**Advanced Machine Learning**

**CIS550**

**Spring ‘24**

Lab Homework 2

Submitted by :

Name : Manas Vishal (01971464)

Email : mvishal@umassd.edu

Title : Homework on exploring datasets and its analysis

Exploring the data

The goal is to explore the dataset and its classification. We also aim to study the data by their correlation values. The data was recorded by Dr. Henrique da Mota during a medical residence period in the Group of Applied Research in Orthopedics (GARO) of the Centre Médico-Chirurgical de Réadaptation des Massues, Lyon, France. The data has been organized in two different, but related, classification tasks.

The features of the data are:

* Pelvic incidence
* Pelvic tilt
* Lumbar lordosis angle
* Sacral slope
* Pelvic radius
* Grade of spondylolisthesis

Importing the data

In this part we first import the data and the modules needed to analyze the various features of it.



Fig1. Importing the data

Exploring the data

We explore the data in this section. We first load the data in arff and then see the shape and print out the column indentifiers.



Fig 2. Exploring the data

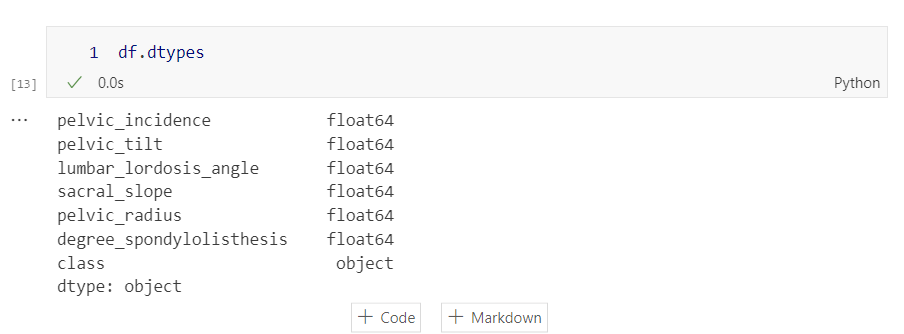


Fig 3. Datatypes

We see the data types of the associated column field.

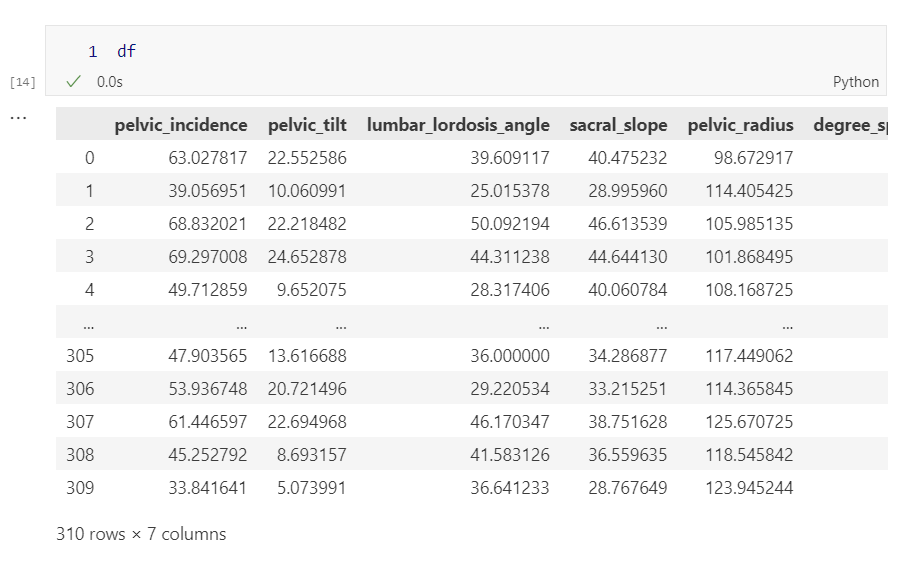


Fig. 4 The dataset

We print the dataset to screen.

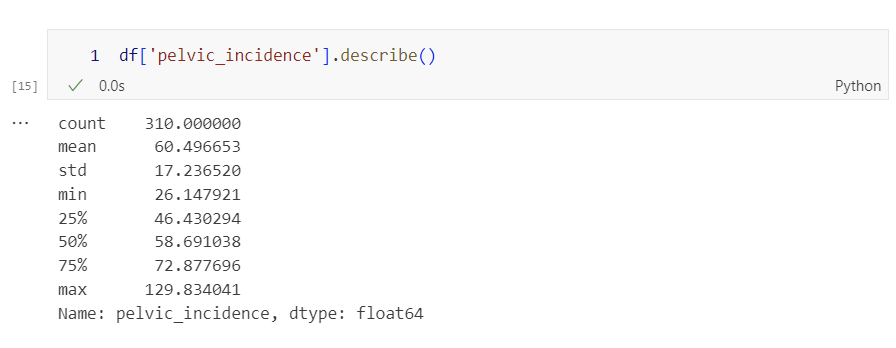
We can use the describe function to get a quick statistic on the full data or the subset of it. 

Fig. 5 Describe function on a Column

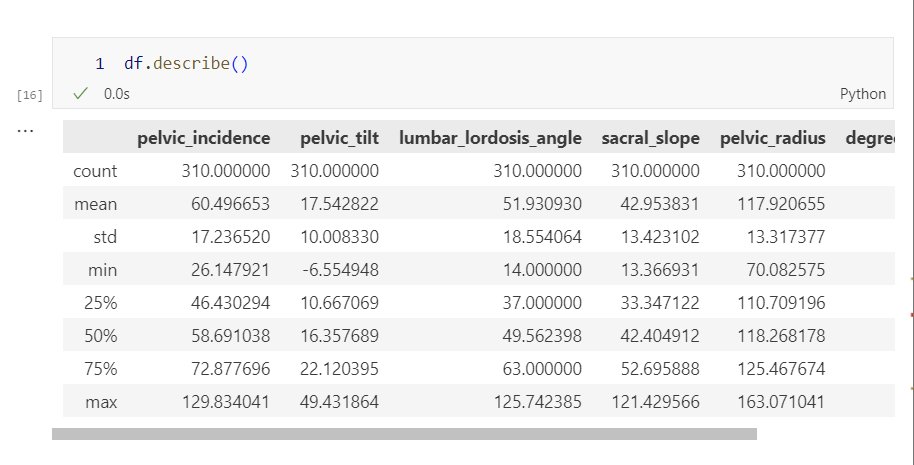


Fig. 6 Describe function on full data

Since we don’t see any anomaly by eye hence we now rely on the metrics and plots.



Fig. 7 Plot of data

We can also use the density plot to visualize the distribution of each column/field/feature.

