Titanic Dataset – Exploratory Data Analysis (EDA) Report

📌 Overview

The Titanic dataset contains details about passengers aboard the RMS Titanic, including socio-economic status, demographics, ticket information, and survival status.

EDA was conducted to understand feature distributions, missing values, relationships with the target variable (Survived), and to identify trends or patterns useful for predictive modeling.

📌 Data Summary

- Total observations: 891 (train.csv)
- Target variable: **Survived** (binary: 0 = No, 1 = Yes)
- Key features:
    - PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

---

📌 Missing Values

- **Age**: Significant missing values (~20%).
- **Cabin**: Highly missing (>75% missing) ➜ Dropped for EDA.
- **Embarked**: 2 missing values ➜ Filled with mode.
- **Fare**: No missing in train.csv but test.csv may have missing ➜ Checked and filled as needed.

✅ *Action taken:*

- Age and Fare imputed with median.
- Embarked imputed with mode.
- Cabin, Name, and Ticket dropped due to high cardinality or irrelevance for basic EDA.

📌 Categorical Features Analysis

- **Sex**:
  - Male ~65%, Female ~35%.
  - Survival higher for females.
- **Embarked**:
  - Majority at 'S'.
  - Smaller numbers at 'C' and 'Q'.

✅ *Observation:*

- Sex is a strong predictor (clear survival difference).
- Embarked may relate to socio-economic status or cabin location.

📌 Numerical Features – Summary Statistics

- **Age**:
  - Median ~28 years.
  - Range ~0.42–80.
  - Slight right-skewed.
- **Fare**:

    - Median ~14.5.
    - Highly right-skewed with extreme outliers.
- **SibSp & Parch**:
    - Most values = 0 (passengers traveling alone).
    - Few large families.

✅ *Observation:*

- Fare and Age need normalization or binning for modeling.
- Family size may be engineered from SibSp and Parch.

📌 Histograms & Boxplots

- **Fare** shows strong right skew and outliers ➜ log-transform may help.
- **Age** has moderate outliers but mostly young to middle-aged adults.
- **SibSp** and **Parch** are sparse, mostly zero.

✅ *Observation:*

- Important to address outliers during preprocessing.

📌 Correlation Analysis

- Positive correlation between **Fare** and **Survived**.
- Negative correlation between **Pclass** and **Survived** (lower class → lower survival).
- Weak correlation of **Age**, **SibSp**, **Parch** with **Survived** individually.

✅ *Observation:*

- Fare and Pclass are strong candidates for modeling.

- Family features may need feature engineering for better signal.

📌 Pairplots

- Clear separation of survival by **Fare** and **Pclass**.
- Weak trends with **Age** and other variables.

✅ *Observation:*

- Suggests simple linear boundaries may not fully capture relationships.

📌 Scatterplots vs Target

- **Fare**: Survivors generally paid higher fares.
- **Age**: Survivors skew slightly younger.
- **Pclass**: Clearer separation; 1st class had higher survival.
- **SibSp**, **Parch**: No strong linear trend with survival, but low family sizes may be safer.

✅ *Observation:*

- Confirms importance of socio-economic status in survival.

📌 Countplots for Categorical Variables

- Sex distribution imbalanced but survival favors females.
- Embarked distribution dominated by 'S'.

✅ *Observation:*

- Sex and Embarked may be useful categorical predictors.

📌 Boxplots by Survival

- **Fare**: Survivors have much higher median Fare.
- **Age**: Slightly younger survivors.
- **Pclass**: Survivors heavily from 1st class.
- **SibSp**, **Parch**: Slight survival benefit with small family groups.

✅ *Observation:*

- Boxplots reveal clear differences in feature distributions between survivors and non-survivors.

📌 Final Insights & Recommendations

✅ **Strong predictors identified**:

- Pclass
- Fare
- Sex
- Embarked

✅ **Moderate predictors**:

- Age (especially after binning)
- Family size (engineered from SibSp + Parch)

✅ **Recommendations for Modeling**:

- Impute missing Age with median or predictive models.
- One-hot encode Embarked.
- Map Sex to binary.

- Log-transform Fare to reduce skew.
- Engineer FamilySize = SibSp + Parch + 1.
- Drop or engineer Cabin/Ticket if using advanced models.

☑ Conclusion:

The EDA confirms socio-economic status, fare price, and gender were major survival determinants. Careful preprocessing and feature engineering will improve model performance on this dataset.

📌 Author:

Exploratory Data Analysis Report prepared in Python using Pandas, Matplotlib, and Seaborn.

PREPARED BY :

CHITRARTH VASDEV