DATS_6450_15

TIME SERIES MODELING & ANALYSIS

Instructor Name: Reza Jafri

Lab#: 2

Submitted by: Dinesh Kumar Padmanabhan

Date: 23-sept-2020

ABSTRACT

In LAB #2, we learned concepts of Scatter Plots, Correlation coefficients and the relationship between them. Using simple datasets, we plotted Scatter plots and indicated the extend of correlation between the observations and visualized the relationship. We also measured the strength of the relationship between two variables. The correlation coefficient between two variables were calculated using python function and manually as well.

INTRODUCTION

Scatter plots are important statistics because they can show the extend of correlation, if any, between observations. Scatter plot helps us visualize relationship between two random variables. If a variable is uncorrelated with the output, it should be removed from the model. To measure the strength of the relationship between two variables. The correlation coefficient between x and y can be calculated by:

$$r = \frac{\sum(x_t - \overline{x})(y_t - \overline{y})}{\sqrt{\sum(x_t - \overline{x})^2}\sqrt{\sum(y_t - \overline{y})^2}}$$

where x & y are the mean value for x and y and the value of r always lies between -1 and 1

Figure 1 shows Scatter plots showing different correlation from clockwise:

Positive relationship – Y_Dataset v/s X_Dataset

Negative relationship – Z_Dataset v/s X_Dataset

No linear relationship - H_Dataset v/s G_Dataset

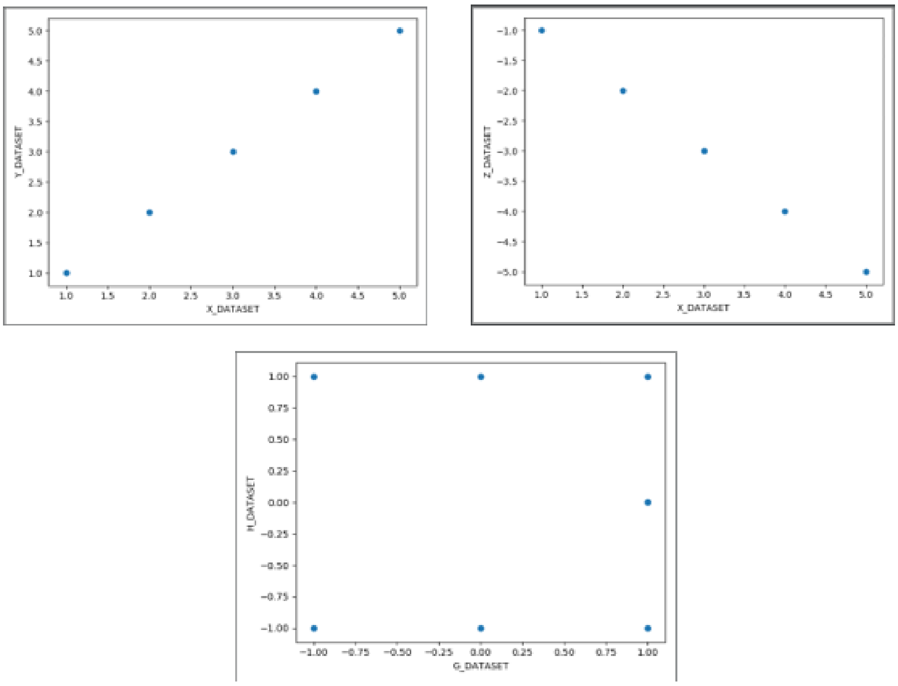


Fig1: Scatter plots showing different correlation

METHOD, THEORY & PROCEDURES

Method:

1. Programming Language: Python

Libraries used: Some basic libraries used for analysis & model building are mentioned below

- *library(Numpy)* - large collection of high-level mathematical functions to operate on these arrays.
- *library (Pandas)* – For Data manipulation and analysis
- *library(Matplotlib)* – is a system for declaratively creating graphics
- *library(Math) –To Compute mathematical calculations*

Theory:

To Plot the scatter plots for the given data set and determine how variables in the dataset are correlated and find the correlation coefficients for the same.

Procedure:

I shall be looking at the variables through scatter plots and infer about it in my analysis. And through my exploration I shall try to identify the how the variables are correlated and draw inferences.

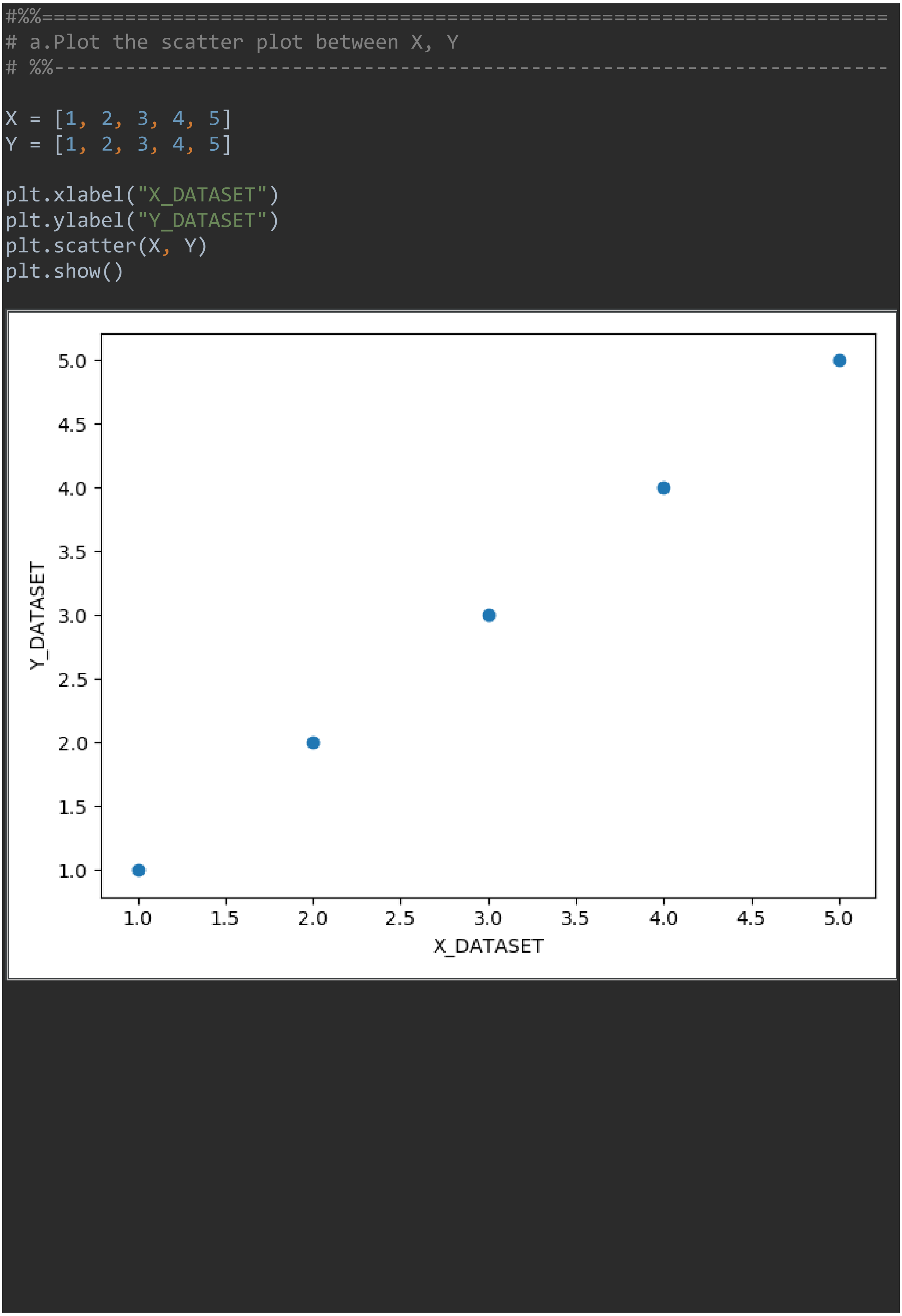
The Dataset will be explored in following stages:

1. **Data Exploration (EDA)** – looking at continuous variables and making inferences about the data.
2. **Data Visualization** – Plotting scatter plots for the variables.
3. **Testing** – Running correlation coefficients function to identify the correlation between them.

