BUSINESS STATISTICS USING R

STOCK MARKET ANALYSIS: BAJAJ FINANCE LTD.

1. Introduction:

The stock market analysis is effective for all the investors since knowing the market condition beforehand and planning to take a position in a particular stock proves to be beneficial. The analysis minimizes the losses of investors and assures consistency. It gives them a better idea about the entry and exit points and this is the smartest way of making money. The Stock Market Analysis reduces and stabilizes the inflation of the country.

2. Problem Statement:

Exploring the basic techniques of visualizing stock market data, calculating moving averages to analyse trends, and computing the returns.

3. Objective:

- To find the trends in market and to estimate the stability of the volume stock and the variations in prices of the stock in market (as on 03-12-21).
- To analyse the trend of variations over the given interval of time.
- To measure the variations in the price, physical volume and value of the stocks on daily basis.

4. Dataset:

Kaggle Source: https://www.kaggle.com/datasets/gianetan/bajaj-finance-limited-bajfinance

National Stock Exchange (NSE) data of Bajaj Finance Limited from 2015 to 2021.

6 columns – Date, Opening and Closing Balance, Highest and Lowest Prices, Volume of Stock.

5. Methods:

5.1 Packages Used:

- Hmisc
- Psych
- Forecast
- ggplot2
- qcc

5.2 Software Used:

The Quantitative Analysis is executed in RStudio using the R Programming Language.

6. Analysis And Prediction:

6.1 BASIC OPERATIONS

- Descriptive Statistics is implemented to show the summary and describe the data using various statistical tools.
- The concept of index number is implemented.
- The covariance function is used to measure the relationship between two variables and their changes together.
- T – Test is implemented which compares the means of two samples.

Analysis (Code and Output):

```
R 4.2.1 · ~/
> data<-read.csv(file="C:\\Users\\SM CORPORATES\\OneDrive\\Desktop\\DCS - SEM 3\\R PROJECT\\DATA .csv", header=TRUE)
> class(data)
[1] "data.frame"
> summary(data)
     Date              Op_Bal          Cl_Bal          Highest          Lowest          Volume
 Length:242         Min.   :406.5   Min.   :406.5   Min.   :406.5   Min.   :406.3   Min.   :     0.0
 Class :character   1st Qu.:412.5   1st Qu.:412.5   1st Qu.:412.6   1st Qu.:412.5   1st Qu.:    22.5
 Mode  :character   Median :421.3   Median :421.2   Median :421.4   Median :421.1   Median :   165.0
                    Mean   :420.6   Mean   :420.6   Mean   :420.9   Mean   :420.4   Mean   :  1390.0
                    3rd Qu.:426.6   3rd Qu.:426.6   3rd Qu.:426.9   3rd Qu.:426.5   3rd Qu.:   747.5
                    Max.   :442.8   Max.   :442.4   Max.   :442.8   Max.   :442.4   Max.   :109240.0
> var(data$Volume)
[1] 54704956
> p1=sum(data$Highest)
> p2=sum(data$Lowest)
> index=(p2/p1)*100
> index
[1] 99.87864
> print(cov(data$Highest, data$Lowest))
[1] 75.96918
> t.test(data$Op_Bal, data$Cl_Bal)

        Welch Two Sample t-test

data:  data$Op_Bal and data$Cl_Bal
t = 0.035324, df = 481.97, p-value = 0.9718
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.530425  1.586458
sample estimates:
mean of x mean of y
 420.6295  420.6014
```

Inference:

The basic operations such as descriptive statistics (summary), variance, index numbers, covariance and t-test are implemented in the RStudio to describe the data statistically.

6.2CORRELATION:

- Correlation is used to establish linear dependency between the data sets.
- Here, Correlation is implemented to measure the relationship between the prices of the stocks over the period of time.

```
R 4.2.1 · ~/
> cor(data$Highest, data$Lowest)
[1] 0.9947332
> cor(data$Highest, data$Volume)
[1] 0.1626951
>
```

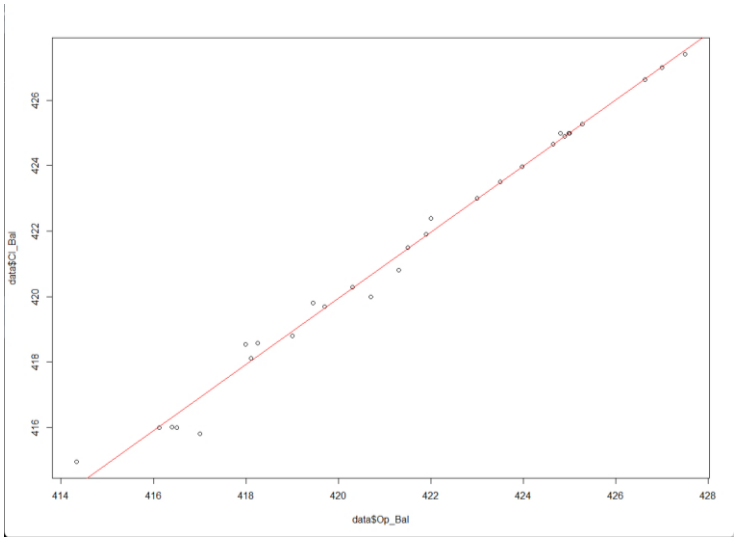
Inference:

The relationship between the Highest and lowest prices and the relationship between the highest price and volume has been derived and it is positive over the time period.

6.3LINEAR REGRESSION:

- Regression is used to determine the strength of the relationship between one dependent variable and one independent variable.
- Here, the linear regression is applied to identify the strength of relationship between the opening and closing prices of the stocks of the company.

```
> data<-read.csv(file="C:\\Users\\SM CORPORATES\\OneDrive\\Desktop\\DCS - SEM 3\\R PROJECT\\2020-22.csv", header=TRUE)
> reg <- lm(data$Cl_Bal ~ data$Op_Bal)
> reg

Call:
lm(formula = data$Cl_Bal ~ data$Op_Bal)

Coefficients:
(Intercept)  data$Op_Bal
     -4.259        1.010

> print(summary(reg))

Call:
lm(formula = data$Cl_Bal ~ data$Op_Bal)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09582 -0.02147  0.01508  0.05414  0.73081

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.25933    7.60055   -0.56     0.58
data$Op_Bal  1.01001    0.01803   56.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3526 on 29 degrees of freedom
Multiple R-squared:  0.9908,    Adjusted R-squared:  0.9905
F-statistic:  3138 on 1 and 29 DF,  p-value: < 2.2e-16

> plot(data$Op_Bal,data$Cl_Bal)
> model2=lm(data$Cl_Bal ~ data$Op_Bal,data = data)
> abline(model2, col="red")
```



Inference:

When the correlation is positive, the regression slope will be positive. The opening and closing prices have strong effects on each other.

