Assignment NO :- 3

Title:-

Write Python code for that loads any data set & plot the graph.

Problem statement:-

Write python code that loads any dataset (from www.data.gov.in), & plot the graph.

Pre-Lab:-

A basic understanding of Computer Programming terminologies. A basic understanding of any of the programming languages will help in understanding the python programming & datascience concepts.

Theory :-

steps in Data science process. We have already seen a simple linear form of data science process including five distinct activities that depend on each others. We are sometime refer to it as geo-spacial query. Namely, explore data & pre-data process data. This is why this step is called explore. Once we know more than the data through exploratory analysis, the next step is preprocessing of data for analysis. Pre-programming includes cleaning data, sub-setting or filtering data, creating data, which programs can read & understand, such as multiple data sets involved this step also includes integration of multiple data source or streams. This steps can take a couple of iterations on its own or might require data scientiests to go back to steps one & two to get more data or package data

in a different way. step four for communicating results includes evaluation of results. presenting them in a visual way, creating reports that includes an evaluation on assessment of results with respect to success criteria. Activities in this step brings us back to the very first reason we do data science, the purpose you initially defined is what we refer to as the act step. We have now see all the steps in a typical data science process. Please note that this is an iterative process & findings from one step may required the previous step to be repeated with the new information.

1) step I. Acquiring Data:-

step one, acquiring data. The first step in the data sciene process is to aquire the data. You need to obtain the source material before analyzing or acting on it. The first step in acquiring data is to determine what data is available. Leaving out even a small amount of important data can lead to incorrect conclusions. Data comes from many place, local & remote, in many verieties, structured & un-structured. A lot of data exists in conventional relational database, like structure big data from organizations. The tools of choice to access data from databases is structured big data from organizations. Data can be also exist in fields such as text files & Excel spreadsheets. scripting languages are generally used to get data from filed. A scripting language is a high level programm

language that can be either general purpose or specialized for specific functions. Common scripting languages with support for processing files like javascript, Python, PHP, Perl, R & MATLAB & are many others. RESET stands for Representational state Transfer. And it is an approach to implementing web services with performance, scalablity & maintainability in mind. NoSQL storage systems are increasingly used to manage a variety of data types in big data. NOSQL data stores provide APIs to allow users to access data. These APIs can be used directly or in an application that needs to access the REST. Once, we start to listening to this services, we recieve weather station measurments as they occurs. This data is then processed & compared to patterns found by our models to determine if a whether station is experiencing santa Ana conditions. The combinations of sensor data & tweet sentiments helps to give us a sence of the urgency of the fire situation. As a summary, big data comes from many places. Finding & evaluating data useful to your big data analysis is important before you start aquiring data. Depending on the source & structure of the data, there are alternative ways to access it.

3) step 2-A: Exploring Data:—

step 2-A:— Exploring Data. After you've put together the data that you need for your application, you might be tempted to immediately build

models to analyze the data. Resist this temptation. Correlational graphs can be used to explore the dependencies betn different variables in the data. Graphing the general trends of variables will show you if their is consistent directions in which the values of these are moving towards, like sales prices going ups & down. Summary statistics are quantities that statistics provide numerical values to describe your data. Looking at these measurment will give you an idea of the nature of your data. A heap map, such as one show here, can quickly give you the idea of where the hotspots are. many other different types of graphs can be used. Histogram show that the distroubution of the data & can show skewness or unusual dispersion. Boxplots are another type of plot for showing data distribution. Line graphs are useful for how values in your data change over time. Spikes in the data are also easy to spot. Scatter plot can you show you correlan between two variables. Overall, their are many type of graph to visualize data. They are very useful to helping you understand the data you have. In summary, what you get by exploring your data is better understanding of the complexity of the data you have to work with. This, in turn, will guide the rest of your process.

step 2 - B: Pre-Processing Data:—

step 2: B:- The raw data that you get directly from your sources are never in the format that you

need to postform analysis on. There are two main goals in the data pre-processing step. Missing customer agent demographics studies. Invalid data like an invalid zip code for example, a six digit code. And outliers like a sense of failure causing values to be much bigger or lower for expected for a period of time. Since get the data downstream we usually have little control over how the data is collected. Perhaps is make sence to retain the never value wherever there's conflict for invalids values, the best estimation for a reasonable value can be used as a replacement. Raw data often has to be manupulating to be in correct formato to analysis. For example. from sample recording daily changes in stock prices, we many wants to capture change prices for a particular market segments like real estate or health care. This would required determining which stock is belongs to which market segments, grouping them together & perhaps computing the mean, range, standard deviation for each group. In summary, Data preperation is very important part in the data science proccess. In fact, this is the where you are spend most of your time on any data science effort. It can be tedious process, but it is a crucial step. Always remember, garbage in, garbage out. If you don't spend the time & effort to create good data for the analysis, you will not get good results no matter how sophisticated the analysis techniques you're using is.

