

Final project cpis-490



Students:

Alaa Al-jedani - 1708633

Ena’am Bernawi - 1808433

**Contents:**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[Task 1: Discovery 2](#_Toc69310837)

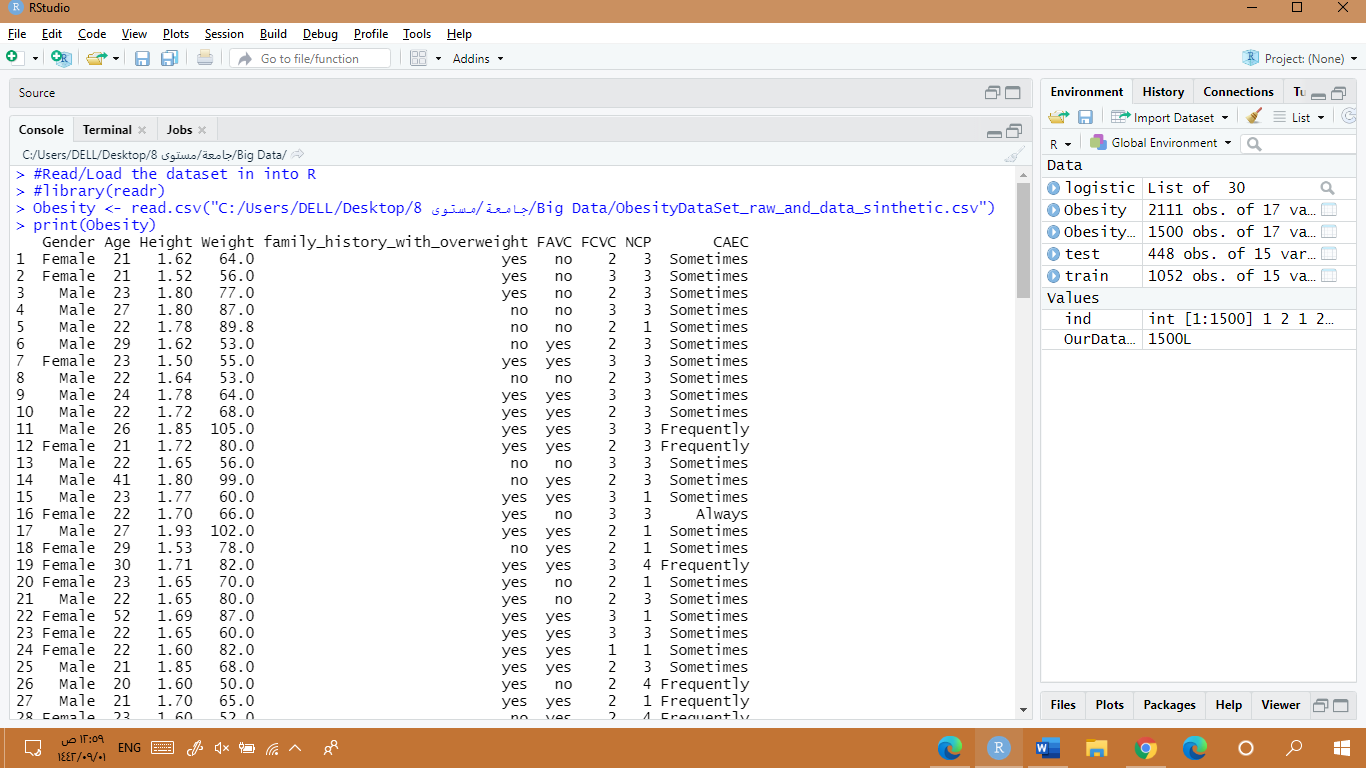
[Task 2: Hypothesis Testing 8](#_Toc69310838)

[Task 3: Model Building& Evaluation 9](#_Toc69310839)

[References 16](#_Toc69310840)

Task 1: Discovery

1. Read/Load the dataset in [2] into R.



صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

1. In preparation of your data:

* صورة تحتوي على نص

  تم إنشاء الوصف تلقائياًSelect a random 1500 instances.
* Use all the attributes except:

1. Group 1, remove MTRANS and CALC.
2. Group 2, remove CAEC and CALC
3. Group 3, remove MTRANS and CAEC
4. Group 4, remove CAEC, CALC, and FAF

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

1. Check the structure of the data identifies the following: # of instances, attributes, data types, missing values, etc. Report the values.

There is 1500 observation of 15 attributes, we use str () to discover the structure of the data set.

The code line "sum(is.na.data.frame(ObesityData)" was used to the number of missing values and we got 0

