PANDAS

- Pandas stand for Panel Data System
- Pandas is an open source library for data analysis, Data manipulation and Data Visualization.
- (OR) Pandas provide powerful data structures for data analysis, time series and statistics.
- Pandas works on the top numpy and matplotlib.

Features of pandas

1. Handling huge amount data
2. Missing Data
3. Cleaning up data
4. Alignment and indexing
5. Merging and joining
6. Grouping and Visualizing data
7. Time Series Functionality
8. Allows to load data from multiple file formats
9. Input and Output Tools

Pandas library is used by scikit-learn for ML

Applications of Pandas

1. Recommendation Systems
2. Stock Prediction
3. Big Data and Data Science
4. NLP (Natural Language Processing)
5. Statistics and Analytics
6. Neuroscience

Important data structures of Pandas are,

1. Series
2. DataFrame

Q: What is data analysis?

Data analysis is process of collecting, transforming, cleaning and modeling the data with goal of discovering required information.
Data analysis process consists of the following steps.

1. Data Requirement Specifications
2. Data Collection
3. Data Processing
4. Data Cleaning
5. Data Analysis
6. Communication

What is Series?
Pandas series is a one dimensional array object, this object can hold data of any type. It can be integers, floats, string or python objects.
Pandas series represents or equal to a column in any data base (MsExcel, Oracle, MySQL, SQLServer,..)

What is DataFrame?
DataFrame is a two dimensional array object or data structure. Data stored tabular format, which is rows and columns.
The Dataframe consist of 3 components.

1. Data
2. Rows
3. Columns

How to install pandas?
Other than jupyter and googlecolab, it is required to install pandas lib.

pip install pandas

Pandas Series
Series is single dimension array like object with homogeneous or heterogeneous data.
Series object can be created in different ways.

1. Using array
2. Using Dictionary
3. Using Scalar values
4. Using other iterables

Series is name of the class or type which is used to construct Series object.

Syntax: Series(data,index,dtype)

Data : the source using which series object is created
Index : index values must hashable and must be unique/access labels
dtype: type of the series is defined using dtype.

Creating Empty Series

```
import pandas as pd
import numpy as np
s1=pd.Series(dtype=np.int8)
print(s1)
```

```
Series([], dtype: int8)
```

Creating Series using List object

```
s2=pd.Series([10,20,30,40,50])
print(s2)
s3=pd.Series([10,20,30,40,50],index=['a','b','c','d','e'])
print(s3)
```

```
0    10
1    20
2    30
3    40
4    50
dtype: int64
a    10
b    20
c    30
d    40
e    50
dtype: int64
```

Creating Series using ndarray

```
a=np.ndarray(shape=(5,))
i=0
for value in range(10,60,10):
    a[i]=value
    i+=1
print(a)
print(type(a))
s=pd.Series(a)
print(s)
```

```
[10. 20. 30. 40. 50.]
<class 'numpy.ndarray'>
0    10.0
1    20.0
2    30.0
3    40.0
4    50.0
dtype: float64
```

Creating Series Using Dictionary

We can create series using dictionary (OR) we can pass the dictionary object to series.

Series object is using dictionary values as data and dictionary keys as index labels.

```
sales_dict={2018:50000,2019:60000,2020:75000}
s=pd.Series(sales_dict)
print(s)
emp_dict={'naresh':5000,'suresh':6000,'kishore':9000}
s=pd.Series(emp_dict)
print(s)
```

```
2018    50000
2019    60000
2020    75000
dtype: int64
naresh     5000
suresh     6000
kishore    9000
dtype: int64
```

Creating Series using Scalar values

If the series is created using scalar values we must define index. This index defines the length of series.

```
s=pd.Series(15,index=[0,1,2,3,4])
print(s)
```

```
0    15
1    15
2    15
3    15
4    15
dtype: int64
```

Accessing Data from Series

Series is index based collection, we can read and manipulate data using index. This index starts with 0.

```
s1=pd.Series([100,200,300,400,500])
print(s1)
print(s1[0],s1[1],s1[2],s1[3],s1[4])
s2=pd.Series([1000,2000,3000,4000,5000],index=['a','b','c','d','e'])
print(s2['a'],s2['b'],s2['c'],s2['d'],s2['e'])
print(s2[0],s2[1],s2[2],s2[3],s2[4])
```

```
0    100
1    200
2    300
3    400
4    500
dtype: int64
100 200 300 400 500
1000 2000 3000 4000 5000
1000 2000 3000 4000 5000
```