6.4 BARPLOT

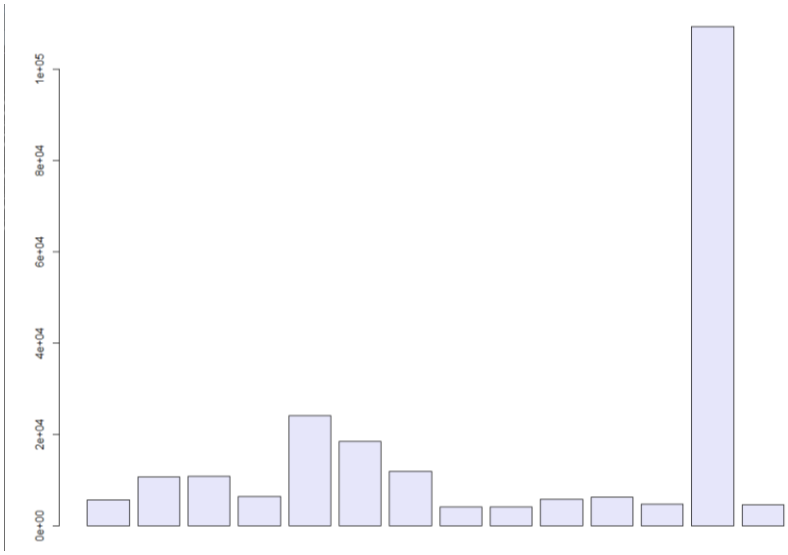
Plotting bar chart using subset to show the variations in prices and volume of stock on a specific condition.

```
> data<-read.csv(file="C:\\Users\\SM CORPORATES\\OneDrive\\Desktop\\DCS - SEM 3\\R PROJECT\\DATA .csv", header=TRUE)
> d2 <- subset(data, data$volume >= 4000)
> d2
          Date Op_Bal Cl_Bal Highest Lowest Volume
3   30-09-2002 433.51 435.03  435.40 433.51   5700
12  30-06-2003 430.10 431.87  431.87 429.00  10680
19  30-01-2004 424.00 423.00  424.00 423.00  10930
33  28-02-2005 429.99 430.00  430.00 429.88   6450
54  30-11-2006 431.00 430.50  431.52 429.16  24190
57  28-02-2007 419.00 418.50  419.00 418.00  18480
68  31-01-2008 417.00 413.02  417.00 413.02  11960
84  29-05-2009 418.51 420.00  420.00 418.51   4160
105 28-02-2011 412.00 412.00  412.00 412.00   4200
157 30-06-2015 409.00 408.02  409.00 408.02   5800
162 30-11-2015 409.13 411.90  412.00 408.98   6380
180 31-05-2017 415.50 415.00  415.50 415.00   4750
181 30-06-2017 432.01 435.47  440.21 432.01 109240
204 31-05-2019 423.01 423.00  423.90 423.00   4630
```

##Plot for Volume

print(barplot(d2$Volume, col='lavender'))

The bar plot representing the volume of stock greater than 4000 over the years.



##Plot for Highest Price

print(barplot(d2$Highest, col='darkmagenta'))

The bar plot representing the highest prices of stock greater than 4000 over the years.

Inference:

The bar plot representing the volume of stock greater than 4000 and the highest prices for the given range has a positive and increasing effect on the stock.

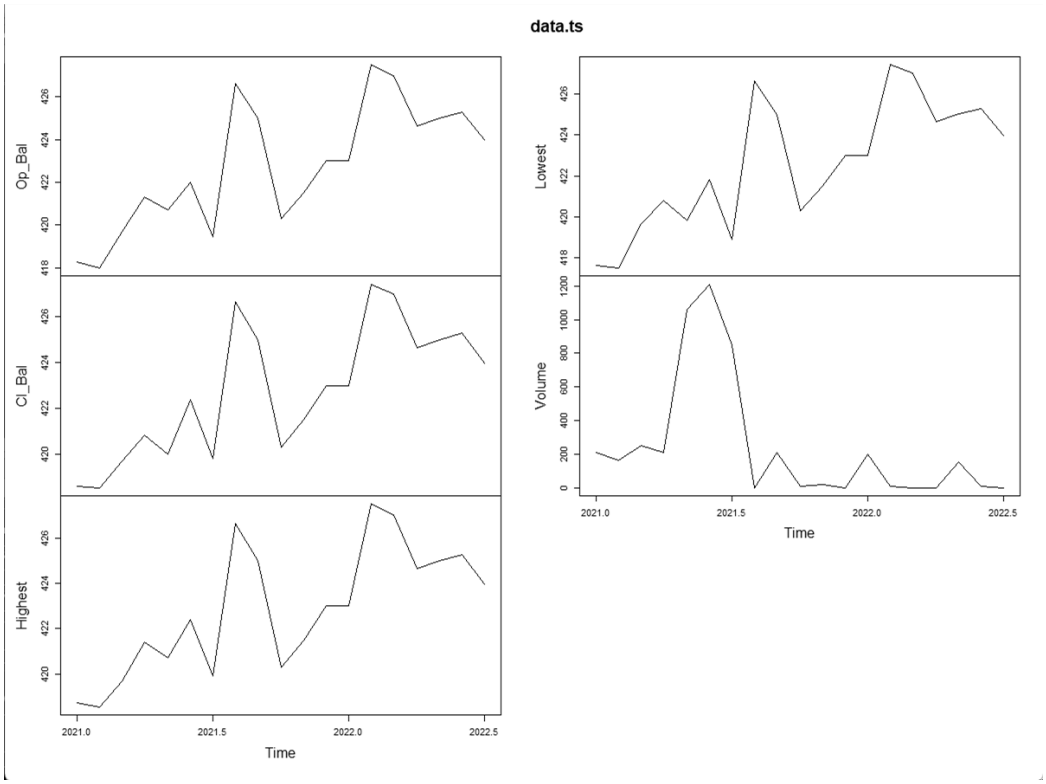
6.5 TIME SERIES ANALYSIS - PLOTTING

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time.

