Investment Data Analysis Report

Table of Contents

Table of Images

Introduction

An investment is anything that is acquired with the goal of profiting financially. An investment always entails the use of some resource today, such as time, money, effort, or an item, with the hope of receiving a return that is higher than the initial investment. Given that the objective is to make money, there are risk considerations.Generally speaking, investment risk is the possibility of losing the money you invest. There is also a danger that you won't earn as much money as you expect. As risk is reduced, the potential gains are also reduced. The level of risk raises the potential profits[1][2].
The provided dataset includes information gathered from a haphazard sample of businesses in the manufacturing and retail sectors.Any person or other organization (such as a company or mutual fund) who invests money in the hopes of making a profit is referred to as an investor.A non-professional investor who buys and sells securities or funds that comprise a variety of assets, such as mutual funds and exchange traded funds, is referred to as a retail investor, sometimes known as an individual investor.Industry investors are any of the following who have a beneficial interest in investments in any Dealer Member or holding company of a Dealer Member corporation.The study's objective is to statistically analyze the supplied data in order to identify the best investment plan and the variables that produce the most accurate investment measurements.

Data and Methods

The database offered contains a random sample of businesses from the manufacturing and retail industries. Each of these data sets underwent a separate exploratory data analysis, which is described below.

Exploratory Data Analysis on Retail Data

The retail dataset includes 100 entries for 12 variables, including MktPrice, TotMktCap, DivYield, PERatio, Beta, TotalSales17 and TotalSales18, CapEmp, Dividend, MktBook, Ret17 and Ret18. All of the variables in the dataset are numerical and can be classified as continuous quantitative variables.The summary of the data set is taken into account as the first phase of analysis, and the outcome is shown below.

```
   MktPrice        TotMktCap         DivYield         PERatio
Min.   : 18.02   Min.   : 43.85   Min.   : 2.520   Min.   : 14.12
1st Qu.: 54.80   1st Qu.: 65.20   1st Qu.: 8.595   1st Qu.: 26.14
Median :101.52   Median : 81.56   Median : 9.910   Median : 30.66
Mean   :156.45   Mean   :131.35   Mean   :10.779   Mean   : 37.77
3rd Qu.:193.88   3rd Qu.:153.12   3rd Qu.:13.057   3rd Qu.: 40.17
Max.   :865.69   Max.   :468.99   Max.   :24.470   Max.   :139.47
                                                  NA's   :26
     Beta         TotalSales17      TotalSales18        CapEmp
Min.   :0.1000   Min.   :  92.48   Min.   : 110.0   Min.   :  24.06
1st Qu.:0.7400   1st Qu.: 134.71   1st Qu.: 124.0   1st Qu.:  41.64
Median :0.9400   Median : 158.04   Median : 135.2   Median :  53.09
Mean   :0.9564   Mean   : 235.42   Mean   : 189.1   Mean   : 118.01
3rd Qu.:1.2175   3rd Qu.: 241.35   3rd Qu.: 173.2   3rd Qu.:  96.78
Max.   :1.6900   Max.   :1250.25   Max.   :1420.8   Max.   :1786.78
NA's   :2
   Dividend        MktBook           Ret17           Ret18
Min.   : 7.03   Min.   : 3.610   Min.   :-34.94   Min.   :-72.31
1st Qu.:14.49   1st Qu.: 7.135   1st Qu.: 46.06   1st Qu.:-29.65
Median :19.11   Median : 9.325   Median : 75.23   Median : -2.17
Mean   :21.59   Mean   :10.454   Mean   : 76.47   Mean   :  1.27
3rd Qu.:25.27   3rd Qu.:11.512   3rd Qu.:105.16   3rd Qu.: 28.00
Max.   :94.74   Max.   :37.510   Max.   :223.00   Max.   :154.81
NA's   :1
```

Image 1

It is clear from image 1 that the dataset's variables PERatio, Beta, and Dividend all contain null values. The solution to these null values is to apply data imputation techniques.

In general, there are 3 data imputation techniques that are frequently used. These are the average, median, and mode. The mean of the numerical column data is used to replace null values when the data is normally distributed. The median was applied if there were any outliers in the data. The mode is chosen when there are more instances of a particular value or when a value is more prevalent. Another approach to dealing with missing values is to remove the rows from the dataset.

Let's analyze the distribution of the data across these columns and the number of outliers.We can choose between mean value imputation and median imputation because the data includes quantitative variables and eliminating the null values will lead to smaller data.

The dataset's normality of the variables and outlier identification are both tested in order to determine which of the aforementioned methods is the best.The test statistic(A) and matching test statistic p-value are returned as the test result (p-value). The data do indeed follow a normal distribution, which is the null hypothesis for the A-D test. Therefore, if our test's p-value falls below our designated level of significance (popular options are 0.10, 0.05, and 0.01), we can reject the null hypothesis and draw the conclusion that there is enough evidence to indicate that our data do not follow a normal distribution[8].

```
        Anderson-Darling normality test

data:  retail_data$MktPrice
A = 7.6082, p-value < 2.2e-16

        Anderson-Darling normality test

data:  retail_data$TotMktCap
A = 10.335, p-value < 2.2e-16

        Anderson-Darling normality test

data:  retail_data$DivYield
A = 0.90235, p-value = 0.0206

        Anderson-Darling normality test

data:  retail_data$PERatio
A = 7.5178, p-value < 2.2e-16

        Anderson-Darling normality test

data:  retail_data$Beta
A = 0.14921, p-value = 0.9626

        Anderson-Darling normality test

data:  retail_data$TotalSales17
A = 11.573, p-value < 2.2e-16

        Anderson-Darling normality test

data:  retail_data$TotalSales18
A = 19.935, p-value < 2.2e-16

        Anderson-Darling normality test

data:  retail_data$CapEmp
A = 20.766, p-value < 2.2e-16

        Anderson-Darling normality test

data:  retail_data$Dividend
A = 4.8833, p-value = 3.593e-12

        Anderson-Darling normality test

data:  retail_data$MktBook
A = 6.6459, p-value < 2.2e-16

        Anderson-Darling normality test

data:  retail_data$Ret17
A = 0.28438, p-value = 0.6232

        Anderson-Darling normality test

data:  retail_data$Ret18
A = 0.42935, p-value = 0.3036
```

Image 2

Since the Beta variable was found to be normally distributed by the Anderson-Darling test(image 2), the mean may be used in place of the null values.Given that PERatio and Dividend contain outliers, as seen by the box plot of the variables, the median will be the most effective data imputation strategy for these variables.

The next images(image 3,image 4,image 5)[5] show the QQ plot of the 10 outlier-containing variables before (red) and after (green) eliminating outliers and reducing the dataset to 50 rows and 12 variables.

