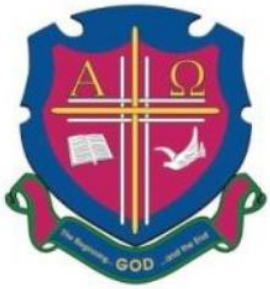


UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

FACULTY OF ENGINEERING DESIGN AND TECHNOLOGY

NAME: MOGA MUZAMIL ABDUL WAHAB

REG NO: S21B23/013

ACCESS NO: A94166

COURSE: BACHELOR OF SCIENCE IN COMPUTER SCIENCE (BSCS)

COURSE UNIT: DATA SCIENCE

LECTURER: DR. DAPHINE NYACHAKI BITALO (PhD GENETICS AND BIOINFORMATICS)

Number 1. Show how you would transform the dataset to exclude missing data in your analysis.

It is part of the R script.

Number 2. A problem was posed wherein the perception of current diamond prices is negative. Is this true?

The perception is false because the sum and the average of the variables is positive.

Sum = 212135217

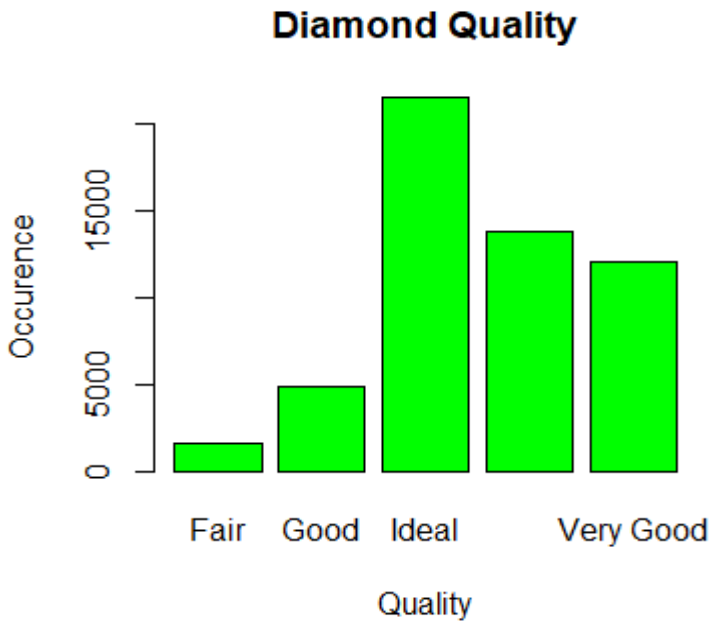
Average = 3933.529

Number 3. It is hypothesized that perceptions on pricing change once the quality of diamonds is understood. Is this true?

This is true because the occurrence changes. And from the hypothesis, the following table was generated in the process of going around the problem.

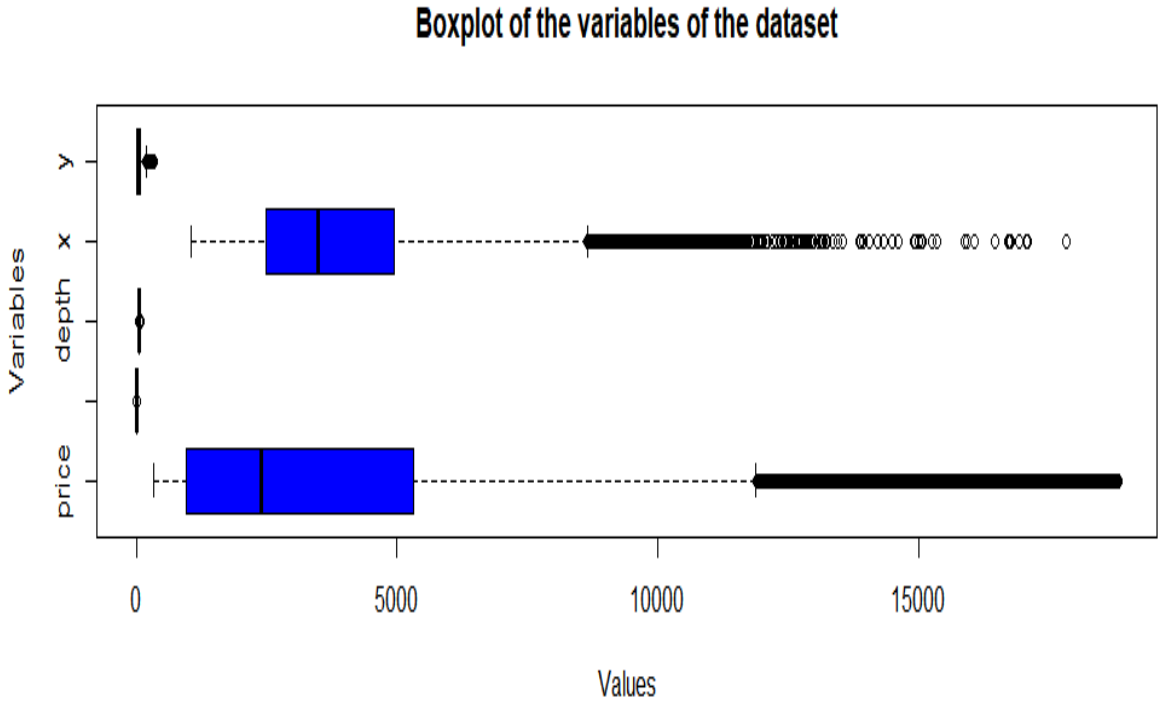
Fair	Good	Ideal	Premium	Very Good
1610	4906	21551	13791	12082

