# Assignment 1 – Introduction to Data Science

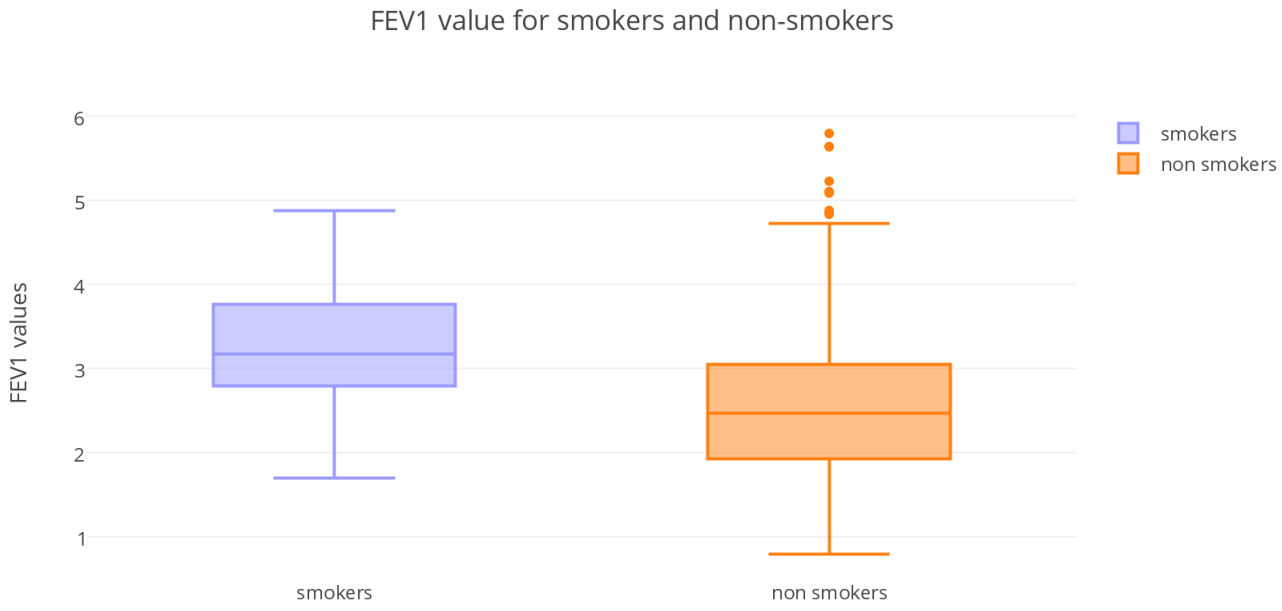
**Modules used:**

The plots in the code are made with the library 'plotly' and once the code is run an html page is opened in which the plots can be explored interactively.
In the file corr.py, uncomment the line from 88 to 90 to show the histograms reported here.

## Exercise 1

The average FEV1 values for smokers and non-smokers are: **3.2768615384615383**, **2.5661426146010187.**
The values show a lower average FEV1 for non-smokers which indicates a decrease in lung function. This doesn't match with what was expected: the FEV1 should be lower for the smokers group as it's widely known that smoking affects the breathing functions negatively.

