MACHINE LEARNING
ASSIGNMENT - 2

Name:Thandarupalli Naveen Goud
UCM ID:nxt46830
700#:700734683

`

Video
link:https://drive.google.com/file/d/1XM34XnVEe-7FRMR31QD9mqWOQcI5HZ_X/view?usp=sharing

1. Numpy:
Using NumPy create random vector of size 15 having only Integers in the range 1-20. 1. Reshape the array to 3 by 5
2. Print array shape.
 3. Replace the max in each row by 0

```
1   import numpy as np
2   array=np.random.randint(1,20,(1,15))
3   print(array)
4   array=np.reshape(array,(3,5))
5   print(array)
6   for i in array:
7       i[np.where(i==i.max())]=0
8   print(array)
```

```
[[19 14 15  6 19  3 10 17  9  9  3  1  2 13  1]]
[[19 14 15  6 19]
 [ 3 10 17  9  9]
 [ 3  1  2 13  1]]
[[ 0 14 15  6  0]
 [ 3 10  0  9  9]
 [ 3  1  2  0  1]]
```
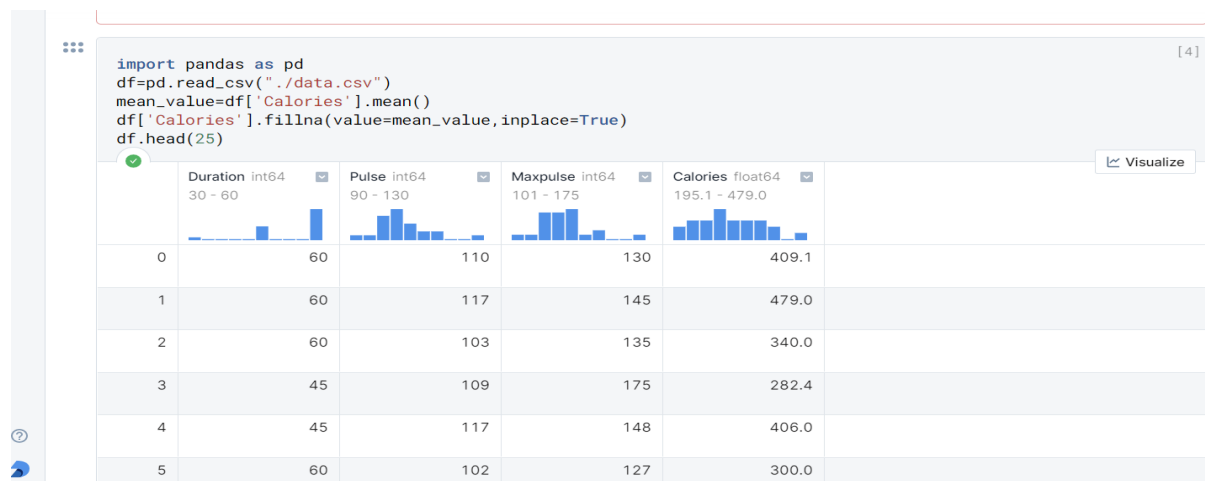
2. Pandas
 1. Read the provided CSV file 'data.csv'.
https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?usp=sharing

2. Show the basic statistical description about the data.

3. Check if the data has null values.

a. Replace the null values with the mean

 4. Select at least two columns and aggregate the data using: min, max, count, mean.

 5. Filter the dataframe to select the rows with calories values between 500 and 1000.

6. Filter the dataframe to select the rows with calories values > 500 and pulse < 100.

7. Create a new "df_modified" dataframe that contains all the columns from df except for "Maxpulse".

8. Delete the "Maxpulse" column from the main df dataframe

 9. Convert the datatype of Calories column to int datatype.

10. Using pandas create a scatter plot for the two columns (Duration and Calories).

[4]

```python
import pandas as pd
df=pd.read_csv("./data.csv")
mean_value=df['Calories'].mean()
df['Calories'].fillna(value=mean_value,inplace=True)
df.head(25)
```

☑ Visualize

	Duration int64 30 - 60	Pulse int64 90 - 130	Maxpulse int64 101 - 175	Calories float64 195.1 - 479.0	
0	60	110	130	409.1	
1	60	117	145	479.0	
2	60	103	135	340.0	
3	45	109	175	282.4	
4	45	117	148	406.0	
5	60	102	127	300.0	

●					▷ Run notebook ⌄ ↻
8	30	109	133	195.1	
9	60	98	124	269.0	

25 rows, showing [10 ⌄] per page   « ‹ Page [1] of 3 › »   ↓

[5]

```python
df.Duration.describe()
```

```
count    169.000000
mean      63.846154
std       42.299949
min       15.000000
25%       45.000000
50%       60.000000
75%       60.000000
max      300.000000
Name: Duration, dtype: float64
```

```
df.Pulse.describe()
```

```
count    169.000000
mean     107.461538
std       14.510259
min       80.000000
25%      100.000000
50%      105.000000
75%      111.000000
max      159.000000
Name: Pulse, dtype: float64
```

[6]

```
df[(df['Calories']>500) & (df['Calories']<1000)]
```

[7]

	Duration int64 80 - 180	Pulse int64 90 - 123	Maxpulse int64 100 - 146	Calories float64 500.3 - 953.2
51	80	123	146	643.1
62	160	109	135	853.0
65	180	90	130	800.4
66	150	105	135	873.4
67	150	107	130	816.0
72	90	100	127	700.0
73	150	97	127	953.2

```
df[(df['Calories']>500 & (df['Pulse']<100))]
```

[8]

	Duration int64 15 - 300	Pulse int64 80 - 159	Maxpulse int64 100 - 184	Calories float64 50.3 - 1860.4
0	60	110	130	409.1
1	60	117	145	479.0
2	60	103	135	340.0
3	45	109	175	282.4
4	45	117	148	406.0
5	60	102	127	300.0
6	60	110	136	374.0
7	45	104	134	253.3

```
df_modified=df.drop("Maxpulse",axis=1)
```
[9]

df_modified [10]

	Duration int64 15 - 300	Pulse int64 80 - 159	Calories float64 50.3 - 1860.4
0	60	110	409.1
1	60	117	479.0
2	60	103	340.0
3	45	109	282.4
4	45	117	406.0
5	60	102	300.0
6	60	110	374.0
7	45	104	253.3

```
df=df.drop("Maxpulse",axis=1)
```
[11]

df [12]

	Duration int64 15 - 300	Pulse int64 80 - 159	Calories float64 50.3 - 1860.4
0	60	110	409.1
1	60	117	479.0
2	60	103	340.0
3	45	109	282.4
4	45	117	406.0
5	60	102	300.0
6	60	110	374.0
7	45	104	253.3

```
df['Calories']=df['Calories'].astype(int)
```

[13]

```
df.dtypes
```
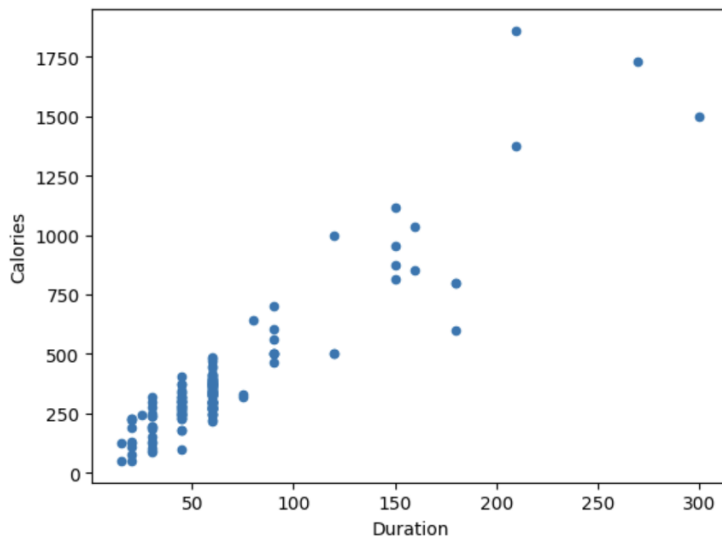
[14]

```
Duration    int64
Pulse       int64
Calories    int64
dtype: object
```

```
df.plot.scatter( x = 'Duration', y = 'Calories')
```

[15]

```
<AxesSubplot:xlabel='Duration', ylabel='Calories'>
```

▷ Run notebook

3. Matplotlib
1. Write a Python programming to create a below chart of the popularity of programming Languages.
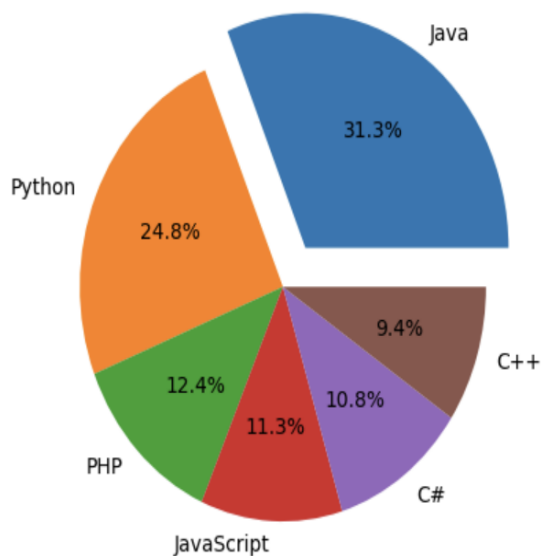 2. Sample data: Programming languages: Java, Python, PHP, JavaScript, C#, C++
Popularity: 22.2, 17.6, 8.8, 8, 7.7, 6.7

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


y = np.array([22.2, 17.6, 8.8, 8, 7.7, 6.7])
mylabels = ["Java", "Python", "PHP", "JavaScript", "C#", "C++"]
myexplode = [0.2, 0, 0, 0,0,0]
plt.pie(y,labels=mylabels, explode = myexplode,autopct='%1.1f%%')
plt.show()
```

```
prgmng_df=pd.DataFrame({"popularity":[22.2, 17.6, 8.8, 8, 7.7, 6.7]},index=['Java','Python','PHP','JavaScript
prgmng_df
```

[16]

	popularity float64
Java	22.2
Python	17.6
PHP	8.8
JavaScript	8.0
C#	7.7
C++	6.7

Visualize

[17]
```
prgmng_df.plot.pie(y='popularity',autopct='%1.1f%%')
```

```
<AxesSubplot:ylabel='popularity'>
```

<AxesSubplot:ylabel='popularity'>