❖ DATA SCIENCE

EXPERIMENT NO.2

AIM: Consider two data sets given i.e., Customer Behaviour and House Price Prediction.

I. Find Bivariate Association between numeric variables using Covariance and Simple Correlation for the given “House Price Prediction” Data set. Represent the results of covariance and correlation into n*n matrices. Where n is the number of numeric variables.

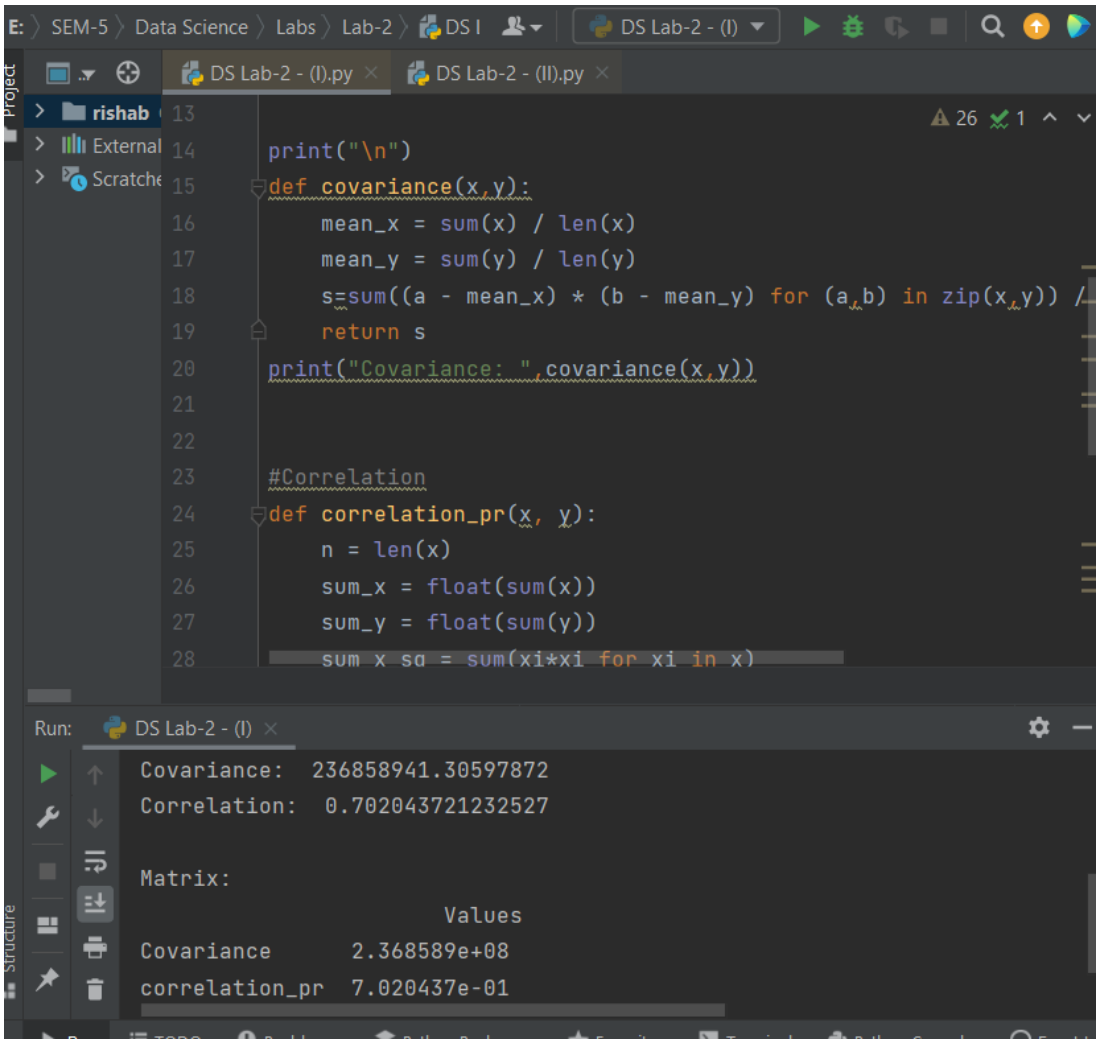
CODE:

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

DF = pd.read_csv("kc_house_data.csv")
print(DF.head())

x = DF.price
y = DF.sqft_living

#Covariance

print("\n")
def covariance(x,y):
    mean_x = sum(x) / len(x)
    mean_y = sum(y) / len(y)
    s=sum((a - mean_x) * (b - mean_y) for (a,b) in zip(x,y)) / len(x)
    return s
print("Covariance: ",covariance(x,y))

#Correlation

def correlation_pr(x, y):
    n = len(x)
    sum_x = float(sum(x))
    sum_y = float(sum(y))
    sum_x_sq = sum(xi*xi for xi in x)
    sum_y_sq = sum(yi*yi for yi in y)
    psum = sum(xi*yi for xi, yi in zip(x, y))
    num = psum - (sum_x * sum_y/n)
    den = pow((sum_x_sq - pow(sum_x, 2) / n) * (sum_y_sq - pow(sum_y, 2) / n), 0.5)
    if den == 0:
        return 0
```

```
    return num / den
print("Correlation: ",correlation_pr(x,y))

ls={'Values':[covariance(x,y),correlation_pr(x,y)]}
matrix=pd.DataFrame(data=ls, index=['Covariance','Correlation_pr'])
print("\nMatrix: \n",matrix)
```

OUTPUT:

```
Covariance:  236858941.30597872
Correlation:  0.702043721232527

Matrix:
                      Values
Covariance        2.368589e+08
Correlation_pr    7.020437e-01
```



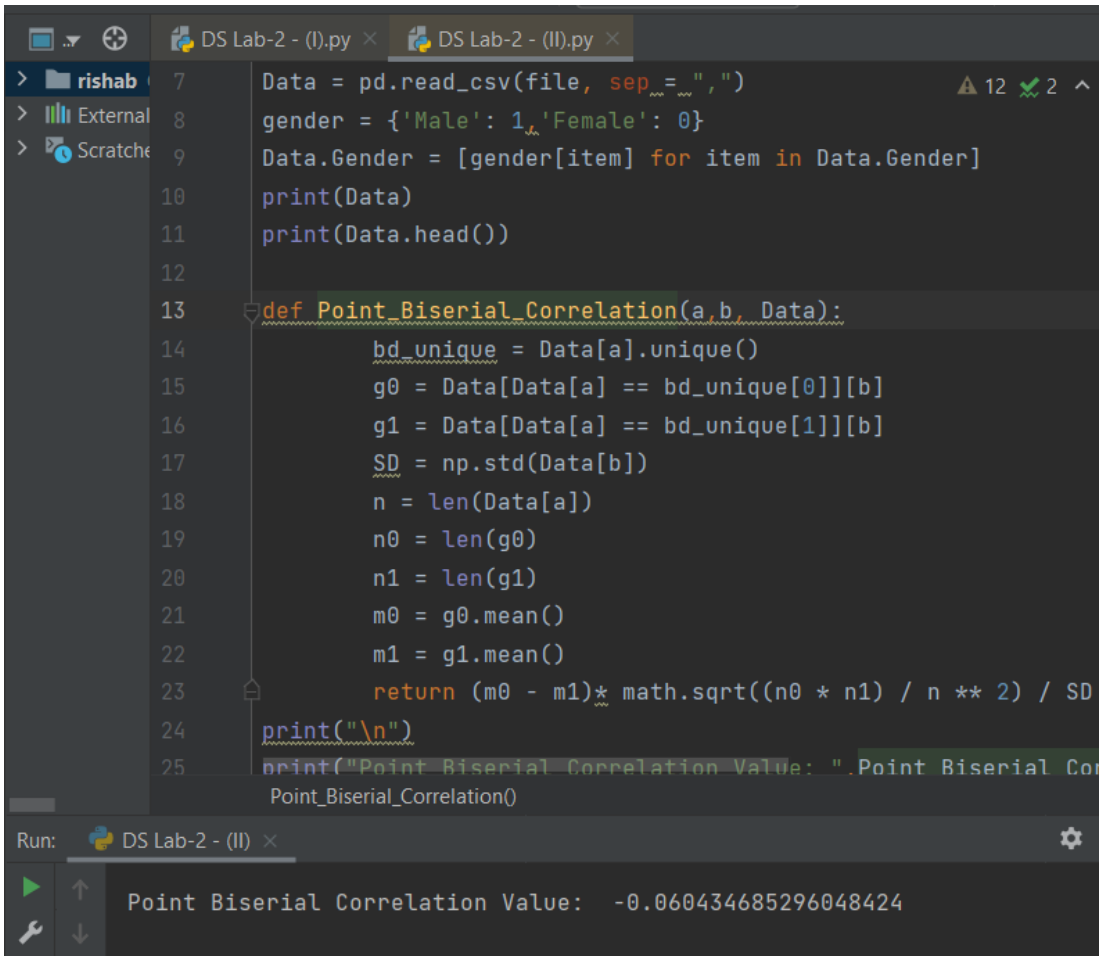
II. Find Bivariate Association between categorical variable “Gender” and numerical variable “Salary” using Point Biserial Correlation for the given Data set i.e., “Customer Behaviour”

CODE:

```
import pandas as pd
import numpy as np
import math
file = open("Customer_Behaviour.csv", "r")

Data = pd.read_csv(file, sep = ",")
gender = {'Male': 1,'Female': 0}
Data.Gender = [gender[item] for item in Data.Gender]
print(Data)
print(Data.head())

def Point_Biserial_Correlation(a,b, Data):

    bd_unique = Data[a].unique()

    g0 = Data[Data[a] == bd_unique[0]][b]
    g1 = Data[Data[a] == bd_unique[1]][b]

    SD = np.std(Data[b])
    n = len(Data[a])
    n0 = len(g0)
    n1 = len(g1)
    m0 = g0.mean()
    m1 = g1.mean()

    return (m0 - m1)* math.sqrt((n0 * n1) / n ** 2) / SD

print("\n")
print("Point Biserial Correlation Value: ",Point_Biserial_Correlation("Gender",
"Salary", Data))
```

OUTPUT:

```
Point Biserial Correlation Value:  -0.060434685296048424
```



CONCLUSION:

Hence, we have learned to find bivariant association between two variable using covariance, correlation and point biserial correlation.
In the first question we have obtained covariance and correlation value which is positive. So, we can say that the bivariant association between two variables is strongly related. Whereas, in the point biserial correlation we obtained a negative value as an output that means the bivariant association between two variables is weakly related in the given dataset.

Name: Rishab Jha
PRN: 20190802072