Here, all the variables are analyzed over a specific period of time i.e., from 1st January, 2021 to 31st July, 2022.

```
> data<-read.csv(file="C:\\Users\\SM CORPORATES\\OneDrive\\Desktop\\DCS - SEM 3\\R PROJECT\\2020-22.csv", header=TRUE)
> data.ts = ts(data, frequency=12, start=c(2021,1), end=c(2022,7))
> data.ts
         Op_Bal Cl_Bal Highest Lowest Volume
Jan 2021 418.26 418.59  418.75 417.65    210
Feb 2021 417.99 418.54  418.54 417.51    160
Mar 2021 419.70 419.70  419.70 419.66    250
Apr 2021 421.30 420.81  421.40 420.81    210
May 2021 420.70 420.00  420.70 419.81   1060
Jun 2021 422.00 422.39  422.39 421.80   1210
Jul 2021 419.45 419.80  419.90 418.89    850
Aug 2021 426.63 426.63  426.63 426.63      0
Sep 2021 424.99 425.00  425.00 424.99    210
Oct 2021 420.30 420.30  420.30 420.30     10
Nov 2021 421.50 421.50  421.50 421.50     20
Dec 2021 423.00 423.00  423.00 423.00      0
Jan 2022 423.00 423.00  423.00 423.00    200
Feb 2022 427.50 427.41  427.50 427.41     10
Mar 2022 427.00 427.00  427.00 427.00      0
Apr 2022 424.65 424.65  424.65 424.65      0
May 2022 425.00 425.00  425.00 425.00    150
Jun 2022 425.27 425.27  425.27 425.27     10
Jul 2022 423.98 423.98  423.98 423.98      0
> plot.ts(data.ts)
```

Plot:



Inference:

The time series data analysis give the interpretation on all the variables in the data. According to it, all the variables in the data have a positive slope and is increasing over years.

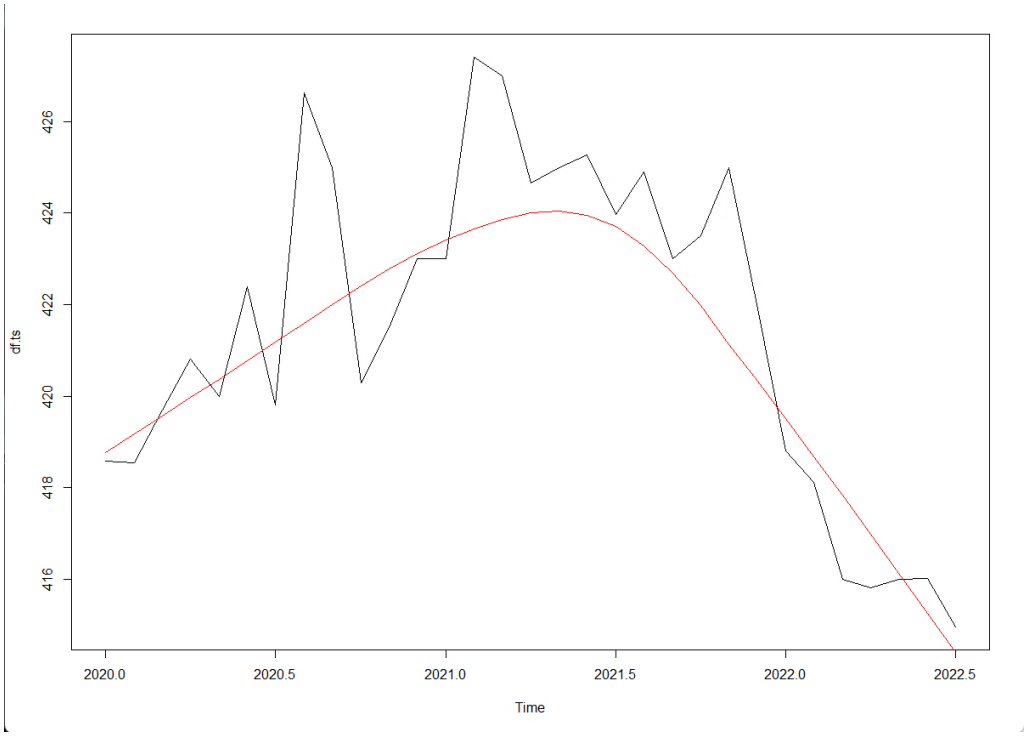
6.6 TRENDLINE ANALYSIS (LINEAR) - PLOTTING

A linear trendline usually shows whether the variable is increasing or decreasing at a steady rate. Data is linear if the pattern in its data points resembles a line.

Here, the linear trend is employed to gain the trendline of the closing prices over the last 3 years (2020-2022).

Code and Output:

```
data<-read.csv(file="C:\\Users\\SM CORPORATES\\OneDrive\\Desktop\\DCS - SEM 3\\R PROJECT\\2020-22.csv", header=TRUE)

df.ts = ts(data$Cl_Bal, frequency=12, start=c(2020,1), end=c(2022,7))

plot.ts(df.ts)

lines(lowess(time(df.ts), df.ts), col="red")
```



Inference:

The trendline analysis gives the clarity that the clossing prices of the company's stock has been decreasing heavily for the past three years.

6.7FORECASTING – PLOTTING

The holt's forecast method is used to predict the future of the stock of the company.

