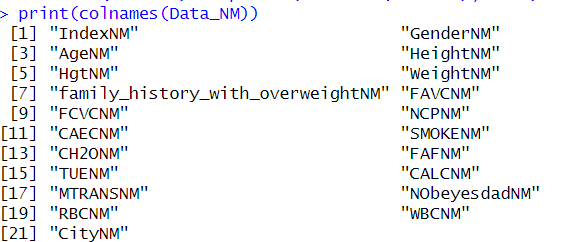
**1. Data Transformation and Preparation**

**Question 1.1: Initial Transformation**

**1)Rename all variables with your initials appended (just as was done in Assignment 1)**



Renamed all the variables and column names, appended them with my initials

**2)Transform character variables to factor variables.**

****

Transformed all the character variables into Factor variables so that we get proper output. Made use of as.factor.

**Question 1.2: Reduce Dimensionality**

**1)** **Drop any variables that do not contribute any useful analytical information at all.**

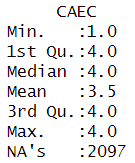
Here Index can be removed as it is only a column to keep track and is used for indexing and it not useful to our analysis.

Moreover, Hgt which is height in Feet can be removed as well since we already have a column which gives us the height in metres so Hgt is just redundant.

Furthermore, many more columns can be removed from this dataset as the questions asked by the Professor do not relate to some of those columns at all. But if we were to take this in a professional setting, we won’t be removing these columns if we were to be asking question in particular which were related to the requirement of those columns. In addition, if we do more analysis, these columns are going to come in use. Therefore, I am only removing columns that don’t contribute at all and columns that are only providing redundant data and will be keeping all the rest of the columns.

**2)Apply the Missing Value Filter to remove appropriate columns of data.**

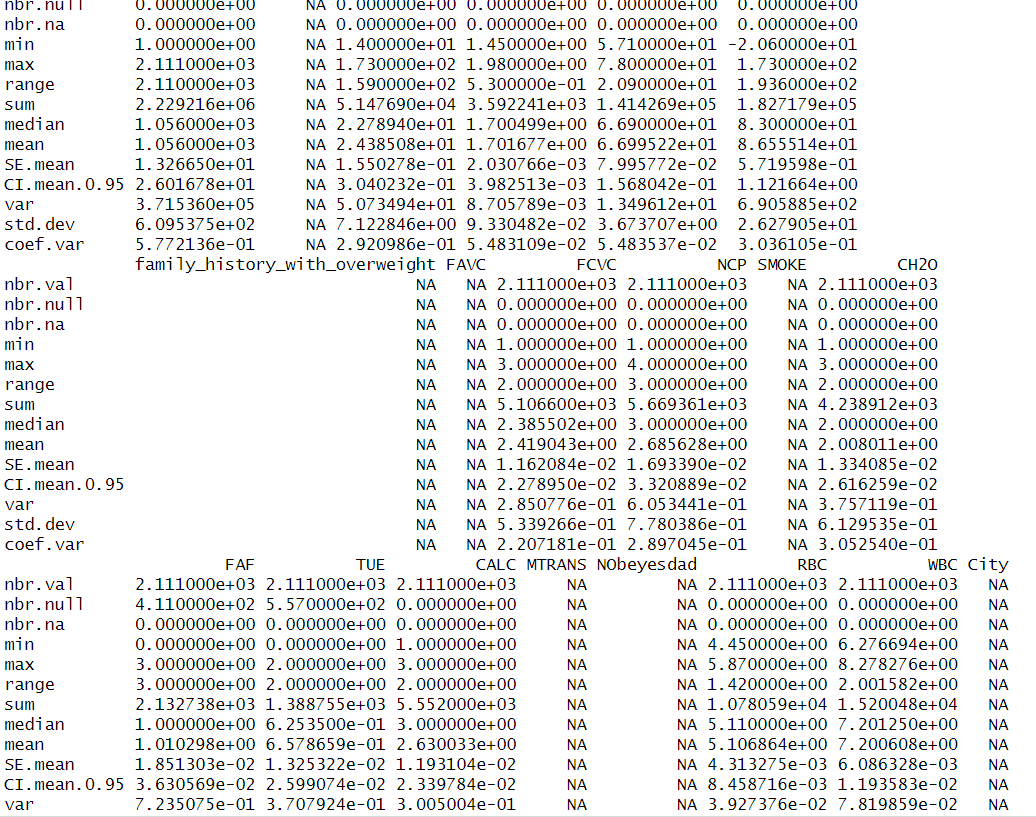




Here as we can see CAEC and FAVCNM have missing values, and CAEC in particular has 2097 missing values. As a result, we can remove those columns according to missing values filter.

