**Module 2 – Report**

**Name: Sunil Raj Thota**

**NUID: 001099670**

**Date: 01/28/2021**

**Title: Inferential Statistics**

**ALY 6015 – Intermediate Analytics**

**Prof. Roseanna Hopper**

**Introduction**

In this Assignment, I used R and R studio to perform Descriptive Statistics, Inferential Statistics, and Regression Analysis on various preloaded datasets provided by CRAN. We can perceive a list of all data sets in R using the function called data().

The dataset used for the analysis is the “mtcars”. It has the information of 11 variables and 32 observations. I have mostly focused on Descriptive Statistics, Exploratory Data Analysis, Linear Regression, Inferential Statistics, and Plots.

Descriptive Statistics provides us the basic information and synopsis of the given data set. Inferential Analysis provides us with insights about the performance of the data and helps us to forecast the movements for the future.

Linear Regression is mainly utilized for performing projecting analysis. The outcomes are for understanding the relation between 1 dependent and 1 or more independent variables. I have used the regression techniques and visualized the data using different plotting techniques that are available in R.

Also done some analysis on one-sample t-test, two-sample t-test, Paired t-test, the test of equal or given proportions, and F-tests.

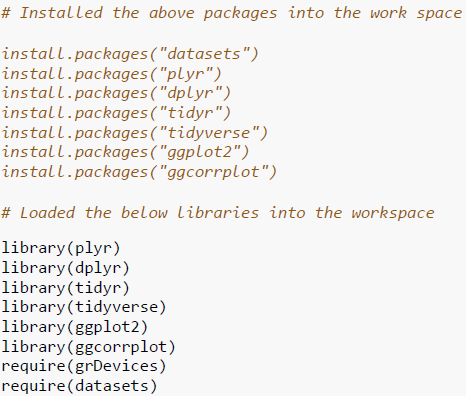
A one-sample t-test states whether an unidentified population means is dissimilar from a definite value. The two-sample t-test is also known as independent samples t-test to test whether the unknown population means of 2 groups are identical or not.

Paired t-test also called the dependent sample t-test to discover whether the mean change between 2 sets is 0. They are validated two times, resulting for pairs of observations.

The test of equal or given proportions will test whether or not a sample from a population represents the true proportion from the entire population. The last test, F-test signifies the linearity gives improved fit.

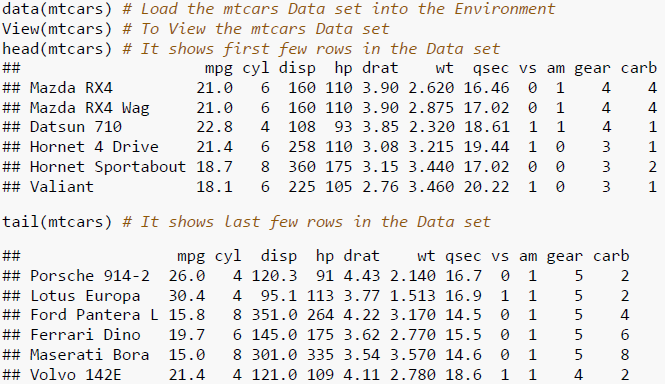
**Analysis**

The “mtcars” dataset is comprised of R’s dataset package and loaded for the Data Analysis. It contains 32 observations and 11 variables.

