superstore_data_analysis

Ziting Liang

2022-04-25

Import packages

```
library(dplyr)
library(car)
library(ggplot2)
```

Read data

```
Data<-read.csv('./Sample - Superstore.csv',header=T)
summary(Data)
```

```
##      Row.ID       Order.ID          Order.Date         Ship.Date
##  Min.   :   1   Length:9994        Length:9994        Length:9994
##  1st Qu.:2499   Class :character   Class :character   Class :character
##  Median :4998   Mode  :character   Mode  :character   Mode  :character
##  Mean   :4998
##  3rd Qu.:7496
##  Max.   :9994
##   Ship.Mode         Customer.ID        Customer.Name        Segment
##  Length:9994        Length:9994        Length:9994        Length:9994
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Country              City              State            Postal.Code
##  Length:9994        Length:9994        Length:9994        Min.   : 1040
##  Class :character   Class :character   Class :character   1st Qu.:23223
##  Mode  :character   Mode  :character   Mode  :character   Median :56430
##                                                           Mean   :55190
##                                                           3rd Qu.:90008
##                                                           Max.   :99301
##     Region           Product.ID          Category         Sub.Category
##  Length:9994        Length:9994        Length:9994        Length:9994
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Product.Name           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
```

```
#apply(is.na(Data),2,sum)#No NA

#check for some key variables
#unique(Data$Variable.name)

#select useful variables
data.clean<-Data%>%select(Order.Date,Ship.Mode,Customer.ID,Segment,State,Region,Category,
                          Sub.Category,Sales,Quantity,Discount,Profit)
summary(data.clean)
```

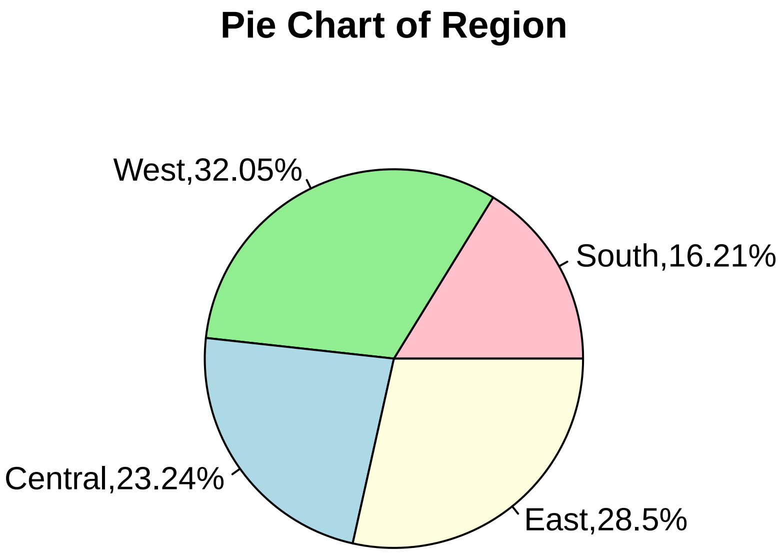
```
##   Order.Date         Ship.Mode         Customer.ID          Segment
##  Length:9994        Length:9994        Length:9994        Length:9994
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     State              Region            Category         Sub.Category
##  Length:9994        Length:9994        Length:9994        Length:9994
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Sales              Quantity        Discount          Profit
##  Min.   :    0.444   Min.   : 1.00   Min.   :0.0000   Min.   :-6599.978
##  1st Qu.:   17.280   1st Qu.: 2.00   1st Qu.:0.0000   1st Qu.:    1.729
##  Median :   54.490   Median : 3.00   Median :0.2000   Median :    8.666
##  Mean   :  229.858   Mean   : 3.79   Mean   :0.1562   Mean   :   28.657
##  3rd Qu.:  209.940   3rd Qu.: 5.00   3rd Qu.:0.2000   3rd Qu.:   29.364
##  Max.   :22638.480   Max.   :14.00   Max.   :0.8000   Max.   : 8399.976
```

Variable recoding

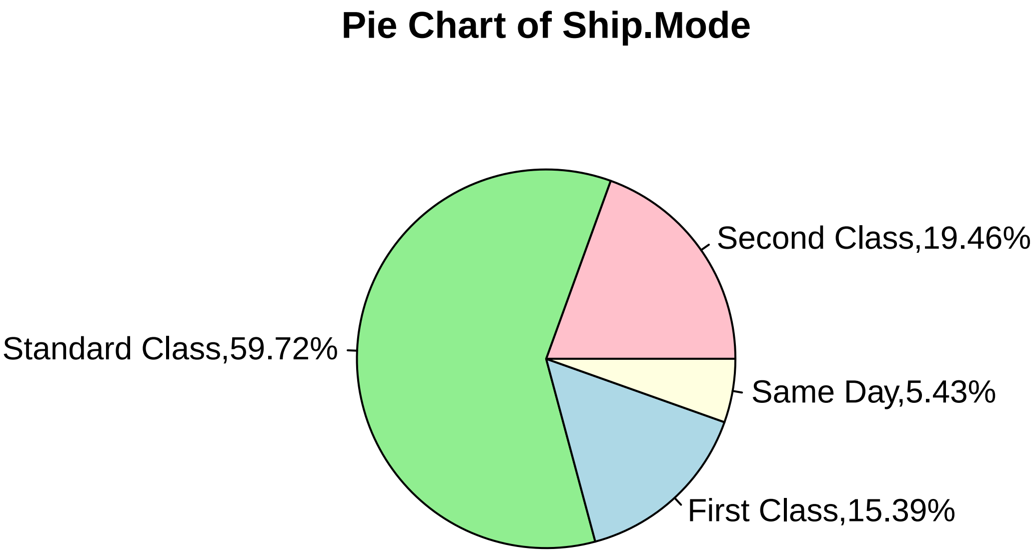
```
#recode 4 new variables for Ship.Mode
data.clean$Ship.Mode[data.clean$Ship.Mode=='Second Class']<-1
data.clean$Ship.Mode[data.clean$Ship.Mode=='Standard Class']<-2
data.clean$Ship.Mode[data.clean$Ship.Mode=='First Class']<-3
data.clean$Ship.Mode[data.clean$Ship.Mode=='Same Day']<-4

#recode 3 new variables for Segment
data.clean$Segment[data.clean$Segment=='Consumer']<-1
data.clean$Segment[data.clean$Segment=='Corporate']<-2
data.clean$Segment[data.clean$Segment=='Home Office']<-3

#recode 4 new variables for Region
data.clean$Region[data.clean$Region=='South']<-1
data.clean$Region[data.clean$Region=='West']<-2
data.clean$Region[data.clean$Region=='Central']<-3
data.clean$Region[data.clean$Region=='East']<-4

#recode 3 new variables for Category
data.clean$Category[data.clean$Category=='Furniture']<-1
data.clean$Category[data.clean$Category=='Office Supplies']<-2
data.clean$Category[data.clean$Category=='Technology']<-3

data.clean$Sales<-round(data.clean$Sales, digits=2)
data.clean$Profit<-round(data.clean$Profit, digits=2)

data.clean$Order.Date<-as.Date(data.clean$Order.Date,format='%m/%d/%Y')
```

Descriptive statistics

```
# pie chart of region
data_region<-data.clean %>% select(Region)
slices <- table(data.clean$Region)
count<-slices/sum(slices)*100
lbls <- c("South", "West", "Central", 'East')
pie(slices, col=c('pink','lightgreen','lightblue','lightyellow'),
    labels = paste0(lbls,',',round(count,2),'%'), main="Pie Chart of Region")
```

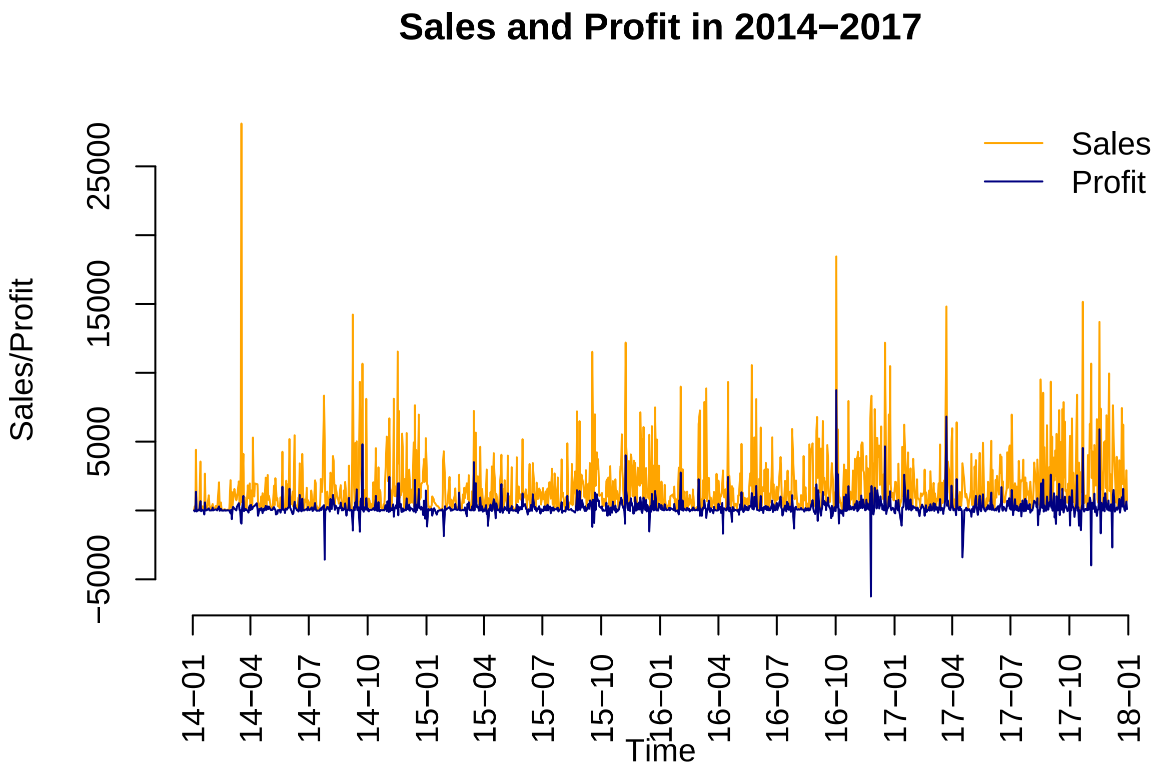


```
# pie chart of Ship.Mode
data_region<-data.clean %>% select(Ship.Mode)
slices <- table(data.clean$Ship.Mode)
count<-slices/sum(slices)*100
lbls <- c("Second Class", "Standard Class", "First Class", 'Same Day')
pie(slices, col=c('pink','lightgreen','lightblue','lightyellow'),
    labels = paste0(lbls,',',round(count,2),'%'), main="Pie Chart of Ship.Mode")
```



