University of Copenhagen

Faculty of Science

Introduction to Data Science

Assignment 1

Dimitrios Galinos (bst265)

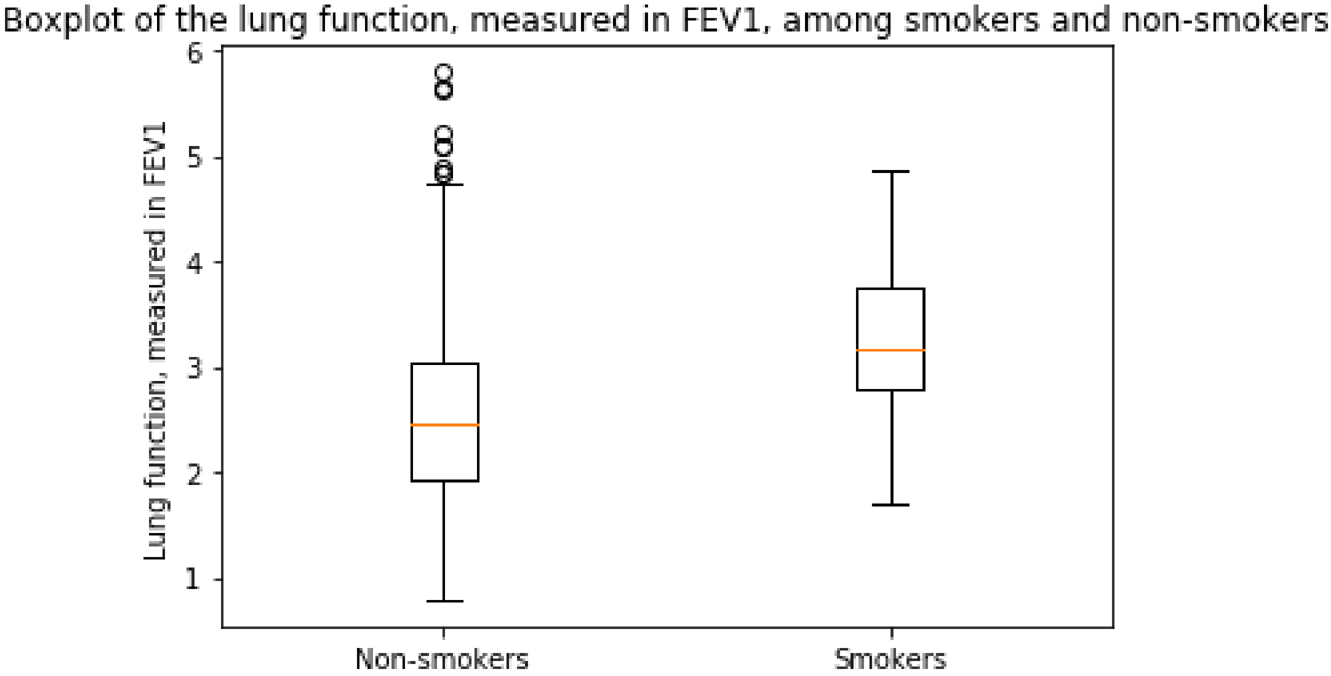
February 17, 2020

1 Reading and processing data

The code is included inside the code.zip and is named "assignment1.py". Please also read the README file.

I have successfully read the data from smoking.txt, I divided the two datasets into groups consisting of the smokers and the non-smokers. My average FEV1 scores are 3.2768615384615383 for smokers and 2.5661426146010187 for non-smokers. At first I was actually surprised that the smokers showed higher lung-function than the non-smokers. This was not something I was expecting since it is generally believed that smokers have lower lung-function.