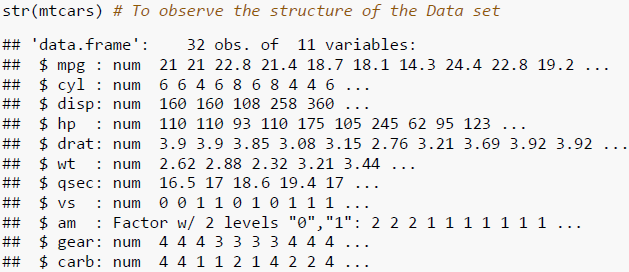


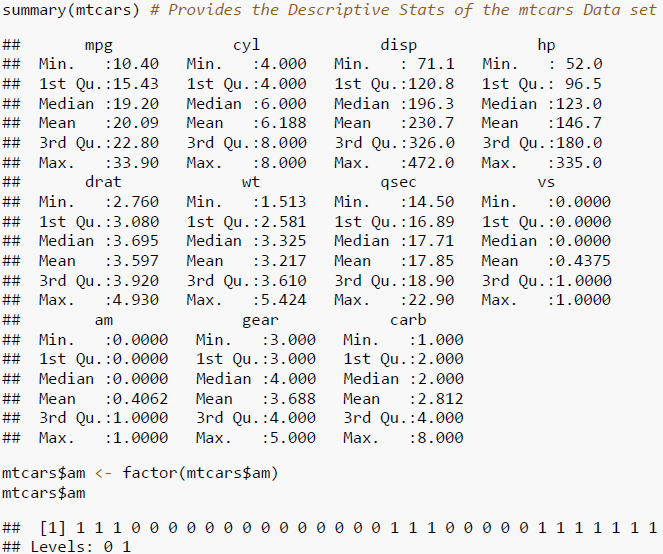
**Steps:**

1. “mtcars” dataset is loaded using the command data(mtcars). We can view the dataset by using the command View(mtcars). The first six rows are printed in the console by using the command head(mtcars) and the last 6 rows by using tail(mtcars)

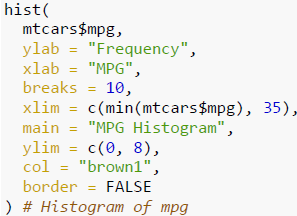


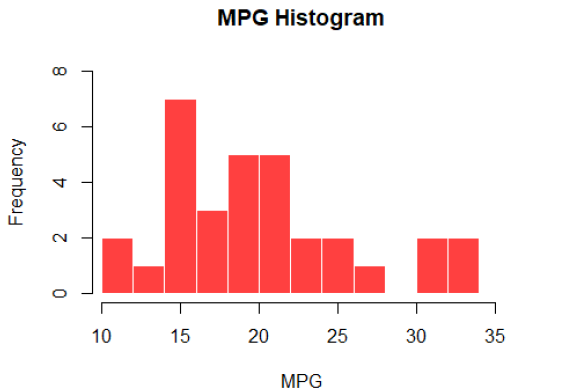
1. Summary of the Dataset has been used to breakdown the complex understanding of the dataset into simple and easy to analyze way. It can be attained using the command summary(mtcars). Also, observed the structure of the dataset



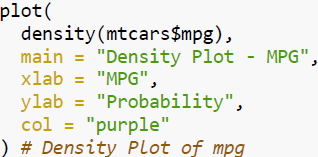


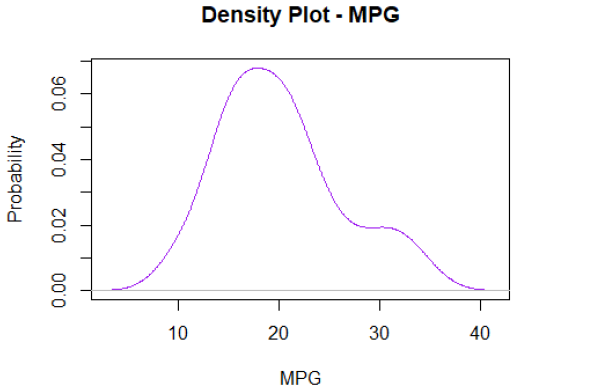
1. Used Histogram plot to find out the Frequency vs MPG relationship and can see that the distribution is normal and I have chosen to include main, xlab, ylab, color, border, ylim, breaks, and xlim as parameters that outline the title, x label, y label, color, border, y limts, breaks, and x limits accordingly