The values are graphically represented To prove the hypothesis and also solve the problem



Number4. Are there any unusual observations within the variables (carat to perception change) of the dataset? Display these outliers if any.

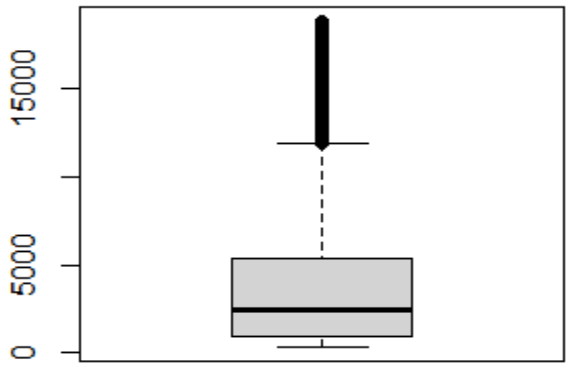
Display of the box plot of all the variable showing the outliers



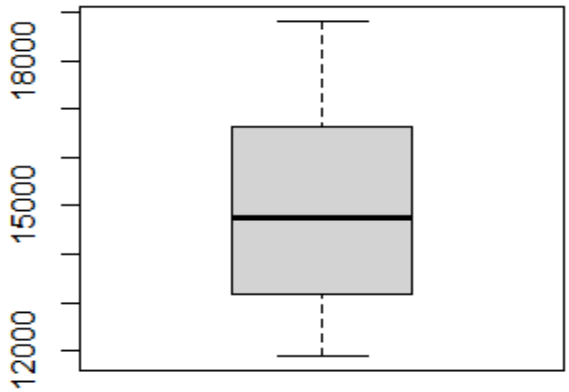
Number 5. How can you filter out the outliers from the dataset?

Taking price as my case study.

The method I used was the Inter Quatile Range method which enabled me remove he outliers using the limits generated to compare with the data.



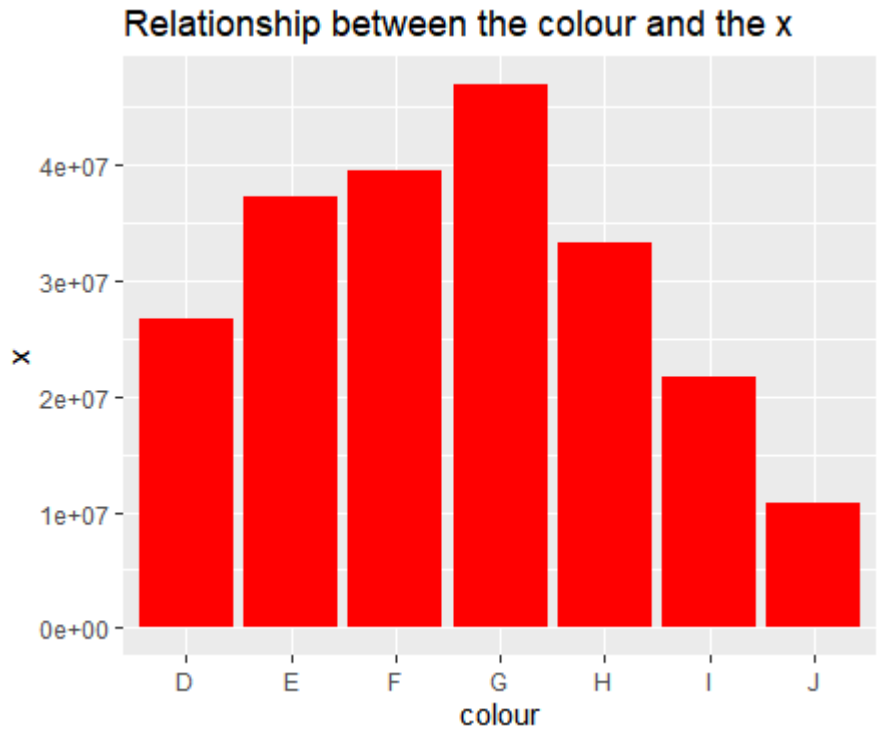
Then it we have to removing the outliers.



