Exploratory Data Analysis (EDA) Report

Date: June 9, 2025 **Author:** Tanisha Singh **Dataset:** Titanic

1. Introduction

This report presents the findings from an Exploratory Data Analysis (EDA) performed on the **Titanic dataset**. The objective of this EDA is to understand the underlying structure of the data, identify patterns, relationships, and anomalies, and prepare the data for further analysis or machine learning model building, particularly for predicting passenger survival.

2. Data Overview

To begin, we'd load the dataset using pandas and get a first look at its structure.

Python

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


# Load the Titanic dataset

df = pd.read_csv('titanic.csv') # Assuming the dataset is named 'titanic.csv'
```

2.1 Dataset Information

Running df.info() would typically reveal the following:

- **Number of Rows:** Approximately 891 (standard for the training set).
- **Number of Columns:** Approximately 12 (e.g., PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked).
- **Data Types:** A mix of numerical (int64, float64) and categorical (object). For example, Age and Fare are numerical, while Sex and Embarked are categorical.
- **Missing Values:** We'd typically find missing values in Age, Cabin, and Embarked. Cabin often has a very high percentage of missing values.

2.2 Descriptive Statistics

Using df.describe() for numerical columns would show:

- **Age:** Mean age around 29-30 years, with a significant standard deviation (approx. 14 years), and values ranging from 0.42 to 80. The 25th percentile is around 20, median around 28, and 75th percentile around 38. This suggests a varied age distribution among passengers.

- **Fare:** Mean fare around $32, but a very high standard deviation (approx. $50) and a maximum value of over $500. The median fare is much lower (around $14), indicating a heavily **right-skewed distribution** with a few passengers paying very high fares.
- **Survived:** (If treated numerically as 0/1) The mean would represent the survival rate, typically around 0.38 (38% survival).
- **Pclass:** (Passenger Class) Mean around 2.3, indicating more passengers in 2nd and 3rd class than 1st.

For categorical columns, df.value_counts() would show:

- **Sex:** Overwhelmingly more 'male' passengers than 'female'.
- **Embarked:** 'Southampton' (S) is the most frequent embarkation port, followed by 'Cherbourg' (C) and 'Queenstown' (Q).
- **Pclass:** '3rd class' has the highest number of passengers, followed by '1st class' and '2nd class'.

---

3. Visual Exploration and Observations

This section would include the visualizations generated in the Jupyter Notebook, along with their respective observations.

3.1 Distribution of Numerical Features

- **Histograms for Age and Fare:**
  - **Observation (Age):** The histogram for Age often shows a relatively normal distribution but with a peak in younger adults (20-40 years) and a noticeable right skew, suggesting more younger individuals. There might be small peaks for children.
  - **Observation (Fare):** The histogram for Fare is **highly right-skewed**, with a large number of passengers paying very low fares and a long tail extending to very high fares, confirming the observation from describe().
- **Boxplots for Age and Fare:**
  - **Observation (Age):** The boxplot for Age would show some **outliers** (very young or very old passengers), but the main body of the data is concentrated.
  - **Observation (Fare):** The boxplot for Fare would clearly illustrate numerous **extreme outliers** on the higher end, reinforcing the skewed distribution and presence of very expensive tickets.

3.2 Distribution of Categorical Features

- **Countplots for Sex, Pclass, and Embarked:**
  - **Observation (Sex):** The countplot for Sex would dramatically show that **males vastly outnumber females** on the Titanic.
  - **Observation (Pclass):** The countplot for Pclass would confirm that **3rd class had the most passengers**, followed by 1st and then 2nd class.

    - **Observation (Embarked):** The countplot for Embarked would highlight **Southampton (S) as the primary embarkation port**, followed by Cherbourg (C) and Queenstown (Q).

3.3 Relationships and Trends

- **Pairplot of Numerical Features (e.g., Age, Fare, SibSp, Parch, Survived):**
    - **Observation:** The pairplot would reveal various relationships. For instance, Fare and Pclass would show an inverse relationship (lower class generally means lower fare). SibSp (siblings/spouse aboard) and Parch (parents/children aboard) would often be low, indicating most people traveled alone or with very few family members. The distributions of Age and Fare for Survived (0 for perished, 1 for survived) would be particularly insightful.
- **Correlation Heatmap:**
    - **Observation:** The heatmap of numerical features (Age, Fare, SibSp, Parch, Pclass, Survived) would likely show:
        - A **moderate negative correlation** between Pclass and Fare (higher class = higher fare).
        - A **negative correlation** between Pclass and Survived (lower class = lower survival rate). This is a crucial finding.
        - A **positive correlation** between Fare and Survived (higher fare = higher survival rate).
        - Weaker correlations for Age, SibSp, and Parch with Survived.
- **Survival Rates by Categorical Features (Bar Plots):**
    - **Survival by Sex:**
        - **Observation:** A bar plot of Survived vs. Sex would unequivocally show that **females had a significantly higher survival rate than males**. This is one of the most impactful findings.
    - **Survival by Pclass:**
        - **Observation:** A bar plot of Survived vs. Pclass would show a clear trend: **1st class passengers had the highest survival rate, followed by 2nd, and then 3rd class with the lowest**. This is consistent with the Pclass-Survived correlation.
    - **Survival by Embarked:**
        - **Observation:** Survival rates might vary slightly by Embarked port, with Cherbourg (C) potentially having a higher survival rate than Southampton (S) or Queenstown (Q), possibly due to the demographics of passengers from those ports (e.g., more 1st class passengers from Cherbourg).
- **Scatterplot: Age vs. Fare with Survived hue:**

    - **Observation:** This plot could show clusters. For example, passengers who paid higher fares and were in certain age groups (e.g., younger adults, older adults) might have had better survival chances. It can visually confirm that higher fares and specific age groups were linked to survival.

---

4. Summary of Findings

- **Key Observations:**
    - The **Titanic dataset** is rich with information about passengers, including demographics, travel class, and survival status.
    - **Missing values** are present, notably in Age and heavily in Cabin, requiring careful handling during preprocessing.
    - **Males significantly outnumbered females**, and **3rd class was the most populated**.
    - **Fare distribution is highly skewed**, indicating a large number of low-fare tickets and a few very high-fare tickets.
    - **Crucial factors influencing survival** were clearly identified:
        - **Sex:** Females had a remarkably higher survival rate than males ("women and children first").
        - **Pclass:** 1st class passengers had the highest survival rate, followed by 2nd, and then 3rd class, suggesting a strong correlation between socio-economic status/location on the ship and survival.
        - **Fare:** Higher fares were associated with higher survival rates, likely due to the correlation between fare and passenger class.
- **Potential Challenges/Areas for Further Investigation:**
    - The high number of **missing Cabin values** makes it difficult to use directly, but perhaps the presence/absence of a cabin number could be a feature.
    - **Age missing values** will need imputation.
    - **Fare's skewed distribution** might benefit from transformation (e.g., log transformation) for modeling purposes.
    - The **strong correlation between Pclass and Fare** should be noted for potential multicollinearity if both are used as features in a model.
- **Next Steps:**
    - **Data Cleaning:** Handle missing values (e.g., impute Age, drop Cabin or create a 'has_cabin' feature, impute Embarked).
    - **Feature Engineering:** Create new features that might capture more information (e.g., FamilySize from SibSp and Parch, IsAlone).
    - **Categorical Encoding:** Convert categorical variables (Sex, Embarked, Pclass if treated as categorical) into numerical representations (e.g., one-hot encoding).

- **Feature Scaling:** Scale numerical features like Age and Fare if using distance-based machine learning algorithms.