SSBT's College of Engineering & Technology,
Bambhori, Jalgaon
Department of Computer Engineering

Name: Rajput Pruthviraj Dineshsing
Class: B.E. Computer
Division : A
Batch: B-2 Roll No: 37
Subject: Advance Technology Lab-I

Date of Performance: _ _ /_ _/20_ _
Date of Completion: _ _ /_ _/20_ _

Experiment No. 2.

Aim: Study and perform Data Analytics and Data Visualization using Numpy, Pandas and Matplotlib

1. Objective: To perform data analysis and visualization on a sample dataset.

2. Background:

2.1 NumPy

It is a Python package. It stands for 'Numerical Python'. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

Numeric, the ancestor of NumPy, was developed by Jim Hugunin. Another package Numarray was also developed, having some additional functionality. In 2005, Travis Oliphant created NumPy package by incorporating the features of Numarray into Numeric package. There are many contributors to this open source project.

Operations using NumPy

Using NumPy, a developer can perform the following operations −

Mathematical and logical operations on arrays.

- Fourier transforms and routines for shape manipulation.
- Operations related to linear algebra. NumPy has in-built functions for linear algebra and random number generation.

NumPy – A Replacement for MatLab

NumPy is often used along with packages like **SciPy** (Scientific Python) and **Mat−plotlib** (plotting library). This combination is widely used as a replacement for MatLab, a popular platform for technical computing. However, Python alternative to MatLab is now seen as a more modern and complete programming language. It is open source, which is an added advantage of NumPy.

2.2 Pandas

Pandas is an open-source Python Library used for high-performance data manipulation and data analysis using its powerful data structures. Python with pandas is in use in a variety of academic and commercial domains, including Finance, Economics, Statistics, Advertising, Web Analytics, and more. Using Pandas, we can accomplish five

typical steps in the processing and analysis of data, regardless of the origin of data — load, organize, manipulate, model, and analyse the data.

Below are the some of the important features of Pandas which is used specifically for Data processing and Data analysis work.

Key Features of Pandas

- Fast and efficient DataFrame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and subsetting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations.
- High performance merging and joining of data.
- Time Series functionality.

Pandas deals with the following three data structures −

- **Series**
- **DataFrame**

These data structures are built on top of Numpy array, making them fast and efficient.

Data Structure	Dimensions	Description
Series	1	1D labeled homogeneous array, size-immutable.
Data Frames	2	General 2D labeled, size-mutable tabular structure with potentially heterogeneously typed columns.

DataFrame is widely used and it is the most important data structures.

Series

Series is a one-dimensional array like structure with homogeneous data. For example, the following series is a collection of integers 10, 23, 56, …

Key Points of Series

- Homogeneous data
- Size Immutable
- Values of Data Mutable

DataFrame

DataFrame is a two-dimensional array with heterogeneous data. For example,

Name	Age	Gender	Rating
Steve	32	Male	3.45
Lia	28	Female	4.6
Vin	45	Male	3.9
Katie	38	Female	2.78

The table represents the data of a sales team of an organization with their overall performance rating. The data is represented in rows and columns. Each column represents an attribute and each row represents a person. Key Points of Data Frame

- Heterogeneous data
- Size Mutable
- Data Mutable

2.3 Matplotlib

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. Matplotlib is written in Python and makes use of NumPy, the numerical mathematics extension of Python. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.

Matplotlib has a procedural interface named the Pylab, which is designed to resemble MATLAB, a proprietary programming language developed by MathWorks. Matplotlib along with NumPy can be considered as the open source equivalent of MATLAB.

Matplotlib was originally written by John D. Hunter in 2003. The current stable version is 2.2.0 released in January 2018.

3. Pre-lab Task:

1. Installation of Ubuntu 18.04
2. Installation of Anaconda 3 with Jupyter Notebook
3. Installation of Numpy Package
4. Installation of Pandas Package
5. Installation of Matplotlib Package

4. In Lab Tasks

Perform the following data analysis and visualization on the given dataset.

1. **Write a Python program to display first 5 rows from COVID-19 dataset. Also print the dataset information and check the missing**

2. **Write a Python program to get the latest number of confirmed, deaths, recovered and active cases of Novel Coronavirus (COVID-19) Country wise.**

3. **Write a Python program to get the latest number of confirmed deaths and recovered people of Novel Coronavirus (COVID-19) cases Country/Region - Province/State wise.**

4. **Write a Python program to get the Chinese province wise cases of confirmed, deaths and recovered cases of Novel Coronavirus (COVID-19).**

5. **Write a Python program to list countries with no cases of Novel Coronavirus (COVID-19) recovered.**

6. **Write a Python program to get the top 10 countries data (Last Update, Country/Region, Confirmed, Deaths, Recovered) of Novel Coronavirus (COVID-19).**

7. **Write a Python program to create a plot (lines) of total deaths, confirmed, recovered and active cases Country wise where deaths greater than 150.**

8. **Write a Python program to visualize the state/province wise combine number of confirmed, deaths, recovered, active Novel Coronavirus (COVID-19) cases in USA.**

9. **Write a Python program to visualize Worldwide Confirmed Novel Coronavirus (COVID-19) cases over time.**

5. Post Lab Tasks

Performed data analysis and visualization on the sample dataset using Numpy, Pandas and Matplotlib and post the GitHub link and the Google Site URL

- GitHub Link: https://github.com/pruthvi7384/Python-ATL-Assignment.git
- Google Site URL: https://pruthvirajrajput.great-site.net/

Questions

1. **Enlist and describe key features of NumPy**
2. **Enlist and describe key features of Pandas**

3. **Enlist and describe key features of Matplotlib**
4. **Describe two important data structures of Pandas**
5. **Describe how Pandas is more powerful as compared to other standard data types in Python**