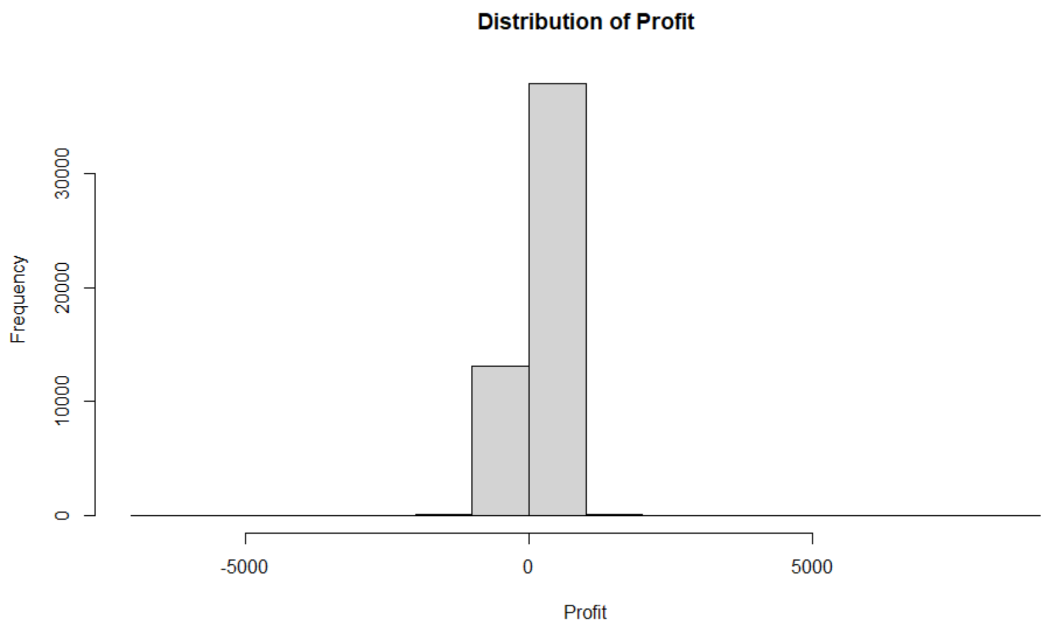
```
hist(df$Sales,main = "Distribution of Sales",xlab = "Sales", ylab =
"Frequency")
```



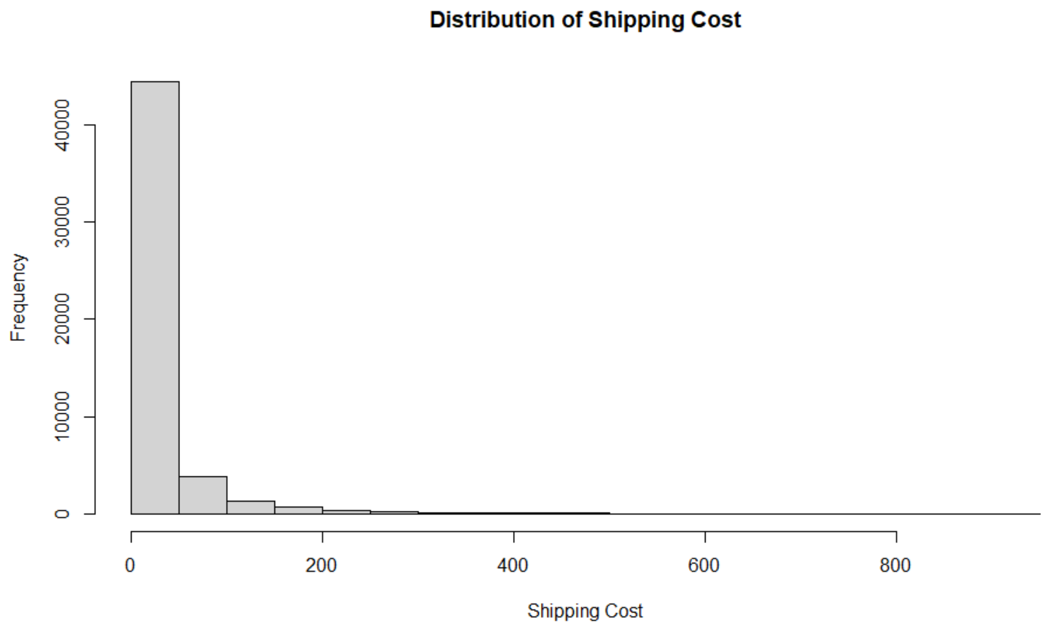
```
hist(df$Discount,main = "Distribution of Discount",xlab = "Discount", ylab =
"Frequency")
```



```
hist(df$Profit,main = "Distribution of Profit",xlab = "Profit", ylab =
"Frequency")
```

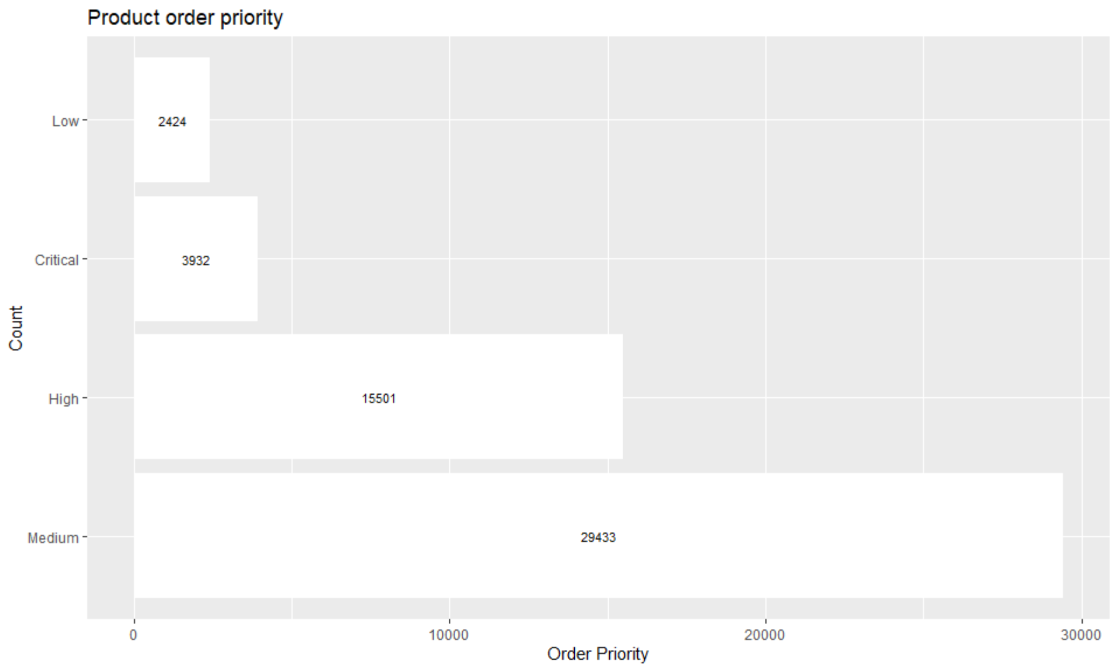


```
hist(df$Shipping.Cost,main = "Distribution of Shipping Cost",xlab = "Shipping
Cost", ylab = "Frequency")
```



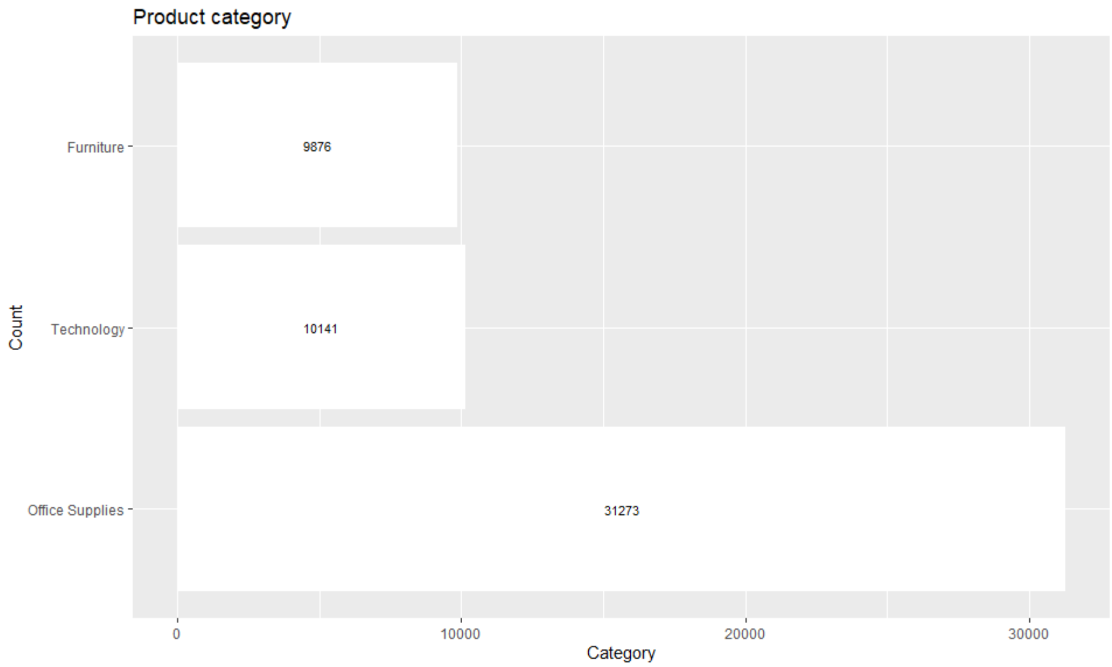
Categorical variable plots

```
df_order = df %>% group_by(Order.Priority) %>% summarise(n=n())

ggplot(df_order, aes(y = reorder(Order.Priority, -n), x = n,label =n)) +
  geom_bar(stat = "identity", fill = "white") +
  xlab("Order Priority") +
  ylab("Count") +
  ggtitle("Product order priority")+geom_text( size = 3,position =
position_stack(vjust = 0.5)) + theme(
    text = element_text(color = "black"),
    plot.title = element_text(color = "black")
  )
```



