Atmiya University
Faculty of Science,
Department of Computer Science & I.T.

Subject Name: **21UFSDE309 Data Science Using Python**

Introduction to Pandas in Python

- Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and naturally.
- It provides various data structures and operations for manipulating numerical data and time series.
- This library is built on top of the NumPy library.
- Pandas is fast and it has high performance & productivity for users.

Introduction to Pandas in Python

- **History:** Pandas were initially developed by Wes McKinney in 2008 while he was working at AQR Capital Management.
- He convinced the AQR to allow him to open source the Pandas.
- Another AQR employee, Chang She, joined as the second major contributor to the library in 2012.
- Over time many versions of pandas have been released. The latest version of the pandas is 1.4.1

Introduction to Pandas in Python

- **Advantages**
- Fast and efficient for manipulating and analyzing data.
- Data from different file objects can be loaded.
- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Data set merging and joining.
- Flexible reshaping and pivoting of data sets
- Provides time-series functionality.
- Powerful group by functionality for performing split-apply-combine operations on data sets.

Installation of Pandas

- If you have Python and PIP already installed on a system, then installation of Pandas is very easy.
- Install it using this command:
- Write below command in command prompt
- C:\Users\Your Name>pip install pandas

- **Checking Pandas Version**
- The version string is stored under __version__ attribute.
- **Example**

```
import pandas as pd
print(pd.__version__)
```

What is Data Structure in Pandas?

- Pandas is divided into three data structures when it comes to dimensionality of an array. These data structures are:
- Series
- DataFrame
- Panel

What is Data Structure in Pandas?

Data Structure	**Dimensions**
Series	1D
DataFrame	2D
Panel	3D

- **Series** and **Data Frames** are the most widely used data structures based on the usage and problem solving sets in data science. If we look at these data structures in terms of a spreadsheet then Series would be a single column of an excel sheet, whereas DataFrame will have rows and columns and be a sheet itself.

What is a Series in Pandas?

- Pandas series is a one dimensional data structure which can have values of integer, float and string. We use series when we want to work with a single dimensional array. It is important to note that series cannot have multiple columns. It only holds one column just like in an excel sheet. Series does have an index as an axis label. You can have your own index labels by customizing the index values.

What is a Series in Pandas?

- This is Series

Name
Dhyey
Krishna
Kishan
Radha
Shyam

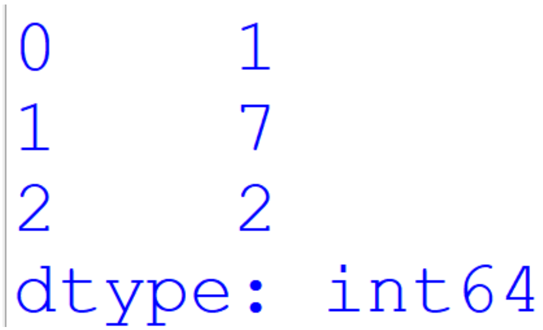
What is a Series in Pandas?

- **Creating a Series in Pandas**
- Pandas Series can be created in different ways from MySQL table, through excel worksheet (CSV) or from an array, dictionary, list etc. Let’s look at how to create a series. Let’s import Pandas first into the python file or notebook that you are working in:
- **import pandas as pd**
- **ps = pd.Series([1,2,3,4,5])**
- **print(ps)**

#Example to print data using Pandas Series.

- import pandas as pd
- a = [1, 7, 2]
- myvar = pd.Series(a)
- print(myvar)

- Output



```
0      1
1      7
2      2
dtype: int64
```

#Example of DataFrames with pandas.

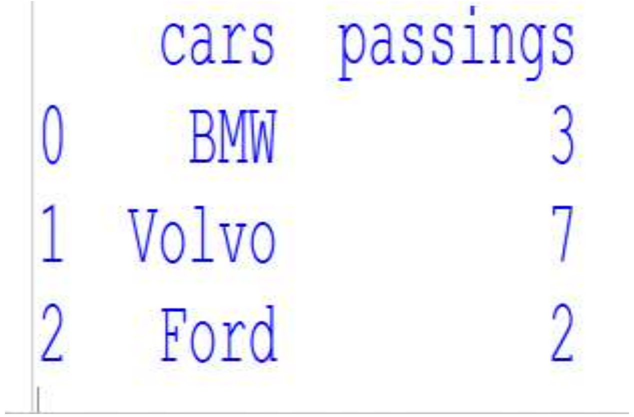
- import pandas as pd

- mydataset = {
-   'cars': ["BMW", "Volvo", "Ford"],
-   'passings': [3, 7, 2]
- }

- myvar = pd.DataFrame(mydataset)

- print(myvar)

- Ouput



```
    cars  passings
0    BMW         3
1  Volvo         7
2   Ford         2
```

#first create data.csv file in Excel and then #read data from CSV file.(Import data from file)

- import pandas as pd

- df = pd.read_csv('data.csv')

- print(df.to_string())

#Exporting data to csv file.

```
import pandas as pd

mydataset = {
  'country': ["India", "Japan", "Dubai"],
  'State': ["Gujarat", "Tokyo", "Sharja"]
}

df = pd.DataFrame(mydataset)

print(df.to_csv('outputfile.csv',index=False))
```

What is Matplotlib?