4) Step 3: Analyzing Data:—

Now that you have your data nicely prepared. The next step is to analyze the data. The main categories of analysis techniques are classification, regression, clustering, association analysis, & graph analysis. A common application of association analysis is known as market basket analysis, which is used to understand customer purchasing behaviors. For example, predicting sunny weather gives very good results, but rainy weather? Perhaps you just need more samples of rainy weather, or perhaps there are some anomalies in these samples. Or maybe there are more samples of rainy weather, or perhaps there are some anomalies in these samples. Or maybe ideal situation would be that your model performs very well with respect to the success criteria that were determined when you defined the problem at the beginning of the project. In that case, you're ready to move on to communicating & acting on the results that you obtained from your analysis. As a summary, data analysis involves selecting the appropriate technique for your problem, building the model, then evaluating the results. As there are different types of problems, there are also different types of analysis techniques.

5) step 4: Communicating Results: —

Reporting insights:

The fourth step in our data science process is reporting the insights gained from our analysis.

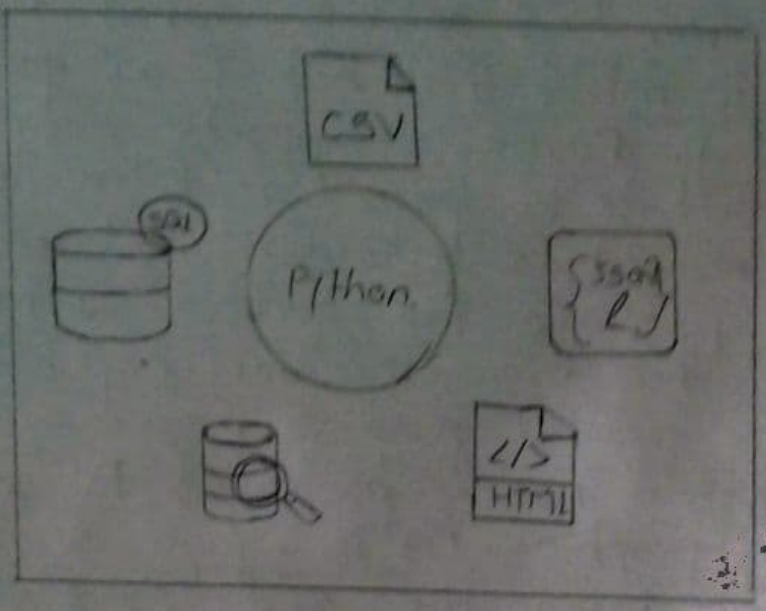
This is very important step to communicate your insights & mack a case for what actions should follow. Keep in mind that not of all your results may be easy. Domain experts may find some of these result to be puzzling, & inconclusive findings may lead to additional analysis. R is a software package for general data analysis. It has powerful visualization capabilities as well.

Python is a general purpose programming language that also has a no. of packages to support data analysis & graphics. D3 is a java script library for producing interactive web based visualizations & data driven documents. Leaflet is a lightweight mobile friendly JavaScript library to create interactive maps. Tableau Public Allows you to create visualization in your public profile, & share them, or put them, on a site, or blog. Google charts provides cross-browser compatibility & closed platform probability to Iphones & Androids. Timeline in a JavaScript library that allows you to create timelines. In summary, you want to report your findings by presenting your results & value-add with graphs using visualization tools.

6) Step 5:- Turning Insights into Action:—

After, turning insights into action. Now that you have evaluated the results from your analysis & generated report on the potential value of the results, the next step is to determine what action or actions

Data Ingestions:-



should be taken, based on the insights gained? Real-time action based on high velocity streaming information. We need to defined what part of our business needs real-time action to be able to influenses the operations or the interaction with the customers. Oneee we define these real time actions, we need to make sure that there are automated systems, or processes to perform such actions & provoid failure recovery in case of problems. Ass summary. big data & data signs are only useful if the insights can be turned into action, & if the actions are carefully defined & evaluated.

Read csv:—

- Input: Path to a comma separated file.
- Output: Pandas Dataframe object containing contents of the file.

read_json:—

- Input: Path to a JSON file or a valid JSON string.
- Output: Pandas Dataframe or a series object containing the contents.

read_html:—

- Input: A URL or a file or a raw HTML string
- Output: A list of Pandas DataFrames.

read_sql_query:—

- InputL: SQL Query.