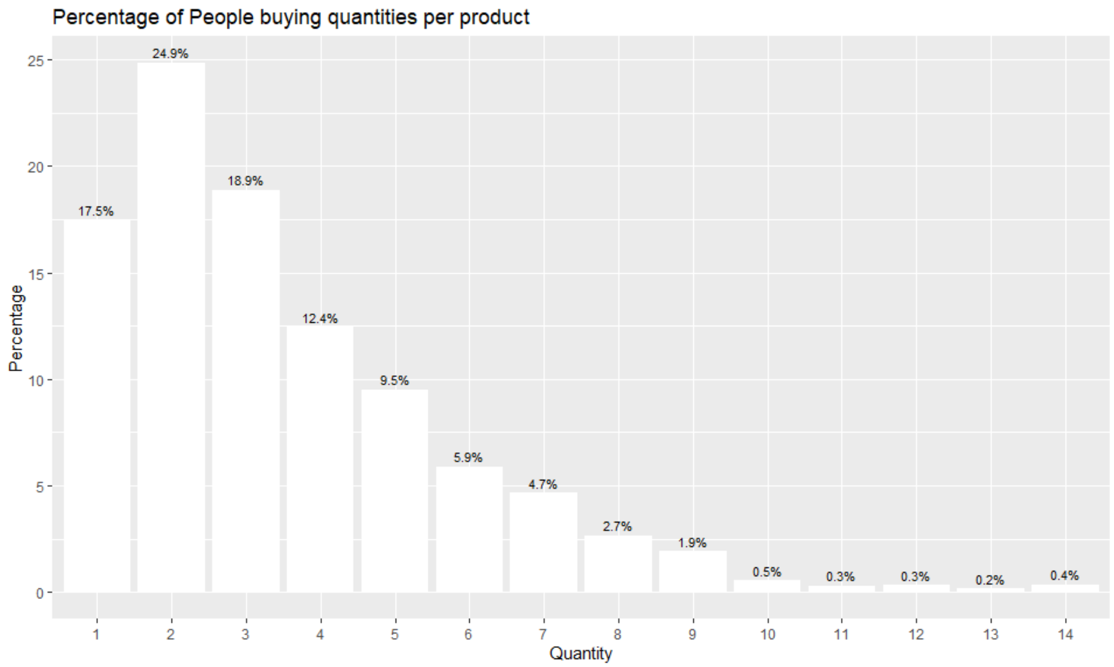
## From the plot we see that the maximum priroity set for the products to buy is Medium

```
df_category = df %>% group_by(Category) %>% summarise(n=n())

ggplot(df_category, aes(y = reorder(Category, -n), x = n,label =n)) +
  geom_bar(stat = "identity", fill = "white") +
  xlab("Category") +
  ylab("Count") +
  ggtitle("Product category")+geom_text( size = 3,position =
position_stack(vjust = 0.5)) +theme(
    text = element_text(color = "black"),
    plot.title = element_text(color = "black")
  )
```



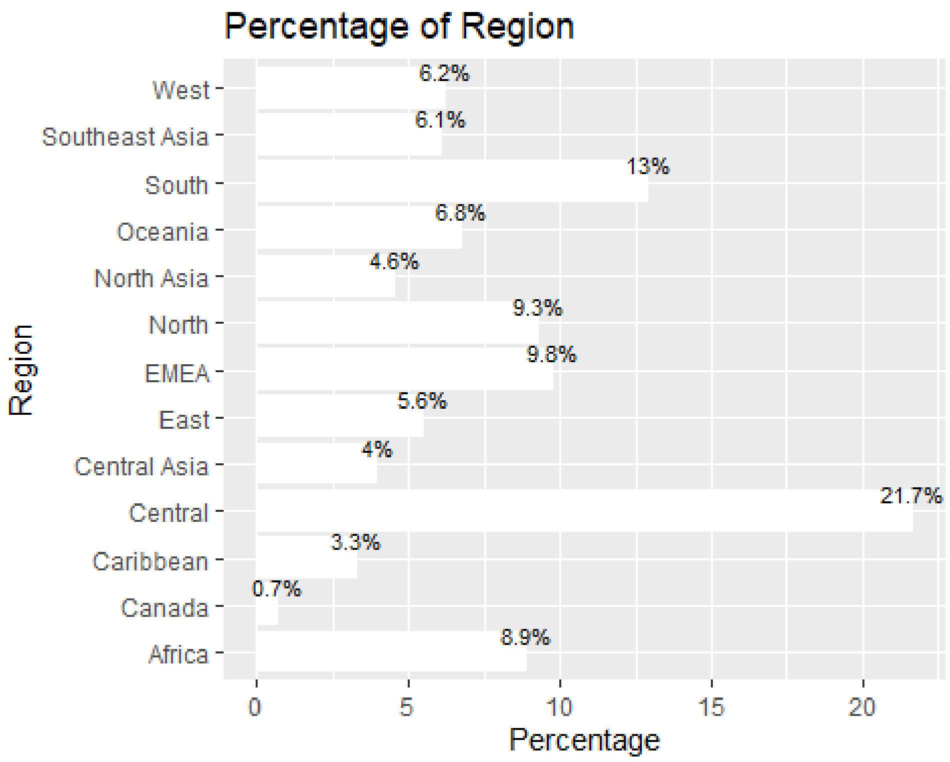
##We see that most of the products brought are Office supplies

```
df$Quantity = as.factor(df$Quantity)

Quantity_percent <- df %>%
  group_by(Quantity) %>%
  summarise(n = n()) %>% mutate(perc = (n / sum(n))* 100)

ggplot(Quantity_percent, aes(x = Quantity, y = perc)) +
  geom_bar(stat = "identity", fill = "white") +
  labs(x = "Quantity", y = "Percentage") +
  ggtitle("Percentage of People buying quantities per
product")+geom_text(aes(label = paste0(round(perc, 1), "%")), vjust = -0.5,
size = 3) + theme(
    text = element_text(color = "black"),
    plot.title = element_text(color = "black")
  )
```



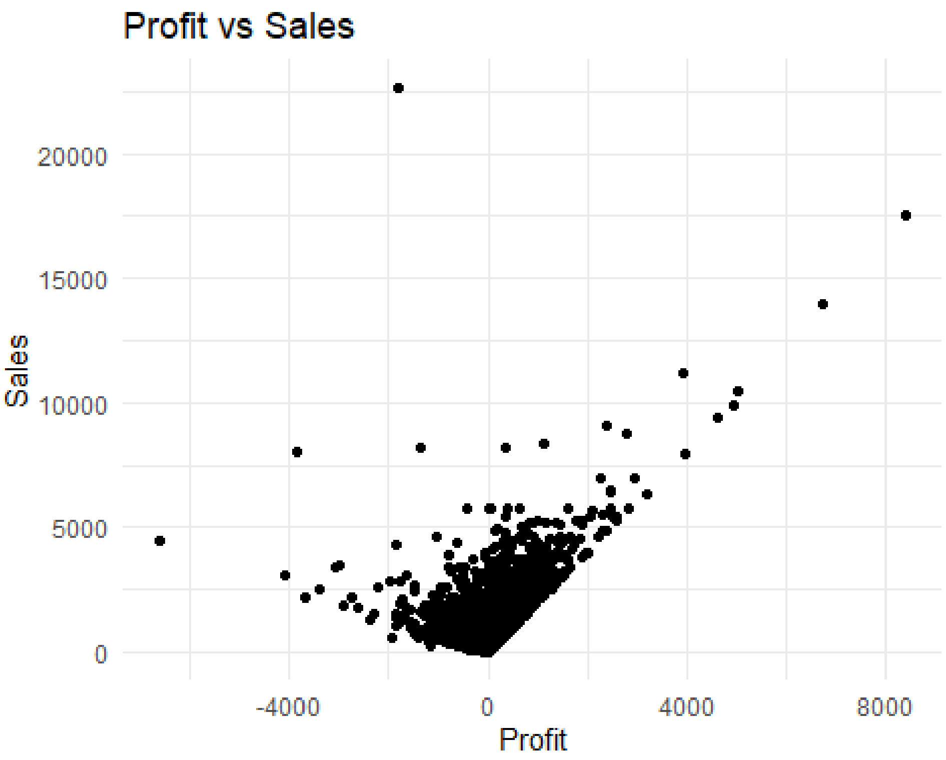
##People prefer buying quantity = 2 the most

```
Region <- df %>%
  group_by(Region) %>%
  summarise(n = n()) %>% mutate(perc = (n / sum(n))* 100)

ggplot(Region, aes(x = Region, y = perc)) +
  geom_bar(stat = "identity", fill = "white") +
  labs(x = "Region", y = "Percentage") +coord_flip()+
  ggtitle("Percentage of Region")+geom_text(aes(label = paste0(round(perc,
1), "%")), vjust = -0.5, size = 3) + theme(
    text = element_text(color = "black"),
    plot.title = element_text(color = "black")
  )
```

