Matea Marinkovic
February 2022
Introduction to Data Science
Assignment 1

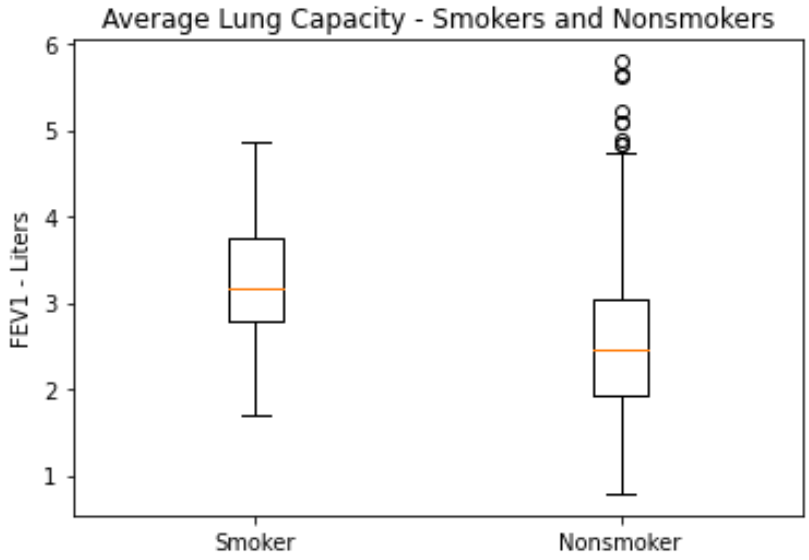
Assignment 1 Report

Exercise 1 – Reading and Processing Data

FEV1 for smokers: 3.2768615384615383
FEV1 for nonsmokers: 2.5661426146010187

I was surprised to see that the average lung functions were so close also that smokers had a higher lung function than non-smokers.

Exercise 2 – Boxplots



From the boxplot you can see that the smokers have a higher average lung function than the non-smokers, but that there were some outliers from the non-smoking group, as well as more even quartiles and larger whiskers.

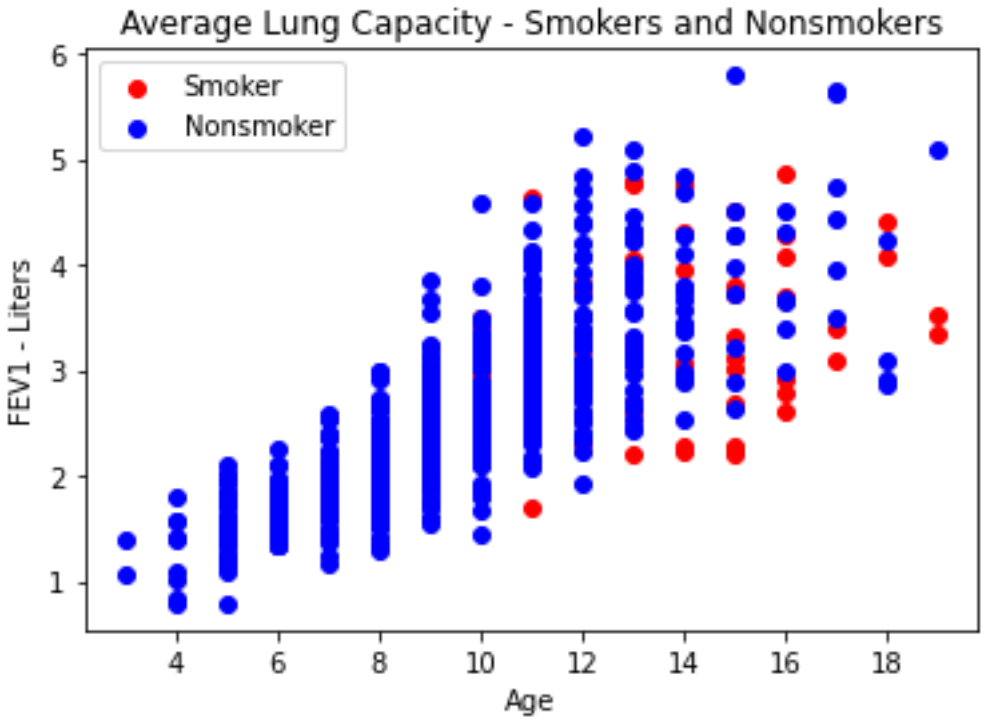
Exercise 3 – Hypothesis Testing

Value of T statistic: 7.149608129503808
Degrees of Freedom: 83.0
Returned P-Value: 3.1173573925292966e-10
Accepted/Rejected hypothesis: Null hypothesis is rejected.

I am not surprised by the outcome that the null hypothesis is false, we could tell initially that the two group means were not equal. By doing this T-test however, we determined that not only are the sample means unequal, but that the difference is statistically significant as out p-value is lower than the significance level of 0.05.

Matea Marinkovic
February 2022
Introduction to Data Science
Assignment 1

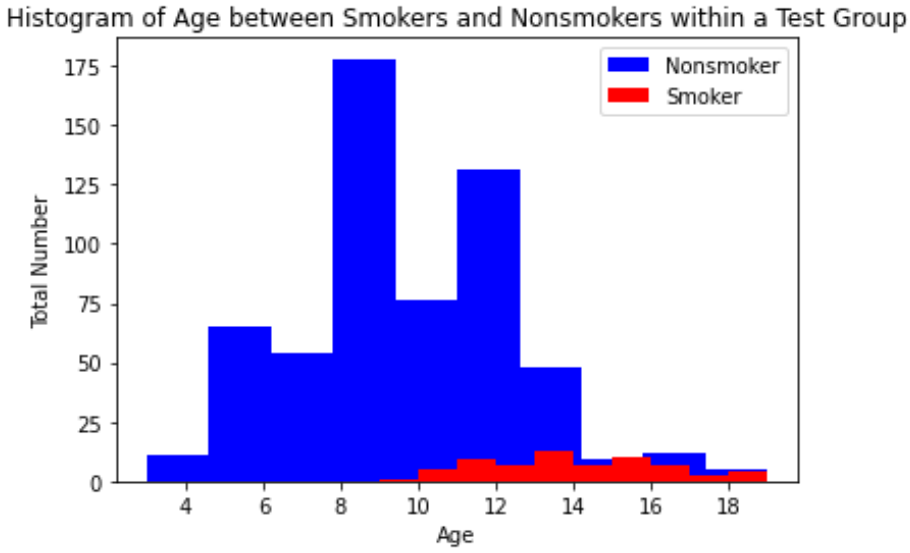
Exercise 4 – Correlation



Correlation value: 0.7564589899895996

We can see that with non-smokers, lung function generally increases with age. With smokers however, there seems to be much less of a distinguishable correlation, perhaps due to the difference in smoking habits amongst the samples.

Exercise 5 – Histograms



We can see from the histograms that there are very few young smokers, and that the majority of smokers are 12 years old and older, while the non-smoker samples are in high numbers

starting at the age of 6. Because the smoker population is consisting of mostly older members than the majority of the non-smokers, their average lung function should be higher based on what we know about lung function increasing with age. In this dataset, we see that perhaps the number of younger members of the non-smoker group might slightly skew the average lung function not because of the smoking status but because of the age.