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

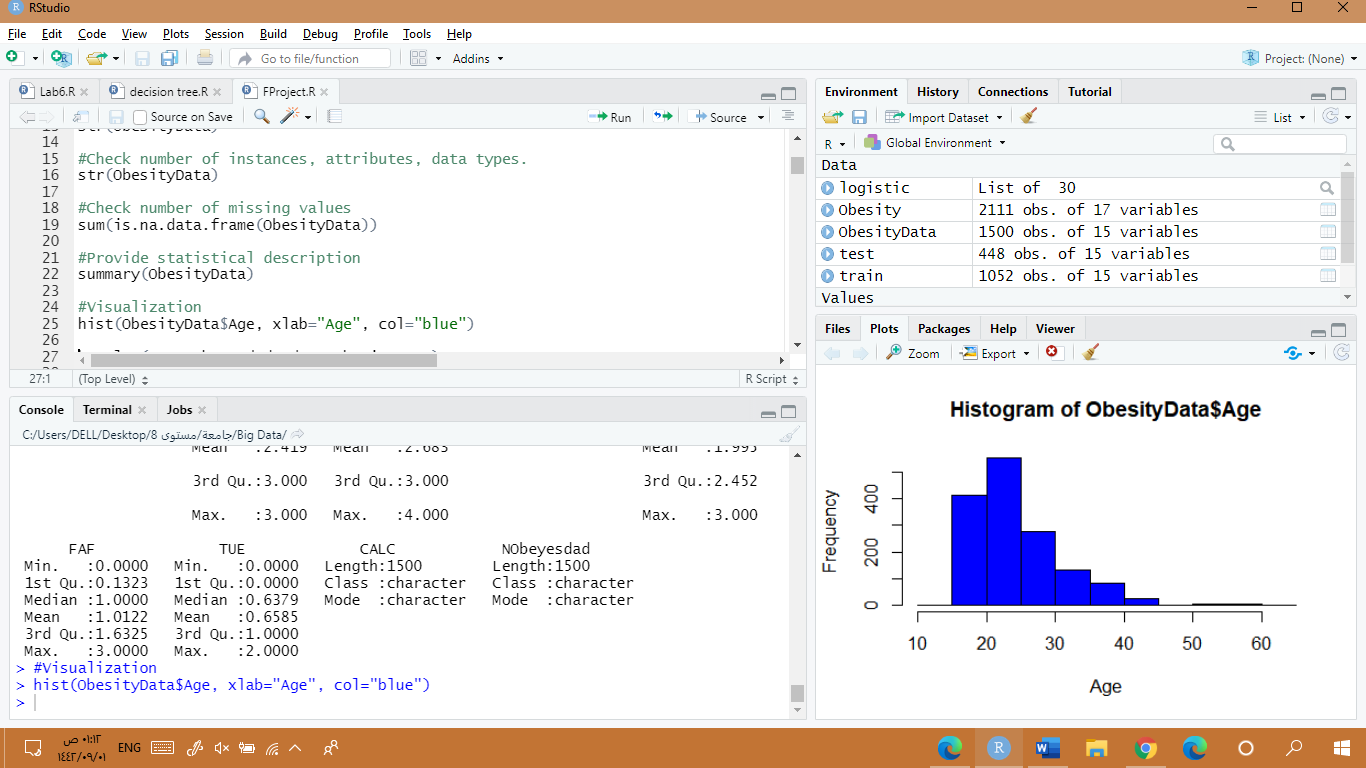
1. Provide a statistical description.

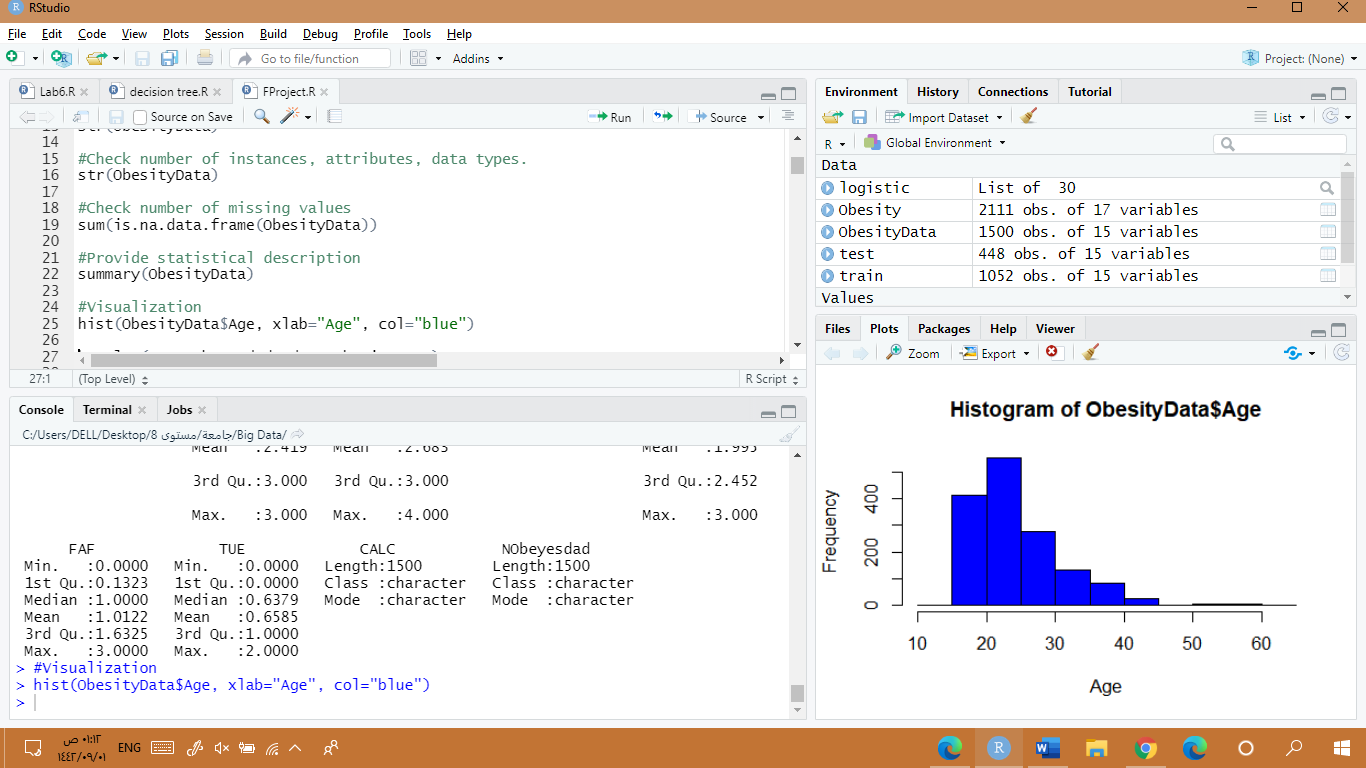
We use the summary () function to display a statistical description of the data which occur in Min, 1st Quartile, Median, Mean, 3rd Quartile, Max.

