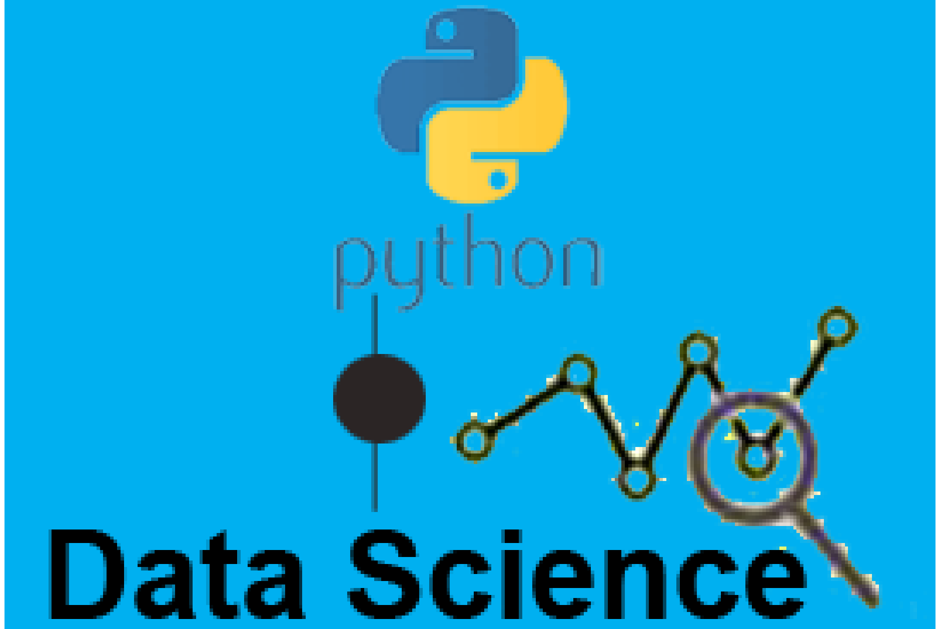
miet

Introduction to Data Science And Data Visualization

Speaker : Nishant Sharma
Date: 13th August, 2018

UDYAT COMMUNITY
Udyat-miet.github.io



@udyat_miet

What is Data Science ?

Data Science is the process of deriving knowledge from a huge and diverse set of data through organizing processing and analysing the data.

What is Data Vizualisation ?

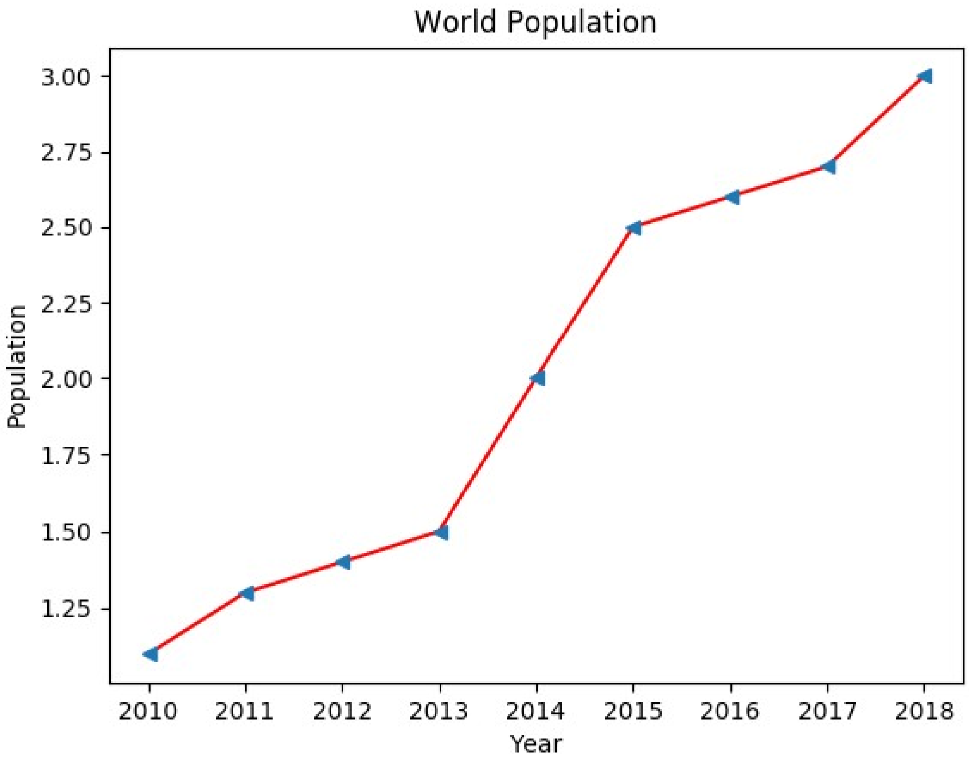
Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context.

