**COMPUTING LAB-3**

**CAT-2 PROJECT**

**DATA ANALYSIS ON PHARMACEUTICAL SALES FOR DIFFERENT YEARS**

**TEAM MEMBERS:**

2032051 - Yazhini Karthik

2032033 - Riya Thomas

2032043 – Sruthi R

**ABSTRACT:**

The analysis of a dataset that has the details of medicine sales and production for different varieties is done in this project. Initially we have calculated the ANOVA table, both one-way and two-way since the dataset has data for variety of medicines. Under one-way ANOVA multiple pairwise comparison tests are carried out such as a small technique to find the quantity index numbers for the data is also used. To check for any differences, present between the sales for two years, non-parametric tests are implemented. Under that, binomial test is used to calculate the percentage of effect and no effect between the two years of data. Another vital method, Wilcoxon rank sum test is used to check whether the means of both the data are significantly different or not. S chart and x bar charts are implemented to see whether the data is under control or not. Finally, Trend line is fitted to the data and the linear equation is obtained. Hence the end of our project.

**CONCEPTS USED:**

* One-way Anova Test
* Two-Way Anova Test
* Index numbers(quantity)
* X bar and S charts
* Non-parametric test
* Wilcoxon rank sum test
* Trend line fitting

**LIBRARIES USED:**

* multcomp
* car
* qicharts2
* IndexNumber
* ggpubr
* MASS
* rmeta

**CONCEPTS USED AND IMPLEMENTATION:**

**1. ONE WAY ANOVA**

**CONCEPT DESCRIPTION:**

The one-way analysis of variance (ANOVA), also known as one-factor ANOVA, is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups. In one-way ANOVA, the data is organized into several groups base on one single grouping variable (also called factor variable).

**ANOVA test hypotheses:**

1. **Null hypothesis:** the means of the different groups are the same
2. **Alternative hypothesis:** At least one sample mean is not equal to the others.

**Assumptions of ANOVA test:**

Here we describe the requirement for ANOVA test. ANOVA test can be applied only when:

1. The observations are obtained independently and randomly from the population defined by the factor levels
2. The data of each factor level are normally distributed.
3. These normal populations have a common variance.

**IMPLEMENTATION:**

**STEP 1: Import your data into R.**

**STEP 2: Computation of one-way anova**

To compute one way anova, the r function aov() can be used. The function summary.aov() is used to summarize the analysis of variance model.

The output includes the columns f value and pr(>f) corresponding to the p-value of the test.

For our dataset the p value is less than 0.05, which means there is significant difference between the means of the groups.

**STEP 3: MULTIPLE PAIRWISE COMPARISON TEST**

In ANOVA test, a significant p-value indicates that some of the group means are different, but we don’t know which pairs of groups are different.

It’s possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of groups is statistically significant.

1. **Tukey multiple pairwise-comparisons**

As the ANOVA test is significant, we can compute Tukey HSD (Tukey Honest Significant Differences, R function: TukeyHSD()) for performing multiple pairwise-comparison between the means of groups.

The function TukeyHD() takes the fitted ANOVA as an argument.

1. **Multiple comparisons using multcomp package**

It’s possible to use the function glht() [in multcomp package] to perform multiple comparison procedures for an ANOVA. glht stands for general linear hypothesis tests. The simplified format is as follow:

1. **Pairwise t-test**

The function pairewise.t.test() can be also used to calculate pairwise comparisons between group levels with corrections for multiple testing.

By performing these tests, we can find out which two pairs of means are significantly different.

INFERENCE:

Hypotheses testing of the equality of the means and simultaneous confidence intervals for all pair-wise differences of the means are inferred.

**2. INDEX NUMBERS:**

CONCEPT DESCRIPTION:

An index number is a statistical device for measuring changes in the magnitude of a group of related variables. It is a measure of the average change in a group of related variables over two different situations.

IMPLEMENTATION:

1. Only the quantity index is found out for the data that is used for one-way anova, since it has the data of volume(quantity).
2. Importing the library IndexNumber, the index numbers for the data is calculated.
3. The results are plotted in a graph for a visual representation of the changes occurred over time in the data.

INFERENCE:

There are significant changes occurred in the data from to time which is clearly inferred from the plotted graph.

**3.TWO WAY ANOVA**

**CONCEPT DISCRIPTION:**

Two-way ANOVA test is used to evaluate simultaneously the effect of two grouping variables (A and B) on a response variable.

**Two-way ANOVA test hypotheses:**

1. There is no difference in the means of factor A
2. There is no difference in means of factor B
3. There is no interaction between factors A and B

The alternative hypothesis for cases 1 and 2 is: the means are not equal.

The alternative hypothesis for case 3 is: there is an interaction between A and B.

**Assumptions of two-way ANOVA test:**

Two-way ANOVA, like all ANOVA tests, assumes that the observations within each cell are normally distributed and have equal variances. We’ll show you how to check these assumptions after fitting ANOVA.

**IMPLEMENTATION:**

**STEP 1: Import the dataset in R.**

**STEP 2: Check your data**

To get an idea of what the data look like, we need to check the structure pf data

If there are any numeric variable, we need to convert it as factor variable.

**STEP 3: Computation**

To compute two way anova, the r function aov() can be used. The function summary.aov() is used to summarize the analysis of variance model.

The output includes the columns f value and pr(>f) corresponding to the p-value of the test.

For our dataset the p value is greater than 0.05, which means there is no significant difference between the means of the groups.

**STEP 4: Multiple pairwise comparison test**

In ANOVA test, a significant p-value indicates that some of the group means are different, but we don’t know which pairs of groups are different.

It’s possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of groups is statistically significant.

In this dataset, as there is no significant difference between the means of groups there is no need to perform multiple pairwise comparison test.

INFERENCE:

Hypotheses testing and pair- wise differences of the means of different groups are discussed and interpreted.

**4. X BAR CHART AND S-CHART:**

**Concept Description:**

This procedure generates X-bar and s (standard deviation) control charts for variables. The format of the control charts is fully customizable. The data for the subgroups can be in a single column or in multiple columns. This procedure permits the defining of stages. For the X-bar chart, the center line can be entered directly or estimated from the data, or a sub-set of the data. Similarly sigma may be estimated from the data or a standard sigma value may be entered. A list of out-of-control points can be produced in the output, if desired, and means and standard deviation values may be stored to the spreadsheet.

**IMPLEMENTATION:**

* The X-bar chart is a type of Shewhart control chart that is used to monitor the arithmetic means of successive samples of constant size, n. This type of control chart is used for characteristics that can be measured on a continuous scale, such as weight, temperature, thickness etc.
* In this procedure, the data may be in either of two formats. The first data structure option is to have the data in several columns, with one subgroup per row.

