**FOUNDATIONS OF DATA SCIENCE**

**LAB ASSIGNMENT - 2**

**EXPLORATORY DATA ANALYSIS**

Download the Iris dataset from the given link - <https://www.kaggle.com/uciml/iris/version/2>

The Iris dataset was used in R.A. Fisher's classic 1936 paper, [The Use of Multiple Measurements in Taxonomic Problems](http://rcs.chemometrics.ru/Tutorials/classification/Fisher.pdf), and can also be found on the [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/).

It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

The columns in this dataset are:

* Id
* SepalLengthCm
* SepalWidthCm
* PetalLengthCm
* PetalWidthCm
* Species

After downloading the .csv file, load its contents into a dataframe on a Python notebook and perform preprocessing techniques as done in Lab 1, ie, check for and handle null values if any. Now, perform the following operations:

1. Print the head of the dataframe to see what the data looks like.
2. Check all the types of variables that you are dealing with and make sure they are all of the appropriate types so that future mathematical operations to be performed on the data are seamless.
3. Find how many flowers of each type of species are present in the dataframe.
4. Compute basic statistics for all continuous variables in the dataframe using the describe function.
5. Find Pearson’s correlation of the dataframe and plot correlation matrix using the seaborn library. Also, describe your interpretation of the results.
6. Find all the unique values of lengths in ‘PetalWidthCm’ and comment on the range of lengths in the column.
7. Copy columns 'PetalWidthCm' and 'PetalLengthCm' into another dataframe. Then, find the average ‘PetalLengthCm’ for each unique value of ‘PetalWidthCm’.
8. Plots a scatter plot using the regplot function from the seaborn python library and check if the following variables have strong or weak linear relationships, giving your comment on the correlation between the variables based on the plot:
   1. ‘SepalLengthCm’ and ‘PetalLengthCm’
   2. ‘SepalLengthCm’ and ‘SepalWidthCm’

Also, find their p values and Pearson’s coefficient and comment on the variables’ relationship with each other and significance, corroborating your earlier inferences based on the plot.

1. Plot every variable in the dataframe against every other variable using pairplots.
2. Create another dataframe with only ‘PetalLengthCm’ and ‘Species’ columns and group by the latter column. Use the ANOVA test to find if there is a significant difference among the mean petal length of different species of Iris flowers. Comment on the F-score and the P-score generated.