Reading multiple elements/values from series

Series allows reading multiple elements by defining index labels within list.

```
s1=pd.Series(range(100,1000,100))
print(s1)
print(s1[[0,3,6,8]])
s2=pd.Series([100,200,300,400,500],index=['a','b','c','d','e'])
print(s2)
print(s2[['a','c','e']])
```

```
0    100
1    200
2    300
3    400
4    500
5    600
6    700
7    800
8    900
dtype: int64
0    100
3    400
6    700
8    900
dtype: int64
a    100
b    200
c    300
d    400
e    500
```

0s completed at 7:01 PM

Series allows slicing, to read multiple elements/values.

```
s1=pd.Series(range(100,1000,100))
print(s1)
print(s1[:3])
print(s1[-3:])
print(s1[-1::-1])
```

```
0    100
1    200
2    300
3    400
4    500
5    600
6    700
7    800
8    900
dtype: int64
0    100
1    200
2    300
dtype: int64
6    700
7    800
8    900
dtype: int64
8    900
7    800
6    700
5    600
```

DataFrame

DataFrame is two dimensional array object with heterogeneous data. In DataFrame data is stored in the form of rows and columns.

DataFrame represents a table in database.

How to create DataFrame?

DataFrame can be created in different ways.

1. Series
2. Lists
3. Dictionary
4. Numpy array
5. From another dataframe
6. Data can read from files or database

"DataFrame" is type or class name, to create dataframe object

Syntax:

DataFrame(data,index,columns,dtype)

data : data is taken from various sources
Index : row labels
columns : columns labels
dtype: data type of each column

Creating empty dataframe

```
import pandas as pd
#creating empty dataframe
df=pd.DataFrame()
print(df)
```

```
Empty DataFrame
Columns: []
Index: []
```

Creating DataFrame using dictionary

Dictionary consist of key and values.
Dictionary keys as columns headers and values are columns values

```
d={'empno':[1,2,3,4,5],'ename':['naresh','suresh','rajesh','kishore','raman'],'sal':[5000,6000,7000,9000,6000]}
df=pd.DataFrame(d)
print(df)
```

```
   empno    ename   sal
0      1   naresh  5000
1      2   suresh  6000
2      3   rajesh  7000
3      4  kishore  9000
4      5    raman  6000
```

Create DataFrame using List

A nested list represents the content of dataframe.
Each list within list is represented as row.

```
person_list=[['naresh',50],['suresh',45],['kishore',35]]
df=pd.DataFrame(person_list,columns=['name','age'],dtype=float)
print(df)
```

```
      name   age
0   naresh  50.0
1   suresh  45.0
2  kishore  35.0
```

DataFrame created with missing data

Missing data is identified with NaN(Not a Number)