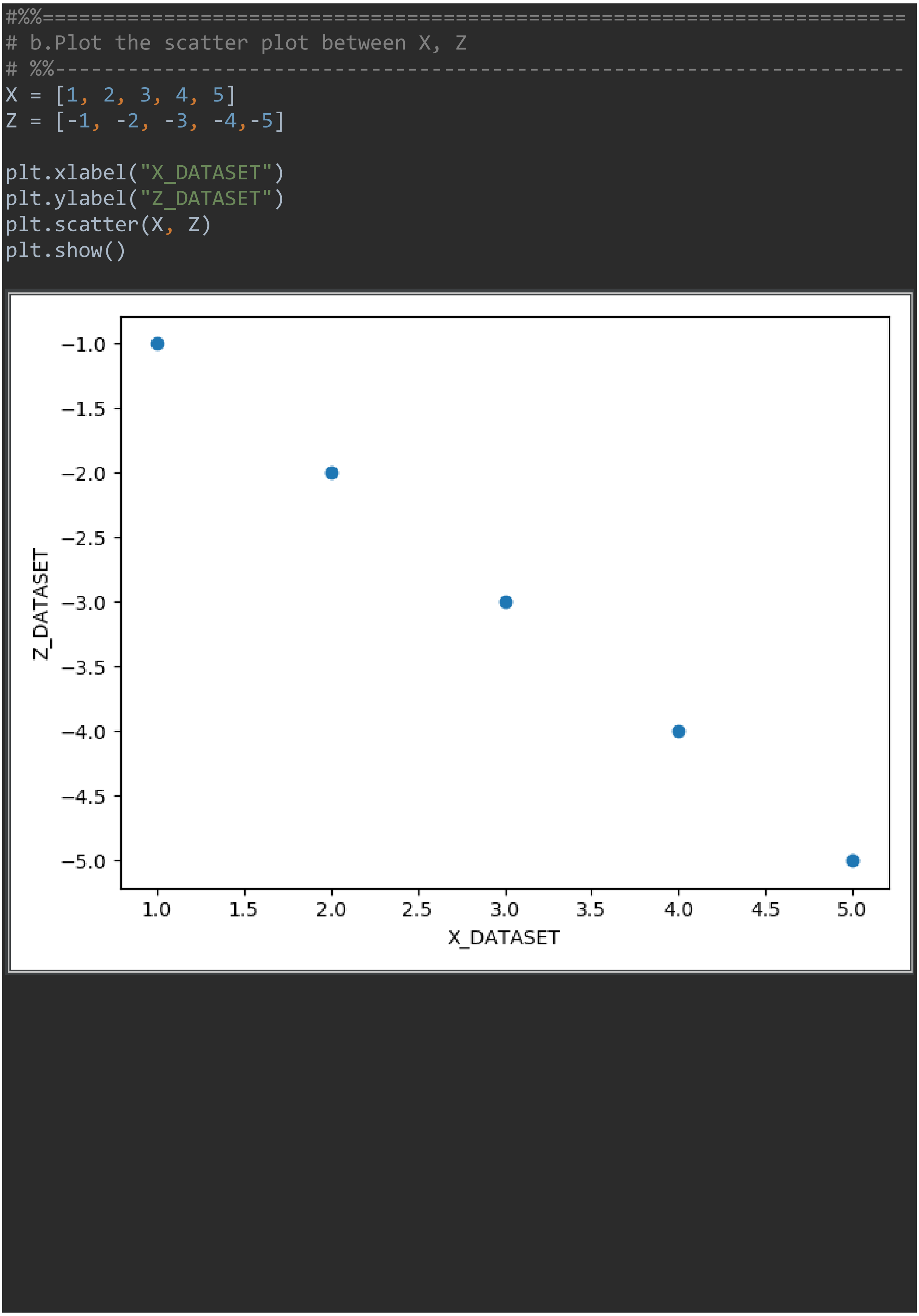
ANSWERS TO ASKED QUESTIONS

```
import pandas as pd
import numpy as np
import math
import matplotlib.pyplot as plt

#Using the Python program and “pandas”,  “matplotlib.pyplot” and “numpy”
library perform the following tasks:
#%%=======================================================================

# 1: Write a python function called “ correlation_coefficent_cal(x,y)”
that implement the correlation coefficient.
# The formula for correlation coefficient is given below.
# The function should be written in a general form than can work for any
dataset x and dataset y.
# The return value for this function is r.
# %%----------------------------------------------------------------------
def correlation_Coefficient_cal(X, Y):
    n = len(X)
    sum_X = 0
    sum_Y = 0
    sum_XY = 0
    squareSum_X = 0
    squareSum_Y = 0
    i = 0
    while i < n:
        sum_X = sum_X + X[i]
        sum_Y = sum_Y + Y[i]
        sum_XY = sum_XY + X[i] * Y[i]

        squareSum_X = squareSum_X + X[i] * X[i]
        squareSum_Y = squareSum_Y + Y[i] * Y[i]

        i = i + 1

    r = (float)(n * sum_XY - sum_X * sum_Y) / \
        (float)(math.sqrt((n * squareSum_X -sum_X * sum_X) * (n *
squareSum_Y -sum_Y * sum_Y)))

    return r
print('\n')
```

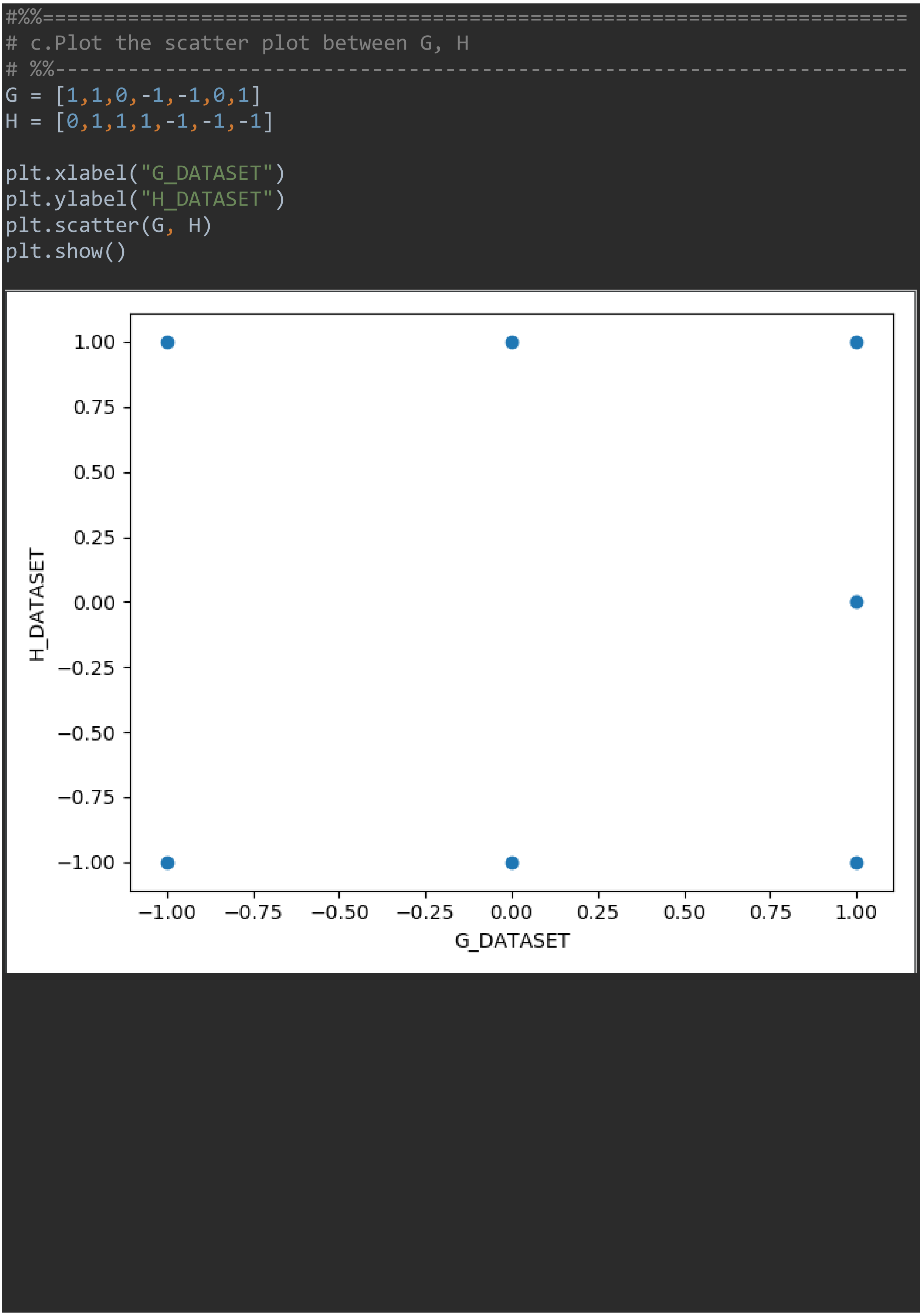
```
#%%=========================================================================
# 2: Test the “ correlation_coefficent_cal(x,y)” function with the
following simple dataset.
# The x and y here are dummy variable and should be replaced by any other
dataset.
# X = [1, 2, 3, 4, 5]
# Y = [1, 2, 3, 4, 5]
# Z = [-1, -2, -3, -4,-5]
# G = [1,1,0,-1,-1,0,1]
# H = [0,1,1,1,-1,-1,-1]
# %%-----------------------------------------------------------------------

#Replacing the dummy variable to below variables

# Input Driver function
X = [15, 18, 21, 24, 27]
Y = [25, 25, 27, 31, 32]

print('Correlation Coeffiecent for X &
Y:{0:.6f}'.format(correlation_Coefficient_cal(X, Y)))
print('\n')

#OUTPUT
Correlation Coeffiecent for X & Y:0.953463
```



```
#%%=========================================================================
# a.Plot the scatter plot between X, Y
# %%-----------------------------------------------------------------------

X = [1, 2, 3, 4, 5]
Y = [1, 2, 3, 4, 5]

plt.xlabel("X_DATASET")
plt.ylabel("Y_DATASET")
plt.scatter(X, Y)
plt.show()
```

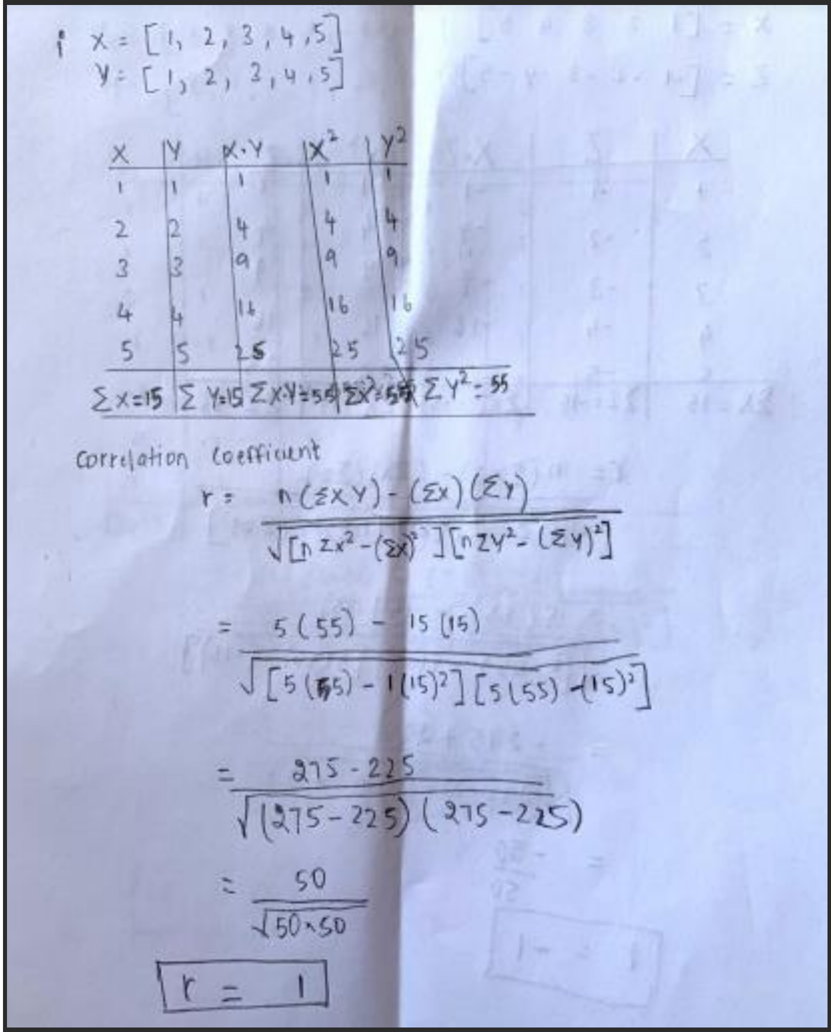


```
#%%=========================================================================
# b.Plot the scatter plot between X, Z
# %%-------------------------------------------------------------------------
X = [1, 2, 3, 4, 5]
Z = [-1, -2, -3, -4,-5]

plt.xlabel("X_DATASET")
plt.ylabel("Z_DATASET")
plt.scatter(X, Z)
plt.show()
```



```
#%%=======================================================================
# c.Plot the scatter plot between G, H
# %%----------------------------------------------------------------------
G = [1,1,0,-1,-1,0,1]
H = [0,1,1,1,-1,-1,-1]

plt.xlabel("G_DATASET")
plt.ylabel("H_DATASET")
plt.scatter(G, H)
plt.show()
```

```
#%%==========================================================================
# d.Without using Python program, implement the above formula to derive
the r_xy, r_xz, r_gh.
# You should NOT use computer to answer this section.
# You need to show all your work for this section on the paper.
# %%--------------------------------------------------------------------------

# r_xy = 1:
See below calculations
```



i $X = [1, 2, 3, 4, 5]$
$Y = [1, 2, 3, 4, 5]$

X	Y	$X \cdot Y$	X^2	Y^2
1	1	1	1	1
2	2	4	4	4
3	3	9	9	9
4	4	16	16	16
5	5	25	25	25
$\Sigma X = 15$	$\Sigma Y = 15$	$\Sigma X \cdot Y = 55$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 55$

Correlation coefficient

$$r = \frac{n(\Sigma XY) - (\Sigma x)(\Sigma Y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma Y^2 - (\Sigma Y)^2]}}$$

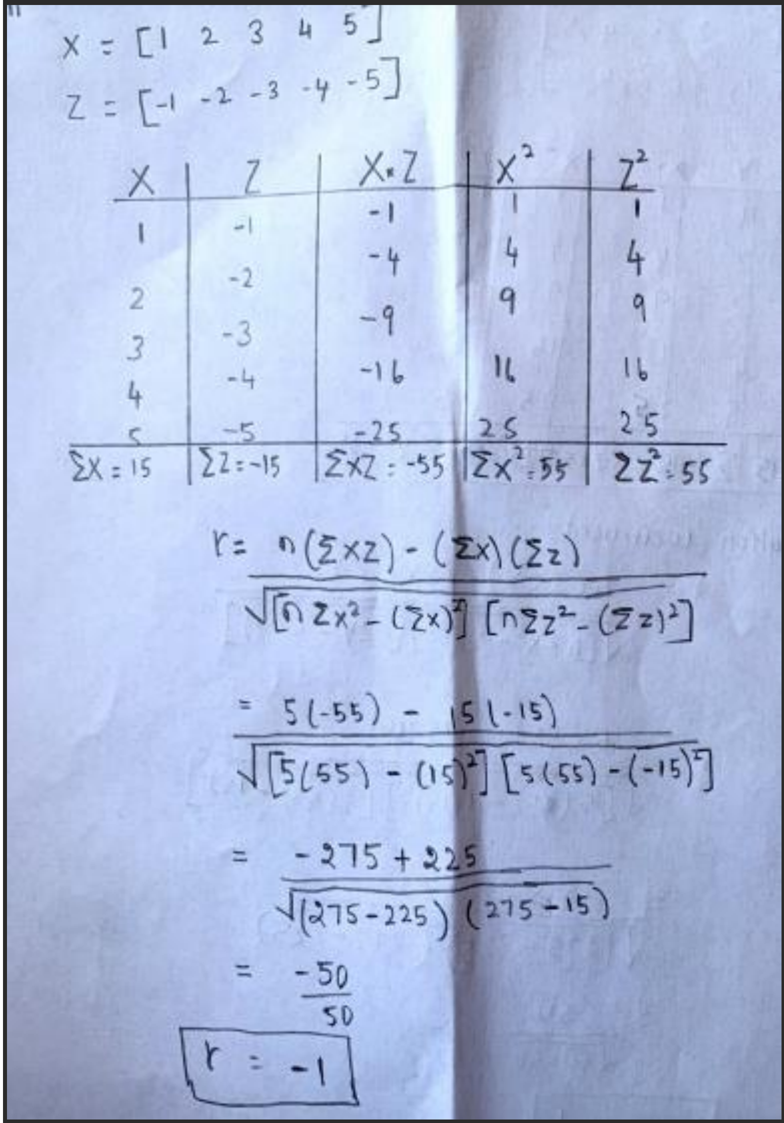
$$= \frac{5(55) - 15(15)}{\sqrt{[5(55) - 1(15)^2][5(55) - (15)^2]}}$$

$$= \frac{275 - 225}{\sqrt{(275-225)(275-225)}}$$

$$= \frac{50}{\sqrt{50 \times 50}}$$

$$\boxed{r = 1}$$

```
#r_xz = -1
See below calculations
```



X = [1 2 3 4 5]

Z = [-1 -2 -3 -4 -5]

X	Z	X.Z	X^2	Z^2
1	-1	-1	1	1
2	-2	-4	4	4
3	-3	-9	9	9
4	-4	-16	16	16
5	-5	-25	25	25
$\Sigma X = 15$	$\Sigma Z = -15$	$\Sigma XZ = -55$	$\Sigma X^2 = 55$	$\Sigma Z^2 = 55$

$$r = \frac{n(\Sigma xZ) - (\Sigma x)(\Sigma z)}{\sqrt{[n\,\Sigma x^2 - (\Sigma x)^2]\,[n\,\Sigma z^2 - (\Sigma z)^2]}}$$

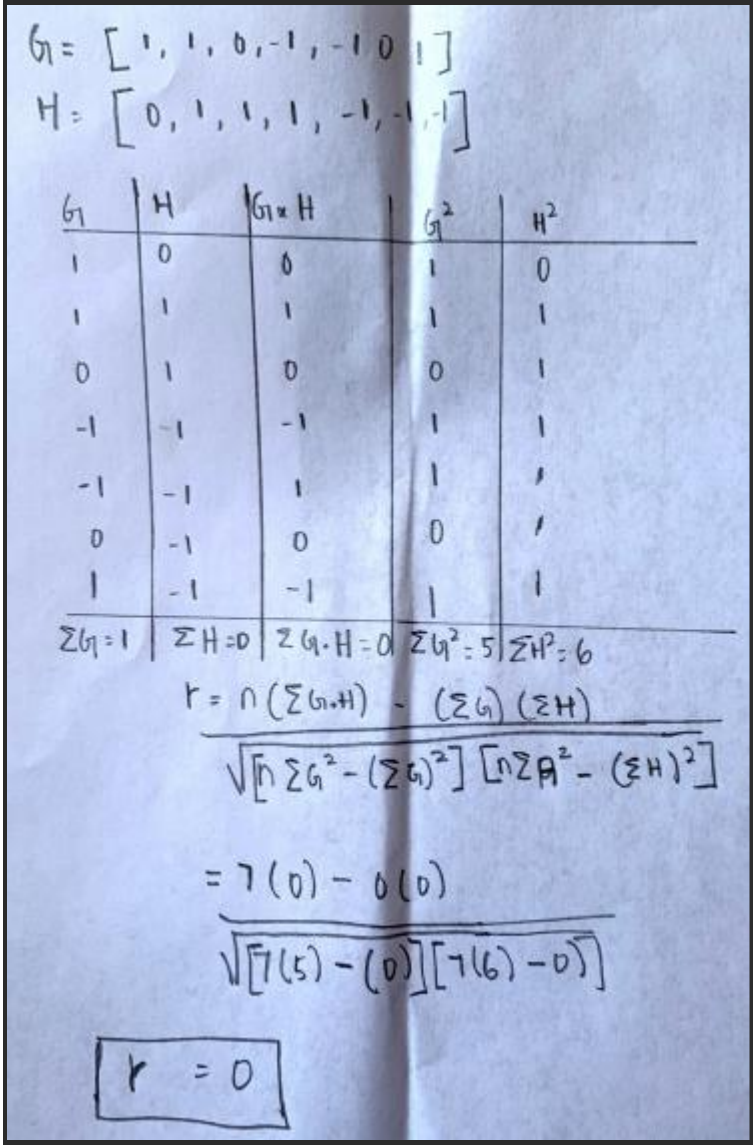
$$= \frac{5(-55) - 15(-15)}{\sqrt{[5(55) - (15)^2]\,[5(55) - (-15)^2]}}$$

$$= \frac{-275 + 225}{\sqrt{(275-225)\,(275 - 15)}}$$

$$= \frac{-50}{50}$$

$$\boxed{r = -1}$$

```
#r_gh = 0

See below calculations.
```



$G = [1, 1, 0, -1, -1\ 0\ 1]$

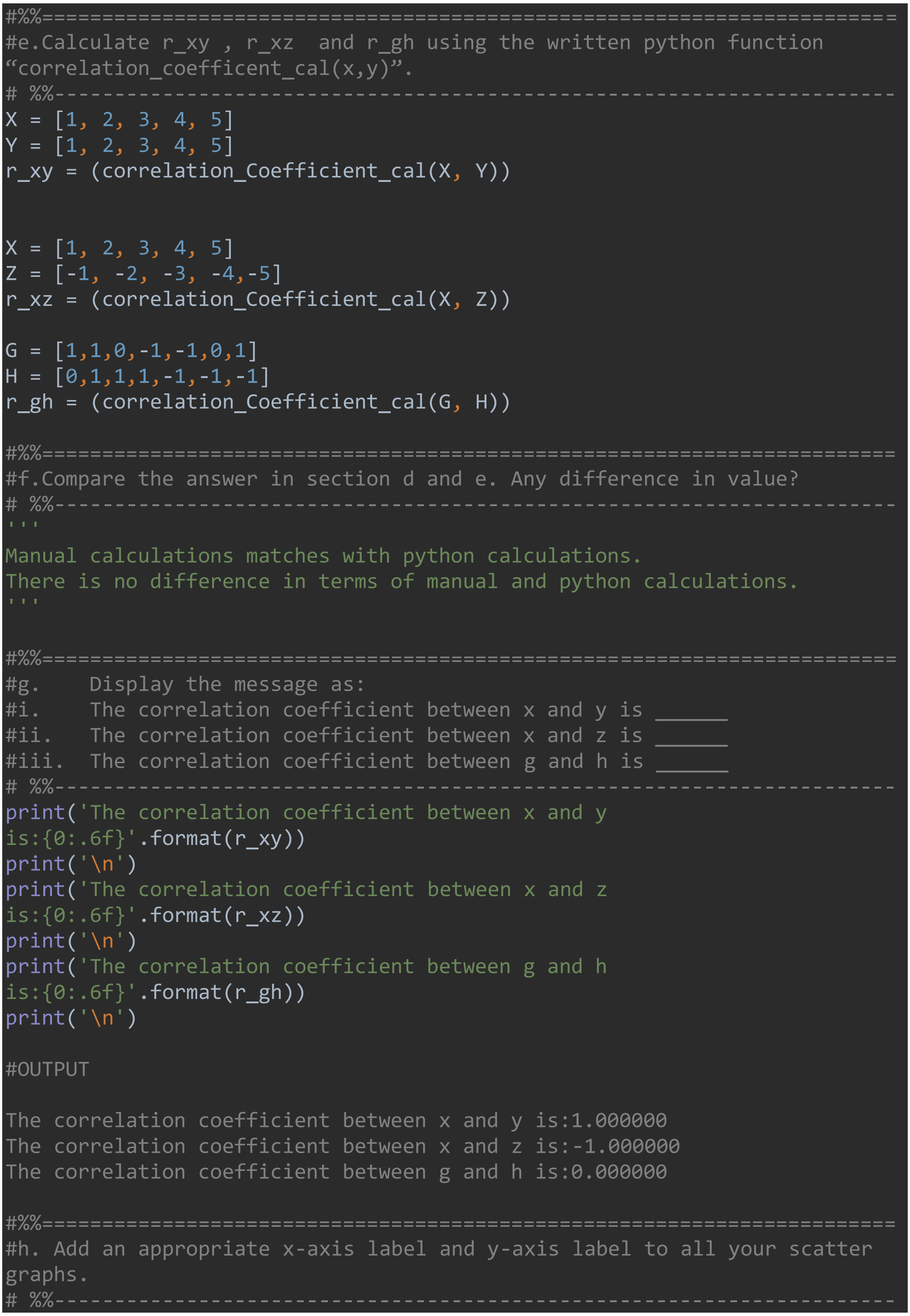
$H = [0, 1, 1, 1, -1, -1, -1]$

G	H	G×H	G^2	H^2
1	0	0	1	0
1	1	1	1	1
0	1	0	0	1
-1	-1	-1	1	1
-1	-1	1	1	1
0	-1	0	0	1
1	-1	-1	1	1
$\Sigma G = 1$	$\Sigma H = 0$	$\Sigma G \cdot H = 0$	$\Sigma G^2 = 5$	$\Sigma H^2 = 6$

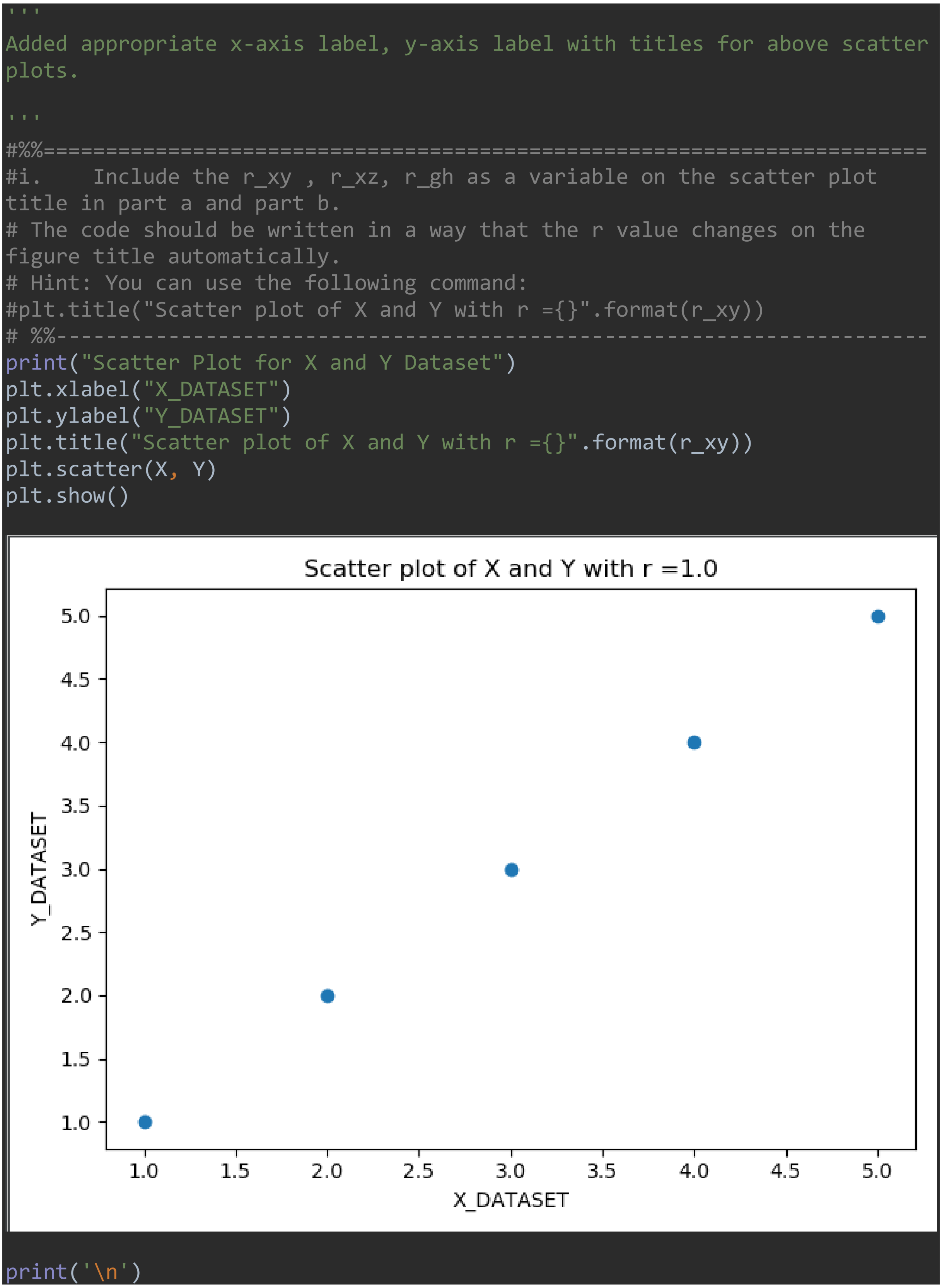
$$r = \frac{n(\Sigma G \cdot H) - (\Sigma G)(\Sigma H)}{\sqrt{[n\Sigma G^2 - (\Sigma G)^2][n\Sigma A^2 - (\Sigma H)^2]}}$$

$$= \frac{7(0) - 0(0)}{\sqrt{[7(5) - (0)][7(6) - 0]}}$$

$$\boxed{r = 0}$$

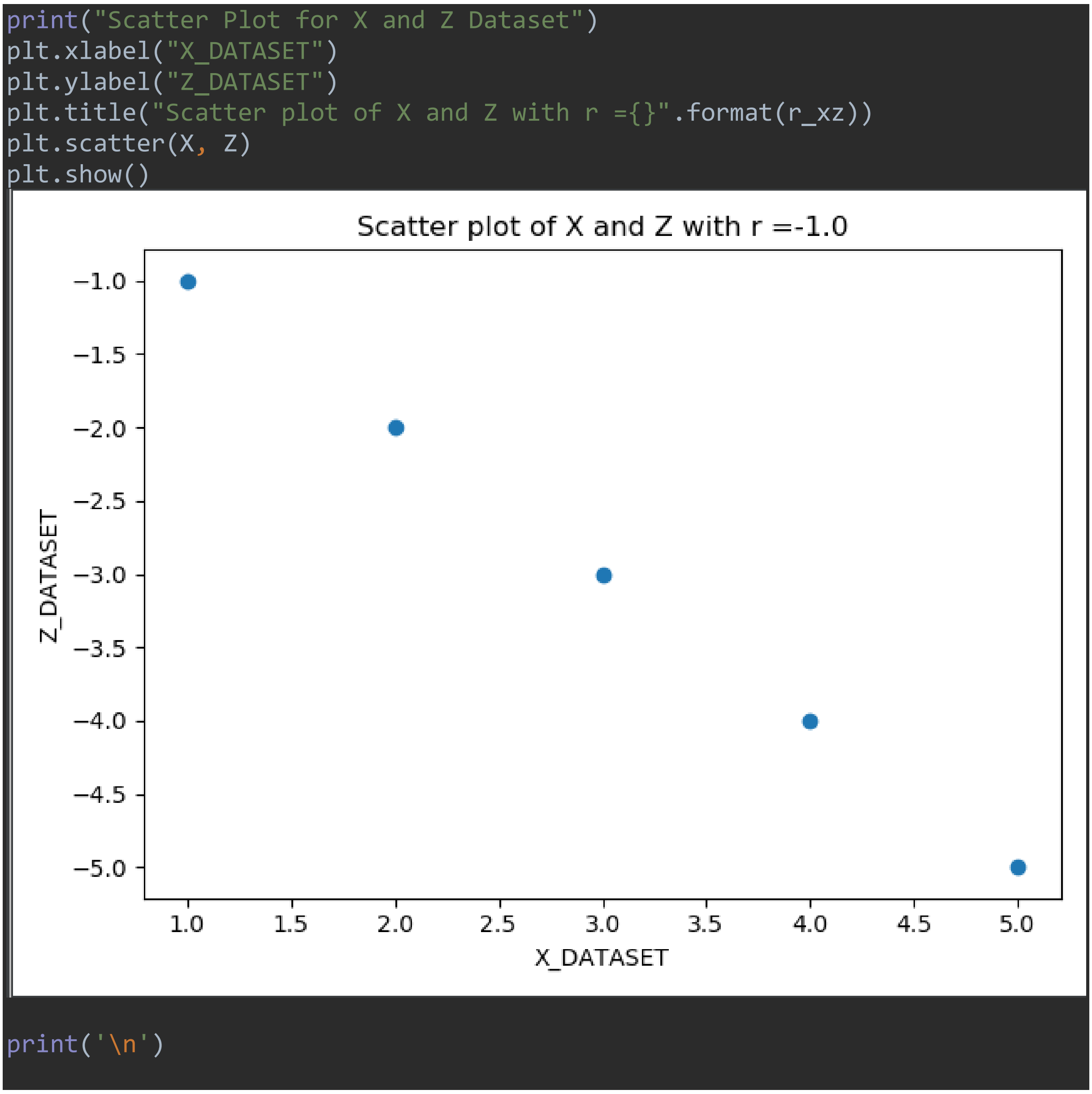


```
#%%===========================================================================
#e.Calculate r_xy , r_xz  and r_gh using the written python function
“correlation_coefficent_cal(x,y)”.
# %%--------------------------------------------------------------------------
X = [1, 2, 3, 4, 5]
Y = [1, 2, 3, 4, 5]
r_xy = (correlation_Coefficient_cal(X, Y))


X = [1, 2, 3, 4, 5]
Z = [-1, -2, -3, -4,-5]
r_xz = (correlation_Coefficient_cal(X, Z))

G = [1,1,0,-1,-1,0,1]
H = [0,1,1,1,-1,-1,-1]
r_gh = (correlation_Coefficient_cal(G, H))

#%%===========================================================================
#f.Compare the answer in section d and e. Any difference in value?
# %%--------------------------------------------------------------------------
'''
Manual calculations matches with python calculations.
There is no difference in terms of manual and python calculations.
'''

#%%===========================================================================
#g.    Display the message as:
#i.    The correlation coefficient between x and y is ______
#ii.   The correlation coefficient between x and z is ______
#iii.  The correlation coefficient between g and h is ______
# %%--------------------------------------------------------------------------
print('The correlation coefficient between x and y
is:{0:.6f}'.format(r_xy))
print('\n')
print('The correlation coefficient between x and z
is:{0:.6f}'.format(r_xz))
print('\n')
print('The correlation coefficient between g and h
is:{0:.6f}'.format(r_gh))
print('\n')

#OUTPUT

The correlation coefficient between x and y is:1.000000
The correlation coefficient between x and z is:-1.000000
The correlation coefficient between g and h is:0.000000

#%%===========================================================================
#h. Add an appropriate x-axis label and y-axis label to all your scatter
graphs.
# %%--------------------------------------------------------------------------
```



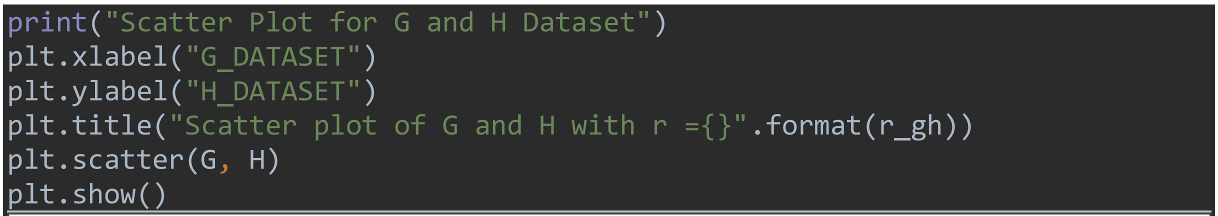
```
'''
Added appropriate x-axis label, y-axis label with titles for above scatter
plots.

'''
#%%=========================================================================
#i.    Include the r_xy , r_xz, r_gh as a variable on the scatter plot
title in part a and part b.
# The code should be written in a way that the r value changes on the
figure title automatically.
# Hint: You can use the following command:
#plt.title("Scatter plot of X and Y with r ={}".format(r_xy))
# %%-------------------------------------------------------------------------
print("Scatter Plot for X and Y Dataset")
plt.xlabel("X_DATASET")
plt.ylabel("Y_DATASET")
plt.title("Scatter plot of X and Y with r ={}".format(r_xy))
plt.scatter(X, Y)
plt.show()
```

```
print('\n')
```

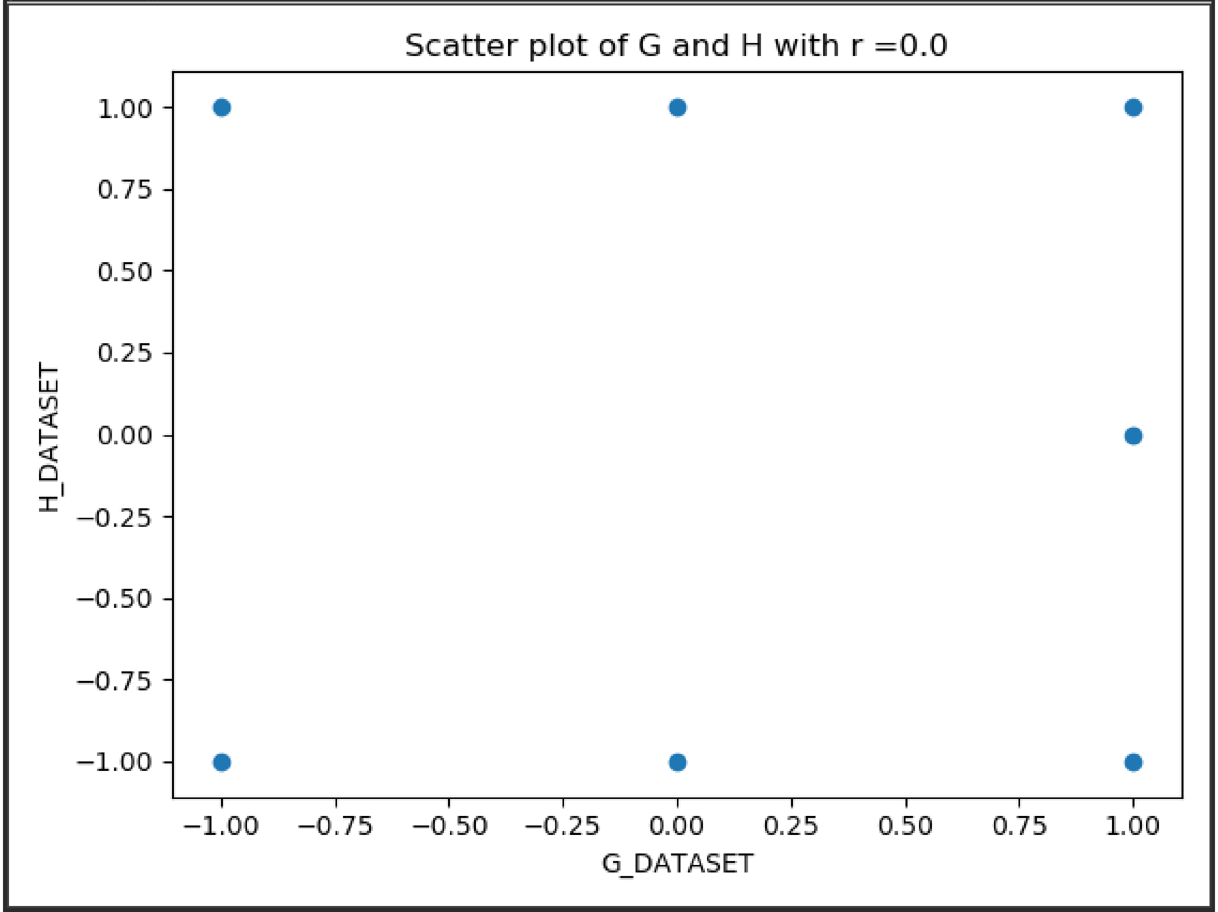


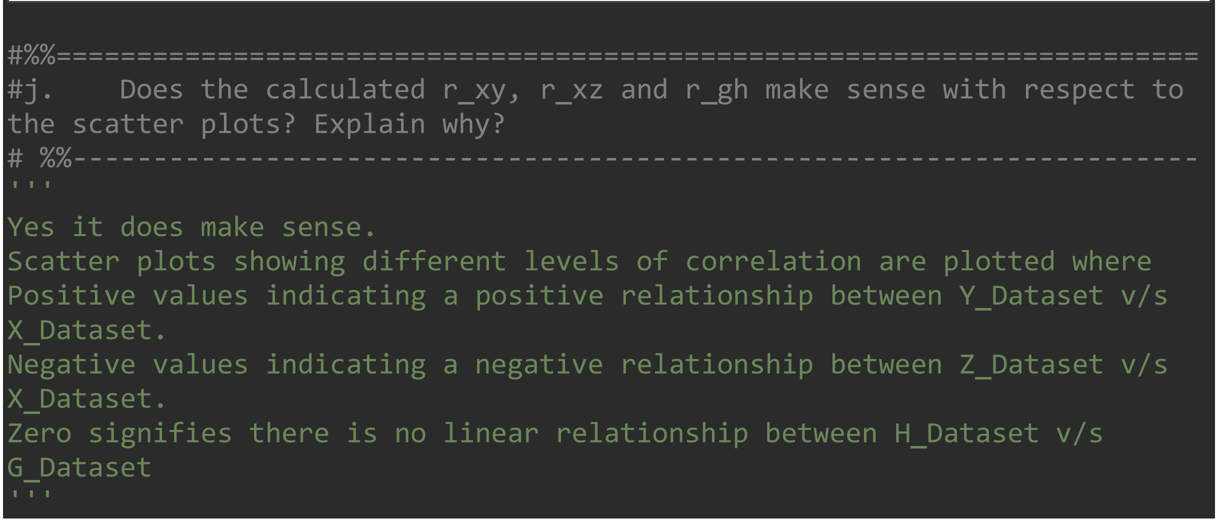
```python
print("Scatter Plot for X and Z Dataset")
plt.xlabel("X_DATASET")
plt.ylabel("Z_DATASET")
plt.title("Scatter plot of X and Z with r ={}".format(r_xz))
plt.scatter(X, Z)
plt.show()
```

```python
print('\n')
```



```
print("Scatter Plot for G and H Dataset")
plt.xlabel("G_DATASET")
plt.ylabel("H_DATASET")
plt.title("Scatter plot of G and H with r ={}".format(r_gh))
plt.scatter(G, H)
plt.show()
```





```
#%%==================================================================
#j.    Does the calculated r_xy, r_xz and r_gh make sense with respect to
the scatter plots? Explain why?
# %%------------------------------------------------------------------
'''
Yes it does make sense.
Scatter plots showing different levels of correlation are plotted where
Positive values indicating a positive relationship between Y_Dataset v/s
X_Dataset.
Negative values indicating a negative relationship between Z_Dataset v/s
X_Dataset.
Zero signifies there is no linear relationship between H_Dataset v/s
G_Dataset
'''
```

CONCLUSION

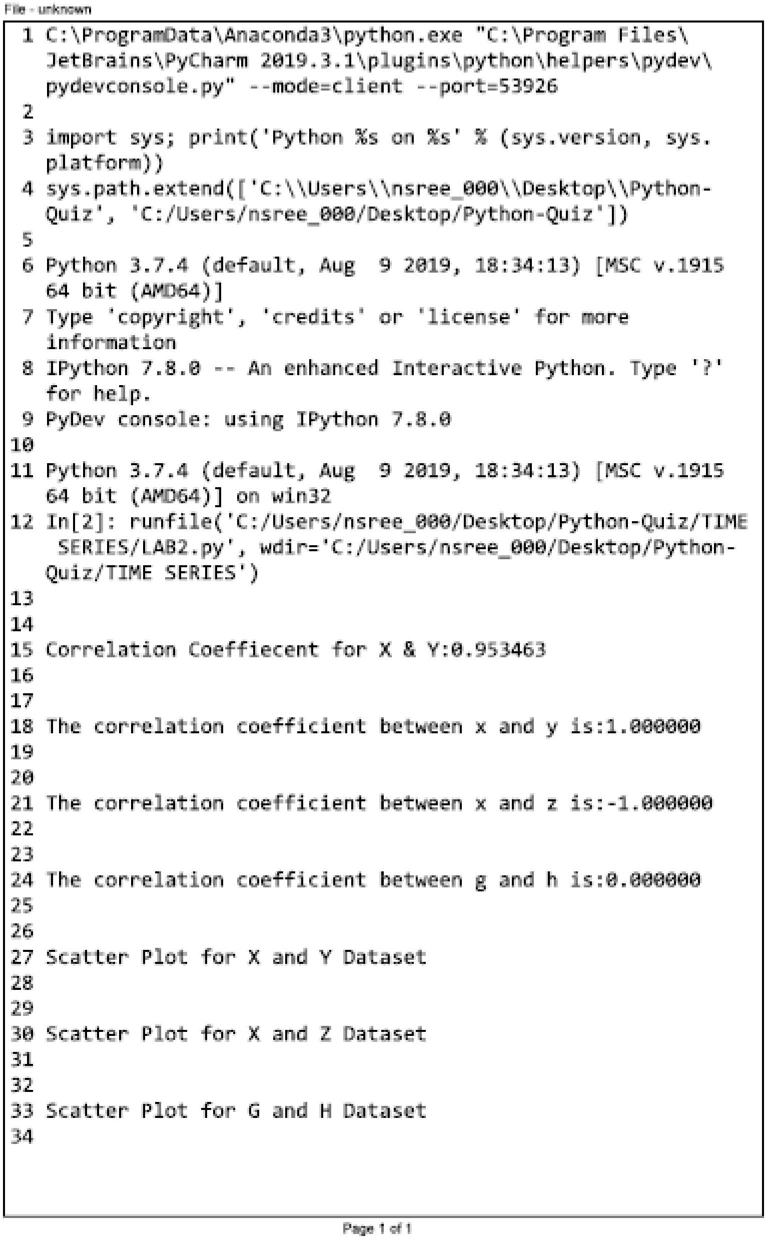
Scatter plots showing different levels of correlation are plotted and the calculated correlation coefficients of xy,xz and gh makes sense with scatter plots. Positive values indicating a positive relationship between Y_Dataset v/s X_Dataset. Negative values indicating a negative relationship between Z_Dataset v/s X_Dataset. Zero signifies there is no linear relationship between H_Dataset v/s G_Dataset. Results can be summarized from below table.

CORRELATION		
Y v/s X	**Z v/s X**	**H v/s G**
Relationship: ‘+’ ve	Relationship: ‘-‘ ve	Relationship: No
r-value: 1	r-value: -1	r-value: 0

CHALLENGE

There was no challenge since it was fairly a simple dataset.

APPENDIX



File - unknown

```
C:\ProgramData\Anaconda3\python.exe "C:\Program Files\JetBrains\PyCharm 2019.3.1\plugins\python\helpers\pydev\pydevconsole.py" --mode=client --port=53926

import sys; print('Python %s on %s' % (sys.version, sys.platform))
sys.path.extend(['C:\\Users\\nsree_000\\Desktop\\Python-Quiz', 'C:/Users/nsree_000/Desktop/Python-Quiz'])

Python 3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)]
Type 'copyright', 'credits' or 'license' for more information
IPython 7.8.0 -- An enhanced Interactive Python. Type '?' for help.
PyDev console: using IPython 7.8.0

Python 3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)] on win32
In[2]: runfile('C:/Users/nsree_000/Desktop/Python-Quiz/TIME SERIES/LAB2.py', wdir='C:/Users/nsree_000/Desktop/Python-Quiz/TIME SERIES')


Correlation Coeffiecent for X & Y:0.953463


The correlation coefficient between x and y is:1.000000


The correlation coefficient between x and z is:-1.000000


The correlation coefficient between g and h is:0.000000


Scatter Plot for X and Y Dataset


Scatter Plot for X and Z Dataset


Scatter Plot for G and H Dataset

```

Page 1 of 1

REFERENCES

https://otexts.com/fpp2/#