Assignment no. 5

Aim: Implementation of Statistical Hypothesis Correlation Test.

LO Mapped: LO2

Theory:

a. What is correlation and covariance?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related. It is a value that ranges from -1 to 1. A correlation of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Covariance, on the other hand, is a measure of the joint variability of two random variables. It tells us whether two variables tend to increase or decrease together. While correlation indicates the strength and direction of a linear relationship, covariance provides a measure of how much two variables change together, but it is not normalized, so its value is influenced by the scale of the variables.

b. What is the role of correlation in data science?

In data science, **correlation** plays a crucial role in understanding relationships between variables. It helps in identifying patterns and associations that can be important for predictive modeling. By analyzing correlation, data scientists can:

1. **Identify Relationships**: Determine whether and how strongly two variables are related, which can guide feature selection in machine learning models.
2. **Feature Engineering**: Highly correlated features may be redundant and can be removed or combined, improving model efficiency and preventing multicollinearity.
3. **Predictive Modeling**: Correlation can guide the selection of predictor variables, ensuring that models leverage relevant relationships.
4. **Data Visualization**: Correlation matrices are often used in exploratory data analysis (EDA) to visualize relationships and uncover patterns.
5. **Decision Making**: In business contexts, understanding correlation can aid in identifying key drivers or predictors of outcomes, influencing strategy or decision-making.