Data Science Lab-I

Frequency Analysis and Probability in Data Science

Objective:

This lab aims to provide hands-on experience with:

1. Creating and analyzing frequency tables
2. Calculating joint, marginal, and conditional probabilities from contingency tables
3. Understanding and computing correlation between variables

Dataset:

Use the **Titanic Dataset** (available via `seaborn` or `Kaggle`). It contains demographic and survival information of passengers on the Titanic.

URL of Titanic Dataset: https://raw.githubusercontent.com/mwaskom/seaborn-data/master/titanic.csv

```
import seaborn as sns

df = sns.load_dataset('titanic')
```

Part I: Frequency Table

Task 1: Frequency Table of Categorical Variable

- Create a frequency table of the `class` variable (First, Second, Third).
- Include:
  - Absolute frequencies
  - Relative frequencies (%)
  - Cumulative frequencies

Part II: Joint, Marginal, and Conditional Probabilities

Task 2: Two-Way Table of sex vs survived

- Construct a two-way table (contingency table) between `sex` and `survived`.

	Survived = 0	**Survived = 1**	**Total**
Male			
Female			
Total			

Task 3: Compute the Following Probabilities:

1. **Joint Probability**: P(Sex = female, Survived = 1)
2. **Marginal Probability**:
    - P(Sex = female)
    - P(Survived = 1)
3. **Conditional Probability**:
    - P(Survived = 1 | Sex = female)
    - P(Sex = female | Survived = 1)

Use pandas `crosstab` and probability formulas.

Part III: Correlation Analysis

Task 4: Numerical Correlation

- Choose two numeric variables:
    - `age` and `fare`
- Clean the data (handle missing values).
- Compute Pearson correlation between them.
- Visualize using:
    - `sns.heatmap()` or `sns.pairplot()`
    - Scatter plot with `plt.scatter()`

Task 5: Interpretation

- Interpret the strength and direction of correlation.
- What does the sign of the coefficient indicate?

Bonus Task (Optional):

Use the `class` and `survived` variables:

- Create a stacked bar chart to visualize survival by class.
- Comment on which class had the highest survival rate.

Deliverables:

- Jupyter Notebook / Python script with:
    - Code
    - Visualizations
    - Explanation and interpretation for each task
- Submit the report in PDF