Lets take a example

2010	1.1
2011	1.3
2012	1.4
2013	1.5
2014	2.0
2015	2.5
2016	2.6
2017	2.7
2018	3.0

Example 1

Data without visualisation



Example 2

Data with visualisation

What we cover

2 Python Packages

1. Numpy
2. Matplotlib

How to use this package in your program?

How to install Package

- Open the Terminal and write

```
pip install package_name
```

for example:

```
pip install numpy
```

How to import into Program

- In your program write

```
import package_name
```

Numpy Stands For Numerical Python

It is a library consisting of multidimensional array objects.

OPERATIONS USING NUMPY

- Mathematical and logical operation on arrays.

NUMPY

How to initialise array using numpy and perform operation on array.

BY using the function

np.array(list_name)

np.arange(start, end, inc)

How to perform operation on array.

np.mean(array) #mean

np.median(array) #median

np.std(array) #standard deviation

np.shape(array) #type of array

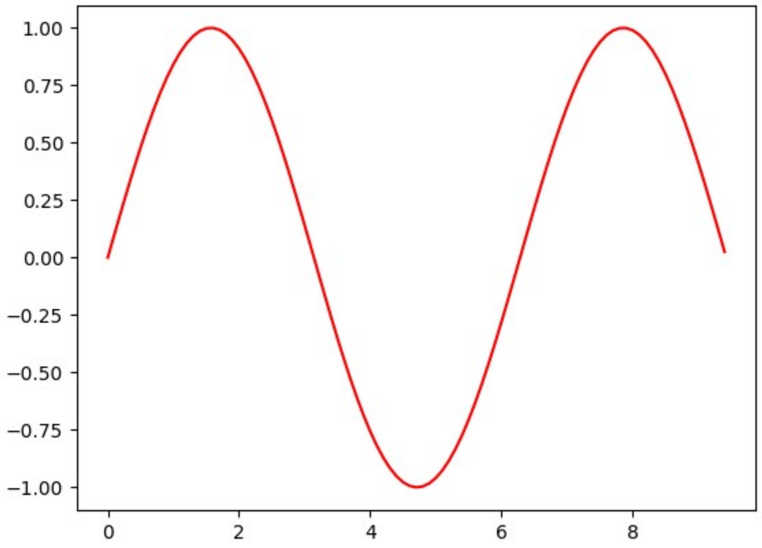
Matplotlib

Matplotlib is a python library used to create 2D Graphs and plots by using python scripts.

IT Supports a very wide variety of graphs and plots namely – histogram, bar charts, power spectra, error charts etc.

Lets draw a simple sin graph

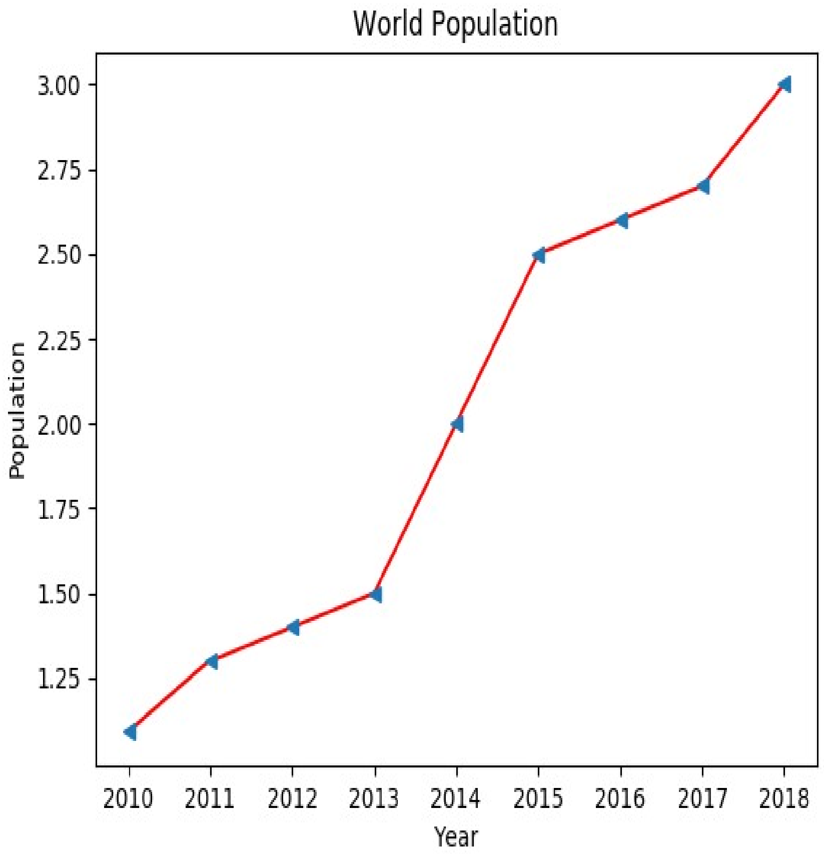
```
import  matplotlib.pyplot as plt
import numpy as np

x =  np.arange(0, 3* np.pi , 0.1)
y = np.sin(x)
plt.plot( x,y )
plt.show()
```