صورة تحتوي على نص

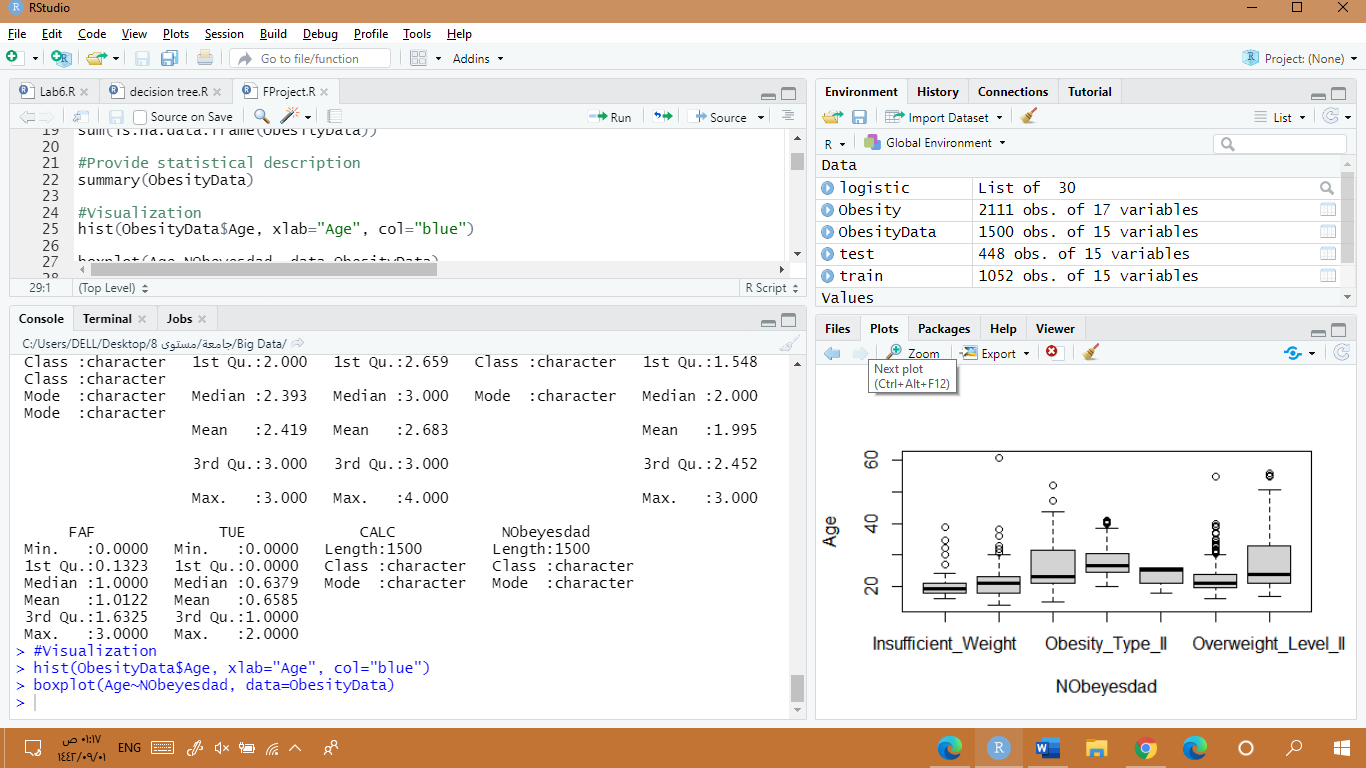
تم إنشاء الوصف تلقائياً

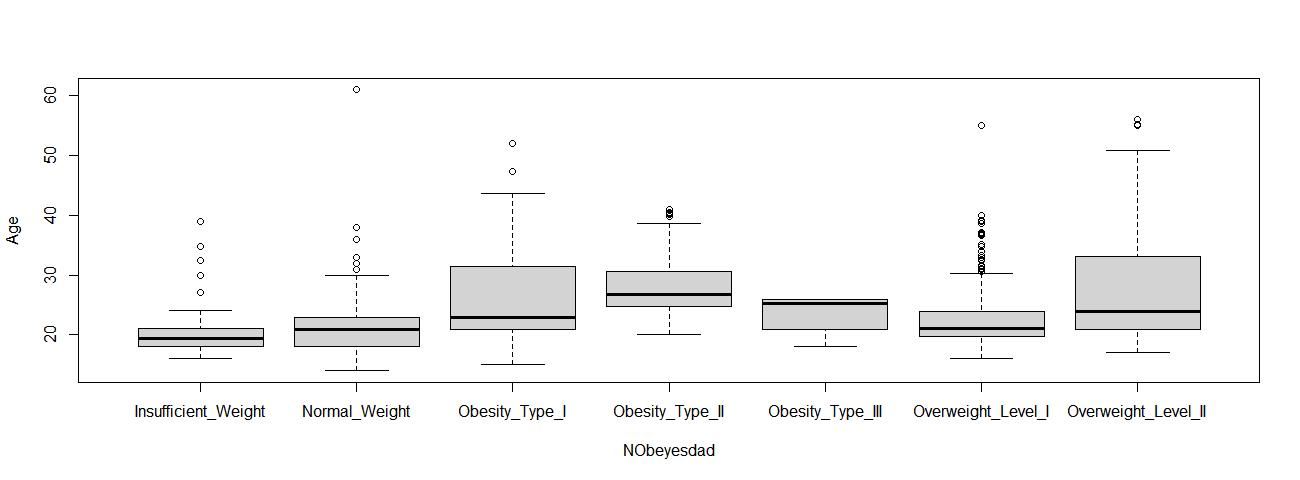
1. Visualize the data in three different ways. Describe the plots in detail.
2. Histogram: This graph shows the frequency of the age attribute in the data set. We can see that most of the people in this data are between 20-25 and the fewest are between 50-60. Also, the range of the is 15-60 years old and we can say it has right-skewed distribution.