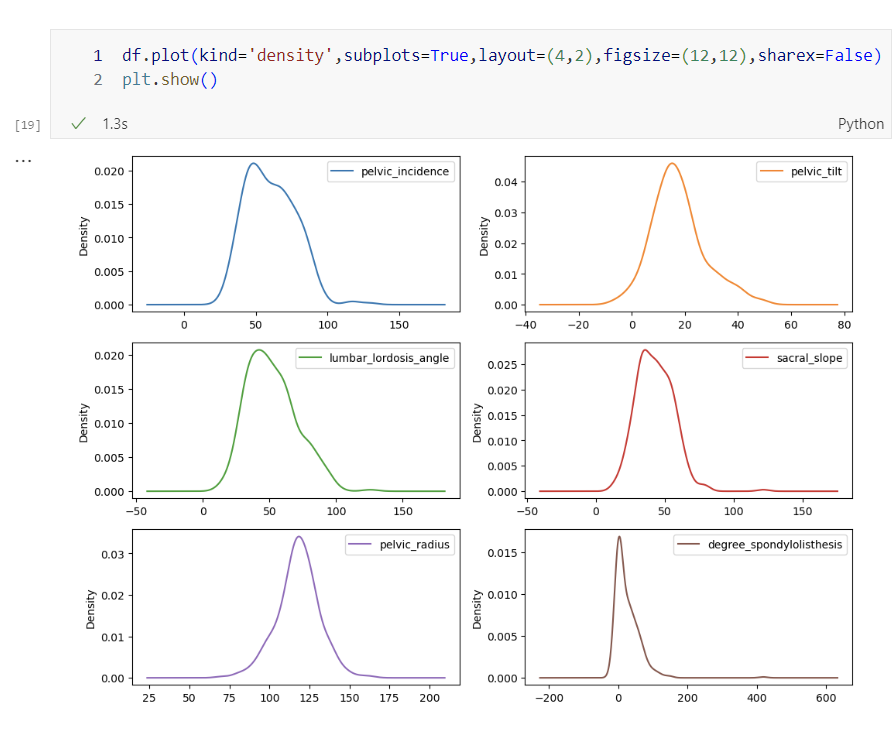


Fig. 8 Density plot of data

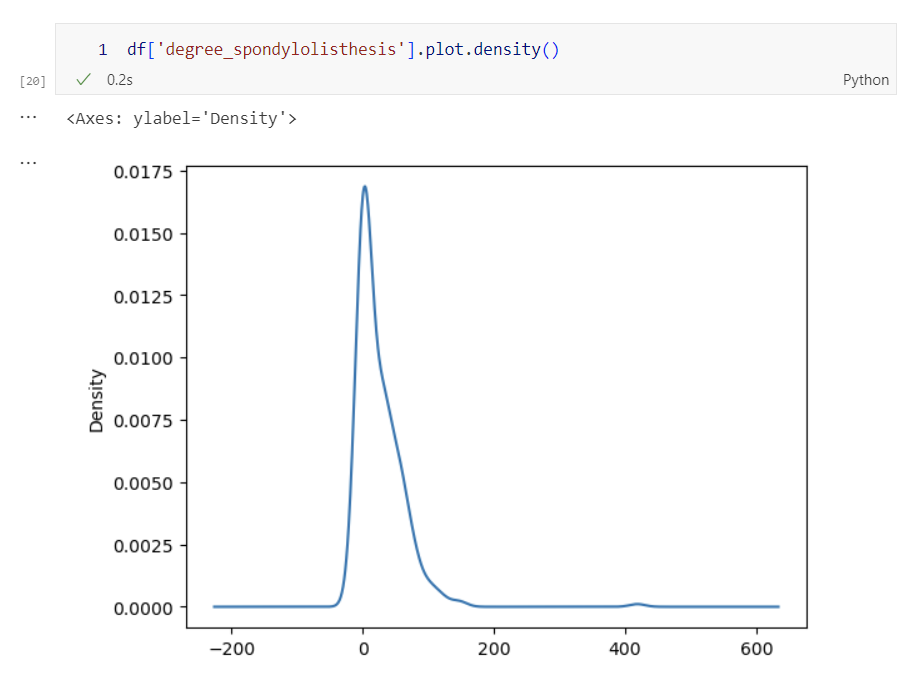


Fig. 9 Density plot of the Spondylolisthesis feature

We also plot histogram to see the skewness of the distribution

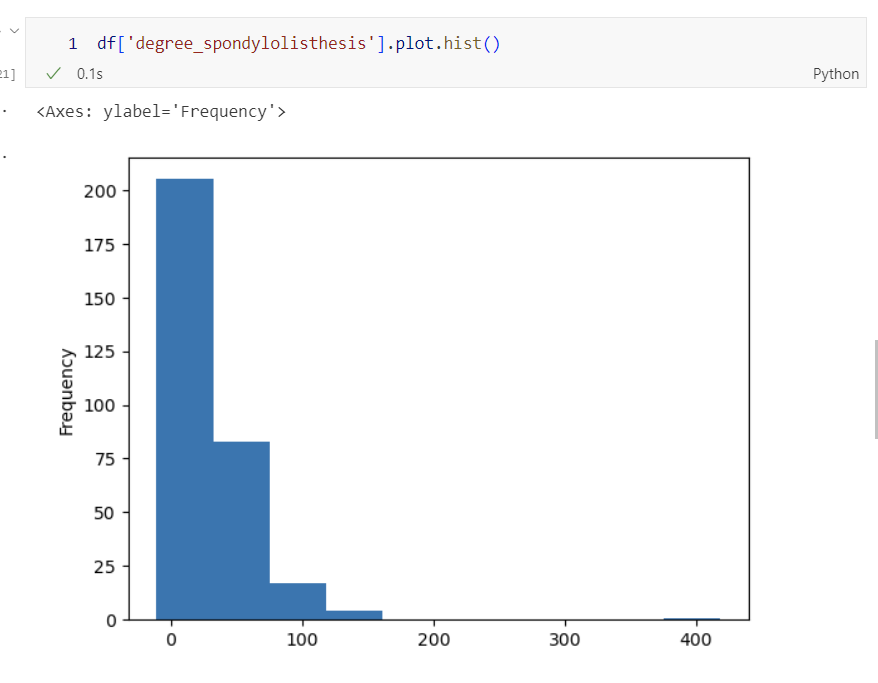


Fig. 10 Histogram plot of the feature

We then use the whisker box plot to see any outlier(s). From the box plot we can see that there seems to be quite a big value near 400.

