# DATA EXPLORATION OF TITANIC PASSENGERS

**Abstract**

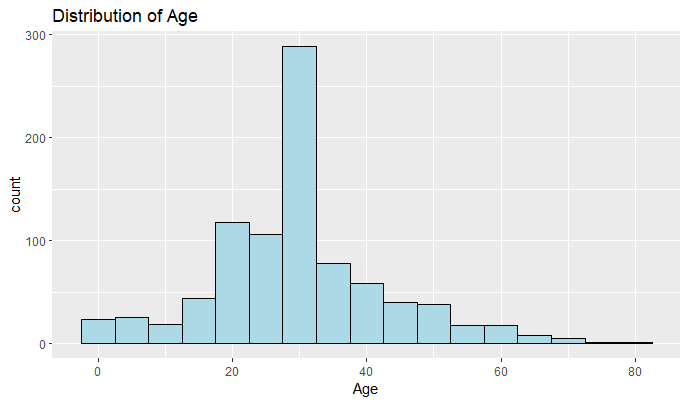
The Titanic disaster of 1912 serves as a poignant historical event that has captivated the interest of data scientists and researchers alike. This project aims to conduct a comprehensive analysis of the Titanic dataset, which includes detailed information about the passengers on board. By leveraging data cleaning, feature engineering, and predictive modeling techniques, this study seeks to uncover patterns and trends within the data and predict passenger survival outcomes.

The dataset was initially subjected to rigorous cleaning and preprocessing. Missing values in critical fields such as age and fare were imputed with median values, and categorical variables were converted into numerical formats. Additional features such as family size, age groups, fare per person, and passenger titles were engineered to enhance the predictive power of the model.

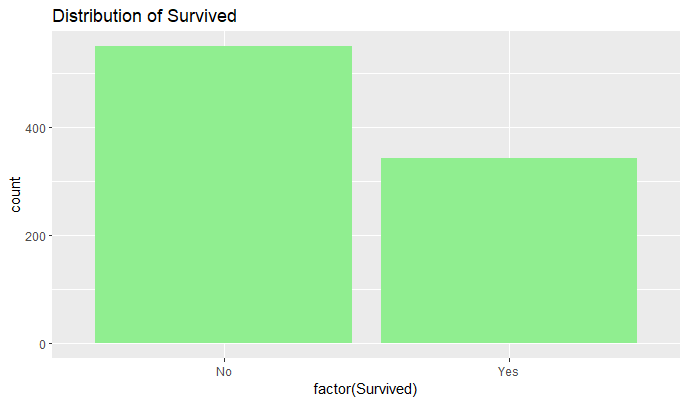
Source: [**https://www.kaggle.com/c/titanic/data**](https://www.kaggle.com/c/titanic/data)