## Exercise 2

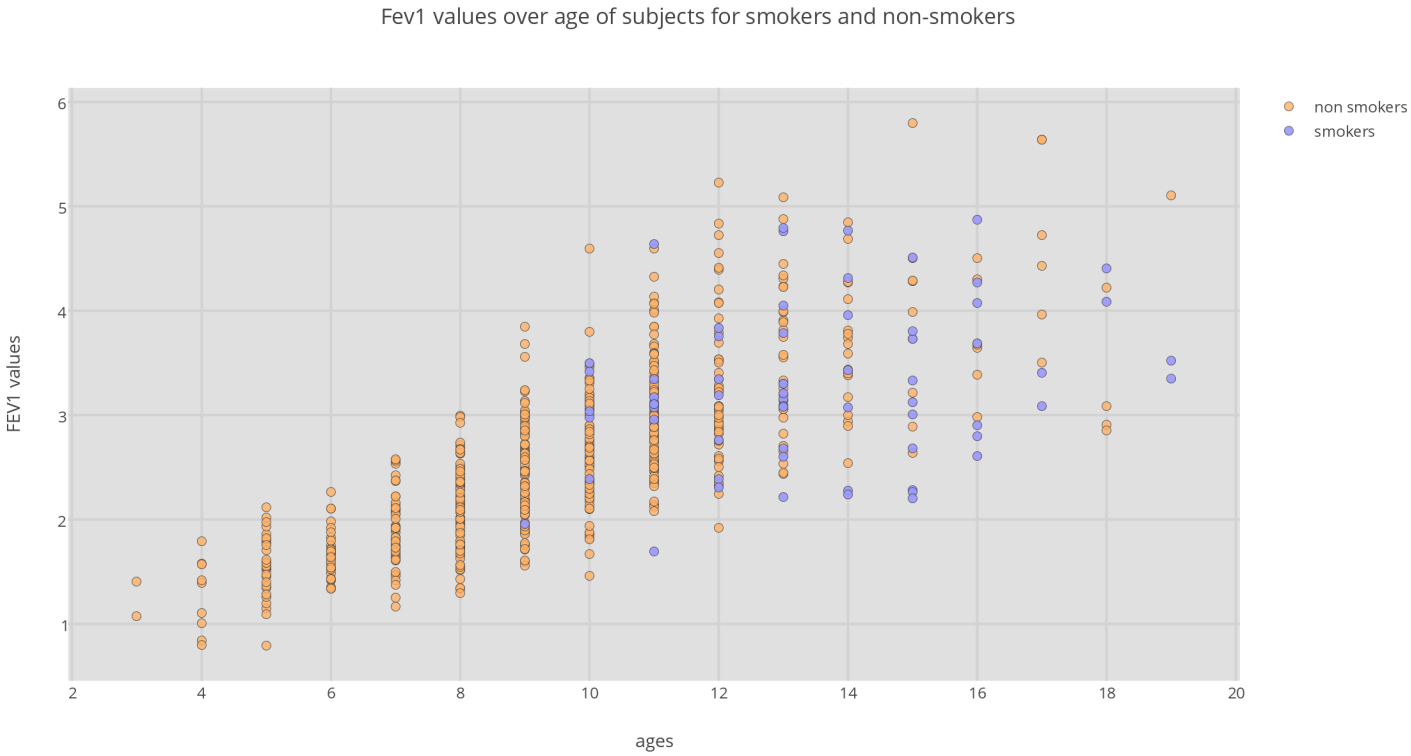



The plot shows that the FEV1 average values are higher for the smokers group than for the non-smokers.
The median (middle quartile) marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less. The top box in the boxplots is the third quartile of the sample, 75% of the scores fall below the upper quartile. The bottom of the box is the second quartile, 25% of scores fall below the lower quartile. The upper and lower whiskers represent scores outside the middle 50%.
In the non-smokers boxplot, there are some outliers plotted with circles, they have FEV1 values higher than the other subjects in their group.
Moreover, according to description of the FEV! function, the value of the median for non-smokers should be higher than the one for smokers. Whereas from the plot we observe the opposite behaviour

**Exercise 3**

The t value is: **7.1496081295**
Degrees of freedom: **83.0**
The p value computed manually is: **3.11735739253e-10**
The p value computed with the built-in formula is: **3.07381274488e-10**
The null hypothesis is that the two populations have the same mean.
The hypothesis is rejected because the p-value resulting from the T-test is smaller than the significance level of α = 0.05. From this we can conclude that the mean of the two populations are different.