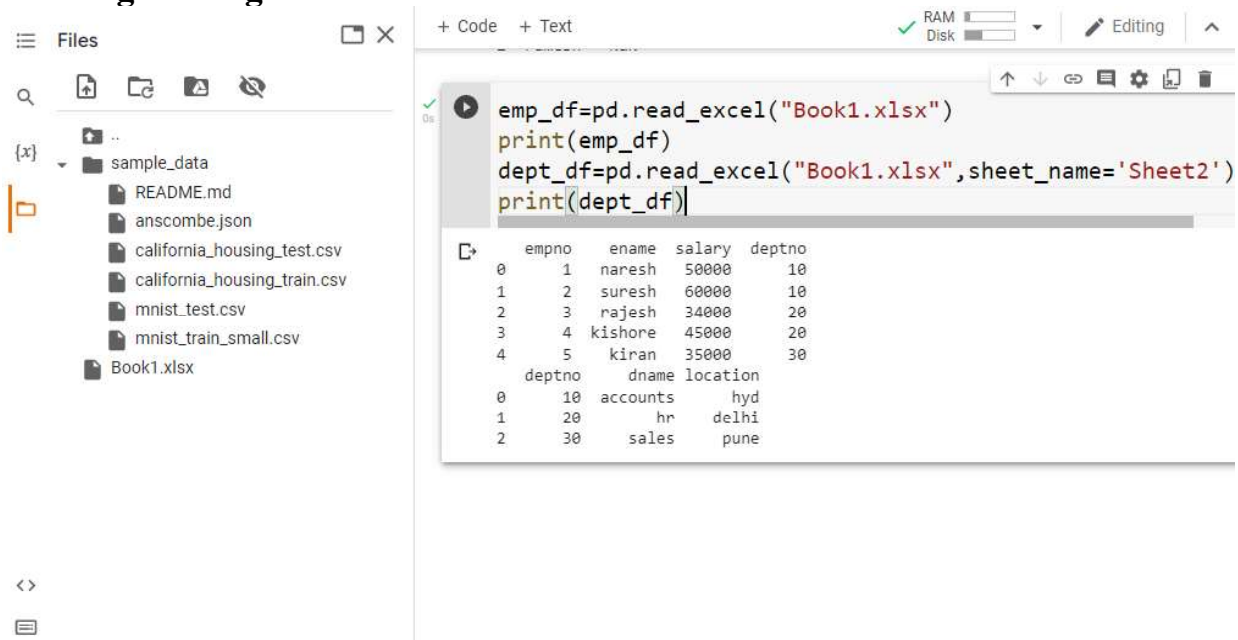
```
data=[['naresh',45],['suresh',56],['kishore',65],['rajesh']]
df=pd.DataFrame(data,columns=['name','age'])
print(df)
```

```
      name   age
0   naresh  45.0
1   suresh  56.0
2  kishore  65.0
3   rajesh   NaN
```

```
data=[{'name':'naresh','age':45},{'name':'kishore'},{'name':'suresh'},{'age':50},{}]
df=pd.DataFrame(data,index=['p1','p2','p3','p4','p5'])
print(df)
```

```
      name   age
p1   naresh  45.0
p2  kishore   NaN
p3   suresh   NaN
p4      NaN  50.0
p5      NaN   NaN
```

Reading/loading data from ms-excel



Selecting Data

1. Row Selection

2. Column Selection

Column Selection

Selecting columns from DataFrame can be done using column header.

```
data=[{'name':'naresh','age':45},{'name':'kishore'},{'name':'suresh'},{'age':50},{}]
df=pd.DataFrame(data)
print(df)
c1=df['name']
c2=df['age']
print(type(c1),type(c2))
print(c1,c2)
```

```
      name   age
0   naresh  45.0
1  kishore   NaN
2   suresh   NaN
3      NaN  50.0
4      NaN   NaN
<class 'pandas.core.series.Series'> <class 'pandas.core.series.Series'>
0     naresh
1    kishore
2     suresh
3        NaN
4        NaN
Name: name, dtype: object 0    45.0
1     NaN
2     NaN
3    50.0
4     NaN
Name: age, dtype: float64
```

Reading multiple columns from DataFrame

In order to read multiple columns, the column names must be defined as a list. It return multiple columns as a dataframe.

single column it read as a series.

```
data={'a':[1,2,3,4,5],'b':[100,200,300,400,500],'c':[1000,2000,3000,4000,5000],'d':[10000,20000,30000,40000,5000]}
df=pd.DataFrame(data)
print(df)
print(df[['a','c']])
r=df[['a','c']]
print(r)
print(type(r))
```

```
   a    b     c      d
0  1  100  1000  10000
1  2  200  2000  20000
2  3  300  3000  30000
3  4  400  4000  40000
4  5  500  5000   5000
   a     c
0  1  1000
1  2  2000
2  3  3000
3  4  4000
4  5  5000
   a     c
0  1  1000
1  2  2000
2  3  3000
3  4  4000
4  5  5000
<class 'pandas.core.frame.DataFrame'>
```

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.

Types of Matplotlib

Matplotlib comes with a wide variety of plots. Plots help to understand trends, and patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. Some of the sample plots are covered here.

- Matplotlib Line Plot
- Matplotlib Bar Plot
- Matplotlib Histograms Plot
- Matplotlib Scatter Plot
- Matplotlib Pie Charts
- Matplotlib Area Plot

Import Matplotlib

```
import matplotlib
```

Checking Matplotlib Version

```
import matplotlib

print(matplotlib.__version__)

import matplotlib.pyplot as plt
```

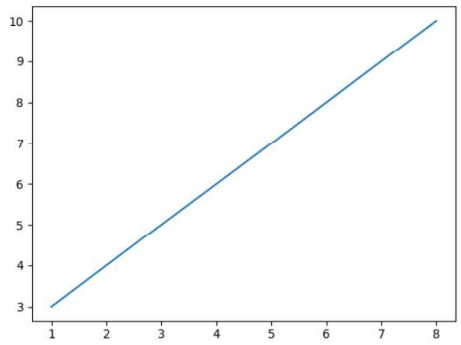
LinePlot

Draw a line in a diagram from position (0,0) to position (6,250):

```
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()
```