Pie Chart of Ship.Mode

```
#A comparison of Sales and Profit in 2014-2017
chrono_data<-data.clean%>%group_by(Order.Date)%>%summarize(Sales=sum(Sales),Profit=sum(Profit))
matplot(chrono_data$Order.Date, cbind(chrono_data$Sales,chrono_data$Profit),
        type="l",col=c("orange","navyblue"),lty=c(1,1),xlab='Time',ylab='Sales/Profit',
        main='Sales and Profit in 2014-2017',axes=F,
        xlim=c(as.Date('2014-01-01'),as.Date('2018-01-01')))
Axis(side=2)
legend(x='topright',legend=c('Sales','Profit'),lty=1,col=c('orange','navyblue'),bty='n')
axis.Date(1,at=seq(from=as.Date('2014-01-01'),to=as.Date('2018-01-01'),by='3 months'),
          format='%y-%m',las=3)
```



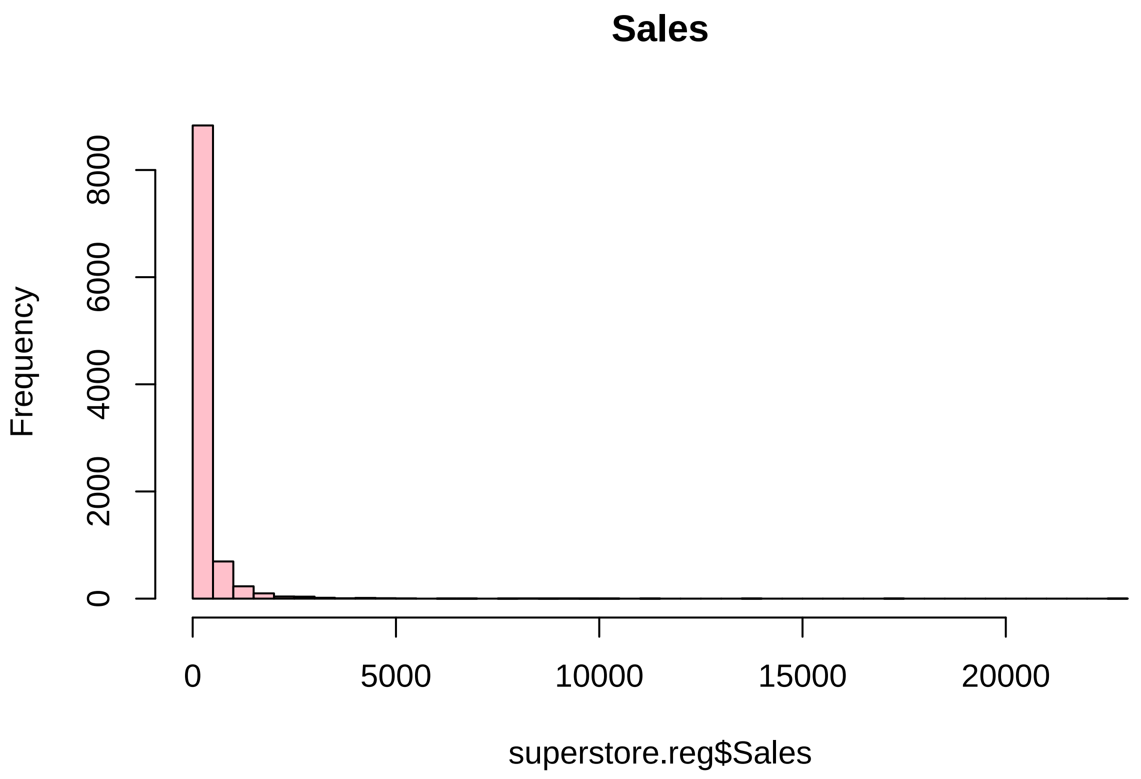
Sales and Profit in 2014−2017

Convert to superstore.reg: preparation for regression

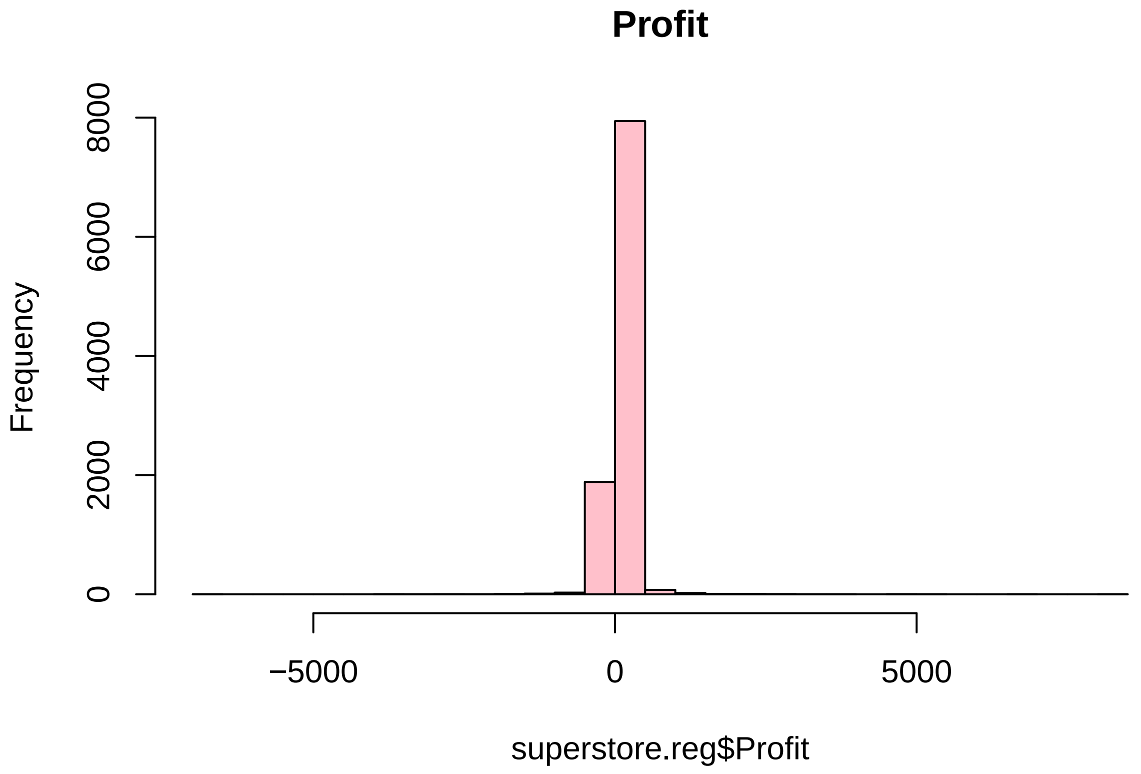
```
#data.clean$Order.Date<-as.Date(data.clean$Order.Date,format='%m%d%Y')

#turn category variables to numeric variables
data.clean$Ship.Mode<-as.numeric(data.clean$Ship.Mode)
data.clean$Segment<-as.numeric(data.clean$Segment)
data.clean$Region<-as.numeric(data.clean$Region)
data.clean$Category<-as.numeric(data.clean$Category)

superstore.reg<-data.clean%>%select(Ship.Mode,Segment,Region,Category,
                                    Sales,Quantity,Discount,Profit)

#create 4 new variables for Ship.Mode for regression in case of multicollinearity,
#and then remove Ship.Mode.
superstore.reg<-superstore.reg%>%
  mutate(SM_second_class=as.numeric(Ship.Mode==1),
         SM_standard_class=as.numeric(Ship.Mode==2),
         SM_first_class=as.numeric(Ship.Mode==3))%>%
  select(-Ship.Mode)

#create 3 new variables for Segment for regression in case of multicollinearity,
#and then remove Segment.
superstore.reg<-superstore.reg%>%mutate(Seg_cons=as.numeric(Segment==1),
                                        Seg_corp=as.numeric(Segment==2))%>%select(-Segment)

#create 4 new variables for Region for regression in case of multicollinearity,
#and then remove Region.
superstore.reg<-superstore.reg%>%mutate(Reg_south=as.numeric(Region==1),
                                        Reg_west=as.numeric(Region==2),
                                        Reg_central=as.numeric(Region==3))%>%select(-Region)

#create 3 new variables for Category for regression in case of multicollinearity,
#and then remove Category.
superstore.reg<-superstore.reg%>%mutate(Cat_furniture=as.numeric(Category==1),
                                        Cat_office=as.numeric(Category==2))%>%select(-Category)
```

Regression before scale (including Correlation Analysis)

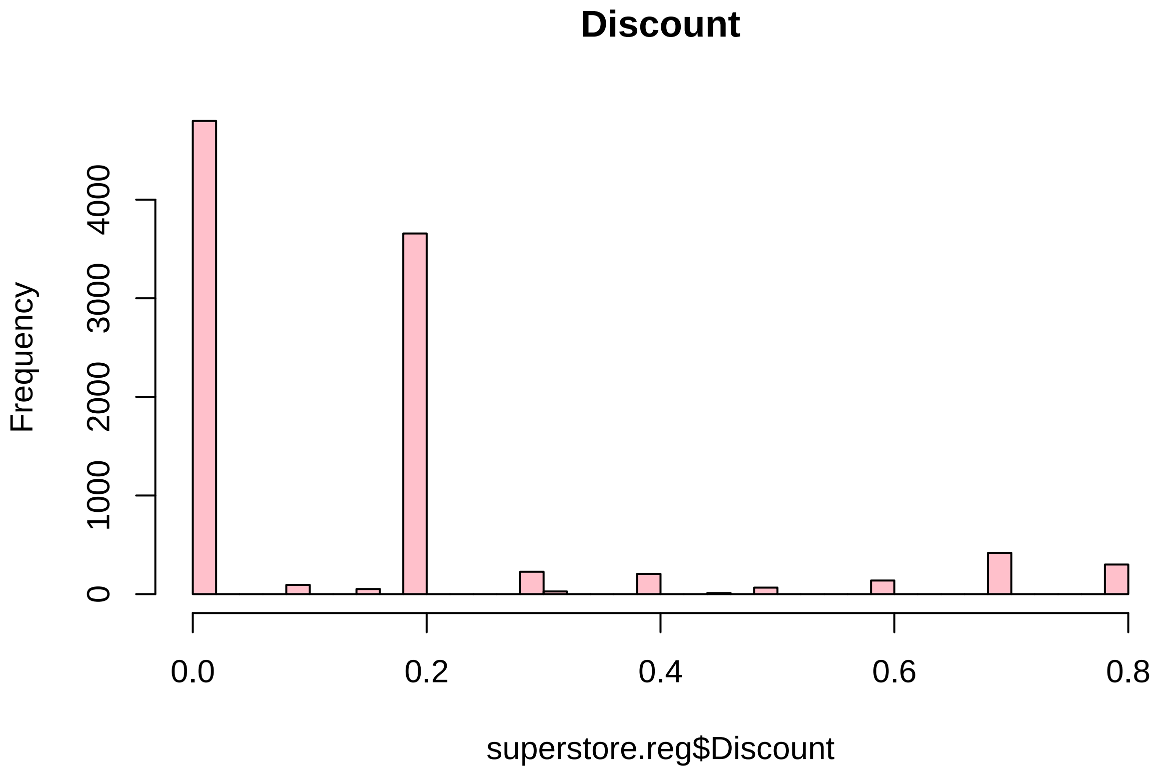
```
###Regression

#Do a normality test before the regression
hist(superstore.reg$Sales, col='pink', main='Sales', breaks=50)
```