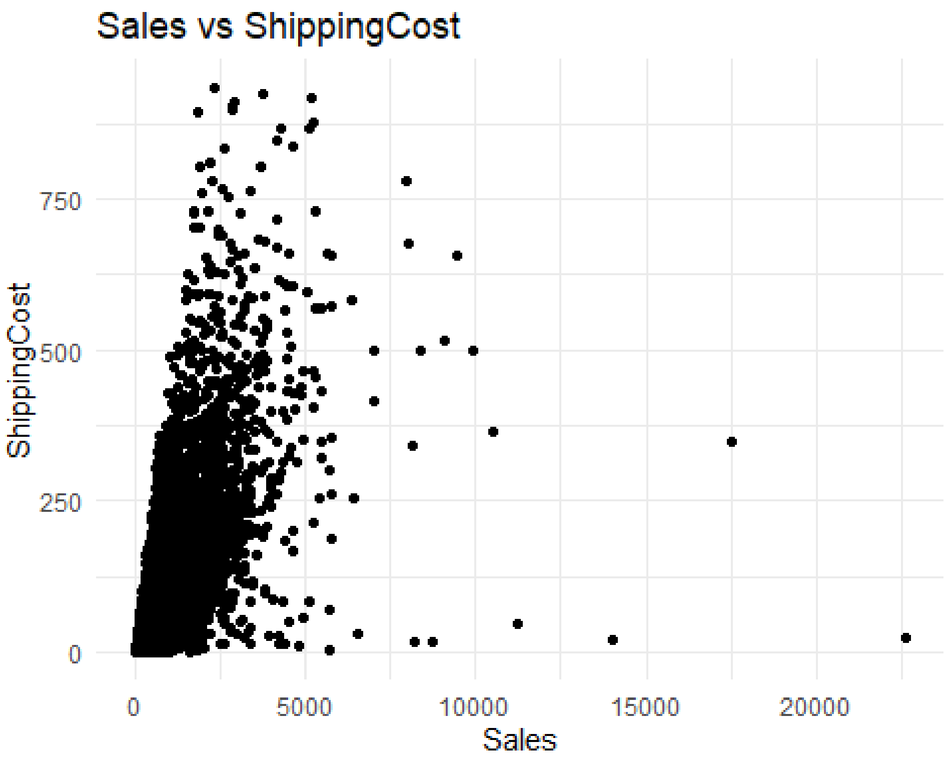


Scatterplots

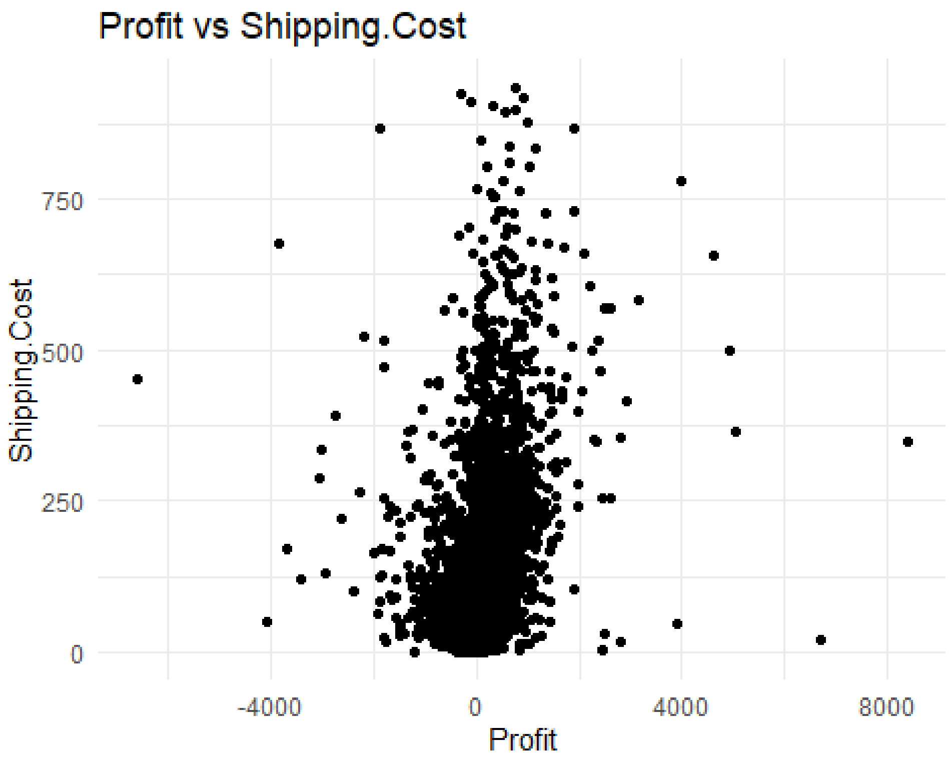
```
ggplot(df, aes(Profit, Sales))+ geom_point()+ xlab("Profit") +
  ylab("Sales") +
  ggtitle("Profit vs Sales") +
  theme_minimal()
```



```
ggplot(df, aes(Sales, Shipping.Cost))+ geom_point()+ xlab("Sales") +
  ylab("ShippingCost") +
  ggtitle("Sales vs ShippingCost") +
  theme_minimal()
```



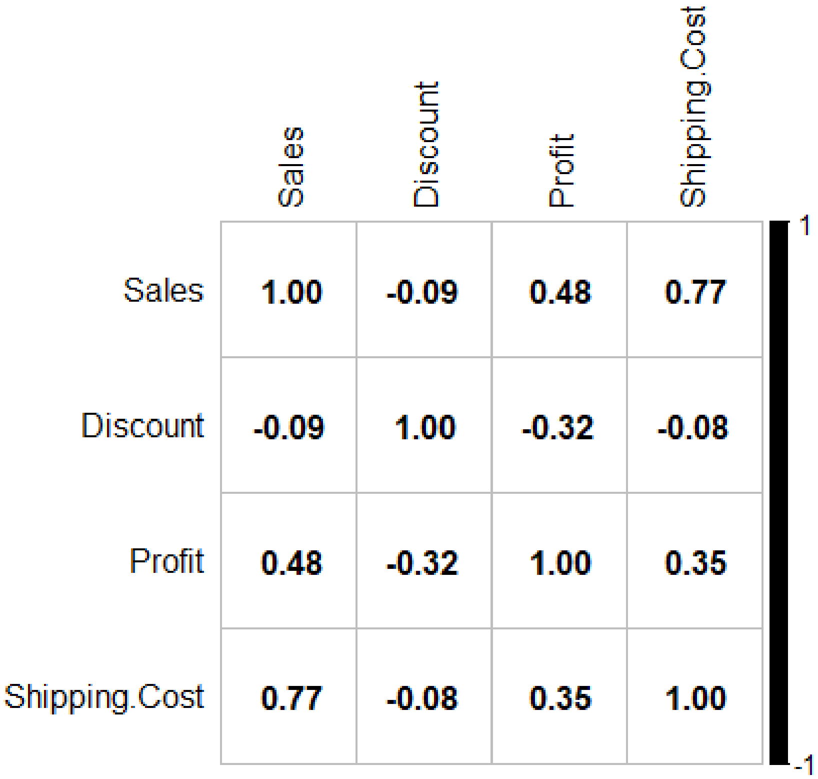
```
ggplot(df, aes(Profit, Shipping.Cost))+ geom_point()+ xlab("Profit") +
  ylab("Shipping.Cost") +
  ggtitle("Profit vs Shipping.Cost") +
  theme_minimal()
```



###The multi-scatterplot ###1. The first plot shows the relationship between the “Profit” and “Sales” variables. ###2. The second plot displays the relationship between the “Sales” and “Shipping.Cost” variables. ###3. The third plot illustrates the relationship between the “Profit” and “Shipping.Cost” variables. ###Each plot includes points representing the data and is formatted with x-axis and y-axis labels, a title, and a minimalistic theme.

