Task-5

Titanic Dataset

**Step 1 – Import libraries & load data**

- Imported **Pandas**, **NumPy**, **Matplotlib**, and **Seaborn**.
- Loaded titanic.csv into a DataFrame.
- Kept a copy of the original dataset for reference.
- Checked dataset **shape** and displayed the first 10 rows.

**Step 2 – Basic exploration**

- Used:
  - df.info() → structure, column types, non-null counts ○
    df.describe() → summary statistics (numeric) ○
    df.describe(include='all') → summary for all columns ○
    df.isnull().sum() → missing values count and percentage
  - df.value_counts() → counts for key categorical columns (Survived, Pclass, Sex, Embarked)

**Step 3 – Data preparation**

- Cleaned column names (.str.strip()).
- Converted categorical variables (Survived, Pclass, Sex, Embarked) to category dtype.

**Step 4 – Univariate analysis (numerical features)**

- **Histograms** for all numeric columns → distribution check.
- **Boxplots**:
  - Age by Survived → survival distribution across ages ○
    Fare by Survived → fare patterns for survivors vs non-survivors

**Step 5 – Categorical analysis**

- **Countplots** for Sex, Pclass, Embarked → passenger distribution.
- **Barplots** for survival rates by each category.

**Step 6 – Feature engineering for deeper insights**

- Extracted **Title** from passenger Name and grouped rare titles.
- Created **FamilySize** = SibSp + Parch + 1
- Created **IsAlone** = 1 if FamilySize = 1, else 0
- Plotted survival rates for new features.

**Step 7 – Bivariate & correlation analysis** •

**Correlation heatmap** for numeric columns.

- **Pairplot** for selected numeric variables .
- **Scatterplot**: Fare vs Age with hue = Survived.

**Step 8 – Missing values inspection**

- Plotted Age distribution before and after **median imputation**.
- Checked Embarked missing values and found mode for imputation.

**Step 9 – Observations & notes**

- Added markdown cells to note:
  - Survival patterns by gender, class, fare, age. ○ Correlation insights (Pclass vs Fare and survival).
  - Impact of family size and titles.

**Step 10 – Summary of findings**

- Wrote final **summary section** describing:
  - ○ Dataset structure and missingness ○ Demographics ○ Survival trends ○

Correlation observations ○ Possible next steps for modeling