Superstore Sales

Daphney

2024-07-21

Overview

To better understand the factors influencing sales I conducted a regression analysis and used the regression models to forecast sales and demand.

EDA

```
#Read in data
data <- read.csv("Superstore.csv")
head(data, 3)
```

```
##   Row.ID       Order.ID Order.Date  Ship.Date    Ship.Mode Customer.ID
## 1      1 CA-2016-152156  11/8/2016 11/11/2016 Second Class    CG-12520
## 2      2 CA-2016-152156  11/8/2016 11/11/2016 Second Class    CG-12520
## 3      3 CA-2016-138688  6/12/2016  6/16/2016 Second Class    DV-13045
##     Customer.Name   Segment       Country        City      State
Postal.Code
## 1     Claire Gute  Consumer United States   Henderson   Kentucky
42420
## 2     Claire Gute  Consumer United States   Henderson   Kentucky
42420
## 3 Darrin Van Huff Corporate United States Los Angeles California
90036
##   Region      Product.ID        Category Sub.Category
## 1  South FUR-BO-10001798       Furniture    Bookcases
## 2  South FUR-CH-10000454       Furniture       Chairs
## 3   West OFF-LA-10000240 Office Supplies       Labels
##                                                 Product.Name  Sales
Quantity
## 1                           Bush Somerset Collection Bookcase 261.96
2
## 2 Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.94
3
## 3   Self-Adhesive Address Labels for Typewriters by Universal  14.62
2
##   Discount   Profit
## 1        0  41.9136
## 2        0 219.5820
## 3        0   6.8714
```

```
#Load libraries
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: ggplot2

## Loading required package: lattice

library(ggplot2)
library(car)

## Loading required package: carData

library(MASS)

## Warning: package 'MASS' was built under R version 4.3.3
```

EDA

```
str(data)

## 'data.frame':    9994 obs. of  21 variables:
##  $ Row.ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Order.ID     : chr  "CA-2016-152156" "CA-2016-152156" "CA-2016-138688"
"US-2015-108966" ...
##  $ Order.Date   : chr  "11/8/2016" "11/8/2016" "6/12/2016" "10/11/2015"
...
##  $ Ship.Date    : chr  "11/11/2016" "11/11/2016" "6/16/2016" "10/18/2015"
...
##  $ Ship.Mode    : chr  "Second Class" "Second Class" "Second Class"
"Standard Class" ...
##  $ Customer.ID  : chr  "CG-12520" "CG-12520" "DV-13045" "SO-20335" ...
##  $ Customer.Name: chr  "Claire Gute" "Claire Gute" "Darrin Van Huff" "Sean
O'Donnell" ...
##  $ 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" ...
##  $ Product.ID   : chr  "FUR-BO-10001798" "FUR-CH-10000454" "OFF-LA-
10000240" "FUR-TA-10000577" ...
##  $ Category     : chr  "Furniture" "Furniture" "Office Supplies"
"Furniture" ...
##  $ Sub.Category : chr  "Bookcases" "Chairs" "Labels" "Tables" ...
##  $ Product.Name : chr  "Bush Somerset Collection Bookcase" "Hon Deluxe
Fabric Upholstered Stacking Chairs, Rounded Back" "Self-Adhesive Address
Labels for Typewriters by Universal" "Bretford CR4500 Series Slim Rectangular
Table" ...
##  $ 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 ...
```

```
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 :56431
##                                                           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
```

```r
#Remove unecessary columns
data <- data[-c(1,2,3,4,6,7,9,10,14)]
head(data, 2)
```

```
##       Ship.Mode  Segment    State Postal.Code Region  Category Sub.Category
## 1 Second Class Consumer Kentucky       42420  South Furniture    Bookcases
## 2 Second Class Consumer Kentucky       42420  South Furniture       Chairs
##                                                 Product.Name  Sales
Quantity
## 1                          Bush Somerset Collection Bookcase 261.96
2
## 2 Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.94
3
##   Discount   Profit
## 1        0  41.9136
## 2        0 219.5820
```

