

## 1. Data Wrangling

Perform the following operations using Python on any open source dataset (e.g., data.csv)

1. Import all the required PythonLibraries.
2. Locate an open source data from the web (e.g. <https://www.kaggle.com>). Provide a clear description of the data and its source (i.e., URL of the website).
3. Load the Dataset into pandas dataframe.
4. Data Preprocessing: check for missing values in the data using pandas `isnull()`, `describe()` function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the dataframe.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper typeconversions.
6. Turn categorical variables into quantitative variables inPython.

In addition to the codes and outputs, explain every operation that you do in the above steps and explain everything that you do to import/read/scrape the data set.

## 2.Data Wrangling

Create an “Academic performance” dataset of students and perform the following operations using Python.

1. Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal withthem.
2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal withthem.
3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normaldistribution.

Reason and document your approach properly.

## 3. Descriptive Statistics - Measures of Central Tendency and variability

Perform the following operations on any open source dataset (e.g., data.csv)

Provide summary statistics (mean, median, minimum, maximum, standard deviation) for a dataset (age, income etc.) with numeric variables grouped by one of the qualitative (categorical) variable. For example, if your categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups. Create a list that contains a numeric value for each response to the categoricalvariable.

Provide the codes with outputs and explain everything that you do in this step.

4. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris- versicolor' of iris.csvdataset. Provide the codes with outputs and explain everything that you do in this step.

#### **5. Data Analytics :**

Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset (<https://www.kaggle.com/c/boston-housing>). The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variables in this dataset.

The objective is to predict the value of prices of the house using the given features.

#### **6. Data Analytics :**

Implement logistic regression using Python/R to perform classification on Social\_Network\_Ads.csv dataset. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

#### **7. Data Analytics**

Implement Simple Naïve Bayes classification algorithm using Python/R on iris.csv dataset. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

#### **8. Text Analytics**

1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.
2. Create representation of document by calculating Term Frequency and Inverse Document Frequency.

## 9. Data Visualization

1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see if we can find any patterns in the data.
2. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.

## 10. Data Visualization

1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age')

Write observations on the inference from the above statistics.

## 11. Data Visualization

Download the Iris flower dataset or any other dataset into a DataFrame.  
(e.g., <https://archive.ics.uci.edu/ml/datasets/Iris> ). Scan the dataset and give the inferences as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.
2. Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Create a box plot for each feature in the dataset.

Compare distributions and identify outliers.

12. Write a simple program in SCALA using Apache Spark framework.

13. Write a code in JAVA for a Simple WordCount application that counts the number of occurrences of each word in a given input set using the Hadoop MapReduce framework on local-standalone set-up.

14. Design a distributed application using MapReduce which processes a log file of a system.

15. Locate dataset (e.g Sample\_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.