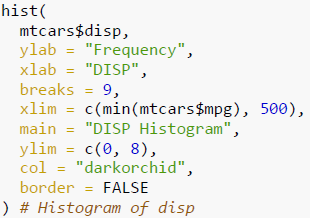


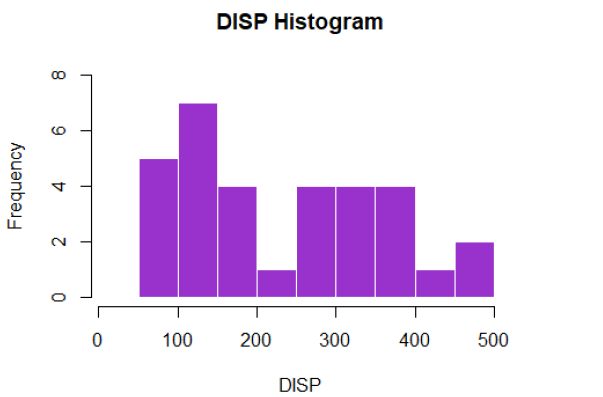
1. Used Density plot to find out the Probability vs MPG relationship and observed that the MPG levels from 12 to 24 has more density and I have main, xlab, ylab, and color as parameters that outline the title, x label, y label, and color accordingly



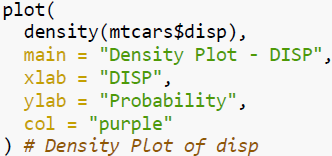


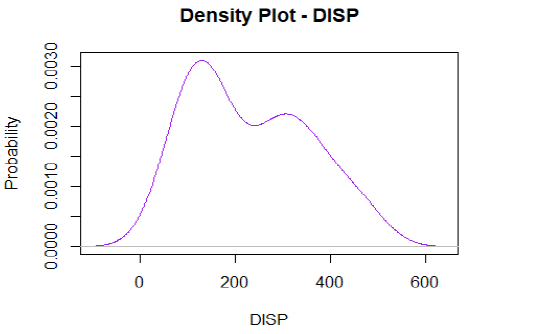
1. Used Histogram plot to find out the Frequency vs DISP relationship and I have included main, xlab, ylab, color, breaks, xlim, ylim, and border as parameters that outline the title, x label, y label, color, breaks (bins), x limts, y limits, and border accordingly



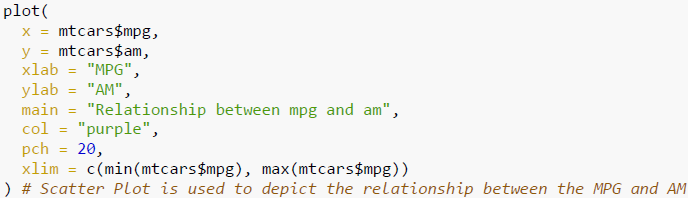


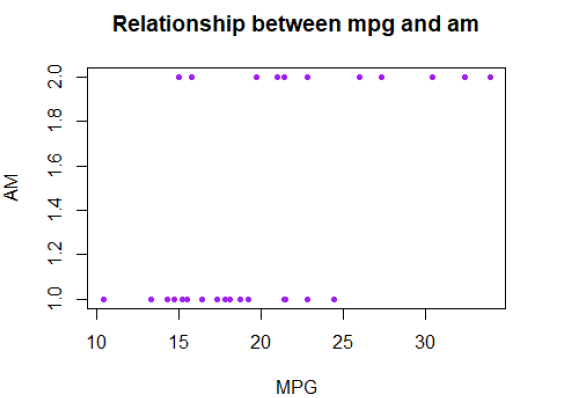
1. Used Density plot to find out the Probability vs DISP relationship and I have chosen main, xlab, ylab, and color as parameters that outline the title, x label, y label, and color accordingly



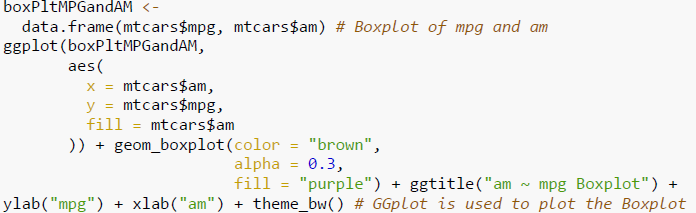


1. By using the Scatter Plot, let's depict the relationship between the mpg and am variables. I have chosen scatter plot and included main, xlab, ylab, color, pch, and xlim as parameters that outline the title, x label, y label, color, point shape, and x limits accordingly.