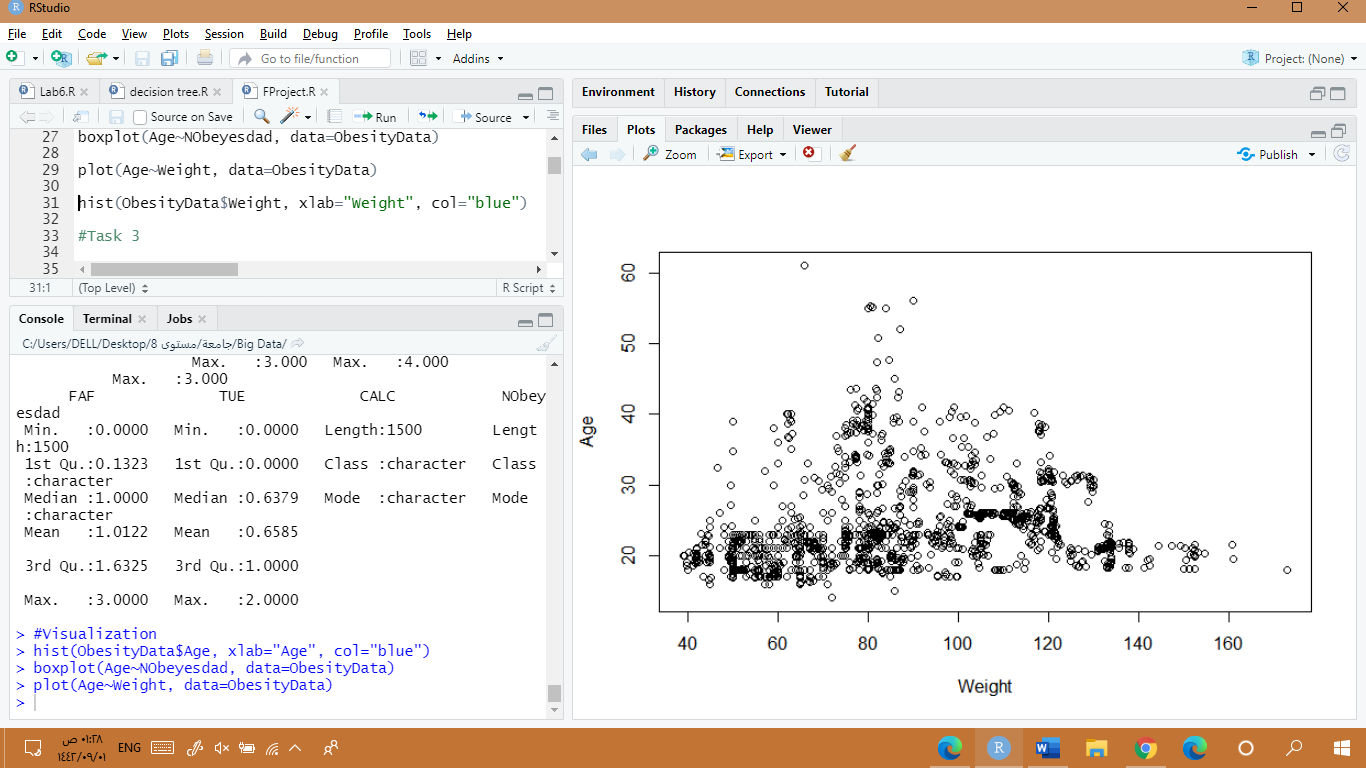


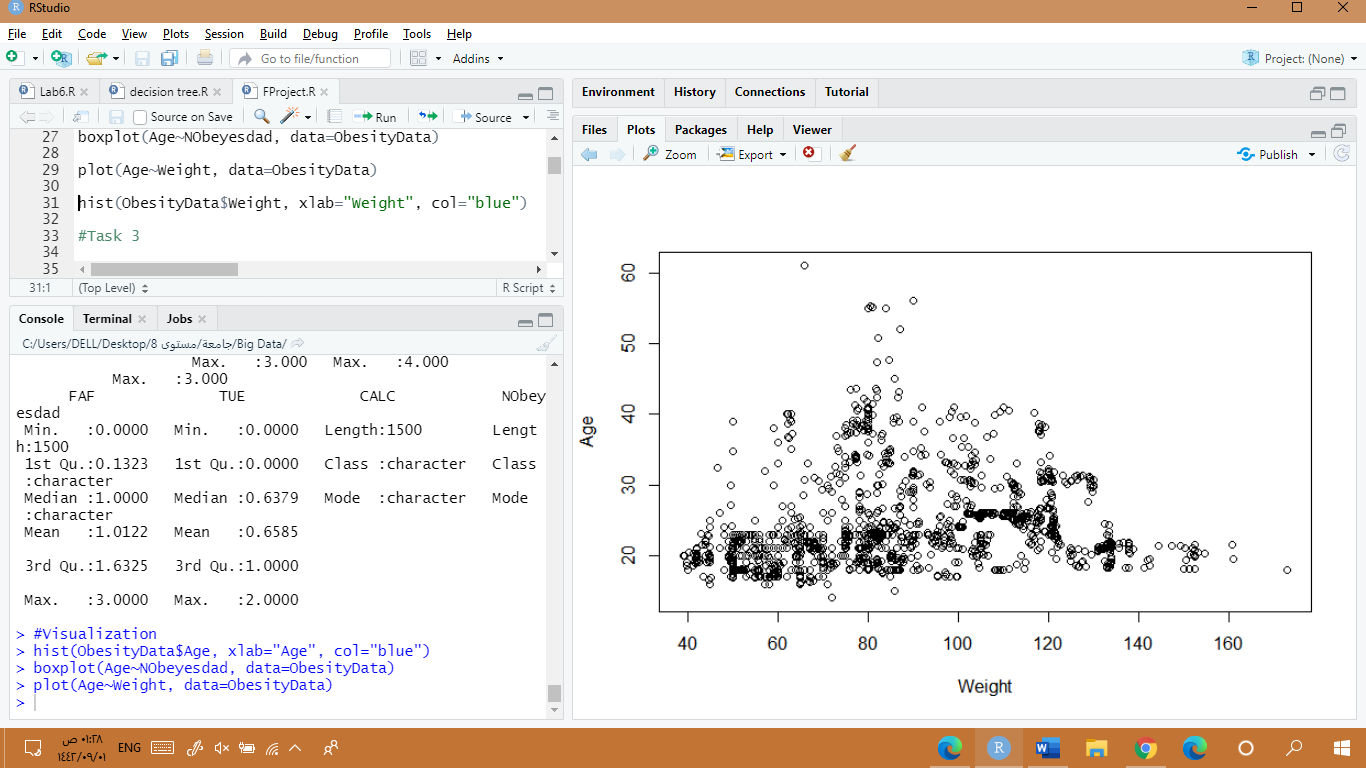
1. Boxplot: This boxplot represents the 7 types of obese for different ages in the data. The lowest median is in (Insufficient\_Weigh) which is 20 years old where the highest median is in (Obesity\_Type\_II). Also, we can see that all types have outliers in different ages except for (Obesity\_Type\_III)





1. Plot: Most of the data is presented in the age from 15 to 25. We use the scatter plot to see whether there is a correlation between the “Age” and the “Weight” but the plot shows that there is no correlation between them.