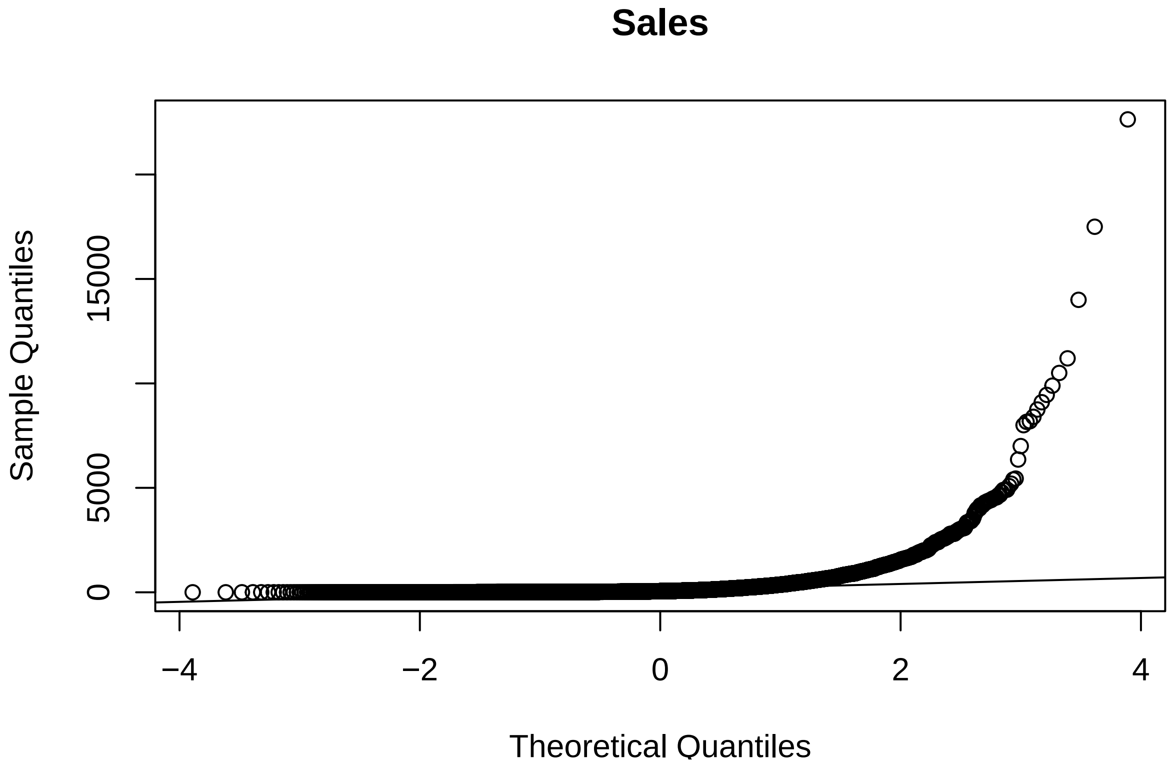
```
hist(superstore.reg$Profit, col='pink', main='Profit', breaks=50)
```



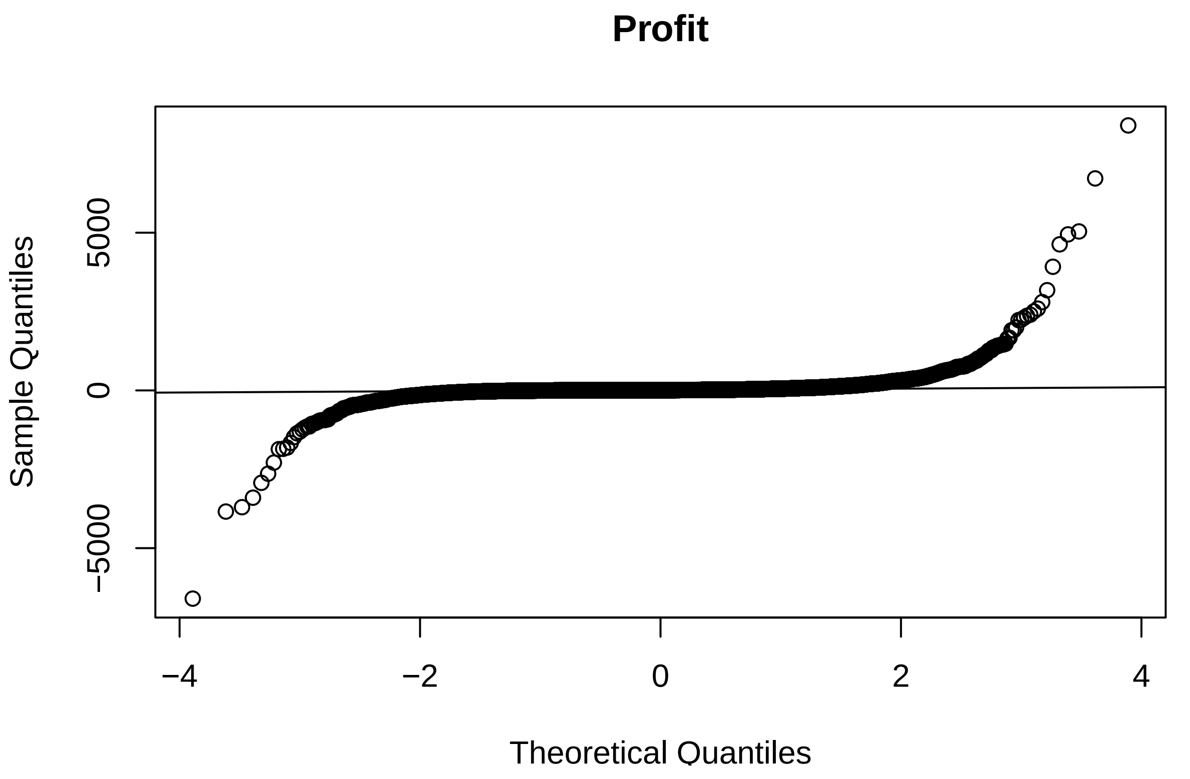
```
hist(superstore.reg$Discount, col='pink', main='Discount',breaks=50)
```



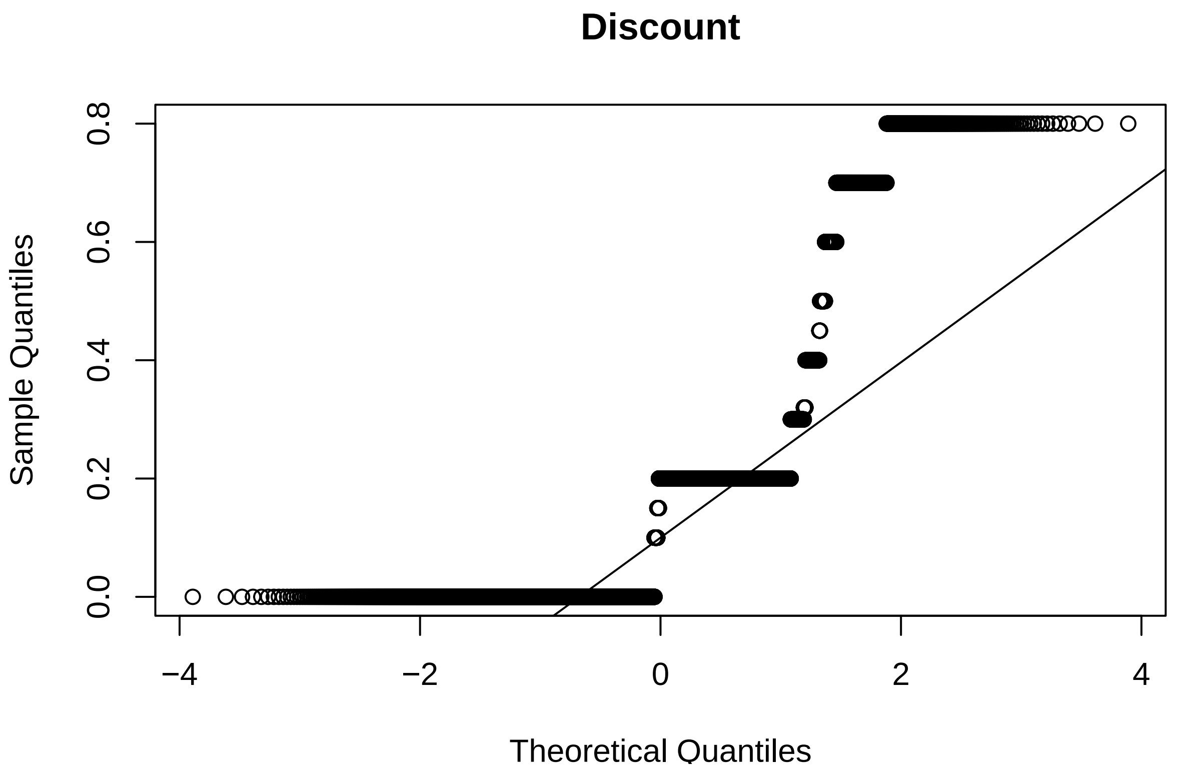
```
qqnorm(superstore.reg$Sales, main='Sales')
qqline(superstore.reg$Sales)
```



```
qqnorm(superstore.reg$Profit, main='Profit')
qqline(superstore.reg$Profit)
```



```
qqnorm(superstore.reg$Discount, main='Discount')
qqline(superstore.reg$Discount)
```



```
#Do regression and see what's going on
lm_fit<-lm(Sales~.,superstore.reg)
summary(lm_fit)#not fit well
```

```
##
## Call:
## lm(formula = Sales ~ ., data = superstore.reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -803.8  -183.5   -55.4    40.5 24357.3
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        173.98862   29.70371   5.857 4.85e-09 ***
## Quantity            47.63140    2.34676  20.297  < 2e-16 ***
## Discount           240.59881   26.67493   9.020  < 2e-16 ***
## Profit               1.26358    0.02294  55.082  < 2e-16 ***
## SM_second_class     -8.89256   25.29193  -0.352   0.7251
## SM_standard_class  -15.82996   23.35325  -0.678   0.4979
## SM_first_class     -21.56301   26.00676  -0.829   0.4071
## Seg_cons            -9.72844   14.29031  -0.681   0.4960
## Seg_corp            -9.37050   15.56014  -0.602   0.5470
## Reg_south            5.37439   16.19935   0.332   0.7401
## Reg_west           -12.47639   13.43810  -0.928   0.3532
## Reg_central        -26.42512   14.77570  -1.788   0.0737 .
## Cat_furniture      -25.57845   16.65284  -1.536   0.1246
## Cat_office        -267.58060   13.90668 -19.241  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 520.3 on 9980 degrees of freedom
## Multiple R-squared:  0.304,  Adjusted R-squared:  0.3031
## F-statistic: 335.4 on 13 and 9980 DF,  p-value: < 2.2e-16
```