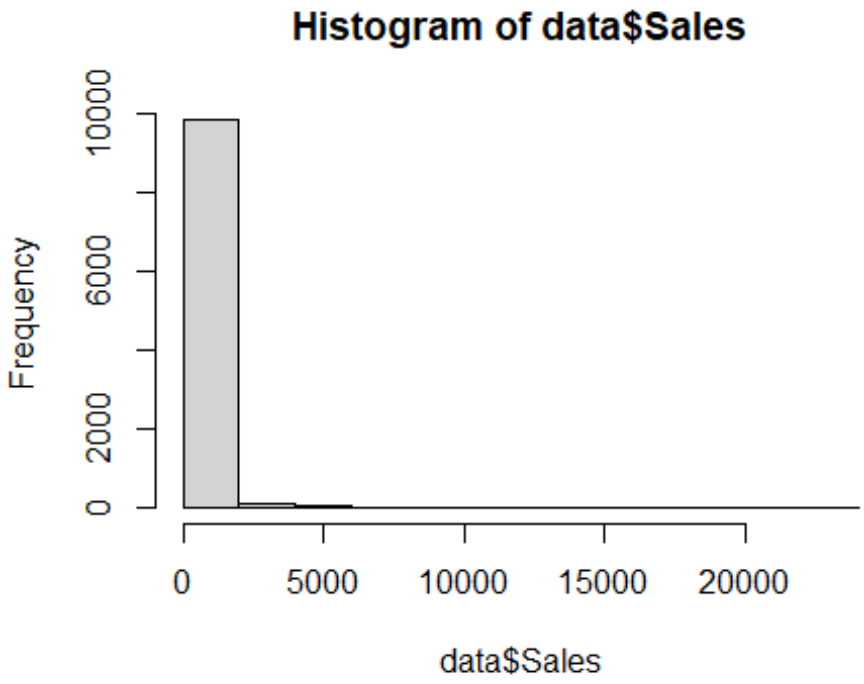
```r
#Check for missing values
any(is.na(data))
```

```
## [1] FALSE
```

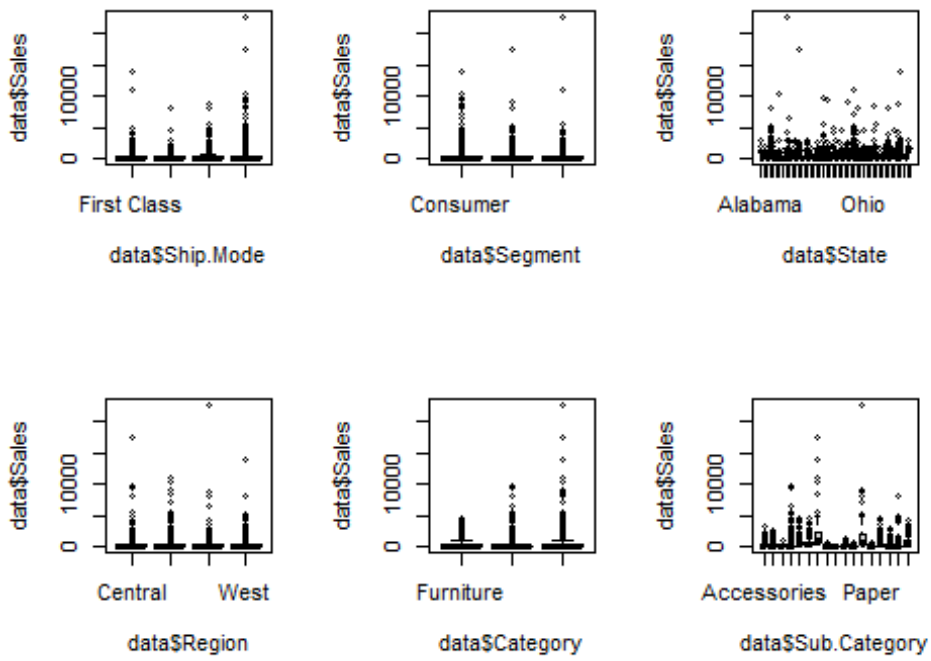
```r
sum(is.na(data))
```

```
## [1] 0
```

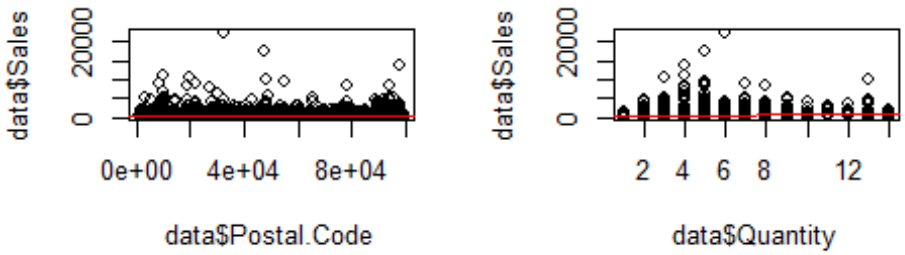
```r
#Convert categorical data type
data$Ship.Mode <-as.factor(data$Ship.Mode)
data$Segment <-as.factor(data$Segment)
data$State <-as.factor(data$State)
data$Region <-as.factor(data$Region)
data$Category <-as.factor(data$Category)
data$Sub.Category <-as.factor(data$Sub.Category)
data$Product.Name <-as.factor(data$Product.Name)

# view the distribution of Sales
hist(data$Sales)
```

