Introduction to Data Science - Assignment 1

Shahriyar Mahdi Robbani

February 14, 2020

Exercise 1

```
Average FEV1 for nonsmokers: 2.5661426146010187
Average FEV1 for smokers: 3.2768615384615383
```

Since FEV1 is an indicator of lung function, it is quite surprising smokers have a higher average FEV1.

Exercise 2

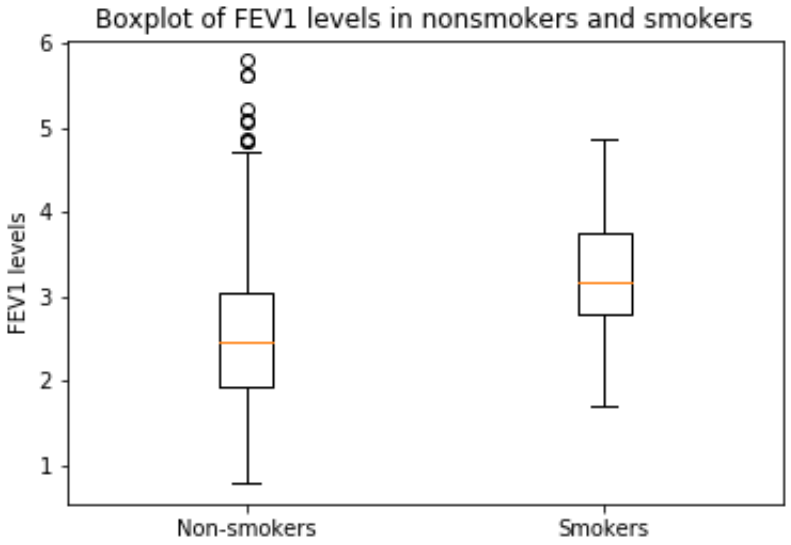

Figure 1: Boxplot of FEV1 levels in nonsmokers and smokers

The non-smokers have a much larger range of FEV1 values and many extreme values that are larger than the Smoker FEV1 values.

Exercise 3

```
T value: 7.1990318609997095
P value: 2.4945644815274697e-10
Degrees of freedom: 83.0
The null hypothesis is rejected at alpha = 0.05
```

The p value is much smaller than alpha (0.05) which indicates the difference between the two means is very statistically significant. This implies our strange results did not occur by chance so the FEV1 levels of our smoker sample is indeed higher than that of the nonsmokers.

Exercise 4

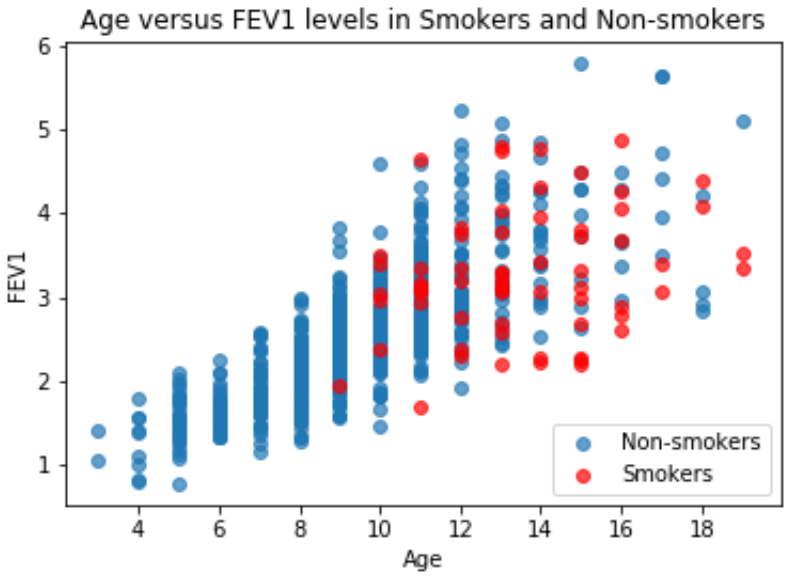

Figure 2: Age versus FEV1 levels in Smokers and Non-smokers

```
Pearson coefficient: 0.7564589899895999
Spearman coefficient: 0.7984229001546537
```

The age and FEV1 levels are positively correlated for both the Pearson and Spearman coefficients. This means older individuals will have a higher FEV1 level.

Exercise 5

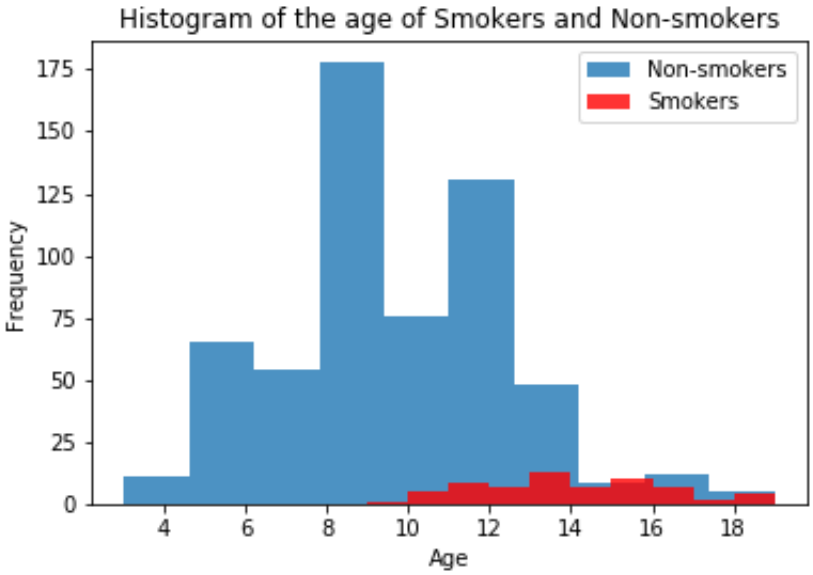

Figure 3: Histogram of the age of Smokers and Non-smokers

The histogram confirms that the Non-smoker group consists of many more individuals who are much younger than the smoker group. Since smaller children will have a smaller lung capacity, the lower FEV1 values obtained from the Non-smoker group is logically consistent.