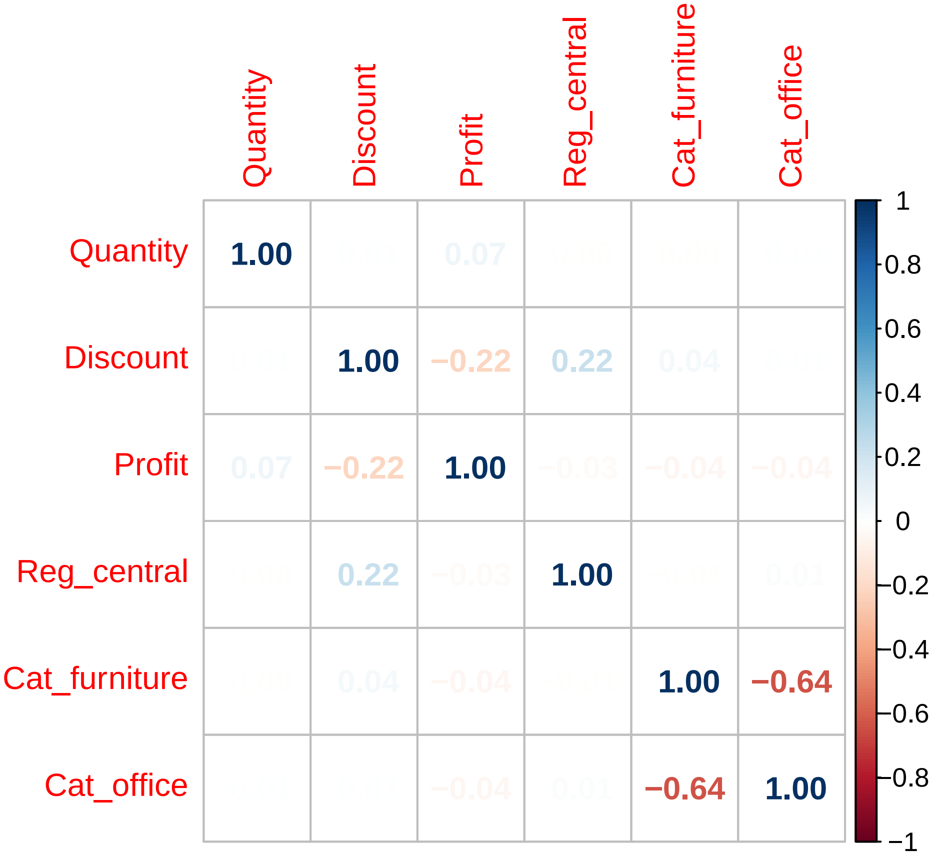
```
reduced<-lm(Sales~1,superstore.reg)
full<-lm(Sales~.,superstore.reg)
step(reduced,scope=c(lower=reduced,upper=full),direction='forward',trace=F)
```

```
##
## Call:
## lm(formula = Sales ~ 1, data = superstore.reg)
##
## Coefficients:
## (Intercept)
##       229.9
```

```
step(full,scope=c(lower=reduced,upper=full),direction='backward',trace=F)
```

```
##
## Call:
## lm(formula = Sales ~ Quantity + Discount + Profit + Reg_central +
##     Cat_furniture + Cat_office, data = superstore.reg)
##
## Coefficients:
##   (Intercept)       Quantity       Discount         Profit    Reg_central
##       147.417         47.596        241.770          1.264        -22.172
## Cat_furniture     Cat_office
##       -25.751       -267.500
```

```
#after doing regression, we want to do a correlation analysis to see if all variables are uncorrelated:
corrplot::corrplot(cor(
  superstore.reg%>%select(c(Quantity,Discount,Profit,Reg_central,Cat_furniture,Cat_office))),
  method='number')
```



```
#we want to see what will the model perform if we delete one of cat_furniture or cat_office
summary(lm(Sales~Quantity+Discount+Profit+Reg_central+Cat_furniture,data=superstore.reg))
```

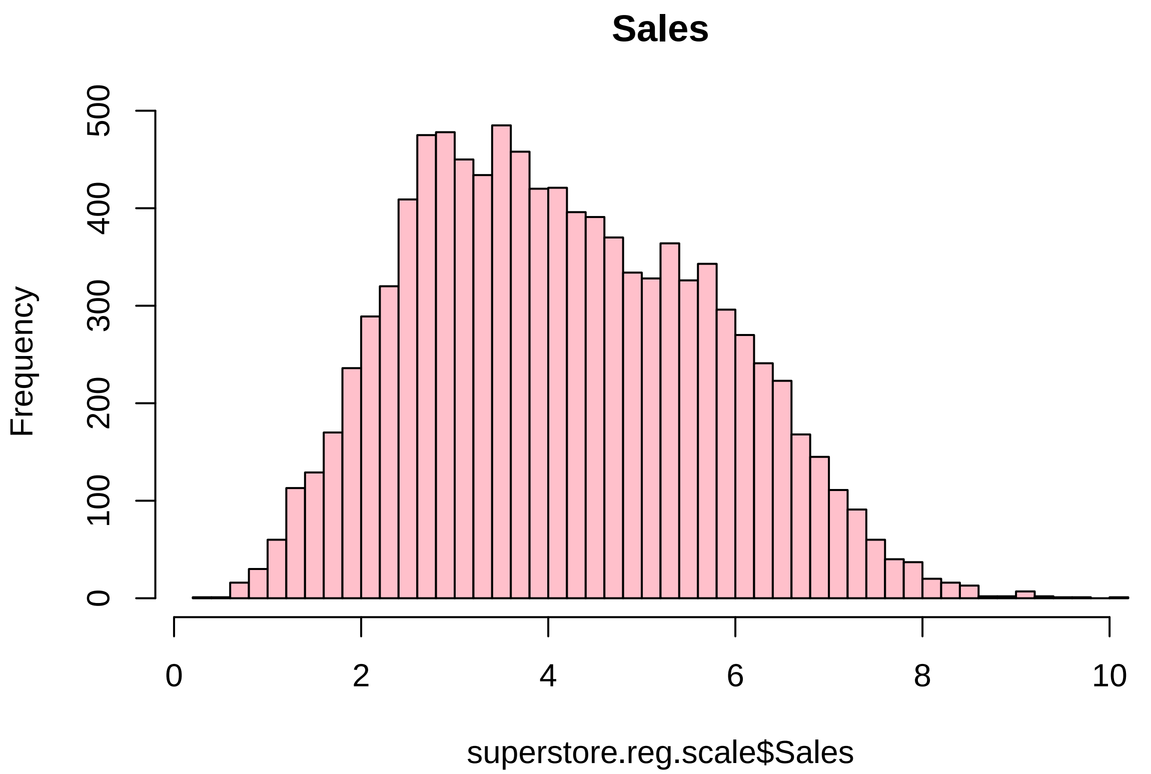
```
##
## Call:
## lm(formula = Sales ~ Quantity + Discount + Profit + Reg_central +
##     Cat_furniture, data = superstore.reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -786.2  -172.1   -74.9    14.6 24654.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -54.48072   11.71655  -4.650 3.36e-06 ***
## Quantity       47.00130    2.38735  19.688  < 2e-16 ***
## Discount      229.01944   27.02121   8.476  < 2e-16 ***
## Profit          1.30210    0.02326  55.975  < 2e-16 ***
## Reg_central   -22.01683   12.87787  -1.710   0.0874 .
## Cat_furniture 180.24926   12.98139  13.885  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 529.7 on 9988 degrees of freedom
## Multiple R-squared:  0.278,  Adjusted R-squared:  0.2777
## F-statistic: 769.2 on 5 and 9988 DF,  p-value: < 2.2e-16
```

```
summary(lm(Sales~Quantity+Discount+Profit+Reg_central+Cat_office,data=superstore.reg))
```

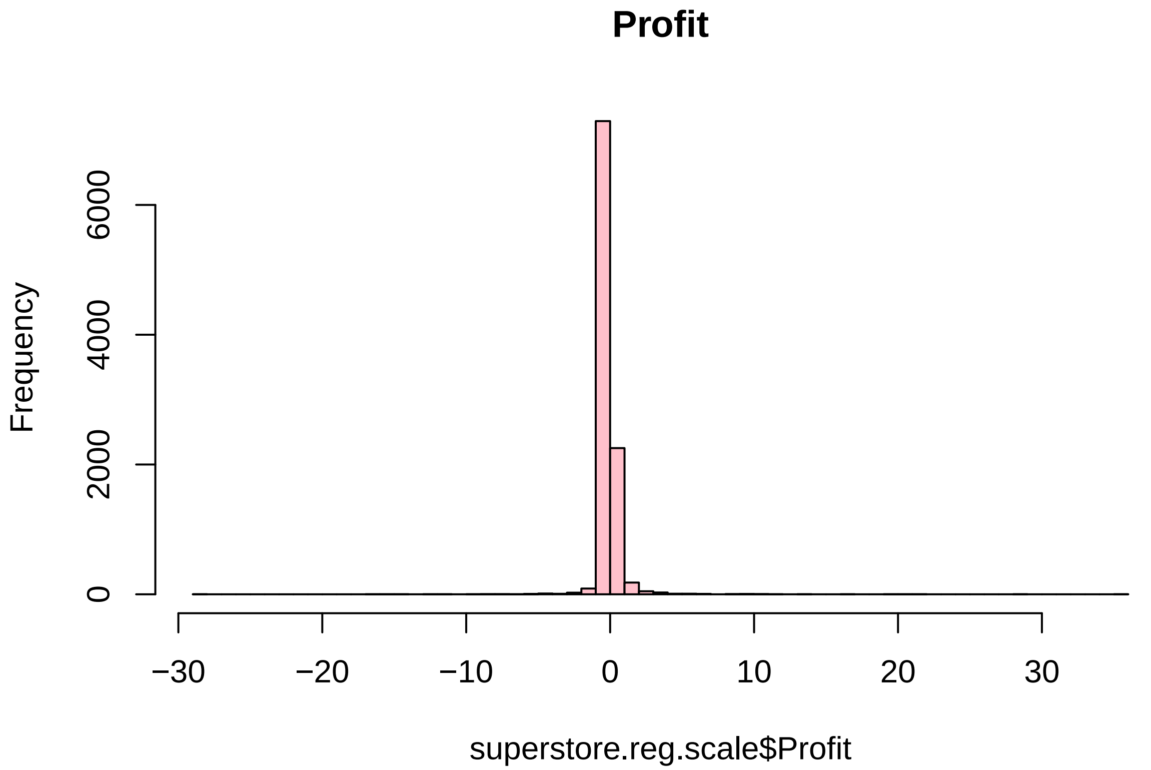
```
##
## Call:
## lm(formula = Sales ~ Quantity + Discount + Profit + Reg_central +
##     Cat_office, data = superstore.reg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -804.1  -182.0   -58.7    35.4 24393.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  133.88847   12.88445  10.391   <2e-16 ***
## Quantity      47.56316    2.34468  20.286   <2e-16 ***
## Discount     239.90021   26.51773   9.047   <2e-16 ***
## Profit         1.26672    0.02285  55.427   <2e-16 ***
## Reg_central  -21.91578   12.64602  -1.733   0.0831 .
## Cat_office  -253.67093   10.64651 -23.827   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 520.2 on 9988 degrees of freedom
## Multiple R-squared:  0.3037, Adjusted R-squared:  0.3033
## F-statistic: 871.1 on 5 and 9988 DF,  p-value: < 2.2e-16
```