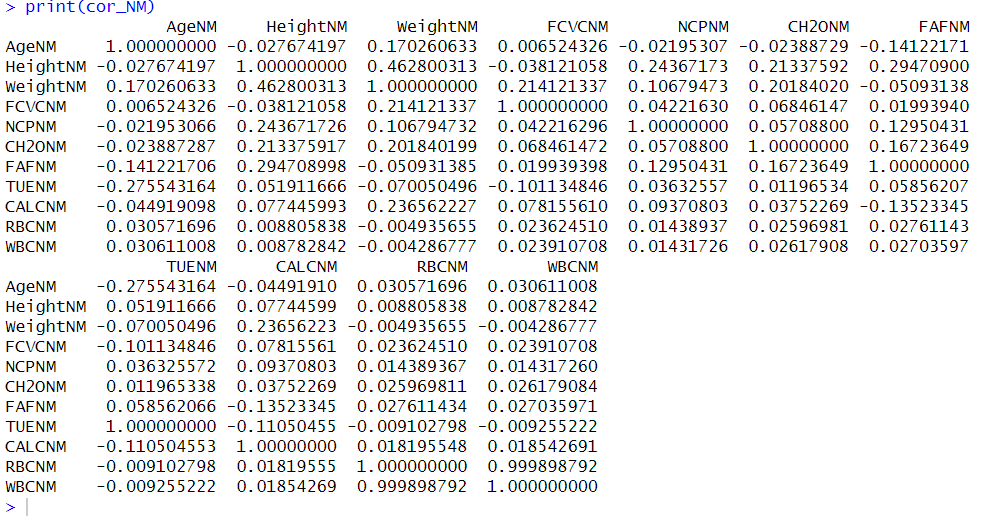
The reason for not removing FAVC is that there are not a lot of missing values in that column and we can fill those columns with either the mode or ffill.

**3) Apply the Low Variance Filter to remove the appropriate columns of data.**

****

Here we can infer that Height has a low variance but since it is an important column, I won’t be removing it. Therefore, the next lowest variance is of RBC but we are making use of RBC in later stages so we will be keeping it. As a result, the next lowest variance belongs to WBC and the removal of WBC is supported by High Correlation Filter as well.

**4) Apply the High Correlation Filter to remove the appropriate columns**

****

Here based on the High Correlation Filter we can see how WBC and RBC are very highly correlated. As a result, we will be removing WBC since when two columns that are highly correlated, they don’t contribute much to the analysis and both of these columns can be thought of as one entity in itself. We could also remove RBC but we will be using RBC later on.

**5) Based on our discussions in class, what are some specific benefits of reducing the dimensionality of this particular dataset? Be specific. For example, if it increases computational efficiency, specify how much of an improvement.**

Well, when it comes to reducing dimensionality, a major benefit is definitely computational efficiency, but this comes more into play when your datasets are larger and you are working with large amounts of columns and rows of data. When it comes to the goal of Dimension Reduction we are aiming for maximum reduction and minimal information loss.

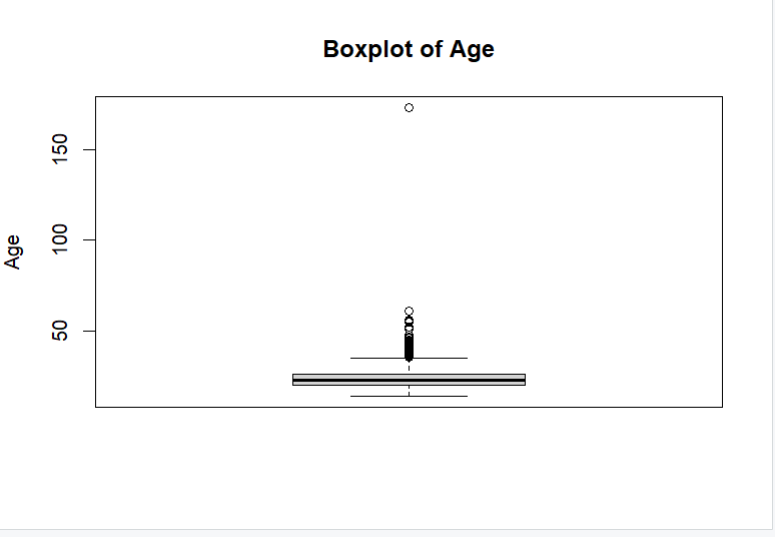
Another benefit of dimensionality reduction is that we are able to focus more on the important parts of the data. As there are times when the dataset is too large and cluttered, it becomes a bit difficult to deal with data and to focus on the important part which you want to analyse. With the help of Dimensionality Reduction, we can reduce those less important data columns and focus more on the important ones.

