Problem Statement 1:
Perform the following operations using Python on any open source dataset (e.g., data.csv)
1. Import all the required Python Libraries.
2. Provide a clear description of the data and its source (i.e., URL of the web site).
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 data frame.
5.Turn categorical variables into quantitative variables in Python using any of the technique.

Problem Statement 2:
Perform the following operations using Python on any open source dataset (e.g., data.csv)
1.Provide a clear description of the data and its source (i.e., URL of the web site).
2. Load the Dataset into pandas dataframe.
3. 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 data frame.
4. 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 type conversions.

Problem Statement 3:
Perform the following operations using Python on "Academic performance" dataset
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 with them.
2. Scan all numeric variables for outliers.
3. Identify outliers using any of the identification techniques and remove them.

Problem Statement 4:
Perform the following operations using Python on "Academic performance" dataset
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 with them.
2. 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 normal distribution.
3. Identify the skewness using any technique and recover from skewness.

Problem Statement 5:

Perform the following operations on any open source dataset

1. 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.)
2. Create a list that contains a numeric value for each response to the categorical variable.
3. 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.csv dataset.

Problem Statement 6:

Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset. 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.
Compute MSE.

Problem Statement 7:

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

Problem Statement 8:

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

Problem Statement 9:

1. Create any 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.

Problem Statement 10:

1. Use the inbuilt dataset 'titanic'.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, distplot,histplot,jointplot,rug plot,barplot.

Problem Statement 11:

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') using boxplot and violinplot.

Problem Statement 12:

Use the Iris flower dataset and perform the following operations. Scan the dataset and give the inference 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 boxplot for each feature in the dataset.
4. Show the demonstration of the subplot.

Problem Statement 13:

Use the Iris flower dataset and perform the following operations. Scan the dataset and give the inference 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. Compare distributions and identify outliers.
4. Show the demonstration of the subplot.

Problem Statement 14:

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 year wise/month-wise.

Problem Statement 15:

Write a simple program in SCALA using Apache Spark framework to perform 4 arithmetic operations.