**Titanic Survival Analysis: Enhanced Report**

**Date:** August 25, 2025

**1. Executive Summary**

This report presents a comprehensive analysis of the enhanced Titanic dataset to identify the primary factors influencing passenger survival. Through detailed data cleaning, feature engineering, and exploratory analysis, this investigation confirms that survival was not arbitrary but was significantly dictated by socio-demographic status. The most critical predictors of survival were **gender** and **passenger class**, with women in first and second class having the highest survival rates. Additional engineered features, such as cabin assignment and family size, provided further nuanced insights, revealing that passengers with cabins and those in small to medium-sized family groups had a greater chance of survival. The overall survival rate was a grim **31.2%**, underscoring the magnitude of the tragedy.

**2. Data and Methodology**

The analysis was conducted on an enhanced version of the Titanic dataset, utilizing the train\_enhanced.csv and test\_enhanced.csv files. The methodology, guided by the process in the titanicc.html notebook, involved three key phases:

* **Data Preparation and Cleaning:** The raw data was meticulously cleaned to ensure quality. This involved addressing a significant number of missing values in the Cabin feature by creating a new binary feature, HasCabin, to indicate whether a cabin was assigned. Data types for Age and Fare were converted to integers for consistency, and the dataset was verified to be free of duplicate entries.
* **Feature Engineering:** To uncover deeper patterns, several features were engineered from the original data:
  + **HasCabin**: A binary feature indicating cabin assignment.
  + **Ticket\_No**: A standardized version of the Ticket feature.
  + **Descriptive Categories**: Pclass and Embarked codes were converted to descriptive text (e.g., '1st', 'Cherbourg').
  + **Interaction Features**: New features like GenderClass (combining gender and class) and FamilyClass were created to analyze combined effects.
  + **Binned Features**: Continuous variables like Age and Fare were grouped into categories (AgeGroup, FareGroup) for clearer analysis.
* **Exploratory Data Analysis (EDA):** A comprehensive EDA was performed, including the generation of a detailed profile report using ydata\_profiling. This was followed by a targeted analysis to quantify the impact of various factors on survival rates, visualized through bar charts and pie charts.

**3. Detailed Analysis and Key Findings**

**3.1. Overall Survival Rate**

The baseline analysis of the combined passenger data revealed that only **31.2%** of individuals survived the disaster. This stark figure highlights that nearly seven out of every ten people on board did not survive.

**3.2. Primary Factors Influencing Survival**

Survival was strongly correlated with several key demographic and socio-economic factors:

* **Gender:** Gender was the most significant survival predictor. **Females had a survival rate of 60.5%**, while **males had a survival rate of only 15.3%**. This confirms the historical accounts of a "women and children first" evacuation policy.
* **Passenger Class (Pclass):** A clear socio-economic hierarchy was evident in survival outcomes. **First-class passengers had the highest survival rate at 49.5%**, followed by second-class at 39.1%, and a stark drop to **20.2% for third-class passengers**.
* **Cabin Assignment:** The presence of an assigned cabin was a strong indicator of survival. Passengers with a cabin (Had Cabin) had a **52.0%** survival rate, more than double the **25.0%** rate for those without. This is strongly correlated with passenger class, as first-class passengers were more likely to have assigned cabins.

**3.3. Analysis of Engineered and Interaction Features**

The engineered features provided a more granular understanding of survival dynamics:

* **Age Group:** Age played a critical role. **Children had the highest survival rate at 42.5%**, followed closely by teenagers at 39.8%. Adults had a 32.7% chance of survival, while seniors had the lowest rate at just 18.2%.
* **Embarkation Port:** The port of embarkation also showed a surprising correlation. Passengers who boarded at **Cherbourg had the highest survival rate (44.6%)**, compared to Queenstown (36.4%) and Southampton (27.1%). This may be linked to the higher proportion of first-class passengers boarding at Cherbourg.
* **Gender and Class Interaction (GenderClass):** Combining gender and class revealed that **second-class females had the highest survival rate of all groups at 76.3%**, slightly edging out first-class females (74.5%). In contrast, males in second and third class had the lowest survival rates (13.0% and 10.7%, respectively).

**4. Conclusion and Further Analysis**

The analysis unequivocally demonstrates that survival on the Titanic was not a matter of chance but was heavily stratified by social class and gender. First-class passengers, women, and children were given clear priority during the evacuation. The feature engineering process was instrumental in uncovering nuanced relationships, such as the high survival rate of second-class women and the impact of cabin assignments.

For further analysis, it is recommended to use these engineered features to build and train a predictive machine learning model. A classification model, such as a Logistic Regression or a Random Forest, could be developed to predict the survival of passengers from the test dataset, providing a practical application of the insights gained from this exploratory analysis.