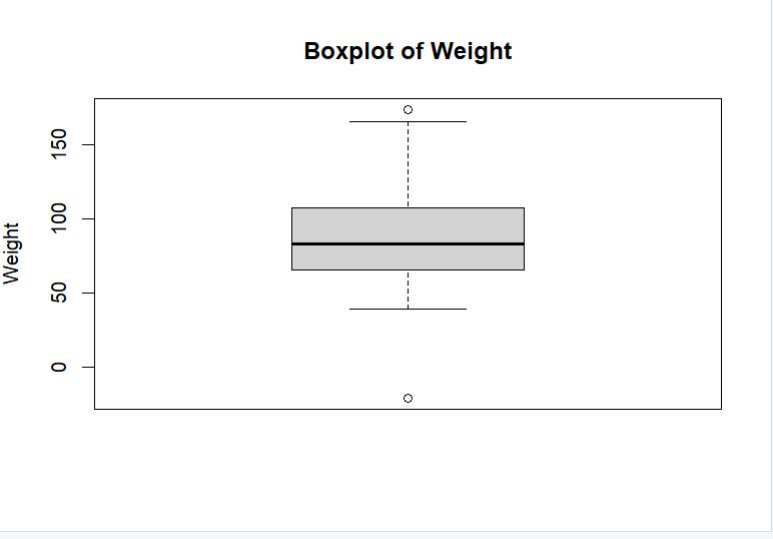
When it comes to this dataset, the obesity levels on citizens of waterloo based on their eating habits and their physical conditions. With the reduction of columns such as Index and Hgt which were contributing nothing to the analysis and were just redundant data respectively we were making our lives easier with having to work with less data and focusing more on the other important data points.

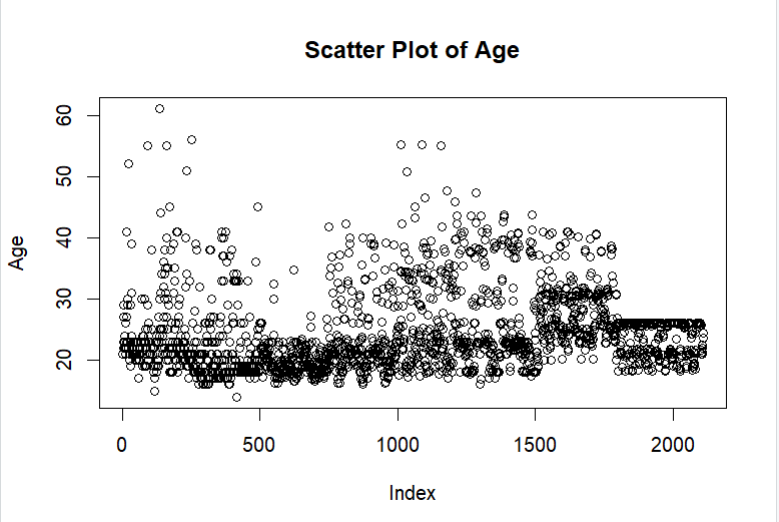
Furthermore, we also removed columns which had large amounts of non-available data since it wasn’t contributing to the analysis and would just cause problems if left there. In addition, we removed columns which are highly correlated with which other since they affect accuracy and analysis of the data. Moreover, with the low variance filter we removed the columns that have low variance and don’t change too much within the dataset itself. These columns can be takes as one value and removed leading to more storage space and faster and easier execution.