Task 2: Hypothesis Testing

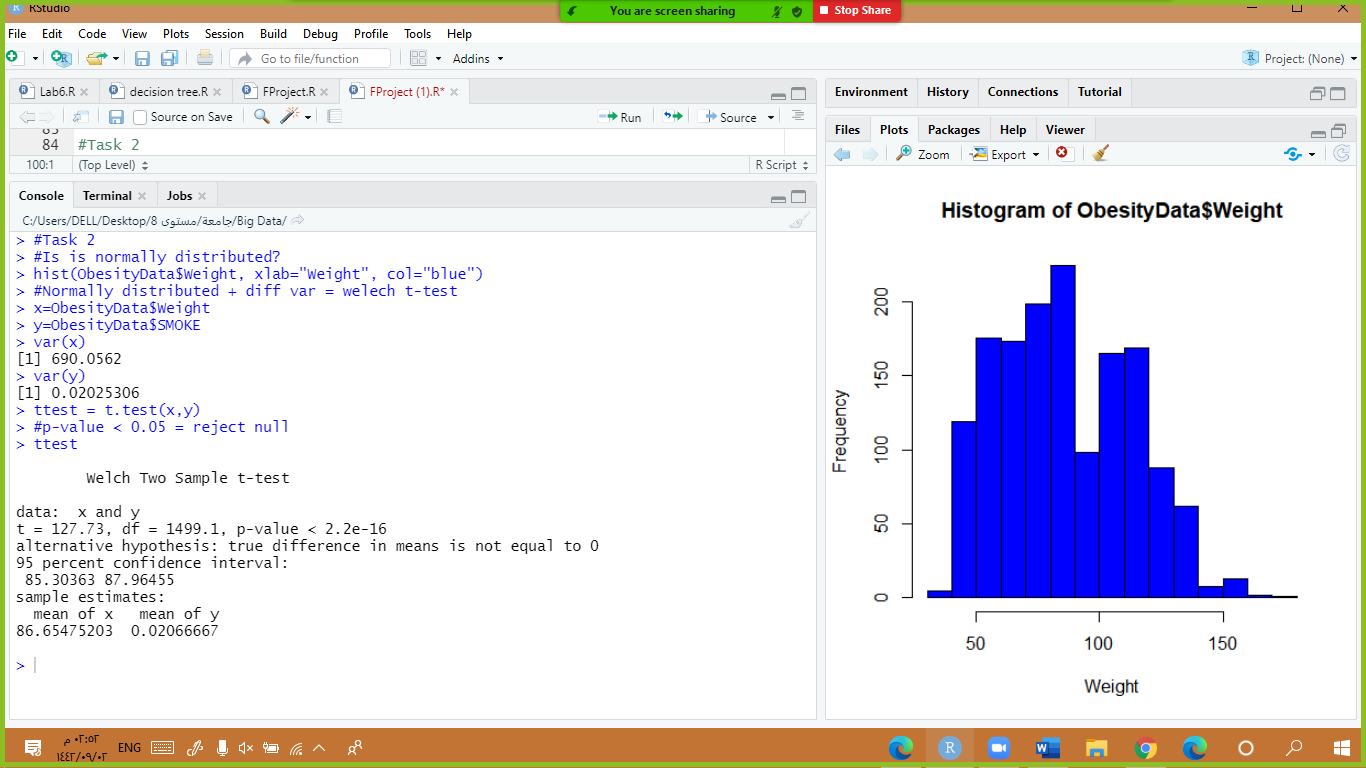
1. Write a hypothesis that you would like to verify. For example, there is no statistically significant difference in the mean of weight between the male and female.

Null hypothesis: there is no statistically significant difference in the mean of the weight between the people who smoke and the people who do not smoke.

Alternative hypothesis: there is a statistically significant difference in the mean of the weight between the people who smoke and the people who do not smoke.

1. Test your hypothesis using an appropriate test.

To choose which test should be used, we first examine whether the weight is normally distributed by using the histogram. Since it is normally distributed we calculate the variance for both variables. The result is different variance so we will use the Welch t-test.

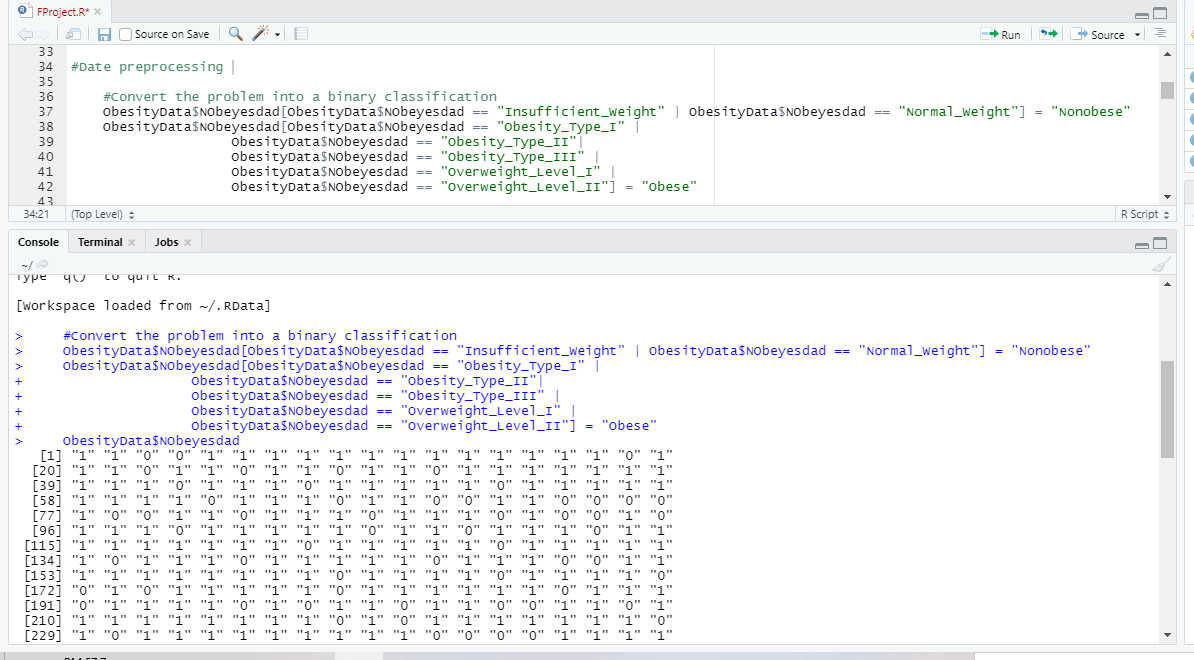


1. What conclusion can you infer from the sample?

Since the p-value is less than 0.05 we will reject the null hypothesis. And say there is a statistically significant difference in mean of the weight between the people who smoke and the people who do not smoke.