STEP 1: First we need to input the data.

STEP 2: Making a plot from the data is quite simple. First we will need to call the ‘qcc’ package, and if this is the first time we’ve used it we’ll need to install it.

STEP 3: Once we’ve called the qcc library, we can use qcc to make an x-bar chart of the data as follows. Simply call the data set, the type of chart to generate.

**INFERENCE:**

Library qcc generates a chart as well as provided information about the chart including the control limits, mean and standard deviation.Both the x bar and s charts infer that all the values of the data are under control.

**5. NON-PARAMETRIC TESTS**:

CONCEPT DESCRIPTION:

Non-parametric tests are methods of statistical analysis that do not require a distribution to meet the required assumptions to be analyzed (especially if the data is not normally distributed). Due to this reason, they are sometimes referred to as distribution-free tests.

IMPLEMENTATION:

We are taking the data for the years 2017 and 2018 and compare them both to check for any differences that happened in those two years in the sales of medicines.

STEP 1: The 2017 data is considered as the before data and the 2018 data is considered as after.

STEP 2: When the after-value is smaller than the corresponding before-value, the data point is a "success." In the days before computers when the Sign Test was calculated manually, it was common to put a '+' sign next to successes and a '-' sign next to a failure.

STEP 3: A function is used to count the number of successes and the number of failures, and displays those counts.

STEP 4: The counts of successes and failures are used to perform a Sign Test, using the built-in R binom.test function, and then prints the percentage of effect and no-effect.

INFERENCE:

Executing the binomial test on number of successes and failures, the percentage of effect is found to be 0.983 and percentage of no-effect is 0.017 between the two data is interpreted. Hence, It is inferred that the there is an effect is in 2018 data compared to 2017 data.

**6. WICOXON TEST(NON-PARAMETRIC TEST):**

The ‘Wilcoxon Rank Sum test’ (also called ‘Mann-Whitney test’), is a distribution-free alternative to the t-test, and is used to test the hypothesis that the distributions in the two groups have the same median.

It is a method of non-parametric test to check the whether there is any significant difference between the two data using hypothesis testing

ASSUMPTIONS:

* If the data is completely normally distributed Wilcoxon rank sum test cannot be carried out.
* But if the normal distribution is irregular we can carry out the test.

IMPLEMENTATION:

STEP 1: Here we’ve taken the data for the medicine M01AB sales alone in 2017 and 2018.Both the data are plotted for a normal distribution. Both of them have an irregular normal distribution. So next steps can be proceeded.

STEP 2: A function wilco.test is used to check whether the calculated p-value is less or greater than the table value.

If p-value greater than table value, it says null hypothesis (i.e) there is no significant change in both the data.On the contrary, the program opts for alternate hypothesis since there is a significant change.

INFERENCE:

The program opts for alternate hypothesis representing there is a true location shift which means there is a significant difference between both the data.

**7.TREND LINES:**

**CONCEPT DESCRIPTION:**

A **trend line**, often referred to as a line of best fit, is a line that is used to represent the behavior of a set of data to determine if there is a certain pattern. A trend line is an analytical tool used most often in conjunction with a scatter plot (a two-dimensional graph of ordered pairs) to see if there is a relationship between two variables.

**IMPLEMENTATION:**

1. First, we need to create some data. Next, we can draw our data.
2. Then we visualize the output of our linear regression by adding the regression line to our line plot.
3. We extract the coefficients of our model using the coefficient function.
4. Finally we print the equation of our linear model using the coefficients.

**INFERENCE:**

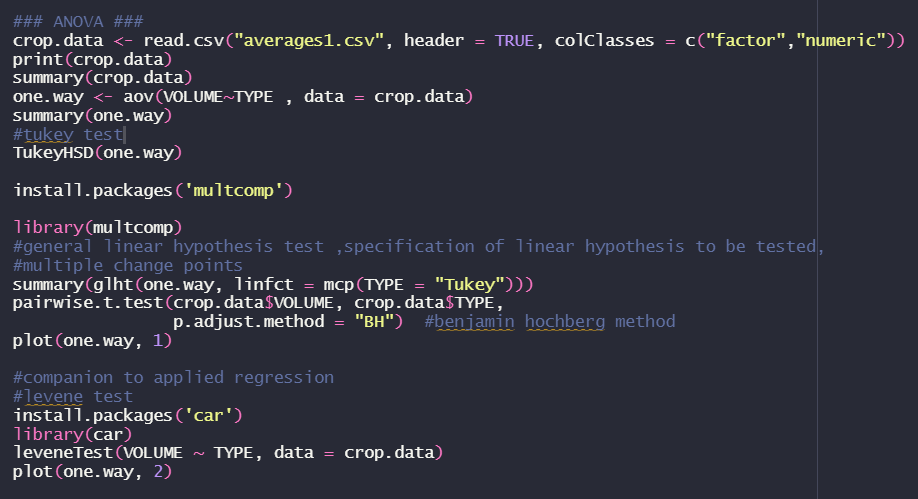
A trend line is fitted using a linear model for the data. The linear equation is also obtained.

