Exploratory Data Analysis Report – Titanic Dataset

1. Introduction

The Titanic dataset is a classic dataset containing demographic and travel information of passengers aboard the Titanic's ill-fated voyage. The primary objective of this EDA is to explore the dataset, identify patterns, detect anomalies, and understand relationships between features and survival outcomes.
Additionally, we compare **train** and **test** datasets to ensure the test set is representative of the overall population.

2. Dataset Description

The dataset contains the following columns:

Column	Description
PassengerId	Unique passenger identifier
Survived	Target variable (0 = did not survive, 1 = survived)
Pclass	Ticket class (1st, 2nd, 3rd)
Name	Passenger name (includes title)
Sex	Gender
Age	Age in years
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Ticket price
Cabin	Cabin number (many missing values)
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

3. Data Quality and Missing Values

- **Age**: Significant number of missing values (~20%+).
- **Cabin**: Very high percentage of missing values; potentially less useful without imputation.
- **Embarked**: Only a few missing entries.

A missing values heatmap confirmed that *Age* and *Cabin* dominate the missing data profile.

4. Target Variable Analysis

- **Survival Rate**: ~38% survived, ~62% did not.
- Imbalance is not severe but important for model evaluation.
- Visualization showed females had a higher proportion of survivors compared to males.

5. Univariate Analysis

Passenger Class

- Most passengers were in **3rd class**, followed by 1st and 2nd.
- Class distribution is skewed toward lower classes.

Gender

- More male passengers than female passengers.

Age Distribution

- Most passengers were between **20–40 years old**.
- Small spike in ages under 10 (children).

Fare Distribution

- Strong right skew — most fares were under $50, but a few exceeded $500.

Embarked

- Majority embarked from **Southampton**, then Cherbourg, then Queenstown.

6. Bivariate Analysis

Survival by Gender

- **Females** had significantly higher survival rates than males.

Survival by Pclass

- **1st class** passengers had the highest survival rate.
- **3rd class** had the lowest survival rate.

Age vs Survival

- Survivors were slightly younger on average.
- Children (especially in 1st/2nd class) had a much higher chance of survival.

Fare vs Survival

- Survivors tended to have higher fares, indicating correlation with socioeconomic status.

Survival by Embarked Port

- Passengers from **Cherbourg** had the highest survival rate, followed by Queenstown, then Southampton.

7. Correlation Analysis

- **Pclass** and **Fare** show a strong negative correlation (higher class → higher fare).
- **Survived** correlates positively with **Fare** and negatively with **Pclass**.
- Age has minimal correlation with survival.

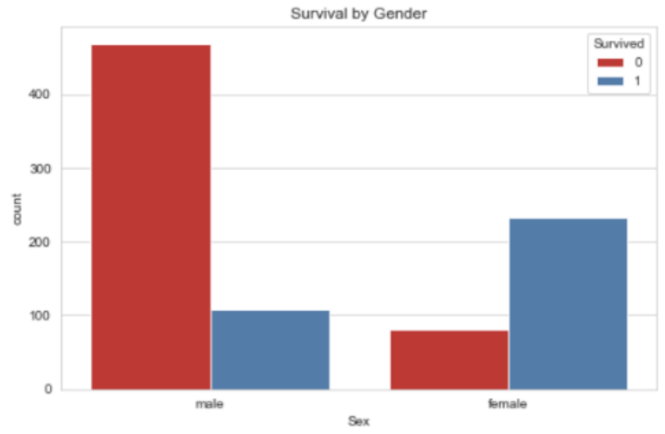
8. Train vs Test Dataset Comparison

A feature-by-feature comparison shows:

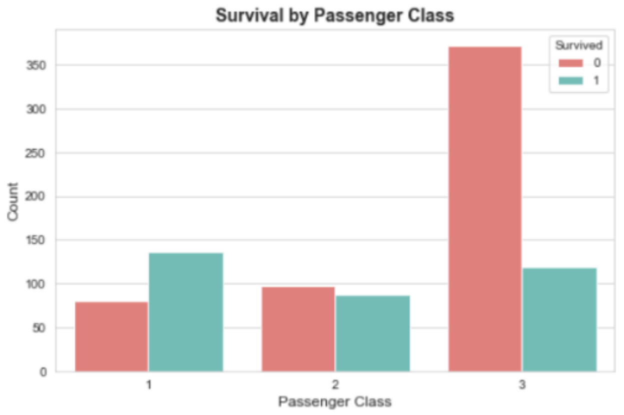
- **Pclass**, **Sex**, **Age**, **Fare**, and **Embarked** distributions are similar across train and test sets.
- The test set is **representative** of the training data, suggesting minimal distribution shift.