2 Boxplots



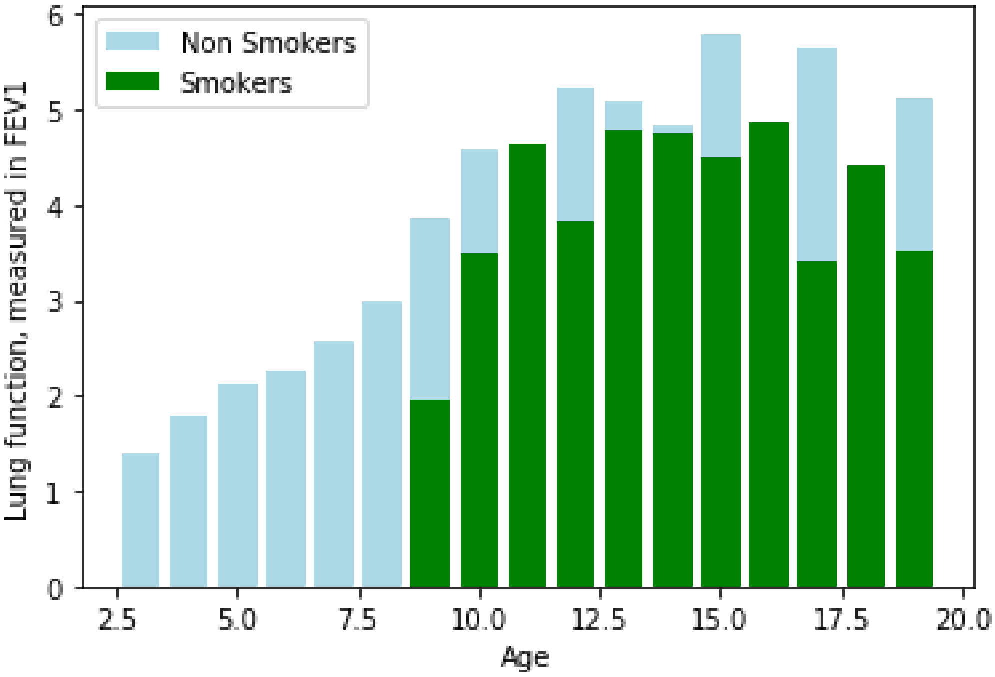
With the boxplot I can see that the non-smokers have in general lower lung-function with the exception of some corner cases whose lung function is actually greater than that of the smokers.

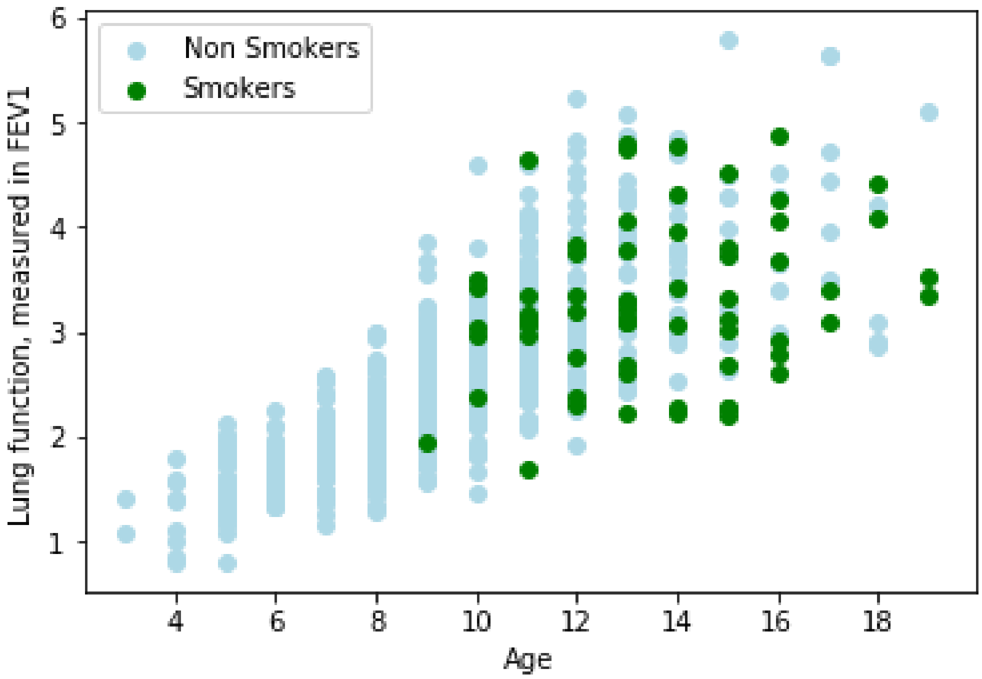
3 Hypothesis testing

I tested the automated way to do the two-sided t-test using scipy's ttest_ind and I also did my own implementation of the whole two-sided t-test as requested. The value of the t-statistic from my own implementation is 7.149608129503808. The floor rounded degrees of freedom are $\nu = 83$. The returned p-value is $3.1173573925292966e - 10$. My response for the hypothesis is that I reject the null hypothesis. Since our significance level is $\alpha = 0.05$ and our p-value is way lower than 0.05 we have to reject our null hypothesis of the two populations having the same mean. What is also a bit interesting is that scipy's ttest_ind doesn't give me exactly the same result but it has an error margin of $1e - 11$, I have also tested scipy's ttest_ind for our example from the slides (the happy pills) and found that it had a very small margin error there as well. Finally I am not surprised by the result of rejecting the null hypothesis because I already knew that the mean of those two populations had a significant difference.

