KJ's Educational Institutes

Trinity Academy of Engineering, Pune

Department of Computer Engineering

LABORATORY MANUAL

S.E. (Computer Engineering)

DATA SCIENCE AND BIG DATA ANALYTICS

Staff: Mrs. Nilufar Zaman

(2019 course)

Teaching Scheme:	Examination Scheme:	Credit:
Practical: 4Hrs/ Week/Batch	**Term Work: 25 Marks**	**Practical: 4**

DATA SCIENCE AND BIG DATA ANALYTICS

List of Assignments

1. Data Wrangling, I

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

1. Import all the required Python Libraries.
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 web site).
3. Load the Dataset into pandas data frame.
4. Data Preprocessing: check for missing values in the data using pandas insult(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
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 type conversions.
6. Turn categorical variables into quantitative variables in Python.

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 II

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 with them.
2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.
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 normal distribution.

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)

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.
2. 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.

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

4. Data Analytics I

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.

5. Data Analytics II

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.

6. Data Analytics III

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,
3. Recall on the given dataset.

7. 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.

8. Data Visualization I

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.

9. Data Visualization II

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.

10. 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 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.

Compare distributions and identify outliers.

11. Locate dataset (e.g., sample_weather.txt) for working on weather data which reads the textinput files and finds average for temperature, dew point and wind speed.

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

13. Use the following dataset and classify tweets into positive and negative tweets.
https://www.kaggle.com/ruchi798/data-science-tweets

14 . Develop a movie recommendation model using the scikit-learn library in python.
Refer dataset:
https://github.com/rashida048/Some-NLP-Projects/blob/master/movie_dataset.csv

Assignment 03:

Title of the Assignment: Descriptive Statistics - Measures of Central Tendency and variability
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) variables. 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.
2. 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.

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

Objective of the Assignment: Students should be able to perform the Statistical operations using Python on any open source dataset.

Prerequisite:
1. Basic of Python Programming
2. Concept of statistics such as mean, median, minimum, maximum, standard deviation etc.

Theory:

Terminologies associated with statistics

- **Population:** It is an entire pool of data from where a statistical sample is extracted. It can be visualized as a complete data set of items that are similar in nature.
- **Sample:** It is a subset of the population, i.e. it is an integral part of the population that has been collected for analysis.
- **Variable:** A value whose characteristics such as quantity can be measured, it can also be addressed as a data point, or a data item.
- **Distribution:** The sample data that is spread over a specific range of values. ·
- **Parameter:** It is a value that is used to describe the attributes of a complete data set (also known as „population"). Example: Average, Percentage
- **Quantitative analysis:** It deals with specific characteristics of data- summarizing some part of data, such as its mean, variance, and so on.
- **Qualitative analysis:** This deals with generic information about the type of data, and how clean or structured it is.

Descriptive statistics:

- **Inferential statistics:** It deals with drawing inferences/conclusions on the sample data set which is obtained from the population (entire data set) based on the relationship identified between data points in the data set. It helps in generalizing the relationship to the entire dataset. It is important to remember that the dataset drawn from the population is relevant and represents the population accurately.

- **Regression:** The term „regression“ which is a part of statistics and machine learning, talks about how data can be fit to a line, and how every point from the straight line gives some insights. In terms of machine learning, it can be understood as tasks that can be solved without explicitly being programmed. They discuss how a line can be fit to a given set of data points, and how it can be further extrapolated for the predictions to be done.
- **Maximum likelihood**: It is a method that helps in finding values of parameters for a specific model. The values of the parameters have to be such that the likelihood of the predictions that occur have to be maximum in comparison to the data values that were actually observed. This means the difference between the actual and predicted value has to be less, thereby reducing the error and increasing the accuracy of the predictions.

Measures of Central Tendency:

- When we work with numerical data, it seems apparent that in most set of data there is a tendency for the observed values to group themselves about some interior values; some central values seem to be the characteristics of the data. This phenomenon is referred to as central tendency.

Arithmetic Mean:

- The mean or average is the most popular and well known measure of central tendency.

Median

- Median represents the middle value for any group.
- It is the point at which half the data is more and half the data is less.

Measures of dispersion

- The measures of central tendency are not adequate to describe data. Two data sets can have the same mean but they can be entirely different. Thus to describe data, one needs to know the extent of variability. This is given by the measures of dispersion.
- The measure of dispersion shows the scatterings of the data.

Range

- Range is calculated by subtracting the minimum value in the data set from the maximum value.
- Range = (Maximum Value - Minimum Value)

Mid Range

- Midrange in layman terms is the middle of any data set or the simply the average, mean of the data.
- Midrange = (Maximum Value + Minimum Value) / 2

Quartile Deviation

- The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The second quartile, (Q2) is the median of the data set. The third quartile, (Q3) is the middle number between the median and the largest number.
- Quartile deviation or semi-inter-quartile deviation is

$$Q = ½ \times (Q3 – Q1)$$

Mean Deviation

- Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency.

Variance

- A variance measures the degree of spread (dispersion) in a variable‟s values. · Theoretically, a population variance is the average squared difference between a variable‟s values and the mean for that variable.

Standard Deviation

- Standard deviation is a squared root of the variance to get original values.

Covariance

- Covariance is a measure of how much two random variables vary together. It‟s similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

Assignment Questions:

1. Explain Measures of Central Tendency with examples.
2. What are the different types of variables? Explain with examples.
3. Which method is used to display statistics of the data frame? write the code.