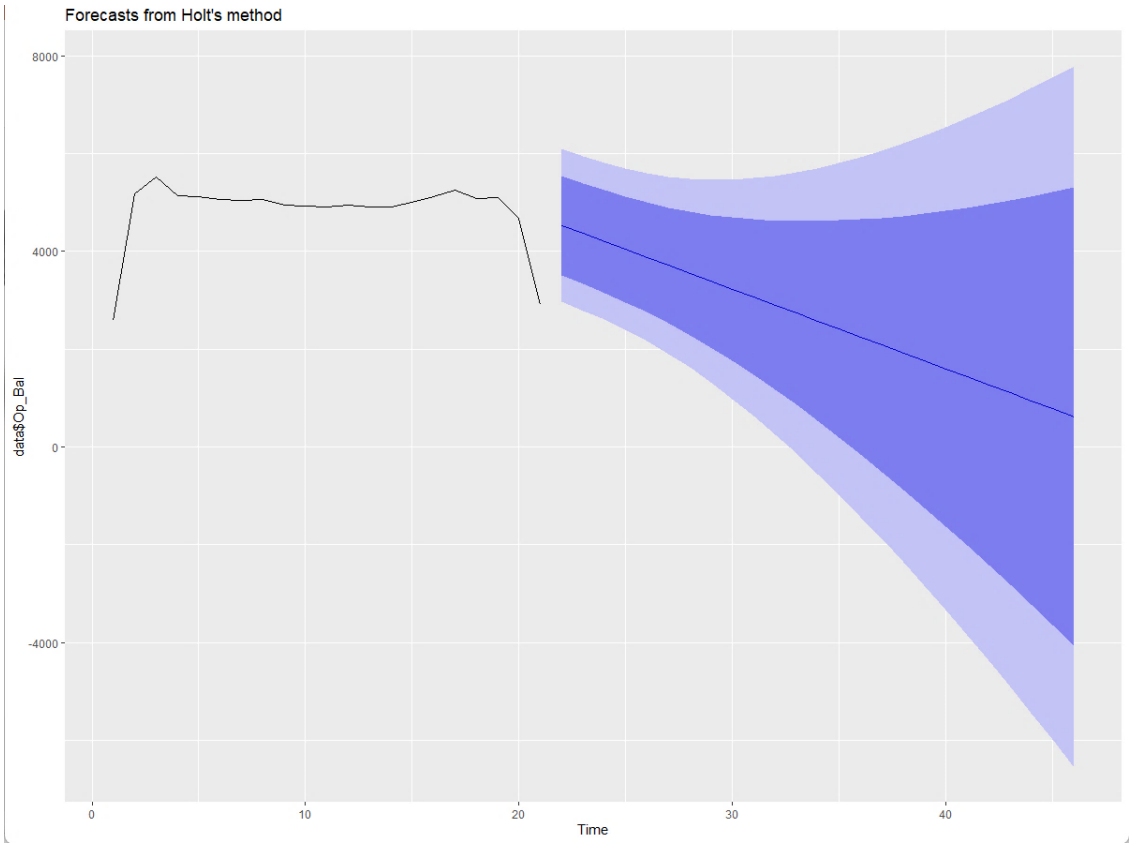
The forecast method depicts that the opening price for the stock will get reduced drastically in the next 20 years.

Code and Output:

```
library(forecast)

data<-read.csv(file="C:\\Users\\SM CORPORATES\\OneDrive\\Desktop\\DCS - SEM 3\\R PROJECT\\Forecast.csv", header=TRUE)
```

##Predict Opening Price

```
holt_mod <- holt(data$Op_Bal, h = 25)

summary(holt_mod)

autoplot(holt_mod)
```

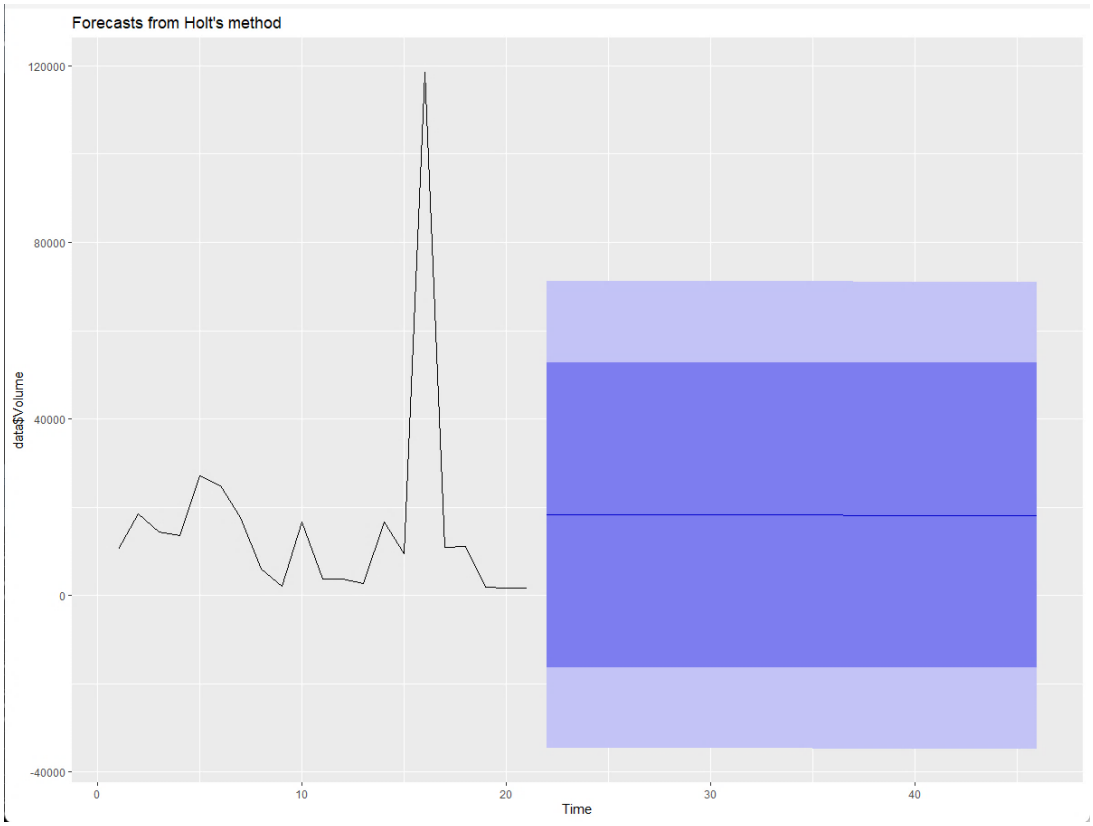


Inference:

The forecasting method which represents the opening prices of the stock states that the opening prices of the stock is going to decrease gradually in the next 25 years.

The forecast method depicts that the volume of the stock will be stable for the next 20 years.

```
##Predict Volume
holt_mod <- holt(data$Volume, h = 25)
summary(holt_mod)
autoplot(holt_mod)
```



Inference:

The forecasting method which represents the volume of the stock states that the volume of the stock is going to stable for the next 25 years.

6.8 Wilcoxon sign-rank test:

It is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to estimate whether their population means ranks differ e.g it is a paired difference test.

Opening_Balance and Closing_Balance → Variables Used

H0: The Stock Analysis for Opening_Balance and Closing_Balance are significantly the same across the market trends.

H1: The Stock Analysis for Opening_Balance and Closing_Balance are significantly not the same (different) across the market trends.

Null Hypothesis: The median difference between pairs of observation is zero.

Alternate Hypothesis: The median difference between pairs of observation is not zero.