Regression after scale or preprocessing (including Correlation Analysis)

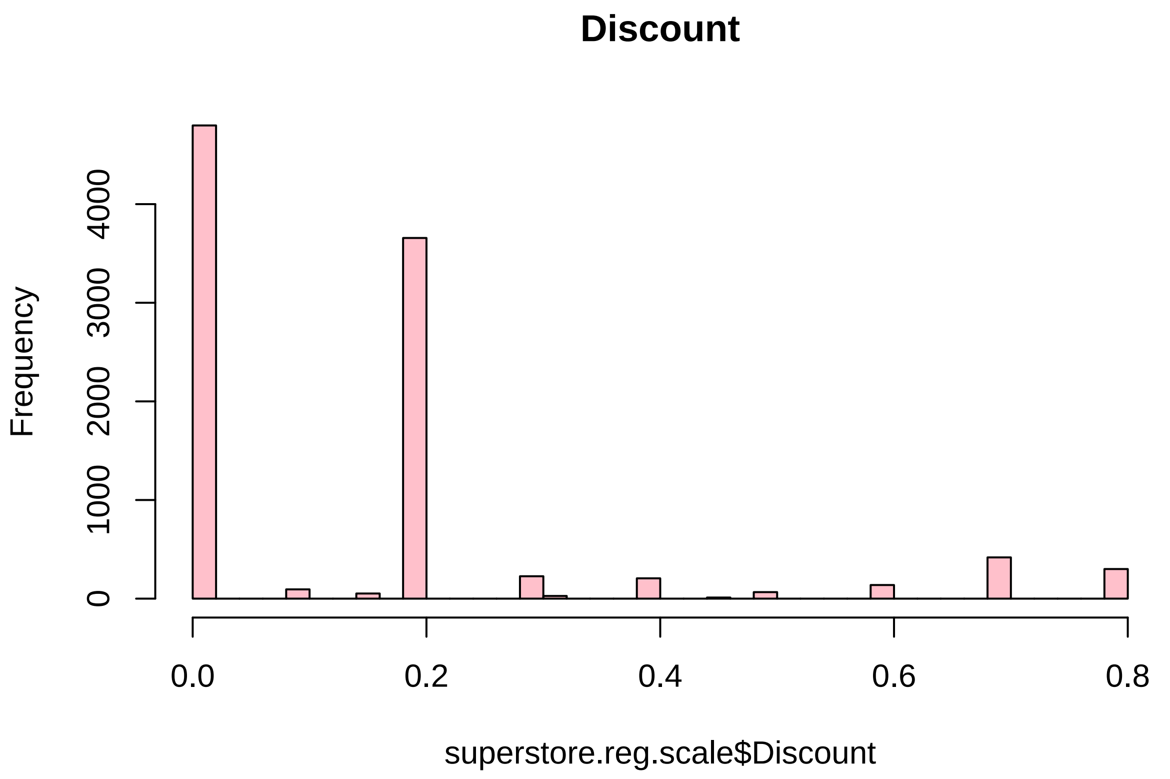
```
#normalization for profit
superstore.reg.scale<-superstore.reg
superstore.reg.scale$Profit<-scale(superstore.reg.scale$Profit)
#take logarithm for sales
superstore.reg.scale$Sales<-log(superstore.reg.scale$Sales+1)

#Do a normality test before the regression
hist(superstore.reg.scale$Sales, col='pink', main='Sales', breaks = 50)
```



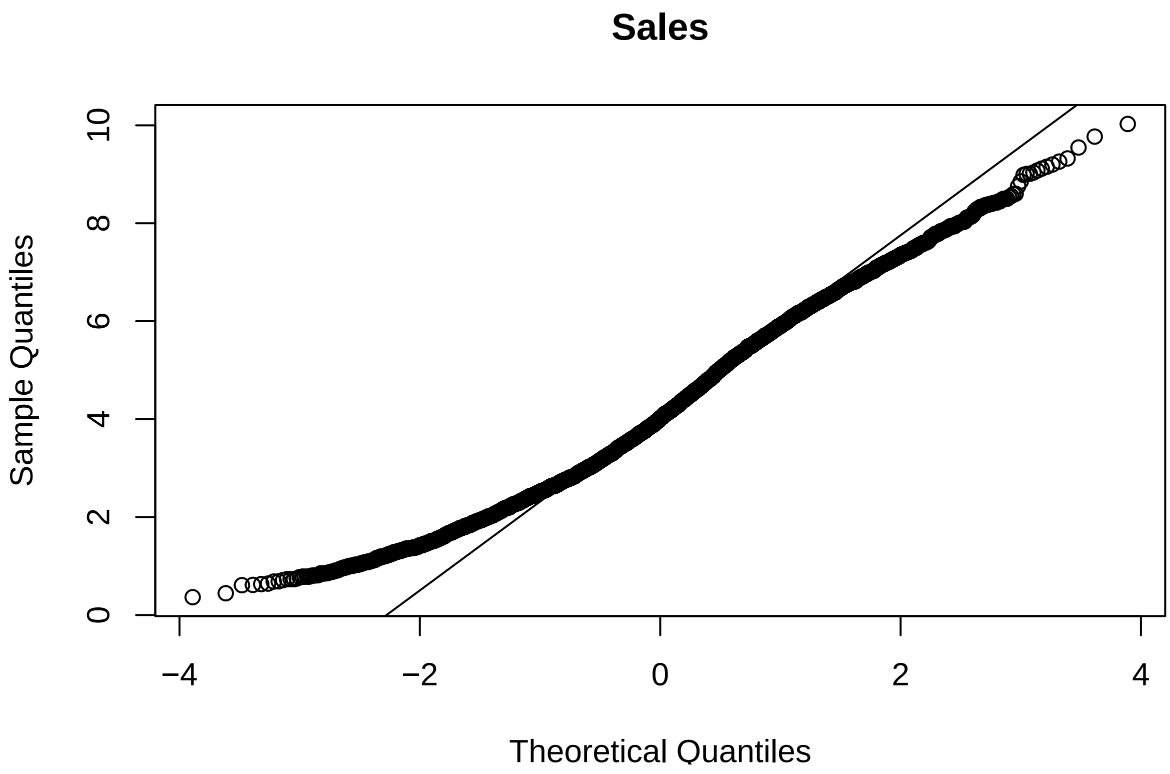
```
hist(superstore.reg.scale$Profit, col='pink', main='Profit',breaks = 50)
```



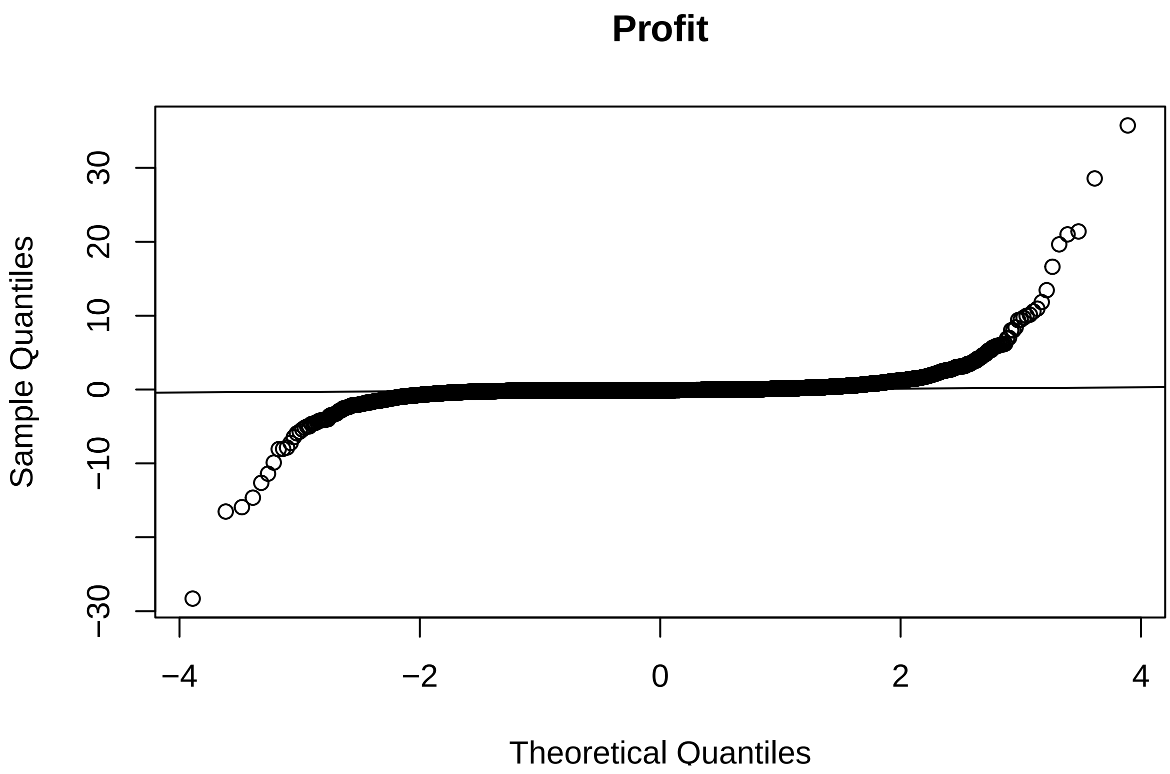
```
hist(superstore.reg.scale$Discount, col='pink', main='Discount',breaks = 50)
```



