Assignment No :- 03

Title : Write a python code for the loads any dataset & plot the graph.

problem statement : Write a python statemen code that loads any dataset & plot the graph.

pre-Lab : A basic understanding of Computer Programming terminologies. A basic understanding of a any of languages will help in understanding the python Programming & data concepts

steps in the data science process we have already seen a simple linear form of data science process including five distinct activitjes that depend on each other let's summarize each Acquire includes anything that makes us rethieve data including finding, accessing acquiring and moving data. It includes identification of authenticated acces to all related data and transportation of data from sources to distributed files System. It includes waste subset data to match the data to regions or times of interest As we sometimes refer to it as a geo-special query. the next activity is prepare data, we divide the pre-data activity. Into two steps. based on the nature of the activity. Namely explore data & pre-process the data. the first step in data preparation involves literally lo

looking at the data to understand its nature. What its means. its quality & format It often make a preliminary analysis of data. or sample of data to understand it. this is why this step is called explore. once we know more about the data through exploratory analysis the next step is pre-processing of data for analysis.

1. step1. Aquiring data.

step one, acquiring data. the first step in the data science process is to acquire the data you need to obtain the source materi-al before analyzing or acting on it. the first step in acquiring data is to determine what is finding the right data sources you want to identify suitable data related to your problem & make use of all data that is relevant to your problem for analysis leaving out even a small amount of important data can read to incorrect conclusion. Data comes from many places. local & remote in many varieties structured & un-structured and with different velocities. Their are different many techni-ques & technologies to acess these different type of data let's discuss a few examples A lot of data exists in Conventional relational database like structure big data from organization like the tool of choice to access data from databases is structured query languages or

or data base Systems Comes with a graphical application environment that allows you to query & explore the dataset in the database.

2. Step2. A : Exploring Data:

After you have put together the data that you need for you application you might be templated to immediatetly for your application. you might be templated to build ~~your~~ models to analyze the data Resist the temptation.

The first step after getting your data is to explore it. Exploring data is a part of the two-step data preparation process you want to do some preliminary investigation in order to gain a better understanding of the specific characteristics of your data. In this step you will not be able to use the data effectively - correlation graphs can be ~~able~~ to use the ~~data~~ explore the dependencies between different variable in the data graphing the general trends of Variables will show you if there is a consistent direction in which the value of these Variables are moving towards, like sales prices going up or down In statistics. an outlier is a data pt that distant from other data points

3. Step2-B. pre-processing Data.

The row data that you get directly from your sources are never in the format that you need to perform analysis on. There are two main goals in the data preprocessing. b

step: the 1st is to clean the data to address data quality issues & the second is to transfo -orm the raw data to make it suitable for analy -sis A s very important part of data preparation is to address quality of issue in your data Real-world data is messy. there are many examples of quality issues with data from real application including inconsistent data like a customer with two different addresses. duplicate customer records, for example, customer addresses, recorded at two different sales location and the two recordings don't agree missing customer agent demographics or studies missing values like missing a customer agent depart agent studies invalid data like an invalid pip code for example, a six digit code and outliers like a sense of failure causing values to be much bigher of lower than expected for a period of time since we get the data downstream we usually have a little control over how the data is collected preventing data quality problems as the data is being collected is not often an option.

4 Step: Analyzing Data.

Now that you have data nicely prepa- -red the next step is to analyze the data analysis involves bulding a model from your data which is called i/p data. the input data is used by the anaysis technique to buil

a model. What your model generates the output data. There are different types of problems and so there are different types analysis techniques. the main categories of analysis techniques are classification, regression, clustering association analysis & graph analysis we will describe each one In classification. the goals to predict the category of the input data An example of this is predicting the whether as being sunny, rainy, windy or cloudy in the this case another example is to classify a tumor as either begin or malignant. In this Case the classification is reffered to as binary classification. since there are only two categories But you can have many categories as well as the weather prediction problem is shown here having four categories. another example is to identify handwritten digits as being is one of the ten Categories from zero to nine.
When your model has to predict a numeric Value instead of a category. then task become a regression problem.

5 Steps: Communicating Results:

The fourth step in our data science process is reporting the insights gained from our analysis. This is a very important step to communicate your insights & make a case for what action should follow. It can change shape based on your audience & should not.

taken lightly So how do you get started? the first thing to do is to look at your analysis results and decide what to present or report as the biggest value or biggest set of values in deciding what to present you should ask yourself this questions. What is punching. In other words What are the main results? What added value to do these results provide or how can model add to application? How do the results compare to the success criteria determined at the begining of the project? Answer to the questions are the items you need to include in your report or presentation.

6 step: Turning Insights into Action.

Now that you have evaluated the results from your analysis & generated reports on potential value of the results, the next step is to determine what action or action should be taken, based on the insights gained? To find actionable insights within all these data sets. to answer questions or for improving buisiness process. for ex. is there something in your process that should change to remove battle neeks? Is there data that should be added to your appln to make it more accurate? should change to remove your population into more well defined groups for as a summary, big data & data signs are only useful. if the insights. Can be turned into action & if the actions are carefully define & evaluated

Describe()
- Syntax: data_frame corr().describe()
- output: shows summary statistics of dataframe

Corr()
- Syntax: data_frame corr()
- computes pairwise pearson coefficient (r) of columns
- other coefficients available kendall pearson

$$p(x,y) = \frac{cov(X,Y)}{\sigma x . \sigma y}$$

cov → covariance
σ → 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 (column)

mean()
- Syntax: data_frame.mean(axis = {0 or 1})
- Axis = 0 Index
- Axis = 1 columns
- output: series or Data frame with standard deviation Normalized by N-1

std()
- Axis=0: Index
- Axis=1 columns
o/p = series or Dataframe with Standard deviation
Normalized by N-1.

any()
- o/p: Returns whether Any ele is true
- Benefits: Can detect if a cell matches a cond^n very quickly.

all()
- o/p: Returns whether ALL element is true
- Benefits Can detect if a column or row matches a condition very quickly

Some other func^n that we are worth exploring
- count()
- clip()
- Rank()
- Round()

Data Visualization:
df.plot bar()
df.plot box()
df.plot hist()
df.plot()

Conclusion :-

Thus student can implement note book for 8 step 2 data science steps for any data by using python tools like pandas, matplotlib, numpy etc.