NOTE:

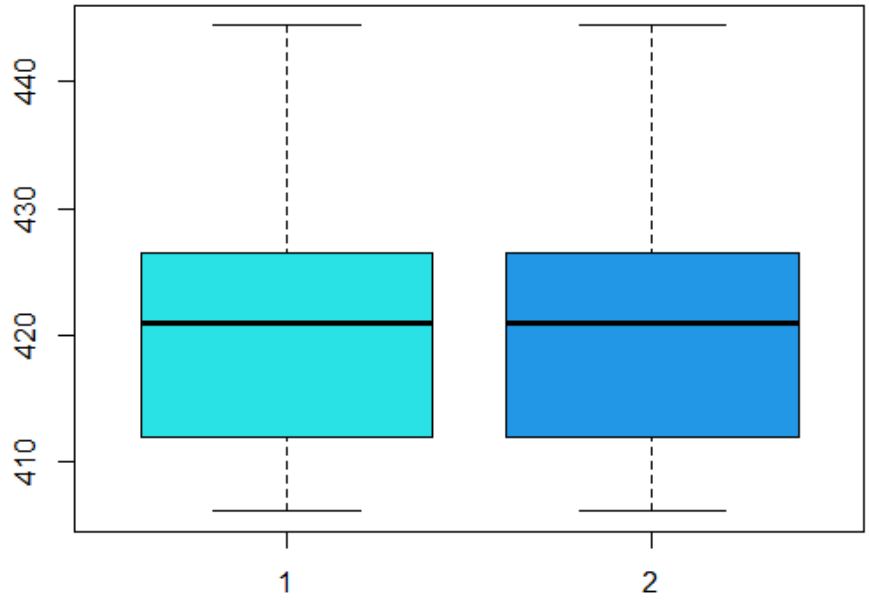
1) If [p_value < 0.05] H0(Null Hypothesis) is rejectd and H1(Alternate Hypothesis) is accepted.
2) If [p_value > 0.05] H0(Null Hypothesis) is accepted and H1(Alternate Hypothesis) is rejectd.

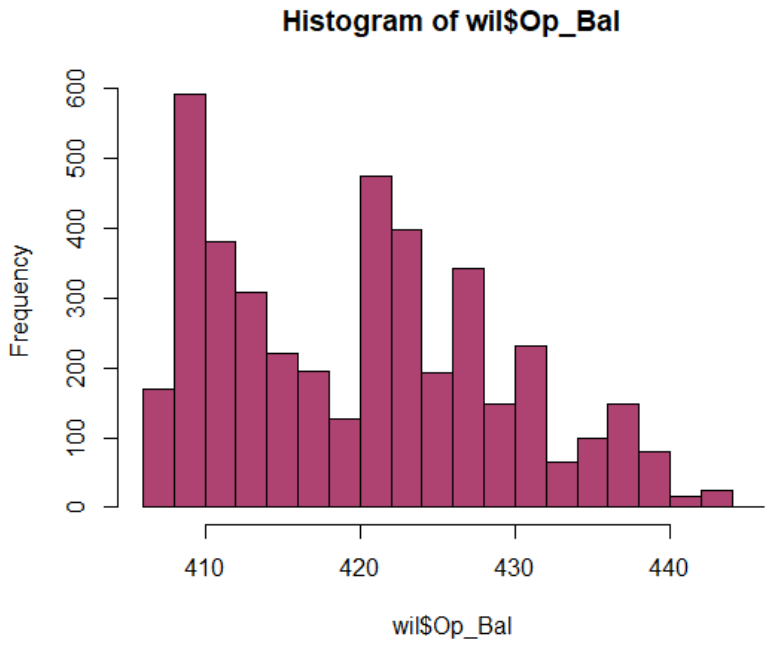
Inference:

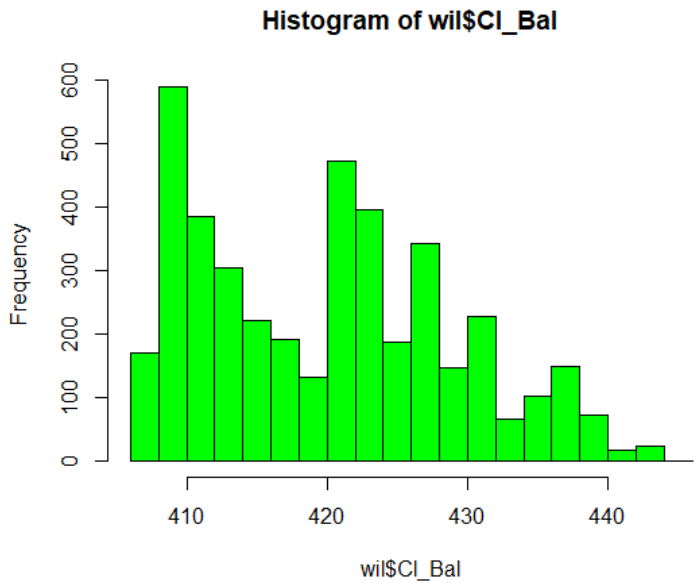
> H0(Null Hypothesis) is accepted and H1(Alternate Hypothesis) is rejectd.

```
        Wilcoxon signed rank test with continuity correction

data:  first and second
V = 1622108, p-value = 0.1099
alternative hypothesis: true location shift is not equal to 0
```

> To check whether the data is normally distributed **Boxplot and histogram** graphs have been used.







- Boxplot and Histogram are used to test the normality of the variables used for analysis.
- If the median line in the boxplot is at the centre, then the normality is accepted.
- If the median line in the boxplot lies below or above the centre line then the normality is rejected.
- Here, the normality condition is rejected.

Therefore, the Wilcoxon signed ranked test reflects that there is **no significant difference** in the **Opening_Balance and Closing_Balance** of the stocks.

```
> library("psych")
> describe(first)
   vars    n  mean   sd median trimmed   mad    min   max range skew kurtosis   se
X1    1 4202 420.1 9.06 420.91  419.48 11.73 406.11 444.4 38.29 0.39    -0.76 0.14
> describe(second)
   vars    n   mean   sd median trimmed   mad    min    max range skew kurtosis   se
X1    1 4202 420.08 9.07  420.9  419.47 11.71 406.11 444.39 38.28 0.39    -0.77 0.14
```

Hence, **Null Hypothesis** is accepted.

6.9Mann-Whitney-U Test:

>It is used **to** compare whether **there is a difference in the dependent variable for two independent groups.**

>It compares whether the distribution of the **dependent variable is the same for the two groups.**

Highest and Lowest -> variables used

H0: The Stock Analysis for **Highest and Lowest** are significantly the same across the market trends.