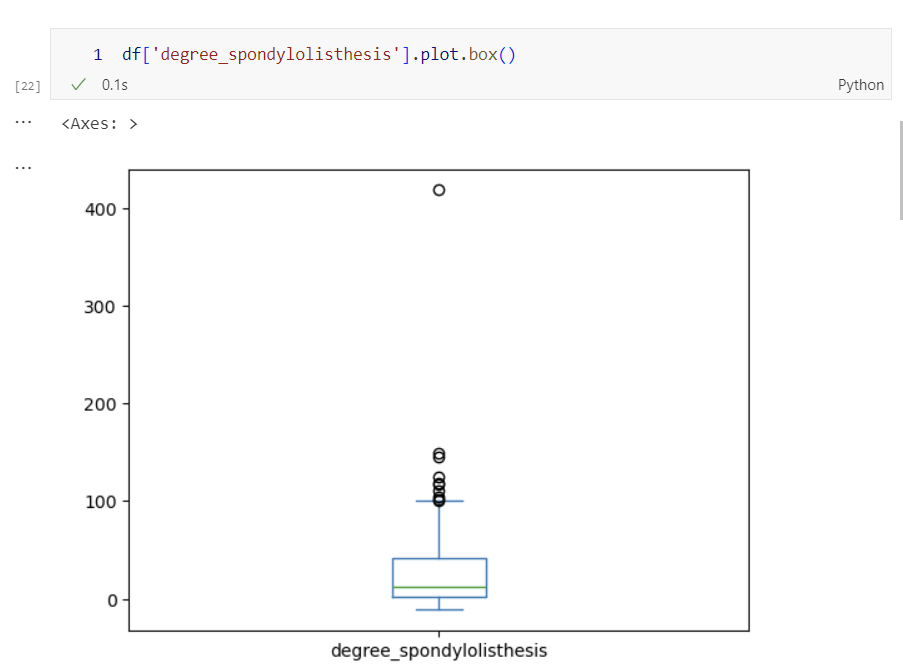


Fig. 11 Box plot of data

Analyzing the target

We check the number of abnormal and normal values. We cannot use strings for analysis, hence we convert abnormal to 1 and normal to 0.

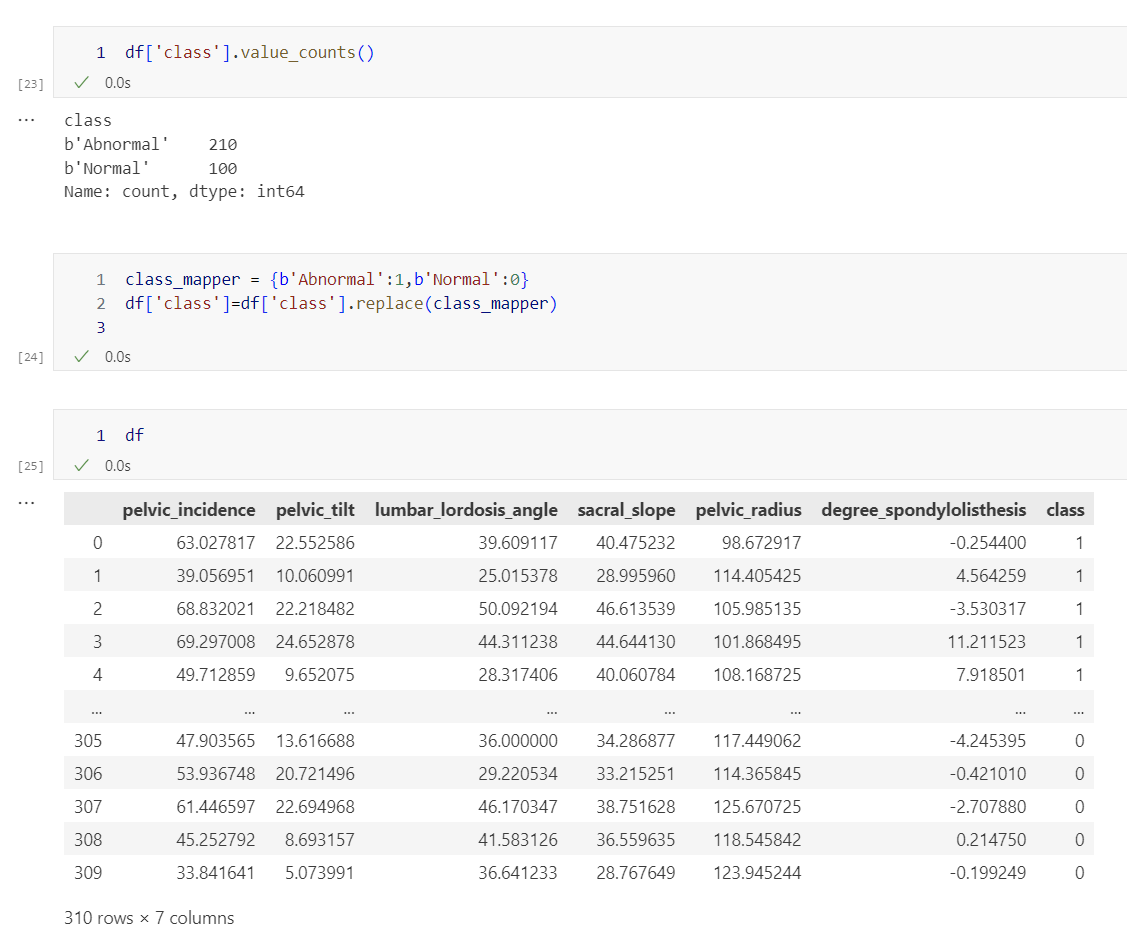


Fig. 12 Updated class values according to number

We then make a scatter plot of the data according to class map we did earlier

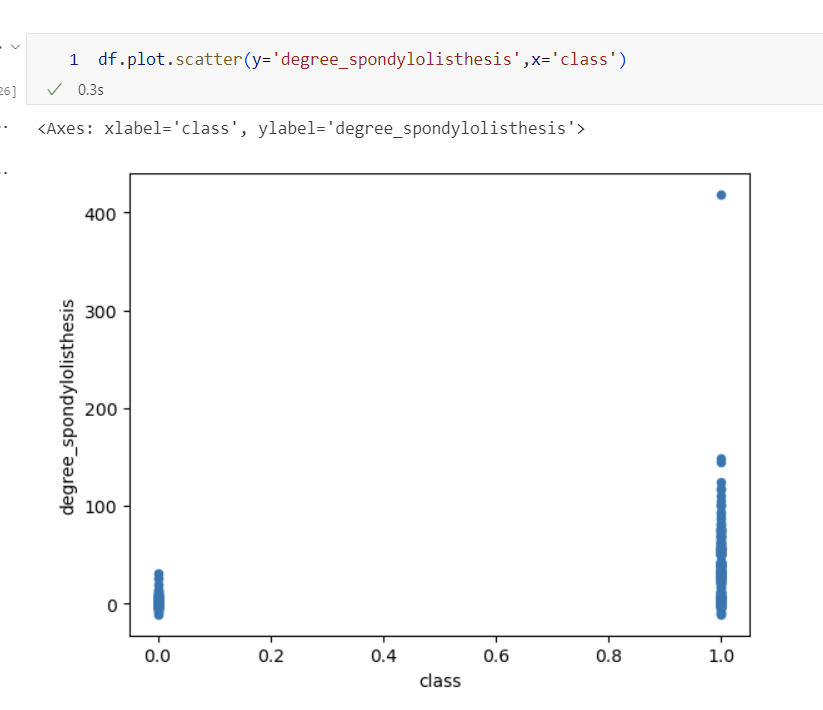


Fig. 13 Scatter Plot of abnormal and normal class

(Challenge Task)

