Nigam Pranay Mehta

EDA Report: Titanic Dataset Analysis

Objective

To extract insights using visual and statistical exploration of the Titanic dataset.

1. Data Overview

The dataset contains information about passengers on the Titanic, including their age, sex, passenger class, fare, and survival status. - **Total Rows:** 891 - **Columns:** 12 (including 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked')

2. Statistical Findings

- **Survival Rate:** Approx. 38% of passengers in the dataset survived.
- **Gender Distribution:** There were more males (approx. 65%) than females (35%).
- **Class Distribution:** The majority of passengers were in 3rd Class.

3. Visual Comparisons & Observations

Correlation

- **Pclass vs Fare:** Strong negative correlation. As expected, better classes (lower number) cost more.
- **Survived vs Fare:** Positive correlation. Passengers who paid more were more likely to survive.
- **Survived vs Pclass:** Negative correlation. Lower class number (1st class) associated with higher survival.

Pairplot Analysis

- Separating by 'Survived' shows distinct clusters in Age vs Fare.
- Younger passengers (children) appear more likely to survive across all classes.

Distributions

- **Age:** The age distribution is slightly right-skewed, with a peak around 20-30 years. However, there is a small secondary peak for young children.
- **Fare:** Highly right-skewed with many low fares and few very high fares.

Boxplots

- **Age by Class:** 1st Class passengers are generally older than 3rd Class passengers.
- **Fare by Survival:** The median fare for survivors is visibly higher than for non-survivors.