H1: The Stock Analysis for **Highest and Lowest** are significantly not the same (different) across the market trends.

Null Hypothesis: The median difference between pairs of observation is zero.

Alternate Hypothesis: The median difference between pairs of observation is not zero.

NOTE:

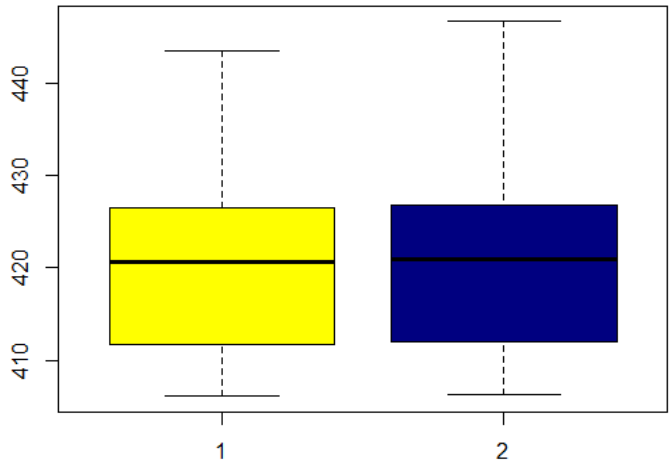
1) If [p_value < 0.05] H0(Null Hypothesis) is rejectd and H1(Alternate Hypothesis) is accepted.
2) If [p_value > 0.05] H0(Null Hypothesis) is accepted and H1(Alternate Hypothesis) is rejectd.

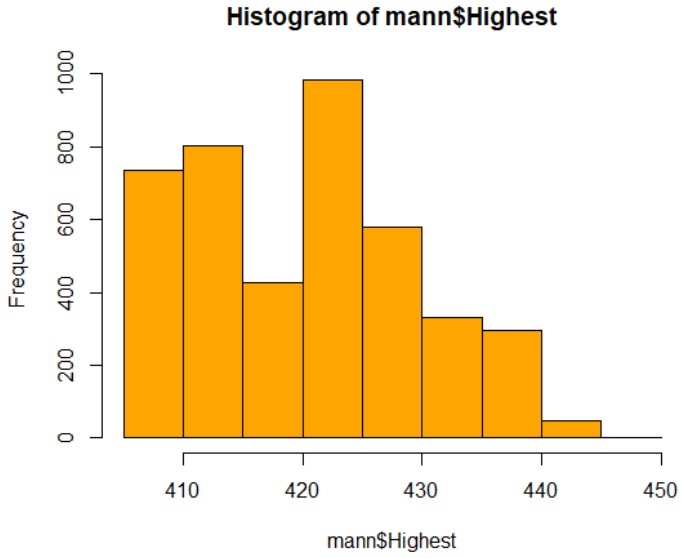
Inference:

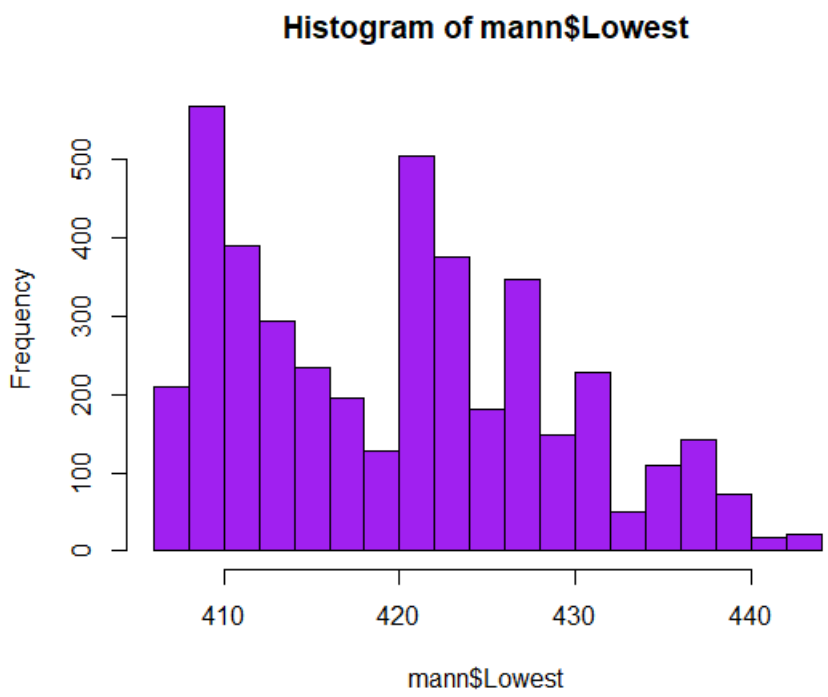
```
        Wilcoxon rank sum test with continuity correction

data:  low_val and high_val
W = 8538143, p-value = 0.009051
alternative hypothesis: true location shift is not equal to 0
```

> H0(Null Hypothesis) is accepted and H1(Alternate Hypothesis) is rejected.

- To check whether the data is normally distributed **Boxplot and histogram** graphs have been used.







- Boxplot and Histogram are used to test the normality of the variables used for analysis.
- If the median line in the boxplot is at the centre, then the normality is accepted.
- If the median line in the boxplot lies below or above the centre line then the normality is rejected.
- Here, the normality condition is rejected.

Therefore, Wilcoxon signed ranked test reflects that is **no significant difference** in the **Opening_Balance and Closing_Balance** from the Stock Analysis.

```
> describe(high_val)
   vars    n   mean   sd median trimmed   mad    min   max range skew kurtosis   se
X1    1 4202 420.32 9.11    421   419.7 11.86 406.21 446.7 40.49 0.39    -0.75 0.14
> describe(low_val)
   vars    n   mean   sd median trimmed   mad    min   max range skew kurtosis   se
X1    1 4202 419.86 9.01 420.62  419.25 11.59 406.06 443.5 37.44 0.39    -0.77 0.14
>
```

