ASSINGMENT 4

Name: varna nemulla

ID: 700744920

Question 1 :

a. Read the provided CSV file ‘data.csv’.

b. https://drive.google.com/drive/folders/1h8C3mLsso-R-sIOLsvoYwPLzy2fJ4IOF?usp=sharing

c. Show the basic statistical description about the data.

d. Check if the data has null values.

i. Replace the null values with the mean

e. Select at least two columns and aggregate the data using: min, max, count, mean. f. Filter the dataframe to select the rows with calories values between 500 and 1000.

g. Filter the dataframe to select the rows with calories values > 500 and pulse < 100. h. Create a new “df_modified” dataframe that contains all the columns from df except for “Maxpulse”.

i. Delete the “Maxpulse” column from the main df dataframe

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

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

```
import numpy as np

import pandas as pd

# 1(a) Import the given "Data.csv"


Data = pd.read_csv('C:/Users/Pavanisodar/Downloads/data.csv')

Data.info()
```

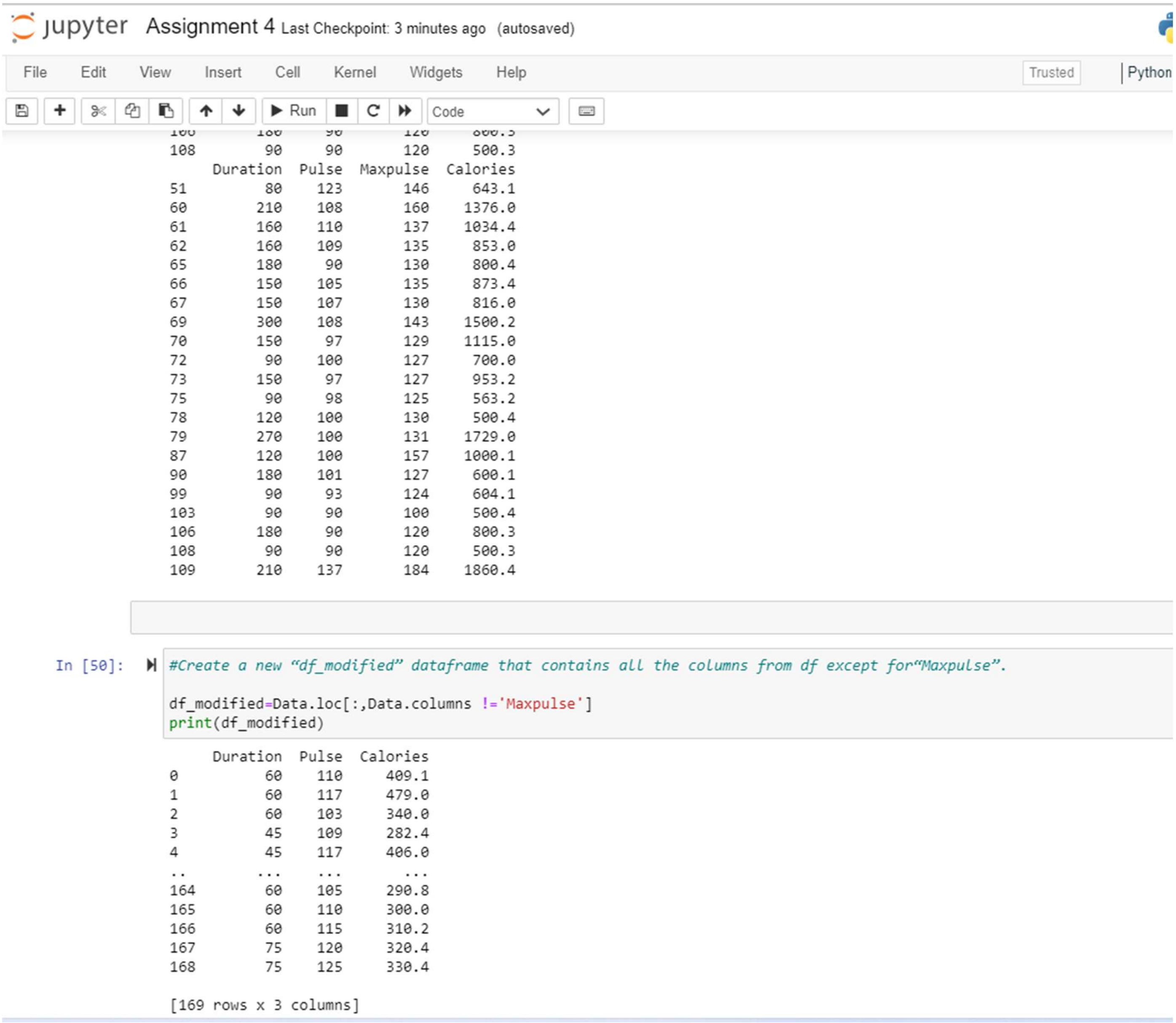
#Show the basic statistical description about the data.

```
Data.head()

#Check if the data has null values.


Data.isnull().any()


Data.fillna(Data.mean(), inplace=True)

Data.isnull().any()

# Replace the null values with the mean

column_means = Data.mean()

print(column_means)

Data=Data.fillna(column_means)

print(Data.head(20))

#(e)select at least two columns and aggregate the data using: min, max, count, mean.

res = Data.agg({'Calories':['mean','min', 'max','count'], 'Pulse': ['mean', 'min','max', 'count']})

print(res)

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


filter_first_Data=Data[(Data['Calories'] >500)&(Data['Calories']<1000)]

print(filter_first_Data)

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

filter_second_Data=Data[(Data['Calories'] >500)&(Data['Pulse']<1000)]
```

```
print(filter_second_Data)

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


df_modified=Data.loc[:,Data.columns !='Maxpulse']

print(df_modified)


#Delete the “Maxpulse” column from the main df dataframe

Data.drop('Maxpulse', inplace=True, axis=1)

print(Data.dtypes)

#Convert the datatype of Calories column to int datatype.


Data["Calories"]=Data["Calories"].astype(float).astype(int)

print(Data.dtypes)

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

as1=Data.plot.scatter(x='Duration', y='Calories')

print(as1)
```

Description: In the first part of the program it is able to read the data from data.csv and it is checking if the data has null values. also Replacing the null values with the mean. Selecting least two columns and aggregate the data using: min, max, count, mean. Filtering the dataframe to select the rows with calories values between 500 and 1000. Filtering the dataframe to select the rows with calories values > 500 and pulse < 100. Then Creating a new “df_modified” dataframe that contains all the columns from df except for “Maxpulse”and Deleting the “Maxpulse” column from the main df dataframe. In the end converting the datatype of Calories column to int datatype. It will create pandas and a scatter plot for the two columns (Duration and Calories).

Screenshot of source code and output:

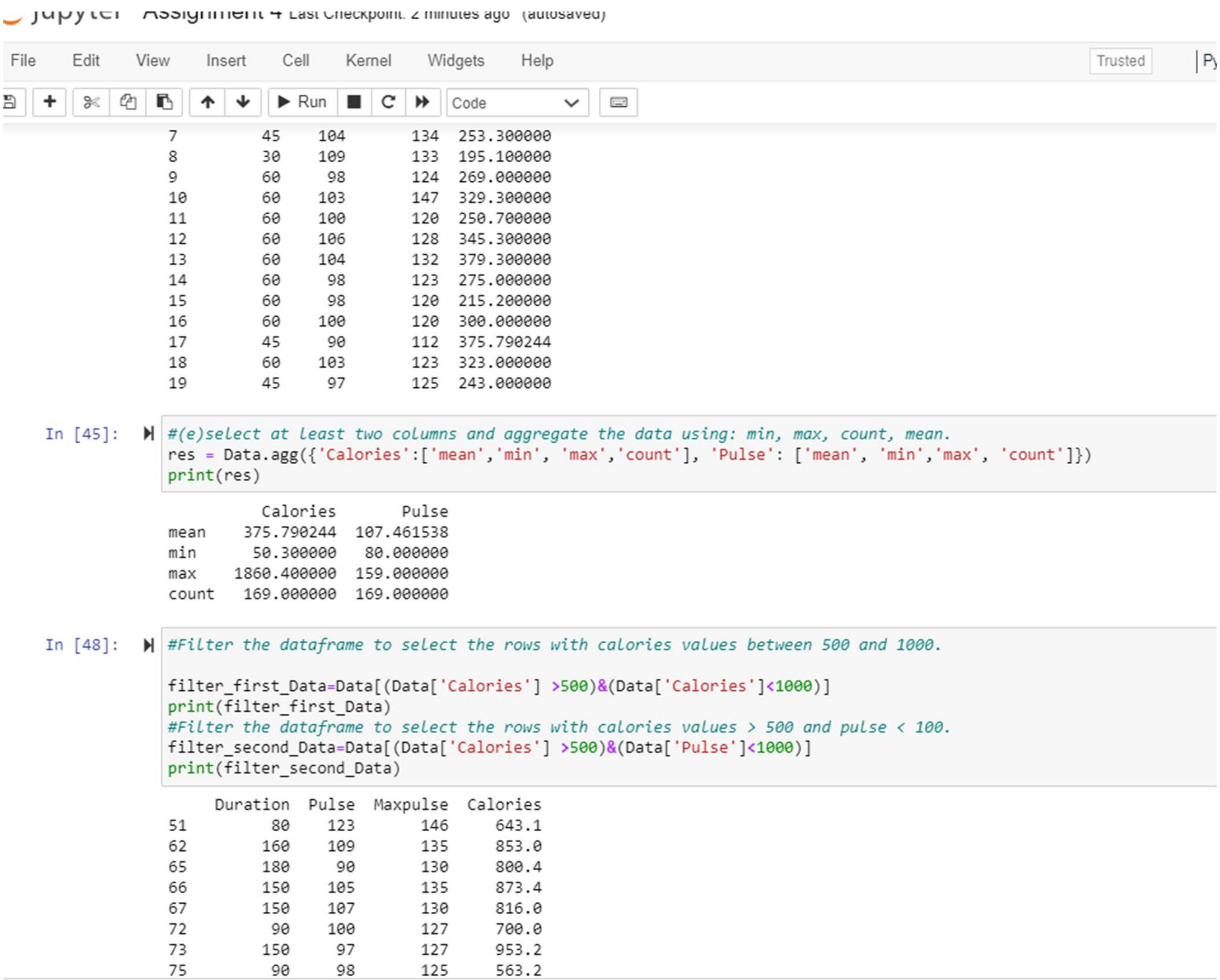


```
106       180     90       120     800.3
108        90     90       120     500.3
     Duration  Pulse  Maxpulse  Calories
51         80    123       146     643.1
60        210    108       160    1376.0
61        160    110       137    1034.4
62        160    109       135     853.0
65        180     90       130     800.4
66        150    105       135     873.4
67        150    107       130     816.0
69        300    108       143    1500.2
70        150     97       129    1115.0
72         90    100       127     700.0
73        150     97       127     953.2
75         90     98       125     563.2
78        120    100       130     500.4
79        270    100       131    1729.0
87        120    100       157    1000.1
90        180    101       127     600.1
99         90     93       124     604.1
103        90     90       100     500.4
106       180     90       120     800.3
108        90     90       120     500.3
109       210    137       184    1860.4
```