For other classes we do the same by just changing the string

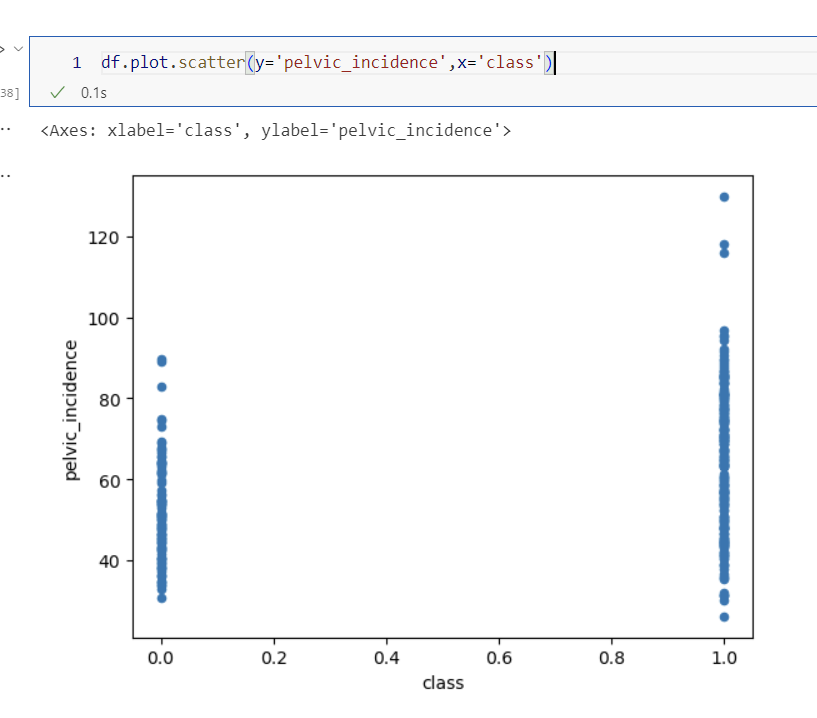


Fig. 14 Scatter plot for different class

To analyze multiple features at once we use the boxplot

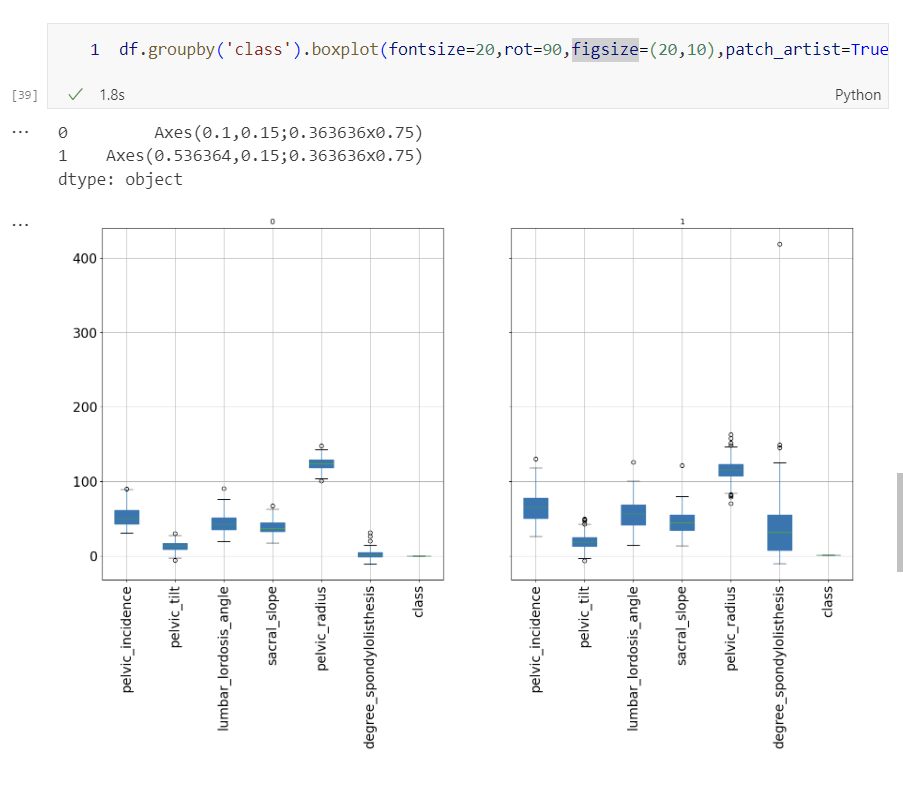


Fig. 15 Boxplot for all variable

.corr function gives us the correlation value beetween each features

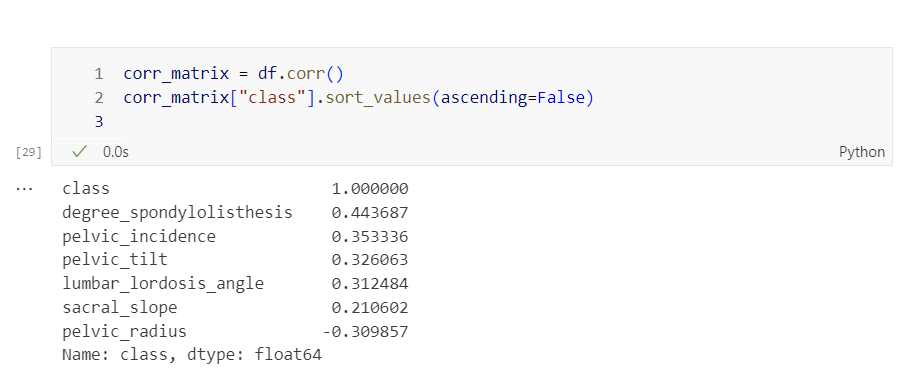


Fig. 16 Correlation matrix and values

We make a scatter plot to visualize the correlation matrix and then followup with heatmap of correlation matrix.

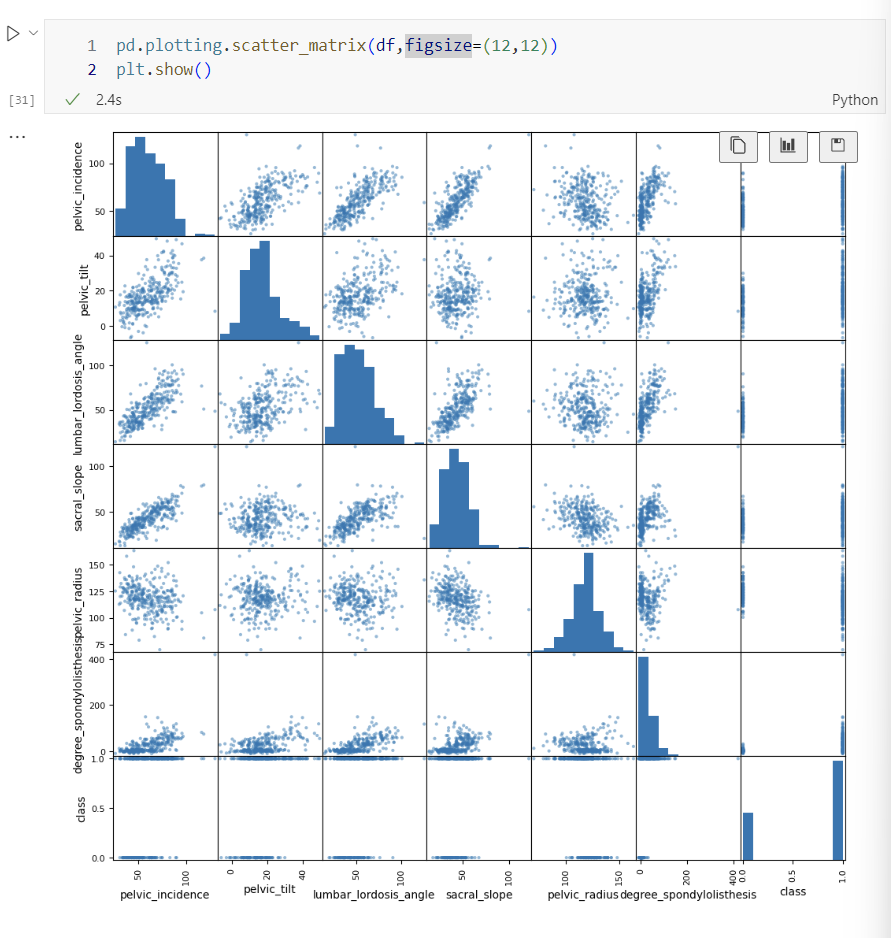


Fig. 17 Scatterplot for correlation matrix

My laptop did not have seaborn installed so I install seaborn in my conda environment with the conda install command.



Fig. 18 Seaborn installation

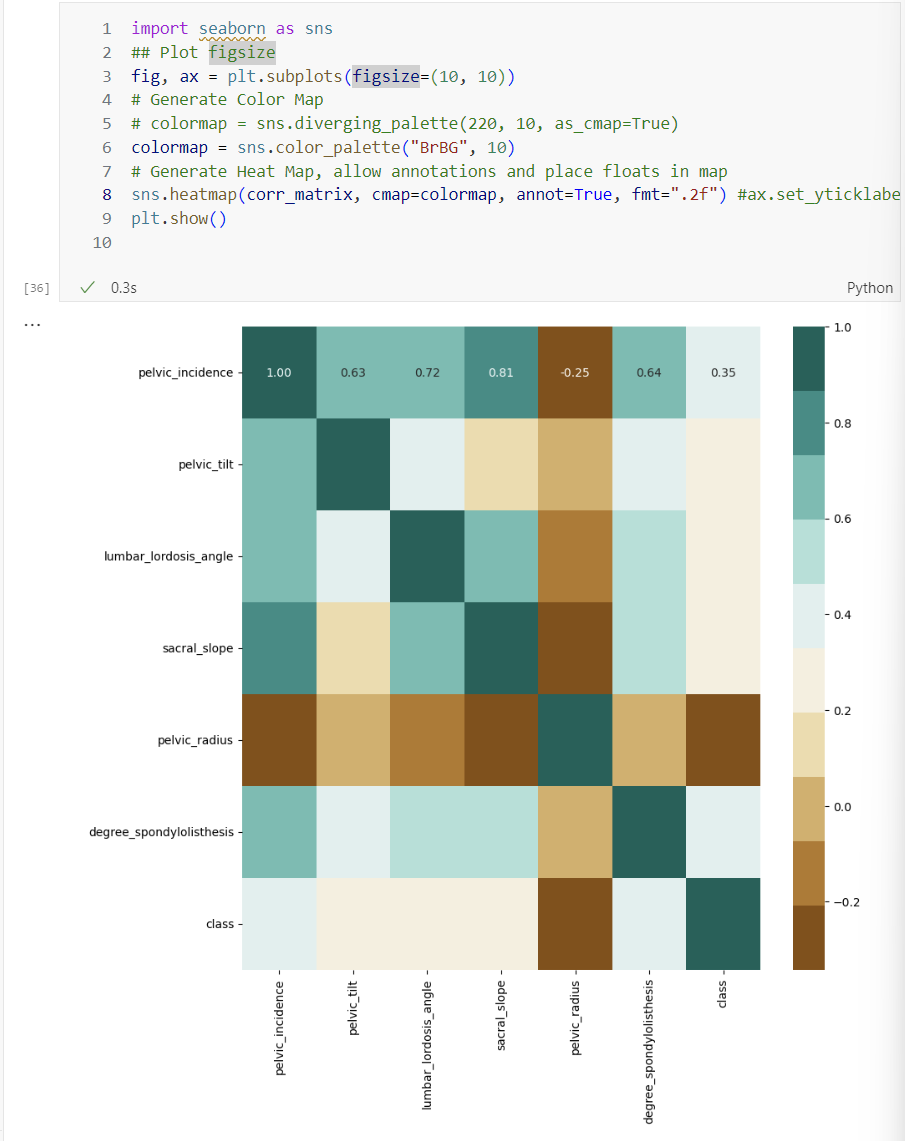


Fig. 19 Heatmap of correlation matrix

(Challenge Task)

Working on a different dataset.