Correlation plot

```
library(corrplot)

## corrplot 0.92 loaded

df_new = df[,c(7,9,10,11)]
M<-cor(df_new)
corrplot(M, method="number",col = "black",tl.col = 'black')
```

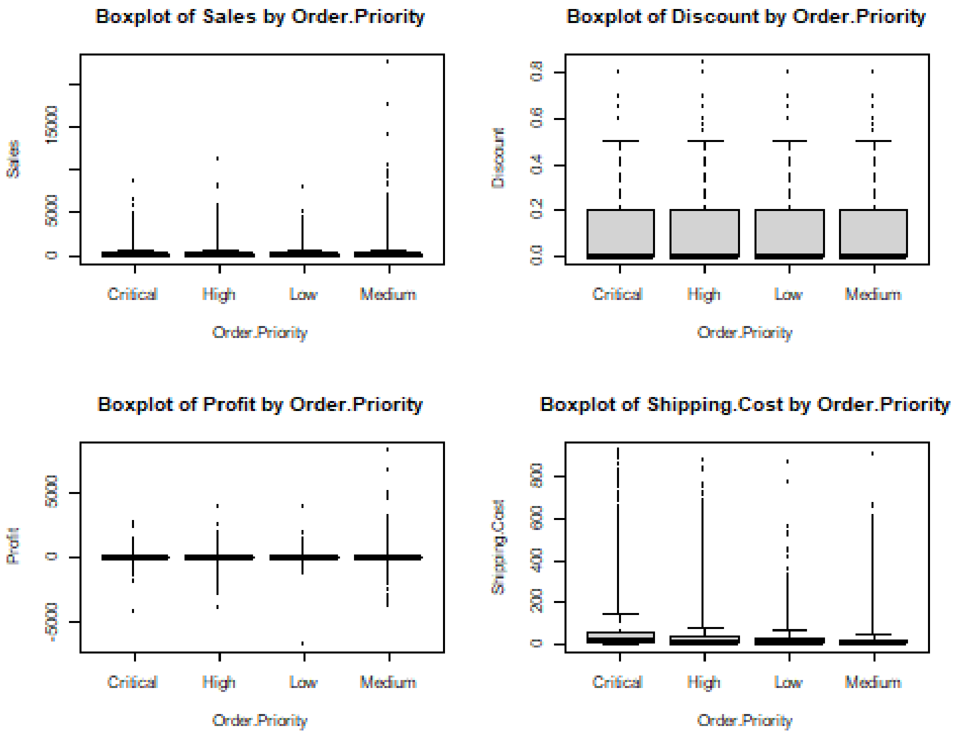


	Sales	Discount	Profit	Shipping.Cost
Sales	1.00	-0.09	0.48	0.77
Discount	-0.09	1.00	-0.32	-0.08
Profit	0.48	-0.32	1.00	0.35
Shipping.Cost	0.77	-0.08	0.35	1.00

###The corrplot function displays correlation values visually, making it easier to understand and identify strong or weak correlations between variables. The corrplot function was used to build a correlation matrix plot for the df_new dataframe, which only includes certain columns. ##Sales has high correlation with Profit and Shipping cost and weak with Discount ##Discount has high correlation with Profit and weak with Shipping cost ##Profit has high correlation with Shipping cost

Outliers

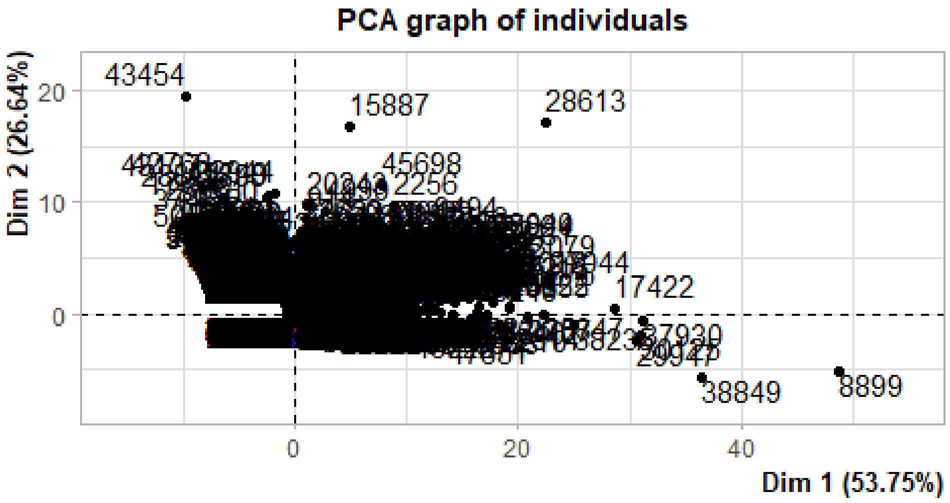
```
my_data <- df[,c(7,9,10,11,12)]
par(mfrow = c(2,2),cex = 0.5)
 for (i in 1:(ncol(my_data) - 1)) {
  boxplot(my_data[, i] ~ my_data[, ncol(my_data)],
          main = paste("Boxplot of", names(my_data)[i], "by",
names(my_data)[ncol(my_data)]),
          xlab = names(my_data)[ncol(my_data)], ylab = names(my_data)[i])
}
```

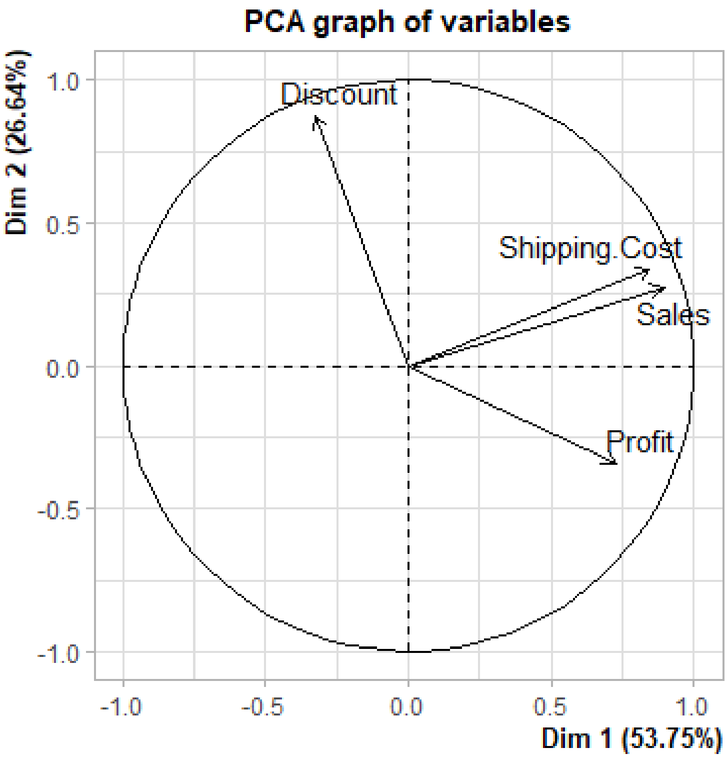


##the box plots offer valuable insights into the distribution of variables within different groups and help identify any potential differences or patterns. Further analysis and statistical tests can be conducted to investigate these findings in more depth.

Principal Component Analysis

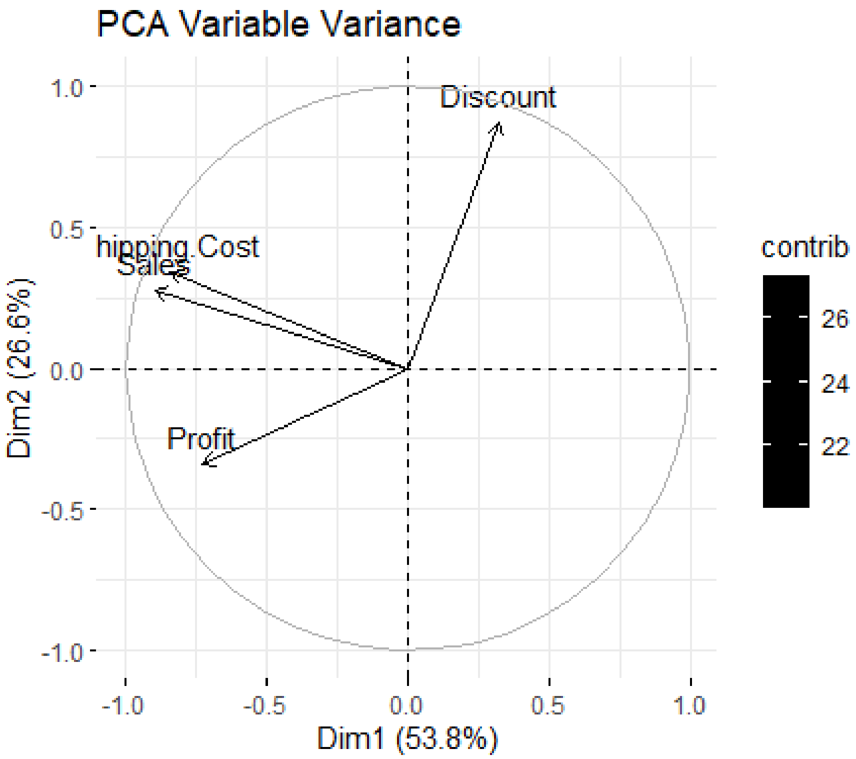
```
library(FactoMineR)
pca <- PCA(df[,c(7,9,10,11)])
```





```
pca <- prcomp(df[,c(7,9,10,11)], scale = TRUE)
# extract loadings
loadings <- pca$rotation
```

```
# print loadings for the first two PCs
print(loadings[, 1:2])

##                     PC1        PC2
## Sales         -0.6105855  0.2673152
## Discount       0.2221577  0.8441832
## Profit        -0.4960958 -0.3303356
## Shipping.Cost -0.5759515  0.3267656

library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

var <- get_pca_var(pca)
fviz_pca_var(pca, col.var="contrib",
gradient.cols = c("black","black","black","black"),ggrepel = TRUE ) + labs(
title = "PCA Variable Variance")
```



#### feature selection using step wise logistic regression as the prediction model is of classification

```
loadings[,c(1:2)]

##                     PC1        PC2
## Sales         -0.6105855  0.2673152
## Discount       0.2221577  0.8441832
## Profit        -0.4960958 -0.3303356
## Shipping.Cost -0.5759515  0.3267656
```

###PC1 (Principal Component 1):

###Sales has a negative coefficient (-0.6106), indicating an inverse relationship with PC1. As PC1 increases, Sales tends to decrease. ###Discount has a positive coefficient (0.2222), suggesting a positive relationship with PC1. As PC1 increases, Discount tends to increase. ###Profit has a negative coefficient (-0.4961), indicating an inverse relationship with PC1. As PC1 increases, Profit tends to decrease. ###Shipping.Cost has a negative coefficient (-0.5760), suggesting an inverse relationship with PC1. As PC1 increases, Shipping.Cost tends to decrease. ###Overall, PC1 can be interpreted as a component that captures the variation in the data related to a decrease in Sales, decrease in Profit, decrease in Shipping.Cost, and increase in Discount.

