**Practical-1**

**Aim**: **Prepare synthetic data set for student data, consisting of Enrollment**

**number, name, gender, semester wise, subject wise marks, difficulty**

**level of the subject, SPI(Semester Index) , address with geographical**

**location.**

1. **(i) Write a program to find correlation between gender and**

**Semester marks.**

**(ii) Write a program to find correlation between geographical location and semester marks. Analyze which two are highly correlated.**

1. **Write a program to calculate correlation between difficulty level**

**and subject marks. The higher the difficulty level the marks should be less. The two should be negatively correlated. Analyze the correlation.**

**Program:**

# To import libraries

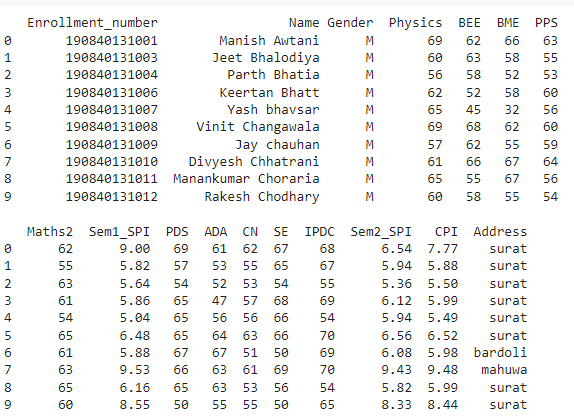
from scipy.stats import spearmanr

import pandas as pd

# To import File and print top 10 lines

file=pd.read\_excel("/content/DAV\_1\_Data.xlsx")

print(file.head(10))



1. **(i)-Correlation between Gender and Marks**

**#** To extract columns

gender=file.loc[:,'Gender']

location=file.loc[:,'Address']

sem1\_marks=file.loc[:,'Sem1\_SPI']

sem2\_marks=file.loc[:,'Sem2\_SPI']

**#** To find spearman relation between gender and sem1 marks

coef1=spearmanr(gender,sem1\_marks)

print("Spearman rank correlation between Gender and Semester 1 marks is:"+str(coef1[0]))



**#** To find spearman relation between gender and sem2 marks

coef2=spearmanr(gender,sem2\_marks)

print("Spearman rank correlation between Gender and Semester 2 marks is:"+str(coef2[0]))



1. **(ii)-Correlation between Location and Marks**

**#** To find spearman relation between location and sem1 marks

coef3=spearmanr(location,sem1\_marks)

print("Spearman rank correlation between Location and Semester 1 marks is:"+str(coef3[0]))



**#** To find spearman relation between location and sem2 marks

coef4=spearmanr(location,sem2\_marks)

print("Spearman rank correlation between Location and Semester 2 marks is:"+str(coef4[0]))



1. **Correlation between Difficulty level and Subject marks**

cn\_marks=file.loc[:,'CN']

difficulty\_levels=pd.Series([])

for i in range(0,len(cn\_marks)):

  if(cn\_marks[i]<=25):

    difficulty\_levels[i]=2

  if(cn\_marks[i]>25 and cn\_marks[i]<55):

    difficulty\_levels[i]=1

  if(cn\_marks[i]>=55):

    difficulty\_levels[i]=0

coef5=spearmanr(difficulty\_levels, cn\_marks)

print("Spearman rank correlation between difficulty level and CN marks is:"+str(coef5[0]))