**REFERENCES:**

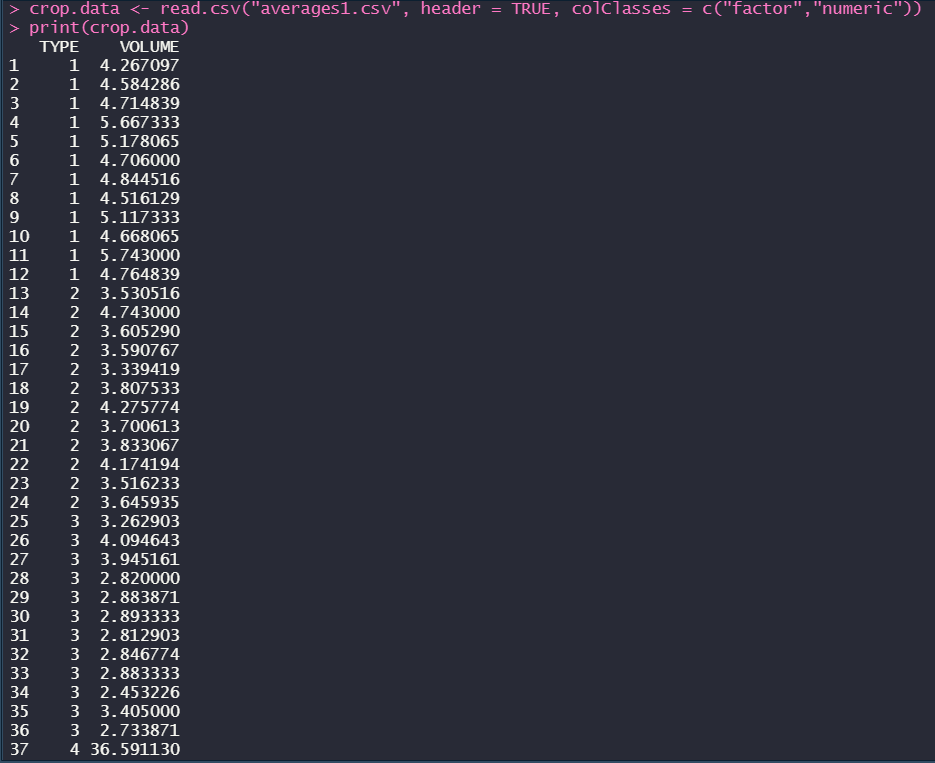
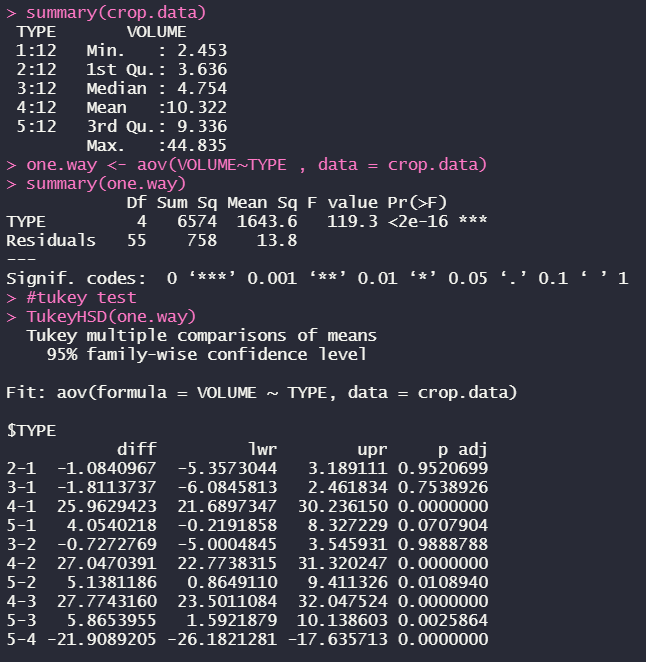
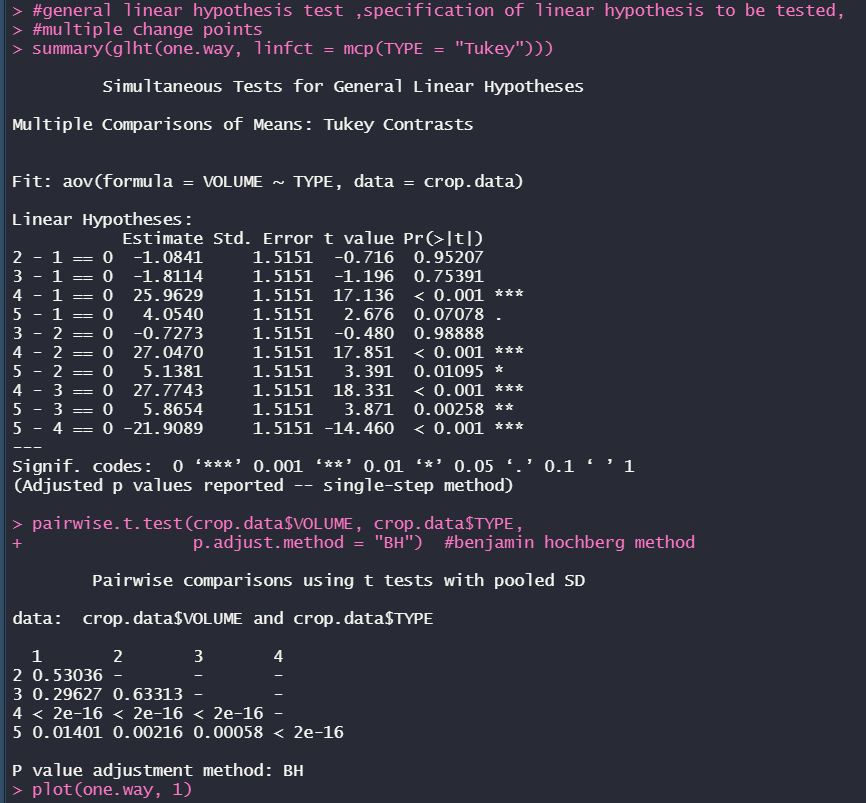
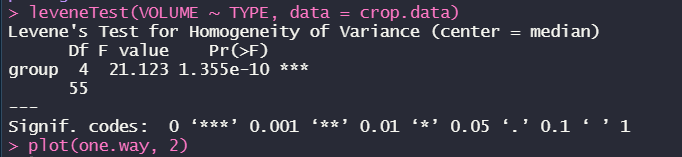
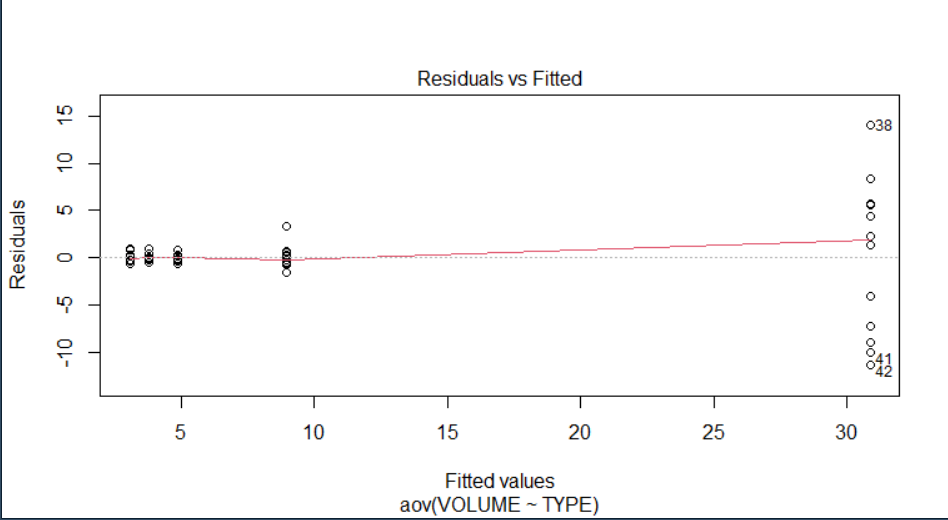
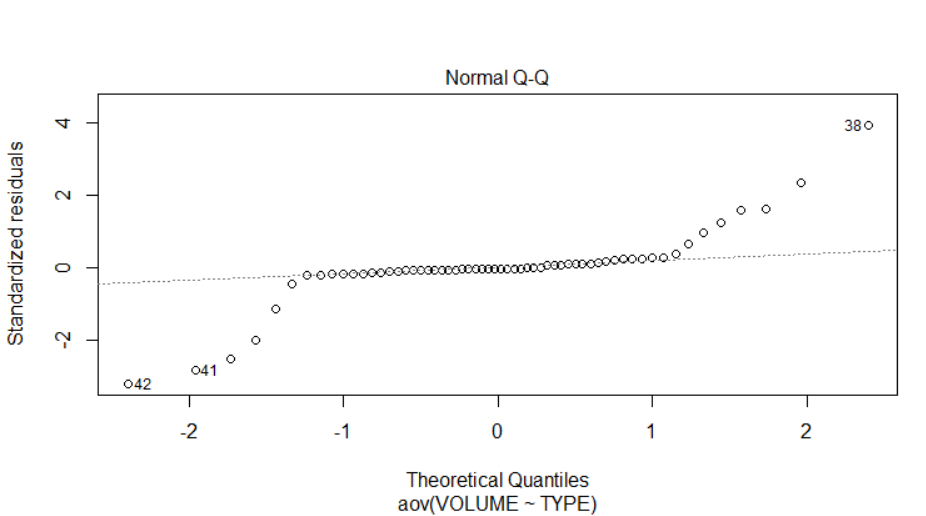
* <http://www.sthda.com/english/wiki/one-way-anova-test-in-r>
* <http://www.sthda.com/english/wiki/two-way-anova-test-in-r>
* <https://rstudio-pubs-static.s3.amazonaws.com/239508_52b8e34ddb9240e1bd3db23f62cf3f98.html>
* <https://statisticsglobe.com/get-equation-of-linear-trend-line-in-r>
* <https://bookdown.org/danieljcarter/r4steph/non-parametric-tests.html>
* <https://visualstudiomagazine.com/articles/2016/10/01/sign-test-using-r.aspx#:~:text=The%20R%20language%20can%20be,%22before%20and%20after%22%20data.&text=The%20Sign%20Test%20is%20most,an%20effect%20of%20some%20sort>