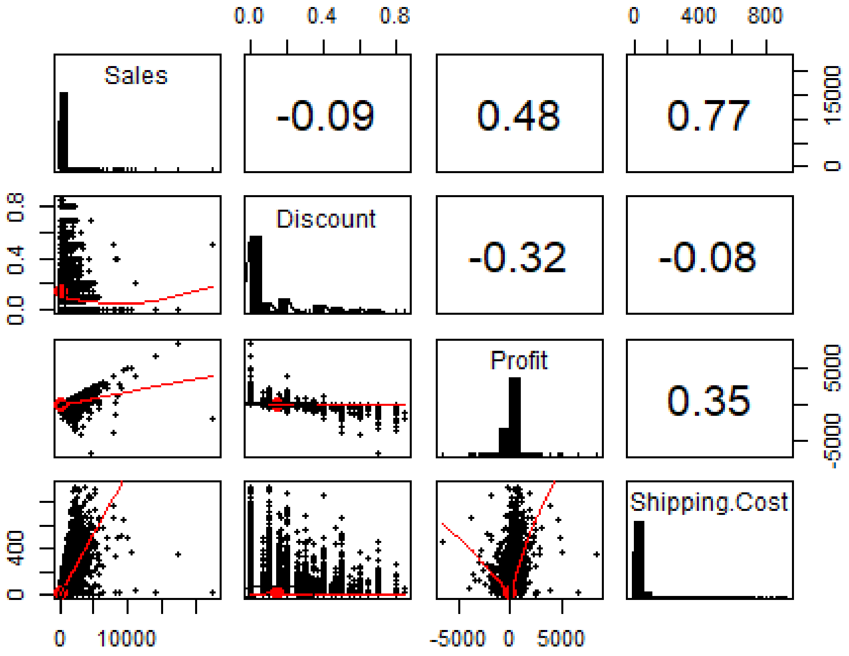
###PC2 (Principal Component 2):

###Sales has a positive coefficient (0.2673), suggesting a positive relationship with PC2. As PC2 increases, Sales tends to increase. ###Discount has a positive coefficient (0.8442), indicating a strong positive relationship with PC2. As PC2 increases, Discount tends to increase. ###Profit has a negative coefficient (-0.3303), indicating an inverse relationship with PC2. As PC2 increases, Profit tends to decrease. ###Shipping.Cost has a positive coefficient (0.3268), suggesting a positive relationship with PC2. As PC2 increases, Shipping.Cost tends to increase.

###PC2 can be interpreted as a component capturing the variation in the data related to an increase in Sales, increase in Discount, decrease in Profit, and increase in Shipping.Cost.

Pairs plot

```
library(psych)

## Warning: package 'psych' was built under R version 4.2.3

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

pairs.panels(df[,c(7,9,10,11)],
             method = "pearson", # correlation method
             hist.col = "black",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
             )
```



###Each cell in the scatterplot matrix will contain a scatterplot of two variables, and the correlation coefficient will be displayed within each cell or represented by an ellipse. The diagonal panels will show histograms or density plots of individual variables.

```
library(fastDummies)

df = df %>% mutate(
   Order.Priority = case_when(
      Order.Priority == "Low" ~ 0,
      Order.Priority == "Medium" ~ 1,
      Order.Priority == "High" ~ 2,
      Order.Priority == "Critical" ~ 3,
      TRUE                        ~ 5
     )
   )


df = dummy_columns(df,select_columns = c("Ship.Mode","Segment","Category","Sub.Category","Quantity"),remove_selected_columns = TRUE)
```

###Converting Categorical variables to numerical variables for the feature selection
###Just be aware that this conversion will change the categorical variables into numerical variables by creating dummy variables

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
Norm_model <- preProcess(df, method = c("center", "scale"))
data_1 <-predict(Norm_model,df)
head(data_1)
```

```
##   Market  Region      Sales   Discount       Profit Shipping.Cost
## 1 Africa  Africa  0.3318723 -0.6732033  0.444697633     0.1585444
## 2   APAC Oceania -0.2586824 -0.2021272  0.042589057    -0.2906954
## 3   EMEA    EMEA -0.3699413 -0.6732033  0.005902328    -0.3177475
## 4     EU   North -0.4135355  1.6821772 -0.313557862    -0.3762150
## 5   APAC Oceania -0.2724160 -0.2021272  0.052535084    -0.3783093
## 6   APAC Oceania -0.3922522 -0.2021272 -0.076109375    -0.4289230
##   Order.Priority Ship.Mode_First Class Ship.Mode_Same Day
## 1     -0.5835974            -0.4140077         -0.2357703
## 2     -0.5835974            -0.4140077         -0.2357703
## 3      0.8457856            -0.4140077         -0.2357703
## 4      0.8457856            -0.4140077         -0.2357703
## 5     -0.5835974            -0.4140077         -0.2357703
## 6     -0.5835974            -0.4140077         -0.2357703
##   Ship.Mode_Second Class Ship.Mode_Standard Class Segment_Consumer
## 1             -0.5015483                0.8164555         0.966509
## 2             -0.5015483                0.8164555         0.966509
## 3              1.9937871               -1.2247827         0.966509
## 4              1.9937871               -1.2247827        -1.034631
## 5             -0.5015483                0.8164555         0.966509
## 6             -0.5015483                0.8164555         0.966509
##   Segment_Corporate Segment_Home Office Category_Furniture
## 1        -0.6559239          -0.4719418         -0.4883292
## 2        -0.6559239          -0.4719418         -0.4883292
## 3        -0.6559239          -0.4719418         -0.4883292
## 4        -0.6559239           2.1188638         -0.4883292
## 5        -0.6559239          -0.4719418          2.0477589
## 6        -0.6559239          -0.4719418         -0.4883292
##   Category_Office Supplies Category_Technology Sub.Category_Accessories
## 1                0.8000378          -0.4964283               -0.2525383
## 2                0.8000378          -0.4964283               -0.2525383
## 3                0.8000378          -0.4964283               -0.2525383
## 4                0.8000378          -0.4964283               -0.2525383
## 5               -1.2499166          -0.4964283               -0.2525383
## 6                0.8000378          -0.4964283               -0.2525383
##   Sub.Category_Appliances Sub.Category_Art Sub.Category_Binders
## 1              -0.1882254        -0.324375           -0.3691754
## 2              -0.1882254        -0.324375           -0.3691754
## 3              -0.1882254        -0.324375           -0.3691754
## 4              -0.1882254        -0.324375           -0.3691754
## 5              -0.1882254        -0.324375           -0.3691754
## 6              -0.1882254        -0.324375           -0.3691754
##   Sub.Category_Bookcases Sub.Category_Chairs Sub.Category_Copiers
## 1             -0.2220922          -0.2678722           -0.2128486
## 2             -0.2220922          -0.2678722           -0.2128486
## 3             -0.2220922          -0.2678722           -0.2128486
## 4             -0.2220922          -0.2678722           -0.2128486
## 5             -0.2220922          -0.2678722           -0.2128486
## 6             -0.2220922          -0.2678722           -0.2128486
##   Sub.Category_Envelopes Sub.Category_Fasteners Sub.Category_Furnishings
## 1             -0.2232496             -0.2225268               -0.2566626
## 2             -0.2232496             -0.2225268               -0.2566626
## 3             -0.2232496             -0.2225268               -0.2566626
## 4             -0.2232496             -0.2225268               -0.2566626
## 5             -0.2232496             -0.2225268                3.8960897
## 6             -0.2232496             -0.2225268               -0.2566626
##   Sub.Category_Labels Sub.Category_Machines Sub.Category_Paper
## 1          -0.2313608            -0.1727321         -0.2721942
## 2          -0.2313608            -0.1727321         -0.2721942
## 3          -0.2313608            -0.1727321         -0.2721942
## 4          -0.2313608            -0.1727321          3.6737757
## 5          -0.2313608            -0.1727321         -0.2721942
## 6          -0.2313608            -0.1727321          3.6737757
##   Sub.Category_Phones Sub.Category_Storage Sub.Category_Supplies
## 1          -0.2646392            3.0229438            -0.2227679
## 2          -0.2646392           -0.3307969             4.4888888
## 3          -0.2646392            3.0229438            -0.2227679
## 4          -0.2646392           -0.3307969            -0.2227679
## 5          -0.2646392           -0.3307969            -0.2227679
## 6          -0.2646392           -0.3307969            -0.2227679
##   Sub.Category_Tables Quantity_1 Quantity_2 Quantity_3 Quantity_4
Quantity_5
## 1          -0.1306644 -0.4601651  1.7387689 -0.4823807  -0.377076 -
0.3243383
## 2          -0.1306644 -0.4601651 -0.5751083  2.0730112  -0.377076 -
0.3243383
## 3          -0.1306644 -0.4601651 -0.5751083 -0.4823807   2.651934 -
0.3243383
## 4          -0.1306644 -0.4601651 -0.5751083  2.0730112  -0.377076 -
0.3243383
## 5          -0.1306644 -0.4601651 -0.5751083 -0.4823807  -0.377076
3.0831404
## 6          -0.1306644 -0.4601651  1.7387689 -0.4823807  -0.377076 -
0.3243383
##   Quantity_6 Quantity_7 Quantity_8 Quantity_9 Quantity_10 Quantity_11
## 1  -0.250127 -0.2208327 -0.1651005  -0.140074 -0.07355389 -0.05523358
## 2  -0.250127 -0.2208327 -0.1651005  -0.140074 -0.07355389 -0.05523358
## 3  -0.250127 -0.2208327 -0.1651005  -0.140074 -0.07355389 -0.05523358
## 4  -0.250127 -0.2208327 -0.1651005  -0.140074 -0.07355389 -0.05523358
## 5  -0.250127 -0.2208327 -0.1651005  -0.140074 -0.07355389 -0.05523358
## 6  -0.250127 -0.2208327 -0.1651005  -0.140074 -0.07355389 -0.05523358
##   Quantity_12 Quantity_13 Quantity_14
## 1 -0.05867893 -0.04025966 -0.06032881
## 2 -0.05867893 -0.04025966 -0.06032881
## 3 -0.05867893 -0.04025966 -0.06032881
## 4 -0.05867893 -0.04025966 -0.06032881
## 5 -0.05867893 -0.04025966 -0.06032881
## 6 -0.05867893 -0.04025966 -0.06032881
```

Linear Regression