**Question 1.3: Outliers**

**Use an appropriate technique (or techniques) demonstrated in class to identify outliers.**

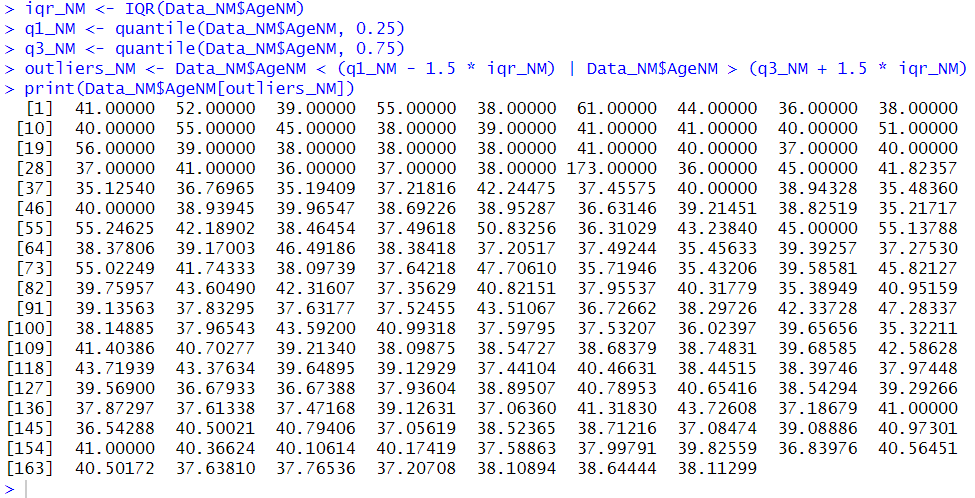
****

****

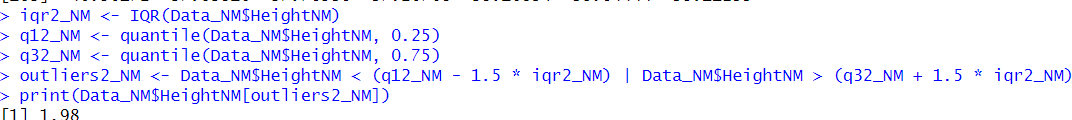
****

With the help of Box plots, Scatter Plots and IQRs we were able to figure out there were outliers in Weight, Age, Height, NCP and many more columns

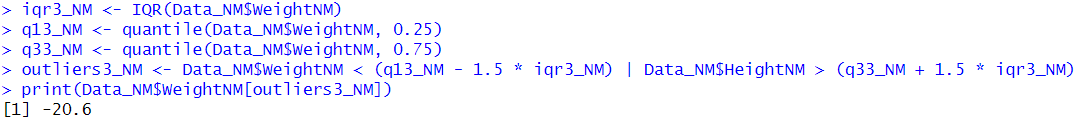
**2)** **Comment on any outliers you see and deal with them appropriately. Make sure you explain why you dealt with them the way you decided to.**



With the help of IQR calculations we can see the outliers in the Age column. We will be removing the outlier age “173” since that is very impractical and doesn’t make sense.



Similarly, we used IQR calculations on the Height column, but there are some people who do reach 200 cm of height so we will be keeping this outlier as it does make sense.



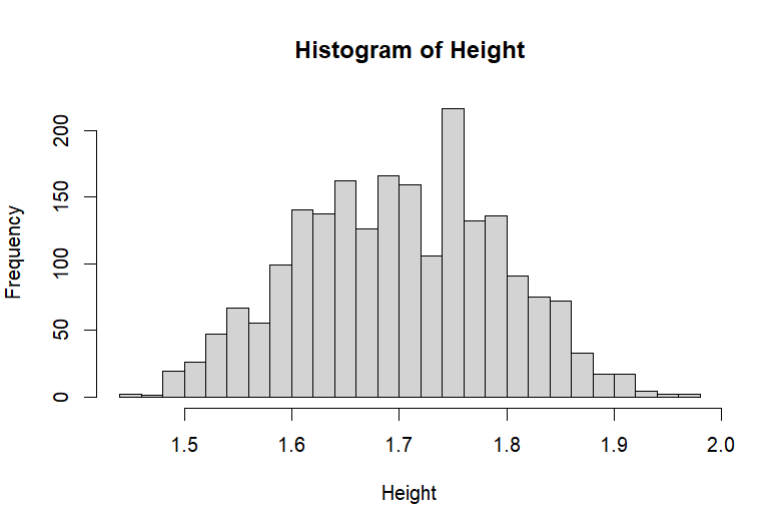
Similarly, for weight we will be removing this outlier as it doesn’t contribute much to the analysis and Weight cannot be in negative.