- Matplotlib is a low level graph plotting library in python that serves as a visualization utility.
- Matplotlib was created by John D. Hunter.
- Matplotlib is open source and we can use it freely.
- Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility.

Installation of Matplotlib

- **Installation of Matplotlib**
- If you have Python and PIP already installed on a system, then installation of Matplotlib is very easy.
- Install it using this command:
- C:\Users\Your Name>pip install matplotlib

Import Matplotlib

- **Import Matplotlib**
- Once Matplotlib is installed, import it in your applications by adding the import *module* statement:
- **import matplotlib**

- **Checking Matplotlib Version**
- The version string is stored under __version__ attribute.
- **Example**
- import matplotlib
  print(matplotlib.__version__)

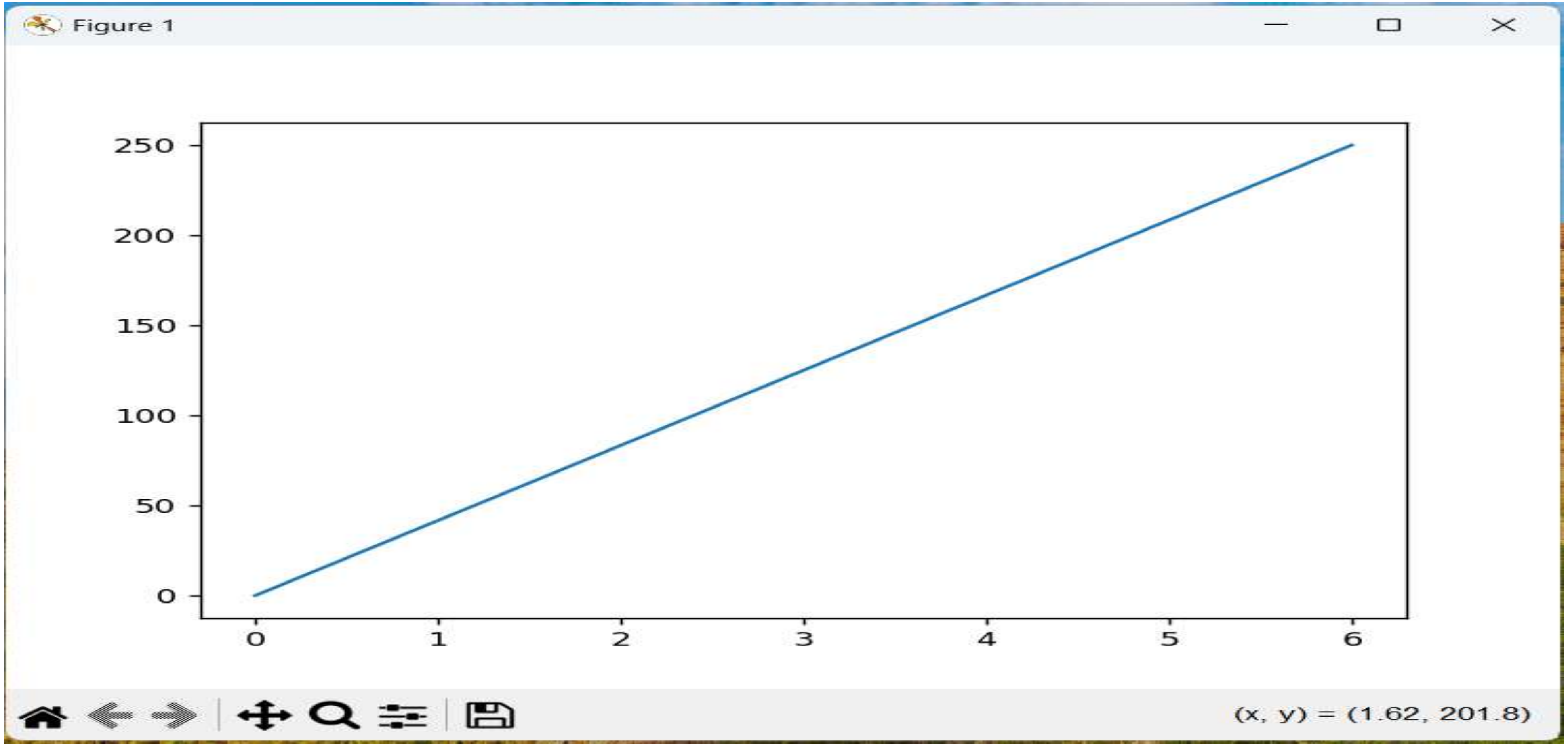
#draw graph of straight line using Matplotlib.