•Input: Database connection.
•Output: Pandas Dataframe object containing contents of the file.

describe():-

Syntax:- data.frame.describe().
Output:- show summary statistics of the dataframe

ratings['rating'].describe().

count 2.000026e+07.
mean 3.525529e+00
std 1.051989e+00.
min 5.000000e-01
25% 3.000000e+00
50% 3.500000e+00
75% 4.000000e+00
max 5.000000e+00
Name: rating, dtype: float64.

corr():-

Syntax: data.frame.corr()
Computes pairwise pearson coefficient (P) of colu
other coefficients available: Kendall, Spearman.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

covariance.

standard deviation.

func = min(), max(), mode(), median().
The general syntax for calling these functions is
data_frame.func()
Frequently used optional parameters:
axis = 0(rows) or 1(columns).

mean()

Syntax: data_frame.mean(axis={0 or 1})
Axis = 0: Index.
Axis = 1: Columns.
Output: Series or Dataframe with the mean values
Syntax: data_frame.

std()

Axis = 0: Index.
Axis = 1: Columns.
Output: Series or Dataframe with the std. Deviation values.
Normalized by N-1.

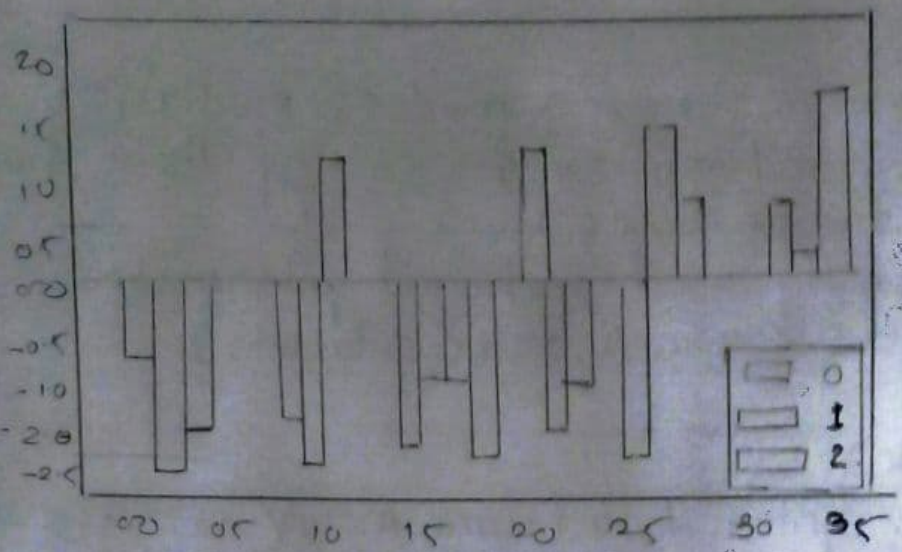
any()

Output: Returns whether ANY element is True.
Benefits:
Can detect if a cell matches a condition very quickly

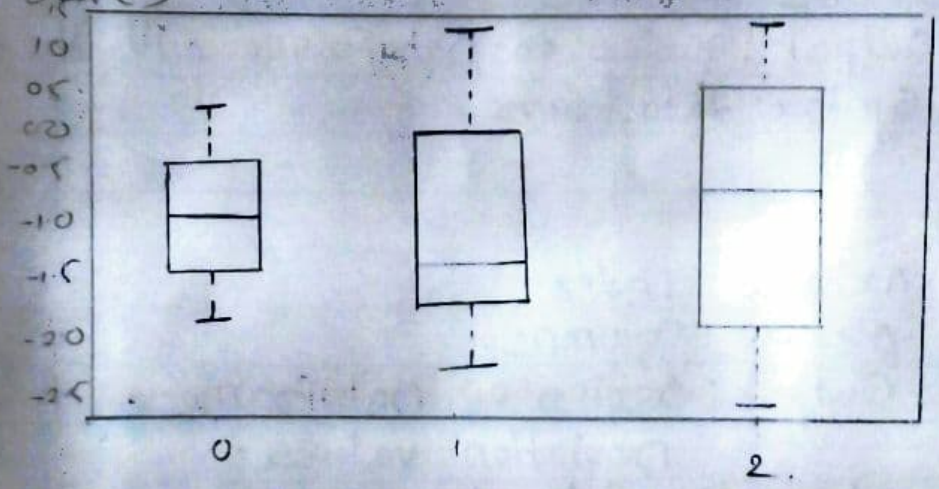
all()

Output: Returns whether ANY element is True.
Benefits:
Can detect if a cell matches a condition very quickl