**CODE IMPLEMETATION IN R SCRIPT:**

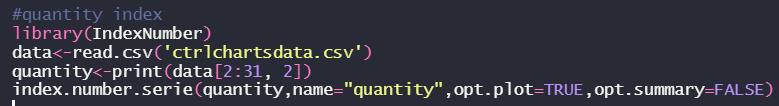
1. **ONE-WAY ANOVA-CODE:**



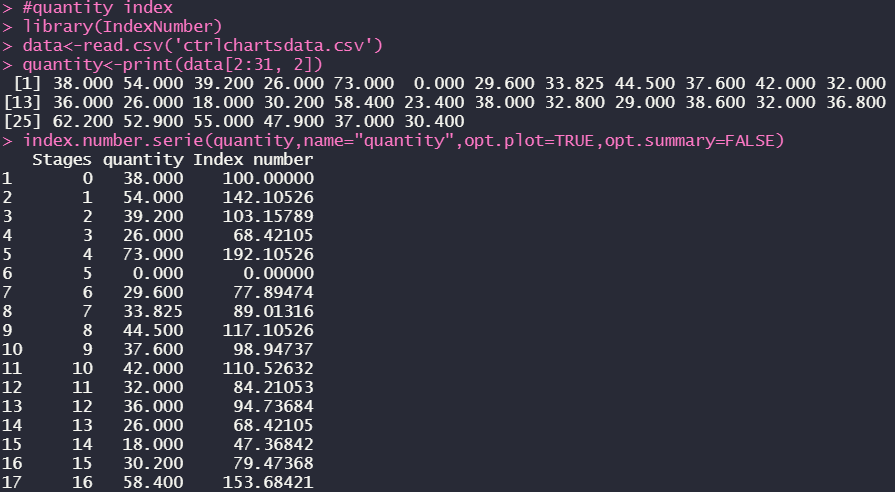
OUTPUT:

