**ASSINGMENT 4**

**Name: BhuvanaNandhimalla**

**ID: 700744203**

**Question 1 :**

1. **Read the provided CSV file ‘data.csv’.**
2. **https://drive.google.com/drive/folders/1h8C3mLsso-R- sIOLsvoYwPLzy2fJ4IOF?usp=sharing**
3. **Show the basic statistical description about the data.**
4. **Check if the data has null values.**

**i. Replace the null values with the mean**

1. **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”.**

1. **Delete the “Maxpulse” column from the main df dataframe**
2. **Convert the datatype of Calories column to int datatype.**
3. **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:**

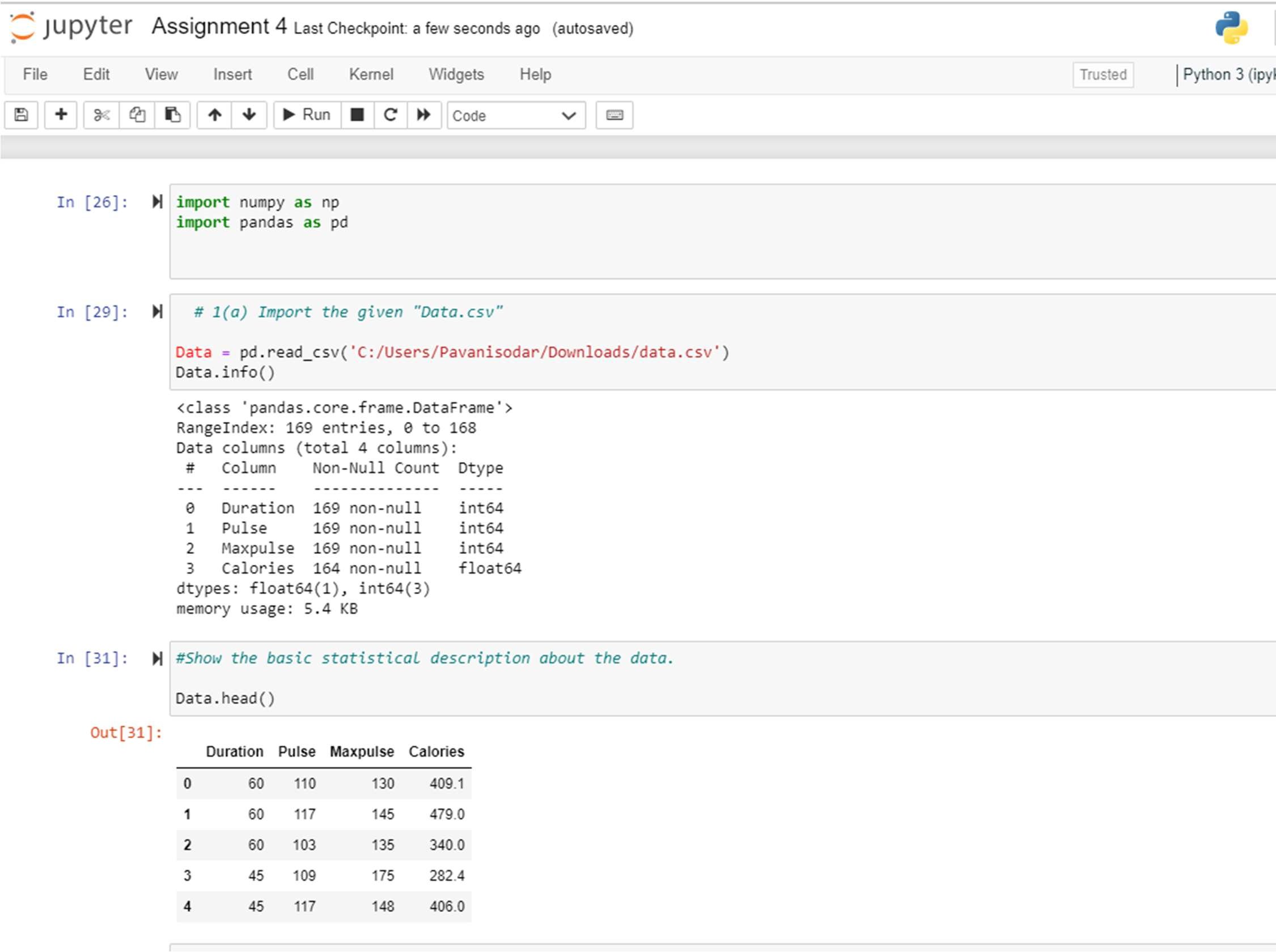
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated







**Question 2:**

1. **Linear Regression**
   1. **Import the given “Salary\_Data.csv”**
   2. **Split the data in train\_test partitions, such that 1/3 of the data is reserved as test subset.**
   3. **Train and predict the model.**
   4. **Calculate the mean\_squared error**
   5. **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')

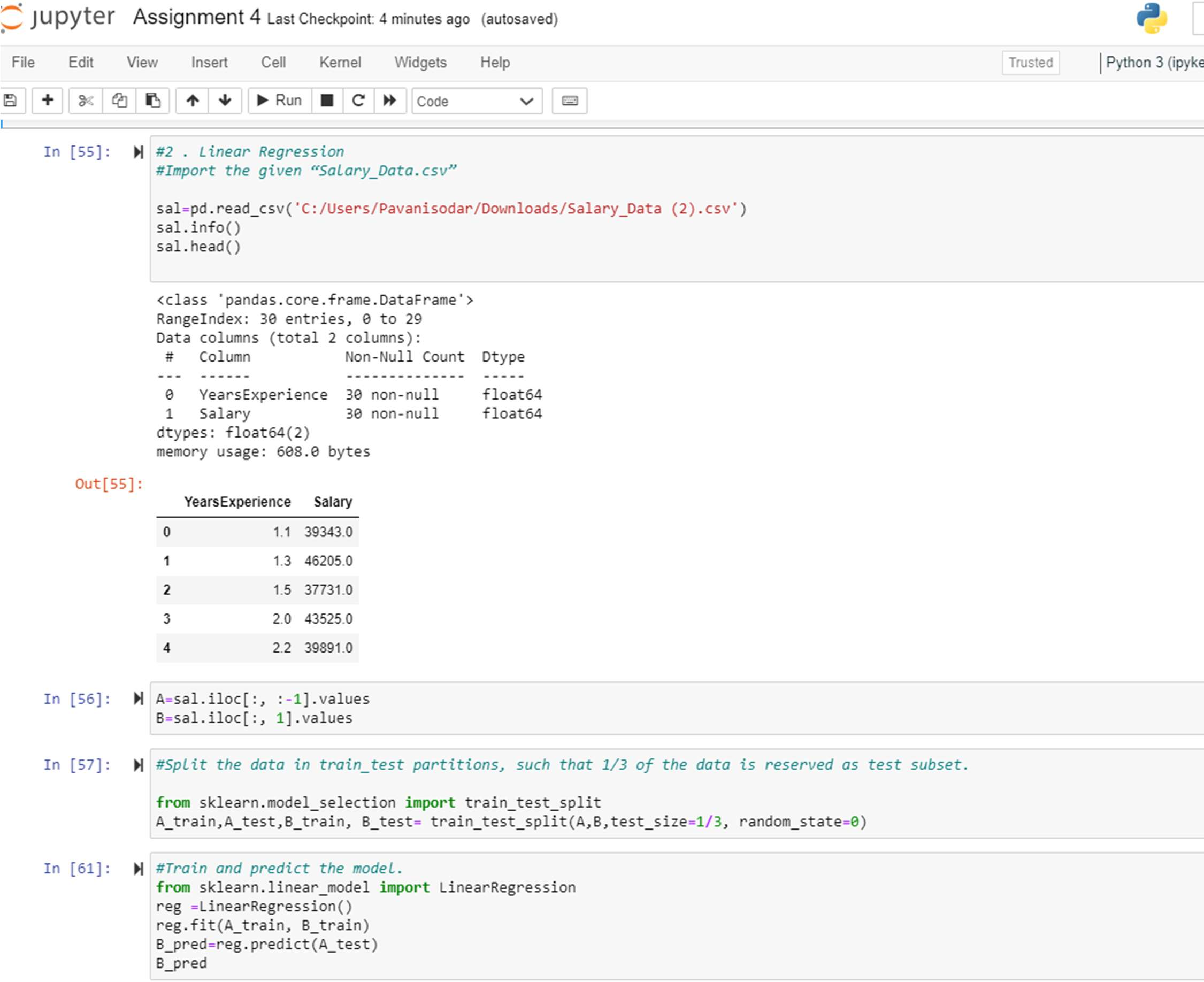
plt.show()

**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:**

A screenshot of a computer

Description automatically generated





**Video Link**: https://drive.google.com/file/d/17TS7wvogfTugXeUfOaSmXWU1x-oXLkaL/view?usp=sharing

**GitHub Link**: https://github.com/BhuvanaNandhimalla/NeuralNetworks\_Assignment4