Exploratory Data Analysis Part-1

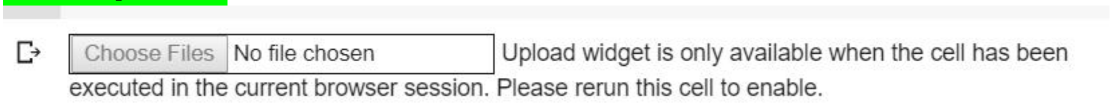
Hi Everyone. Hope you are doing well. To continue further with this tutorial, you need to have a basic understanding of Python Programming and have Anacondas in your System. I explained it in the tutorial before this. You can go back and check it. If you don't have enough time for the installations, you can use Google colab.

In this tutorial, we will first learn about Exploratory data analysis, which is the primary step of data science and machine learning. We will use various techniques and plots to study the structure of data and determine what variables in the data are more important for our end goal. Data cleaning is a significant part of applying any machine learning algorithm.

We are going to perform EDA on Wisconsin Breast Cancer Database. You can download it from here.
(https://www.kaggle.com/roustekbio/breast-cancer-csv)
Our goal is given all the attributes, and we should determine whether a tumour is benign or malignant. We will see what the vital features are in predicting this, although predicting is out of the scope of this tutorial. You can read the description of data here.
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)
The code is highlighted in green colour.

(If you are using Google colab)

```
from google.colab import files
files.upload()
```



Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Choose files and upload the dataset that you downloaded into the system.

If in your local system:
Open the folder in which your dataset is present and open command prompt. Type Jupyter notebook and press enter.

2. Importing necessary modules

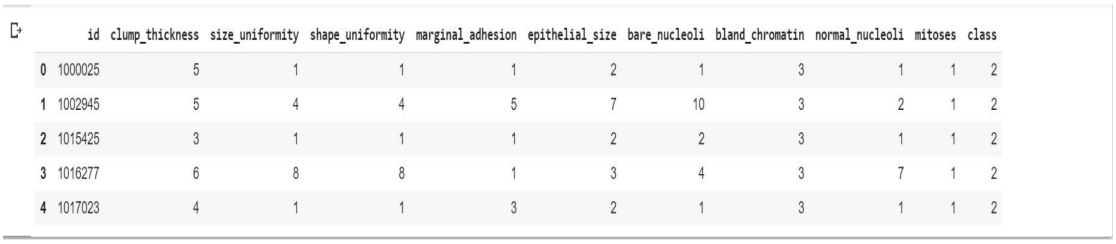
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

3. Reading the data

```
breastCancer=pd.read_csv('breastCancer.csv')
```

4. Visualizing data

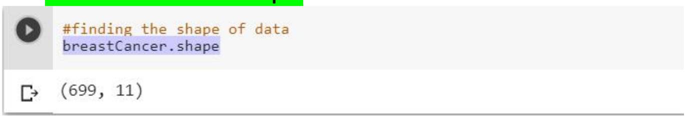
```
breastCancer.head(5)
```



	id	clump_thickness	size_uniformity	shape_uniformity	marginal_adhesion	epithelial_size	bare_nucleoli	bland_chromatin	normal_nucleoli	mitoses	class
0	1000025	5	1	1	1	2	1	3	1	1	2
1	1002945	5	4	4	5	7	10	3	2	1	2
2	1015425	3	1	1	1	2	2	3	1	1	2
3	1016277	6	8	8	1	3	4	3	7	1	2
4	1017023	4	1	1	3	2	1	3	1	1	2

5. Finding the shape of data

breastCancer.shape



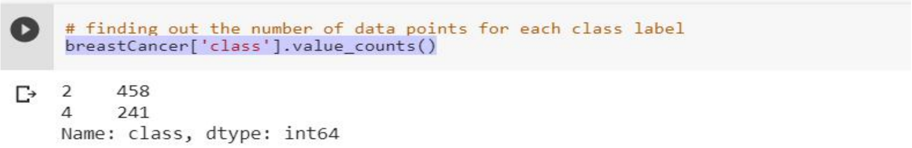
```
#finding the shape of data
breastCancer.shape
```

```
(699, 11)
```

Data has 699 rows and 11 columns. In other words, there are 11 features and 699 datapoints.