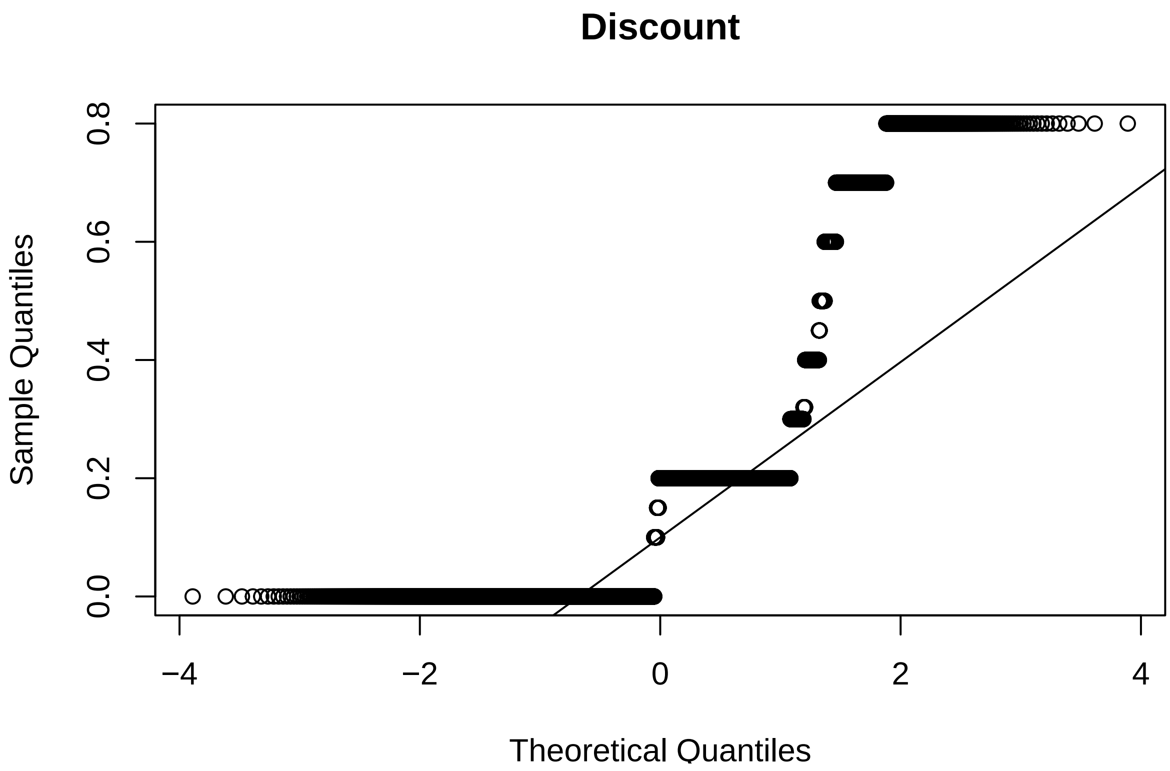
```
qqnorm(superstore.reg.scale$Sales, main='Sales')
qqline(superstore.reg.scale$Sales)
```



```
qqnorm(superstore.reg.scale$Profit, main='Profit')
qqline(superstore.reg.scale$Profit)
```



```
qqnorm(superstore.reg.scale$Discount, main='Discount')
qqline(superstore.reg.scale$Discount)
```



```
#Do regression and see what's going on
lm_fit<-lm(Sales~.,superstore.reg.scale)
summary(lm_fit)#not fit well
```

```
##
## Call:
## lm(formula = Sales ~ ., data = superstore.reg.scale)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0166 -0.9369 -0.1405  0.8618  9.5778
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.4330899  0.0723895  61.239   <2e-16 ***
## Quantity           0.2265526  0.0057265  39.562   <2e-16 ***
## Discount          -0.9312537  0.0650919 -14.307   <2e-16 ***
## Profit             0.2133245  0.0131134  16.268   <2e-16 ***
## SM_second_class   -0.0412880  0.0617171  -0.669   0.5035
## SM_standard_class -0.0587225  0.0569864  -1.030   0.3028
## SM_first_class    -0.0368149  0.0634614  -0.580   0.5619
## Seg_cons           0.0152750  0.0348711   0.438   0.6614
## Seg_corp           0.0243837  0.0379697   0.642   0.5208
## Reg_south         -0.0119671  0.0395295  -0.303   0.7621
## Reg_west          -0.0007855  0.0327915  -0.024   0.9809
## Reg_central       -0.0451756  0.0360555  -1.253   0.2103
## Cat_furniture     -0.0798258  0.0406361  -1.964   0.0495 *
## Cat_office        -1.5373295  0.0339349 -45.302   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.27 on 9980 degrees of freedom
## Multiple R-squared:  0.3631, Adjusted R-squared:  0.3623
## F-statistic: 437.7 on 13 and 9980 DF,  p-value: < 2.2e-16
```

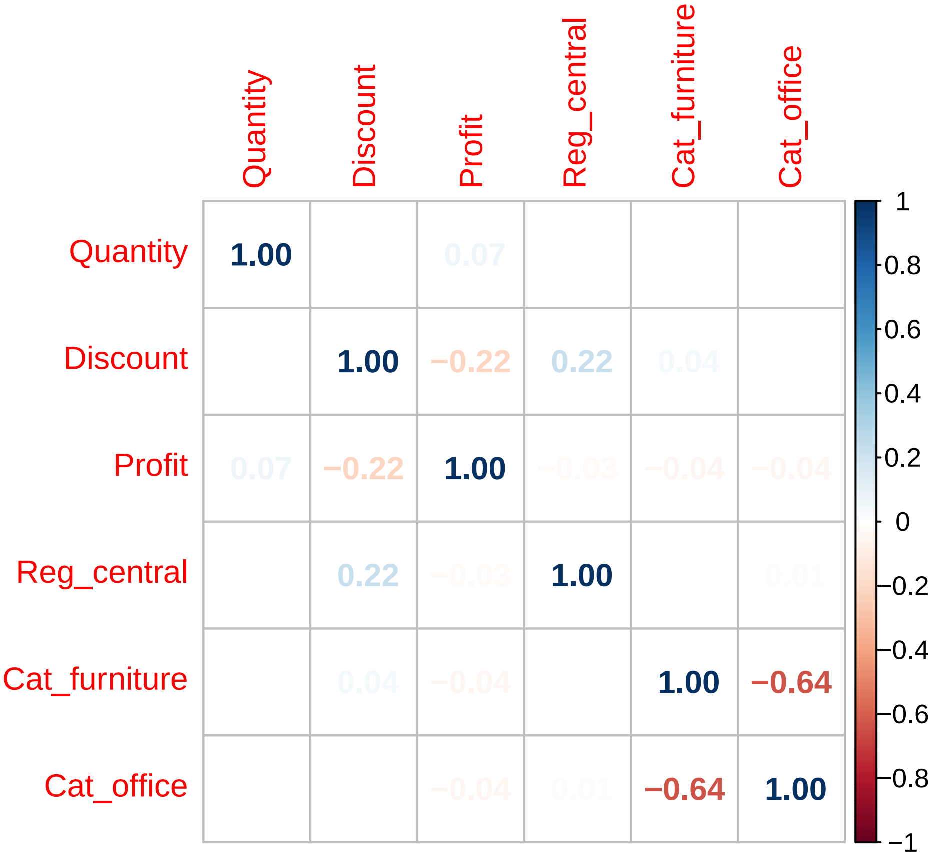
```r
reduced<-lm(Sales~1,superstore.reg.scale)
full<-lm(Sales~.,superstore.reg.scale)
step(reduced,scope=c(lower=reduced,upper=full),direction='forward',trace=F)
```

```
##
## Call:
## lm(formula = Sales ~ 1, data = superstore.reg.scale)
##
## Coefficients:
## (Intercept)
##       4.156
```

```r
step(full,scope=c(lower=reduced,upper=full),direction='backward',trace=F)
```

```
##
## Call:
## lm(formula = Sales ~ Quantity + Discount + Profit + Cat_furniture +
##     Cat_office, data = superstore.reg.scale)
##
## Coefficients:
##   (Intercept)       Quantity       Discount         Profit  Cat_furniture
##       4.39045        0.22646       -0.95207        0.21295       -0.07872
##    Cat_office
##      -1.53757
```

```
#after doing regression, we want to do a correlation analysis to see if all variables are uncorrelated:
corrplot::corrplot(cor(
  superstore.reg.scale%>%
    select(c(Quantity,Discount,Profit,Reg_central,Cat_furniture,Cat_office))),
  method='number')
```



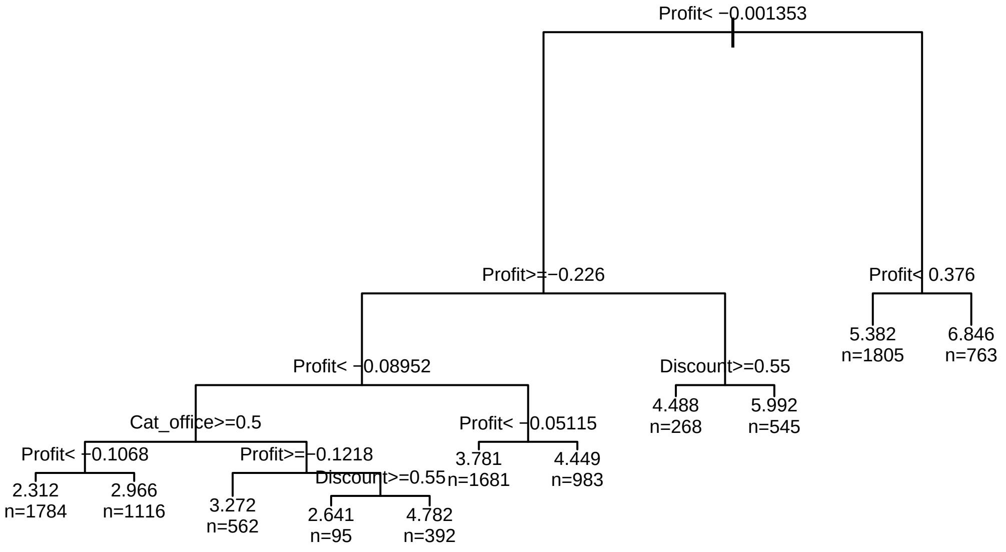
```
#we want to see what will the model perform if we delete one of cat_furniture or cat_office
summary(lm(Sales~Quantity+Discount+Profit+Cat_furniture,data=superstore.reg.scale))
```

```
##
## Call:
## lm(formula = Sales ~ Quantity + Discount + Profit + Cat_furniture,
##     data = superstore.reg.scale)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0346 -1.0806 -0.1099  0.9922 12.2631
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.236413   0.030206  107.15   <2e-16 ***
## Quantity       0.223038   0.006281   35.51   <2e-16 ***
## Discount      -1.024945   0.069281  -14.79   <2e-16 ***
## Profit         0.264531   0.014334   18.45   <2e-16 ***
## Cat_furniture  1.105334   0.034150   32.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.394 on 9989 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.2315
## F-statistic: 753.6 on 4 and 9989 DF,  p-value: < 2.2e-16
```

```
summary(lm(Sales~Quantity+Discount+Profit+Cat_office,data=superstore.reg.scale))
```

```
##
## Call:
## lm(formula = Sales ~ Quantity + Discount + Profit + Cat_office,
##     data = superstore.reg.scale)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9489 -0.9269 -0.1386  0.8712  9.6861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.349469   0.030956  140.50   <2e-16 ***
## Quantity     0.226357   0.005722   39.56   <2e-16 ***
## Discount    -0.957423   0.063068  -15.18   <2e-16 ***
## Profit       0.215052   0.013061   16.46   <2e-16 ***
## Cat_office  -1.495285   0.025980  -57.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.269 on 9989 degrees of freedom
## Multiple R-squared:  0.3626, Adjusted R-squared:  0.3624
## F-statistic:  1421 on 4 and 9989 DF,  p-value: < 2.2e-16
```

```
#Regression tree for scaled
library(rpart)
rt<-rpart(Sales~Quantity+Discount+Profit+Cat_office,superstore.reg.scale)
par(xpd = TRUE)
plot(rt, compress = TRUE)
text(rt, use.n = TRUE,cex=0.55)
```

