**Experiment No. 1**

**Title: Exploratory data analysis using NUMPY**

Batch: B2 Roll No: 1914078 Experiment No.:1

### Aim: To perform exploratory data analysis using python NUMPY

Resources needed: Python IDE

### Theory:

### Data Analysis is basically where you use statistics and probability to figure out trends in the data set. It helps you to sort out the “real” trends from the statistical noise

### Exploratory Data Analysis (EDA) in Python is the first step in your data analysis process developed by “John Tukey” in the 1970s.

### In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

### The main aim of exploratory data analysis is to obtain confidence in your data to an extent where you’re ready to engage a machine learning algorithm.

Basically we do following things in EDA.

1) Quickly describe a dataset; number of rows/columns, missing data, data types, preview.

2) Clean corrupted data; handle missing data, invalid data types, incorrect values.

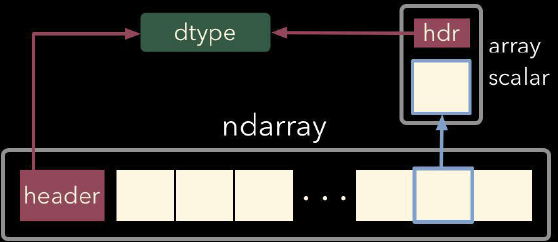
3) Visualize data distributions; bar charts, histograms, box plots.

4) Calculate and visualize correlations (relationships) between variables;

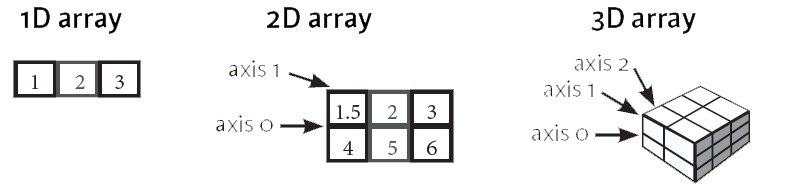
NUMPY(Numeric or Numerical Python):NumPy is a Python library that is the core library for scientific computing in Python.

It contains a collection of tools and techniques that can be used to solve on a computer mathematical models of problems in Science and Engineering.

One of these tools is a high-performance multidimensional array object, ndarray, that is a powerful data structure for efficient computation of arrays and matrices. Memory layout of ndarray is shown in figure below.



Memory layout of ndarrary of python



1D, 2D and 3D arrays in numpy

To work with these arrays, there’s a vast amount of high-level mathematical functions operate on these matrices and arrays.

NumPy’s main object is the homogeneous multidimensional array. It is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive integers. In NumPy dimensions are called *axes*.

For example, the coordinates of a point in 3D space [1, 2, 1] has one axis. That axis has 3 elements in it, so we say it has a length of 3. In the example pictured below, the array has 2 axes. The first axis has a length of 2, the second axis has a length of 3.

[[ 1., 0., 0.], [ 0., 1., 2.]]

NumPy’s array class is called ndarray. It is also known by the alias array.

numpy.array is not the same as the Standard Python Library class array.array, which only handles one-dimensional arrays and offers less functionality. ndarray.ndim the number of axes (dimensions) of the array.

The more important attributes of an ndarray object are:

ndarray.ndim the number of axes (dimensions) of the array.

ndarray.shape the dimensions of the array. This is a tuple of integers indicating the size of the array in each dimension. For a matrix with *n* rows and *m* columns, shape will be (n,m). The length of the shape tuple is therefore the number of axes, ndim.

ndarray.size the total number of elements of the array. This is equal to the product of the elements of shape.

ndarray.dtype an object describing the type of the elements in the array. One can create or specify dtype’s using standard Python types.

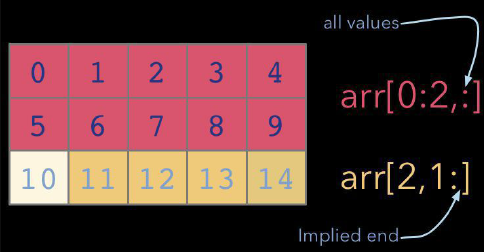
ndarray.itemsize the size in bytes of each element of the array.