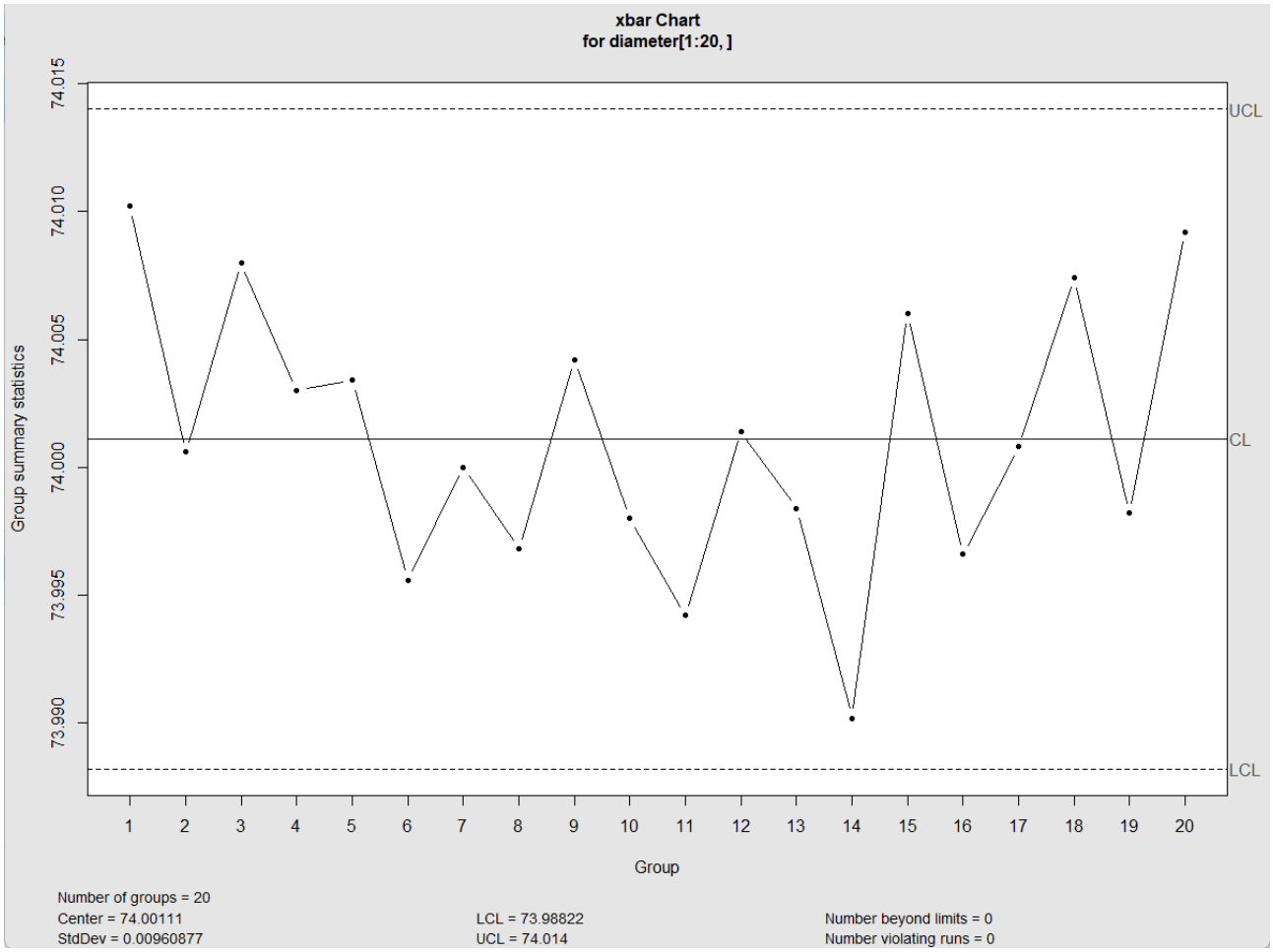
Hence, **Null Hypothesis** is accepted.

6.10 Control Charts Using qcc package:

Code and Output:

- **X-bar Chart:**

```
library(qcc)

data<-read.csv(file="C:\\Users\\SM CORPORATES\\OneDrive\\Desktop\\DCS - SEM 3\\R PROJECT\\2020-22.csv", header=TRUE)

qcc(data = diameter[1:20, ], type = "xbar")
```