Moreover, many of the medical data and categorical data also had some outliers but since they were not diabolically weird or weren’t like very out of the line. We have decided to keep those values.

**2. Organizing Data**

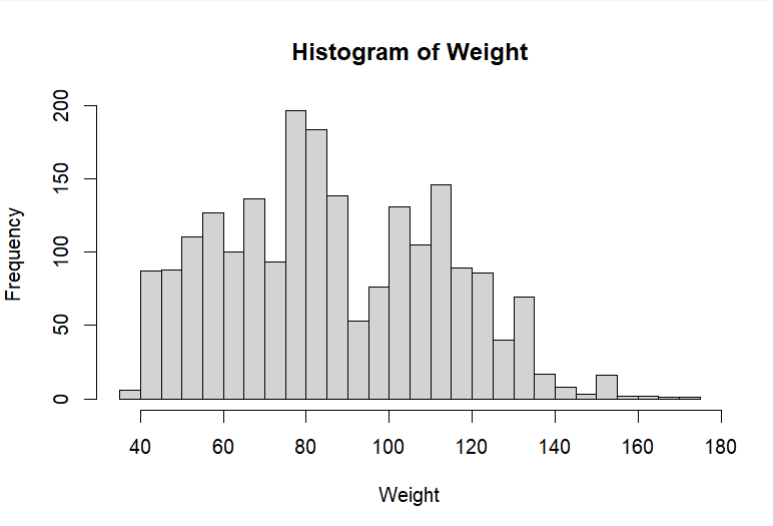
**Question 2.1: Scatter Plot**

**a) Create a Histogram for Height.**

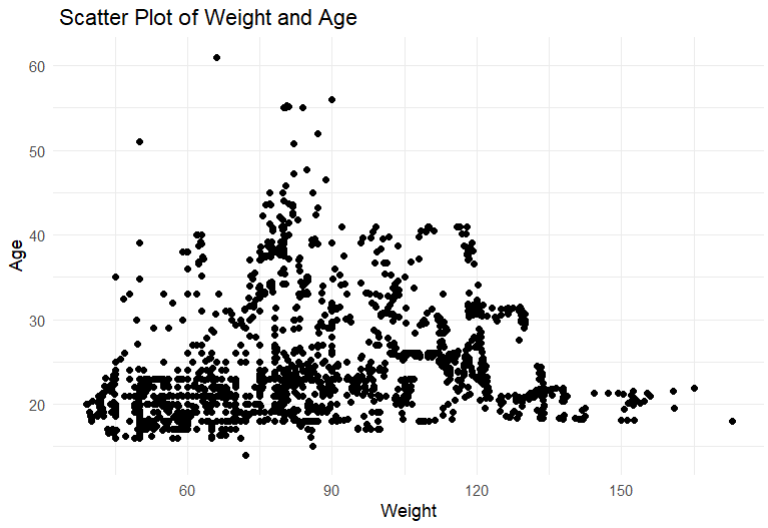


This showcases the Histogram for Height. We can infer that this is a normally distributed chart.

**b) Create a Histogram for Weight**

****This showcases the Histogram for Weight. We can see that most people range from around 40kg to 120kg for their weight with the average being around 80kg.

**3) Create a Scatter plot showing the relationship between Age and Weight**

****

This is a Scatter plot showing the relationship between Age and Weight

Based on this plot we can see that most of the people are in their 20s and 30s and have given their data. Moreover, people in their early 20s and 30s are the only age ranges who are also more on the overweight side of the spectrum. As we move higher and higher, we can see that people in their 40s, 50s and 60s are not overweight and have stable body weight. A conclusion comes to mind from this is healthy diet of older people and high calorific diet of younger adults could be a major factor to this data spread we cant be sure about this but this seems to come to mind.

