Exploratory Data Analysis – Titanic Dataset

Enhanced Summary of Findings

🔍 Dataset Overview

Our exploration of the Titanic dataset, comprising **891 passengers**, reveals a rich set of features including demographic information (**Age, Sex, SibSp, Parch**), socioeconomic indicators (**Pclass, Fare**), and travel details (**Embarked, Cabin**). The primary focus of our analysis is the binary target variable, **survived** (0 = No, 1 = Yes), allowing us to understand the factors influencing passenger outcomes.

📊 Key Statistical Insights

A preliminary statistical overview provides crucial context:

- **Age:** The average age of passengers was approximately **29.7 years**, but the distribution exhibits **outliers** and a significant proportion of **missing values (~20%)**. This missingness will require careful consideration during the data preprocessing stage.
- **Fare:** The fare distribution is **highly right-skewed**, indicating that most passengers paid relatively lower fares, with a few paying significantly more. The median fare is likely a more representative measure of central tendency.
- **Sex Distribution:** The passenger population was predominantly **male (~65%)**, with females comprising approximately **35%**. It is important to consider this imbalance in relation to survival rates.
- **Survival Rate:** Overall, only about **38%** of passengers survived the disaster, while **62%** did not. This sets the baseline for understanding which groups fared better.

📈 Key Patterns & Trends

Our analysis reveals compelling relationships between various features and the likelihood of survival:

- **Sex vs Survival:** The starkest contrast in survival rates is observed between genders. **Females had a remarkably higher survival rate of approximately 75%**, compared to a significantly lower **~20% survival rate for males**. This strongly suggests a "women and children first" protocol was in effect.
- **Pclass vs Survival:** A clear hierarchical trend emerges with passenger class:
    - **1st Class:** Approximately **63%** survival rate.
    - **2nd Class:** Approximately **47%** survival rate.
    - **3rd Class:** Approximately **24%** survival rate. This unequivocally demonstrates a strong positive correlation between higher socioeconomic status (as indicated by passenger class) and improved survival odds.
- **Age vs Survival:** Examining age patterns reveals:
    - **Children (< 10 years old) exhibited higher survival rates**, suggesting prioritization during the evacuation.
    - **Elderly passengers (> 60 years old) were less likely to survive**, possibly due to frailty and slower mobility. Further investigation into specific age ranges could reveal more nuanced patterns.
- **Fare vs Survival:** Passengers who paid **higher fares generally had better survival chances**. This likely correlates with their passenger class, reinforcing the impact of socioeconomic status.
- **Embarked Port:** The majority of passengers boarded from **Southampton**. Interestingly, passengers who embarked from **Cherbourg showed a slightly higher survival rate**. This could be attributed to the distribution of passenger classes originating from each port, warranting further investigation into the class composition per embarkation point.

Correlation Analysis (Implicit): While not explicitly stated, the observed relationships between Pclass and Fare, and their subsequent impact on survival, suggest underlying correlations that could be further quantified using methods like Pearson or Spearman correlation coefficients

🩹 Missing Values

Addressing missing data is crucial for robust modeling:

- **'Age':** The significant **~20% of missing 'Age' values** necessitates imputation using appropriate strategies (e.g., mean, median, or more sophisticated techniques based on other features).
- **'Cabin':** The **heavily missing 'Cabin' information** presents a challenge. It might be beneficial to engineer a new feature indicating the presence or absence of a cabin assignment, or potentially extracting information from the prefix of the cabin number if a pattern exists related to location on the ship. Alternatively, given the high degree of missingness, it might be dropped.
- **'Embarked':** The **2 missing 'Embarked' values** are easily imputable using the mode or by examining patterns in other related features.

📌 Conclusion & Next Steps

Our exploratory data analysis reveals that **Sex, Passenger Class (Pclass), and Fare are strong predictors of survival on the Titanic**. The "women and children first" protocol appears to have significantly influenced survival outcomes. Furthermore, socioeconomic status, as reflected by Pclass and Fare, played a critical role. The embarkation port also shows a potential link to survival, likely mediated by passenger class distribution.

Moving forward, the following steps are recommended:

- **Implement appropriate data cleaning techniques** to handle missing values in 'Age' and 'Embarked'. Strategically address the 'Cabin' variable through feature engineering or removal.
- **Further explore the relationships between features** through visualization techniques (e.g., histograms, box plots, scatter plots) and statistical tests (e.g., chi-squared tests for categorical variables, t-tests or ANOVA for numerical vs. categorical).
- **Investigate potential interactions between features** (e.g., the combined effect of Sex and Pclass on survival).
- **Consider creating new features** based on existing ones (e.g., family size from SibSp and Parch, age groups).