1. Analyse the given dataset using python by applying following methods.
2. Use various types of **Data Pre-processing** steps.
3. Use different methods for **Analysing Data**.
4. Use various methods for **Visualizing Data**.
5. “Classification of 1994 Census Income Data”,

Build a model using python, which can predict if the income of any individual in the US is greater than or less than USD 50,000 based on the data available about that individual.

**Data Set Description:** This Census Income dataset was collected by Barry Becker in 1994 and given to the public site (use this url to download the dataset *http://archive.ics.uci.edu/ml/datasets/Census+Income)*. This data set will help you understand how the income of a person varies depending on various factors such as the education background, occupation, marital status, geography, age, number of working hours/week, etc.

Here’s a list of the independent or predictor variables used to predict whether an individual earns more than USD 50,000 or not:

* Age
* Work-class
* Final-weight
* Education
* Education-num (Number of years of education)
* Marital-status
* Occupation
* Relationship
* Race
* Sex
* Capital-gain
* Capital-loss
* Hours-per-week
* Native-country

The dependent variable is the “income-level” that represents the level of income. This is a categorical variable and thus it can only take two values:

1. <=50k
2. >=50k
3. Analyse the Aggregation and Data Wrangling steps using python for the given two Data frames.

* Data Wrangling

Data Frame 1

1. Emp\_id= 101,102,103,104,105,106
2. Dept\_Name=’CSE’,‘IT’,‘Mechanical’,‘Civil’,‘Biotechnology’,’ECE’
3. Aisle= 1,2,3, NaN,5,6

Data Frame 2

1. Emp\_id= 107,108,109,110,111,112
2. Dept\_Name=’BCA’, ‘BSC’, ‘MCA’, ‘MSc’, ‘Mtech’,’BBA’
3. Aisle= 7,8,9,10,NaN,12

* Aggregation

1. Create a Dataframe (8,4)
2. Date\_range should be 17-11-2019 and periods = 8
3. Based on your own data, write the python code environment for analysing various plots and features present in matplotlib.
4. Based on given description, write the python program and analyse the prediction.

Working with Twitter data has become an integral part of sentiment analysis problems. The dataset is 3MB in size and has 31,962 tweets.

**Problem:**Identify the tweets which are hate tweets and which are not.

1. Retail is another industry which extensively uses analytics to optimize business processes. Tasks like product placement, inventory management, customized offers, product bundling, etc. are being smartly handled using data science techniques. As the name suggests, this data comprises of transaction records of a sales store. This is a regression problem. The data has 8523 rows of 12 variables.

**Problem:** Predict the sales of a store.

1. Based on your own data, analyse the basic Statistics and Descriptive Statistics.
2. Based on your own data, elaborate Python Multiprocessing Module
3. Based on your own data, elaborate Python Data Cleansing by Pandas and Numpy
4. Using python, predict the rainfall using Linear Regression