

Session 7: Exploratory Data Analysis using R

Jamuna S Murthy
Assistant Professor
Department of CSE

Academic Year - 2025-26

1. Kaggle: Titanic Dataset

Dataset Description: The Titanic dataset contains information about passengers on the Titanic ship, including features like age, sex, passenger class, fare, and whether the passenger survived the disaster. It is often used for classification tasks to predict whether a passenger survived or not based on the available features.

Source: Kaggle Titanic Dataset

Key Analysis Tasks:

- Perform exploratory data analysis (EDA) to analyze missing values, correlations, and patterns.
- Apply logistic regression to predict survival.
- Use feature engineering to enhance the model (e.g., creating new features like "family size" by combining family and sibling columns).

Suggested Visualizations:

- Bar chart showing the distribution of survivors vs. non-survivors.
- Heatmap of correlations between features.
- Box plot showing fare distribution by passenger class or survival status.
- Survival rate comparison by sex, age group, and class.