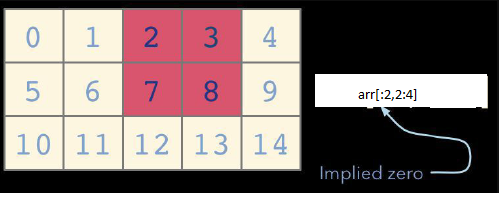
ndarray.data the buffer containing the actual elements of the array. Normally, we won’t need to use this attribute because we will access the elements in an array using indexing facilities.

Using numpy.ones(), numpy.zeros(), numpy.empty() we can create standard arrays of ones, zeros and uninitialized numbers respectively.

We can create array from list of homogeneous numbers as well.

Slicing and indexing in numpy arrays: figure below gives idea about slicing and indexing in numpy array.





In order to use ndarray and its related attributes and functions, we first have to make sure that numpy is installed. Since numpy is basic library of python it comes along with most of the python IDE. In case it is not installed we can download latest wheel of numpy and install it using pip install.

One it is installed using following statement it can be import and its functionalities can be used.

import numpy as nd

#creating array of zeros

np.zeros(5, float)

similarly we can use following functions to find statistical measures using ndarray.

x.sum(),x.mean(),x.min(0,x.max() etc

one can pass axis=0 or axis=1 to do columnwise and rowwise operations.

reshape() function will resize array as per new dimensions passed as an arguments to it.

vstack() and vstack() for concatenation of two compatible arrays

various matrix operations like add(), subtract(),multiply(), divide(), dot() can be performed on 2D arrays in numpy. Numpy allows broadcasting of arrays for uncompatible dimensions which will help while performing these operations.

### Activities:

### 1. Download data set with atleast 1500 rows and 10-20 columns(numeric and non numeric) from valid data sources

### 2. Perform in detail Exploratory data analysis of this dataset

### 3. Write down description of your dataset based on analysis done in activity

### 4. Write atleast 5 different types of conclusions on your dataset

### Result: My dataset contains all the terraform and earthquakes recorded post the invention of instruments to do the same.

### Importing and getting data

* import numpy as np
* data = np.genfromtxt("./database.csv",skip\_header=1,delimiter=",")
* print(data)

### 

### To get the text

import numpy as np

import csv

with open('./database.csv','r') as f:

    df = list(csv.reader(f,delimiter=","))

df1 = np.array(df[1:])

print(df1)

### 

### What is the maximum magnitude of earthquake?

### 

### What is the minimum magnitude of earthquake?

### 

### What is average magnitude of all earthquakes?

### 

### What are the total number of earthquakes that have hit US and it’s average magnitude?

### 

### What is average magnitude error?

### 

### What is average horizontal error?

### 

### What is standard deviation for depth of the epicenter of the earthquake?

### 

### What is the depths of maximum magnitude earthquakes from the entire dataset?

### 

### From the dataset, we can conclude that the worst earthquake ever faced by humans recorded a maximum of 9.1 magnitude and the lowest one being 5.5 magnitude.

### From the dataset, we can conclude that on average an earthquake has a magnitude of 5.88 with magnitude and horizontal calculation errors being 0.071 and 7.66.

### From the above data, we can conclude total number of earthquakes that have hit the country of US and the average of all these earthquakes causing mass devastation.

### From the above data, we can conclude the standard deviation of depth of the epicenter of the earthquake being 122.65 giving us a common scale to compare it the rest of the data which can be further used to plot data visually and predict future depths of earthquakes using this information.

### From the above data, we can conclude that it’s not necessary to have the maximum depth at maximum magnitude as at maximum magnitude of 9.1,

### The depth is 30 and 29 metres below the crust

### Outcomes: Use python libraries like matplotlib, numpy, pandas, scipy for data visualization and scientific-mathematical data computing.

Conclusion: We performed exploratory data analysis using python NUMPY.

References:

1. <https://www.geeksforgeeks.org/python-numpy/>