Task 3: Model Building& Evaluation

1. Since the goal here is to identify obesity, convert the problem into a binary classification problem. Consider Insufficient\_Weight and Normal\_Weight as ‘Nonobese’ while other categories as ‘Obese’.



1. Build a classification model using:

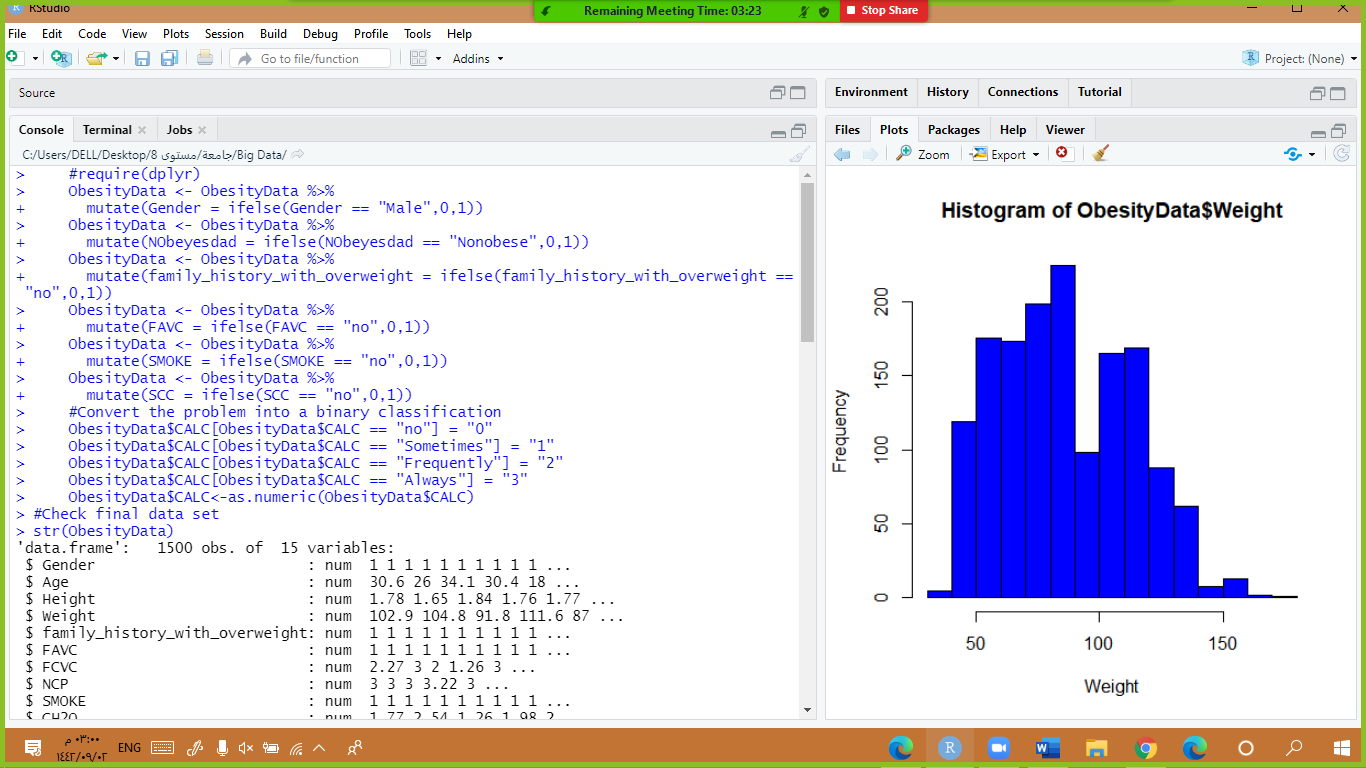
## 3.1 Regression method

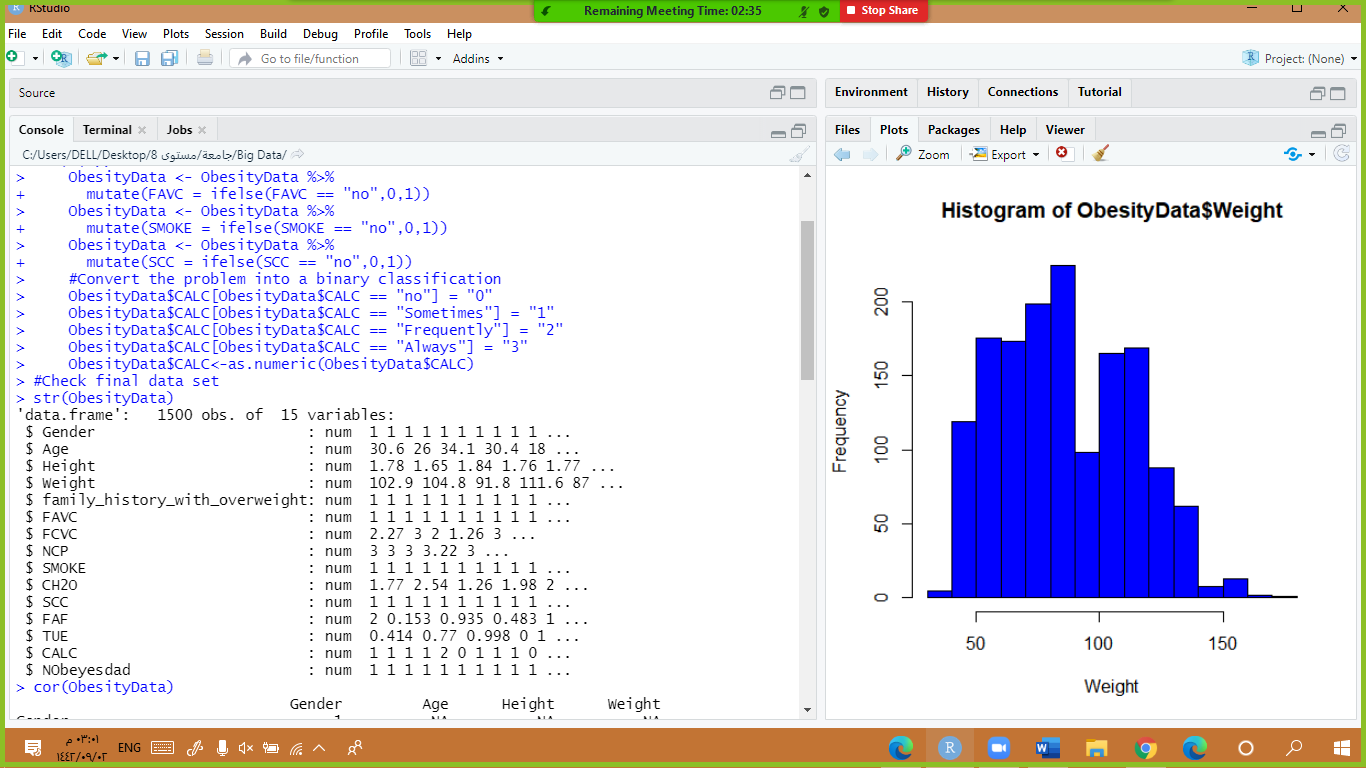
a. What type of regression will you use? why?

We will use logistic regression because we want to predict a categorical output, not continuous, and we want to classify the output as obese or non-obese.

b. Check the assumptions.

We convert the data into numerical.

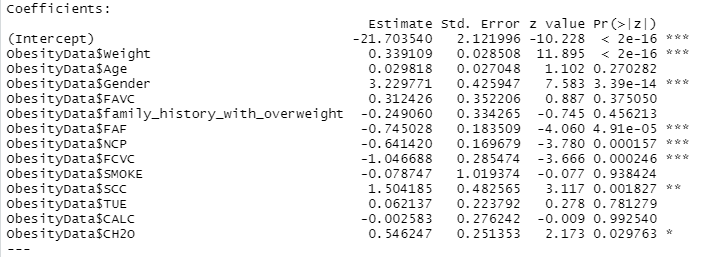




c. Discuss the result.