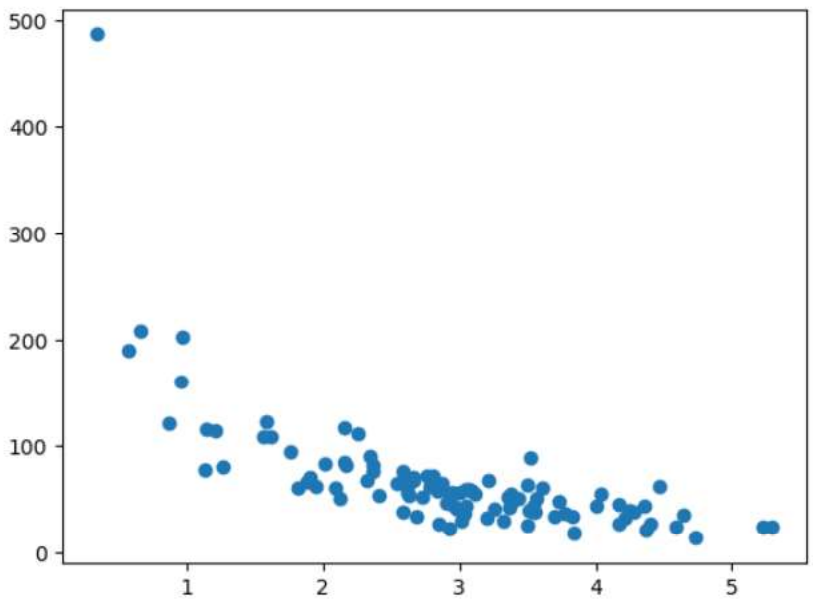


SCATTER PLOT

```
import numpy
import matplotlib.pyplot as plt
numpy.random.seed(2)

x = numpy.random.normal(3, 1, 100)
y = numpy.random.normal(150, 40, 100) / x

plt.scatter(x, y)
plt.show()
```



BOXPLOT

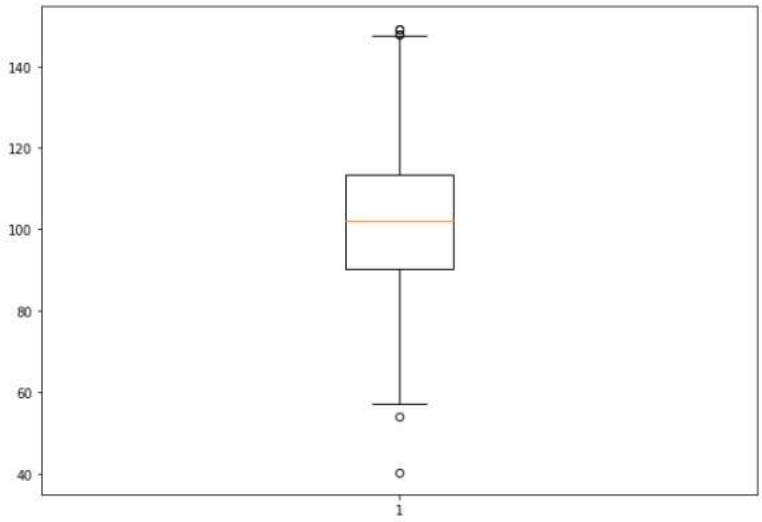
```
# Import libraries
import matplotlib.pyplot as plt
import numpy as np


# Creating dataset
np.random.seed(10)
data = np.random.normal(100, 20, 200)

fig = plt.figure(figsize =(10, 7))

# Creating plot
plt.boxplot(data)

# show plot
plt.show()
```



```
# Import libraries
import matplotlib.pyplot as plt
import numpy as np


# Creating dataset
np.random.seed(10)

data_1 = np.random.normal(100, 10, 200)
data_2 = np.random.normal(90, 20, 200)
data_3 = np.random.normal(80, 30, 200)
data_4 = np.random.normal(70, 40, 200)
data = [data_1, data_2, data_3, data_4]

fig = plt.figure(figsize =(10, 7))

# Creating axes instance
ax = fig.add_axes([0, 0, 1, 1])

# Creating plot
bp = ax.boxplot(data)

# show plot
plt.show()
```

Output: