**Data Wrangling, I**

Perform the following operations using Python on dataset(e.g., employee.csv)
1.Import all the required Python Libraries. And Load the Dataset into pandas data frame.
2. Data Preprocessing: find the missing values in the data columnwise and display statistical information.
3. Provide variable descriptions. Types of variables etc.Check the dimensions of the data frame
4. Data Formatting 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.
5. Data Normalization:Perform Z-Score transformation and plot box plot for any column
6.Turn categorical variables into quantitative variables in Python.

**Data Wrangling, I**

Perform the following operations using Python on dataset(e.g., student.csv)
1.Import all the required Python Libraries. And Load the Dataset into pandas data frame.
2. Data Preprocessing: find the missing values in the data and display statistical information.
3. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame
4. Data Normalization: Perform min max normalization and plot box plot for any column
5.Turn categorical variables for PG column into quantitative variables in Python.

**Data Wrangling II operations using Python.** (e.g., Academic_Performance.csv)

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(using mean and mode).Apply for single column and whole dataset.
2. Scan all numeric variables for outliers. If there are outliers,any of the suitable techniques to deal with them.(using z score)
3. Display and Remove the outliers
4. Apply data transformations on at least one of the variables **Create bins and Labels**.
5. Draw box plot

---

**Data Wrangling II operations using Python..**(e.g., Academic_Performance.csv)

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(using median and 0).Apply for single column and whole dataset.
2. Scan all numeric variables for outliers. If there are outliers,any of the suitable techniques to deal with them.(using IQR)
3.Display and Remove the outliers show q1 and q3
4. Apply aggregation function (max,avg). 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.
5. Draw Scatter plot

**Descriptive Statistics**–

Measures of Central Tendency and variability Perform the following operations on any open source dataset (e.g., employee_2.csv/data.csv) 1.Provide summary statistics (mean, median, minimum) for a dataset (age, salary etc.) with numeric variables

2.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.

3. Perform grouping on and display JOB_ID and it's count

4.Show data visualization for any column

**Descriptive Statistics**–

Measures of Central Tendency and variability Perform the following operations on any open source dataset (e.g., employee_2.csv/data.csv) 1.Provide summary statistics (**maximum, standard deviation, covaiance)** for a dataset (age, salary etc.) with numeric variables

2. 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.

3. Perform grouping on and display JOB_ID and it's count

4.Show data visualization for any column

**Data Analytics I:**

**https://www.youtube.com/watch?v=QcPycBZomac**

Show linear regression technique for user values.

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

**Data Analytics I:**

**https://www.youtube.com/watch?v=QcPycBZomac**

Show linear regression technique for user values.

Create a Linear Regression Model using Python/R to predict salary of 15 years of experience using salary_csv file.

| Experience | Salary |
|---|---|
| 5 | 20000 |
| 7 | 25000 |
| 9 | 40000 |
| 12 | 60000 |
| 18 | 80000 |
| 20 | 110000 |

**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.

**Data Analytics II**

**https://www.youtube.com/watch?v=UCOm-LFKX9E**

1. Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset.

| STUDY HOURS | PASS/FAIL |
|---|---|
| 29 | 0 |
| 15 | 0 |
| 33 | 1 |
| 48 | 1 |
| 39 | 1 |

2.Find the logistic regression for student if they study 25 and 42 hours

3. Compute Confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall on the given dataset.

**Data Analytics III**

**For naïve bayes https://www.youtube.com/watch?v=XzSlEA4ck2I**

For confusion matrix :https://www.youtube.com/watch?v=_CGTbkHwUHQ

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.

**Text Analytics**

https://www.youtube.com/watch?v=8F_ERPqN0T0

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.

**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.

**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')
2.Write observations on the inference from the above statistics.

**Data Visualization III**

Use the Iris flower dataset or any other dataset into a DataFrame. 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