Number 6. create a new dataset with the outliers and missing data removed save the new dataset as a csv file removing the outliers from the whole dataset for the quantitative variables

Done within script

Number 7. Display the relationship between one qualitative variable and one finite variable in the dataset. (Define the variables you have chosen and link these to the research hypothesis).



Number 8. Show statistical tests to assess for normal distribution of all the variables (carat to PC) in the newly formulated datasheet.

I used Kolmogorov-Smirnov test because the data set has values that are greater than 5000. The values are also ties for this case and then aimed at removing the ties from the data by removing duplicates.

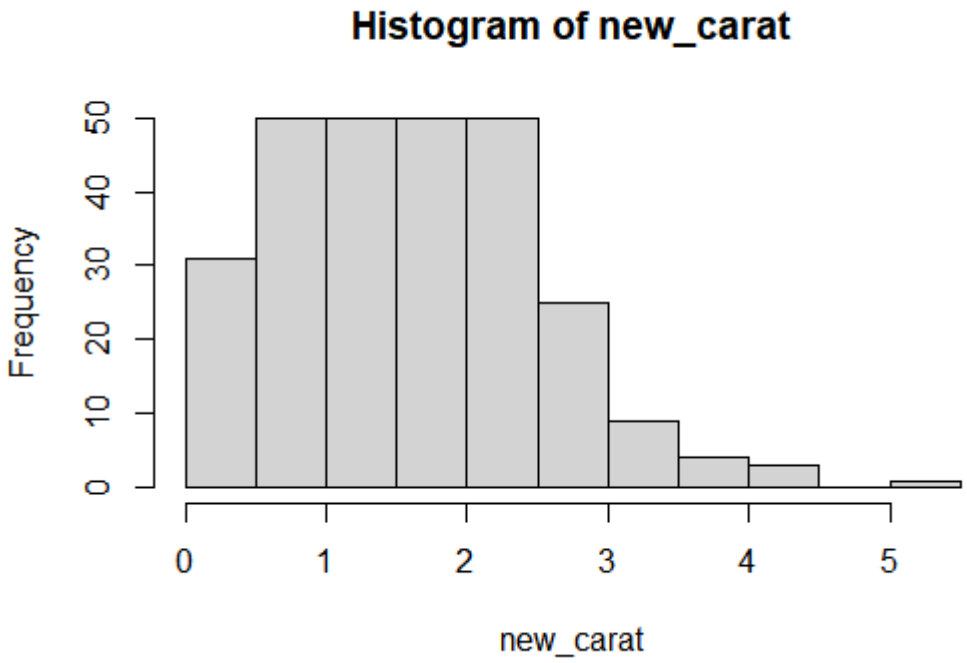
The variables where all called new_<variable name > after removing the duplicates.

The Kolmogorov-Smirnov test was used hence the results are;

For **carat** the p value is 0.4397 thus normally distributed.

The p-value is 0.4397 which is greater than 0.05 and therefore we fail to reject the null hypothesis

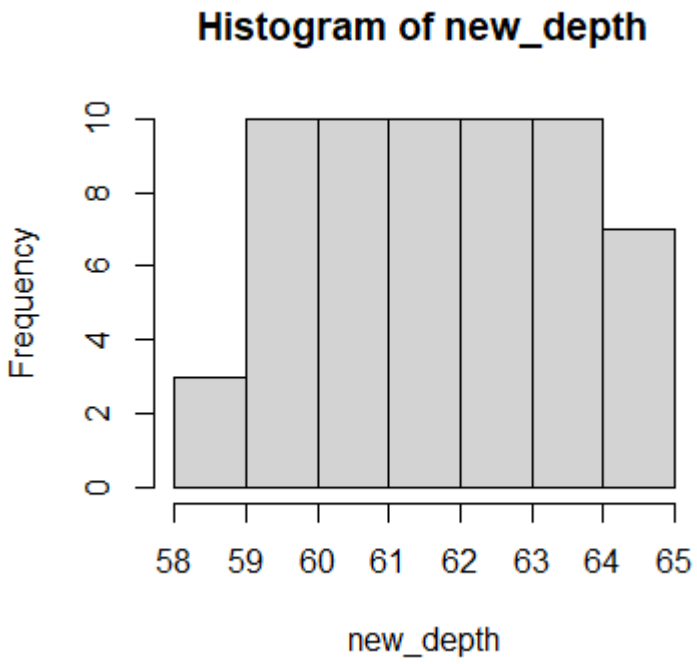
Histogram



For **depth** the p value is 0.9558 thus normally distributed.

The p-value is 0.9558 greater than 0.05and therefore we fail to reject the null hypothesis

Histogram



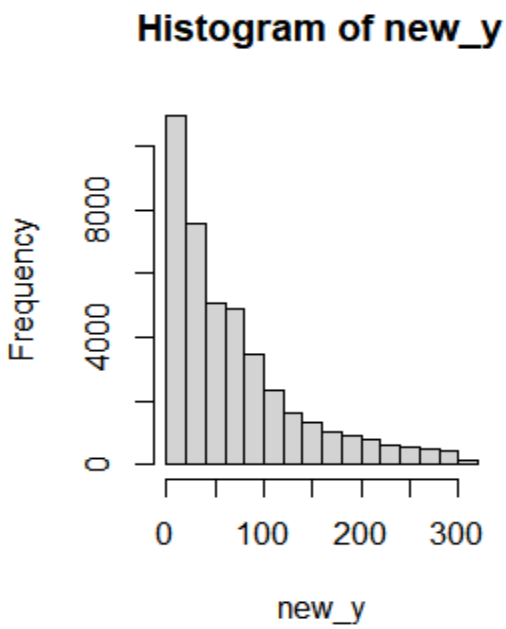
For **price**, the p value is 2.2e-16 thus not normally distributed.

The value of the p-value is 2.2e-16 which is less than 0.05 and therefore the null hypothesis is rejected

Histogram

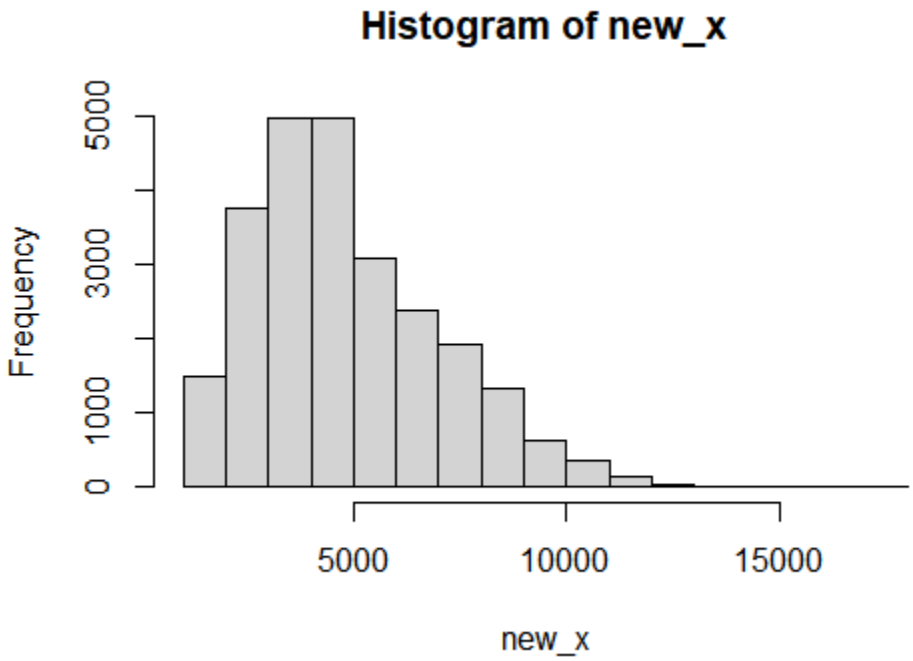
For y, the p value is 2.2e-16 which is less than 0.05 and therefore the null hypothesis is rejected. Thus not normally distributed.