6. Finding out the number of data points for each class label

breastCancer['class'].value_counts()



```
# finding out the number of data points for each class label
breastCancer['class'].value_counts()
```

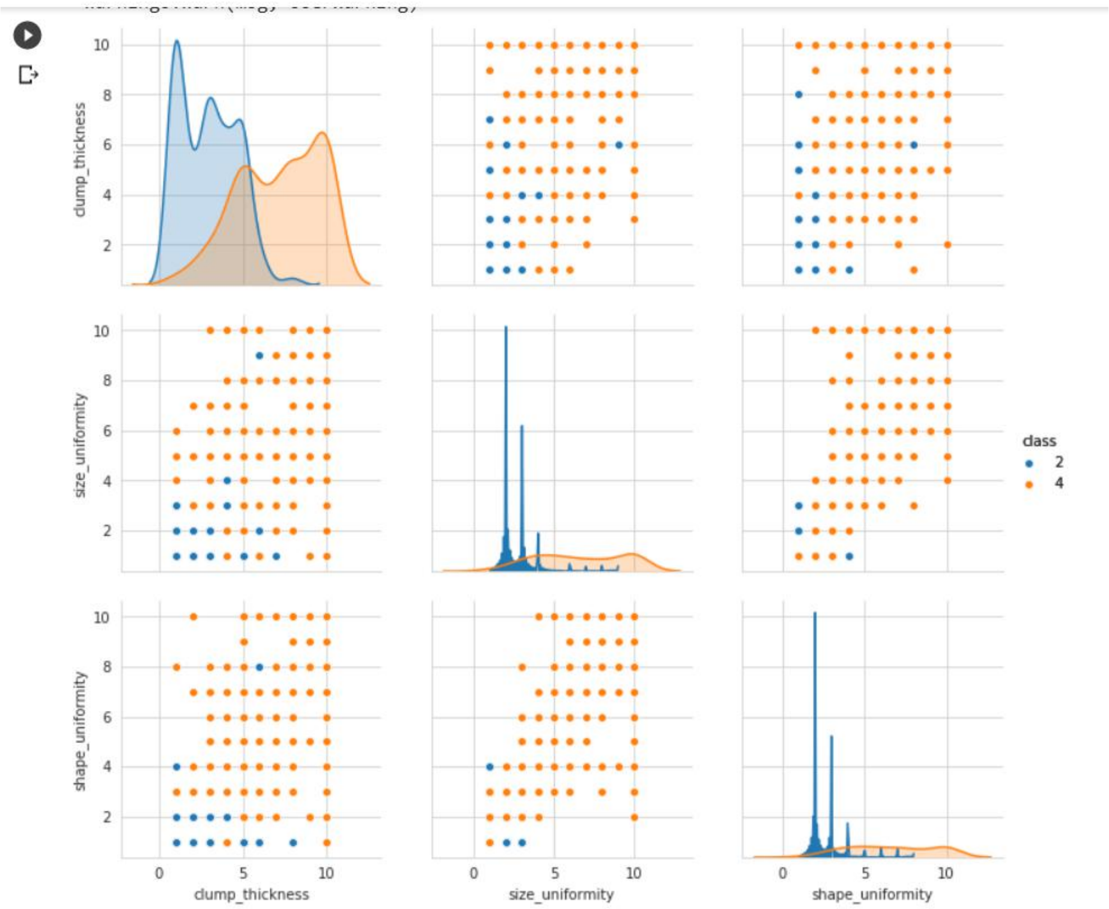
```
2    458
4    241
Name: class, dtype: int64
```

As you can see the data is unbalanced even though the difference is not too much.
There are 458 data points for benign tumours and 241 data points for malignant ones.

Pair Plots

Pair plots don't work if the dimensionality of the data is high even though they are suitable for low dimensionality data. The total number of plots we get are nc2, which is a high number if we have 100 dimensions. Going through those many plots and observing the structure of the data will take up a lot of time. This is a drawback in using pair plots.

```
sns.set_style('whitegrid')
sns.pairplot(breastCancer,hue='class',vars=['clump_thickness','size_uniformity','shape_uniformity'],size=3).add_legend()
plt.show()
```



From the above plot, we can tell that there is a fair degree of overlap between the two class labels. But we can deduce that size_uniformity and shape_uniformity are more important as the degree of overlap is around a small region. I only took three variables as it would be easy to explain. Please consider all the 11 features for this pair plot and find out which features are more important. In any machine learning problem, it is essential to write your observation in plain English.

Coming up next: PDF(probability density function), CDF(cumulative density function)