```
model <- lm(Order.Priority ~., data = data_1)
summary(model)
```

```
##
## Call:
## lm(formula = Order.Priority ~ ., data = data_1)
##
## Residuals:
##     Min       1Q  Median       3Q      Max
## -4.2824 -0.2662 -0.1688  0.4734  8.4481
##
## Coefficients: (11 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.0031259  0.0134907  -0.232 0.816765
## MarketAPAC                 -0.0405545  0.0208578  -1.944 0.051861 .
## MarketCanada                0.0342339  0.0457768   0.748 0.454557
## MarketEMEA                  0.0016796  0.0175711   0.096 0.923848
## MarketEU                   -0.0235411  0.0265257  -0.887 0.374824
## MarketLATAM                -0.0240369  0.0269306  -0.893 0.372103
## MarketUS                    0.0122094  0.0207732   0.588 0.556702
## RegionCanada                       NA         NA      NA       NA
## RegionCaribbean             0.0769195  0.0310620   2.476 0.013278 *
## RegionCentral              -0.0134435  0.0212771  -0.632 0.527500
## RegionCentral Asia          0.0817373  0.0247439   3.303 0.000956 ***
## RegionEast                  0.0158341  0.0221595   0.715 0.474890
## RegionEMEA                         NA         NA      NA       NA
## RegionNorth                 0.0126545  0.0251069   0.504 0.614245
## RegionNorth Asia            0.0497053  0.0238782   2.082 0.037383 *
## RegionOceania               0.0829017  0.0213277   3.887 0.000102 ***
## RegionSouth                 0.0431701  0.0220360   1.959 0.050110 .
## RegionSoutheast Asia               NA         NA      NA       NA
## RegionWest                         NA         NA      NA       NA
## Sales                      -0.1879275  0.0068872 -27.286  < 2e-16 ***
## Discount                    0.0001682  0.0041708   0.040 0.967837
## Profit                      0.0014717  0.0046771   0.315 0.753025
## Shipping.Cost               0.2695090  0.0061694  43.685  < 2e-16 ***
## `Ship.Mode_First Class`     0.3556766  0.0039982  88.960  < 2e-16 ***
## `Ship.Mode_Same Day`        0.2551700  0.0038960  65.496  < 2e-16 ***
## `Ship.Mode_Second Class`    0.2829203  0.0039495  71.634  < 2e-16 ***
## `Ship.Mode_Standard Class`         NA         NA      NA       NA    
## Segment_Consumer            0.0197793  0.0051713   3.825 0.000131 ***
## Segment_Corporate           0.0002103  0.0051715   0.041 0.967556    
## `Segment_Home Office`              NA         NA      NA       NA    
## Category_Furniture         -0.0320041  0.0132389  -2.417 0.015634 *  
## `Category_Office Supplies`  0.0127444  0.0114027   1.118 0.263715    
## Category_Technology                NA         NA      NA       NA    
## Sub.Category_Accessories    0.0093582  0.0051387   1.821 0.068593 .  
## Sub.Category_Appliances     0.0011437  0.0049901   0.229 0.818725    
## Sub.Category_Art            0.0024128  0.0063056   0.383 0.701982    
## Sub.Category_Binders        0.0046601  0.0067419   0.691 0.489436    
## Sub.Category_Bookcases      0.0144376  0.0073554   1.963 0.049670 *  
## Sub.Category_Chairs         0.0118494  0.0083903   1.412 0.157877    
## Sub.Category_Copiers       -0.0027493  0.0048261  -0.570 0.568894    
## Sub.Category_Envelopes      0.0071462  0.0052474   1.362 0.173254    
## Sub.Category_Fasteners      0.0053249  0.0052402   1.016 0.309556    
## Sub.Category_Furnishings    0.0291234  0.0083088   3.505 0.000457 ***
## Sub.Category_Labels         0.0072230  0.0053326   1.354 0.175587    
## Sub.Category_Machines      -0.0007941  0.0045149  -0.176 0.860378    
## Sub.Category_Paper          0.0104774  0.0057981   1.807 0.070760 .  
## Sub.Category_Phones                NA         NA      NA       NA    
## Sub.Category_Storage        0.0040277  0.0063691   0.632 0.527137    
## Sub.Category_Supplies              NA         NA      NA       NA    
## Sub.Category_Tables                NA         NA      NA       NA    
## Quantity_1                 -0.0196845  0.0243872  -0.807 0.419576    
## Quantity_2                 -0.0194771  0.0276057  -0.706 0.480473    
## Quantity_3                 -0.0258413  0.0250564  -1.031 0.302392    
## Quantity_4                 -0.0268005  0.0211910  -1.265 0.205982    
## Quantity_5                 -0.0212369  0.0189229  -1.122 0.261745    
## Quantity_6                 -0.0217088  0.0153217  -1.417 0.156526    
## Quantity_7                 -0.0251731  0.0138203  -1.821 0.068543 .  
## Quantity_8                 -0.0143426  0.0108144  -1.326 0.184762    
## Quantity_9                 -0.0127189  0.0094581  -1.345 0.178708    
## Quantity_10                -0.0001522  0.0059697  -0.025 0.979663    
## Quantity_11                -0.0082984  0.0051451  -1.613 0.106778    
## Quantity_12                -0.0062544  0.0052900  -1.182 0.237092    
## Quantity_13                -0.0044283  0.0045656  -0.970 0.332083    
## Quantity_14                        NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8596 on 51237 degrees of freedom
## Multiple R-squared:  0.2618, Adjusted R-squared:  0.261 
## F-statistic: 349.4 on 52 and 51237 DF,  p-value: < 2.2e-16
```

###The lm() function was used to fit a linear regression model to the data. This model predicts the Order.Priority variable using various predictor variables, such as market, region, sales, discount, profit, shipping cost, ship mode, segment, category, subcategory, and quantity. ###The linear regression model found that several predictor variables have a significant impact on the Order.Priority variable. ###Variables with low p-values (e.g., < 0.05) are typically considered statistically significant and may be useful for predicting the response variable. On the other hand, variables with high p-values (e.g., > 0.05) are usually not statistically significant and may not contribute much to the model. ###The F-statistic tests the overall significance of the linear regression model. It assesses whether there is a significant linear relationship between the predictor variables and the Order.Priority variable. ###The associated p-value (p-value: < 2.2e-16) is extremely small, indicating strong evidence against the null hypothesis of no relationship between the predictor variables and Order.Priority. ### variables that appear to be statistically significant: ###MarketAPAC, RegionCaribbean, RegionCentral Asia, RegionNorth Asia, RegionOceania, RegionSouth, Sales, Shipping.Cost, Ship.Mode_First Class, Ship.Mode_Same Day, Ship.Mode_Second Class, Segment_Consumer, Category_Furniture, Sub.Category_Furnishings, Sub.Category_Bookcases

Anova

```
anova_results<- aov(Order.Priority ~ ., data = data_1)
summary(anova_results)
```

