Department of Computer Science and Engineering (Data Science)
Sy.B.Tech. Sem: IV
Subject: Statistics For Data Science (DJS23DLPC403)
Experiment 2

Name: Smayan Kulkarni **SAP ID: 60009230142**

Date:	**Experiment Title: Correlation**
Aim	Given a data set of 10 rows. Calculate Karl Pearson's coefficient of correlation, Spearman's rank correlation coefficient (using repeated ranks) manually. Then write a python program to calculate both coefficients and match it with the manually calculated values. Solve the real world problem statements.
Software	Google Colab, Visual Studio Code, Jupyter Notebook
Theory To Be written	What is a strong monotonic relationship? State the types of monotonic relationship with examples.
Implementation	Data: X = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 Y = 5, 6, 7, 8, 7, 9, 10, 10, 11, 12 Plot a scatter plot of the above data. **Step 2: Karl Pearson's Coefficient of Correlation (r)** The formula for Karl Pearson's correlation coefficient is: $$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$ Where: • n is the number of data points (in this case 10), • x and y are the individual data points of the variables X and Y.

Step 3: Spearman's Rank Correlation Coefficient (ρ)

Spearman's rank correlation is based on the ranks of the data. The formula is:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where:

- d is the difference between the ranks of the corresponding values of X and Y,
- n is the number of data points.

Python Code :

1, Write a function to calculate pearson_correlation(X, Y).

2. Write a function to calculate spearman_rank_correlation(X, Y).

3. Use scipy to verify Spearman's rank.

4. Print all the three results.

Real world problems.

Q1. *The following table gives the data on weekly family consumption expenditure*(Y) *and weekly family income*(X)

$Y:$	70	65	90	95	110	115	120	140	155	150
$X:$	80	100	120	140	160	180	200	220	240	260

(i) *Compute the coefficient of correlation between* X *and* Y.

(ii) *Test the significance of the coefficient of correlation between* X *and* Y *at* 5 *percent level of significance.*

Q2. *The following table gives the per capita household expenditure on food* (Y) *and per capita total household expenditure* (X)

$Y:$	60	90	110	125	150	170	180	200	220	230	240	250	255	260	260
$X:$	100	150	200	250	300	350	400	450	500	550	600	650	700	750	800

(i) *Compute the coefficient of correlation between* X *and* Y.

(ii) *Test the significance of the coefficient of correlation between* X *and* Y *at* 5 *percent level of significance.*

Q3. A company wants to analyze the **factors affecting employee productivity**. The HR department wants to know:

1. Which independent variable (**X1, X2, X3**) has the **strongest correlation** with employee productivity (**Y**)?
2. Is the correlation **statistically significant** at a **5% level**?
3. Can we visualize the relationships using scatter plots and a heatmap?

Employee	Hours Worked (X1)	Experience (X2)	Training Programs (X3)	Productivity Score (Y)
1	35	2	1	50
2	40	3	2	55
3	45	5	3	65
4	50	7	2	70
5	52	9	3	78
6	55	10	4	85
7	60	12	4	88
8	62	14	5	90
9	65	15	6	92
10	68	18	6	94
11	70	20	7	96
12	75	22	8	98

Conclusion

Hence, we have calculated Karl Pearson's coefficient of correlation, Spearman's rank correlation coefficient (using repeated ranks) manually and have implemented a python program to calculate both. Both the coefficients are matching.

Conclusion of real world problem 1

Conclusion of real world problem 2

Conclusion of real world problem 3

Colab Link

https://colab.research.google.com/github/SmayanKulkarni/AI-and-ML-Course/blob/master/SDS/exp_2.ipynb

Signature of Faculty