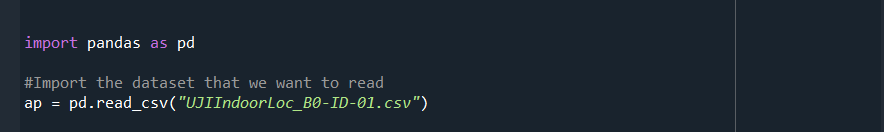
***Week 1: Exploratory Data Analysis, Dataset UJIIndoorLoc\_B0-ID-01***

1 - Are the values numerical or nominal? In this case you can load them in both ways.

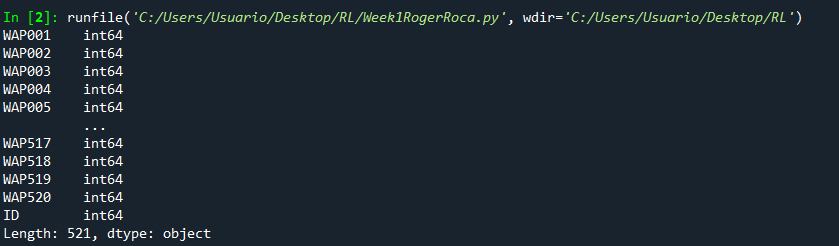
The first step in order to be able to classify the following data is to read the file. The file is UJIIndoorLoc\_B0-ID-01.csv. Once the file is loaded into Python, we will have to see the attributes and the class of the stored data. There are different ways to read a csv file and we will use the one that requires pandas.

First, we opened the file (File – Open File - UJIIndoorLoc\_B0-ID-01.csv) and then we ran the code:



After the file has been read, we have to see which type of attributes we can find in it.





The console shows us that all the data stored in the loaded file has attribute and type int64 so we can affirm that all of them are numerical values. The ID also belongs to the int64 category so it can also be a numerical value.

The access points can be 2 possible values. These values just mean if the access points cover a room or do not (0 and 1, indistinctly) so it can also be represented as a nominal value.

2 - Do we have enough samples for each ID class label? An easy way to look at it

is by counting how many instances there are of each tag.

Now we will focus on the amount of data that each tag has.



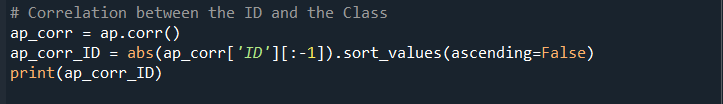
We decided to order them in descending order to have the tag with the most access points on top of the list.

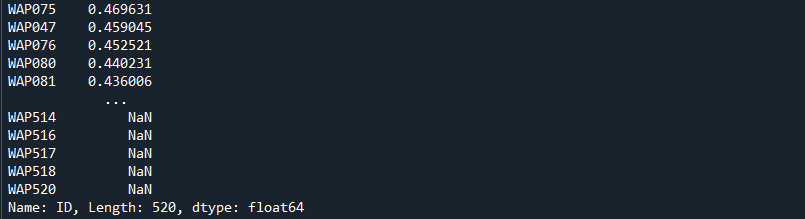


By observing the console, we can see that the ID 120 is the one that has the most access points with 42. Then some of them have 30 the last ones have 10. Considering the amount of access points we previously found (521 was the length but there is the ID part so there are 520 access points) and the number of classes seen in the last image (256 classes) we can tell that we might not have enough to determine in which room we are by looking at the access points that cover a certain room.

3- Do we have WAPXXX that can help us predict ID? Indicates 5 access points with greater potential in prediction. Indicate 5 access points with no impact on the prediction. A simple way to look at it is with the correlation of each of the attributes with the class.

In this exercise we will have to see the correlation between the ID and the attribute. If the correlation is low, it means that the access point we are looking at does not give much information. On the other hand, if the correlation is high, it means that that access point gives a lot of information about the room we are in. The correlation has to be analysed in absolute value.





Since we have ordered the access points in descending order the first 5 on the list (WAP075, WAP047, WAP076, WAP080 and WAP081) are the ones that will help the most and the last 5 (WAP514, WAP516, WAP517, WAP518 and WAP520) are the ones with the lowest correlation so they will not give enough information.

If we look a little bit further and decide to check the values of these last ones:

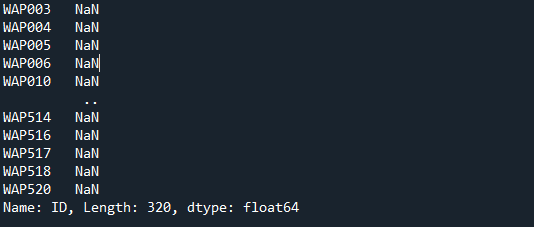




So, we can see that all the values are 0 which means this particular access point does not cover any room.

And if we wanted to see how many of those access points have null correlation (we were only asked to find 5 previously):





And here we can see that out of the total there are 320 access points that have null correlation so these ones will not help us.