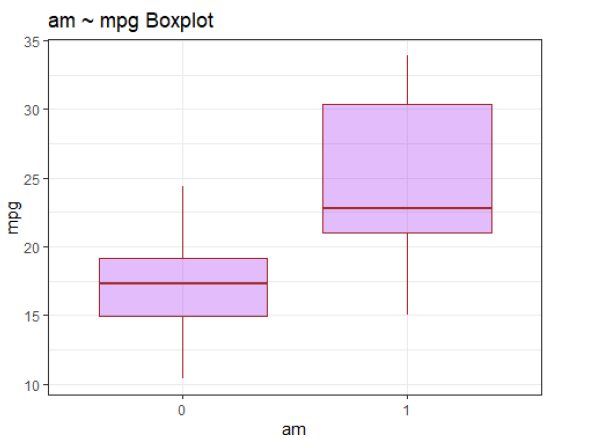


1. Now let’s implement Box Plot for mpg and am variables of the “mtcars” dataset



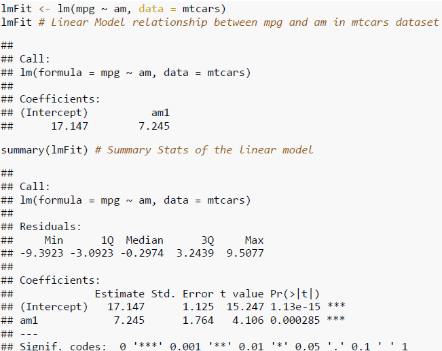
**Box Plot – MPG vs AM:**

Box Plot can be plotted by using the boxplot() function. In this, the plot function calls the mtcars$mpg and mtcars$am data. I have chosen GGplot to plot the box plot graph and included ggtitle, xlab, ylab, color, fill, and theme as parameters that outline the title, x label, y label, color, colors to be filled in the boxplot, and theme accordingly.

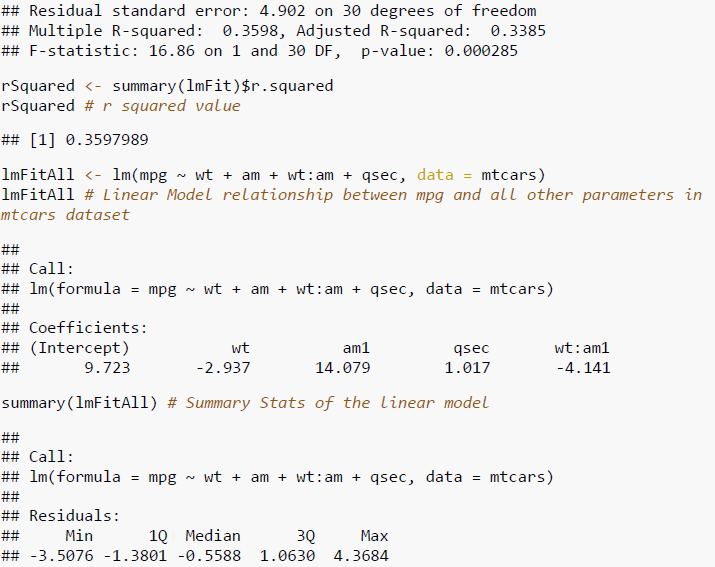


It demonstrates the Min, 1st, 3rd Quartile, Median, and Max values.

