

# Machine Learning Assginment -2

**Name** : Sachit Kumar Tadishetty  
**700#** : 700734682  
**UCM ID** : SXT46820

**VideoLink:**

<https://drive.google.com/drive/folders/1MnDsneFjuKIdMMcHG8EFSIvQiiMWfjTQ?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

SACHIT / 🌟 Welcome / notebook

Run notebook

```
import numpy as np
a=np.random.randint(1,20,15)
a=a.reshape(3,5)
a
```

array([[ 1, 13, 17, 8, 13],  
 [ 1, 11, 12, 8, 18],  
 [13, 9, 12, 7, 19]])

```
a.shape
```

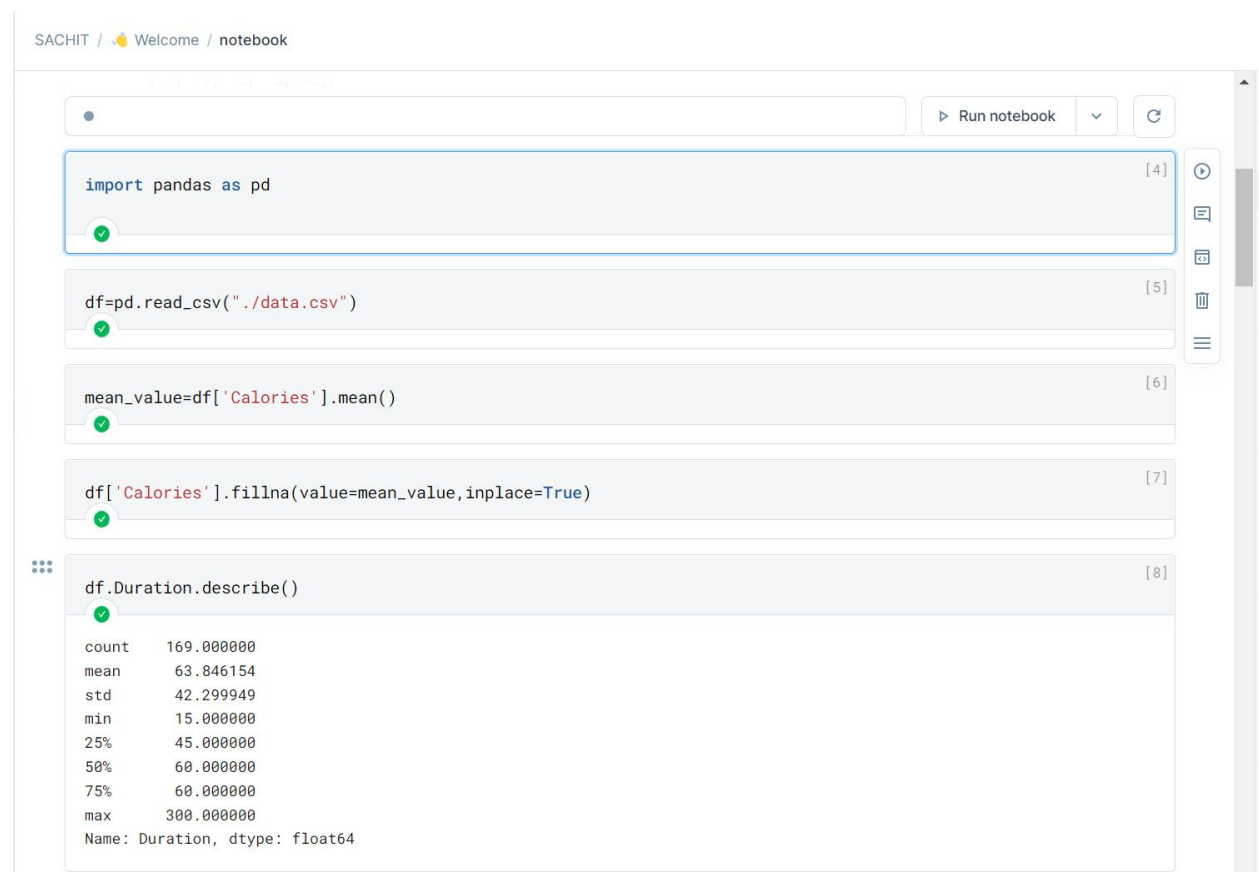
(3, 5)

```
a[np.where(a==np.max(a))]=0
a
```

array([[ 1, 13, 17, 8, 13],  
 [ 1, 11, 12, 8, 18],  
 [13, 9, 12, 7, 0]])

## 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).



```
import pandas as pd
```

```
df=pd.read_csv("./data.csv")
```

```
mean_value=df['Calories'].mean()
```

```
df['Calories'].fillna(value=mean_value,inplace=True)
```

```
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	

Name: Duration, dtype: float64

Run notebook



df.Pulse.describe()

[9]

```
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
```



df[(df['Calories']&gt;500) &amp; (df['Calories']&lt;1000)]

[10]

Visualize

	Duration int64 80 - 180	Pulse int64 90 - 123	Maxpulse int64 100 - 146	Calories float64 500.3 - 953.2
51				643.1
62				853.0
65				800.4
66				873.4
67				816.0
72				700.0
73				953.2
75				563.2

78	120	100	130	500.4
90	180	101	127	600.1

14 rows, showing 10 per page

<< < Page 1 of 2 > >>



df[(df['Calories']&gt;500 &amp; (df['Pulse']&lt;100))]

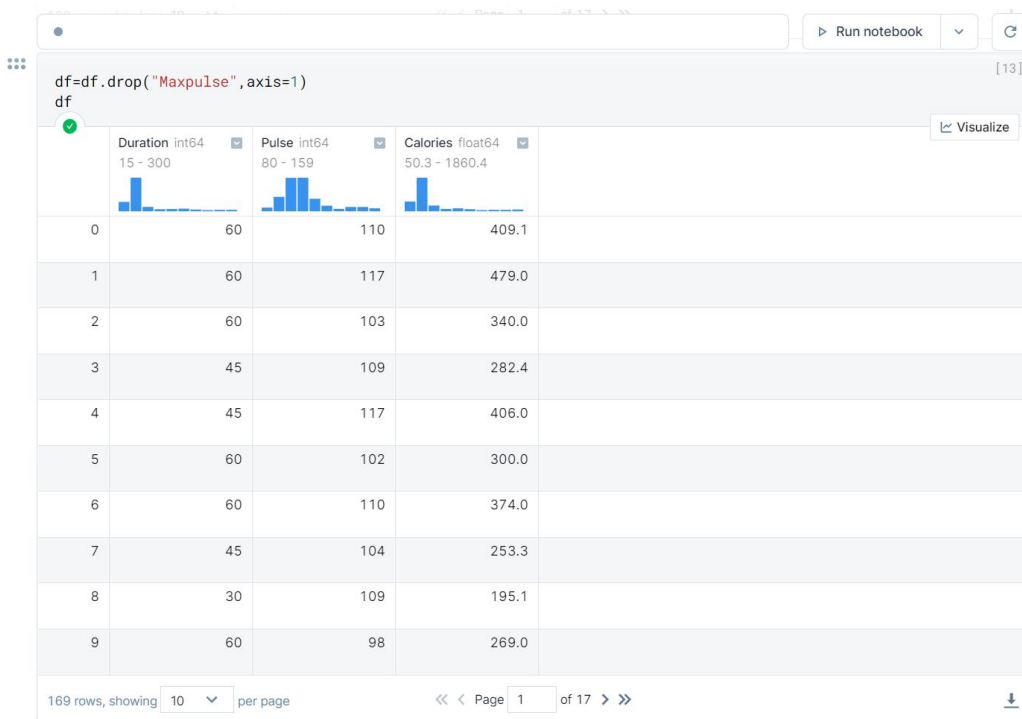
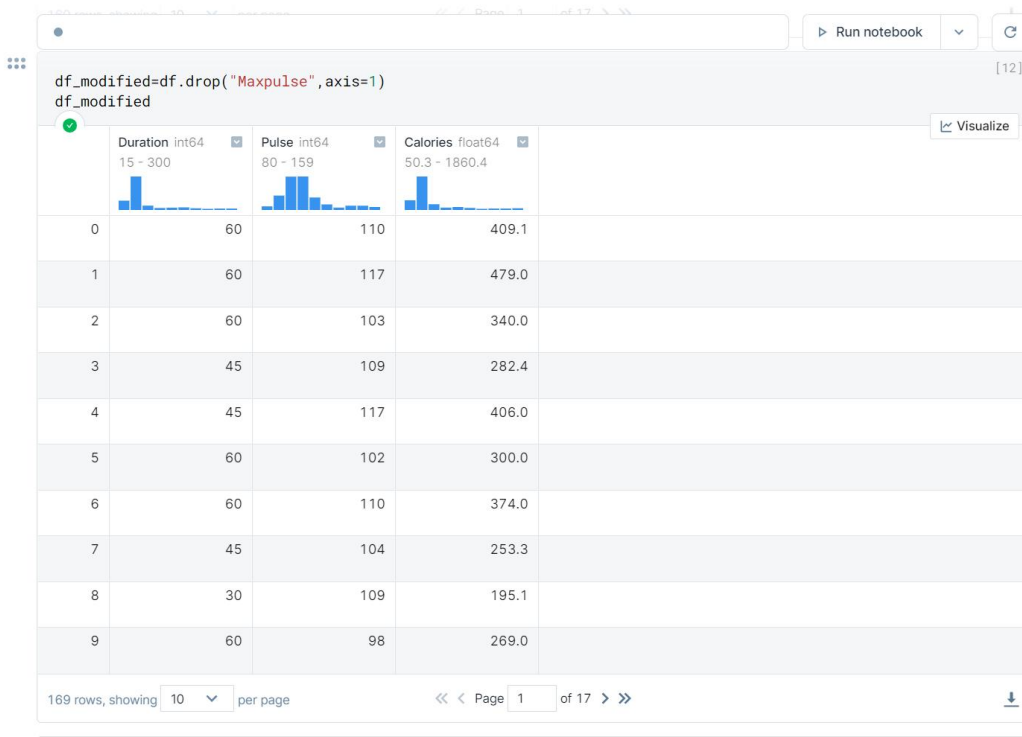
[11]

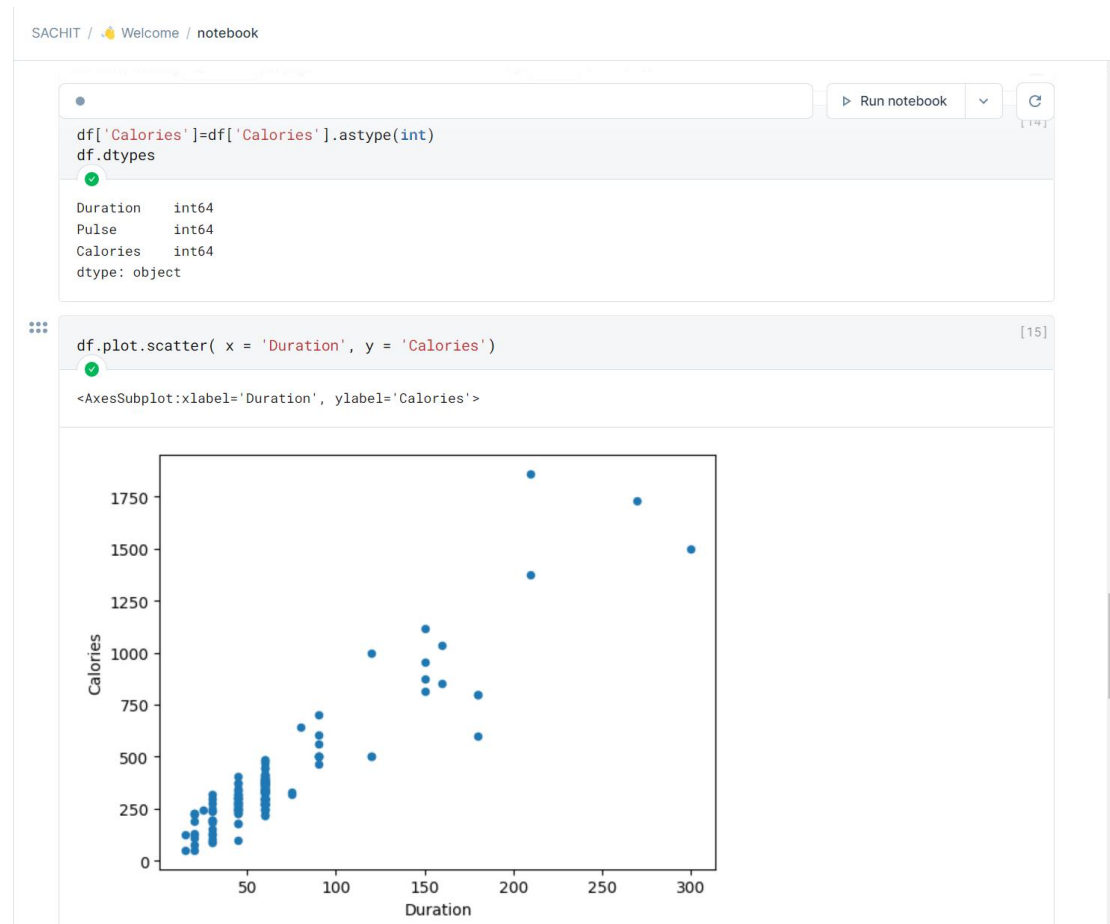
Visualize

	Duration int64 15 - 300	Pulse int64 80 - 159	Maxpulse int64 100 - 184	Calories float64 50.3 - 1860.4
0				409.1
1				479.0
2				340.0
3				282.4
4				406.0
5				300.0
6				374.0
7				253.3
8				195.1
9				269.0

169 rows, showing 10 per page

&lt;&lt; &lt; Page 1 of 17 &gt; &gt;&gt;





### 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

SACHIT / Welcome / notebook Disconnected

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

popularity float64 Visualize

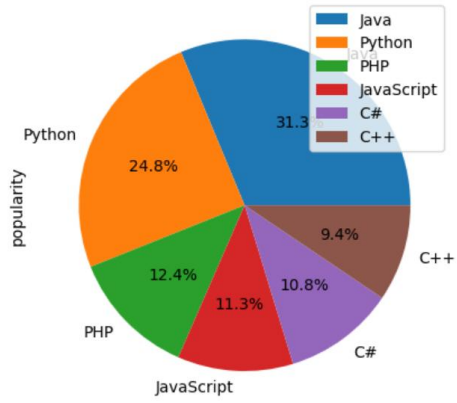
	popularity
Java	22.2
Python	17.6
PHP	8.8
JavaScript	8.0
C#	7.7
C++	6.7

6 rows, showing 10 per page << < Page 1 of 1 > >>

6 rows, showing 1 to 6 per page Page 1 of 1 Run notebook [17]

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

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



<> Code Text SQL Chart Input