```
##                                Df Sum Sq Mean Sq  F value   Pr(>F)    
## Market                          6     20       3    4.546 0.000129 ***
## Region                          8     53       7    8.961 2.31e-12 ***
## Sales                           1      0       0    0.620 0.430946    
## Discount                        1      0       0    0.441 0.506764    
## Profit                          1      0       0    0.177 0.673627    
## Shipping.Cost                   1   3765    3765 5095.212  < 2e-16 ***
## `Ship.Mode_First Class`         1   3444    3444 4659.928  < 2e-16 ***
## `Ship.Mode_Same Day`            1   2258    2258 3055.630  < 2e-16 ***
## `Ship.Mode_Second Class`        1   3811    3811 5157.484  < 2e-16 ***
## Segment_Consumer                1     20      20   26.945 2.10e-07 ***
## Segment_Corporate               1      0       0    0.000 0.997796    
## Category_Furniture              1     11      11   14.595 0.000133 ***
## `Category_Office Supplies`      1      5       5    7.368 0.006643 ** 
## Sub.Category_Accessories        1      3       3    3.894 0.048477 *  
## Sub.Category_Appliances         1      0       0    0.011 0.916181    
## Sub.Category_Art                1      1       1    0.983 0.321527    
## Sub.Category_Binders            1      0       0    0.650 0.420124    
## Sub.Category_Bookcases          1      0       0    0.002 0.968558    
## Sub.Category_Chairs             1      4       4    4.941 0.026235 *  
## Sub.Category_Copiers            1      0       0    0.127 0.721250    
## Sub.Category_Envelopes          1      0       0    0.348 0.555459    
## Sub.Category_Fasteners          1      0       0    0.006 0.940855    
## Sub.Category_Furnishings        1      7       7    9.989 0.001576 ** 
## Sub.Category_Labels             1      0       0    0.370 0.543229    
## Sub.Category_Machines           1      0       0    0.020 0.888496    
## Sub.Category_Paper              1      2       2    2.920 0.087482 .  
## Sub.Category_Storage            1      0       0    0.548 0.459083    
## Quantity_1                      1      1       1    1.199 0.273554    
## Quantity_2                      1      9       9   12.462 0.000416 ***
## Quantity_3                      1      2       2    3.143 0.076282 .  
## Quantity_4                     1      0       0    0.210 0.646469
## Quantity_5                     1      2       2    2.400 0.121373
## Quantity_6                     1      0       0    0.086 0.769886
## Quantity_7                     1      2       2    2.555 0.109979
## Quantity_8                     1      0       0    0.163 0.686168
## Quantity_9                     1      1       1    0.687 0.407224
## Quantity_10                    1      1       1    1.785 0.181516
## Quantity_11                    1      1       1    1.202 0.272910
## Quantity_12                    1      1       1    0.767 0.381288
## Quantity_13                    1      1       1    0.941 0.332083
## Residuals                  51237  37862       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

###The ANOVA table displays the significance levels (p-values) and statistics for each predictor variable. The p-values can be used to assess which predictors have a significant impact on the Order.Priority variable. ###The smaller p-values (typically below a pre-defined significance level, such as 0.05) are considered statistically significant. ###These factors that are marked *** can be selected for feature selection. ###factors with larger p-values (> 0.05) are considered non-significant and may be removed from the model during the feature selection process. ###Market, Region, Shipping.Cost, Ship.Mode_First Class, Ship.Mode_Same Day, Ship.Mode_Second Class, Segment_Consumer, Category_Furniture, Category_Office Supplies, Sub.Category_Accessories, Sub.Category_Chairs, Sub.Category_Furnishings, Sub.Category_Paper, Quantity_2, Quantity_3

Step Wise Regression

```
train.control <- trainControl(method = "cv", number = 15)
# Train the model
step.model <- train(Order.Priority ~ ., data = data_1,
                    method = "leapBackward",
                    tuneGrid = data.frame(nvmax = 1:5),
                    trControl = train.control
                    )

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:
```

```
## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:
```

```
## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:

## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
11
## linear dependencies found

## Reordering variables and trying again:
```

##backward stepwise regression with cross-validation to train a model for predicting Order.Priority based on the variables in the data_1 dataset.

```
step.model$results

##   nvmax      RMSE     Rsquared       MAE      RMSESD   RsquaredSD
MAESD
## 1     1 0.9998644 0.0003761039 0.8597905 0.008148773 0.0004352627
0.006306285
## 2     2 0.9998611 0.0003651073 0.8597743 0.008165764 0.0003600903
0.006314623
## 3     3 0.9998490 0.0004219598 0.8597355 0.008164663 0.0004321954
0.006308688
## 4     4 0.9997407 0.0008268008 0.8595648 0.008072135 0.0011369734
0.006220061
## 5     5 0.9534077 0.0912967793 0.7831148 0.007183741 0.0099344588
0.004874602
```

obtain insights into which variables were included in the final model, their respective coefficients, and an assessment of the model's fit or predictive performance.

```
summary(step.model$finalModel)

## Subset selection object
## 63 Variables  (and intercept)
##                             Forced in Forced out
## MarketAPAC                      FALSE      FALSE
## MarketCanada                    FALSE      FALSE
## MarketEMEA                      FALSE      FALSE
## MarketEU                       FALSE      FALSE
## MarketLATAM                    FALSE      FALSE
## MarketUS                       FALSE      FALSE
## RegionCaribbean                FALSE      FALSE
## RegionCentral                  FALSE      FALSE
## RegionCentral Asia             FALSE      FALSE
## RegionEast                     FALSE      FALSE
## RegionNorth                    FALSE      FALSE
## RegionNorth Asia               FALSE      FALSE
## RegionOceania                  FALSE      FALSE
## RegionSouth                    FALSE      FALSE
## Sales                          FALSE      FALSE
## Discount                       FALSE      FALSE
## Profit                         FALSE      FALSE
## Shipping.Cost                  FALSE      FALSE
## `Ship.Mode_First Class`        FALSE      FALSE
## `Ship.Mode_Same Day`           FALSE      FALSE
## `Ship.Mode_Second Class`       FALSE      FALSE
## Segment_Consumer               FALSE      FALSE
## Segment_Corporate              FALSE      FALSE
## Category_Furniture             FALSE      FALSE
## `Category_Office Supplies`     FALSE      FALSE
## Sub.Category_Accessories       FALSE      FALSE
## Sub.Category_Appliances        FALSE      FALSE
## Sub.Category_Art               FALSE      FALSE
## Sub.Category_Binders           FALSE      FALSE
## Sub.Category_Bookcases         FALSE      FALSE
## Sub.Category_Chairs            FALSE      FALSE
## Sub.Category_Copiers           FALSE      FALSE
## Sub.Category_Envelopes         FALSE      FALSE
## Sub.Category_Fasteners         FALSE      FALSE
## Sub.Category_Furnishings       FALSE      FALSE
## Sub.Category_Labels            FALSE      FALSE
## Sub.Category_Machines          FALSE      FALSE
## Sub.Category_Paper             FALSE      FALSE
## Sub.Category_Storage           FALSE      FALSE
## Quantity_1                     FALSE      FALSE
## Quantity_2                     FALSE      FALSE
## Quantity_3                     FALSE      FALSE
## Quantity_4                     FALSE      FALSE
## Quantity_5                     FALSE      FALSE
## Quantity_6                     FALSE      FALSE
## Quantity_7                     FALSE      FALSE
## Quantity_8                     FALSE      FALSE
## Quantity_9                     FALSE      FALSE
## Quantity_10                    FALSE      FALSE
## Quantity_11                    FALSE      FALSE
## Quantity_12                    FALSE      FALSE
## Quantity_13                    FALSE      FALSE
## RegionCanada                   FALSE      FALSE
## RegionEMEA                     FALSE      FALSE
## RegionSoutheast Asia           FALSE      FALSE
## RegionWest                     FALSE      FALSE
## `Ship.Mode_Standard Class`     FALSE      FALSE
## `Segment_Home Office`          FALSE      FALSE
## Category_Technology            FALSE      FALSE
## Sub.Category_Phones            FALSE      FALSE
## Sub.Category_Supplies          FALSE      FALSE
## Sub.Category_Tables            FALSE      FALSE
## Quantity_14                    FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: backward
##          MarketAPAC MarketCanada MarketEMEA MarketEU MarketLATAM MarketUS
## 1  ( 1 ) " "        " "          " "        " "      " "         " "     
## 2  ( 1 ) " "        " "          " "        " "      " "         " "     
## 3  ( 1 ) " "        " "          " "        " "      " "         " "     
## 4  ( 1 ) " "        " "          " "        " "      " "         " "     
## 5  ( 1 ) " "        " "          " "        " "      " "         " "     
## 6  ( 1 ) " "        " "          " "        " "      " "         " "     
##          RegionCanada RegionCaribbean RegionCentral RegionCentral Asia
## 1  ( 1 ) " "          " "             " "           " "               
## 2  ( 1 ) " "          " "             " "           " "               
## 3  ( 1 ) " "          " "             " "           " "               
## 4  ( 1 ) " "          " "             " "           " "               
## 5  ( 1 ) " "          " "             " "           " "               
## 6  ( 1 ) " "          " "             " "           " "               
##          RegionEast RegionEMEA RegionNorth RegionNorth Asia RegionOceania
## 1  ( 1 ) " "        " "        " "         " "              " "          
## 2  ( 1 ) " "        " "        " "         " "              " "          
## 3  ( 1 ) " "        " "        " "         " "              " "          
## 4  ( 1 ) " "        " "        " "         " "              " "          
## 5  ( 1 ) " "        " "        " "         " "              " "          
## 6  ( 1 ) " "        " "        " "         " "              " "          
##          RegionSouth RegionSoutheast Asia RegionWest Sales Discount Profit
## 1  ( 1 ) " "         " "                  " "        " "   " "      " "   
## 2  ( 1 ) " "         " "                  " "        " "   " "      " "   
## 3  ( 1 ) " "         " "                  " "        " "   " "      " "   
## 4  ( 1 ) " "         " "                  " "        " "   " "      " "   
## 5  ( 1 ) " "         " "                  " "        "*"   " "      " "   
## 6  ( 1 ) " "         " "                  " "        "*"   " "      " "   
##          Shipping.Cost `Ship.Mode_First Class` `Ship.Mode_Same Day`
## 1  ( 1 ) " "           "*"                     " "                 
## 2  ( 1 ) " "           "*"                     " "                 
## 3  ( 1 ) " "           "*"                     "*"                 
## 4  ( 1 ) "*"           "*"                     "*"                 
## 5  ( 1 ) "*"           "*"                     "*"                 
## 6  ( 1 ) "*"           "*"                     "*"                 
##          `Ship.Mode_Second Class` `Ship.Mode_Standard Class`
Segment_Consumer
## 1  ( 1 ) " "                      " "                        " "
## 2  ( 1 ) "*"                      " "                        " "
## 3  ( 1 ) "*"                      " "                        " "
## 4  ( 1 ) "*"                      " "                        " "
## 5  ( 1 ) "*"                      " "                        " "
## 6  ( 1 ) "*"                      " "                        "*"
##          Segment_Corporate `Segment_Home Office` Category_Furniture
## 1  ( 1 ) " "               " "                   " "
## 2  ( 1 ) " "               " "                   " "
## 3  ( 1 ) " "               " "                   " "
## 4  ( 1 ) " "               " "                   " "
## 5  ( 1 ) " "               " "                   " "
## 6  ( 1 ) " "               " "                   " "
##          `Category_Office Supplies` Category_Technology
## 1  ( 1 ) " "                        " "
## 2  ( 1 ) " "                        " "
## 3  ( 1 ) " "                        " "
## 4  ( 1 ) " "                        " "
## 5  ( 1 ) " "                        " "
## 6  ( 1 ) " "                        " "
##          Sub.Category_Accessories Sub.Category_Appliances Sub.Category_Art
## 1  ( 1 ) " "                      " "                     " "
## 2  ( 1 ) " "                      " "                     " "
## 3  ( 1 ) " "                      " "                     " "
## 4  ( 1 ) " "                      " "                     " "
## 5  ( 1 ) " "                      " "                     " "
## 6  ( 1 ) " "                      " "                     " "
##          Sub.Category_Binders Sub.Category_Bookcases Sub.Category_Chairs
## 1  ( 1 ) " "                  " "                    " "
## 2  ( 1 ) " "                  " "                    " "
## 3  ( 1 ) " "                  " "                    " "
## 4  ( 1 ) " "                  " "                    " "
## 5  ( 1 ) " "                  " "                    " "
## 6  ( 1 ) " "                  " "                    " "
##          Sub.Category_Copiers Sub.Category_Envelopes
Sub.Category_Fasteners
## 1  ( 1 ) " "                  " "                    " "
## 2  ( 1 ) " "                  " "                    " "
## 3  ( 1 ) " "                  " "                    " "
## 4  ( 1 ) " "                  " "                    " "
## 5  ( 1 ) " "                  " "                    " "
## 6  ( 1 ) " "                  " "                    " "
##          Sub.Category_Furnishings Sub.Category_Labels
Sub.Category_Machines
## 1  ( 1 ) " "                      " "                 " "
## 2  ( 1 ) " "                      " "                 " "
## 3  ( 1 ) " "                      " "                 " "
## 4  ( 1 ) " "                      " "                 " "
## 5  ( 1 ) " "                      " "                 " "
## 6  ( 1 ) " "                      " "                 " "
##          Sub.Category_Paper Sub.Category_Phones Sub.Category_Storage
## 1  ( 1 ) " "                " "                 " "
## 2  ( 1 ) " "                " "                 " "
## 3  ( 1 ) " "                " "                 " "
## 4  ( 1 ) " "                " "                 " "
## 5  ( 1 ) " "                " "                 " "
## 6  ( 1 ) " "                " "                 " "
##          Sub.Category_Supplies Sub.Category_Tables Quantity_1 Quantity_2
## 1  ( 1 ) " "                   " "                 " "        " "
## 2  ( 1 ) " "                   " "                 " "        " "
## 3  ( 1 ) " "                   " "                 " "        " "
## 4  ( 1 ) " "                   " "                 " "        " "
## 5  ( 1 ) " "                   " "                 " "        " "
## 6  ( 1 ) " "                   " "                 " "        " "
##          Quantity_3 Quantity_4 Quantity_5 Quantity_6 Quantity_7 Quantity_8
## 1  ( 1 ) " "        " "        " "        " "        " "        " "
## 2  ( 1 ) " "        " "        " "        " "        " "        " "
## 3  ( 1 ) " "        " "        " "        " "        " "        " "
## 4  ( 1 ) " "        " "        " "        " "        " "        " "
## 5  ( 1 ) " "        " "        " "        " "        " "        " "
## 6  ( 1 ) " "        " "        " "        " "        " "        " "
##          Quantity_9 Quantity_10 Quantity_11 Quantity_12 Quantity_13
Quantity_14
## 1  ( 1 ) " "        " "         " "         " "         " "         " "
## 2  ( 1 ) " "        " "         " "         " "         " "         " "
## 3  ( 1 ) " "        " "         " "         " "         " "         " "
## 4  ( 1 ) " "        " "         " "         " "         " "         " "
## 5  ( 1 ) " "        " "         " "         " "         " "         " "
## 6  ( 1 ) " "        " "         " "         " "         " "         " "
```