I chose the Rice dataset [Rice (Cammeo and Osmancik)] from [Rice (Cammeo and Osmancik) - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik)

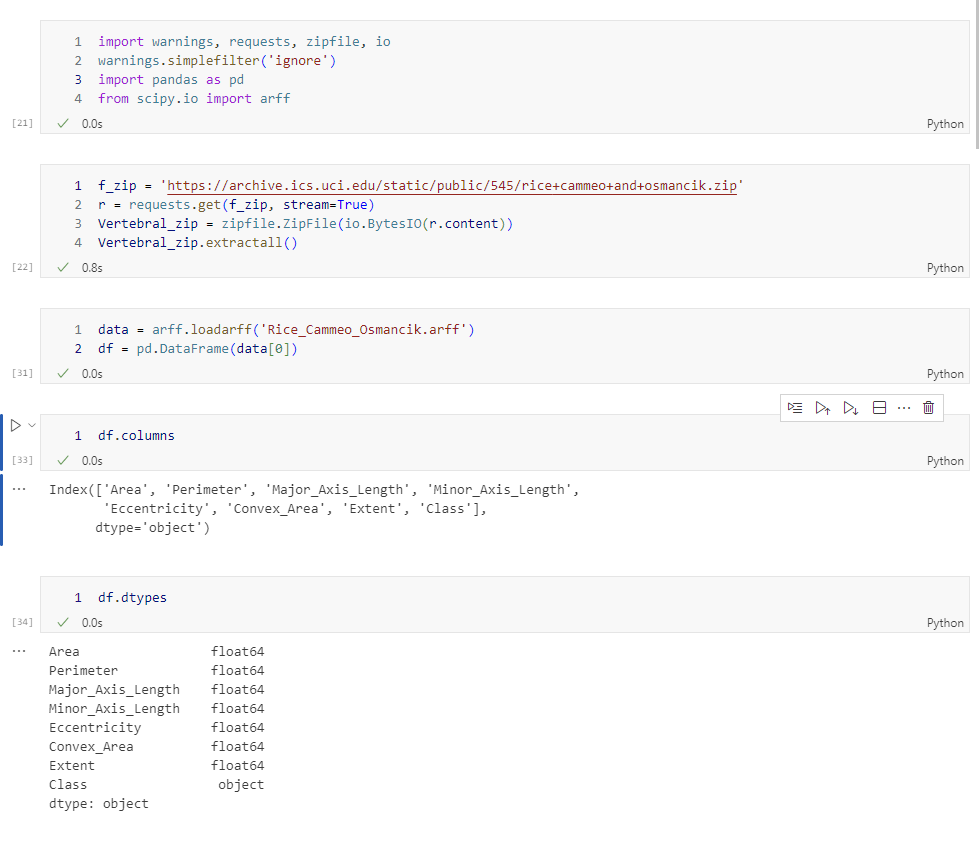


Fig. 20 Importing New Dataset

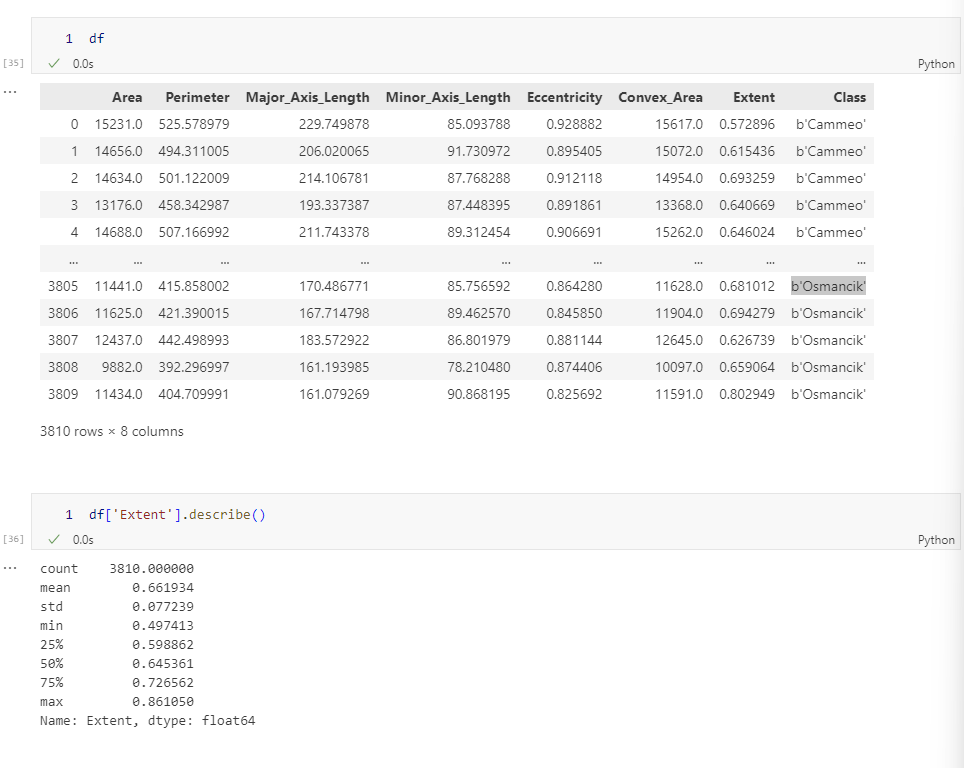