We build the first model with all attributes in it to see the important attributes. Also, we build the second model and used the attribute which has a high correlation with NObeyesdad and it gives us higher accuracy than the first model.

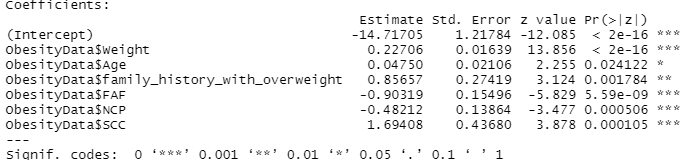
i . Which variables are important/non-important?



The important data: Weight, FAVC, FAF, NCP, FCVC, SCC, CH2O.

Non-important data: Age, Gender, family\_history\_with\_overweight FAVC, SMOKE, CH2O, TUE, CALC.

ii. Which variables have a positive/negative impact on the outcome?



Weight, Age, family\_history\_with\_overweight, and SCC has a positive impact.

While FAF and NCP have a negative impact on the outcome.

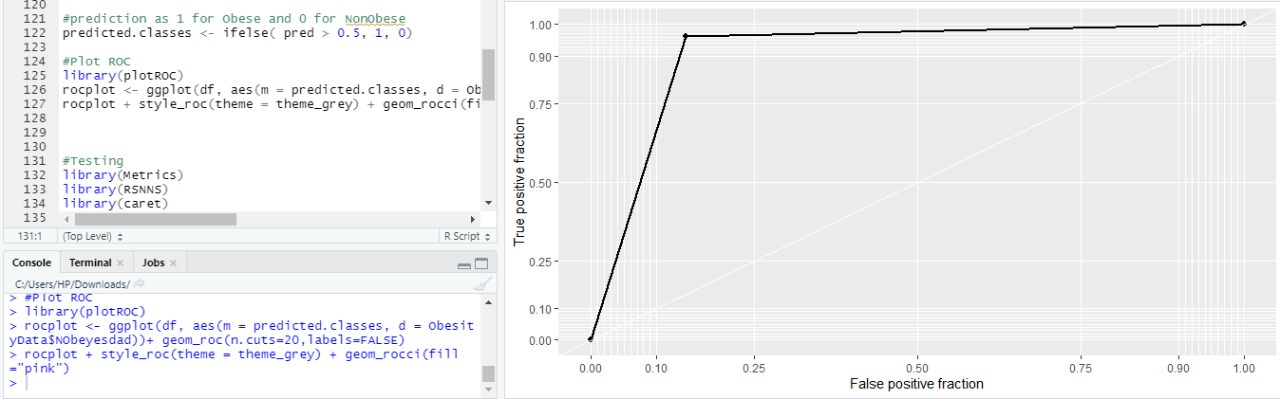
iii. Pick two variables and write a complete interpretation/explanation for their importance and their effects on the outcome.