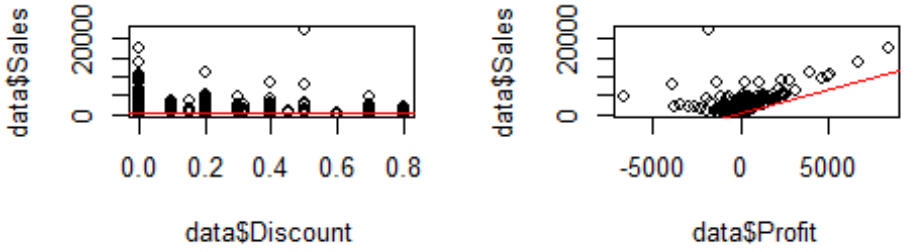



```r
#boxplot of categorical data
par(mfrow=c(2,3))
boxplot(data$Sales ~ data$Ship.Mode)
boxplot(data$Sales ~ data$Segment)
boxplot(data$Sales ~ data$State)
boxplot(data$Sales ~ data$Region)
boxplot(data$Sales ~ data$Category)
boxplot(data$Sales ~ data$Sub.Category)
```




```r
#scatterplots of numerical data
par(mfrow=c(2,2))
plot(data$Postal.Code, data$Sales)
abline(lm(Sales ~ Postal.Code, data = data), col = "red")

plot(data$Quantity, data$Sales)
abline(lm(Sales ~ Quantity, data = data), col = "red")

plot(data$Discount, data$Sales)
abline(lm(Sales ~ Discount, data = data), col = "red")

plot(data$Profit, data$Sales)
abline(lm(Sales ~ Profit, data = data), col = "red")
```






```
#Correlation Matrix
cor(data[c(4,9,10,11,12)])
```

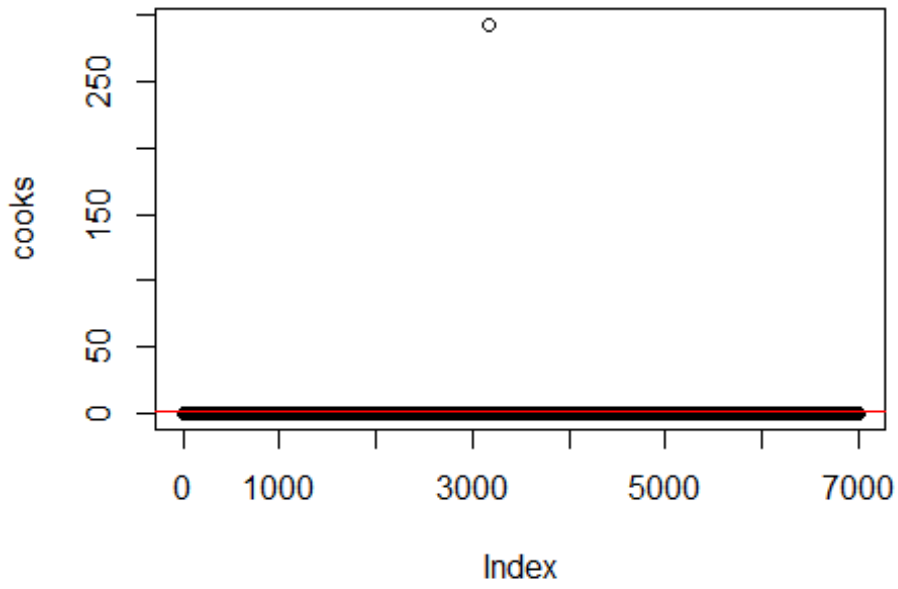
```
##             Postal.Code       Sales    Quantity    Discount      Profit
## Postal.Code  1.00000000 -0.02385377 0.01276071  0.05844306 -0.02996119
## Sales       -0.02385377  1.00000000 0.20079477 -0.02819012  0.47906435
## Quantity     0.01276071  0.20079477 1.00000000  0.00862297  0.06625319
## Discount     0.05844306 -0.02819012 0.00862297  1.00000000 -0.21948746
## Profit      -0.02996119  0.47906435 0.06625319 -0.21948746  1.00000000
```

Findings: Sales and Profit have a moderately strong correlation. All other variables appear to show a weak correlation between dependent variables. No strong correlations observed between the predicting variables. No evidence of multicolinearity.

Analysis:

```
#Split data
set.seed(123)
train_index <- createDataPartition(data$Sales, p=0.7, list = FALSE)
train_data <- data[train_index,]
test_data <- data[-train_index,]

#Model fitting
full_model <- lm(Sales ~ ., data = train_data)
#summary(full_model)

#Check for outliers
cooks <- cooks.distance(full_model)
```

```r
plot(cooks)
abline(h=1, col= "red")
```




```r
which.max(cooks)
```

```
## 4489
## 3171
```

An outlier is detected. There is a point significantly above the threshold line of h=1.

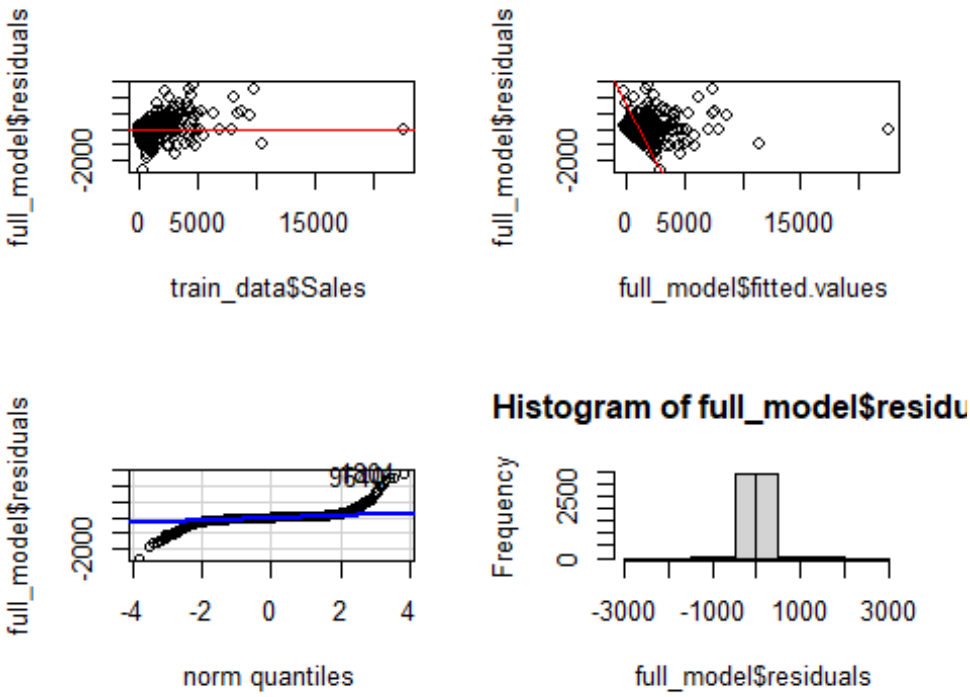
Residual Analysis

```r
#Checking Homoscedascity Assumption:
par(mfrow=c(2,2))
plot(train_data$Sales, full_model$residuals)
abline(h=0, col="red")

#Checking homoscedascity and Linearity Assumption:
plot(full_model$fitted.values, full_model$residuals)
abline(full_model, col = "red")
```

```
## Warning in abline(full_model, col = "red"): only using the first two of
1881
## regression coefficients
```

```r
#Checking Normality Assumption:
qqPlot(full_model$residuals)
```

```
## 1804 9640
## 1264 6761

hist(full_model$residuals)
```