Fig. 21 Printing the dataset on screen

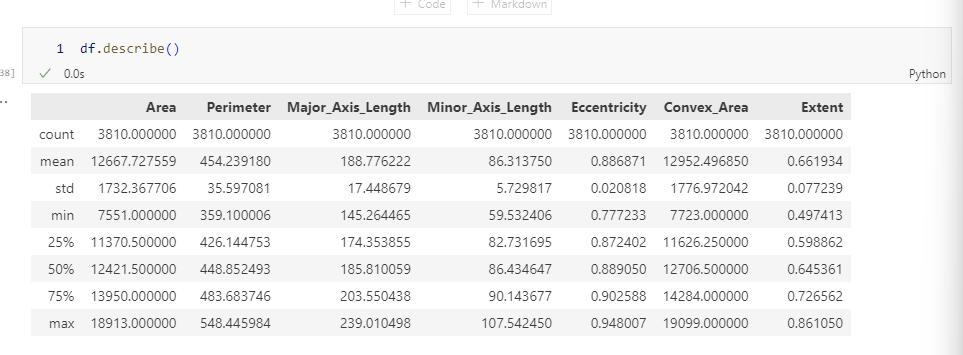


Fig. 22 Find a simple statistic of the data

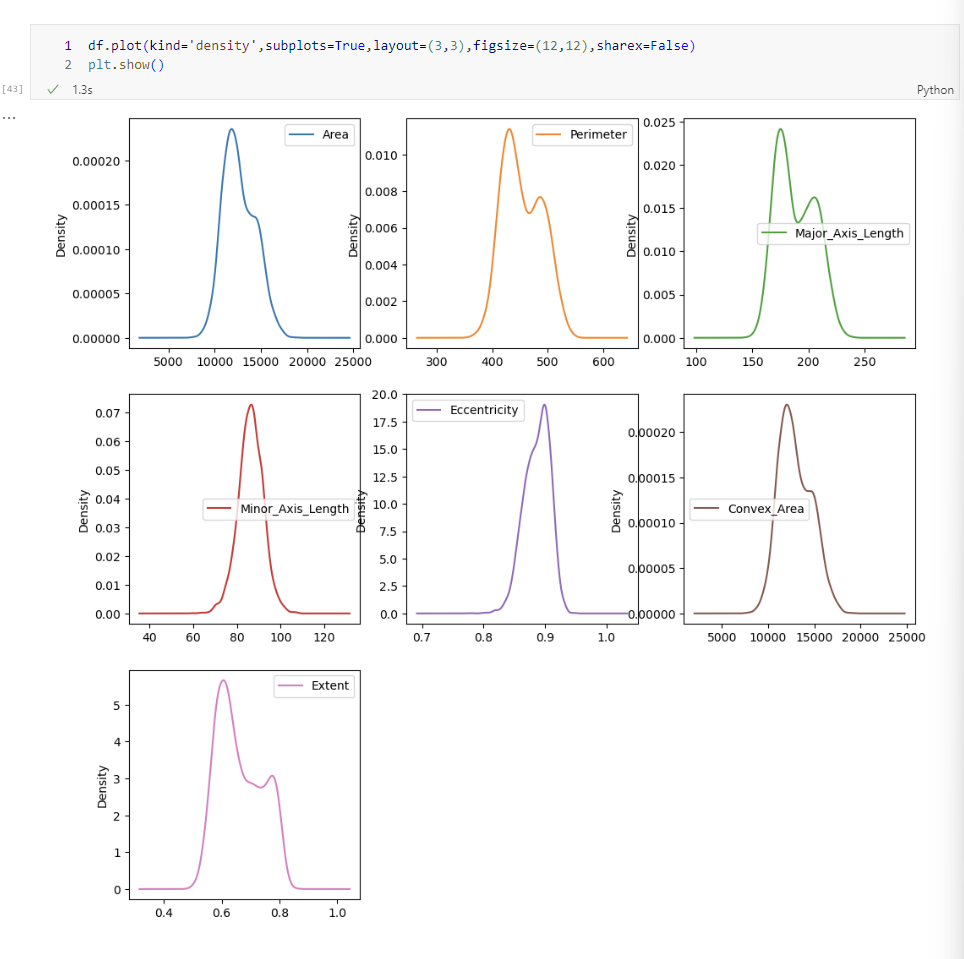
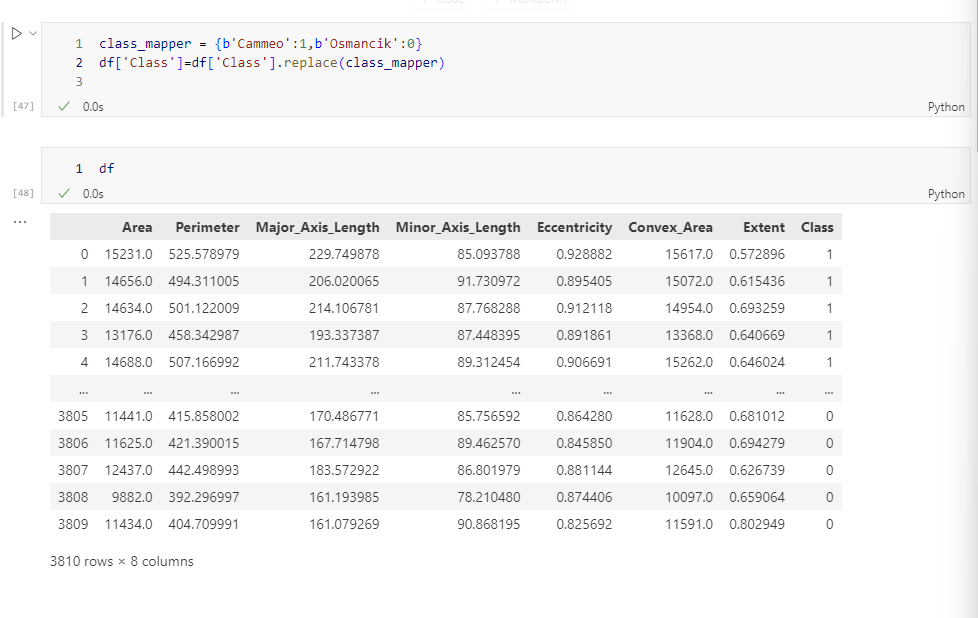
We now visualize the dataset by plotting the density plot

