
📊 **Exploratory Data Analysis Report: Titanic Passenger Survival Trends**

Prepared by: Tadishetty Sriram

Role: Data Analyst Intern

Date: October 26, 2023

---

1. Executive Summary

The objective of this analysis was to investigate the train.csv dataset to identify the primary factors influencing passenger survival on the Titanic. By utilizing statistical summaries and advanced visualizations (heatmaps, pairplots), we have determined that **socio-economic status (Fare/Class)** and **demographics (Gender/Age)** were the most significant predictors of survival.

---

2. Data Structure & Quality Inspection

Before generating visuals, we performed a statistical health check using .info() and .describe() methods.

- **Dataset Volume:** The analysis was performed on 891 passenger records with 12 features.
- **Missing Values:** Significant gaps were identified in the **Age** (~19% missing) and **Cabin** columns.
    - *Action Taken:* Missing ages were imputed using the median age (28.0) to maintain statistical integrity for visualization. Two missing 'Embarked' values were filled with the mode.
- **Statistical Spread:**
    - **Survival Rate:** Only ~38% of passengers in this training set survived.
    - **Age:** Ranged from infants (0.42 years) to elderly (80 years), with an average age of roughly 29 years.
    - **Fares:** Highly skewed; the median fare was $14.45, but the maximum was $512.32, indicating extreme wealth disparity.

---

3. Visual Analysis & Observations

A. Univariate Analysis (Distributions)

Method: Histograms & Value Counts

- **Age Distribution:** The histogram reveals a roughly normal distribution with a slight right skew. A notable peak exists in the 20-30 age range, indicating a young passenger demographic.
- **Fare Distribution:** The fare histogram is heavily right-skewed. The vast majority of tickets were sold for under $50, with very few outliers purchasing tickets above $200.
- **Survival Counts:** The .value_counts() analysis confirms a mortality imbalance: 549 deaths vs. 342 survivors.

B. Bivariate Analysis (Relationships)

Method: Boxplots & Scatterplots

- **Age vs. Passenger Class (Boxplot):**
    - *Observation:* There is a clear relationship between Age and Class. First-class passengers were significantly older (median age ~37) compared to Third-class passengers (median age ~24). This suggests that wealth and status were accumulated with age.
- **Age vs. Fare (Scatterplot):**
    - *Observation:* While there is no linear correlation between Age and Fare across the board, the scatterplot colored by 'Survival' shows a dense cluster of survivors in the high-fare region, regardless of their age.

C. Multivariate Analysis (Correlations)

Method: Pairplots & Heatmaps

- **Correlation Heatmap:**
    - *Key Finding 1:* **Fare** has the strongest positive correlation with **Survival** (+0.26). This indicates that paying more for a ticket directly increased the likelihood of rescue.
    - *Key Finding 2:* **Pclass** has a strong negative correlation with **Survival** (-0.34). Since 1st class is represented by the number '1' and 3rd class by '3', this

mathematically confirms that *lower* class numbers (higher status) had *higher* survival rates.

    - *Key Finding 3:* **Age** has a weak negative correlation with **Survival**, implying that being younger offered a slight advantage, likely due to the "children first" protocol.

- **Pairplot Analysis:**
    - The pairplot visually separates the dataset by survival status. It highlights that passengers with **small family sizes** (1-3 siblings/parents) had better survival outcomes than large families or solo travelers.

---

4. Summary of Findings

Based on the visual and statistical evidence, we conclude the following:

1. **Wealth was the #1 Survival Factor:** The strongest data trend is that First Class passengers (who paid higher fares) were prioritized during the evacuation.
2. **Gender & Protocol Adherence:** The "Women and Children First" maritime protocol was strictly followed, evidenced by the high survival rate of females and the slight advantage for younger passengers.
3. **Family Size Risk:** Large families faced a disadvantage, likely due to the logistical difficulty of staying together and moving quickly during the disaster.