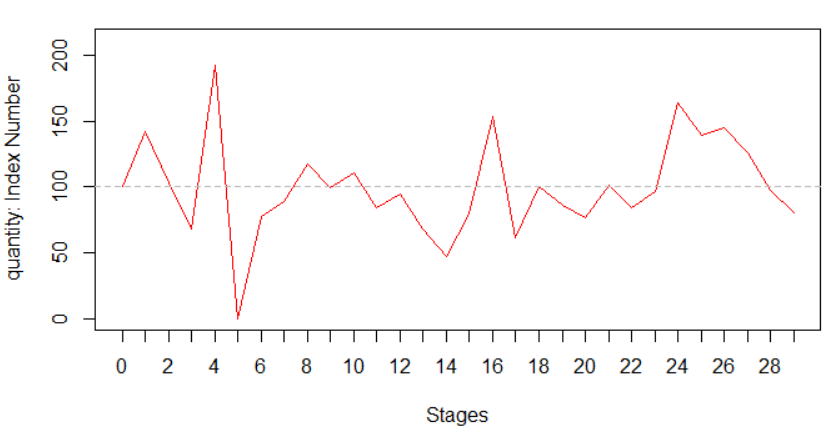
    

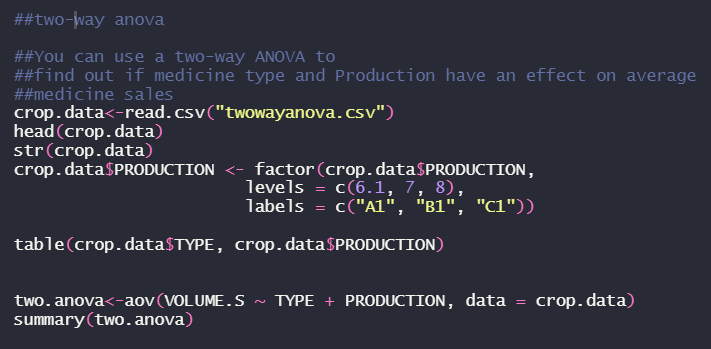
1. **INDEX NUMBERS CODE:**



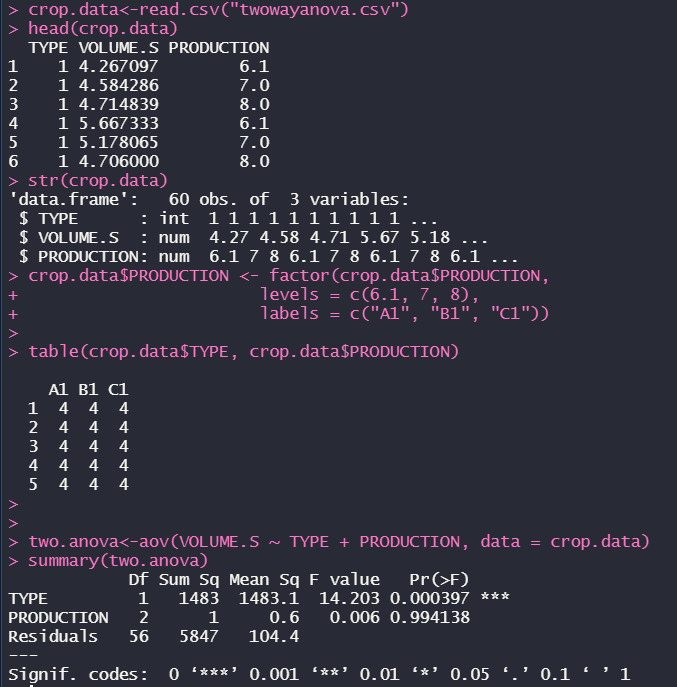
**OUTPUT:**



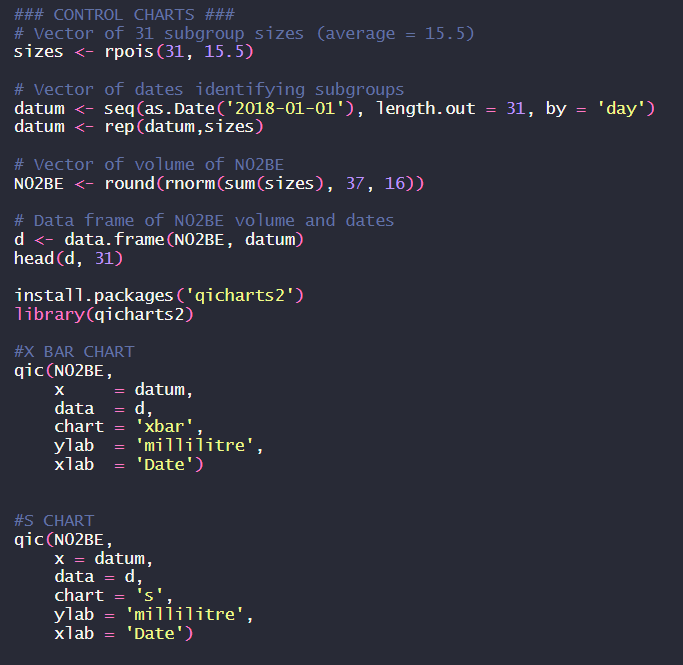


1. **TWO – WAY ANOVA CODE**:

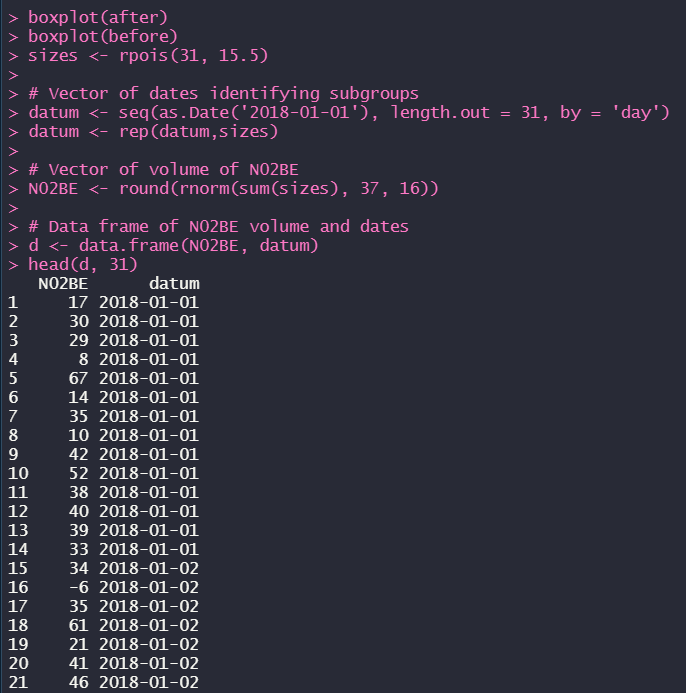
**OUTPUT:**

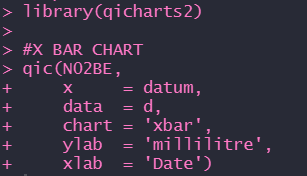


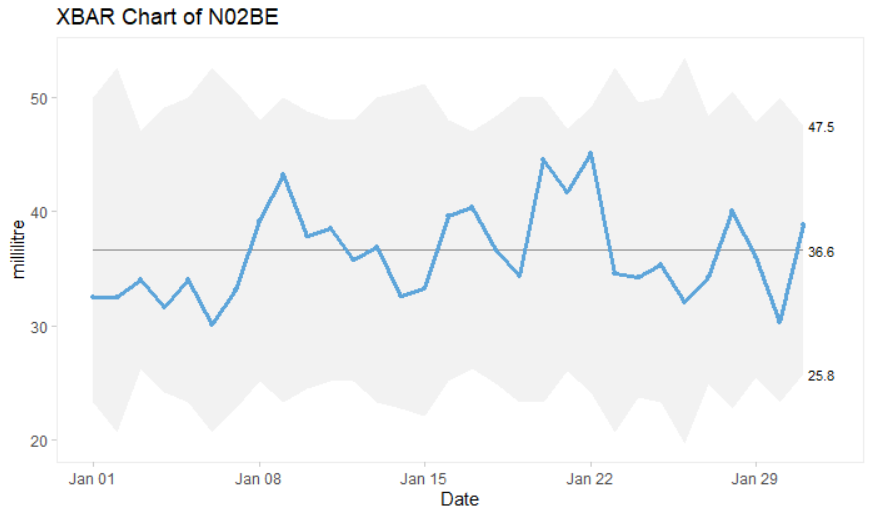
**4. CONTROL CHARTS-CODE:**

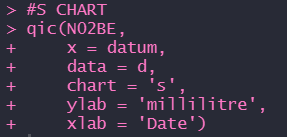


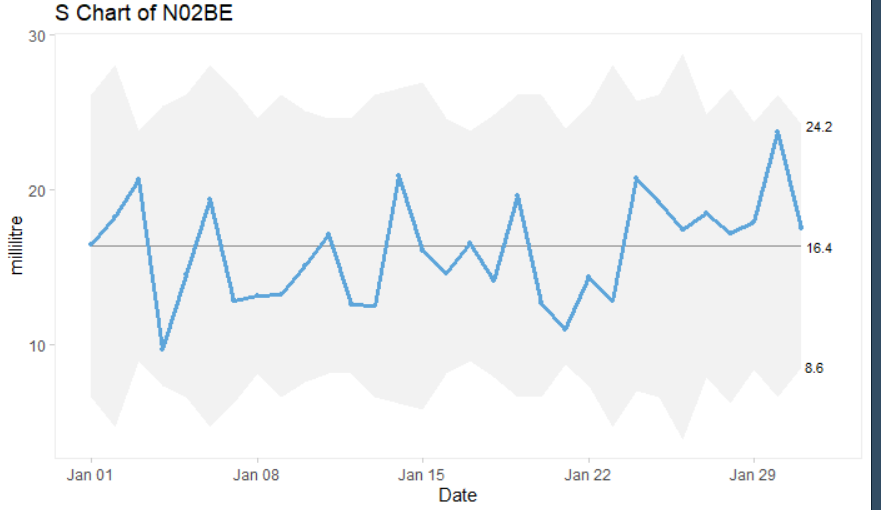
**OUTPUT:**

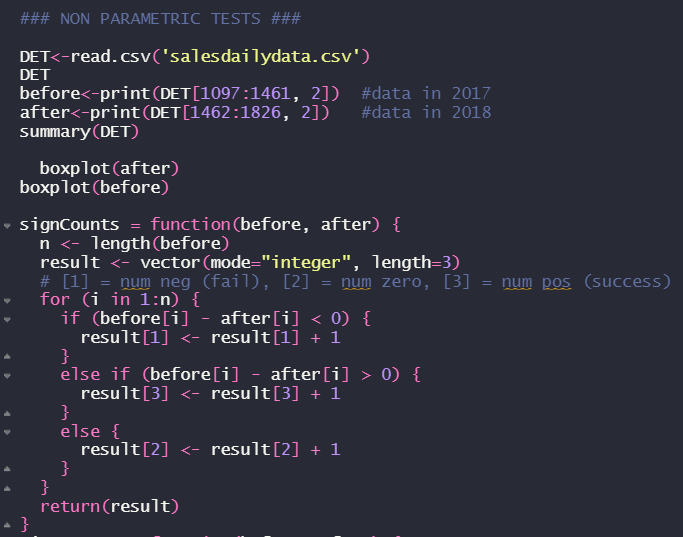
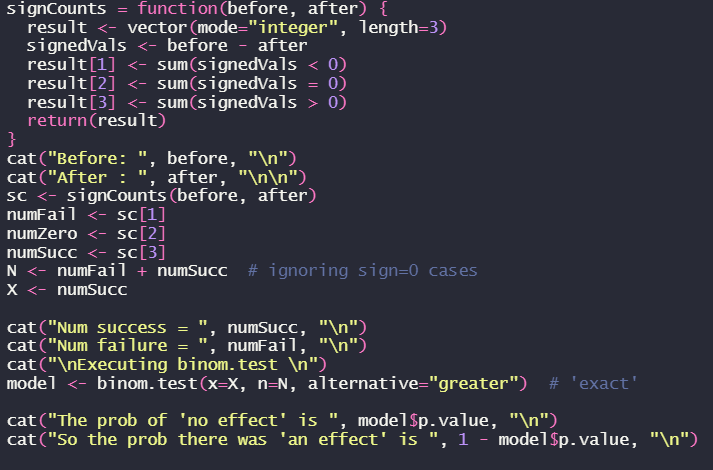




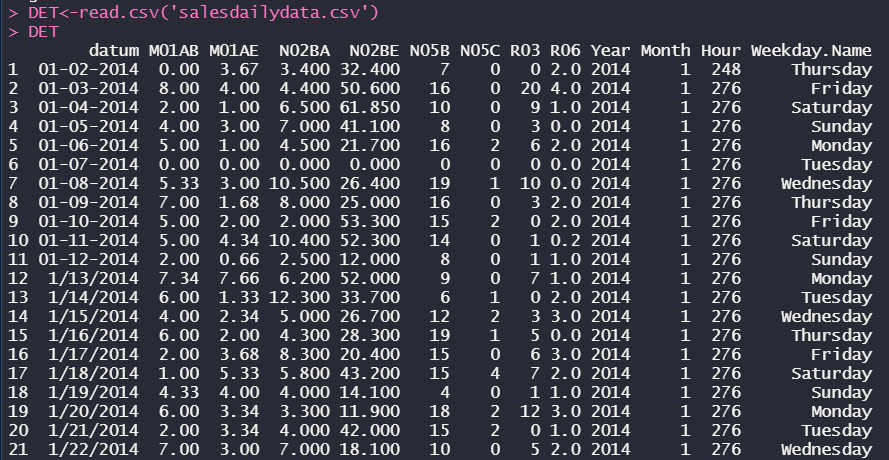


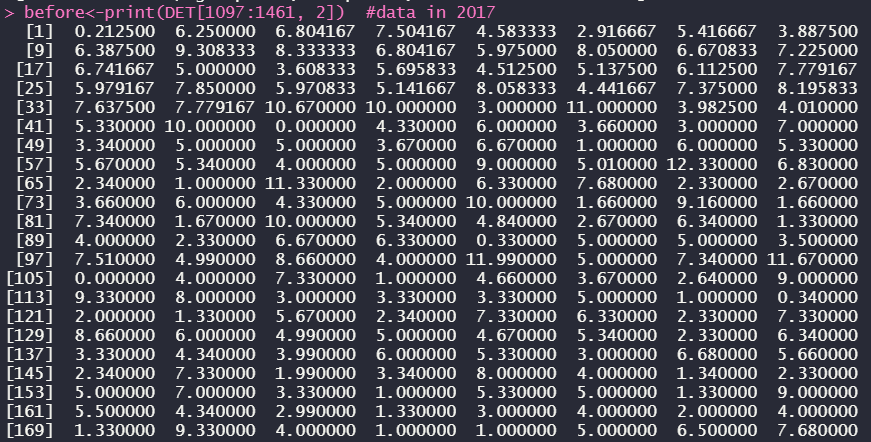


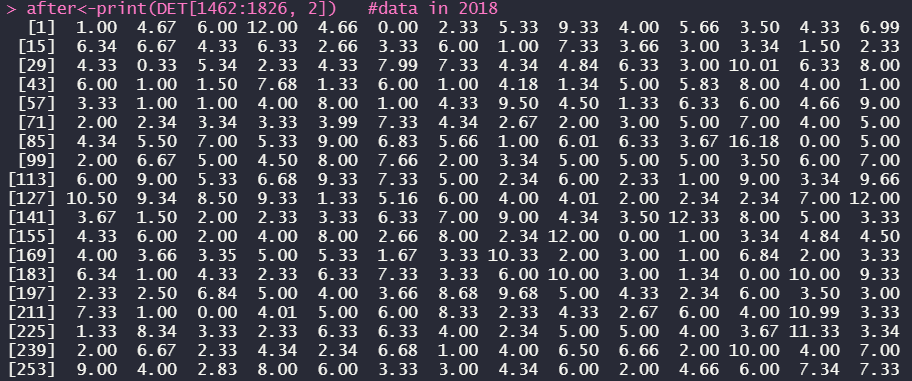


**5.NON-PARAMETRIC TESTS CODE**: 

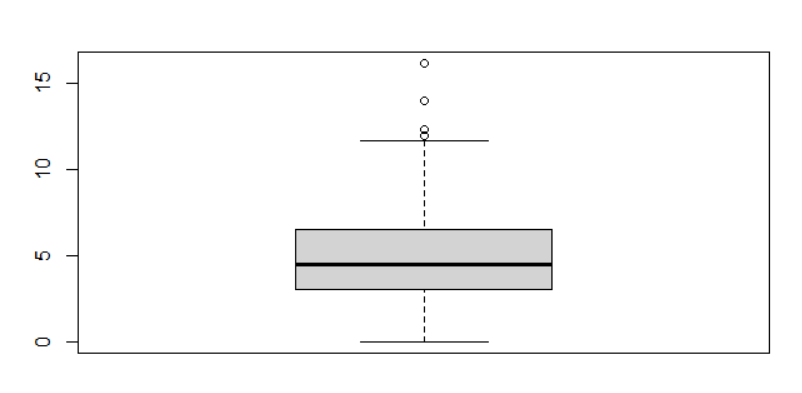
**OUTPUT:**



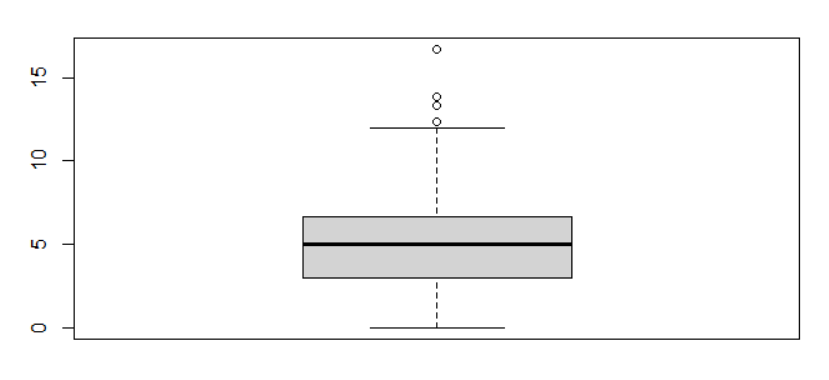


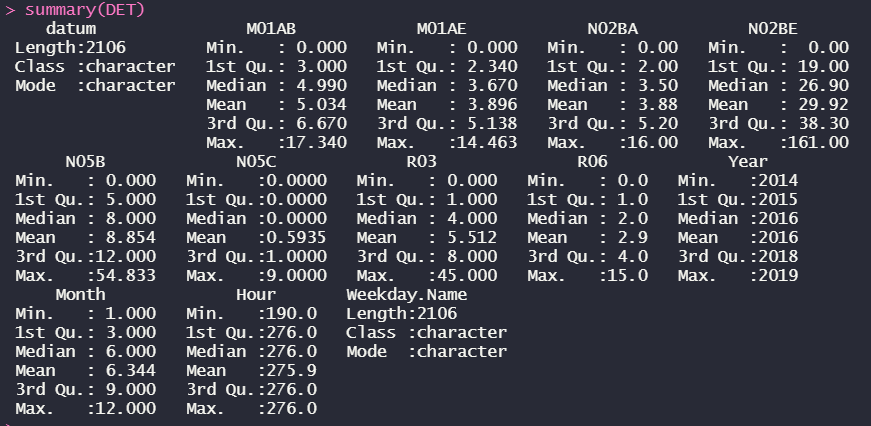


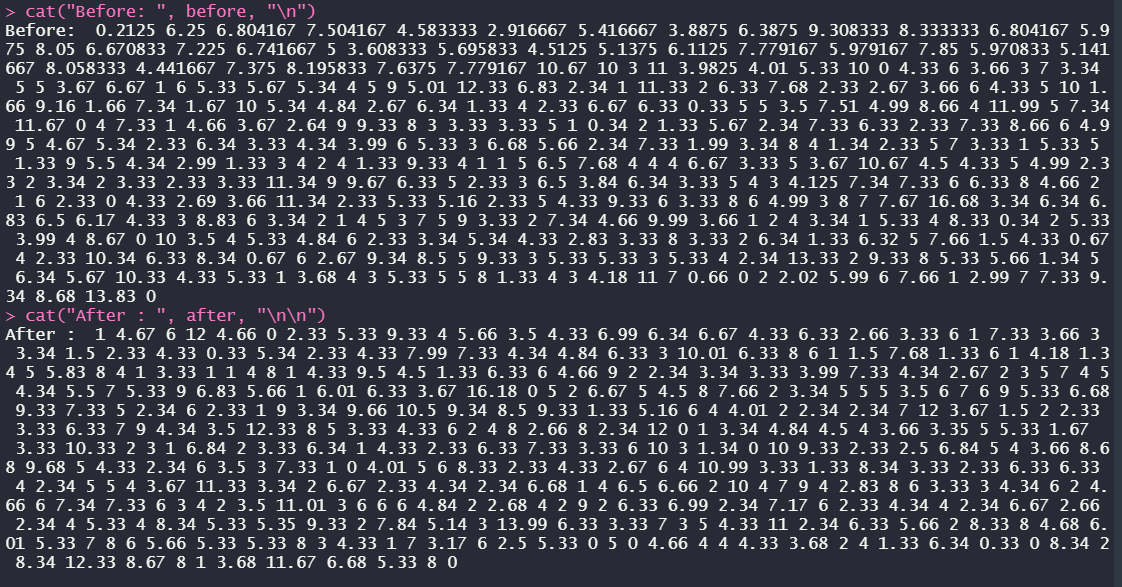
BOX-PLOT FOR BEFORE DATA:

