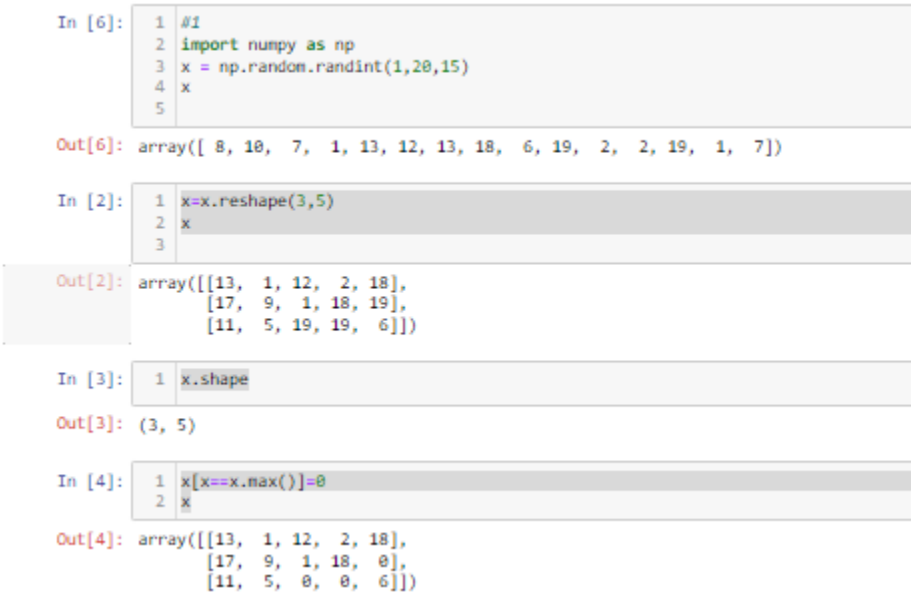
NumPy

1. Used NumPy to create random vector of size 15 having only Integers in the range 1-20.
    1. Reshaped the array to 3 by 5
    2. Printed array shape.
    3. Replaced the max in each row by 0



```
In [6]:
#1
import numpy as np
x = np.random.randint(1,20,15)
x

Out[6]: array([ 8, 10,  7,  1, 13, 12, 13, 18,  6, 19,  2,  2, 19,  1,  7])

In [2]:
x=x.reshape(3,5)
x

Out[2]: array([[13,  1, 12,  2, 18],
               [17,  9,  1, 18, 19],
               [11,  5, 19, 19,  6]])

In [3]:
x.shape

Out[3]: (3, 5)

In [4]:
x[x==x.max()]=0
x

Out[4]: array([[13,  1, 12,  2, 18],
               [17,  9,  1, 18,  0],
               [11,  5,  0,  0,  6]])
```

2. Pandas

1. Read the provided CSV file ‘data.csv’
2. Show the basic statistical description about the data.

In [1]:
```
#2
import pandas as pd
df=pd.read_csv("./data.csv")
df
```

Out[1]:

	Duration	Pulse	Maxpulse	Calories
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
...	...	...	...	...
164	60	105	140	290.8
165	60	110	145	300.0
166	60	115	145	310.2
167	75	120	150	320.4
168	75	125	150	330.4

169 rows × 4 columns

In [2]:
```
mean_value=df['Calories'].mean()
df
```

Out[2]:

	Duration	Pulse	Maxpulse	Calories
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
...	...	...	...	...
164	60	105	140	290.8
165	60	110	145	300.0
166	60	115	145	310.2
167	75	120	150	320.4
168	75	125	150	330.4

169 rows × 4 columns

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.

In [3]:
```
df.isnull().sum()
```

Out[3]:
```
Duration    0
Pulse       0
Maxpulse    0
Calories    5
dtype: int64
```

In [4]:
```
df['Calories'].fillna(value=mean_value,inplace=True)
df
```

Out[4]:

	Duration	Pulse	Maxpulse	Calories
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
...	...	...	...	...
164	60	105	140	290.8
165	60	110	145	300.0
166	60	115	145	310.2
167	75	120	150	320.4
168	75	125	150	330.4

169 rows × 4 columns

In [5]:
```
df.Duration.describe()
```

Out[5]:
```
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
```

In [6]:
```
df.Pulse.describe()
```

Out[6]:
```
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
```

7. Created a new “df_modified” dataframe that contains all the columns from df except for “Maxpulse”.

8. Deleted the “Maxpulse” column from the main df dataframe

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

10. Created a scatter plot for the two columns (Duration and Calories) using Pandas.

In [7]:
```
df[(df['Calories']>500) & (df['Calories']<1000)]
df
```

Out[7]:

	Duration	Pulse	Maxpulse	Calories
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
...	...	...	...	...
164	60	105	140	290.8
165	60	110	145	300.0
166	60	115	145	310.2
167	75	120	150	320.4
168	75	125	150	330.4

169 rows × 4 columns

In [8]:
```
df[(df['Calories']>500 & (df['Pulse']<100))]
df
```

Out[8]:

	Duration	Pulse	Maxpulse	Calories
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
...	...	...	...	...
164	60	105	140	290.8
165	60	110	145	300.0
166	60	115	145	310.2
167	75	120	150	320.4
168	75	125	150	330.4

169 rows × 4 columns

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

Out[9]:

	Duration	Pulse	Calories
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
...	...	...	...
164	60	105	290.8
165	60	110	300.0
166	60	115	310.2
167	75	120	320.4
168	75	125	330.4

169 rows × 3 columns

In [10]:
```
df=df.drop("Maxpulse",axis=1)
df
```

Out[10]:

	Duration	Pulse	Calories
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
...	...	...	...
164	60	105	290.8
165	60	110	300.0
166	60	115	310.2
167	75	120	320.4
168	75	125	330.4

169 rows × 3 columns

In [11]:
```
df["Calories"] = df["Calories"].astype(float).astype(int)
df
```

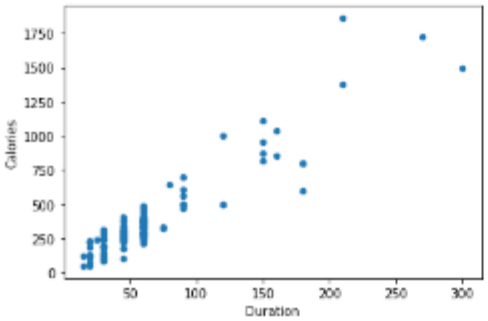
Out[11]:

	Duration	Pulse	Calories
0	60	110	409
1	60	117	479
2	60	103	340
3	45	109	282
4	45	117	406
...	...	...	...
164	60	105	290
165	60	110	300
166	60	115	310
167	75	120	320
168	75	125	330

169 rows × 3 columns

In [13]:
```
df.plot.scatter(x = 'Duration', y = 'Calories')
```

Out[13]: <AxesSubplot:xlabel='Duration', ylabel='Calories'>



3. **Matplotlib**

1. Created the below chart of the popularity of programming Languages using Python programming.

2. Created a piechart using Sample data: Programming languages: Java, Python, PHP, JavaScript, C#, C++
Popularity: 22.2, 17.6, 8.8, 8, 7.7, 6.7

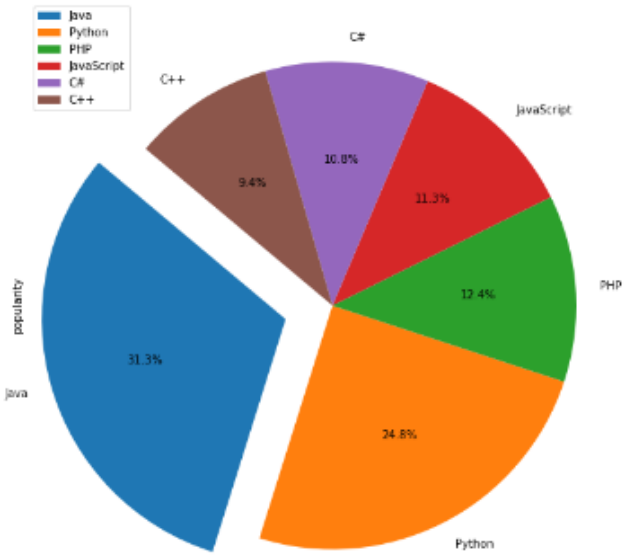
In [16]:
```
#3
prgmng_df=pd.DataFrame({"popularity":[22.2, 17.6, 8.8, 8, 7.7, 6.7]},index=['Java','Python','PHP','JavaScript', 'C#', 'C++'])
prgmng_df.explode('popularity')
prgmng_df
```

Out[16]:

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

In [22]:
```
myexplode = [0.2, 0, 0, 0,0,0]
prgmng_df.plot.pie(y='popularity',autopct='%1.1f%%',explode=myexplode,figsize=[10,10],startangle=140)
```

Out[22]: <AxesSubplot:ylabel='popularity'>



GitHub: https://github.com/Sanjana9791/MachineLearningAssignment2.git

Video Link: https://drive.google.com/file/d/16m0aOlwtW-WfdFSTq-AyaXtOyOaYtI4M/view?usp=sharing