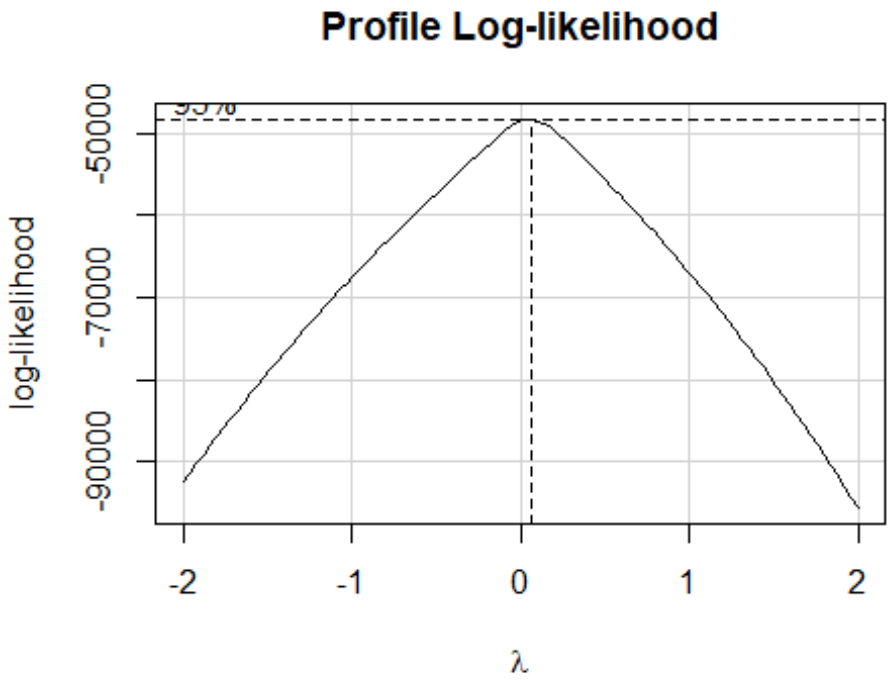
Findings: The residuals should be randomly scattered across the zero line with no clear pattern, however, the residuals appear to be more spread out at larger values of sales, indicating increased variance as sales increases. Potential heteroscedascity present.

The residuals are not randomly distributed. There is a noticeable funnel shape, suggesting heteroscedasticity and non-linearity.

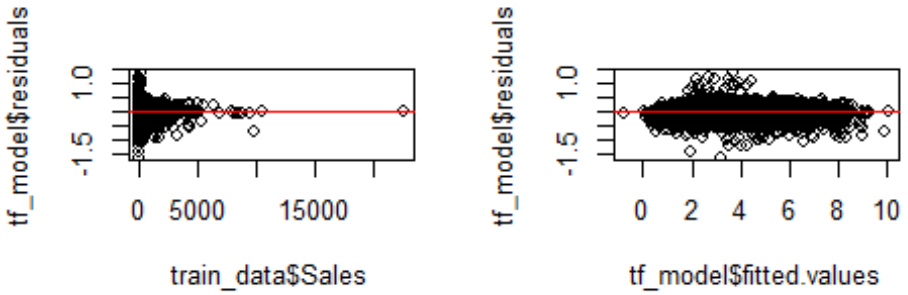
There are significant deviations from the line, especially at the tails. This indicates that the residuals are not normally distributed. The histogram shows a high concentration around zero with a sharp drop-off on both sides, indicating they are not normally distrubted.

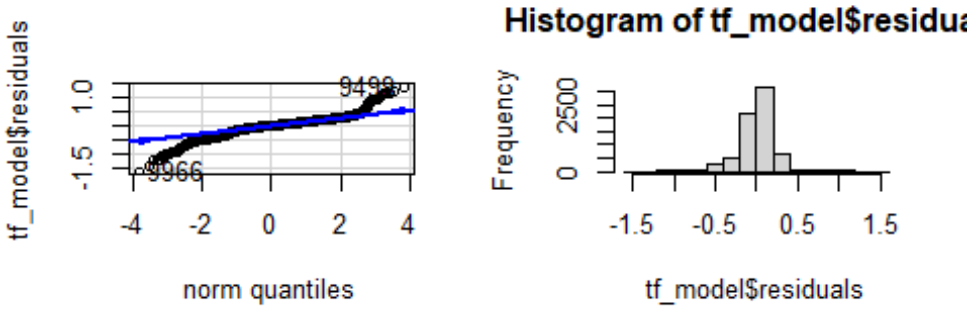
The assumptions do not hold in this case so the model's predictions may not be reliable, so this issue needs to be addressed.

```
#Transformation
#Box-cox to find optimal lambda
bc <- boxCox(full_model)
```




```r
#Log transformation
transformed <- log(train_data$Sales)
tf_model <- lm(transformed ~ ., data = train_data)

#Recheck Assumptions to verify
par(mfrow = c(2,2))
plot(train_data$Sales, tf_model$residuals)
abline(h=0, col="red")
plot(tf_model$fitted.values, tf_model$residuals)
abline(h=0, col = "red")
qqPlot(tf_model$residuals)

## 9966 9499
## 6977 6657

hist(tf_model$residuals)
```



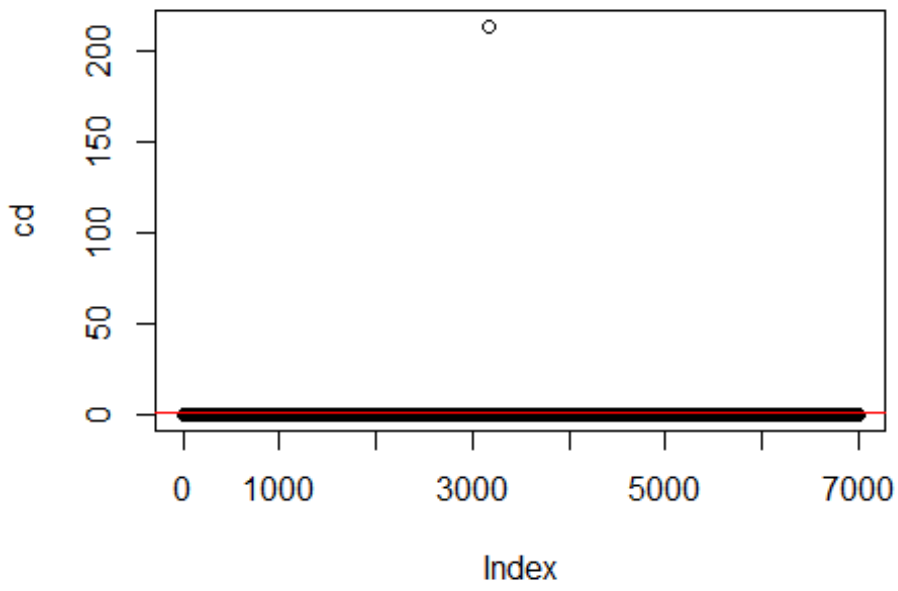


Analysis Significant Improvement in the plots after the transformation.

The plot shows a more consistent spread of residuals around the horizontal line, indicating an improvement in homoscedasticity. The residuals are now more randomly scattered around the horizontal line with less pronounced patterns, indicating an improvement in both linearity and homoscedasticity.

For normality, The histogram shows a bell shaped distrubtion and is not skewed. The residuals in the QQ plot follow the line more closely but there are still some deviations in the tails. Not perfect normality but the deviations are not significant enough to reject the normality assumption.

```
#Outliers
cd <- cooks.distance(tf_model)
plot(cd)
abline(h=1, col= "red")
```



There are no significant outliers at a cooks distance threshold of 1. The influential outlier is no longer present, indicating that the transformation has successfully reduced the influence of outliers on the model.

```
#Compare Models
summary(full_model)$adj.r.squared

## [1] 0.8922954

summary(tf_model)$adj.r.squared

## [1] 0.9797957
```

The transformed model explains more variability than the other model. It explains 98% of the variability in sales.

Prediction

```
#Prediction
train_data1 = train_data[-c(3,8)]
test_data1 = test_data[-c(3,8)]
reduced_model <- lm(Sales ~ ., data = train_data1)
#summary(reduced_model)

prediction <- predict(reduced_model, newdata = test_data1)
actual_values <- test_data1$Sales
mse <-mean((actual_values - prediction)^2)
rmse = sqrt(mse)
cat("R^2 reduced model:",summary(reduced_model)$adj.r.squared)
```

```
## R^2 reduced model: 0.3447719

cat("R^2 full model:",summary(full_model)$adj.r.squared)

## R^2 full model: 0.8922954

cat("R^2 transformed model:",summary(tf_model)$adj.r.squared)

## R^2 transformed model: 0.9797957

#Predict scenario
new_data <- data.frame(
  Ship.Mode = "Standard Class",
  Segment = "Home Office",
  State  = "Florida",
  Postal.Code = 32216,
  Region = "South",
  Category = "Technology",
  Sub.Category = "Machines",
  Product.Name = "
HTC One Mini",
  Quantity = 6,
  Discount = 0.50,
  Profit = -1811.0784
)
new_data$Product.Name <- factor(new_data$Product.Name, levels =
levels(train_data$Product.Name))
prediction <- predict(full_model, newdata = new_data)
```