SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE CODE: DJ19ITL406 **DATE**: 04-06-2022

COURSE NAME: Programing Laboratory 2 (Python) **CLASS:** SYBTECH

EXPERIMENT NO. 8

CO/LO: CO1, CO2,CO4.

AIM / OBJECTIVE:

Write a Python program to implement data analysis using pandas& matplotlib.

DESCRIPTION OF EXPERIMENT:

Describe how to:

1. Create data frame from excelsheet, csv format
   - Read excel files using read_excel() method in pandas.
   - Read csv files using read_csv() method in pandas.

2. Create data frame from objects (list, tuple dictionary) and from excel sheet, .csv files.
   - We can create dataframes from objects by using DataFrame() method

3. Data cleaning for missing data or incomplete data
   - Using the isnull() method, we can confirm that both the missing value and "NA" were recognized as missing values. Then we can use methods like fillna(),etc. to fill the missing values

4. Analyzing data by sorting, finding mean quartiles etc.
   - For sorting the data we use sort_values() method in pandas.
   - For finding quartiles we can use quantile() method in numpy.

5. Extracting best and last five records, sort records and data column wise
   - We sort records and data by sort_values() method.
   - For extracting last five best and worst records we can use head(5) and tail(5) methods.

6. Searching for a record by the id.
   - We can use where(df.id=' ') method to search for a record by id in a dataframe.

QUESTIONS: (We have used Jupyter Notebook)

1. Load the students datasheet shared in a data frame and sort the data as per the marks
   - CODE:

```
import pandas as pd
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

```
import xlrd


#creating dataframe from csv and excel file, and json object.
df1=pd.read_csv(r'\Users\aakas\Downloads\student (1).csv')


#sorting the data as per marks in df1
df1=df1.sort_values(by=['mark'],ascending=[False])
df1
```

- OUTPUT:

Out[65]:

	id	name	class	mark	gender
32	33	Kenn Rein	Six	96	female
11	12	Recky	Six	94	female
31	32	Binn Rott	Seven	90	female
10	11	Ronald	Six	89	female
24	25	Giff Tow	Seven	88	male
15	16	Gimmy	Four	88	male
14	15	Tade Row	Four	88	male
13	14	Bigy	Seven	88	female
12	13	Kty	Seven	88	female
34	35	Rows Noump	Six	88	female
30	31	Marry Toeey	Four	88	male
27	28	Rojj Base	Seven	86	female
7	8	Asruid	Five	85	male
1	2	Max Ruin	Three	85	male
26	27	Big Nose	Three	81	female
22	23	Herod	Eight	79	male

25	26	Crelea	Seven	79	male
8	9	Tes Qry	Six	78	male
6	7	My John Rob	Fifth	78	male
23	24	Tiddy Now	Seven	78	male
0	1	John Deo	Four	75	female
17	18	Honny	Five	75	male
20	21	Babby John	Four	69	female
33	34	Gain Toe	Seven	69	male
19	20	Jackly	Nine	65	female
4	5	John Mike	Four	60	female
3	4	Krish Star	Four	60	female
21	22	Reggid	Seven	55	female
9	10	Big John	Four	55	female
28	29	Tess Played	Seven	55	male
5	6	Alex John	Four	55	male
2	3	Arnold	Three	55	male
16	17	Tumyu	Six	54	male
18	19	Tinny	Nine	18	male

2. Extract three best scoring and worse scoring students
   - CODE:

```
# 3 best scoring students
df1[['name','mark']].head(3)
```

   - OUTPUT:

Out[16]:

	name	mark
32	Kenn Rein	96
11	Recky	94
31	Binn Rott	90

   - CODE:

```
# 3 worst scoring students

df1[['name','mark']].tail(3)
```

- OUTPUT:

Out[17]:

	name	mark
2	Arnold	55
16	Tumyu	54
18	Tinny	18

3. Display the number of students scoring 88 marks
   - CODE:

```
df1[['name','mark']][df1.mark==88]
```

   - OUTPUT:

Out[22]:

	name	mark
24	Giff Tow	88
14	Tade Row	88
30	Marry Toeey	88
12	Kty	88
15	Gimmy	88
13	Bigy	88
34	Rows Noump	88

4. Sort the list in descending marks and people having same marks should be sorted in ascending order of names.
   - CODE:

```
df1.sort_values(by=['mark','name'],ascending=[False,True])
```

- OUTPUT:

Out[27]:

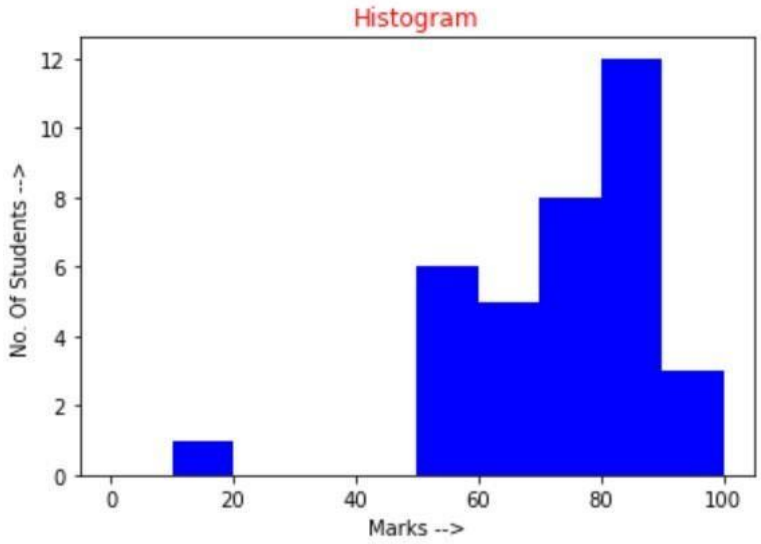
	id	name	class	mark	gender
32	33	Kenn Rein	Six	96	female
11	12	Recky	Six	94	female
31	32	Binn Rott	Seven	90	female
10	11	Ronald	Six	89	female
13	14	Bigy	Seven	88	female
24	25	Giff Tow	Seven	88	male
15	16	Gimmy	Four	88	male
12	13	Kty	Seven	88	female
30	31	Marry Toeey	Four	88	male
34	35	Rows Noump	Six	88	female
14	15	Tade Row	Four	88	male
27	28	Rojj Base	Seven	86	female
7	8	Asruid	Five	85	male
1	2	Max Ruin	Three	85	male
26	27	Big Nose	Three	81	female
25	26	Crelea	Seven	79	male
29	30	Reppy Red	Six	79	female
6	7	My John Rob	Fifth	78	male
8	9	Tes Qry	Six	78	male
23	24	Tiddy Now	Seven	78	male
17	18	Honny	Five	75	male
0	1	John Deo	Four	75	female
20	21	Babby John	Four	69	female
33	34	Gain Toe	Seven	69	male
19	20	Jackly	Nine	65	female
4	5	John Mike	Four	60	female
3	4	Krish Star	Four	60	female
5	6	Alex John	Four	55	male
2	3	Arnold	Three	55	male
9	10	Big John	Four	55	female
21	22	Reggid	Seven	55	female
28	29	Tess Played	Seven	55	male
16	17	Tumyu	Six	54	male
18	19	Tinny	Nine	18	male

5. Generate a histogram of marks scored by students.[give different bin size]
   - CODE:

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

```
#plotting of histogram
x=plt.hist(df1.mark, bins=[0,10,20,30,40,50,60,70,80,90,100],facecolor='b')

plt.title('Histogram',color='r')
plt.xlabel('Marks -->')
plt.ylabel('No. Of Students -->')

plt.show()
```

- OUTPUT:

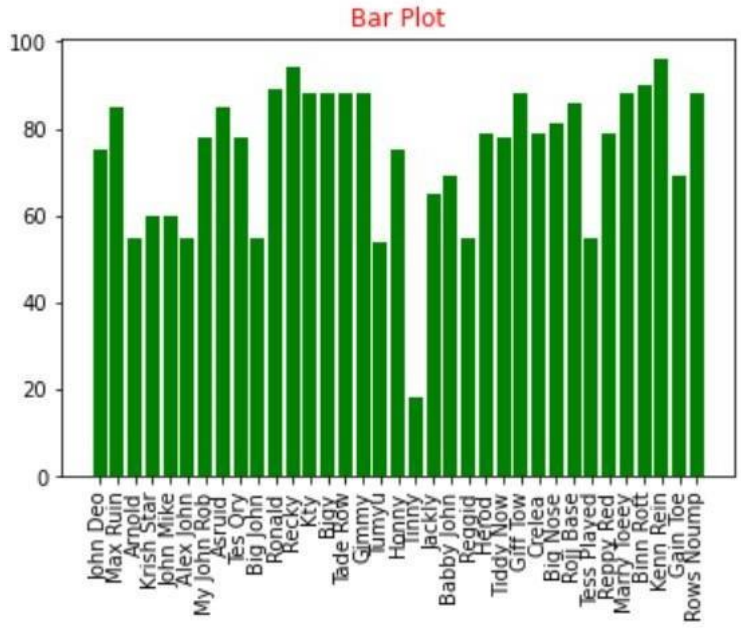


6. Generate a bar plot of student marks
   - CODE:

```
import pandas as pd
import xlrd
import numpy as np
import matplotlib.pyplot as plt

df1=pd.read_csv(r'\Users\aakas\Downloads\student (1).csv')

plt.bar(df1.name,df1.mark, facecolor='g')
plt.tick_params(axis='x', labelrotation = 90)
plt.title("Bar Plot",color='r')
plt.show()
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

- OUTPUT:

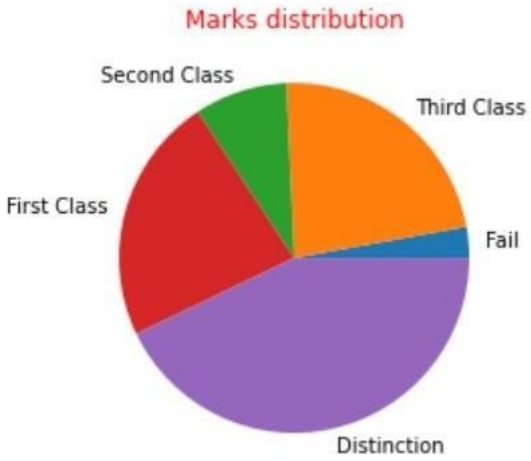


7. Generate a pie chart of marks distribution
   - CODE:

```
list1 = [0,0,0,0,0]

for i in df1.mark:
  if i<50:
    list1[0] = list1[0] + 1
  if i>50 and i<=60:
    list1[1] = list1[1] + 1
  if i>60 and i<=70:
    list1[2] = list1[2] + 1
  if i>70 and i<=80:
    list1[3] = list1[3] + 1
  if i>80 and i<=100:
    list1[4] = list1[4] + 1
print(list1)

labels = ['Fail','Third Class','Second Class','First Class','Distinction']

plt.pie(list1, labels=labels)
plt.title("Marks distribution",color='r')
plt.show()
```

- OUTPUT:

```
[1, 8, 3, 8, 15]
```



OBSERVATIONS / DISCUSSION OF RESULT:

We saw above by using pandas and matplotlib in python, we were able to create dataframes and perform operations on them, and also plot the values to the graphs.

The results showed that we were able to retrieve different types of data easily and plot them to graphs. Just by importing pandas and matplotlib, our work to analyze the data and visualize them became easier.

Thus this experiment demonstrated the implementation of pandas and matplotlib and various ways to achieve desired results.

CONCLUSION:

Thus by using pandas, we can work with relational or labeled data both easily and intuitively. It provide various data structures and operations for manipulating numerical data and time series.

By using matplotlib we can visualize the data. It allows us visual access to huge amounts of data in easily digestible visuals. It comes with a variety of plots like line, bar, scatter, histogram, etc. which helps us to understand trends, patterns and make co-relations.

REFERENCES:

Website References:

[1] https://www.w3schools.com/python

[2] https://www.geeksforgeeks.org/python-introduction-matplotlib/

[3] https://www.geeksforgeeks.org/introduction-to-pandas-in-python/#:~:text=Pandas%20is%20an%20open%2Dsource,top%20of%20the%20NumPy%20library.