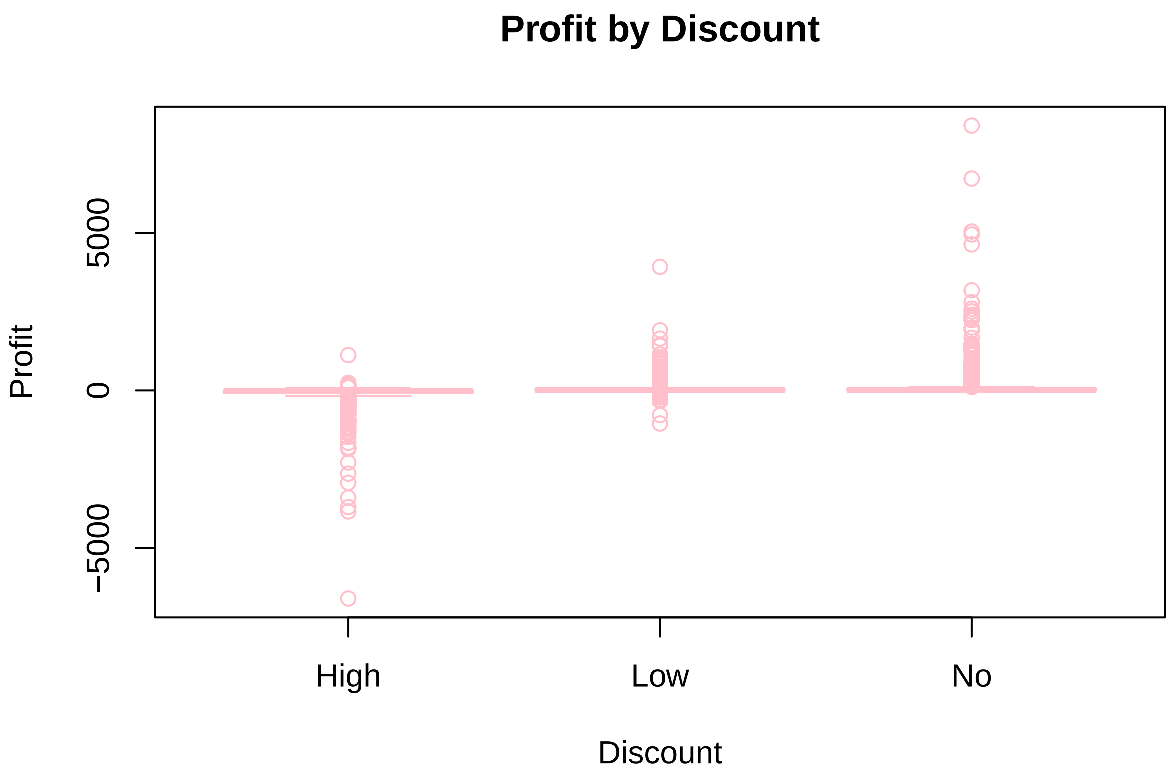


ANCOVA

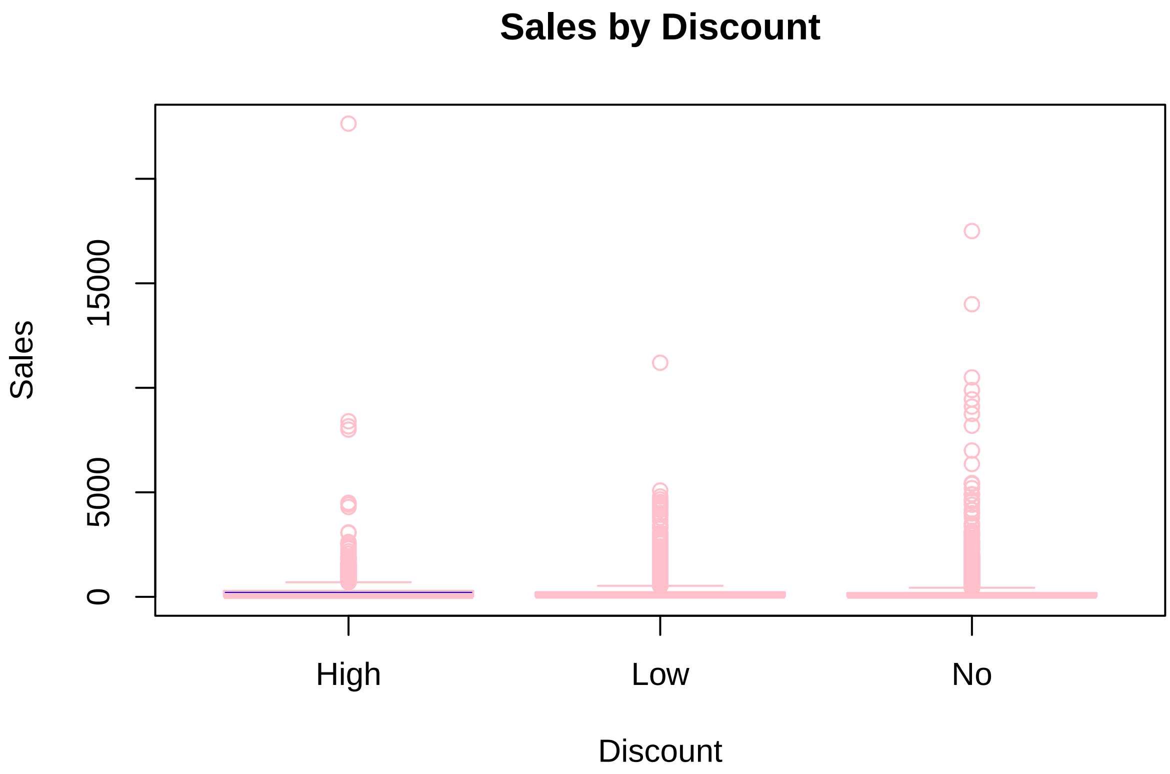
```
#transfer data
ancova_data<-data.clean %>%select(Sales,Profit,Discount)
idx<-1:9994
disc_No<-idx[ancova_data$Discount==0]
disc_Low<-idx[ancova_data$Discount>0&ancova_data$Discount<=0.2]
disc_High<-idx[ancova_data$Discount>0.2]
ancova_data$Discount[disc_No]<-'No'
ancova_data$Discount[disc_Low]<-'Low'
ancova_data$Discount[disc_High]<-'High'

rm(disc_No,disc_Low,disc_High,idx)
#EDA and summary
ancova_data %>%
  group_by(Discount) %>%
  summarise(mean_sales = mean(Sales),
            median_sales = median(Sales),
            sd_sales = sd(Sales),
            mean_profit = mean(Profit),
            median_profit = median(Profit),
            sd_profit = sd(Profit))
```

```
## # A tibble: 3 x 7
##   Discount mean_sales median_sales sd_sales mean_profit median_profit sd_profit
##   <chr>         <dbl>        <dbl>    <dbl>       <dbl>         <dbl>     <dbl>
## 1 High           260.         44.4     823.       -97.2        -18.2       328.
## 2 Low            223.         56.2     489.        26.5          6.74      118.
## 3 No             227.         53.6     650.        66.9         16.0       257.
```

```
#boxplot
boxplot(Profit ~ Discount,data = ancova_data,main = "Profit by Discount",
        xlab = "Discount",ylab = "Profit",col = "blue",border = "pink")
```



```
boxplot(Sales ~ Discount,data = ancova_data,main = "Sales by Discount",
        xlab = "Discount",ylab = "Sales",col = "blue",border = "pink")
```



```
#hypothesis testing:independence btw discount and sales, and equality of variance
summary(aov(Sales ~ Discount,data = ancova_data))
```

```
##               Df    Sum Sq Mean Sq F value Pr(>F)
## Discount       2 1.549e+06  774355   1.994  0.136
## Residuals   9991 3.880e+09  388357
```

```
#The p-value is 0.136 that is greater than 0.05, so Discount and Sales are independent to each other.

#Levene’s Test
leveneTest(Profit~Discount, data = ancova_data)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    2  39.936 < 2.2e-16 ***
##       9991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#The p-value of the test is 2.2e-16, which indicates that the variances among the groups are not equal.

#Fit analysis of covariance model ANCOVA
ancova_model <- aov(Profit ~ Discount + Sales, data = ancova_data)
Anova(ancova_model, type="III")
```

```
## Anova Table (Type III tests)
##
## Response: Profit
##                 Sum Sq   Df F value    Pr(>F)
## (Intercept)   28390734    1  724.88 < 2.2e-16 ***
## Discount      31268197    2  399.18 < 2.2e-16 ***
## Sales        128032014    1 3268.96 < 2.2e-16 ***
## Residuals    391267852 9990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

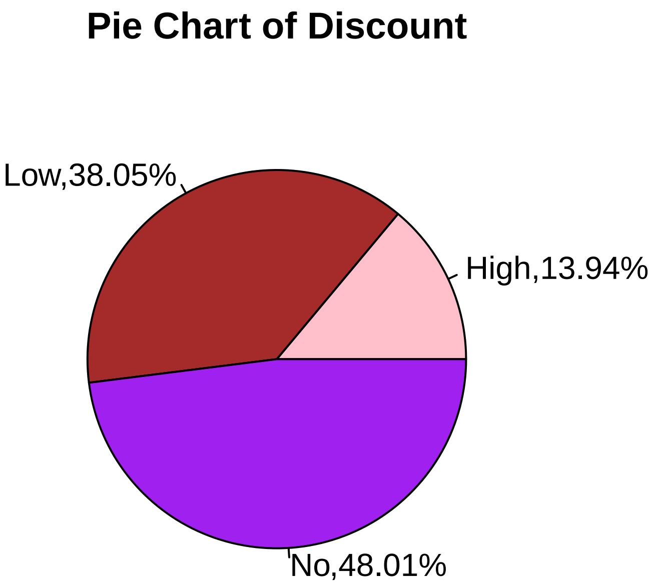
From this result, we can conclude that no matter we control discount or sales still, the other variable is significant in this model, which indicates that they will significantly contribute to the model. From two significant p-value of Discount and Sales, we may say that: (1) different Discount will influence the relationship between Sales and Profit, and (2) sales does have relationship to profit.

```
lm_ancova<-lm(Profit~Sales+Sales:Discount,data=ancova_data)
summary(lm_ancova)
```

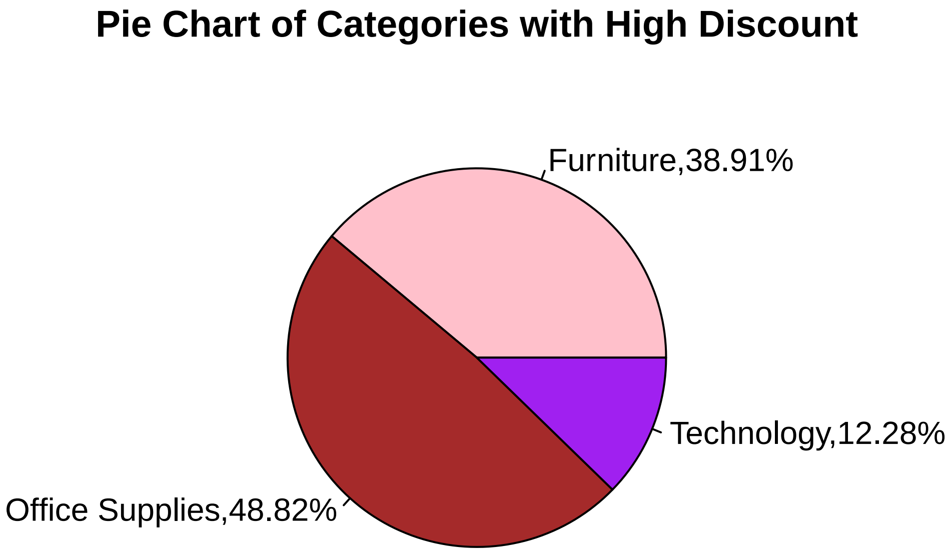
```
##
## Call:
## lm(formula = Profit ~ Sales + Sales:Discount, data = ancova_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5667.7     2.1    17.1    20.9  2844.6
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -18.086540   1.488097  -12.15   <2e-16 ***
## Sales               -0.203156   0.004339  -46.82   <2e-16 ***
## Sales:DiscountLow    0.372645   0.006042   61.68   <2e-16 ***
## Sales:DiscountNo    0.568423   0.005202  109.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138.8 on 9990 degrees of freedom
## Multiple R-squared:  0.6493, Adjusted R-squared:  0.6492
## F-statistic:  6165 on 3 and 9990 DF,  p-value: < 2.2e-16
```

From this linear model with interaction of Sales and Discount, we may say that different discount will contribute different relationship between Sales and Profit. Here, when discount is 0, the coefficient is 0.5684-0.2032=0.3652; when discount is low, the coefficient is 0.3726-0.2032=0.1694; when discount is high, the coefficient is -0.2032.

```
tableone<-data.frame(ancova_data$Discount, data.clean$Category)
slices <- table(tableone$ancova_data.Discount)
count<-slices/sum(slices)*100
lbls <- c("High", "Low", "No")
pie(slices, col=c('pink','brown','purple'),labels = paste0(lbls,',',round(count,2),'%'),
    main="Pie Chart of Discount")
```



```
tableone_high<-tableone %>% filter(ancova_data$Discount=='High')
slices <- table(tableone_high$data.clean.Category)
count<-slices/sum(slices)*100
lbls <- c("Furniture", "Office Supplies", "Technology")
pie(slices, col=c('pink','brown','purple'),labels = paste0(lbls,',',round(count,2),'%'),
    main="Pie Chart of Categories with High Discount")
```



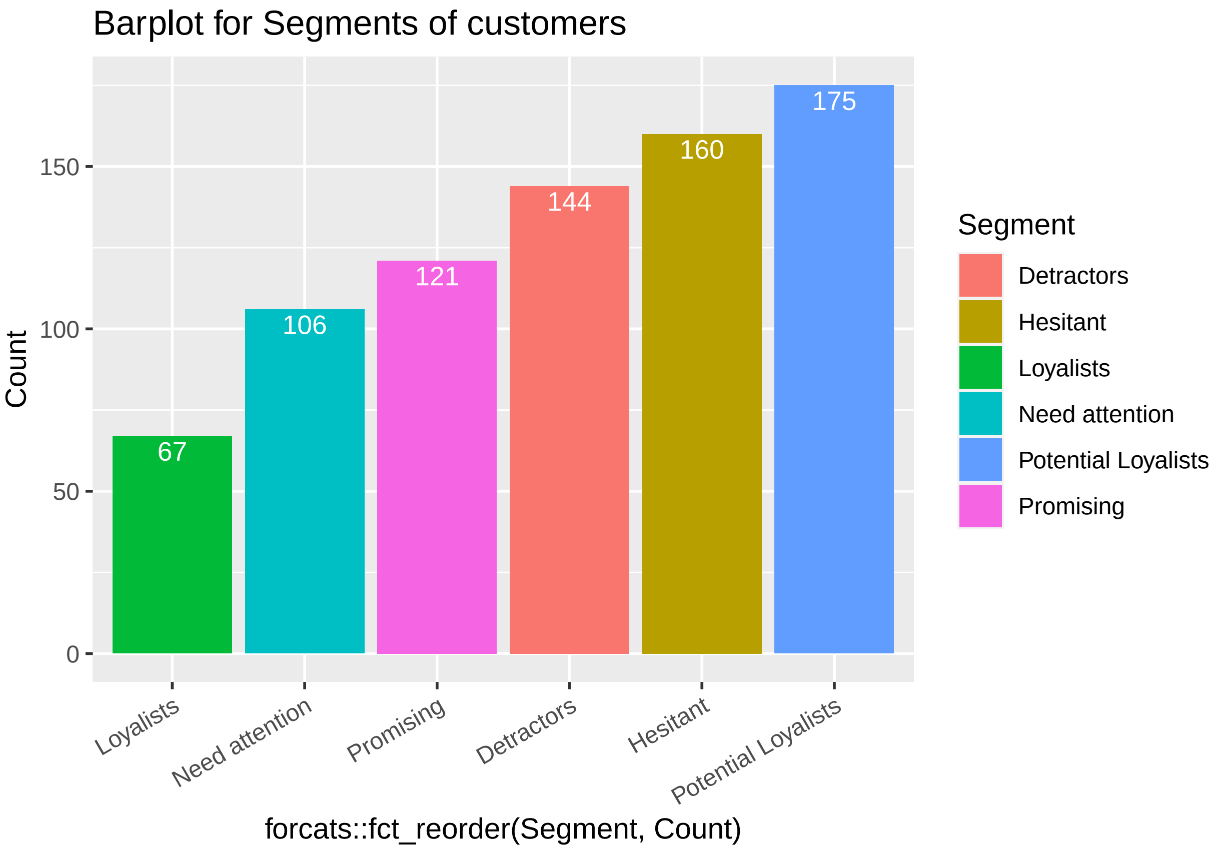
RFM analysis

```r
#create RFM table
RFM.data<-data.clean[lubridate::year(data.clean$Order.Date)%in%c(2016,2017),]
RFM.table<-RFM.data%>%
  group_by(Customer.ID)%>%
  summarize(Recency=max(Order.Date),Frequency=n(),Monetary=sum(Sales))
RFM.table$Recency<-as.numeric((as.Date('2017-12-31')-RFM.table$Recency))
summary(RFM.table)
```

```
##  Customer.ID           Recency        Frequency         Monetary
##  Length:773         Min.   :  1.0   Min.   : 1.000   Min.   :    2.81
##  Class :character   1st Qu.: 30.0   1st Qu.: 4.000   1st Qu.:  519.76
##  Mode  :character   Median : 73.0   Median : 7.000   Median : 1194.96
##                     Mean   :129.3   Mean   : 7.631   Mean   : 1736.64
##                     3rd Qu.:163.0   3rd Qu.:10.000   3rd Qu.: 2216.88
##                     Max.   :720.0   Max.   :27.000   Max.   :18344.05
```

```r
#Scored data
rfm_scored<-RFM.table
rfm_scored$r_score <- rep(0,773)
rfm_scored$r_score[RFM.table$Recency >= 163.0] <- 1
rfm_scored$r_score[RFM.table$Recency >=73.0 & RFM.table$Recency < 163.0] <- 2
rfm_scored$r_score[RFM.table$Recency >=30.0 & RFM.table$Recency < 73.0] <- 3
rfm_scored$r_score[RFM.table$Recency < 30.0] <- 4

rfm_scored$f_score <- rep(0,773)
rfm_scored$f_score[RFM.table$Frequency >= 10.000] <- 1
rfm_scored$f_score[RFM.table$Frequency >=7.000 & RFM.table$Frequency < 10.000] <- 2
rfm_scored$f_score[RFM.table$Frequency >=4.000 & RFM.table$Frequency < 7.000] <- 3
rfm_scored$f_score[RFM.table$Frequency < 4.000] <- 4

rfm_scored$m_score <- rep(0,773)
rfm_scored$m_score[RFM.table$Monetary >= 2216.88] <- 1
rfm_scored$m_score[RFM.table$Monetary >=1194.96 & RFM.table$Monetary < 2216.88] <- 2
rfm_scored$m_score[RFM.table$Monetary >=519.76 & RFM.table$Monetary < 1194.96] <- 3
```

```r
rfm_scored$m_score[RFM.table$Monetary <519.76] <- 4

rfm_scored<-rfm_scored %>%
  mutate(RFM_score=r_score*100+f_score*10+m_score) %>%
  select(Customer.ID,r_score,f_score,m_score,RFM_score)

#segments
rfm_scored$Segment <- "0"
rfm_scored$Segment[which(rfm_scored$RFM_score
  %in% c(444,434,443, 344, 442, 244, 424, 441))] <-"Loyalists"
rfm_scored$Segment[which(rfm_scored$RFM_score
  %in% c(332,333,342,343,334,412,413,414,431,432,441,421,422,423,424,433))]<- "Potential Loyalists"
rfm_scored$Segment[which(rfm_scored$RFM_score
  %in% c(233,234, 241,311, 312, 313,314,321,322,323,324, 331,  341))] <- "Promising"
rfm_scored$Segment[which(rfm_scored$RFM_score
  %in% c(124, 133, 134, 142, 143, 144, 214,224,234, 242, 243, 232 ))] <- "Hesitant"
rfm_scored$Segment[which(rfm_scored$RFM_score
  %in% c(122, 123,131 ,132, 141, 212, 213, 221, 222, 223, 231 ))] <- "Need attention"
rfm_scored$Segment[which(rfm_scored$RFM_score
  %in% c(111, 112, 113, 114, 121, 131, 211, 311, 411 ))] <-"Detractors"

#plot of segments
rfm_scored%>%
  group_by(Segment)%>%
  summarize(Count=n())%>%
  ggplot(aes(x = forcats::fct_reorder(Segment, Count),y=Count,fill = Segment)) +
  geom_bar(stat='identity')+
  geom_text(aes(label=Count),nudge_y=-.5,color="white",size = 3.5,vjust=1.2)+
  theme(axis.text.x=element_text(angle=30,hjust=1))+
  labs(title = "Barplot for Segments of customers")
```


Barplot for Segments of customers
Count
0
50
100
150
67
106
121
144
160
175
Loyalists
Need attention
Promising
Detractors
Hesitant
Potential Loyalists
forcats::fct_reorder(Segment, Count)
Segment
Detractors
Hesitant
Loyalists
Need attention
Potential Loyalists
Promising