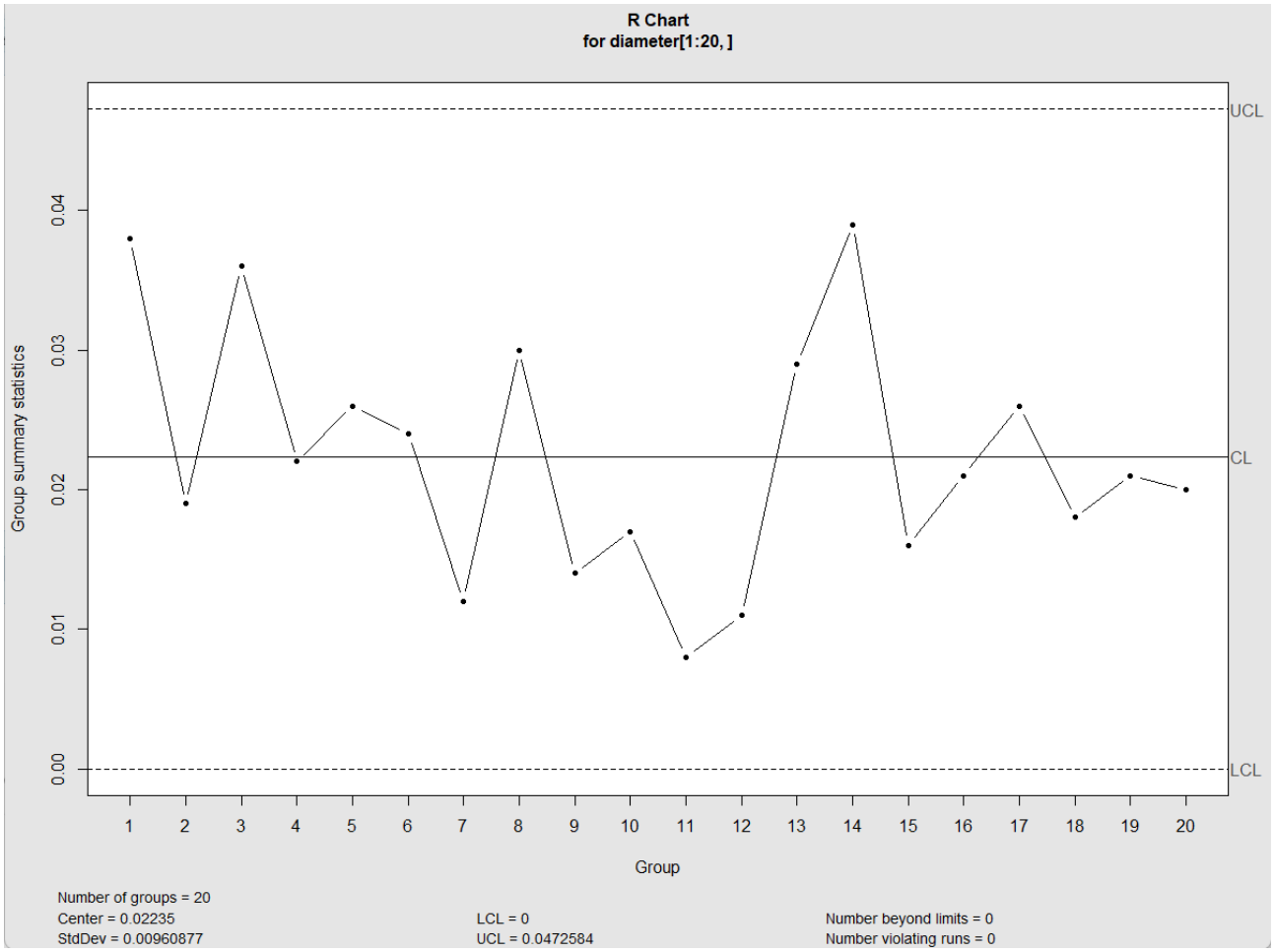


Inference:

The x-bar chart depicting the data of the stocks from 2020-22 is under the statistical control since all the points are lie inside the control limits.

➢ R – Chart:

qcc(data = diameter[1:20,], type = "R")

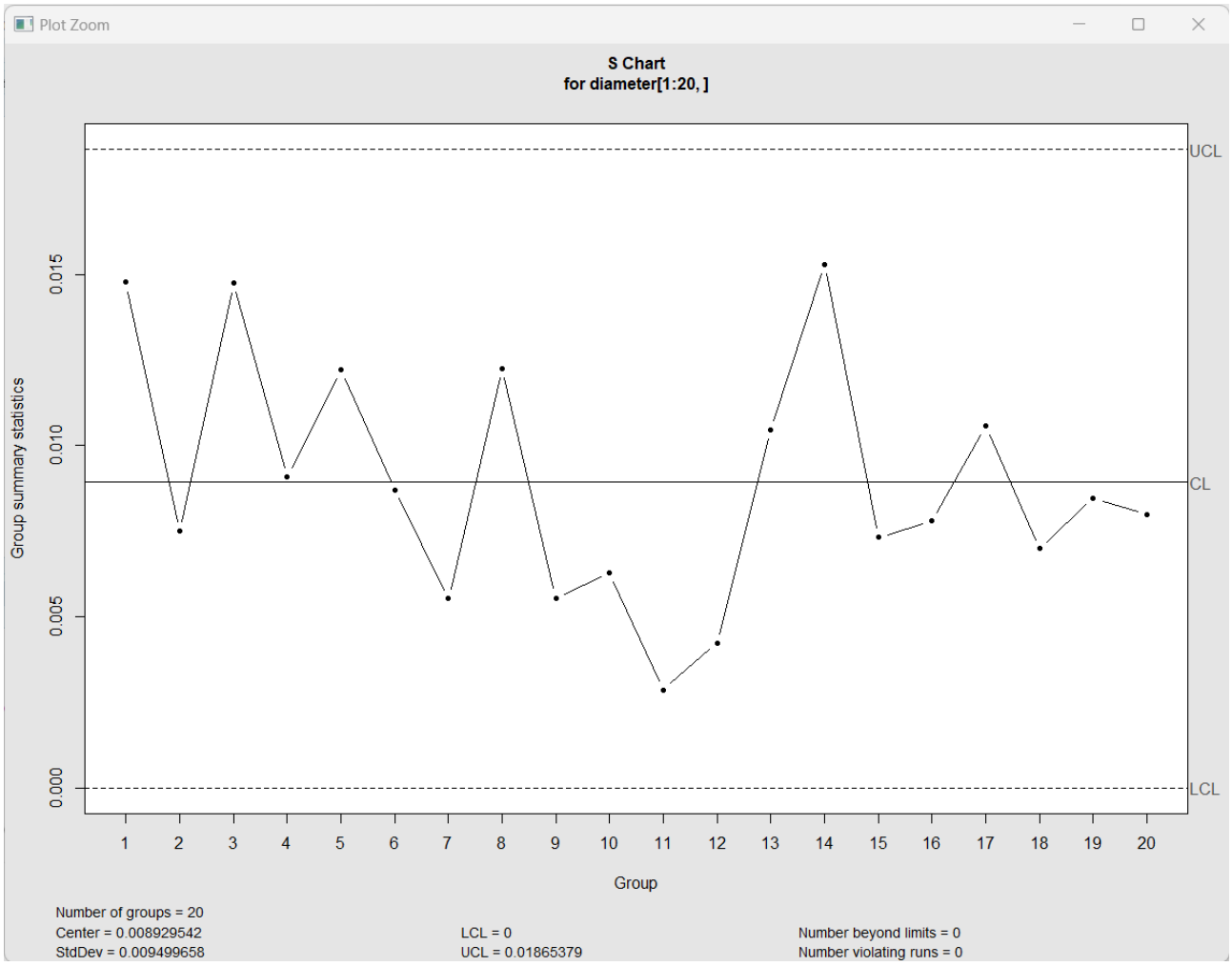


Inference:

The R chart depicting the data of the stocks from 2020-22 is under the statistical control since all the points are lie inside the control limits.

➢ S – Chart:

qcc(data = diameter[1:20,], type = "S")



Inference:

The S chart depicting the data of the stocks from 2020-22 is under the statistical control since all the points are lie inside the control limits.

7. Conclusion:

Proper research in the stock market is never harmful to investment, not just it reduces risk but also guarantees profit to the investors and the company. The statistical forecasting of the future of the stocks and depiction of the current and the past situation of the company has given a clear understanding about the position of the company's stocks in the market. Thus, the analysis and prediction help the investors or traders in making a choice of investing in the market.

8. References:

Basic -> https://towardsdatascience.com/analyzing-stocks-using-r-550be7f5f20d

Control charts -> https://luca-scr.github.io/qcc/articles/qcc.html

Plotting -> https://rpubs.com/markloessi/495609

Examples of various companies -> https://medium.com/codex/stock-market-analysis-with-r-programming-language-c3ab502eb3e7