Fig. 23 Density plot of the data

I chose class by using Cammeo as 1 Osmancik as 0

Fig. 24 Replacing with target class

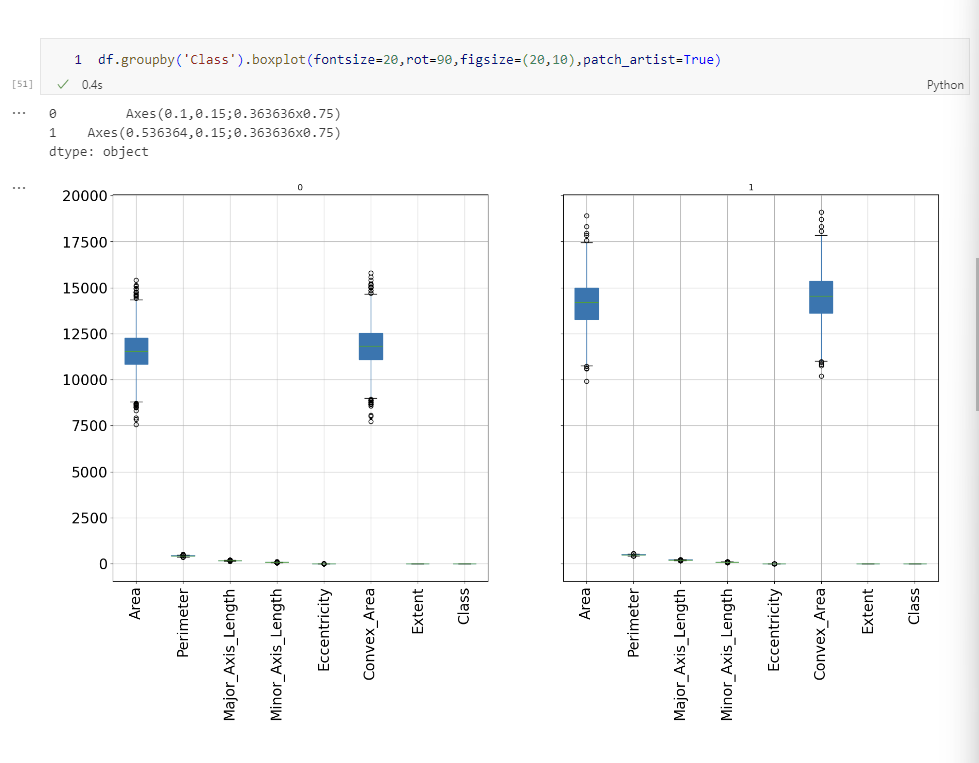


Fig. 25 Boxplots sorted by class

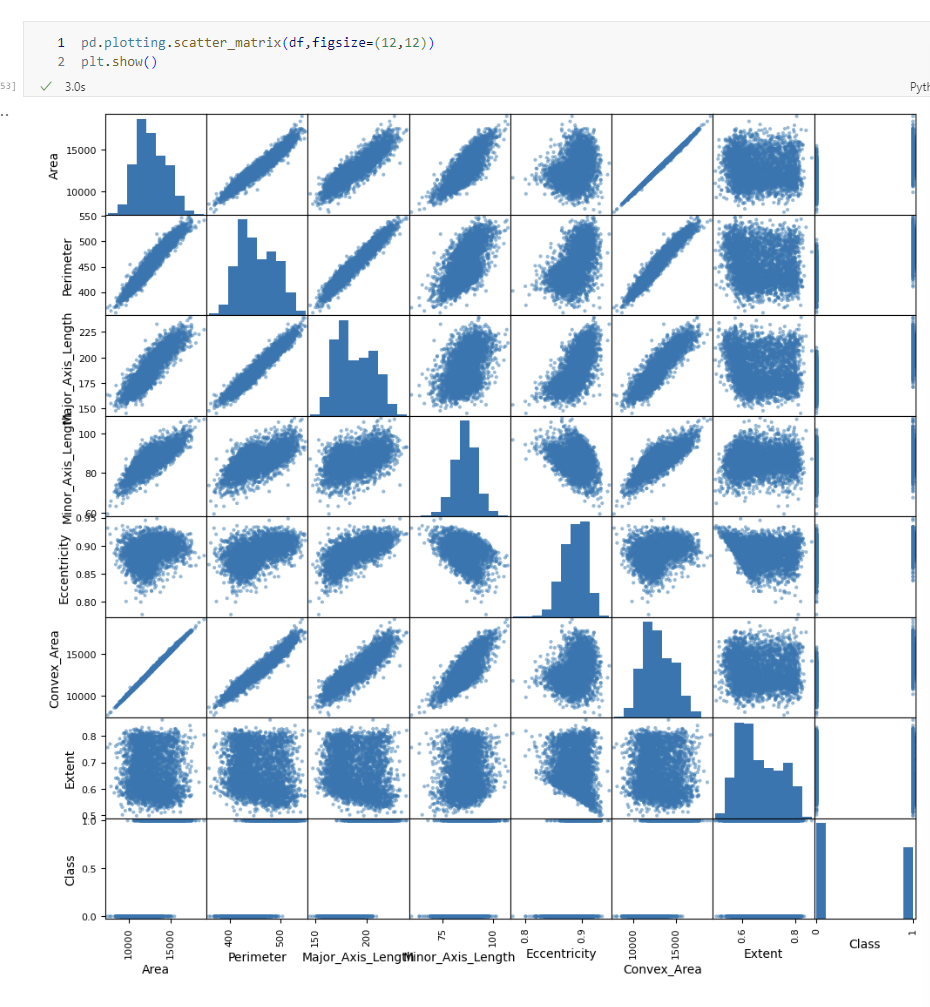
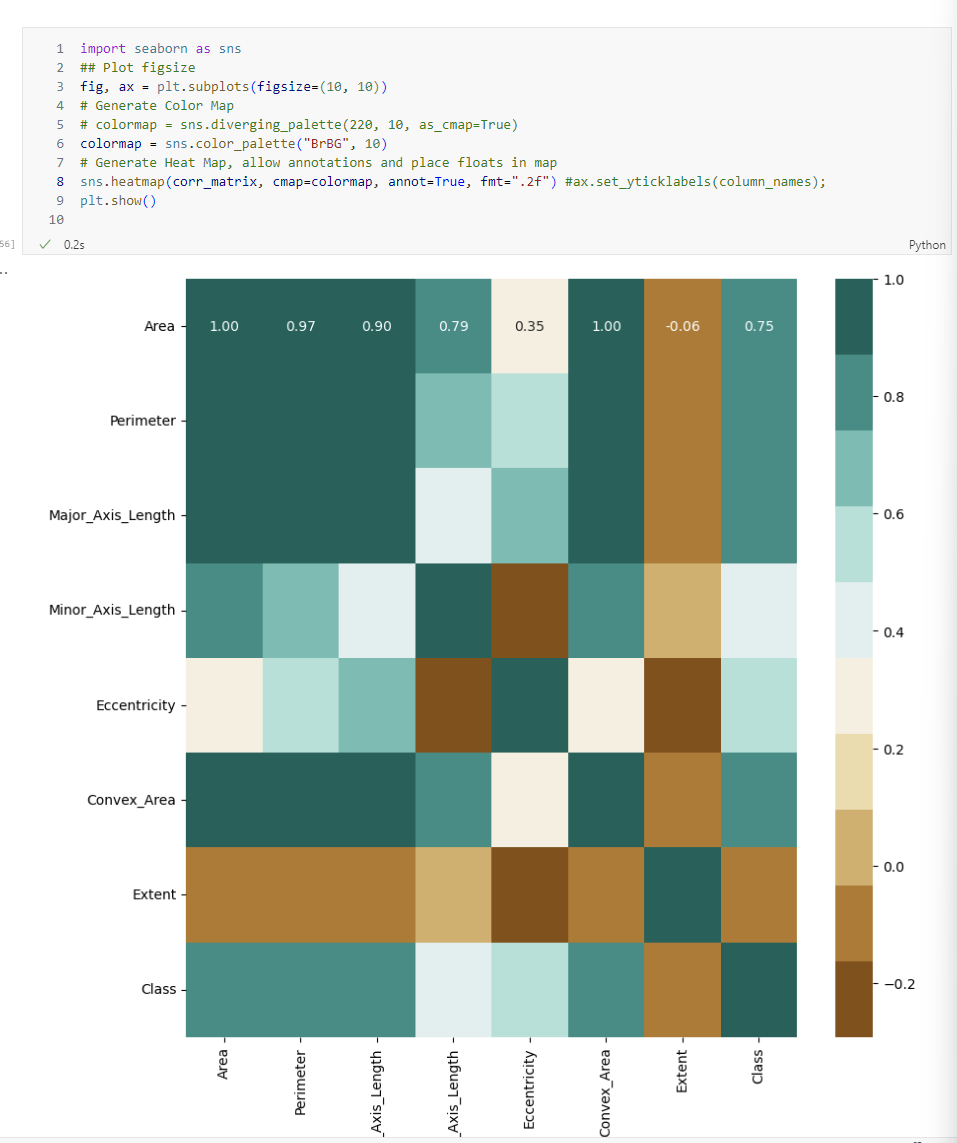


Fig. 26 Scatter Matrix of the new data set

Fig. 27 Heat Map of the new correlation matrix

Conclusion

From this homework, we learn a lot of new techniques in data analysis and visualization. To start with importing and exploring the dataset, we first learnt how to download data in python using the pandas module. We learnt how to visualize the several features by using various statistical tools and techniques. Going forward in this machine learning class, these tools are going to be very helpful in determining the model chosen for the dataset. I look forward to next lectures and assignments.