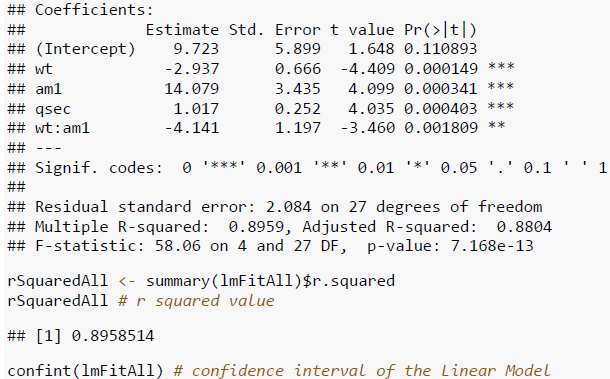
1. Now I have performed some linear regression model analysis

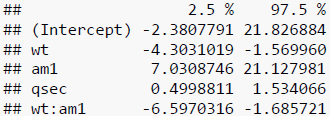


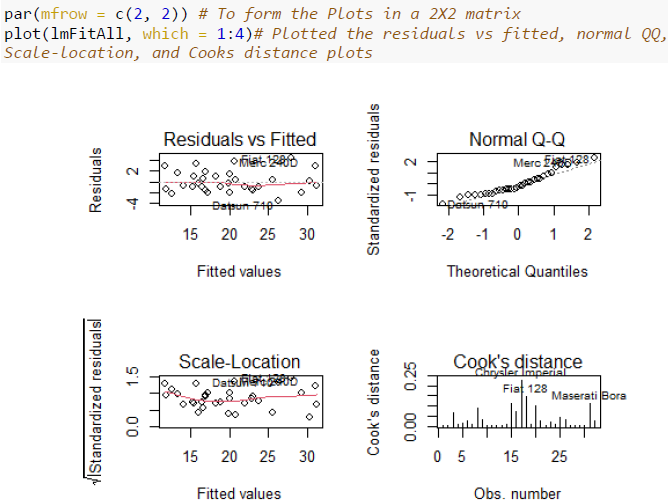
And found out these attributes summary and r squared values. Model between mpg and am



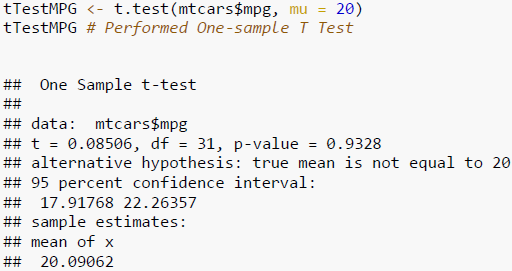
Model between mpg and all other parameters of the “mtcars” dataset and found out the r squared value, confidence interval, and summary values. Plotted the Residuals vs fitted, Normal QQ, Scale-Location, and Cooks distance plots.



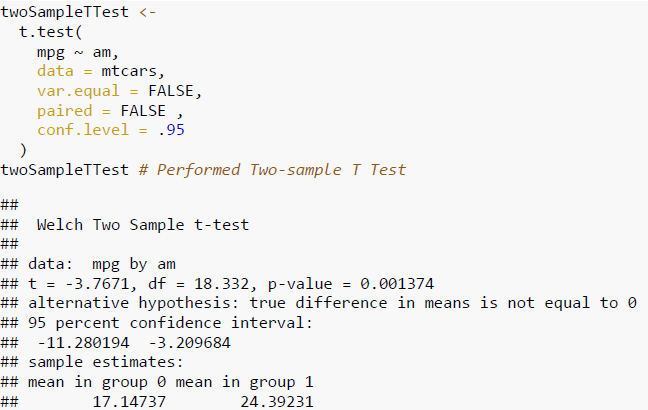




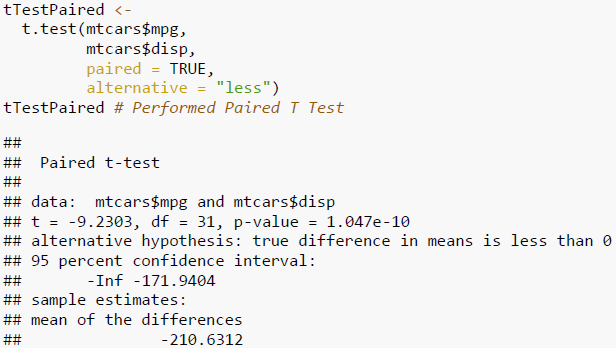
1. Hence performed a one-sample t-test and here is enough indication to reject the Ho. The mean is slightly > than 20. The MPG level is not significantly different than 20



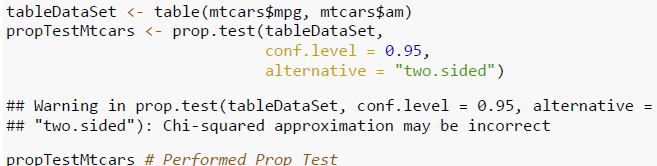
1. Hence performed a 2-sample t-test and here are enough indications to reject the Ho. The p-value is < 0.05. There is no difference in mpg between transmission types. The confidence interval (CI) also defines how much lesser the MPG is in physical cars than in automatic cars. The true difference as shown in the below image is between 3.2 and 11.3.

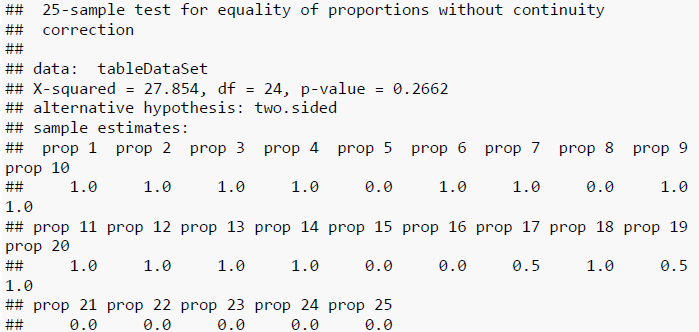


1. Hence performed paired t-test and here is enough indication to reject the Ho. And, mpg is better than disp

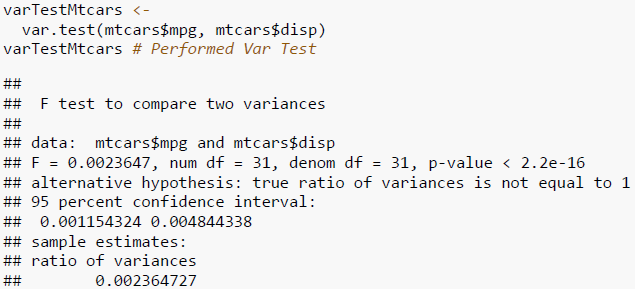


1. Hence performed test of equal or given proportions and found the p-value is (0.2662 ) > the alpha value of 0.05.





1. Hence performed an f-test and here are enough indication to reject the hypothesis where variances of mpg and disp are equal. As, the p-value = 2.2e-16, which is very much less than 0.05



**Conclusion**

To perform multiple analytical operations, we use R on widespread range of data like mtcars dataset. R is a very powerful tool to perform analysis which is mainly built by researchers, statisticians, and developers. In the beginning we did Descriptive Statistics Analysis, Exploratory Data Analysis, and continued with the Linear Regression and Inferential Analysis to get understandings from the provided data. Plotted Histograms, Density, and Box which helps us to understand the data in a clear way. Also, performed the 1-sample test, 2-sample t-test, paired t-test, a test of equal or given proportions, and f-test for the mtcars datasets to solve given business problems. Ho and Ha are essential in performing the Hypothesis testing for the datasets to make and take data-driven business decisions.

**References**

[1] Phil Spector, using t-tests in R Originally for Statistics 133 was retrieved from <https://statistics.berkeley.edu/computing/r-t-tests>

[2] J H Maindonald, Using R for Data Analysis and Graphics was retrieved from <https://cran.r-project.org/doc/contrib/usingR.pdf>

[3] Test of Equal Or Given Proportions was retrieved from <https://www.rdocumentation.org/packages/stats/versions/3.4.1/topics/prop.test>

[4] F-Test: Compare Two Variances in R was retrieved from <http://www.sthda.com/english/wiki/f-test-compare-two-variances-in-r#compute-f-test-in-r>