4 Correlation

I did a barplot and a scatterplot for the visualization of the correlation. Even though the barplot probably looks prettier the scatter plot gives us more insight about our data:

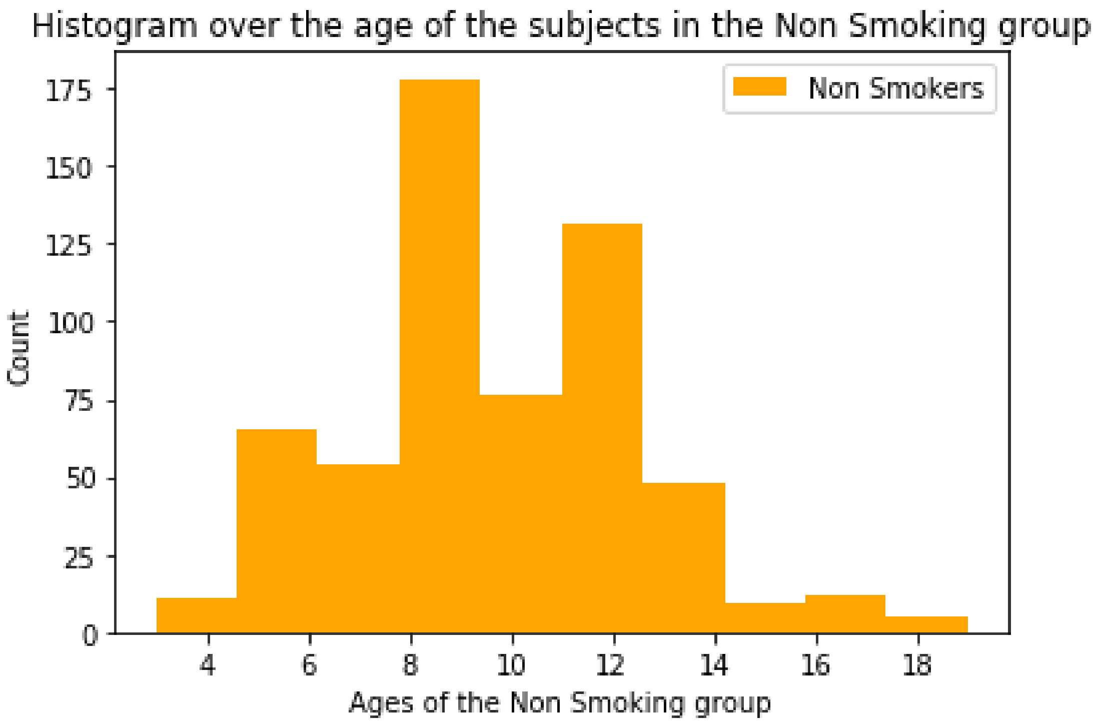




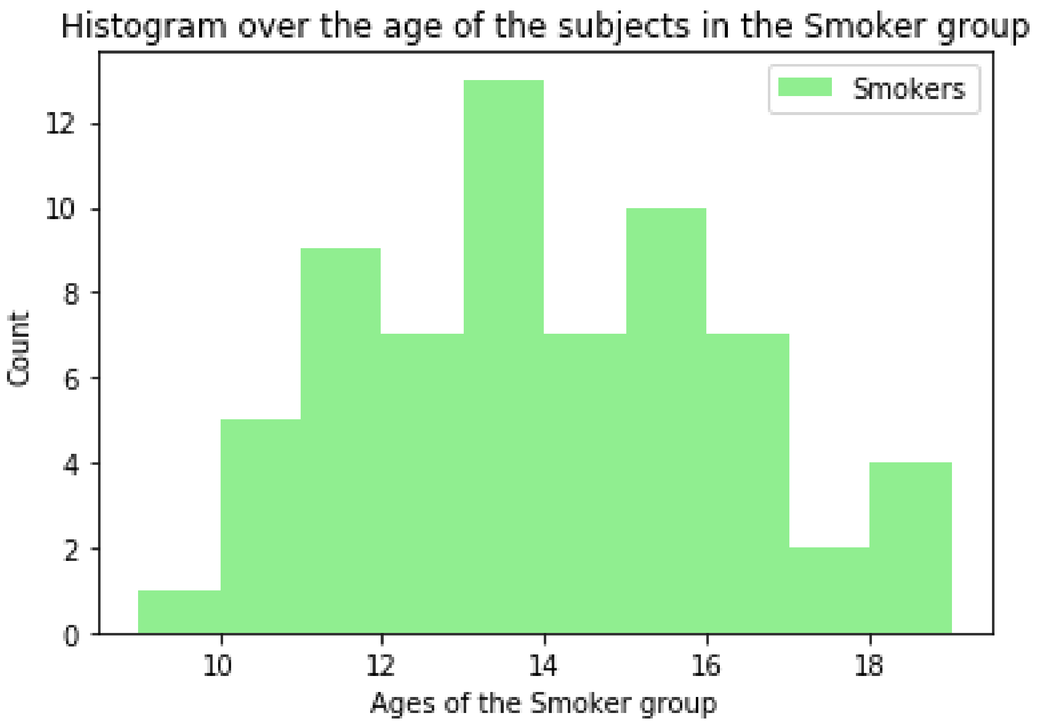
Furthermore I calculated Pearson's correlation coefficient between age and FEV1 for the whole dataset which turned out to be 0.7564589899895997. My main comments is that it is obvious that age is strongly correlated with FEV1 levels and the older somebody is the higher FEV1 levels we can expect from him. Furthermore the fact that the population of the smokers is compromised of people older in average than the population of non-smokers would explain our previous findings of smokers having higher average FEV1 levels than non-smokers, it is because non-smokers in general are younger in our dataset.