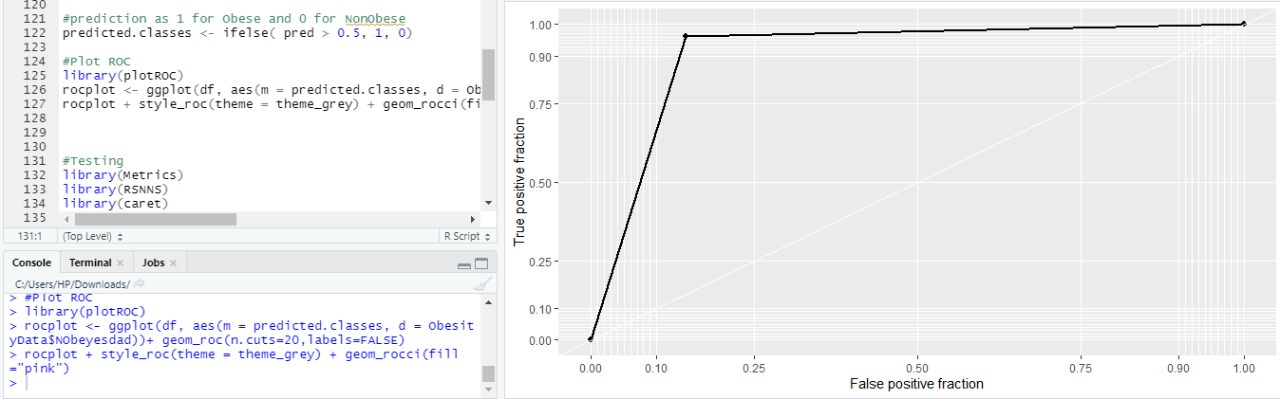
The odds ratio of Obese will be increased by a factor of 1.25 for every unit increased in weight.

the odds ratio of Obese will be decreased by a factor of 0.40 for every unit increased in Physical activity frequency (FAF).

d. Plot ROC curve. Report AUC.

As our roc plot curve is represented, we can say that our model performance is good.





e. Evaluate the developed model using at least three different metrics.

i. Explain each metric. What is it and how is it calculated?

ii. What value did you get?

iii. What does that mean?

**Confusion Matrix**

**Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Actual | | |
|  | Positive 0 | Negative 1 |
| Positive 0 | 354 (TP) | 40 (FP) |
| Negative 1 | 58 (FN)  Precision | 1048 (TN) |

Recall

1. Accuracy

is one of the most popular performance measures to see how good your model is. Simply it is a ratio of correct predicted observations to all observations.

Formula: (TP + TN) / (TP + TN + FP + FN) or #CORRECT\_PREDICTIONS / #TOTAL

We got an accuracy of 93% which mean our model performance is very good.

because it predicted 1402 observation correctly out of 1500!

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

1. Recall (Sensitivity)

Recall (Sensitivity) is the ratio of correctly positive prediction observations to all observations in the positive class. The question recall answer is: Of all the people that truly Nonobese, how many did our mode label as Nonobese?

Formula: (TP) / (TP + FN) or #CORRECT\_POSITIVE\_PREDICTIONS / #TRUE\_TRUTH\_VALUES

We have got a recall of 85% which is good for this model as it is above 50%.

صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

1. Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question Precision answer is of all people that labeled by the model as Nonobese, how many actually Nonobese?

Formula: (TP) / (TP + FP) or #CORRECT\_POSITIVE\_PREDICTIONS / #POSITIVE\_SAMPLES

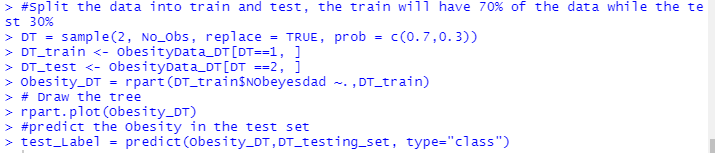
We have got 89% precision which is pretty good since high precision means a low false-positive rate.

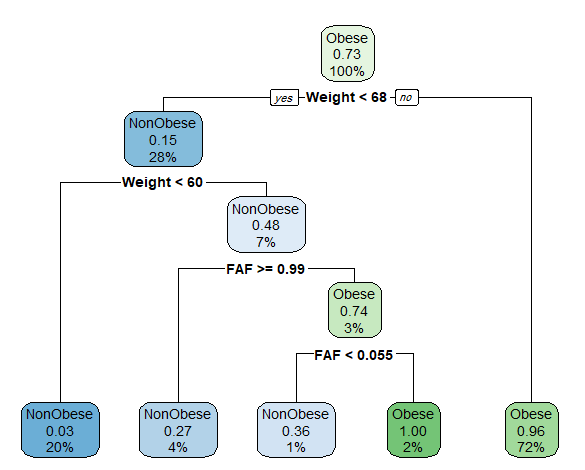
صورة تحتوي على نص

تم إنشاء الوصف تلقائياً

## **3.2 Decision Tree**

1. Draw the tree and write at least two rules.





### Rules:

If the result is less than 0.5 then it is non-obese, otherwise it well be classified as obese.

1. If weight > 68 then it is classified as obese.
2. If weight < 68 and weight <60 it is classified as nonobese.
3. What are the two most important variables?

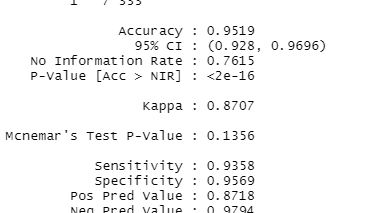
The most important variables are weight and family\_history\_with\_overweight.



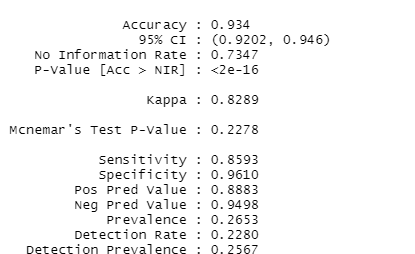
1. Which method performed the best?

Based on the accuracy and the sensitivity the decision tree is the best model

Decision tree evaluation



Logistic regression evaluation



References

* Obesity based on eating habits & physical cond. (2020). Retrieved 14 April 2021, from <https://www.kaggle.com/ankurbajaj9/obesity-levels>
* Housing Prices Competition for Kaggle Learn Users. (2020). Retrieved 14 April 2021, from <https://www.kaggle.com/c/home-data-for-ml-course/discussion/143364>
* Geerinck, X. (2020). Artificial Intelligence — How to measure performance — Accuracy, Precision, Recall, F1, ROC, RMSE…. Retrieved 14 April 2021, from <https://medium.com/@xaviergeerinck/artificial-intelligence-how-to-measure-performance-accuracy-precision-recall-f1-roc-rmse-611d10e4caac>
* Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog. (2016). Retrieved 14 April 2021, from https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/