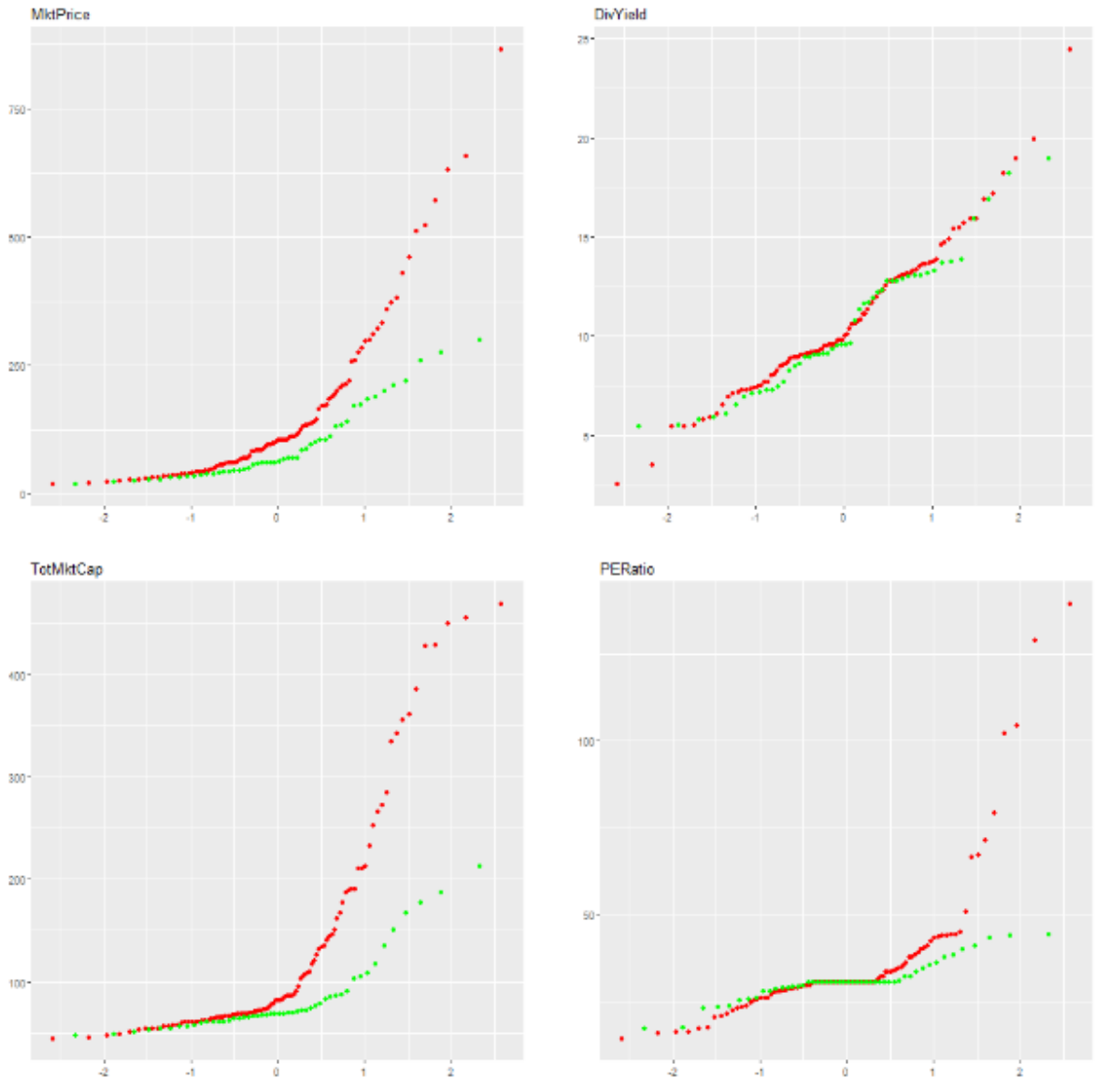

Image 3

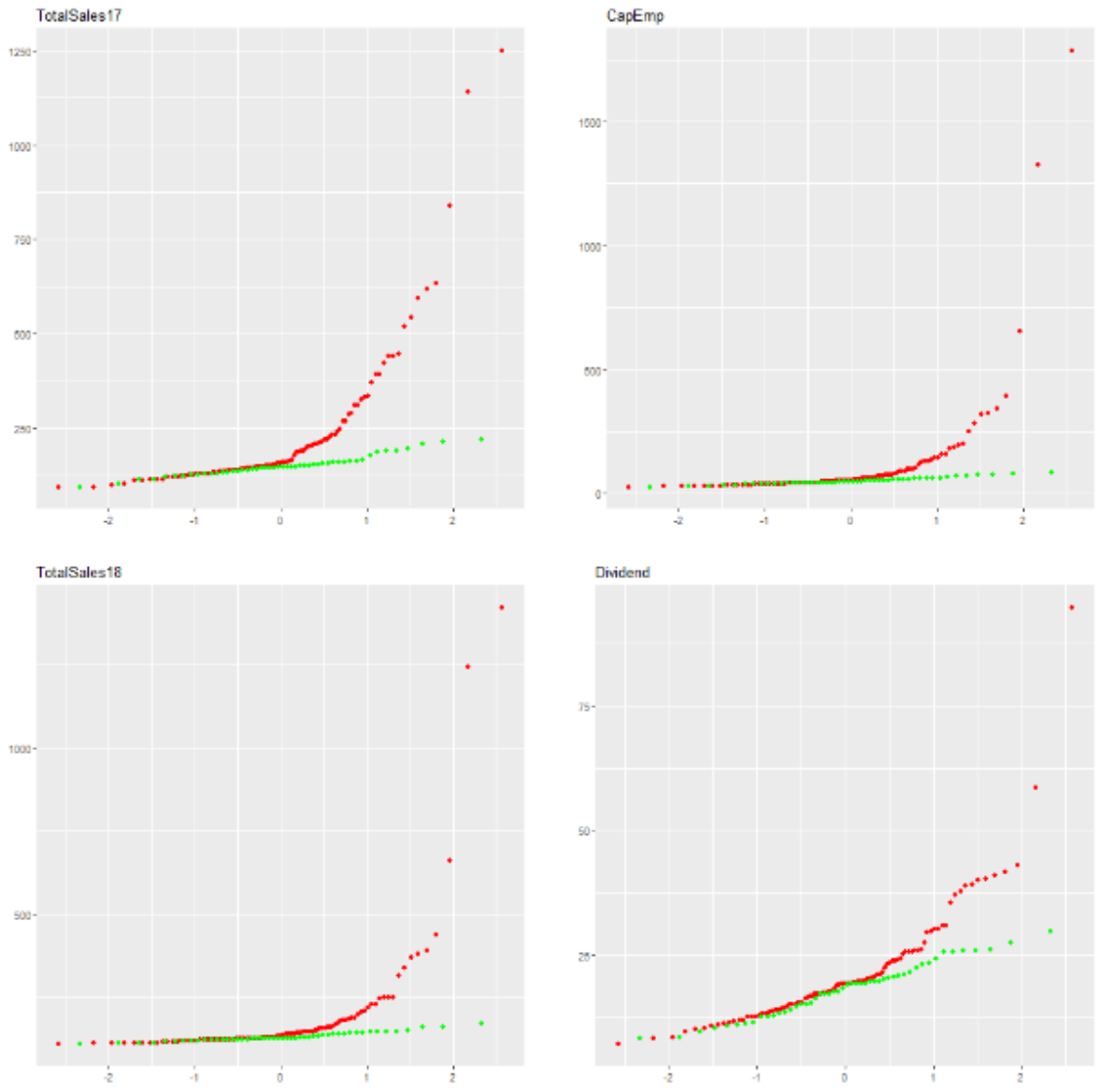
TotalSales17
1250
1000
750
500
250
-2
-1
0
1
2
CapEmp
1500
1000
500
0
-2
-1
0
1
2
TotalSales18
1000
500
-2
-1
0
1
2
Dividend
75
50
25
-2
-1
0
1
2

Image 4

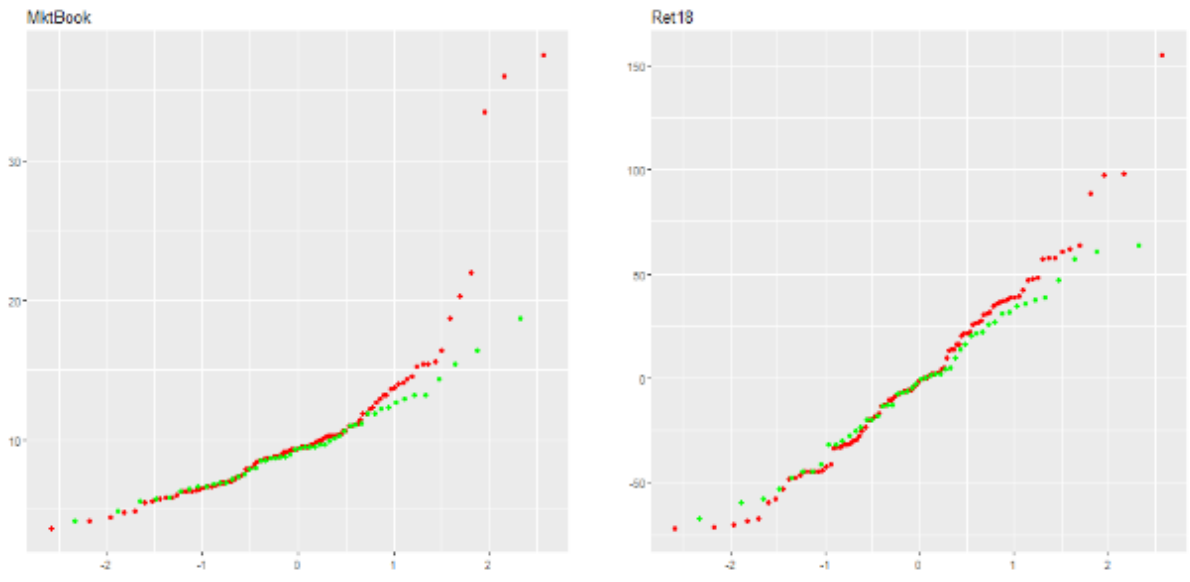

Image 5

Exploratory Data Analysis on Industry Data

The industrial data is comparable to retail data in that it has 100 entries for 12 identical quantitative factors. The dataset's executive summary is provided below(image 6).

```
   MktPrice        TotMktCap          DivYield         PERatio
Min.   : 11.22   Min.   :  43.52   Min.   : 1.550   Min.   :  3.11
1st Qu.: 41.70   1st Qu.:  57.62   1st Qu.: 6.915   1st Qu.: 19.70
Median : 86.22   Median :  72.58   Median : 9.310   Median : 26.22
Mean   :147.41   Mean   : 163.91   Mean   : 9.522   Mean   : 32.86
3rd Qu.:177.67   3rd Qu.: 155.39   3rd Qu.:11.490   3rd Qu.: 34.60
Max.   :862.32   Max.   :2466.74   Max.   :29.090   Max.   :132.80
                                                    NA's   :25
     Beta          TotalSales17      TotalSales18        CapEmp
Min.   :0.0300   Min.   :  88.88   Min.   :  94.73   Min.   :  16.20
1st Qu.:0.4825   1st Qu.: 121.25   1st Qu.: 115.65   1st Qu.:  31.91
Median :0.7600   Median : 147.39   Median : 127.10   Median :  43.95
Mean   :0.7616   Mean   : 236.32   Mean   : 176.76   Mean   : 106.85
3rd Qu.:1.0025   3rd Qu.: 235.37   3rd Qu.: 161.68   3rd Qu.:  88.02
Max.   :1.7100   Max.   :1325.65   Max.   :1347.38   Max.   :1776.38
NA's   :2
   Dividend        MktBook           Ret17            Ret18
Min.   : 5.67   Min.   :-1.010   Min.   :-88.210   Min.   :-56.69
1st Qu.:12.28   1st Qu.: 7.378   1st Qu.:-32.400   1st Qu.: 34.98
Median :15.93   Median : 9.370   Median :  2.510   Median : 56.17
Mean   :20.27   Mean   :10.024   Mean   : -2.133   Mean   : 65.83
3rd Qu.:25.05   3rd Qu.:11.363   3rd Qu.: 24.668   3rd Qu.: 99.69
Max.   :96.12   Max.   :40.190   Max.   :133.990   Max.   :209.55
NA's   :1
```

Image 6

The dataset's normality and outliers are investigated(image 7), as was already discussed, and median imputation is used to filter out null values. The analysis results are shown in the images(images 8,image 9,image 10)[5] below.

```
        Anderson-Darling normality test

data:  industry_data$MktPrice
A = 8.7744, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  industry_data$TotMktCap
A = 20.267, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  industry_data$DivYield
A = 1.0676, p-value = 0.008012
```

```
        Anderson-Darling normality test

data:  industry_data$PERatio
A = 7.7101, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  industry_data$Beta
A = 0.36524, p-value = 0.4298
```

```
        Anderson-Darling normality test

data:  industry_data$TotalSales17
A = 13.271, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  industry_data$TotalSales18
A = 18.847, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  industry_data$CapEmp
A = 20.229, p-value < 2.2e-16
```

```
        Anderson-Darling normality test

data:  industry_data$Dividend
A = 5.8098, p-value = 2.097e-14
```

```
        Anderson-Darling normality test

data:  industry_data$MktBook
A = 4.3041, p-value = 9.118e-11
```

```
        Anderson-Darling normality test

data:  industry_data$Ret17
A = 0.27672, p-value = 0.6483
```

```
        Anderson-Darling normality test

data:  industry_data$Ret18
A = 0.54626, p-value = 0.1563
```

Image 7

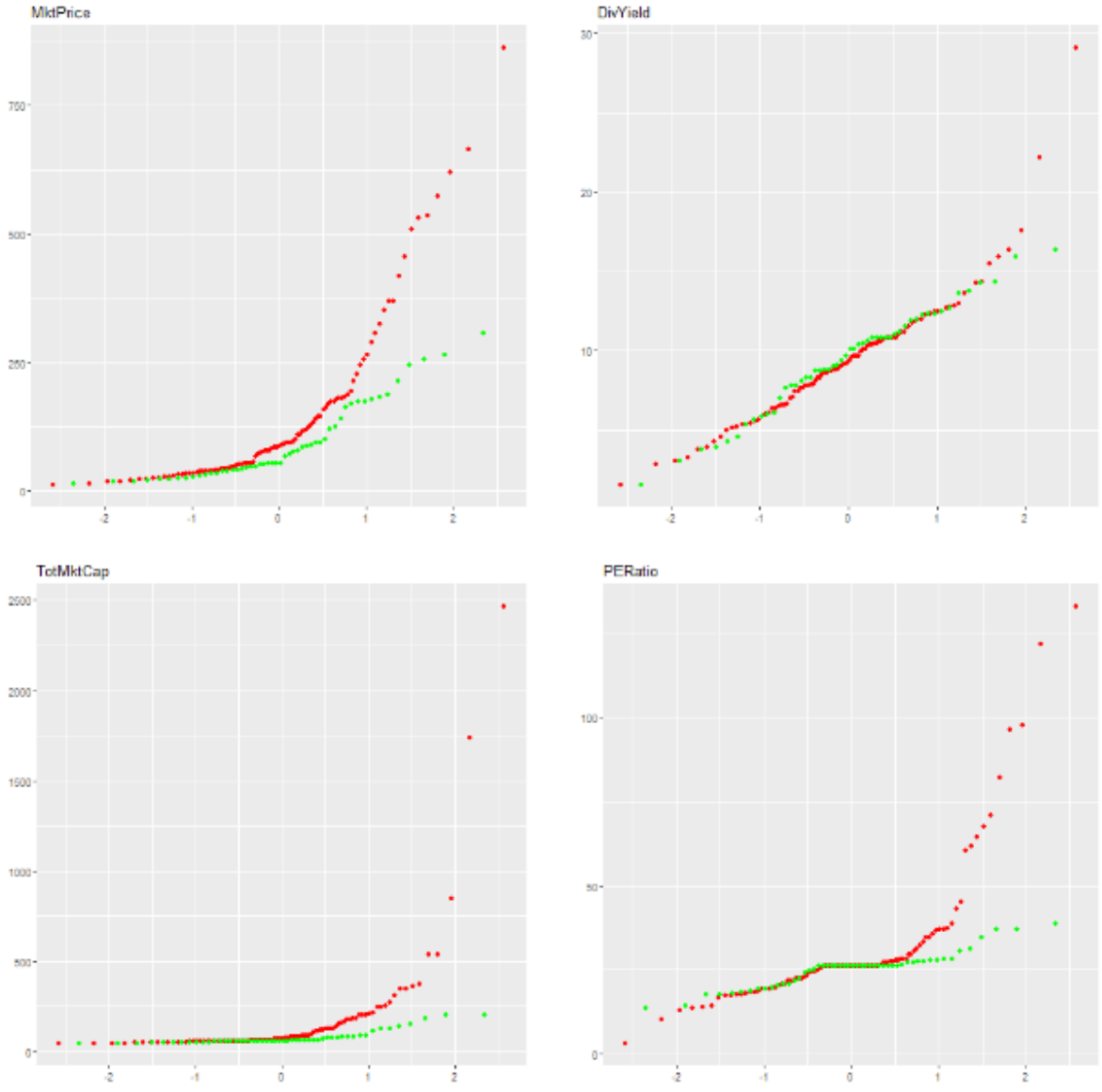

Image 8

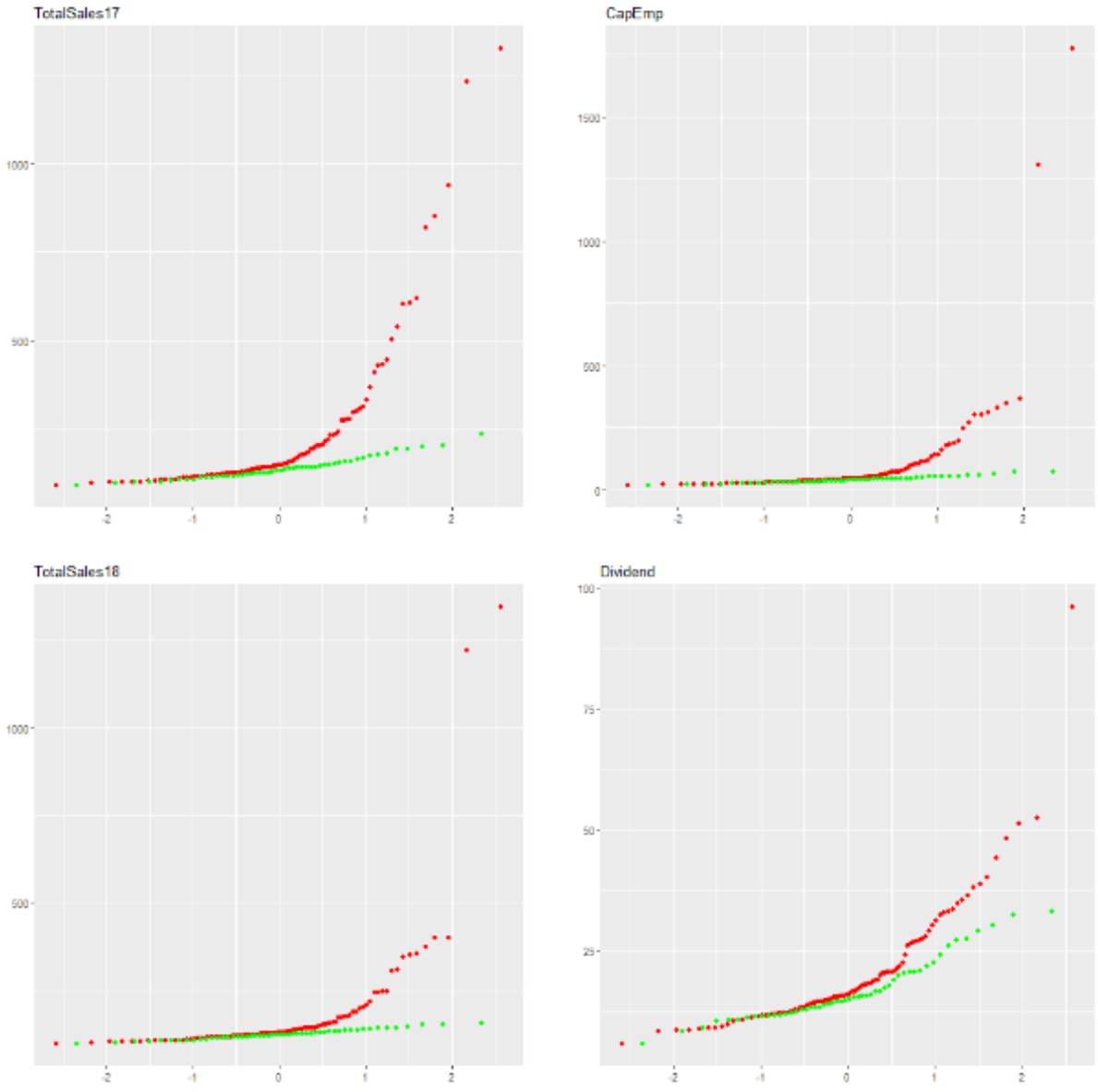
TotalSales17
1000
500
-2
-1
0
1
2
CapEmp
1500
1000
500
0
-2
-1
0
1
2
TotalSales16
1000
500
-2
-1
0
1
2
Dividend
100
75
50
25
-2
-1
0
1
2

Image 9

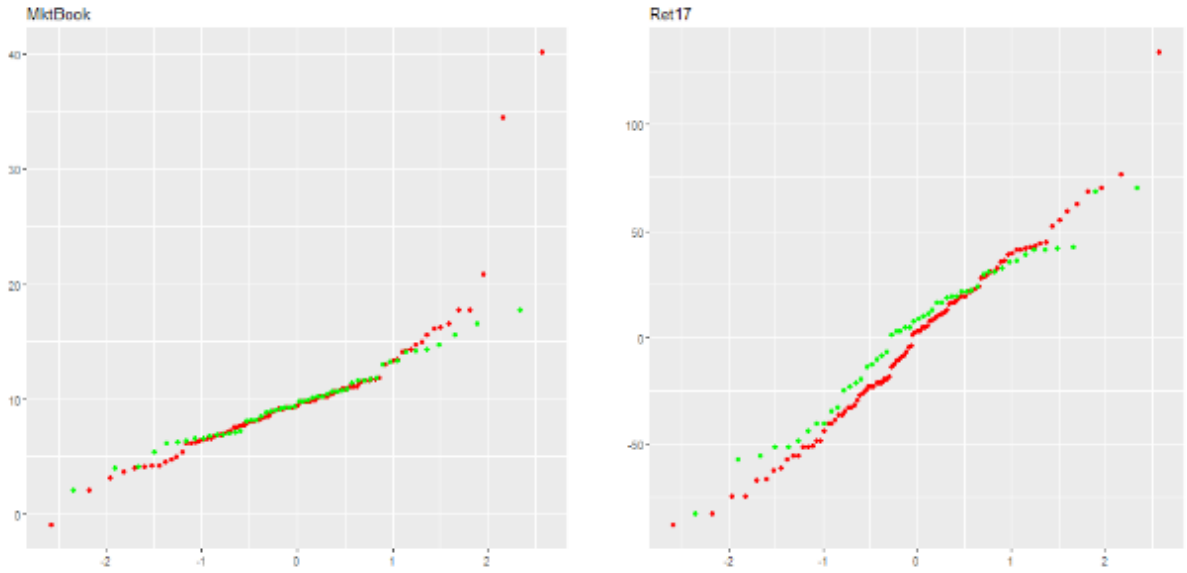

Image 10

Feature selection using Correlation

Feature selection produces a subset of pertinent features by removing the extraneous variables from the dataset. This improves the prediction power of the model and makes the results more accurate.There are many methods for choosing features, and in this case we use the correlation coefficient to do so.

The measurement of correlation is employed to determine the linear relationship between two variables. It always accepts values in the range of -1 and 1, where -1 denotes a perfect negative correlation and 1 denotes a perfect positive correlation. There is no correlation between the variables if the correlation coefficient is zero.In a correlation study, the alternative hypothesis contends that there is an association between the variables, while the null hypothesis contends that the variables are independent.Below are the findings of the correlation analysis performed on retail and industry data.

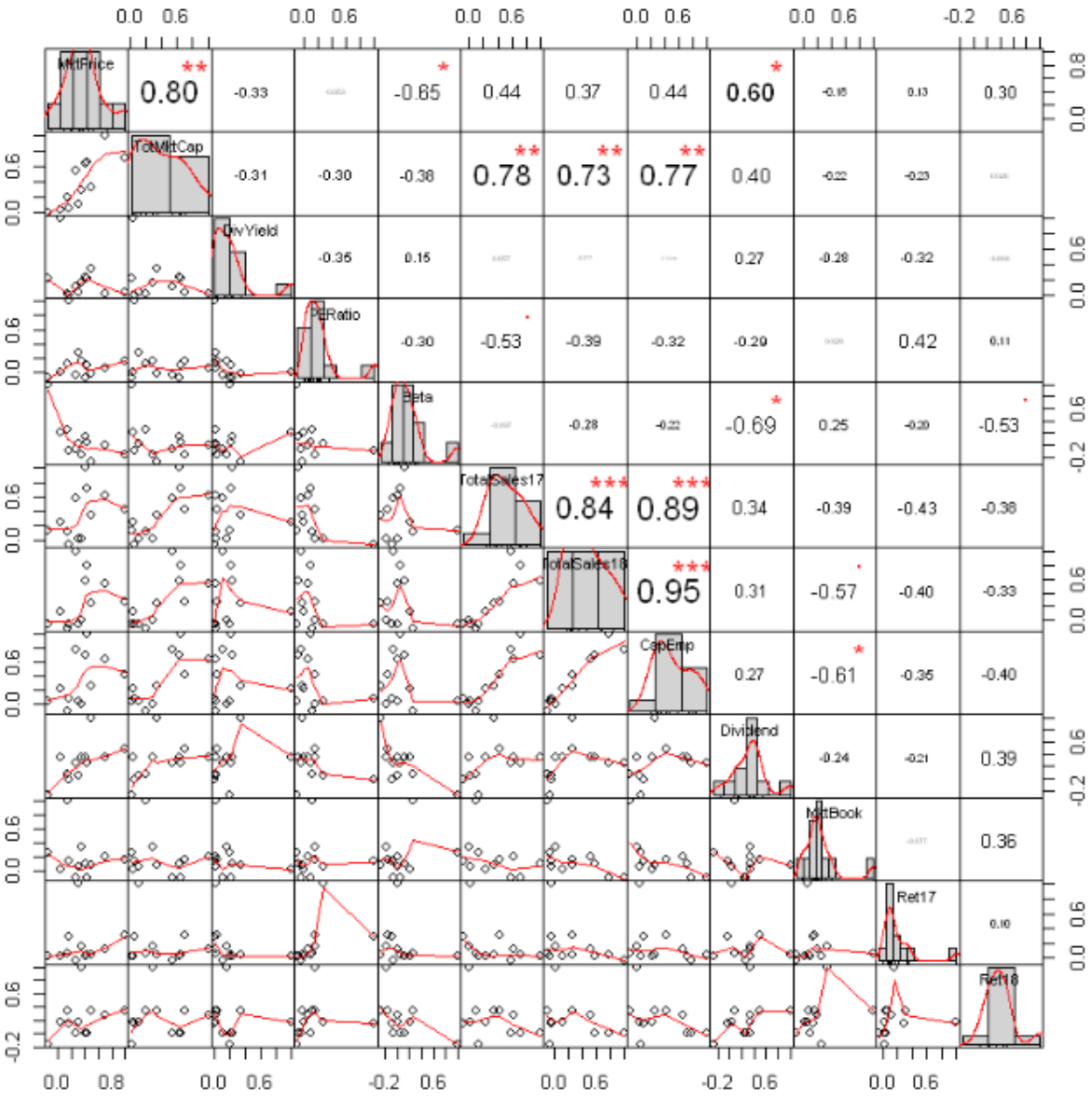

Image 11

The correlation plot of the variables in the retail data is displayed in the figure(image 11)[3] above. It is clear from the correlation graph that six variables—MktPrice, TotMktCap, TotalSales17, TotalSales18, CapEmp, and Dividend—have strong positive correlations. These variables are taken into account when modeling the data.

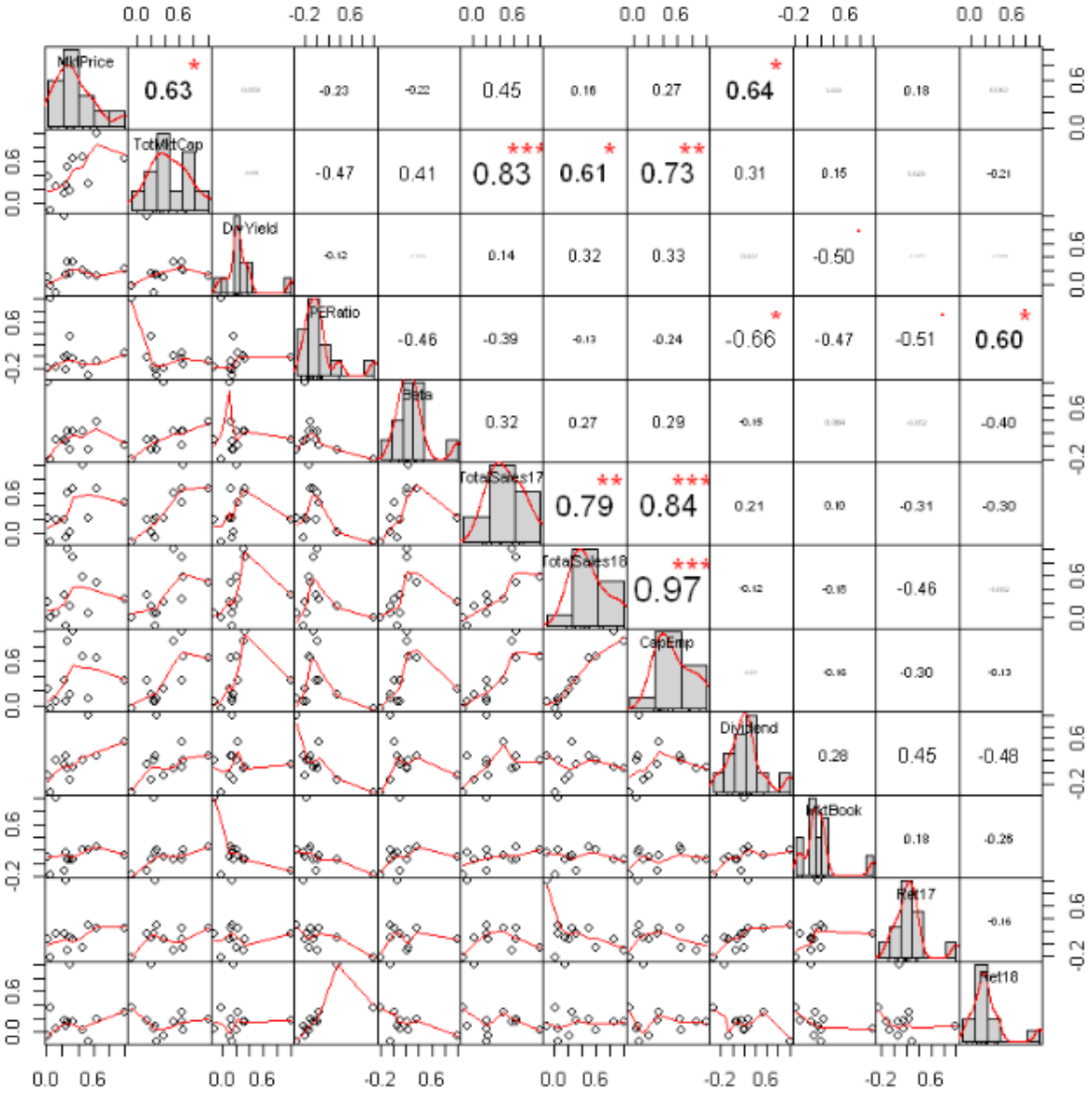

Image 12

Similarly to test the connection between the variables in the industry data, a correlation plot is first constructed(image 12)[3]. The variables MktPrice, TotMktCap, PERatio, TotalSales17, TotalSales18, CapEmp, Dividend, and Ret18 are then chosen for modeling since they exhibit a strong positive correlation.

Multi linear regression modeling

Finding the stock type that delivers the best return on investment is the study's main objective. The dataset contains a variety of quantitative characteristics, therefore multilinear regression modeling[4][6] can be utilized to ascertain the degree of correlation between these variables.Multiple linear regression is a type of regression model that uses a straight line to represent the relationship between two or more independent variables and a quantitative dependent variable.The null and alternative hypotheses for multiple linear regression are as follows: According to the null hypothesis, all model coefficients are equal to zero. In other words, there is no statistically significant correlation between any of the predictor factors and the response variable y. Also, not all coefficients are simultaneously equal to 0 will be the alternative hypothesis.The following four premises govern multi linear regression:
1.Linearity of the data. It is assumed that the connection between the predictor (x) and the result (y) is linear.
2. Normality of residuals. It is assumed that the residual errors are regularly distributed.
3. Homogeneity of residuals variance.It is presummated that the residuals' variance will never change(homoscedasticity).
4.Observations are independent.

Let's choose MktPrice as the target variable since it indicates a product's or service's cost, making it useful for determining the best stock to purchase.In order to compare and identify the best stock for investment, multiple linear regression models are developed using data from both the retail and industrial sectors. The findings are provided in the following session.

Results

Model 1 :Retail Data

Explanatory Variables:TotMktCap,TotalSales17,TotalSales18,CapEmp,Dividend
Target Variable(y): MktPrice

The multi linear regression model's hypotheses are,
Null Hypothesis : The dependent variable and the independent variables have no association, according to the null hypothesis of a multiple regression.

Alternative Hypothesis:When all other factors are equal, or when the levels of the other independent variables remain constant, the dependent variable is associated with the observed independent variable.