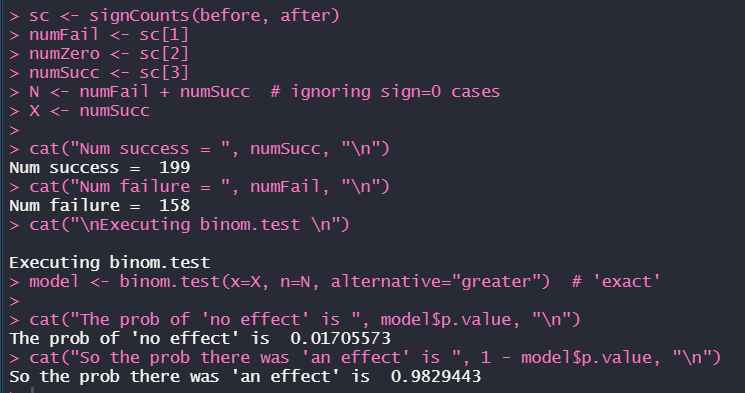


BOX-PLOT FOR AFTER DATA:

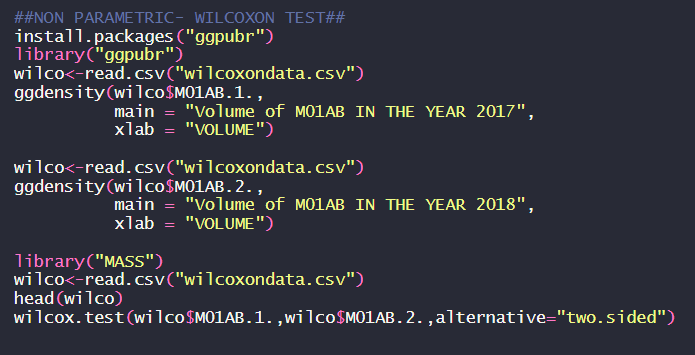




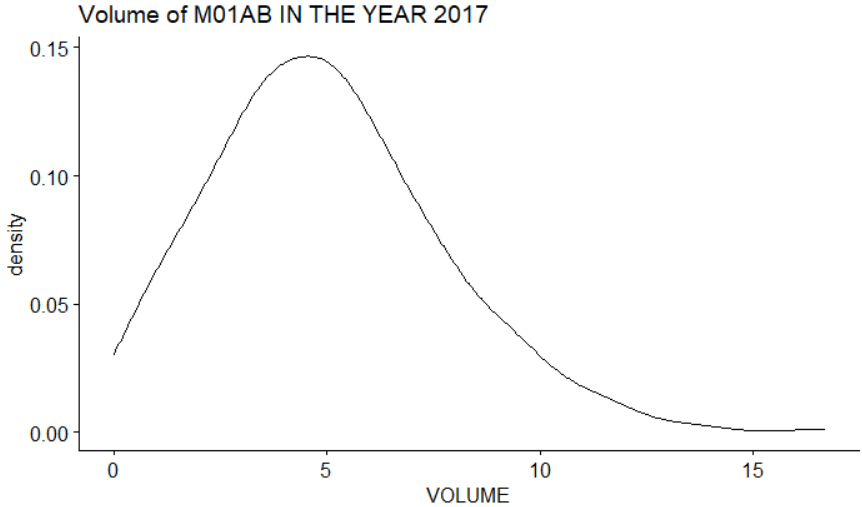


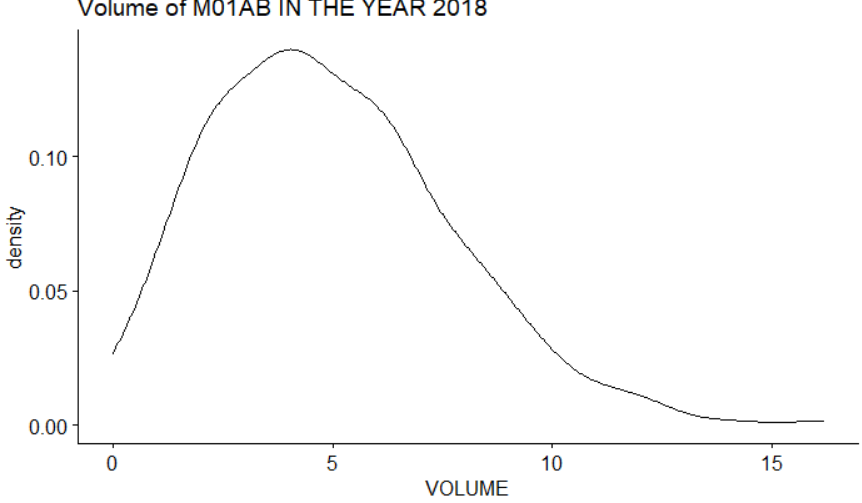


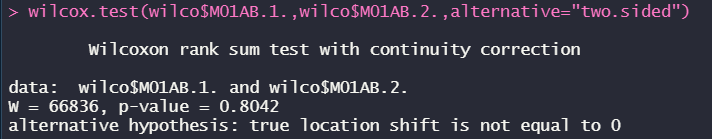
**6.WILCOXON RANK SUM TEST-CODE:**



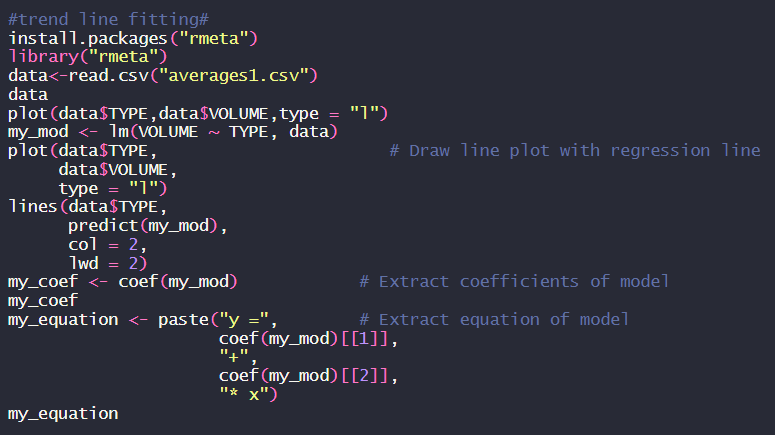
**OUTPUT:**







**7.TREND LINE FITTING-CODE:**



**OUTPUT:**