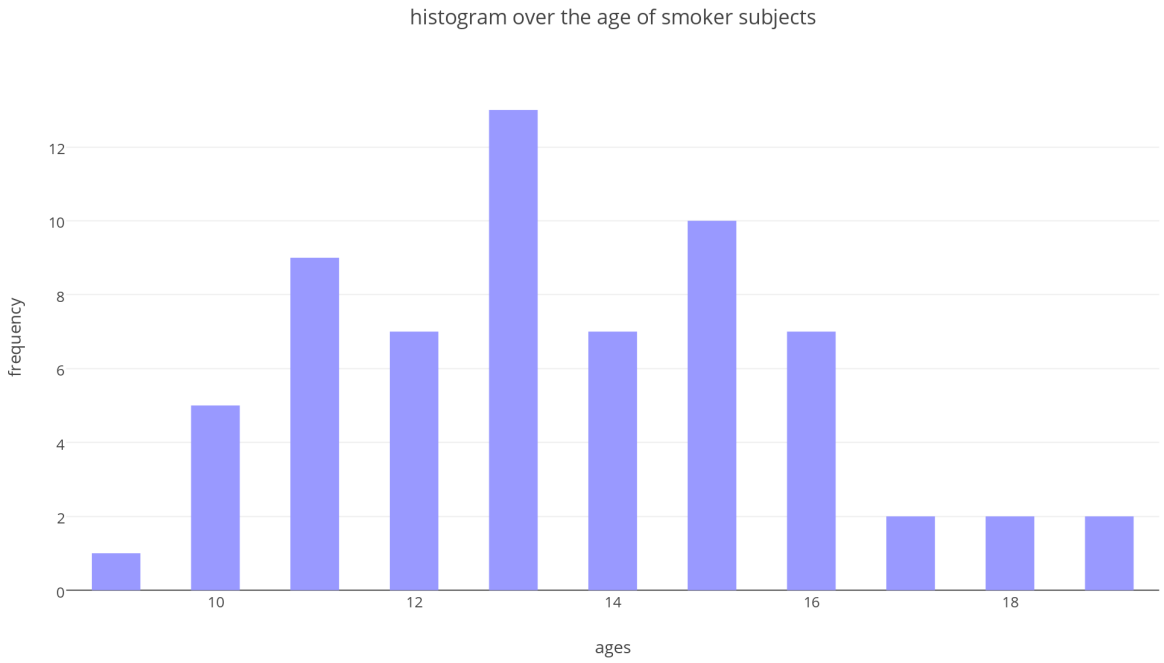
**Exercise 4**



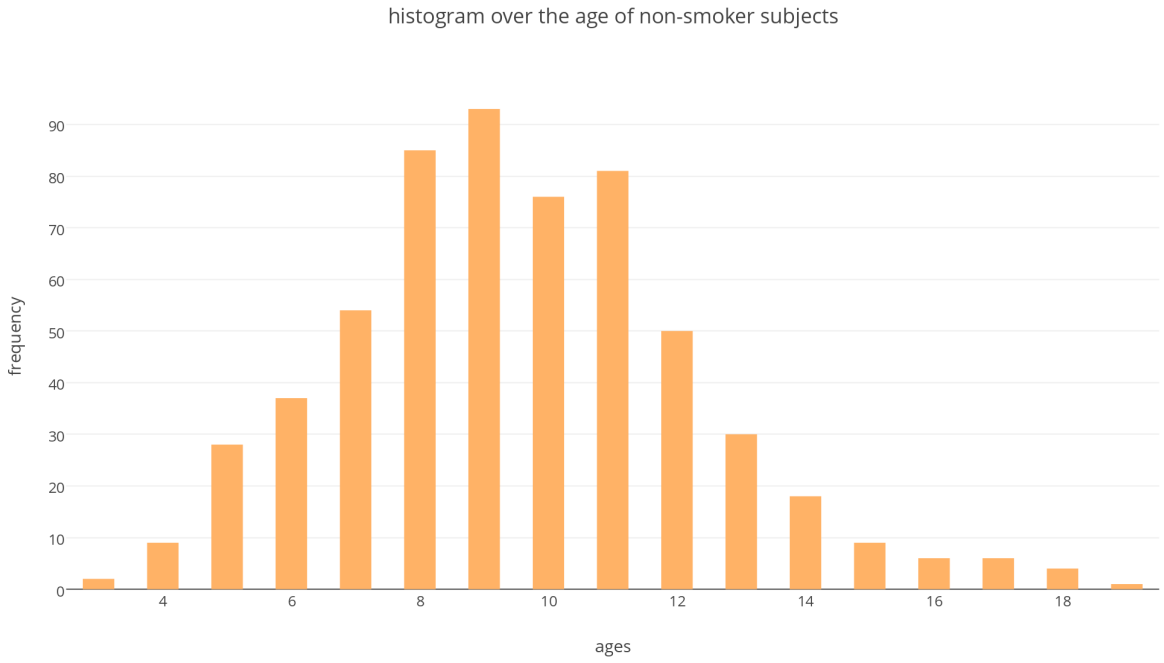

The plot shows the FEV1 values over the ages in all the subjects from the data provided, in orange the values of non-smokers and in purple the values of smokers.
From the plot, it is possible to notice that highest FEV1 values are found in non-smoking subjects starting from age greater than 15.
On the other side, the lowest values are those of children below 7 years old, this could probably be explained because their lungs are not fully developed yet and they result having low values of FEV1.
From this plot, it is also possible to notice that in the data there are more non-smokers than smokers.
The correlation coefficient manually computed is: **0.75645898999**
The Pearson correlation coefficient computed with the built-in formula is: **0.75645898999**
The correlation coefficient ranges from −1 to 1. In this case, the value is closer to 1 than to 0, this means that there is a linear equation describing the relationship between ages and FEV1 values, with all data points lying on a line for which FEV1 increases as age increases. A value of 0 would imply that there is no linear correlation between the variables.