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

xpoints = np.array([0, 6])
ypoints = np.array([0, 250])

plt.plot(xpoints, ypoints)
plt.show()
```

Output



#create graph of multiple plots.

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

- xpoints = np.array([0,2,5,10])
- ypoints = np.array([5,10,3,15])

- plt.plot(xpoints, ypoints)
- plt.show()

Output

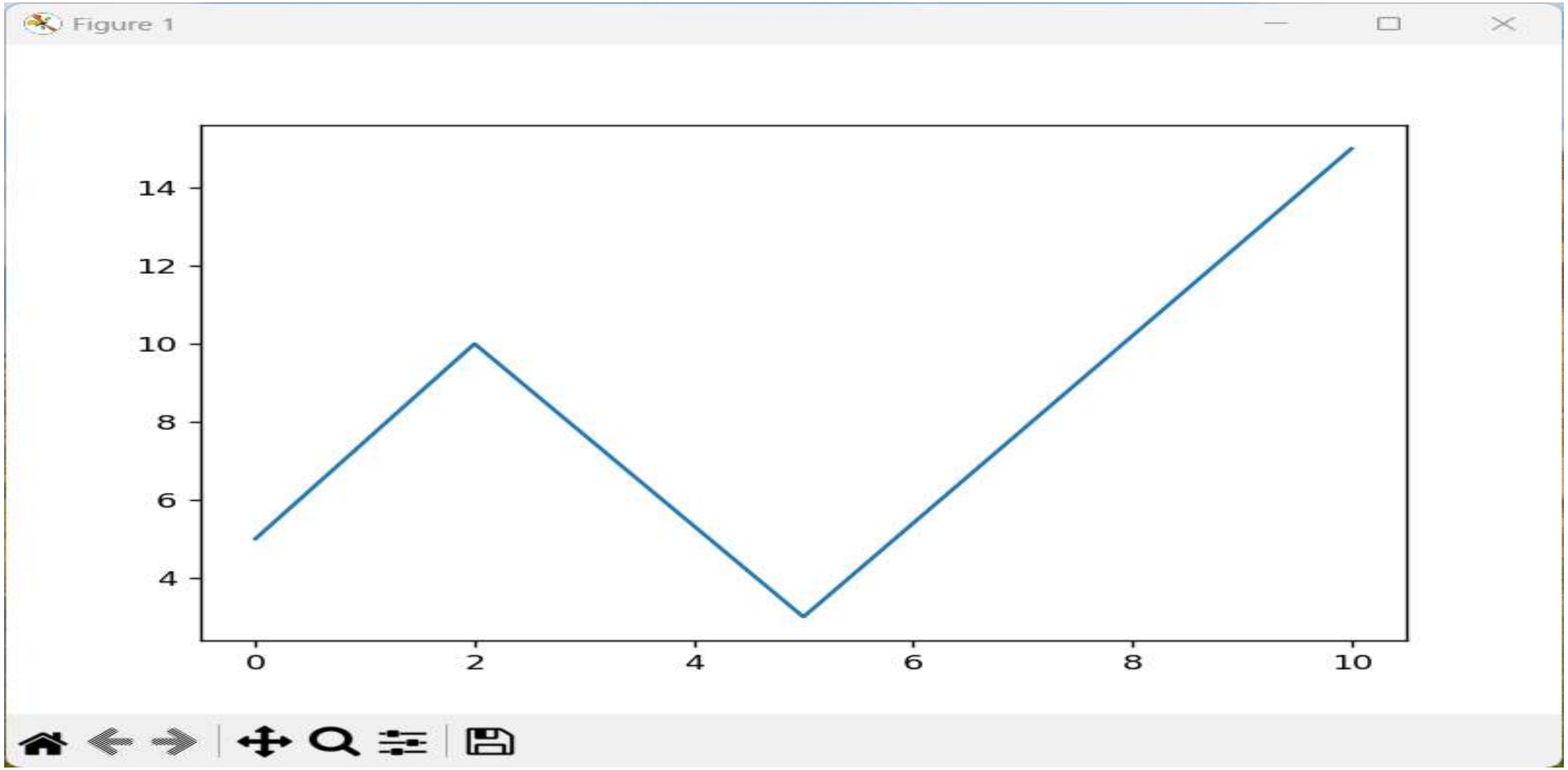


Figure 1
14
12
10
8
6
4
0
2
4
6
8
10