

Titanic Data Analysis Report

Elaborated Summary of Findings

After performing comprehensive Exploratory Data Analysis (EDA) on the Titanic dataset, several important patterns and relationships were discovered that significantly influenced passenger survival outcomes:

1. Higher Survival Rate Among Females:

The data clearly showed that female passengers had a much higher likelihood of surviving compared to males. This pattern aligns with the historical accounts of the Titanic disaster, where evacuation protocols prioritized "women and children first." The survival rate for females was substantially higher across all classes, indicating a strong gender-based prioritization during rescue efforts.

2. First-Class Passengers Had Better Survival Odds:

Passengers traveling in the 1st class had significantly higher survival rates than those in 2nd and 3rd class. This trend suggests that socioeconomic status impacted survival chances, possibly due to better access to lifeboats, proximity to deck areas, or preferential treatment during evacuation. Higher-class passengers may have had cabins located closer to lifeboats and benefited from greater resources or crew assistance.

3. Influence of Age and Fare on Survival:

Age emerged as an important factor; younger passengers, particularly children, were more likely to survive. This supports the notion that younger passengers were given priority during evacuation. Fare paid, which often correlates with passenger class and cabin location, also influenced survival chances - passengers who paid higher fares tended to have higher survival rates. This reflects the interplay between wealth, ticket price, and physical access to lifeboats.

4. Correlation Between Gender, Class, and Survival:

There is a notable combined effect of gender and class on survival probability. For example, females in 1st class had the highest survival rates, while males in lower classes faced the lowest

Titanic Data Analysis Report

chances. This multidimensional correlation highlights how social hierarchy and gender norms together affected survival outcomes.

5. Handling Missing Data:

The dataset had missing values, especially in the Age and Cabin columns. These missing values were carefully handled during analysis, either by imputing with appropriate statistical methods or by excluding from certain calculations to avoid bias. Proper treatment of missing data ensured that the insights and patterns derived were accurate and reliable.

6. Role of Visualizations:

Visual tools such as heatmaps, pairplots, and bar charts were invaluable in revealing clear patterns and relationships between variables. These visualizations helped to quickly identify correlations and trends, supporting deeper understanding beyond mere statistics. For example, heatmaps illustrated correlations between survival and multiple features, while pairplots helped visualize distributions and interactions across categories.

Elaborated Observations for Each Visual

Heatmap:

The heatmap clearly shows that the encoded variable for sex (with females coded higher) and fare have a positive correlation with the survival status. This means that being female and paying a higher fare are both strongly associated with increased chances of survival. The heatmap also helps visualize other variable relationships and highlights that survival is interconnected with multiple factors rather than isolated features.

Pairplot:

The pairplot illustrates the distribution and relationships between multiple variables simultaneously. It reveals that most survivors were females, generally younger passengers, and those who paid higher fares. The visual clusters show distinct groupings separating survivors from non-survivors, emphasizing that gender, age, and fare were key determinants in survival outcomes.

Titanic Data Analysis Report

Age Histogram:

The histogram of passenger ages indicates that the majority of passengers were between 20 and 40 years old. This age group forms the largest segment on the ship. The distribution is skewed toward adults rather than children or elderly, which helps contextualize other observations, such as why children had different survival odds despite being fewer in number.

Age vs Survival Boxplot:

The boxplot comparing age distributions of survivors versus non-survivors shows that survivors tended to be younger. The median age of survivors is lower, and the interquartile range is shifted toward younger ages. This confirms that younger passengers, including children and young adults, were prioritized or had better chances of rescue during the disaster.

Fare vs Age Scatterplot:

The scatterplot depicting the relationship between fare and age alongside survival status indicates that passengers who paid higher fares generally had better survival chances, regardless of their age. There are clusters of survivors in higher fare brackets, while those who paid less had lower survival rates. This visualization reinforces the link between socioeconomic status and access to lifeboats or rescue efforts.