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. Load the Dataset into pandas' data frame.

3. Data Preprocessing: check for missing values in the data using pandas isnull (), describe(),head(),tail(), info(),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 2:
Perform the following operations using Python on any open-source dataset (e.g., data.csv)

1. Import all the required Python Libraries.

2. Load the Dataset into pandas' data frame.

3. Data Preprocessing: check for missing values in the data using pandas isnull (),head(),tail(), info() describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.

4. Turn categorical variables into quantitative variables in Python.

Problem statement 3:
Perform the following operations using Python

1. Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use the following techniques to deal with them.

a)Delete rows or column b)replace missing values with mean c)replace missing values with mode d)replace missing values with median

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.

Problem statement 4:

Perform the following operations using Python

1. Scan all numeric variables for outliers. If there are outliers, use the following techniques to deal with them. a)min max normalization b)z-score normalization c)Box plot

Problem statement 5:

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

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. Create a list that contains a numeric value for each response to the categorical variable.

Problem statement 6:

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- verginica' of iris.csv dataset. Grouped by one of the qualitative (categorical) variable from the dataset

Problem statement 7:

Create a Linear Regression Model using Python/R to predict home prices.The objective is to predict the value of prices of the house using the given features in the dataset.

Problem statement 8:

1. Implement logisticregression using Python /R to perform classification on a given dataset.

2. Compute Confusion Matrix of findTP,FP,TN,FN,Accuracy, Error Rate, Precision,Recall on the given dataset

Problem statement 9:

1. Implement naive bayes classification algorithm using Python /R to perform classification on a given dataset.

2. Compute Confusion Matrix of findTP,FP,TN,FN,Accuracy, Error Rate, Precision,Recall on the given dataset

Problem statement 10:

1. Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.

2. Create representation of documents by calculating Term Frequency and Inverse Document Frequency.

Problem statement 11:

1. Use the inbuilt dataset 'titanic' as used . 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')

2. Write observations on the inference from the above statistics.

Problem statement 12:

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 box plot for each feature in the dataset.
4. Compare distributions and identify outliers.

Problem statement 13:

1.Which configuration files are required for setting the hadoop environment (*mapred-site.xml*, *core-site.xml*, *hdfs-site.xml*)

2.What are the steps involved in setting up a Hadoop cluster ?

3. Draw a hadoop architecture

4.Use hadoop cluster to load a log file or text file (BigData) and execute word count example application that counts the number of occurrences of each word in a given input set using the Hadoop Map-Reduce framework on local-standalone set-up.

Problem statement 14:

1.Which configuration files are required for setting the hadoop environment (*mapred-site.xml*, *core-site.xml*, *hdfs-site.xml*)

2.What are the steps involved in setting up a Hadoop cluster ?

3. Draw a hadoop architecture

4.Use various HDFS commands to load the input number file(Big Data) and Run the sorting example on it

Problem statement 15:

1. Write a simple word count program in SCALA using Apache Spark Framework