Exploratory Data Analysis (EDA) Report

(Titanic Dataset)

1. Objective

This report presents findings from an Exploratory Data Analysis (EDA) of the Titanic training dataset (train.csv) using Python (Pandas, Matplotlib, Seaborn). The goal is to identify patterns, trends, and anomalies related to passenger survival.

2. Dataset Overview

The Titanic training dataset contains 891 passenger records with 12 features:

- **PassengerId**: Unique ID
- **Survived**: 0 (No), 1 (Yes)
- **Pclass**: Passenger class (1, 2, 3)
- **Sex**: Male, Female
- **Age:** Age in years
- **SibSp**: Siblings/spouses aboard
- **Parch**: Parents/children aboard
- **Fare**: Ticket fare
- **Embarked**: Port (C = Cherbourg, Q = Queenstown, S = Southampton)
- **Name, Ticket, Cabin**: Additional details

Missing values: Age (177, 20%), Cabin (687, 77%), Embarked (2, 0.2%)

3. Data Preprocessing

- Dropped Cabin due to 77% missing values.
- Filled Age with median (28).
- Filled Embarked with mode ('S')

4. Statistical Summary

- **Survival**: 38.4% survived (342), 61.6% did not (549).
- **Sex**: 64.8% male (577), 35.2% female (314).
- **Pclass**: 55% 3rd class (491), 24% 1st (216), 21% 2nd (184).
- **Embarked**: 72% Southampton (646), 19% Cherbourg (168), 9% Queenstown (77).
- **Age**: Mean 29.7, median 28, range 0.42–80.
- **Fare**: Mean $32.2, median $14.5, max $512 (skewed).

5. Visual Analysis

- Age and Fare Distributions
  - **Age**: Right-skewed, most passengers 20–40, peak at 28.
  - **Fare**: Highly right-skewed, most $50).
- Age and Fare by Pclass
  - **Age**: 1st class older (median 37), 3rd class younger (median 25).
  - **Fare**: 1st class median $60, 3rd class $8, with outliers in 1st class.
- Age vs. Fare by Survival
  - Higher fares linked to survival, especially in 1st class.
  - Younger passengers have slightly higher survival rates.
- Pairwise Relationships
  - Higher fares correlate with survival (Fare > $50).
  - Small family sizes (1–2 SibSp/Parch) linked to higher survival.
- Correlation Heatmap
  - **Pclass** vs. **Fare**: Strong negative correlation (-0.55).
  - **Survived** vs. **Fare**: Weak positive correlation (0.26).
  - **Pclass** vs. **Survived**: Weak negative correlation (-0.34).
- Survival Analysis
  - **Sex**: Females 74% survival (233/314), males 19% (109/577).
  - **Pclass**: 1st class 63% (136/216), 2nd 47% (87/184), 3rd 24% (119/491).
  - **Embarked**: Cherbourg 55% (93/168), Queenstown 39% (30/77), Southampton 34% (219/646)

6. Summary of Findings

- Survival Patterns: Females, 1st class passengers, and Cherbourg boarders had higher survival rates, likely due to lifeboat prioritization.
- Age and Fare: Younger passengers and higher fares (1st class) linked to survival. Fare is highly skewed with outliers.
- Family Size: 1–2 siblings/parents associated with higher survival.
- Anomalies: Outliers in Fare ($512) and missing Age (20%) and Cabin (77%).
- Trends: Higher fares and better class strongly correlate with survival; younger age and smaller family size improve survival odds.

7. Conclusion

The EDA shows that women, first-class passengers, and those boarding at Cherbourg were more likely to survive. Higher fares and smaller families also increased survival chances. These findings can help create new features and build models to predict survival for the test set.