**c) Calculate a correlation coefficient between these two variables. Why did you choose the correlation coefficient you did? What conclusion you draw from it?**



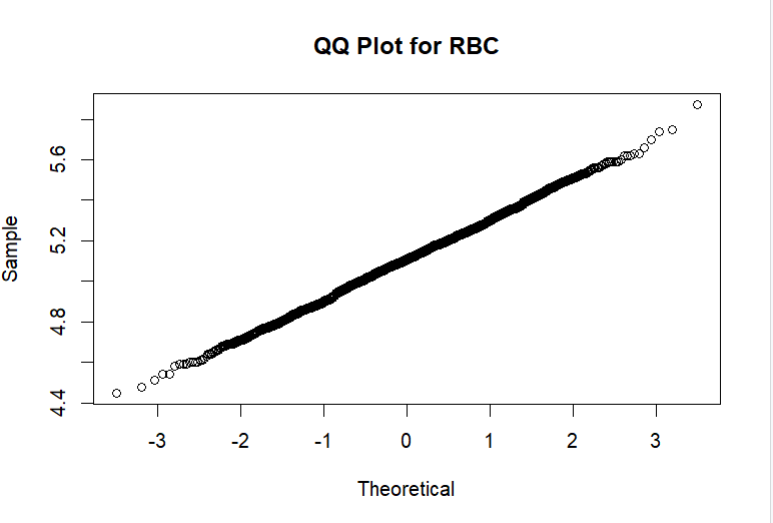
As we can see it is a positive correlation but the two variables aren’t highly correlated with each other.

The chosen correlation coefficient is Pearson’s correlation coefficient which is the basic correlation coefficient that comes with r.

**3. Inference**

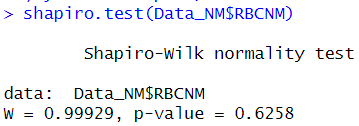
**Question 3.1: Normality**

**a) Create a QQ Normal plot of for Red Blood Cell Count.**



QQ Plot for Red Blood Cell Count, we can see that it is normally distributed.

**b) Conduct a statistical test for normality on Red Blood Cell Count.**



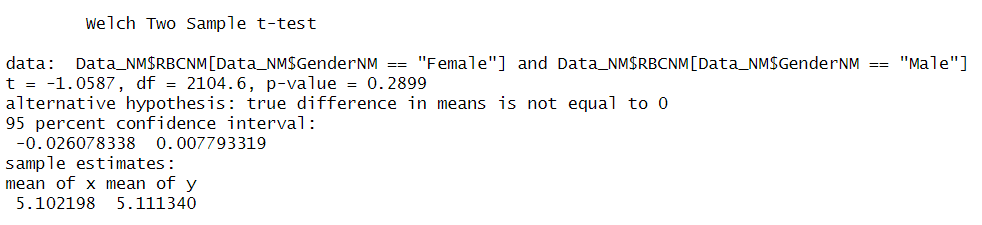
From the Shapiro Wilks test we can see that the p-value is much greater than 0.05. Hence, null hypothesis cannot be rejected and that the distribution is likely normal.

**c) Is Red Blood Cell Count normally distributed? What led you to this conclusion?**

Well based on the QQ plot and the Shapiro Wilks test we can confidently say that the data is normally distributed and follows the curve of a normal distribution graph.

**Question 3.2: Statistically Significant Differences**

**a) Compare Red Blood Cell count between Genders in your dataset using a suitable hypothesis test.**



As we inferred from the last test RBC is normally distributed and we will be using the T-Test to compare the RBC counts between genders as T-Test is used to compare the means between two variables and can be used if the distribution is normal.

For the Hypothesis.

H0: There is no significant difference between the means of the two variables.

H1: There is a significant difference between the means of the two variables.

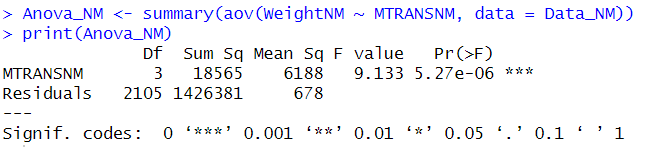
And from the T-Test we can see that the p value is 0.2899 which is very high (>0.05). Hence, we can say that the null hypothesis is correct.