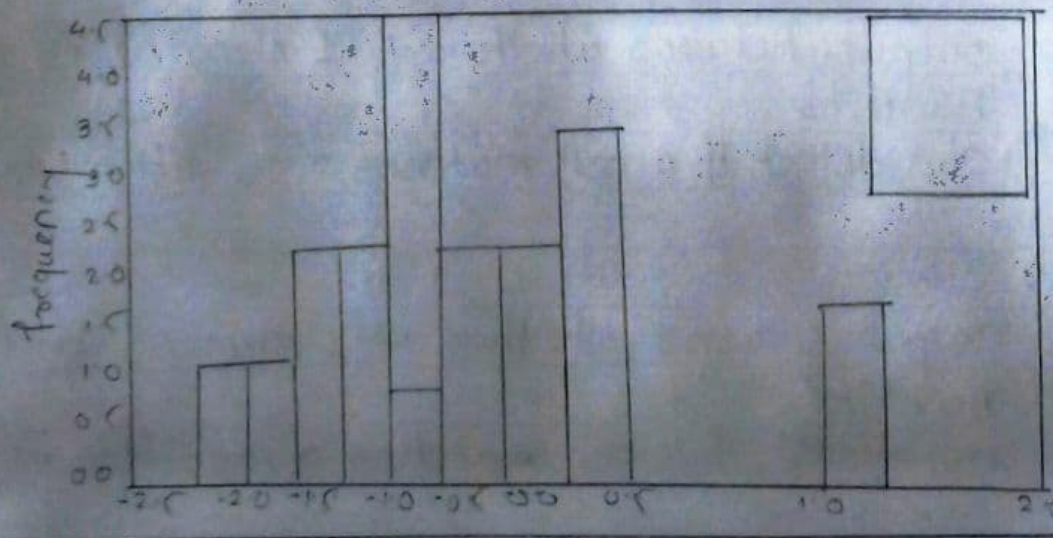
df.plot.bar()



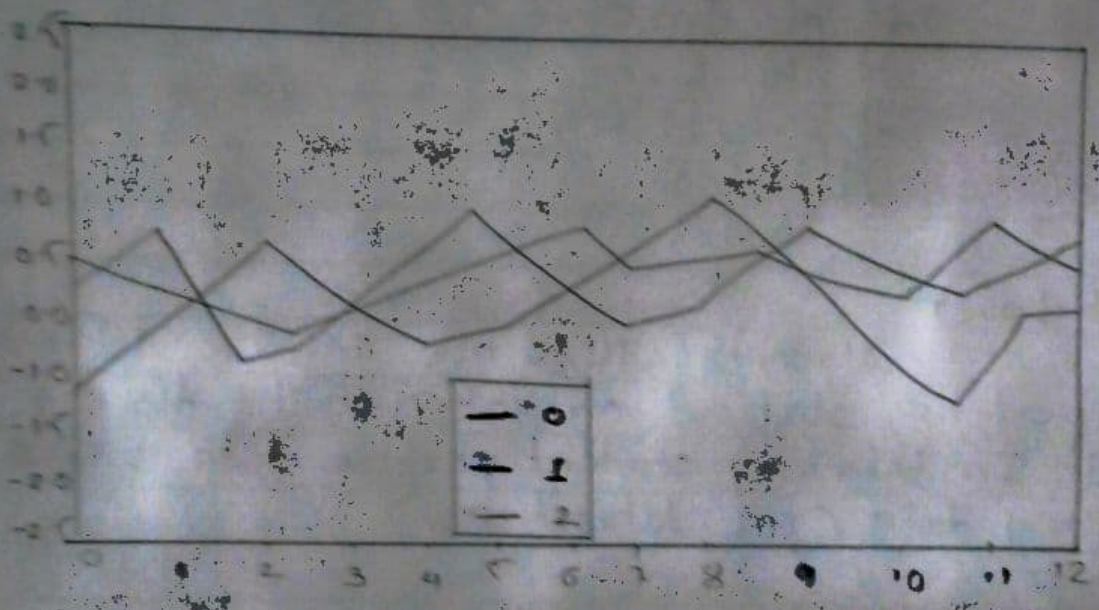
df.plot.box()



df.plot.hist().



plot()



Some other functions that are worth exploring:
count()
Clip()
Rank()
Round()

Data visualization:-

df.plot.bar().
df.plot.box()
df.plot.hist()
df.plot()

Post-Lab:-

Students will be able acquire different data & perform statical operation on it & display graph of it. In this way student perform data science operation on any data (IPL dataset.csv).

Conclusion:-

Thus student can implement note book for (perform step1 & step2 data science steps for any data (IPL-data) by using python tools like pandas, matplotlib, numpy etc.