In [50]:

```
#Create a new “df_modified” dataframe that contains all the columns from df except for“Maxpulse”.

df_modified=Data.loc[:,Data.columns !='Maxpulse']
print(df_modified)
```

```
     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 x 3 columns]
```


Jupyter Assignment 4 Last Checkpoint: 2 minutes ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Py
Run
Code

```
7         45    104       134  253.300000
8         30    109       133  195.100000
9         60     98       124  269.000000
10        60    103       147  329.300000
11        60    100       120  250.700000
12        60    106       128  345.300000
13        60    104       132  379.300000
14        60     98       123  275.000000
15        60     98       120  215.200000
16        60    100       120  300.000000
17        45     90       112  375.790244
18        60    103       123  323.000000
19        45     97       125  243.000000
```

In [45]:

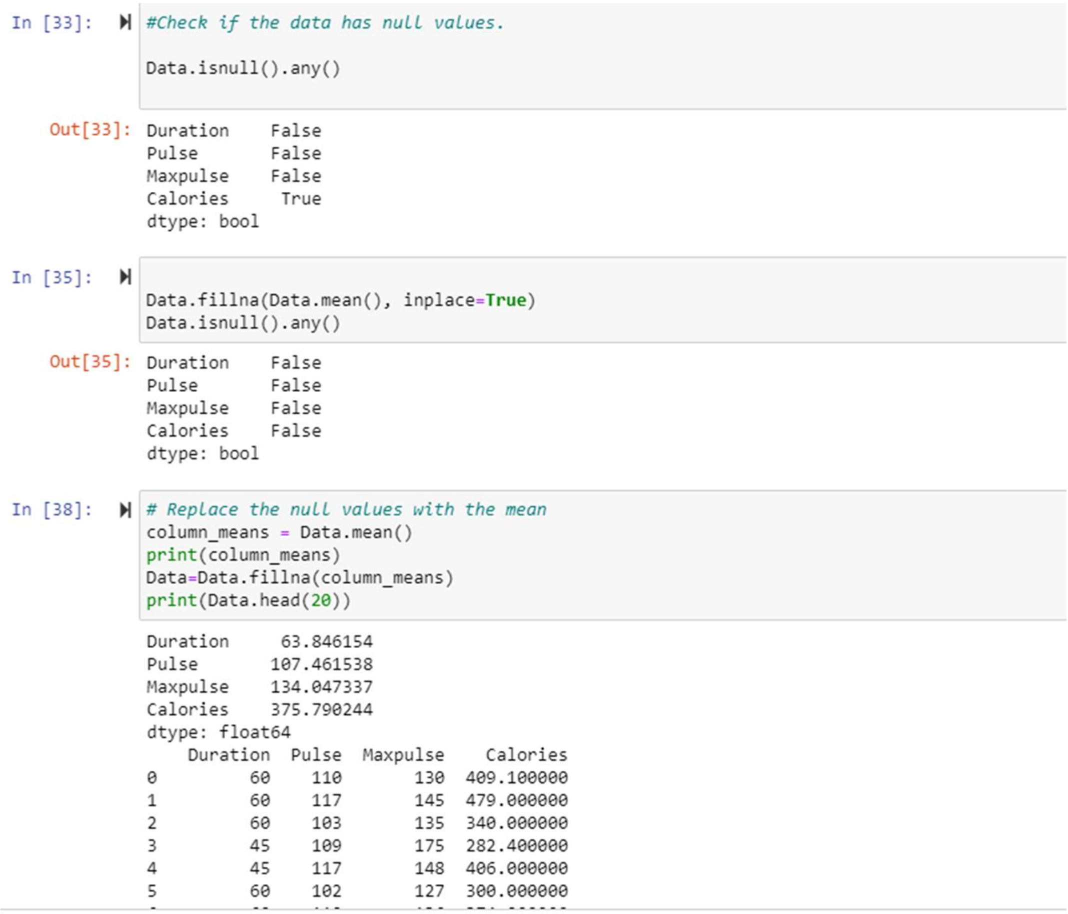
```
#(e)select at least two columns and aggregate the data using: min, max, count, mean.
res = Data.agg({'Calories':['mean','min', 'max','count'], 'Pulse': ['mean', 'min','max', 'count']})
print(res)
```

```
          Calories       Pulse
mean    375.790244  107.461538
min      50.300000   80.000000
max    1860.400000  159.000000
count   169.000000  169.000000
```

In [48]:

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

filter_first_Data=Data[(Data['Calories'] >500)&(Data['Calories']<1000)]
print(filter_first_Data)
#Filter the dataframe to select the rows with calories values > 500 and pulse < 100.
filter_second_Data=Data[(Data['Calories'] >500)&(Data['Pulse']<1000)]
print(filter_second_Data)
```

```
     Duration  Pulse  Maxpulse  Calories
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
75         90     98       125     563.2
```



In [33]:

```
#Check if the data has null values.