Types of Graph

1. Line graph
2. Scatter graph
3. Histogram

Lets Take again example of world population



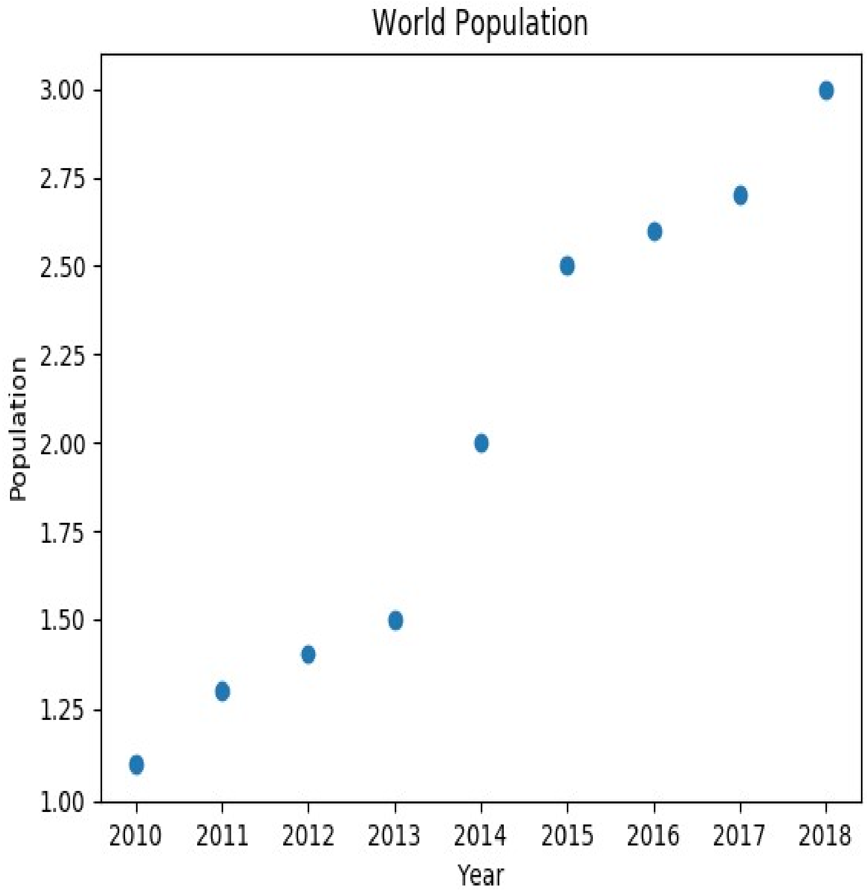
Line Graph

Function Used

Plt.plot(x_axis_array , y_axis_array)
Plt.show() #to display graph
Plt.savefig(‘name.format’ , format= ‘name’) #to save the figure
Plt.xlabel(‘string’)
Plt.ylabel(‘string’)
plt.title(‘string’)
Plt.xticks and plt.yticks

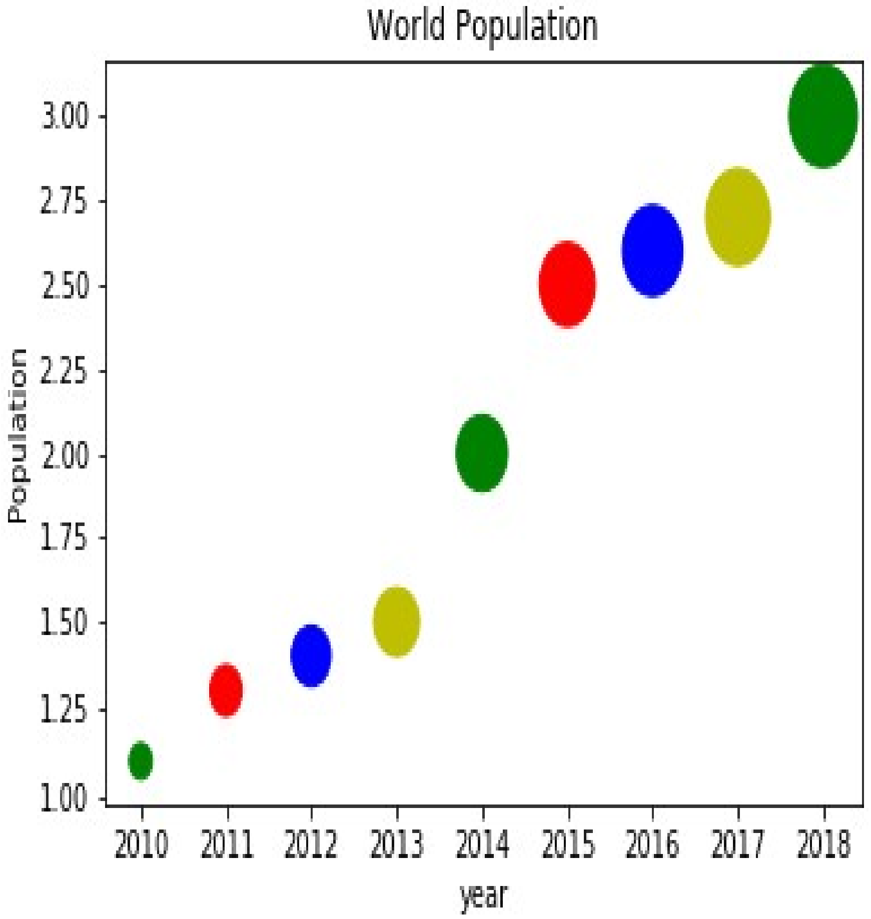
Scatter Graph

```
import numpy as np
import matplotlib.pyplot as plt

pop_list = [1.1,1.3,1.4,1.5,2.0,2.5,2.6,2.7,3 ]

year = np.arange(0,9,1)
pop = np.array(pop_list)

plt.scatter( year, pop)
plt.xlabel(“Year”)
plt.ylabel(“Population”)
plt.title(“World Population”)
plt.display()
```