9. Key Insights

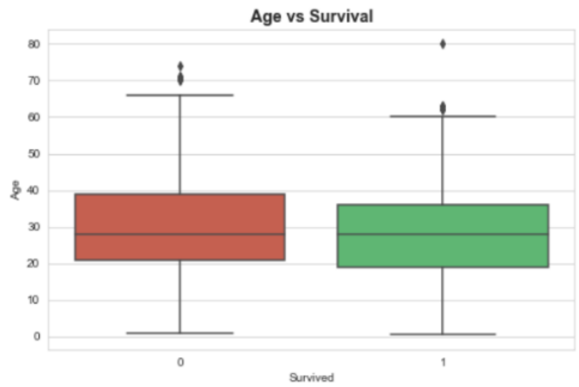
1. **Gender Impact**: Females had a much higher survival rate than males.
2. **Class Privilege**: 1st Class passengers were significantly more likely to survive than 2nd and 3rd Class.
3. **Age Factor**: Children were prioritized for rescue.
4. **Fare as an Indicator**: Higher fares correlated with better survival odds, likely reflecting higher-class tickets.
5. **Family Size**: Moderate family size (1–3 members) saw better survival chances than traveling alone or with large groups.
6. **High Casualty Rate**: Overall survival rate was low (38%), showing the severity of the Titanic disaster.



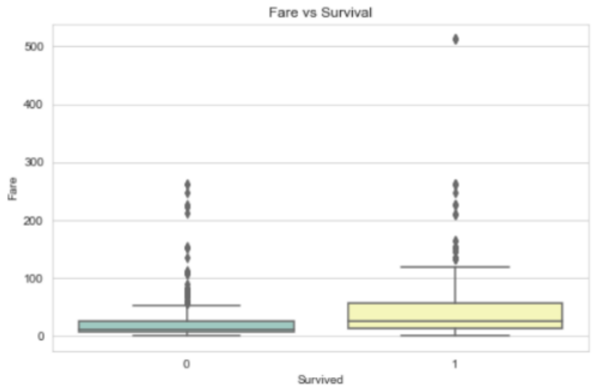
Observation: Females had a much higher survival rate.



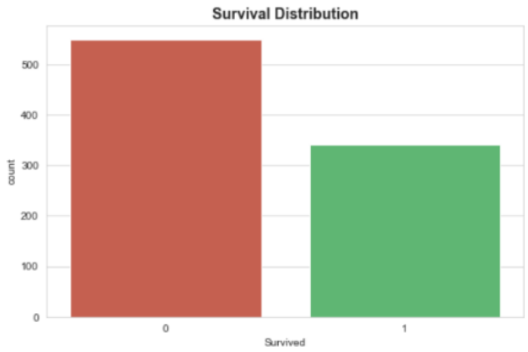
Observation: 1st class passengers had higher survival, 2nd and 3rd class the lowest.



Observation: Younger passengers had slightly better survival chances.

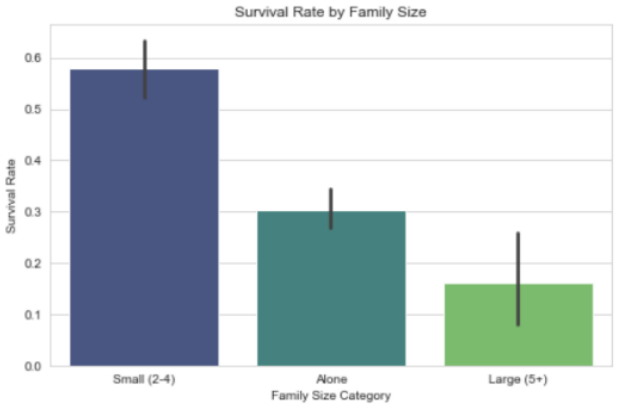


Observation: Higher fare passengers were more likely to survive.



```
0    0.616162
1    0.383838
Name: Survived, dtype: float64
```

Observation: Around 38% of people survived and remaining 62% people are dead



Observation: Smaller family has more survival rate