**b) Do you have strong evidence that Red Blood Cell count is different between genders?**

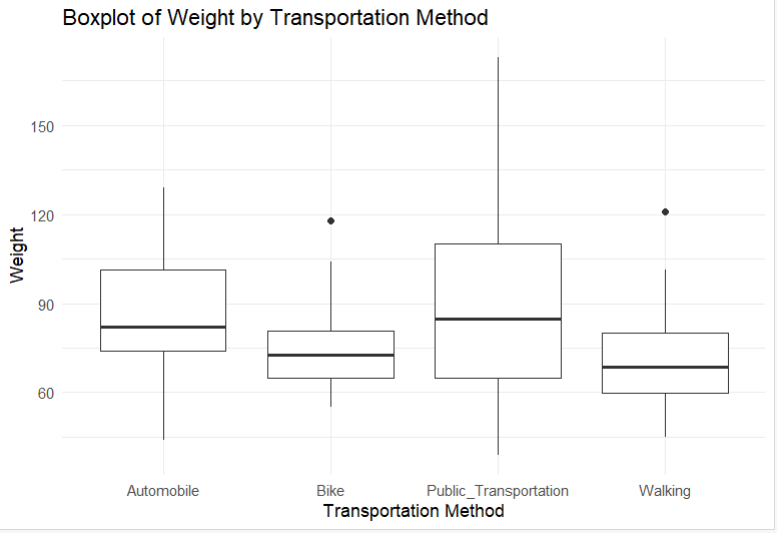
According to the T-Test and the p-value being greater than 0.05, we don’t have significant evidence to reject our Null Hypothesis. As a result, we can say that the means for the two groups are very similar. Furthermore, we can see that the p-value is high but that still doesn’t give us proper mean to negate the null hypothesis but we don’t have significant proof to actually reject is so we have to accept it as of now and this also tells us that there is difference in mean but it could by minor difference or a difference due to luck.

**Question 3.3: Multiple Statistical Differences**

**a) Determine if Weight varies by method of transportation using ANOVA (statistical) and a sequence of boxplots (graphical)**



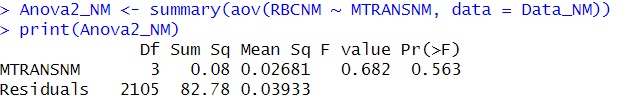
(Statistical)



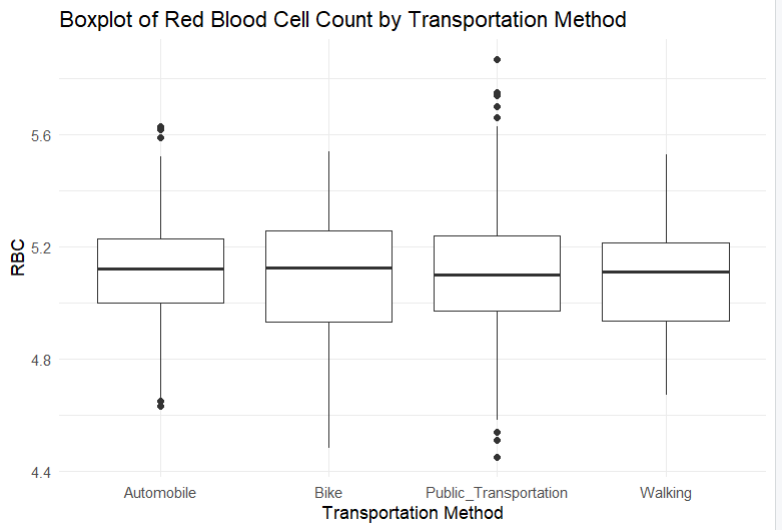
(Graphical)

From the Anova test and the Graphical depiction of different transportation methods based on weight we can see that the p value is very small and less than 0.05 hence we can reject the null hypothesis and accept that there is a significant difference between the weights using different modes of transportation. Moreover, with the stars presented in the Anova Test we can also easily identify that the significant difference is more. It becomes overall clearer when we check the boxplots as well.

**b)** **Determine if red blood count varies by method of transportation using ANOVA and a sequence of boxplots.**

****

(Statistical)



(Graphical)

Here from the Anova test it is clear that the p value is 0.5 which is a very high p-value (> 0.05). So, we can say from the test and the boxplots that there is no significant difference between the RBC blood counts and the different methods of transportation.