###“Forced in” indicates whether a variable was forced to be included in the model. If it is marked as “TRUE”, it means the variable was included in the model regardless of the stepwise regression process. If it is marked as “FALSE”, the variable was selected based on the stepwise regression algorithm.

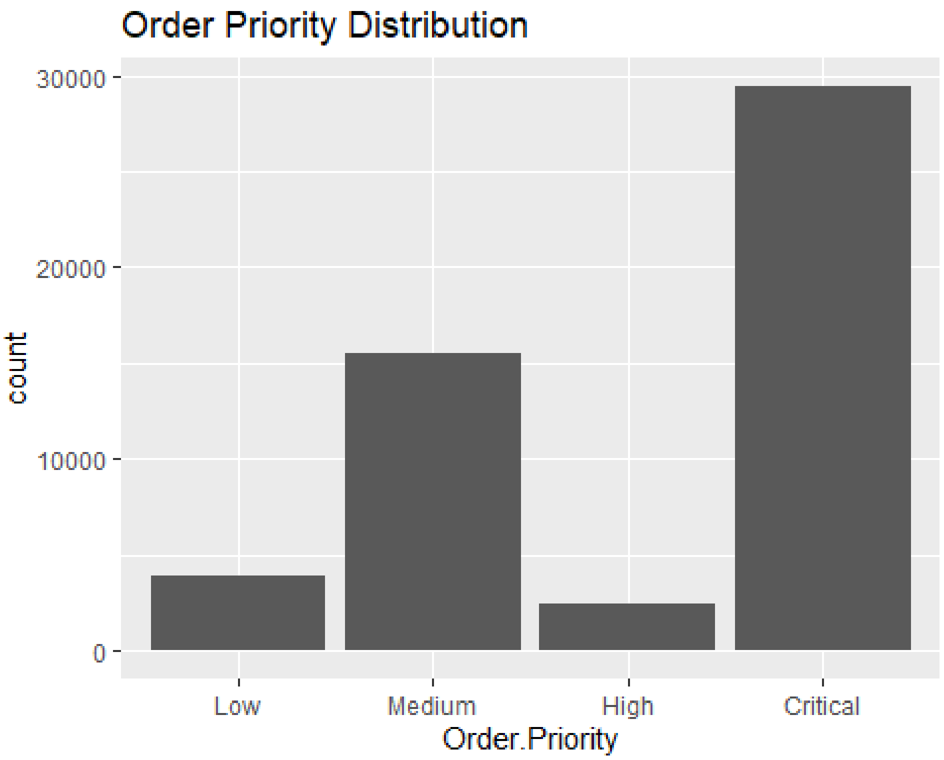
###“Forced out” indicates whether a variable was forced to be excluded from the model. If it is marked as “TRUE”, it means the variable was excluded from the model regardless of the stepwise regression process. If it is marked as “FALSE”, the variable was selected based on the stepwise regression algorithm.

there are a total of 63 variables, including an intercept term. For each variable, “FALSE” is indicated for both “Forced in” and “Forced out”, which means all variables were selected through the stepwise regression process without any forced inclusions or exclusions.

###(” “) means the variable of no use in feature selection where as (”“) means the variable *as part of feature selection activity. ###It is useful as part of feature selection activity

```
library(tidyverse)
library(ggplot2)

# Read the dataset
```

```
data_1 <- read.csv("C:/Users/vamsh/Downloads/superstore_dataset2011-
2015.csv")

# Check the levels of Order.Priority
order_priority_levels <- levels(data_1$Order.Priority)

# Check if Order.Priority has at least two levels
if (length(order_priority_levels) < 2) {
  data_1$Order.Priority <- ifelse(data_1$Order.Priority == "", "Other",
data_1$Order.Priority)
}

# Create a bar plot of Order.Priority
data_1 %>%
  ggplot(aes(x = Order.Priority)) +
  geom_bar() +
  labs(title = "Order Priority Distribution") +
  scale_x_discrete(labels = c("Low", "Medium", "High", "Critical", "Other"))
```



###To draw more specific conclusions or analyze the distribution in more detail.

```
# Convert Order.Priority to binary variable
df$Order.Priority.Binary <- ifelse(df$Order.Priority >= 2, 1, 0)

write.csv(df,"Superstore_Data1.csv", row.names = F)
```