Data.isnull().any()
```

Out[33]:

```
Duration    False
Pulse       False
Maxpulse    False
Calories     True
dtype: bool
```

In [35]:

```

Data.fillna(Data.mean(), inplace=True)
Data.isnull().any()
```

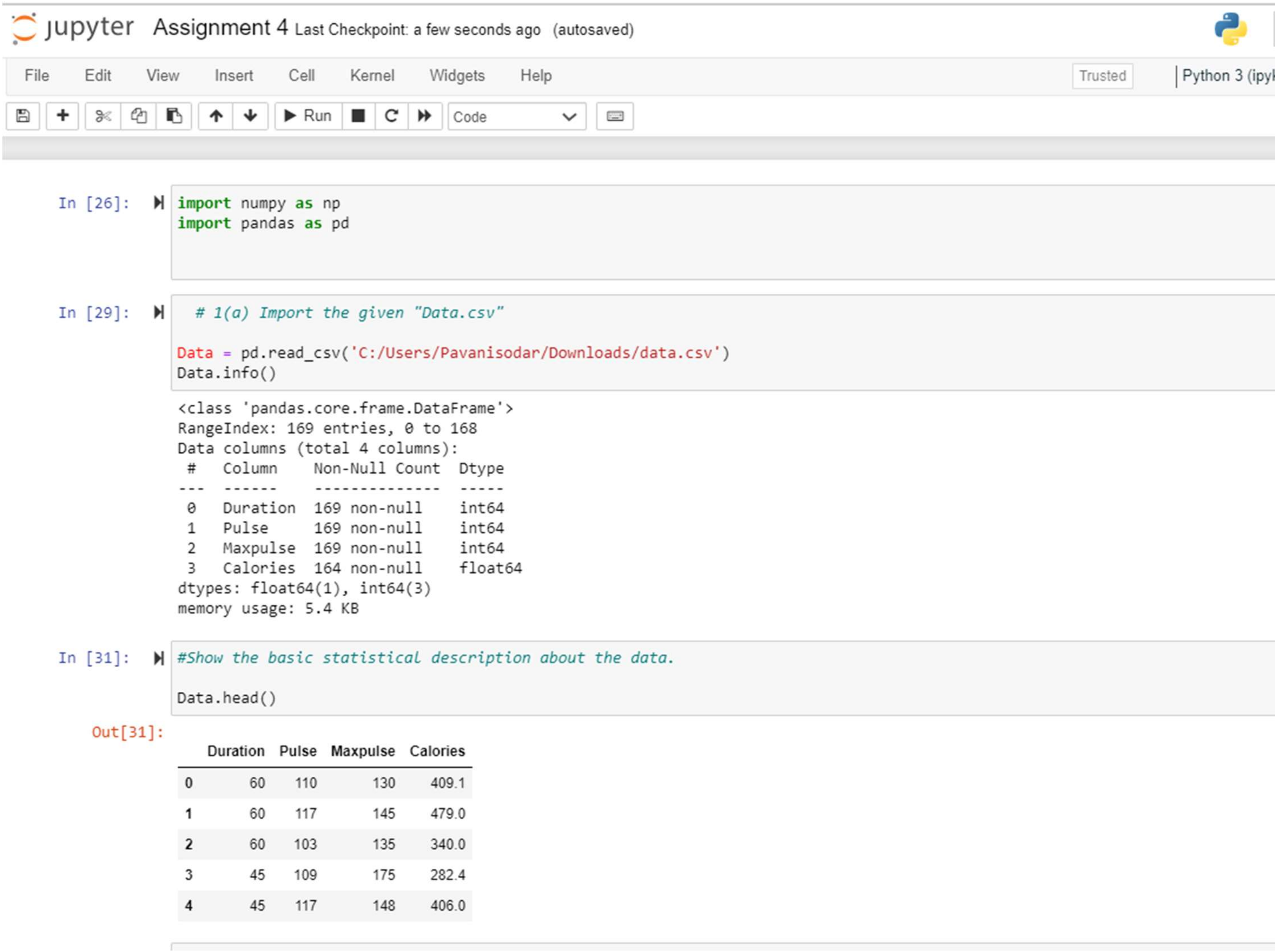
Out[35]:

```
Duration    False
Pulse       False
Maxpulse    False
Calories    False
dtype: bool
```

In [38]:

```
# Replace the null values with the mean
column_means = Data.mean()
print(column_means)
Data=Data.fillna(column_means)
print(Data.head(20))
```

```
Duration     63.846154
Pulse       107.461538
Maxpulse    134.047337
Calories    375.790244
dtype: float64
    Duration  Pulse  Maxpulse    Calories
0         60    110       130  409.100000
1         60    117       145  479.000000
2         60    103       135  340.000000
3         45    109       175  282.400000
4         45    117       148  406.000000
5         60    102       127  300.000000
```



In [26]:

```
import numpy as np
import pandas as pd
```

In [29]:

```
  # 1(a) Import the given "Data.csv"

Data = pd.read_csv('C:/Users/Pavanisodar/Downloads/data.csv')
Data.info()
```

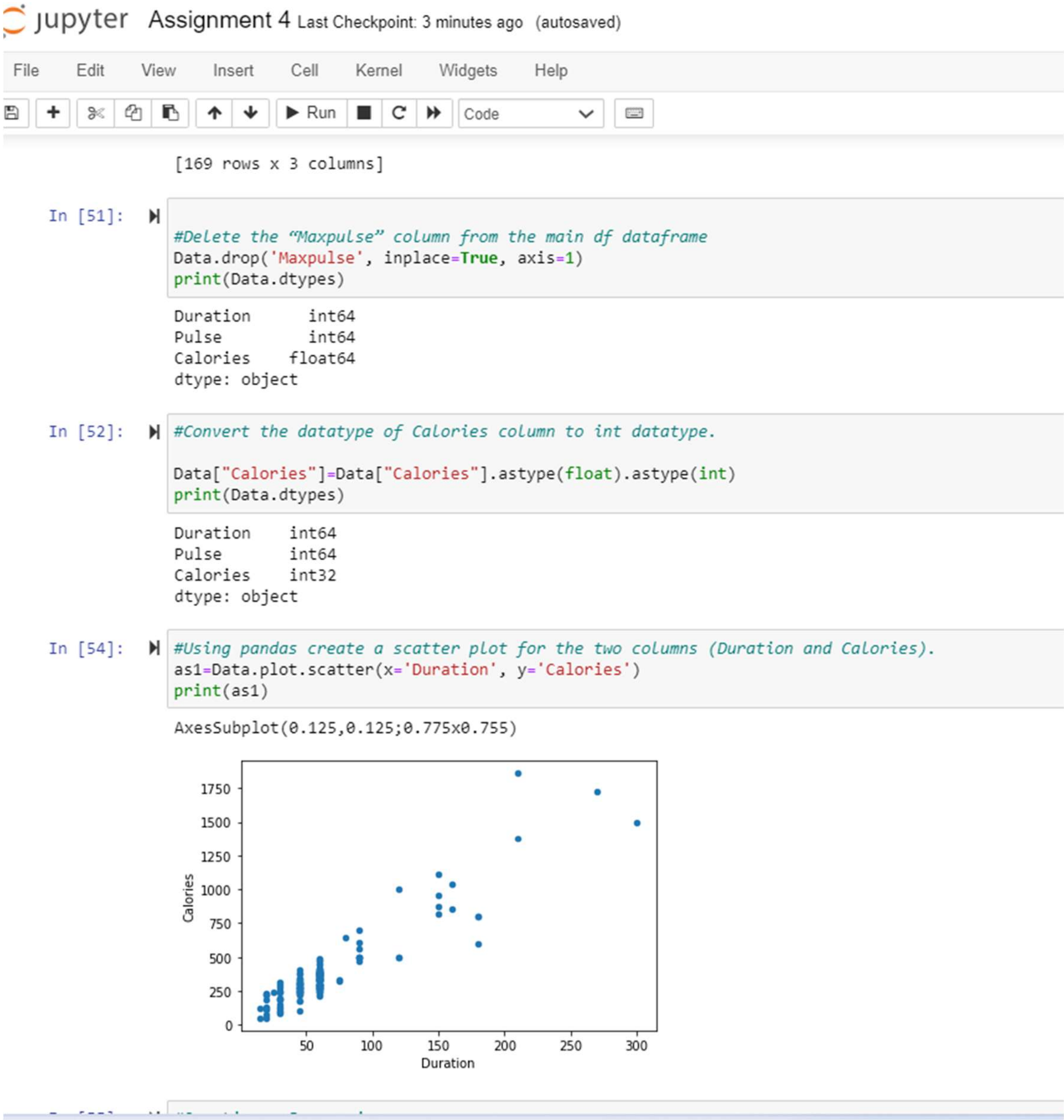
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 169 entries, 0 to 168
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Duration  169 non-null    int64  
 1   Pulse     169 non-null    int64  
 2   Maxpulse  169 non-null    int64  
 3   Calories  164 non-null    float64
dtypes: float64(1), int64(3)
memory usage: 5.4 KB
```

In [31]:

```
#Show the basic statistical description about the data.

Data.head()
```

Out[31]:

	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



```
[169 rows x 3 columns]
```

In [51]:

```
#Delete the "Maxpulse" column from the main df dataframe
Data.drop('Maxpulse', inplace=True, axis=1)
print(Data.dtypes)
```

```
Duration      int64
Pulse         int64
Calories    float64
dtype: object
```

In [52]:

```
#Convert the datatype of Calories column to int datatype.

Data["Calories"]=Data["Calories"].astype(float).astype(int)
print(Data.dtypes)
```

```
Duration    int64
Pulse       int64
Calories    int32
dtype: object
```

In [54]:

```
#Using pandas create a scatter plot for the two columns (Duration and Calories).
as1=Data.plot.scatter(x='Duration', y='Calories')
print(as1)
```

```
AxesSubplot(0.125,0.125;0.775x0.755)
```

Question 2:

2. Linear Regression

a) Import the given "Salary_Data.csv"

b) Split the data in train_test partitions, such that 1/3 of the data is reserved as test subset.

c) Train and predict the model.

d) Calculate the mean_squared error

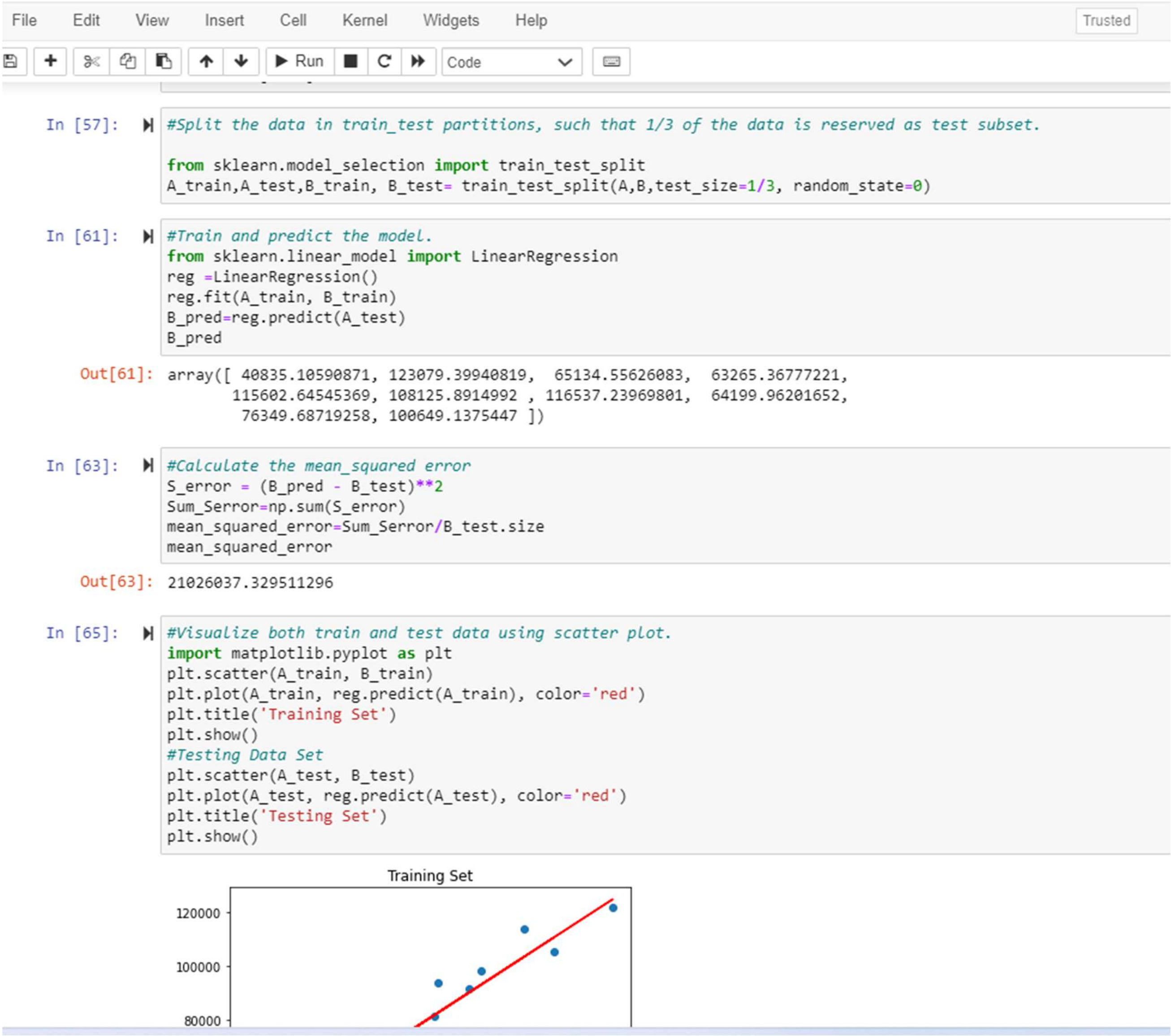
e) Visualize both train and test data using scatter plot.

```
#2 . Linear Regression

#Import the given “Salary_Data.csv”


sal=pd.read_csv('C:/Users/Pavanisodar/Downloads/Salary_Data (2).csv')

sal.info()

sal.head()

A=sal.iloc[:, :-1].values

B=sal.iloc[:, 1].values

#Split the data in train_test partitions, such that 1/3 of the data is reserved as test
subset.


from sklearn.model_selection import train_test_split

A_train,A_test,B_train, B_test= train_test_split(A,B,test_size=1/3,
random_state=0)

#Train and predict the model.

from sklearn.linear_model import LinearRegression

reg =LinearRegression()

reg.fit(A_train, B_train)

B_pred=reg.predict(A_test)

B_pred

#Calculate the mean_squared error

S_error = (B_pred - B_test)**2
```

```
Sum_Serror=np.sum(S_error)

mean_squared_error=Sum_Serror/B_test.size

mean_squared_error

#Visualize both train and test data using scatter plot.

import matplotlib.pyplot as plt

plt.scatter(A_train, B_train)

plt.plot(A_train, reg.predict(A_train), color='red')

plt.title('Training Set')

plt.show()

#Testing Data Set

plt.scatter(A_test, B_test)

plt.plot(A_test, reg.predict(A_test), color='red')

plt.title('Testing Set')

plt.show()
```

Description: Here firstly Importing the given “Salary_Data.csv” and then Spliting the data in train_test partitions so that 1/3 of the data is reserved as test subset. We are Training and predicting the model. After the train and predict part we are Calculating the mean_squared error and Visualizing both train and test data using scatter plot.

Screenshot of source code and output:



In [57]:

```
#Split the data in train_test partitions, such that 1/3 of the data is reserved as test subset.

from sklearn.model_selection import train_test_split
A_train,A_test,B_train, B_test= train_test_split(A,B,test_size=1/3, random_state=0)
```

In [61]:

```
#Train and predict the model.
from sklearn.linear_model import LinearRegression
reg =LinearRegression()
reg.fit(A_train, B_train)
B_pred=reg.predict(A_test)
B_pred
```

Out[61]:

```
array([ 40835.10590871, 123079.39940819,  65134.55626083,  63265.36777221,
       115602.64545369, 108125.8914992 , 116537.23969801,  64199.96201652,
        76349.68719258, 100649.1375447 ])
```

In [63]:

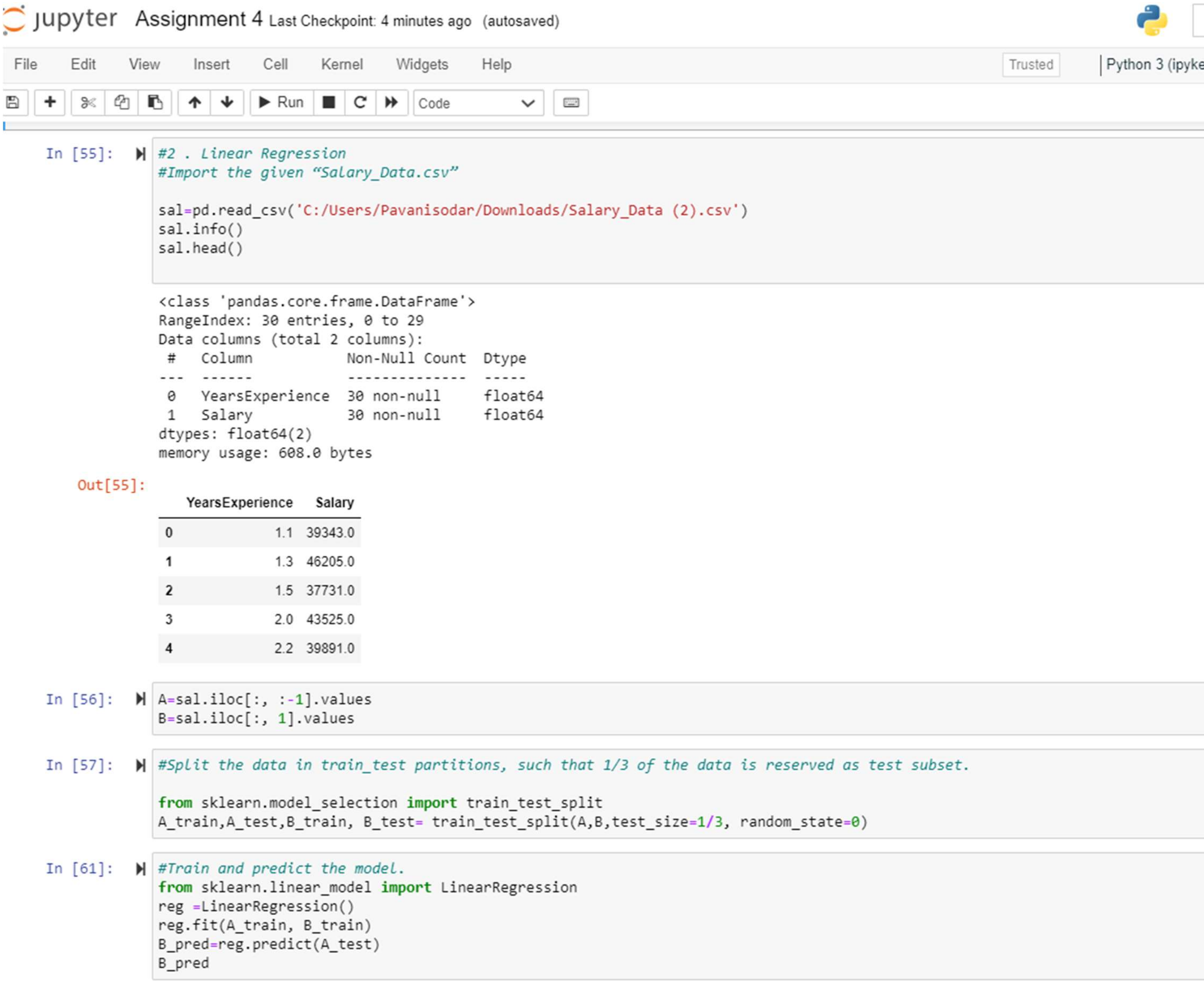
```
#Calculate the mean_squared error
S_error = (B_pred - B_test)**2
Sum_Serror=np.sum(S_error)
mean_squared_error=Sum_Serror/B_test.size
mean_squared_error
```

Out[63]:

```
21026037.329511296
```

In [65]:

```
#Visualize both train and test data using scatter plot.
import matplotlib.pyplot as plt
plt.scatter(A_train, B_train)
plt.plot(A_train, reg.predict(A_train), color='red')
plt.title('Training Set')
plt.show()
#Testing Data Set
plt.scatter(A_test, B_test)
plt.plot(A_test, reg.predict(A_test), color='red')
plt.title('Testing Set')
plt.show()
```



In [55]:

```
#2 . Linear Regression
#Import the given “Salary_Data.csv”

sal=pd.read_csv('C:/Users/Pavanisodar/Downloads/Salary_Data (2).csv')
sal.info()
sal.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   YearsExperience  30 non-null     float64
 1   Salary           30 non-null     float64
dtypes: float64(2)
memory usage: 608.0 bytes
```

Out[55]:

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

In [56]:

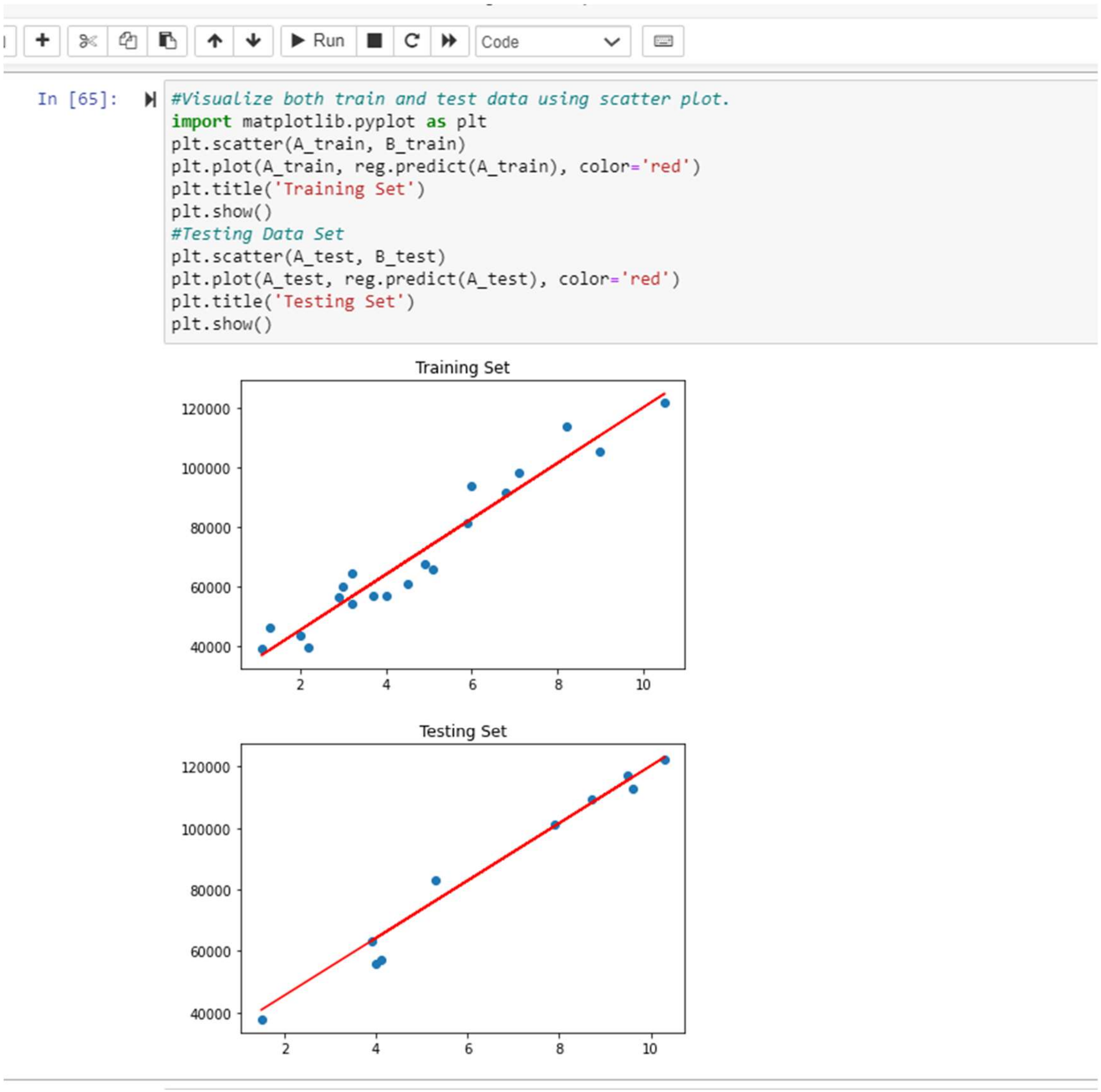
```
A=sal.iloc[:, :-1].values
B=sal.iloc[:, 1].values
```

In [57]:

```
#Split the data in train_test partitions, such that 1/3 of the data is reserved as test subset.

from sklearn.model_selection import train_test_split
A_train,A_test,B_train, B_test= train_test_split(A,B,test_size=1/3, random_state=0)
```

In [61]:

```
#Train and predict the model.
from sklearn.linear_model import LinearRegression
reg =LinearRegression()
reg.fit(A_train, B_train)
B_pred=reg.predict(A_test)
B_pred
```



```
#Visualize both train and test data using scatter plot.
import matplotlib.pyplot as plt
plt.scatter(A_train, B_train)
plt.plot(A_train, reg.predict(A_train), color='red')
plt.title('Training Set')
plt.show()
#Testing Data Set
plt.scatter(A_test, B_test)
plt.plot(A_test, reg.predict(A_test), color='red')
plt.title('Testing Set')
plt.show()
```

Video Link: https://drive.google.com/file/d/1tlLaSqScerXln4o-HKBDIYdJHX5RL-MJ/view?usp=sharing

GitHub Link: