|  |  |
| --- | --- |
| Subject: Data Analytics and Visualization Lab | Course ID: CSL-601 |
| Semester: VI | Course: AI & DS |
| Laboratory: 407 | Name of teacher: Prof. Gitanjali Korgaonkar |
| Name of Student: Ayush Gupta | Roll No: VU2S2324001 |

**EXPERIMENT NO. 4**

**Aim:**

To implement Correlation and Covariance.

**Theory:**

1. **Covariance:**

Covariance is a statistical measure that quantifies the degree to which two variables change together. It determines whether an increase in one variable corresponds to an increase or decrease in another.

**Definition and Calculation of Covariance**

Mathematically, covariance between two random variables X and Y is calculated using the formula:



Where:

* Xi and Yi ​ are individual data points,
* X̄ and Ȳ ate the means of X and Y,
* n is the number of data points.

**Interpreting Covariance Values**

* **Positive Covariance**: If the covariance value is positive, it indicates a direct relationship between the variables—i.e., when one variable increases, the other tends to increase.
* **Negative Covariance**: A negative covariance suggests an inverse relationship—when one variable increases, the other decreases.
* **Zero Covariance**: If the covariance is zero, it implies no relationship between the two variables.

**Limitations of Covariance**

* **Scale Dependency**: Covariance values depend on the scale of measurement and cannot be used for direct comparisons across datasets.
* **Does Not Indicate Strength**: While covariance tells us about direction, it does not indicate the strength of the relationship.

1. **Correlation**

Unlike covariance, correlation measures both the strength and direction of the relationship between two variables. It standardizes covariance values by normalizing them within a fixed range of -1 to +1, making comparisons across datasets easier.

**Definition and Calculation of Correlation**

The correlation coefficient (rrr) is calculated using the Pearson correlation formula:



Where:

* Cov(X,Y) is the covariance between X and Y,
* σX and σY​ are the standard deviations of X and Y.

**Interpreting Correlation Values**

* **r =+ 1:** Perfect positive correlation (as one variable increases, the other increases proportionally).
* **r =− 1**: Perfect negative correlation (as one variable increases, the other decreases proportionally).
* **r = 0**: No correlation (no relationship between variables).

**Pearson Correlation Coefficient and Its Properties**

1. **Strength**: Values close to +1 or -1 indicate strong relationships, while values near 0 indicate weak relationships.
2. **Direction**: A positive correlation implies direct proportionality, whereas a negative correlation indicates an inverse relationship.
3. **Standardized Measure**: Unlike covariance, correlation is unit-free and allows comparisons between different datasets.
4. **Sensitivity to Outliers**: Extreme values can significantly impact the correlation coefficient.
5. **Other Types of Correlation Coefficients:**

Apart from Pearson’s correlation, other correlation coefficients include:

* **Spearman’s Rank Correlation**: Measures the relationship between ranked data and is useful for non-linear associations.
* **Kendall’s Tau**: Measures the ordinal relationship between two variables, useful for small datasets or data with ties.

**Advantages and Disadvantages:**

* **Advantages of Covariance**
* Simple to calculate and interpret.
* Useful in portfolio analysis to measure asset relationships.
* **Disadvantages of Covariance**
* Scale-dependent, making cross-dataset comparisons difficult.
* Does not quantify the strength of relationships.
* **Advantages of Correlation**
* Standardized, allowing for easy interpretation.
* Less affected by measurement units.
* Helps in identifying strong relationships between variables.
* **Disadvantages of Correlation**
* Can be influenced by outliers.
* Only measures **linear** relationships.
* Does not imply causation (correlation ≠ causation).

**Learning Objectives:**

* Understand the concept of covariance and how it indicates the direction of relationships between variables.
* Learn about correlation, its interpretation, and how it standardizes relationships.
* Gain hands-on experience in calculating correlation and covariance using Python (NumPy, Pandas) and R.
* Understand the limitations of covariance and the advantages of correlation in statistical analysis.

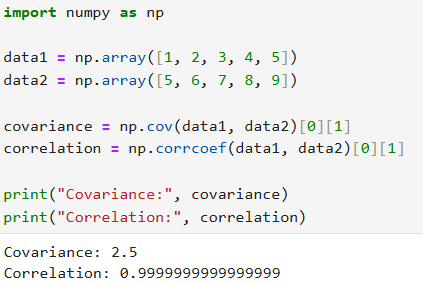
**Conclusion:**

In this experiment, we successfully implemented correlation and covariance using both Python and R. Covariance was used to determine the direction of relationships, while correlation provided both direction and strength of relationships. We also explored different types of correlation and understood the advantages and limitations of both measures. These concepts are crucial in statistical analysis, particularly in fields like finance, economics, and machine learning.

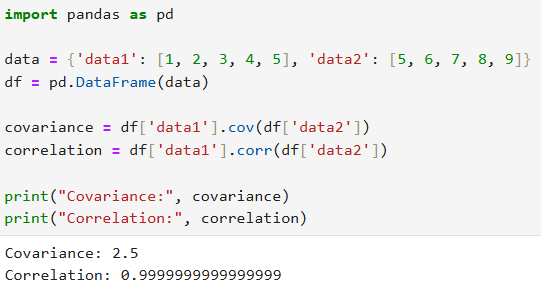


**Program and Output:**

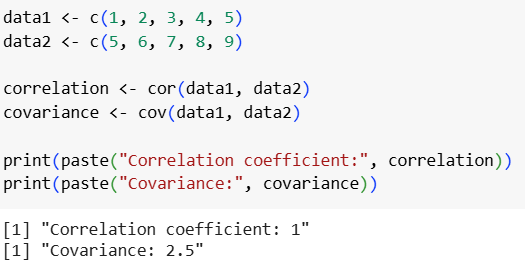
* 1. **Using Numpy:**

****

* 1. **Using Pandas:**

****

* 1. **Using R programming:**

****