1. Is MktPrice associated with TotMktCap at a constant level of TotalSales17,TotalSales18,CapEmp,Dividend
2. Is MktPrice associated with TotalSales17 at a constant level of TotalSales18,CapEmp,Dividend,TotMktCap
3. Is MktPrice associated with TotalSales18 at a constant level of CapEmp,Dividend,TotMktCap ,TotalSales17
4. Is MktPrice associated with CapEmp at a constant level of Dividend,TotMktCap,TotalSales17,TotalSales18
5. Is MktPrice associated with Dividend at a constant level of TotMktCap,TotalSales17,TotalSales18,CapEmp

Testing of Assumptions

1.Linearity of the relationship between y and its explanatory variables

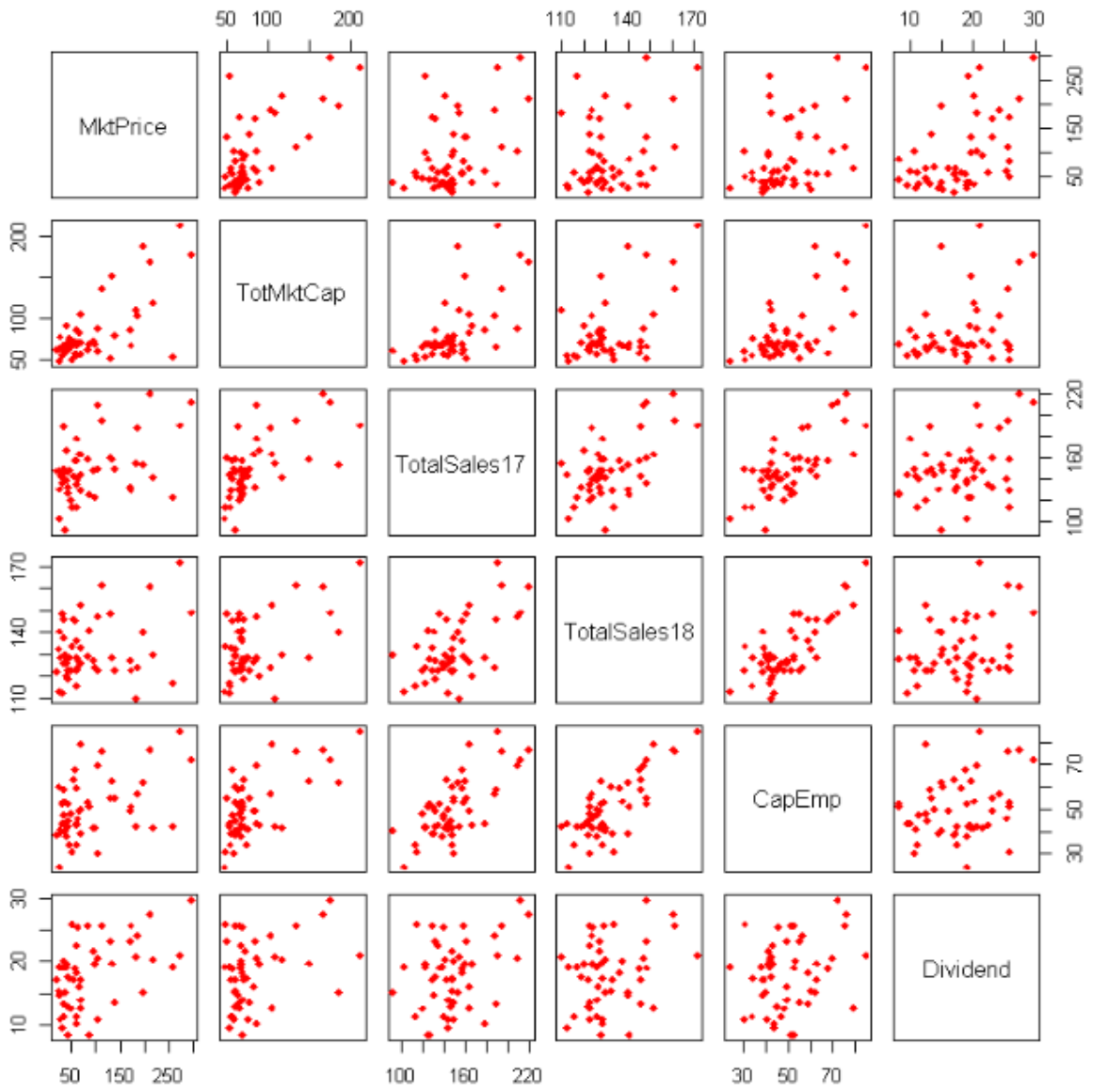

Image 13

The pair plot (image 13)[5][7] demonstrates the positive linear association between the variables MktPrice and TotMktCap. The target variable and the other explanatory factors, such as TotSales17, TotSales18, CapEmp and Divident, appear to have a modestly positive relationship.

2. Normality of residuals. It is assumed that the residual errors are regularly distributed.

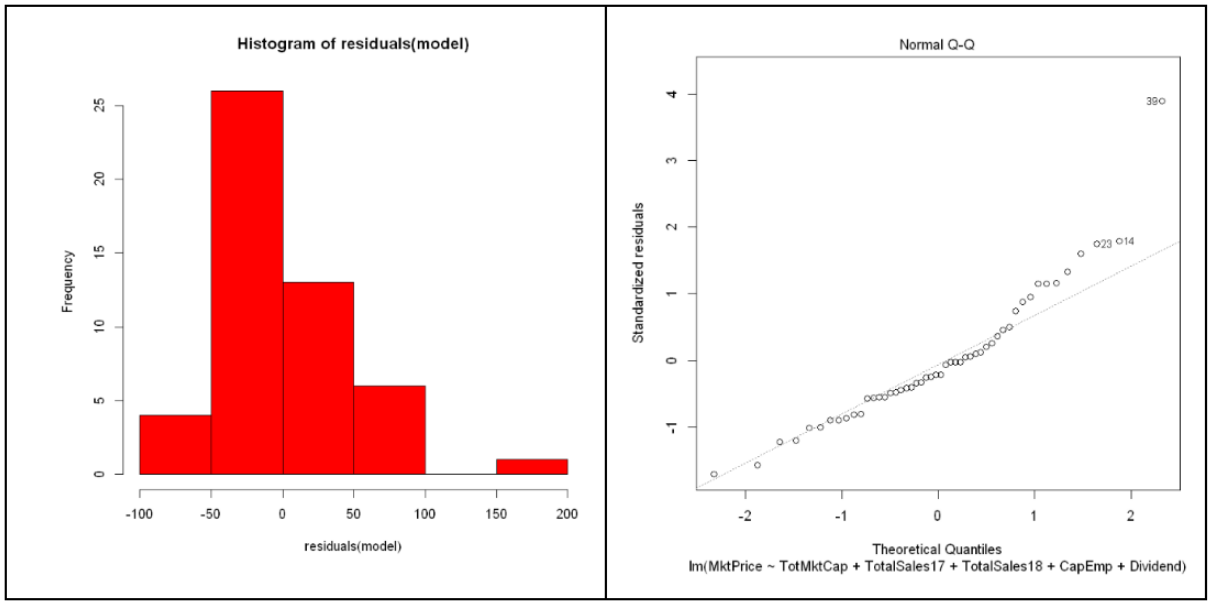

Image 14

The histogram(image 14) shows that the distribution is slightly right-skewed, but it is not sufficiently atypical enough to raise any serious concerns.We can tell from this plot that the residuals seem to support the notion that the residual terms follow a normal distribution.

3. Homogeneity of residuals variance.It is presummated that the residuals' variance will never change(homoscedasticity).

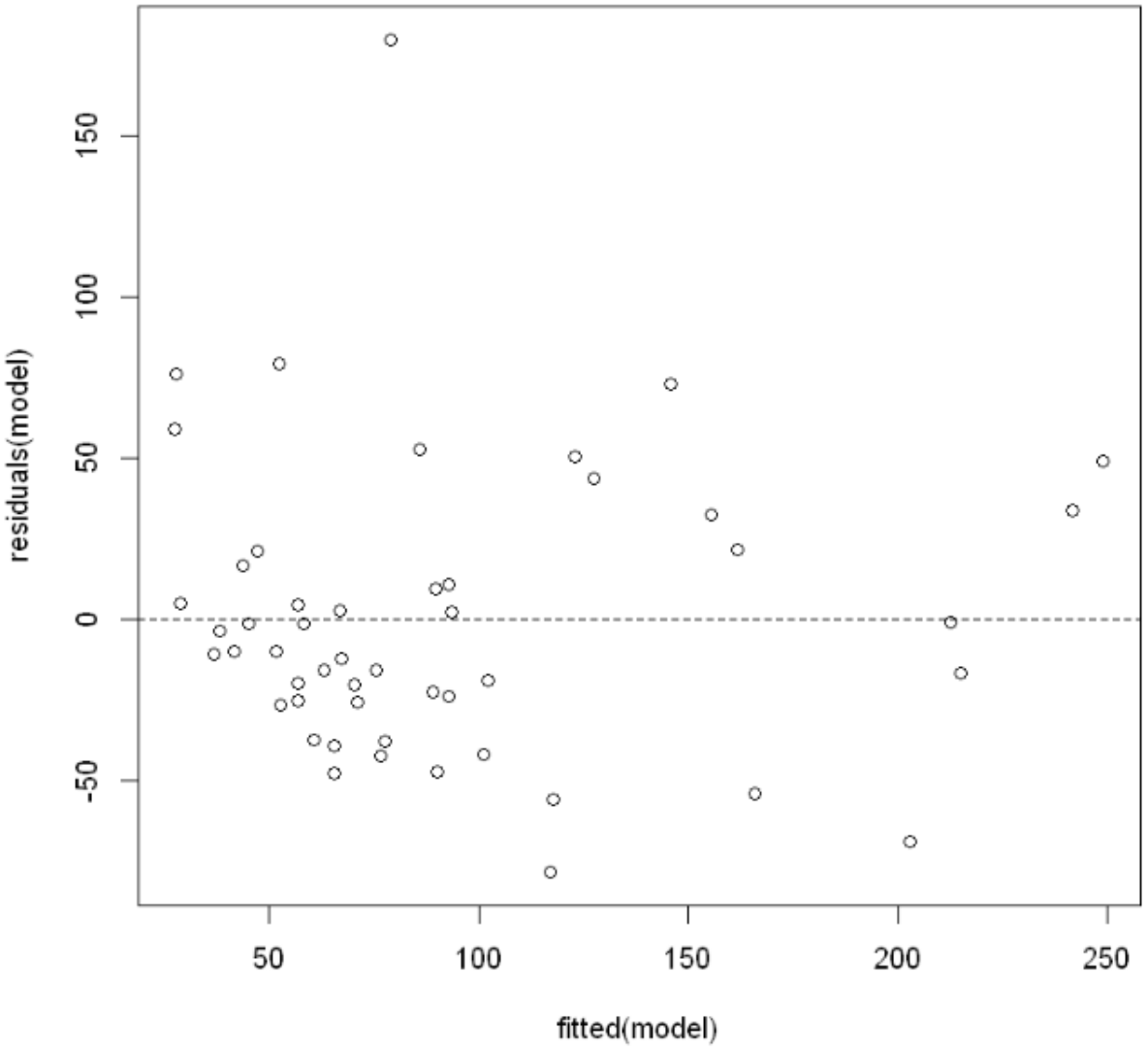

Image 15

At every fitted value, the residuals should ideally be evenly distributed(image 15). The scatter does tend to increase slightly with smaller fitted values, as can be seen from the plot, but this tendency isn't particularly alarming.

4.Observations are independent.

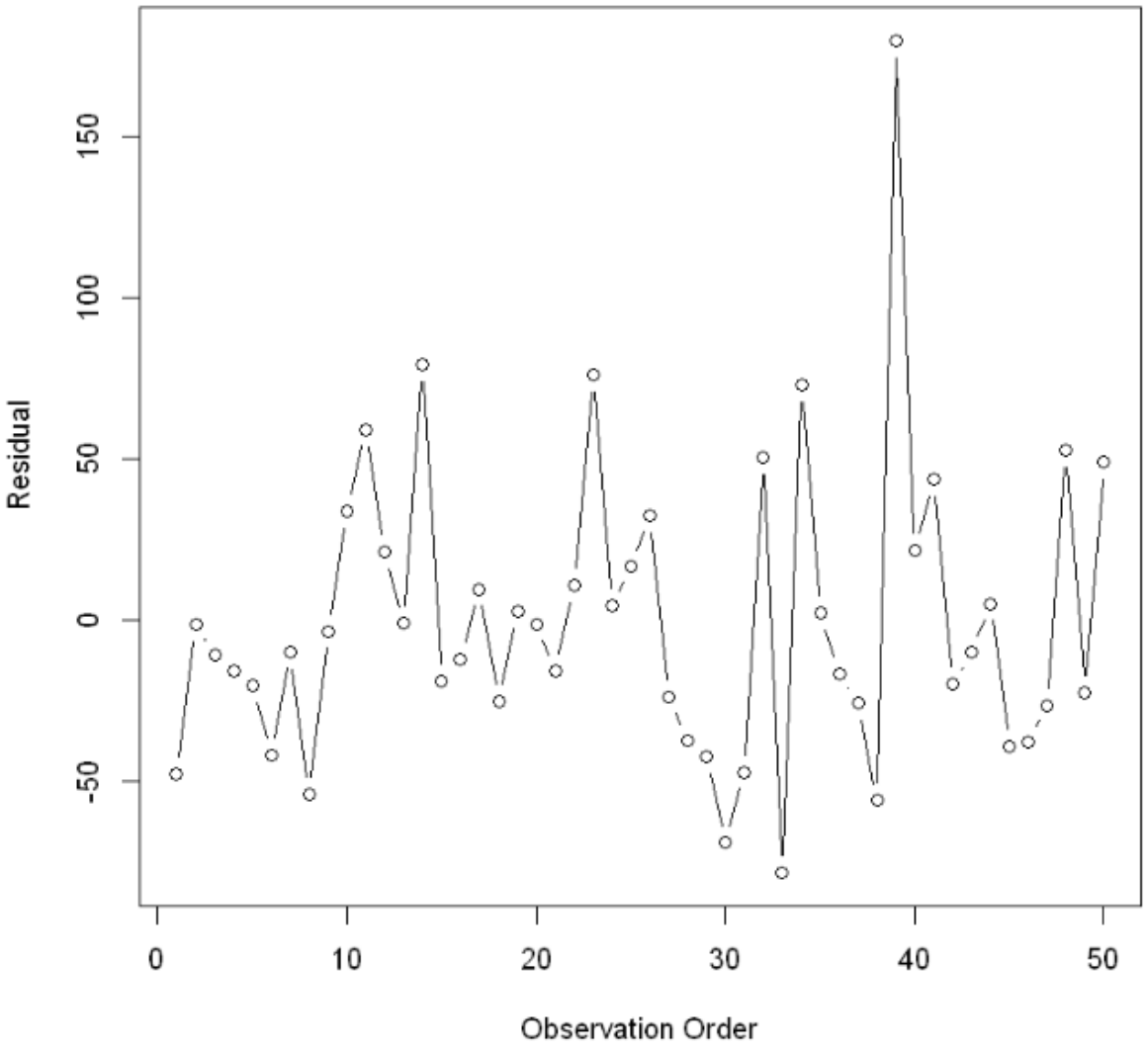

Image 16

Because there is no discernible pattern in the Residual vs. Order Diagnostic Plot (bottom left), we can say that the residuals are independent of one another(image 16).

Model Interpretation

```
Call:
lm(formula = MktPrice ~ TotMktCap + TotalSales17 + TotalSales18 +
    CapEmp + Dividend, data = retail_new)

Residuals:
    Min      1Q  Median      3Q     Max
-78.318 -25.410  -9.755  20.154 180.012

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   79.0173    86.7231   0.911  0.36718
TotMktCap      1.3053     0.2565   5.090 7.17e-06 ***
TotalSales17  -0.1966     0.3889  -0.505  0.61575
TotalSales18  -1.3949     0.8374  -1.666  0.10287
CapEmp         0.8561     1.0060   0.851  0.39943
Dividend       4.2591     1.3916   3.061  0.00376 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.01 on 44 degrees of freedom
Multiple R-squared:  0.6012,    Adjusted R-squared:  0.5559
F-statistic: 13.27 on 5 and 44 DF,  p-value: 6.754e-08
```

Image 17

The model's overall F-statistic is 13.27, and the associated p-value is 6.754e-08. This demonstrates the statistical significance of the model as a whole.The estimate for each parameter is provided in the table Coefficients (column Estimate), along with the p-value for the parameter's nullity(image 17).

At the 0.001 level of significance, TotMktCap is statistically significant. If TotalSales17, TotalSales18, CapEmp, and Dividend are held equal, the coefficient from the model's output specifically states that a one unit rise in TotMktCap is related with a 1.3053 unit increase, on average, in MktPrice.

Similarly at the 0.001 level of statistical significance, dividend is significant. If TotalSales17, TotalSales18, CapEmp, and TotMktCap are held constant, the coefficient from the model's output

specifically states that a one unit rise in Dividend is related with a 4.2591 unit increase, on average, in MktPrice.

Goodness of fit of the model

The magnitude of the linear relationship between the predictor variables and the response variable is gauged by the multiple R squared. A multiple R-squared of 1 denotes the existence of an ideal linear relationship, whereas a multiple R-squared of 0 denotes the complete absence of any linear link. The multiple R-squared from the outcome is 0.6012 This shows that the predictors in the model can account for 36.1% of the variance in MktPrice.The measured values deviate from the regression line by an average of 48.01 units, according to residual standard error.Dropping the variables makes the model even better, and the end result is shown below.

```
Start:  AIC=392.75
MktPrice ~ TotMktCap + TotalSales17 + TotalSales18 + CapEmp +
    Dividend

               Df Sum of Sq    RSS    AIC
- TotalSales17  1       589 102013 391.04
- CapEmp        1      1669 103093 391.57
<none>                     101424 392.75
- TotalSales18  1      6396 107820 393.81
- Dividend      1     21591 123015 400.40
- TotMktCap     1     59709 161133 413.90

Step:  AIC=391.04
MktPrice ~ TotMktCap + TotalSales18 + CapEmp + Dividend

               Df Sum of Sq    RSS    AIC
- CapEmp        1      1157 103170 389.61
<none>                     102013 391.04
- TotalSales18  1      6455 108468 392.11
- Dividend      1     21007 123019 398.40
- TotMktCap     1     60757 162770 412.40

Step:  AIC=389.61
MktPrice ~ TotMktCap + TotalSales18 + Dividend

               Df Sum of Sq    RSS    AIC
<none>                     103170 389.61
- TotalSales18  1      6101 109270 390.48
- Dividend      1     20783 123953 396.78
- TotMktCap     1     84662 187832 417.56


Call:
lm(formula = MktPrice ~ TotMktCap + TotalSales18 + Dividend,
    data = retail_new)

Coefficients:
 (Intercept)     TotMktCap  TotalSales18      Dividend
     39.4050        1.3487       -0.9969        4.1098
```

Image 18

From the above image(image 18) the equation for predicting market price will be

MktPrice = 39.4050+1.3487TotMktCap-0.9969TotSales18+4.1098Divident

```
Call:
lm(formula = MktPrice ~ TotMktCap + TotalSales18 + Dividend,
    data = retail_new)

Residuals:
    Min      1Q  Median      3Q     Max
-83.570 -28.406  -7.015  16.337 185.414

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.4050    72.8139   0.541  0.59100
TotMktCap      1.3487     0.2195   6.144 1.76e-07 ***
TotalSales18  -0.9969     0.6045  -1.649  0.10591
Dividend       4.1098     1.3501   3.044  0.00385 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.36 on 46 degrees of freedom
Multiple R-squared:  0.5943,    Adjusted R-squared:  0.5679
F-statistic: 22.47 on 3 and 46 DF,  p-value: 4.179e-09
```

Image 19

It is obvious from the preceding figure(image 19) that removing the variable has no positive effects on the model. So, for extensive analysis, the initial model is taken into account.

Model 2 :Industry Data

Explanatory Variables:TotMktCap,TotalSales17,TotalSales18,CapEmp,Dividend, PERatio,Ret18
Target Variable(y): MktPrice

The multi linear regression model's hypotheses are

Null Hypothesis : The dependent variable and the independent variables have no association, according to the null hypothesis of a multiple regression.

Alternative Hypothesis:When all other factors are equal, or when the levels of the other independent variables remain constant, the dependent variable is associated with the observed independent variable.

1. Is MktPrice associated with TotMktCap at a constant level of TotalSales17,TotalSales18,CapEmp,Dividend,PERatio,Ret18
2. Is MktPrice associated with TotalSales17 at a constant level of TotalSales18,CapEmp,Dividend,TotMktCap,PERatio,Ret18
3. Is MktPrice associated with TotalSales18 at a constant level of CapEmp,Dividend,TotMktCap ,TotalSales17,PERatio,Ret18
4. Is MktPrice associated with CapEmp at a constant level of Dividend,TotMktCap,TotalSales17,TotalSales18,PERatio,Ret18
5. Is MktPrice associated with Dividend at a constant level of TotMktCap,TotalSales17,TotalSales18,CapEmp,PERatio,Ret18
6. Is MktPrice associated with PERatio at a constant level of TotMktCap,TotalSales17,TotalSales18,CapEmp,Ret18
7. Is MktPrice associated with Ret18 at a constant level of TotMktCap,TotalSales17,TotalSales18,CapEmp,PERatio

Testing of Assumptions

1.Linearity of the relationship between y and its explanatory variables

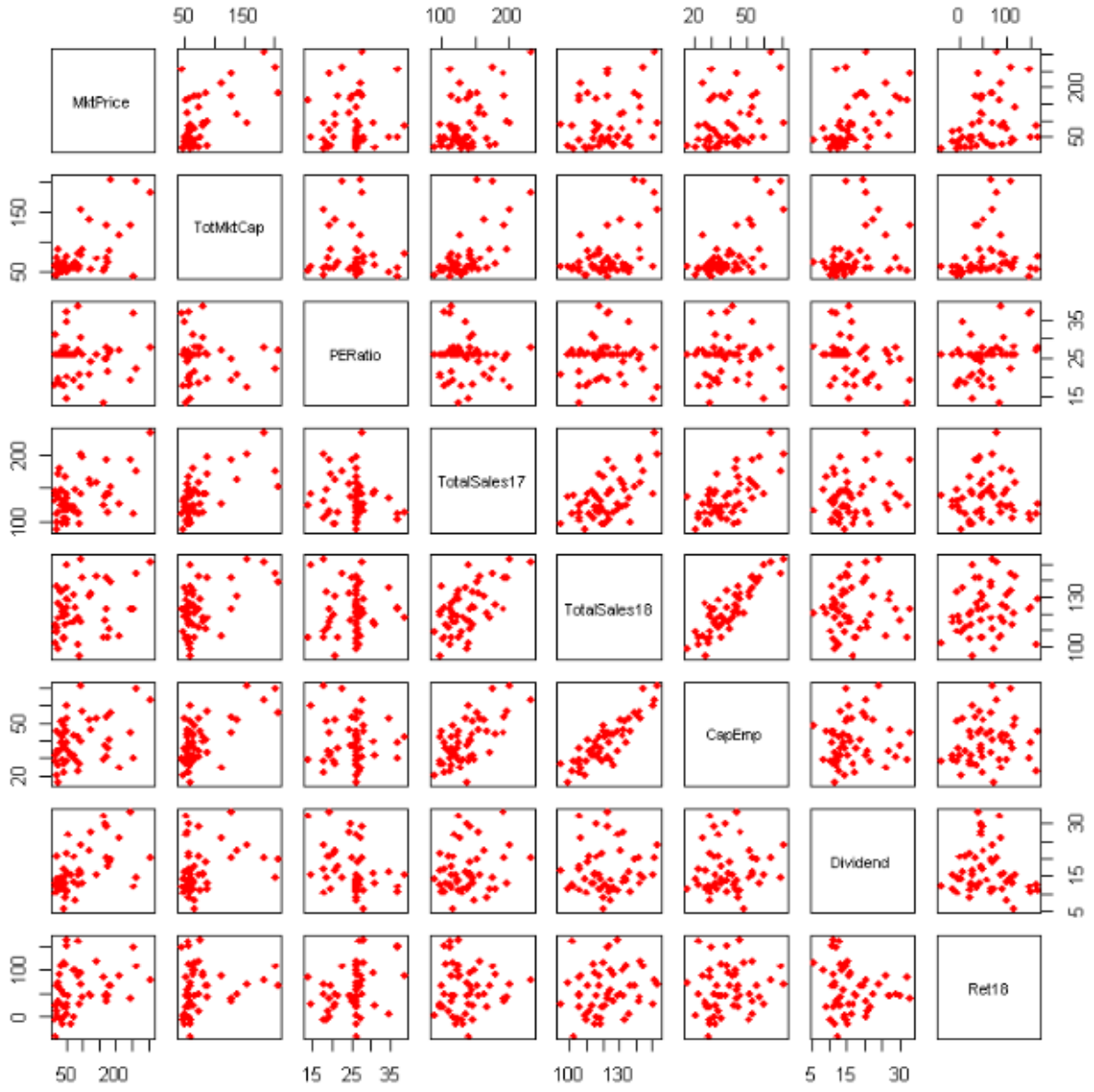

Image 20

The target variable MktPrice and the other explanatory factors, such as TotSales17, TotSales18, CapEmp,PERatio,Ret18 and Divident, appear to have a modestly positive relationship(image 20)[5][7].

2. Normality of residuals. It is assumed that the residual errors are regularly distributed.

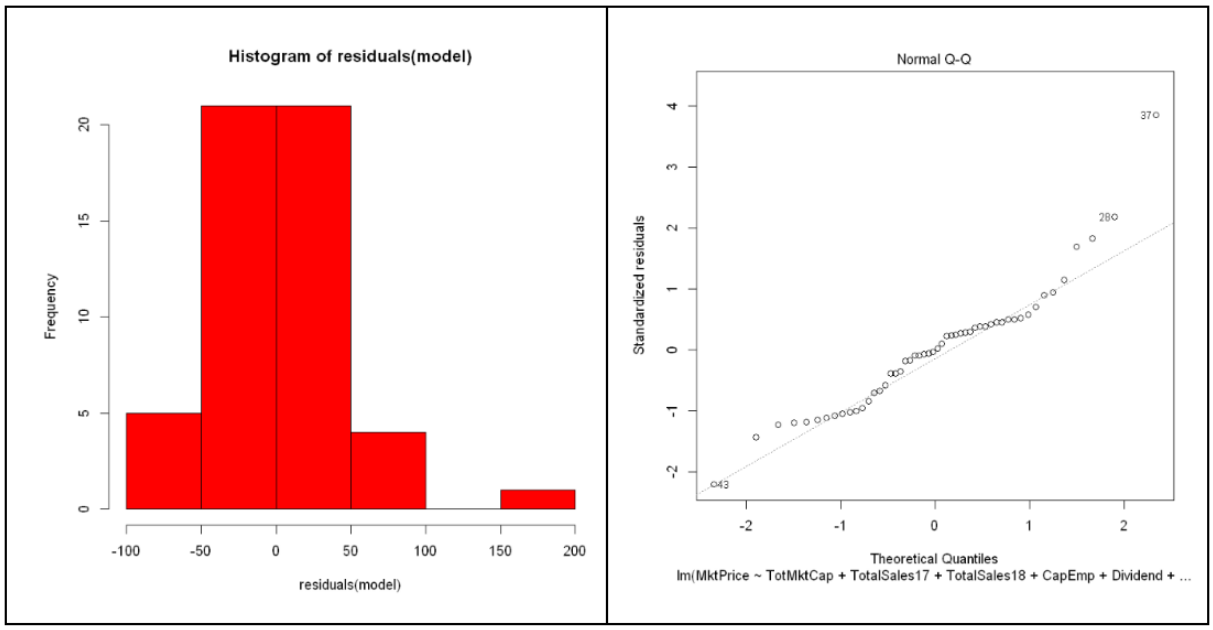

Image 21

The distribution is slightly right-skewed, as indicated by the histogram This plot indicates that the residuals appear to be consistent with the idea that the residual terms have a normal distribution(image 21).

3. Homogeneity of residuals variance.It is presummated that the residuals' variance will never change(homoscedasticity).

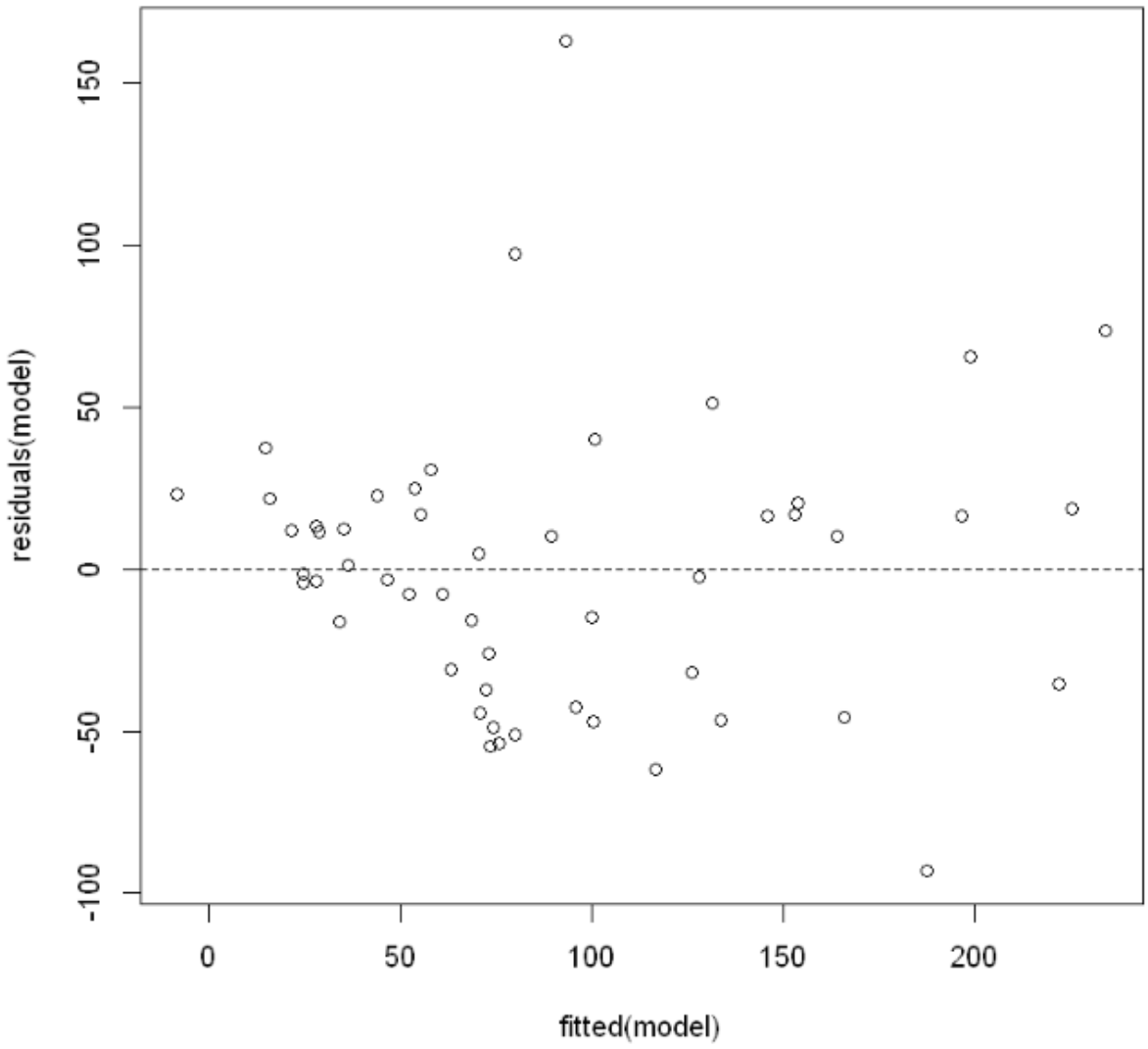

Image 22

The residuals should, in theory, be evenly distributed among all fitted values. The plot (image 22) shows that the scatter does tend to somewhat rise with lower fitted values, but this tendency is not extremely concerning.

4.Observations are independent.

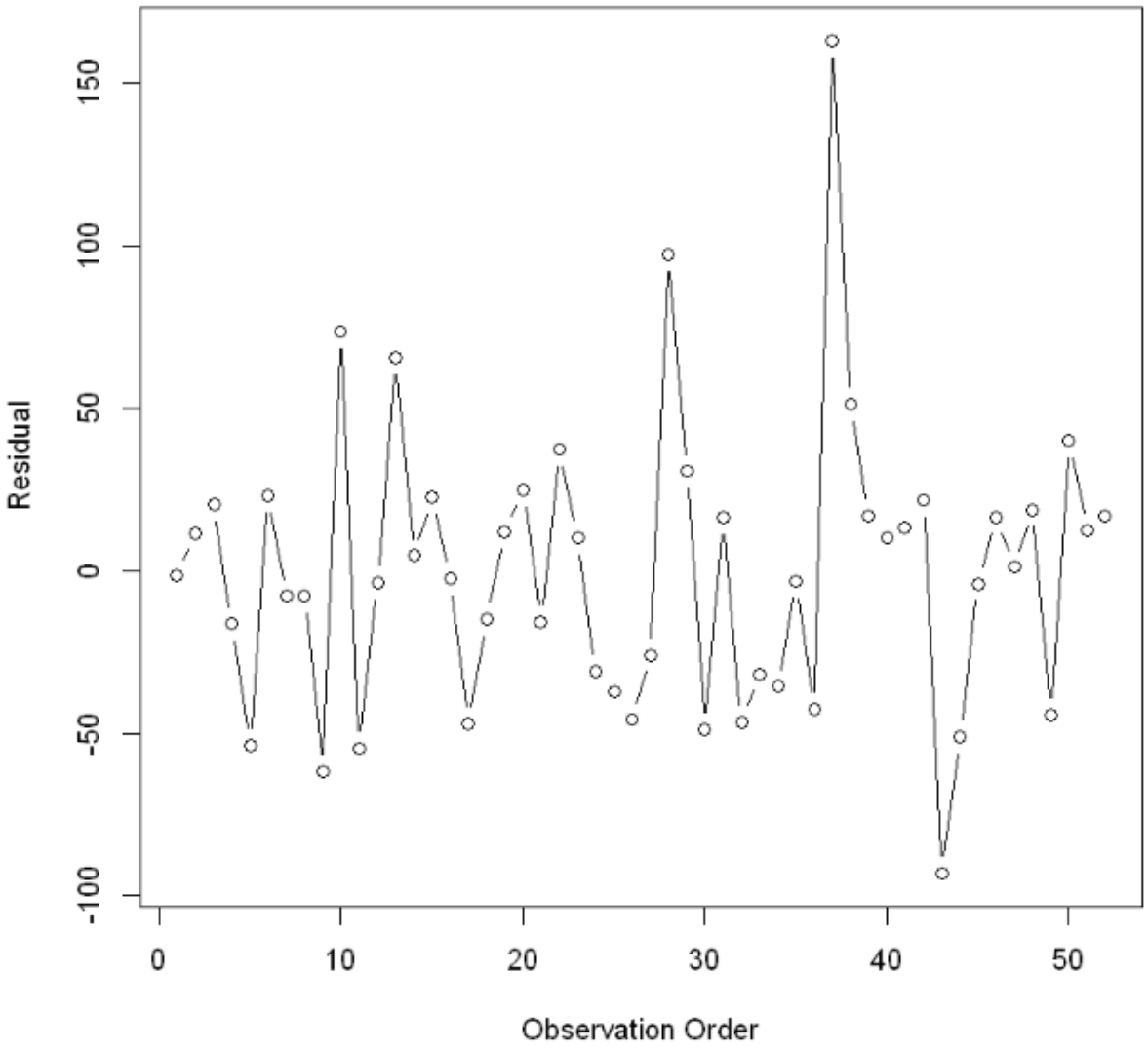

Image 23

We may conclude that the residuals are independent of one another because there is no visible pattern in the Residual vs. Order Diagnostic Plot (bottom left)(image 23).

Model Interpretation

```
Call:
lm(formula = MktPrice ~ TotMktCap + TotalSales17 + TotalSales18 +
    CapEmp + Dividend + PERatio + Ret18, data = industry_new)

Residuals:
    Min      1Q  Median      3Q     Max
-93.176 -32.834  -0.304  19.189 162.946

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -118.6910    91.0041  -1.304   0.1989
TotMktCap      0.8242     0.2570   3.207   0.0025 **
TotalSales17   0.3633     0.3125   1.163   0.2513
TotalSales18  -0.8227     0.9541  -0.862   0.3932
CapEmp         0.2614     1.1940   0.219   0.8277
Dividend       5.7225     1.1562   4.949 1.14e-05 ***
PERatio        2.8061     1.4381   1.951   0.0574 .
Ret18          0.3731     0.1535   2.430   0.0192 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.94 on 44 degrees of freedom
Multiple R-squared:  0.6582,    Adjusted R-squared:  0.6039
F-statistic: 12.11 on 7 and 44 DF,  p-value: 1.698e-08
```

Image 24

The total F-statistic for the model is 12.11, and the corresponding p-value is 1.698e-08. This indicates the model's overall statistical significance.The estimate for each parameter is provided in the table Coefficients (column Estimate), along with the p-value for the parameter's nullity(image 24).

TotMktCap is statistically significant at the level of 0.01 significance. The coefficient from the model's output precisely specifies that, when all other independent variables are held constant, a one unit increase in TotMktCap is correlated with an average increase in MktPrice of 0.8242 units.

Dividend is significant at the 0.001 level of statistical significance. The coefficient from the model's output precisely specifies that, when all other independent variables are held constant, a one unit increase in Dividend is associated to an average increase in MktPrice of 5.7225 units.

At the 0.05 level of statistical significance, Ret18 is significant. When all other independent variables are maintained constant, the coefficient from the model's output accurately states that a one unit rise in Dividend is related with an average increase in MktPrice of 0.3731 units.

Goodness of fit of the model

The multiple R-squared from the outcome is 0.6582 which indicates the predictors in the model can account for 43.3% of the variance in MktPrice.The measured values deviate from the regression line by an average of 46.94 units, according to residual standard error.The improved model is given below.

```
            Df Sum of Sq    RSS    AIC
<none>                    101254 403.86
- PERatio    1      7608 108862 405.62
- Ret18      1     11246 112500 407.33
- TotMktCap  1     55158 156412 424.47
- Dividend   1     56714 157967 424.98


Call:
lm(formula = MktPrice ~ TotMktCap + Dividend + PERatio + Ret18,
    data = industry_new)

Coefficients:
(Intercept)    TotMktCap     Dividend      PERatio        Ret18
  -162.4926       0.9187       5.8053       2.6585       0.3434
```

Image 25

From the above image(image 25) the equation for predicting market price will be

MktPrice= -162.4926+0.9187TotMktCap+5.8053Divident+2.6585PERatio+0.3434Ret18

```
Call:
lm(formula = MktPrice ~ TotMktCap + PERatio + Dividend + Ret18,
    data = industry_new)

Residuals:
    Min      1Q  Median      3Q     Max
-95.159 -27.698  -1.845  22.276 158.837

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -162.4926    44.8049  -3.627 0.000705 ***
TotMktCap      0.9187     0.1816   5.060 6.87e-06 ***
PERatio        2.6585     1.4146   1.879 0.066417 .
Dividend       5.8053     1.1315   5.131 5.40e-06 ***
Ret18          0.3434     0.1503   2.285 0.026882 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.41 on 47 degrees of freedom
Multiple R-squared:  0.6431,    Adjusted R-squared:  0.6127
F-statistic: 21.17 on 4 and 47 DF,  p-value: 4.928e-10
```

Image 26

The previous figure(image 26) makes it clear that the variable's removal has no beneficial effects on the model. Therefore, the initial model is considered for the in analysis.

Conclusion

The accuracy of the models developed using the industrial dataset and the retail dataset can be clearly seen from the study above. When compared to retail data, the model developed with the industry dataset performs better and has a lower error rate, as can be shown after carefully evaluating the model. This leads us to the conclusion that industry sector investing will result in higher returns than retail investing. Dividend and TotMktCap are other factors that have an impact on the model's accuracy, indicating that the size of the business has an impact on its ability to generate profits.We can develop better assumptions if we have daily data on the retail and manufacturing sectors, along with the date variable. A time series analysis of the data will also enable a little safer sector selection and aid in improving estimates of each of these companies' future performance.

References

1. https://www.investopedia.com/terms/r/retailinvestor.asp
2. https://www.lawinsider.com/dictionary/industry-investor
3. http://www.sthda.com/english/wiki/correlation-analyses-in-r
4. https://www.r-bloggers.com/2021/10/multiple-linear-regression-made-simple/
5. https://viz-ggplot2.rsquaredacademy.com/ggplot2-scatter-plot.html
6. https://rpubs.com/ajdowny_student/300663
7. https://r-coder.com/correlation-plot-r/
8. https://www.statology.org/anderson-darling-test-r/