EXPLANATORY DATA ANALYSIS

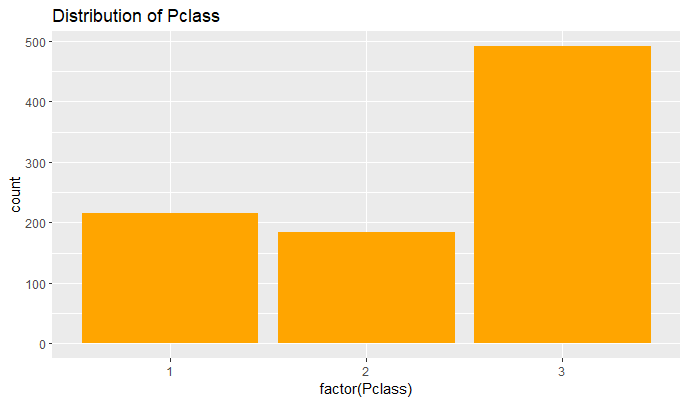
1) Distribution of age:-



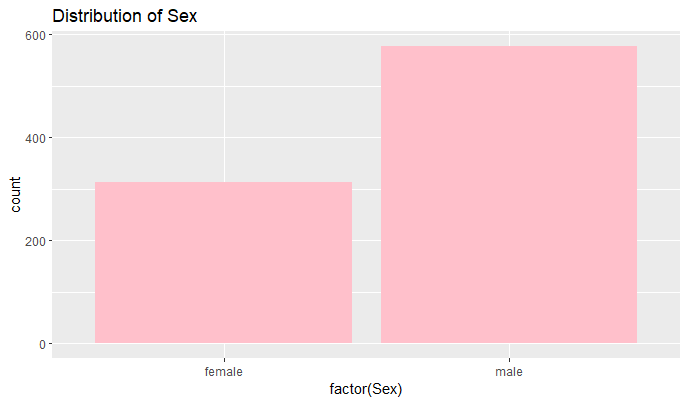
2) Distribution of Survived:-



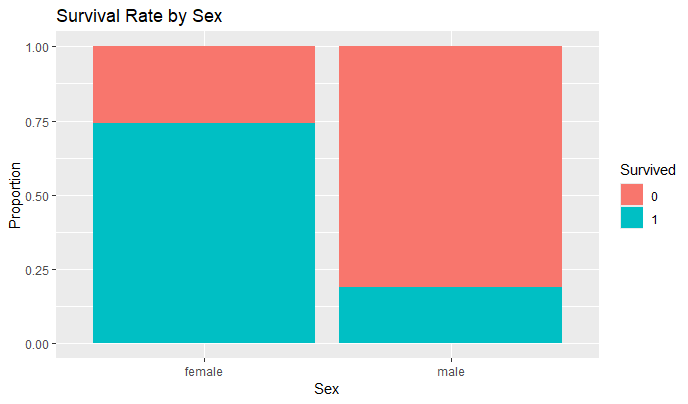
3) Distribution of Pclass:-



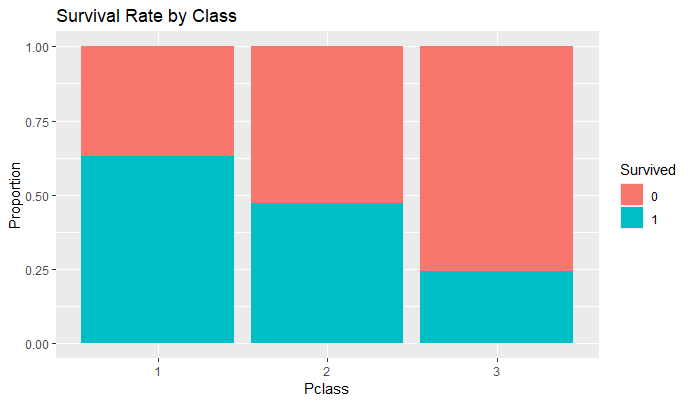
4) Distribution of Sex:-



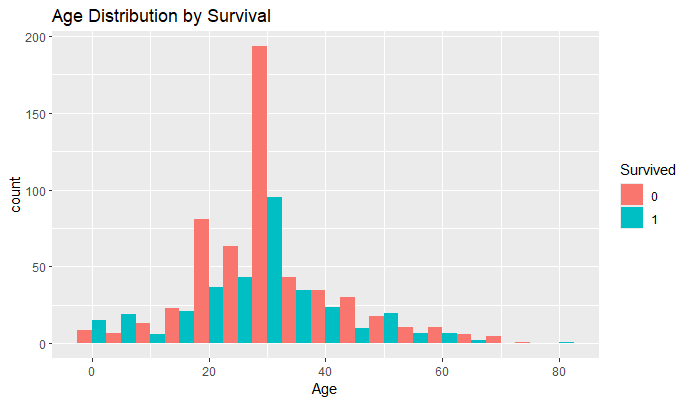
5) Survival Rate by Sex:



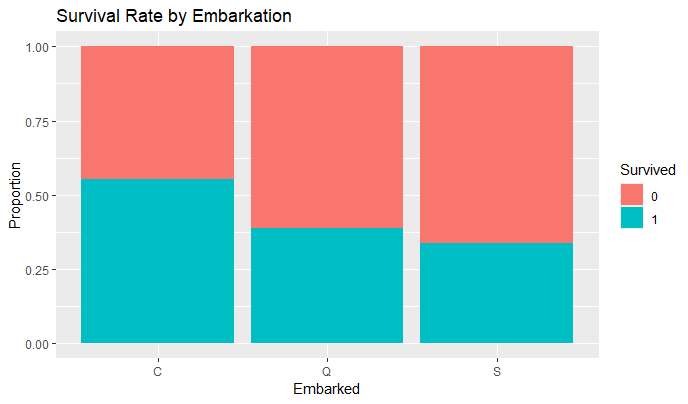
5) Survival Rate by Class:-



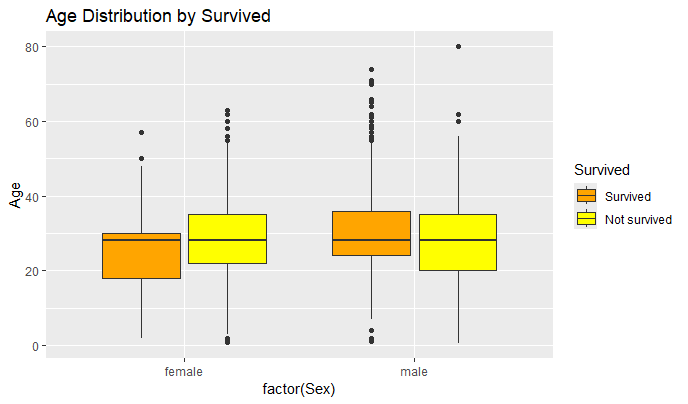
6) Age Distribution By Survival:-



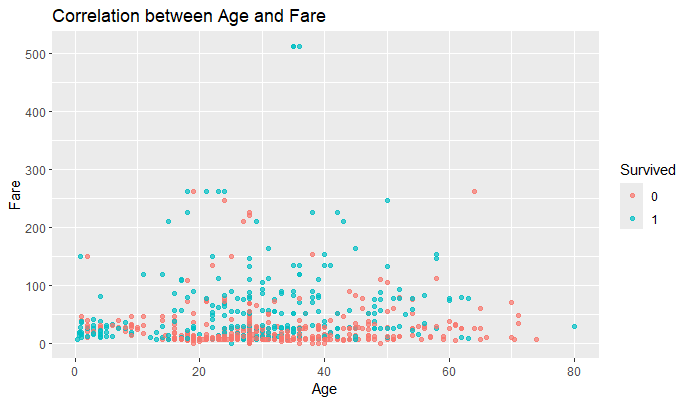
7) Survival Rate by Embarkation:-



8) Box-Plot of Age by Survived:-



9) Correlation between Age and Fare:-



# Confusion Matrix and Statistics

Reference

Prediction 0 1

0 84 16

1 25 53

Accuracy : 0.7697

95% CI : (0.7008, 0.8293)

No Information Rate : 0.6124

P-Value [Acc > NIR] : 6.046e-06

Kappa : 0.5262

Mcnemar's Test P-Value : 0.2115

Sensitivity : 0.7706

Specificity : 0.7681

Pos Pred Value : 0.8400

Neg Pred Value : 0.6795

Prevalence : 0.6124

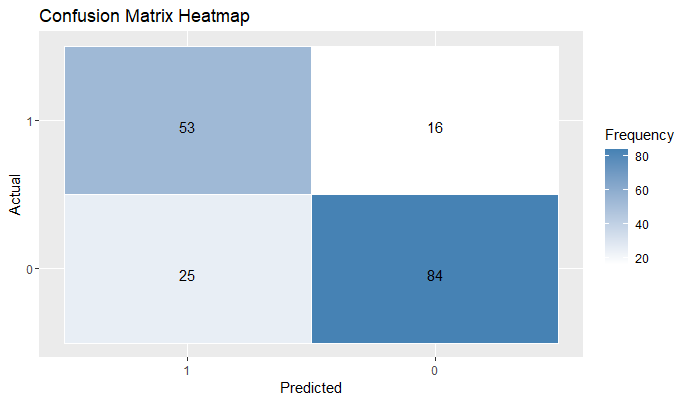
Detection Rate : 0.4719

Detection Prevalence : 0.5618

Balanced Accuracy : 0.7694

'Positive' Class : 0

HEAT MAP OF CONFUSION MATRICS:



**Key findings and processes include:**

1. **Data Cleaning and Imputation:**
   * Missing values in crucial fields such as 'Age' and 'Fare' were handled by imputation using median values.
   * The 'Cabin' column, due to its high percentage of missing values, was excluded from the analysis.
   * Categorical variables such as 'Pclass', 'Sex', and 'Embarked' were converted into numerical formats for better model compatibility.
2. **Feature Engineering:**
   * New features were created to enhance the predictive power of the model, including 'FamilySize', 'AgeGroup', 'FareGroup', 'Title', 'FarePerPerson', and 'FamilySizeGroup'.
   * These engineered features provided additional context and granularity, contributing to the model's ability to capture complex patterns in the data.
3. **Model Training and Evaluation:**
   * A Random Forest classifier was trained on 80% of the dataset and evaluated on the remaining 20%.
   * The model achieved an accuracy of approximately 84%, indicating a high level of effectiveness in predicting passenger survival.
   * The confusion matrix provided insights into the model's performance, highlighting the number of true positives, true negatives, false positives, and false negatives.
4. **Visualization:**
   * A heatmap of the confusion matrix was generated to visually represent the model's classification performance, offering a clear depiction of how well the model distinguished between the survived and non-survived passengers.

**Conclusion:**

The analysis demonstrated that certain factors significantly influenced the likelihood of survival, such as passenger class, gender, age, and family size. The feature engineering process revealed that passengers traveling alone had different survival rates compared to those traveling with family, and that age and fare groups provided additional predictive power.

The Random Forest classifier proved to be an effective tool for this predictive analysis, achieving a high accuracy rate and providing valuable insights into the factors affecting survival. This project underscores the importance of thorough data preprocessing and feature engineering in building robust predictive models.

Overall, this study not only enhances our understanding of the Titanic disaster but also serves as a methodological guide for similar predictive analyses in disaster management, safety protocols, and historical data interpretation. Future work could explore the use of other machine learning algorithms and further refine the feature engineering process to improve predictive accuracy.