5 Histograms

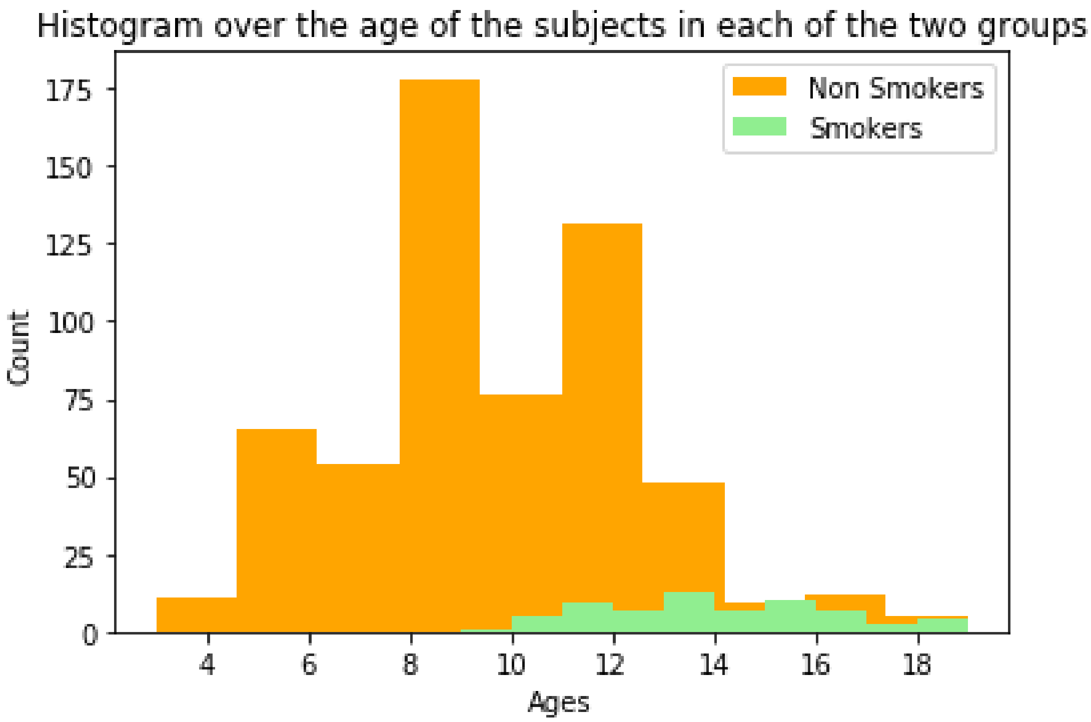
I have created both the histograms requested with one group in each one but I have also created a histogram with both groups in the same because I found it more meaningful and easier to compare values like that.



Histogram over the age of the subjects in the Non Smoking group



Histogram over the age of the subjects in the Smoker group



This explain our results on lung function in the two groups. As mentioned in the previous exercise (whose scatter plot helped a lot understanding the data) the fact that the population of the smokers is compromised of people older in average than the population of non-smokers would explain our previous findings of smokers having higher average FEV1 levels than non-smokers, it is because non-smokers in general are younger in our dataset.