**Exercise 5**

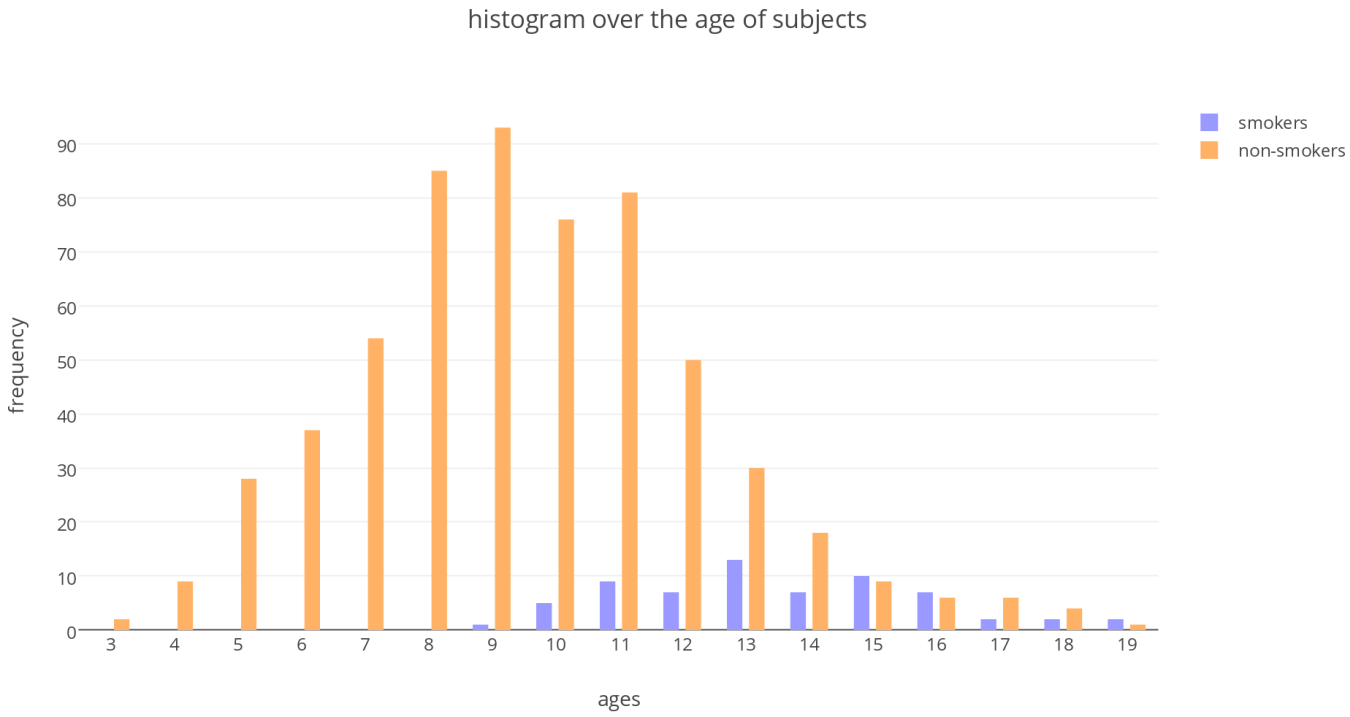
The first histogram shows the age of smoker subjects and their frequency in the data space. From the histogram, it is possible to notice that the highest frequency of smoker subjects has 13 years old. The lowest value recorder is only one subject of 9 years old.




The second histogram shows the age of non-smoker subjects and their frequency in the data. Firstly, it is possible to notice that the highest number of non-smoker of 9 years old is around 90 subjects. The lowest values in the distribution are subjects below 4 and above 18 years old.



To make a better comparison of the results the two histograms are plotted below. From this it is possible to notice that the samples sizes are very different. In the data, there are more non-smoker subjects than smoker.




This explains why the results on lung function in the two groups should have been different. As a matter of facts, the average FEV1 values of the non-smokers result lower because they are computed on a larger sample.