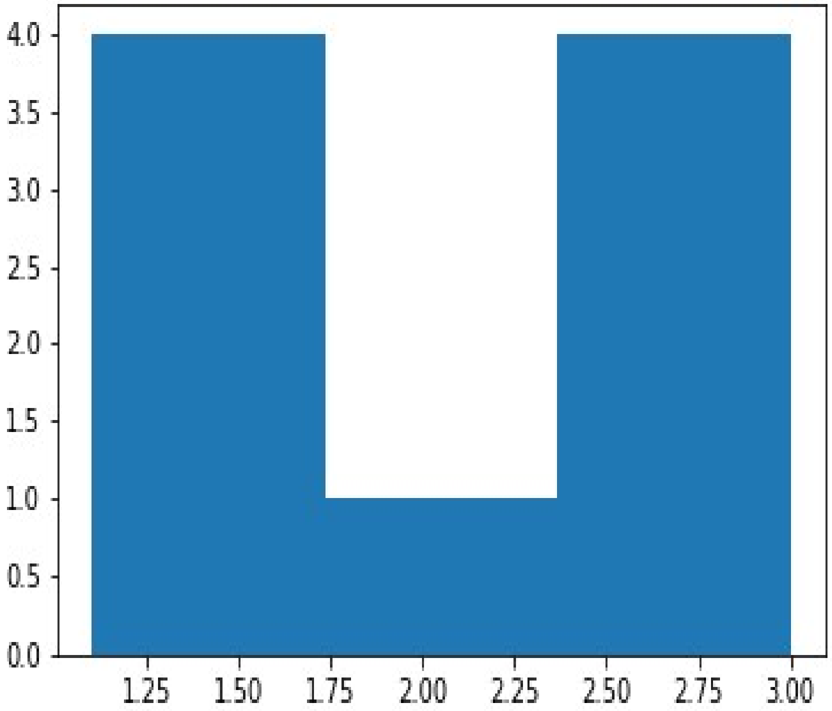
Customisation

```
import numpy as np
import matplotlib.pyplot as plt
pop_list =
[1.1,1.3,1.4,1.5,2.0,2.5,2.6,2.7,3.0 ]

year = np.arange(1,10,1)
pop = np.array(pop_list)

plt.scatter( year,
pop,s=year*100,color=['g','r','b','y'])
plt.xlabel('year')
plt.ylabel('Population')
plt.title('World Population')
plt.xticks(year,['2010','2011','2012','2013
','2014','2015','2016','2017','2018'])
plt.show()
```



Scatter Graph

```
import numpy as np
import matplotlib.pyplot as plt
pop_list =
[1.1,1.3,1.4,1.5,2.0,2.5,2.6,2.7,3.0 ]

pop = np.array(pop_list)

plt.hist( pop)
plt.show()
```



Histogram

https://udyat-github.io/seminars

Thank You

Pandas is a python library used for high-performance data manipulation and data analysis using its powerful data structures.

PANDAS

Key Feature

- Fast and Efficient DataFrame object with default and customized indexing.

PANDAS data Structures

1. Series (1-D labeled homogeneous)

2. Data Frames (General 2D labeled, tabular structure)

PANDAS Series data structure

Initialise using the module

```
pandas.Series( array_name )
```

PANDAS DataFrame data structure

Initialise using the module

pandas.DataFrames(2D_array_name)