Histogram



For x, the p value is 2.2e-16 which is less than 0.05 and therefore the null hypothesis is rejected. Thus not normally distributed.

Histogram

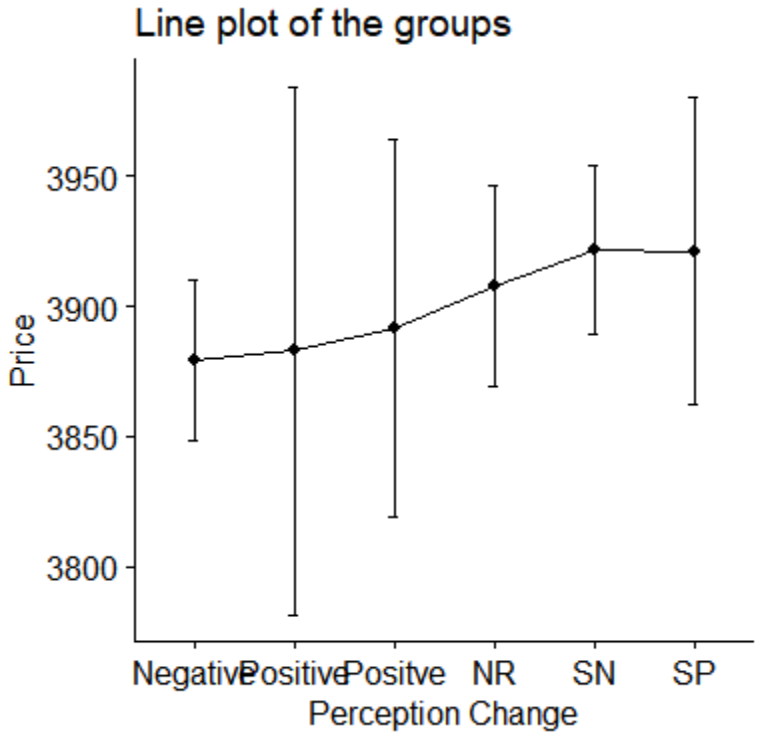


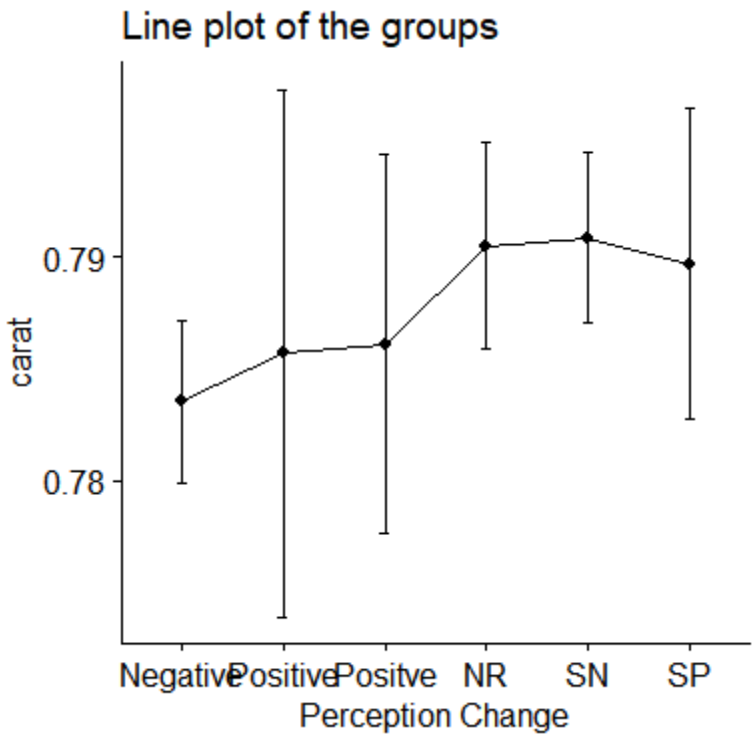
Number9. Compute the variance between three groups; diamond carat, perception change and price view the groups

From the groups given, The null hypothesis is that the groups have the same variance.

The alternative hypothesis is that the groups have different variances by at least one group having a variance not equal to the others groups

The test used is the one-way ANOVA test because the groups are more than two.





The values after running the anova test are as follows;

The P- value is 0.954 which is greater than 0.05 and therefore statistically not significant .