Introduction to Data Science

Assignment 1

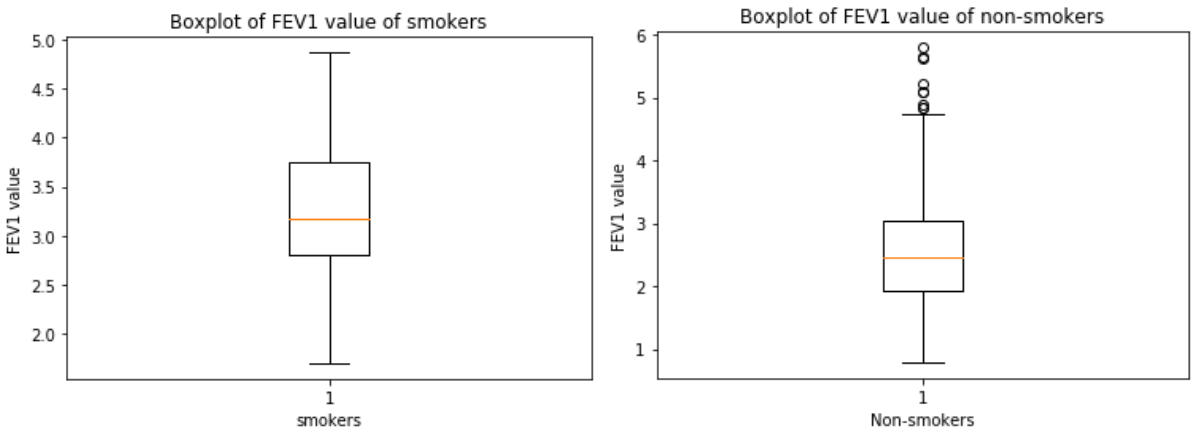
Péter Pölös

Exercise 1

```
The mean FEV1 of smokers: 3.2768615384615383
The mean FEV1 of non-smokers: 2.5661426146010187
```

From the output of Python we can see that the Non-smokers have smaller average FEV1-value compared to the smokers group, meaning that they have worse lung capacities which at first could seem surprising since we would expect the group of smokers to do worse on this test due to the negative effect of smoking on their lungs. But what is worth keep in mind we do not know yet the structure of these two groups so let's dig deeper.

Exercise 2



From the boxplots it can be seen that both the 1st and 3rd quartiles of the non-smokers is below of the corresponding quartile values of the smokers. It can also be seen that the non-smokers group have much more outlier data points and that their values move on a wider spectrum. (spreading from 1 to 6 compared to the smoker's 1 to 5)

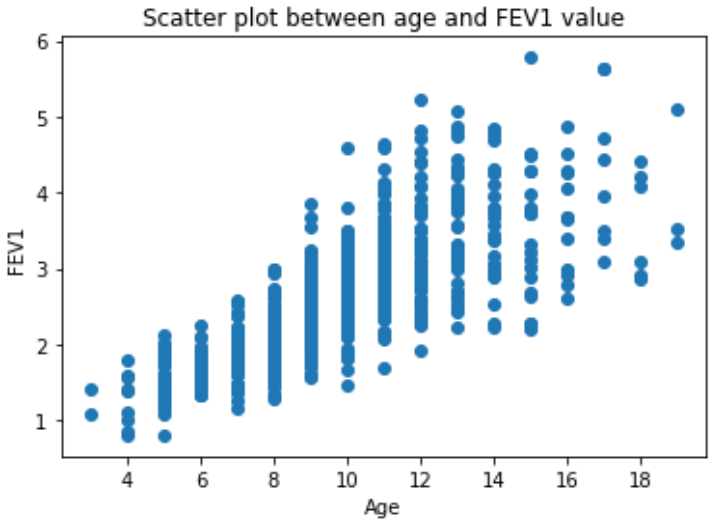
Exercise 3

```
Hypothesis: Reject
p-value: 1.9928459182932803e-10
T-statistic: 6.46445317259654
```

We should reject the null-hypothesis, with the significance level of 0.05. Meaning that there is a statistically significant difference between the mean FEV1 level of smokers and non-smokers.

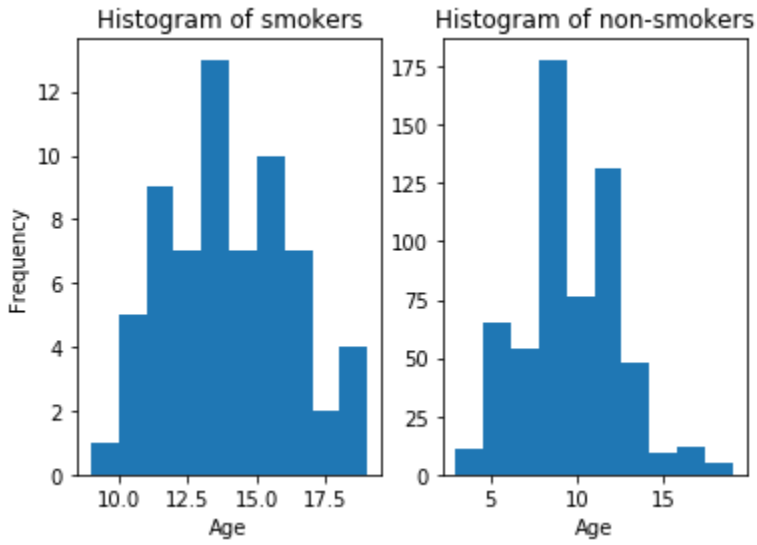
Exercise 4

```
The correlation between age and FEV1 value is 0.7564589899895996
```



As it can be seen form the plot there is a strong positive linear trend between the two variables, so as the age increases the FEV1 value increases with it. Checking the correlation between the two variables, which is pretty high, around 0.76, just confirms this theory.

Exercise 5



It can be seen from the histograms that the smokers are more heavily distributed towards higher age levels compared to the non-smokers. This difference can be seen between their most frequent age groups as well, in case of smokers it is 13 years old, while in case of non-smokers it is 9.

That explains the earlier results we have seen , which first seemed surprising that the smoker group has higher mean FEV1 value, so better lung capacities, but that is only due to the fact that they are in

general older than the kids in the non-smoker group, as we have seen from the previous scatter plot as age increases the FEV1 value increases with it, meaning that as they get older their lungs are more developed, and since all of them is younger than 20 , their lung capacities have not started to deteriorate yet due to their age.