Titanic_Data_Analysis

Norbert Dzikowski

2024-12-03

1. Data Analysis and Cleaning Report

The dataset appears to contain information about Titanic passengers. It includes attributes such as Passenger ID, survival status, class, name, gender, age, number of siblings/spouses aboard, parents/children aboard, ticket, fare, cabin, and embarkation port.

1.1. Data cleaning

Based on an initial inspection, the following data quality issues were identified:
1. Missing Values:
- Cabin: A significant number of missing values.
- Age: Some missing values.
- Embarked: Potentially contains missing values.
2. Data Type Issues:
- Columns such as Fare and Age should be numeric; verification is necessary.
- Embarked should have categorical values (C, Q, S).
3. Data Inconsistencies:
- Possible invalid or non-standard values in Embarked.
- Extreme or unrealistic values in Age (e.g., negative ages or ages above 120).
4. Textual Uniformity:
- Text fields such as Sex may have inconsistent capitalization (e.g., male vs. Male).

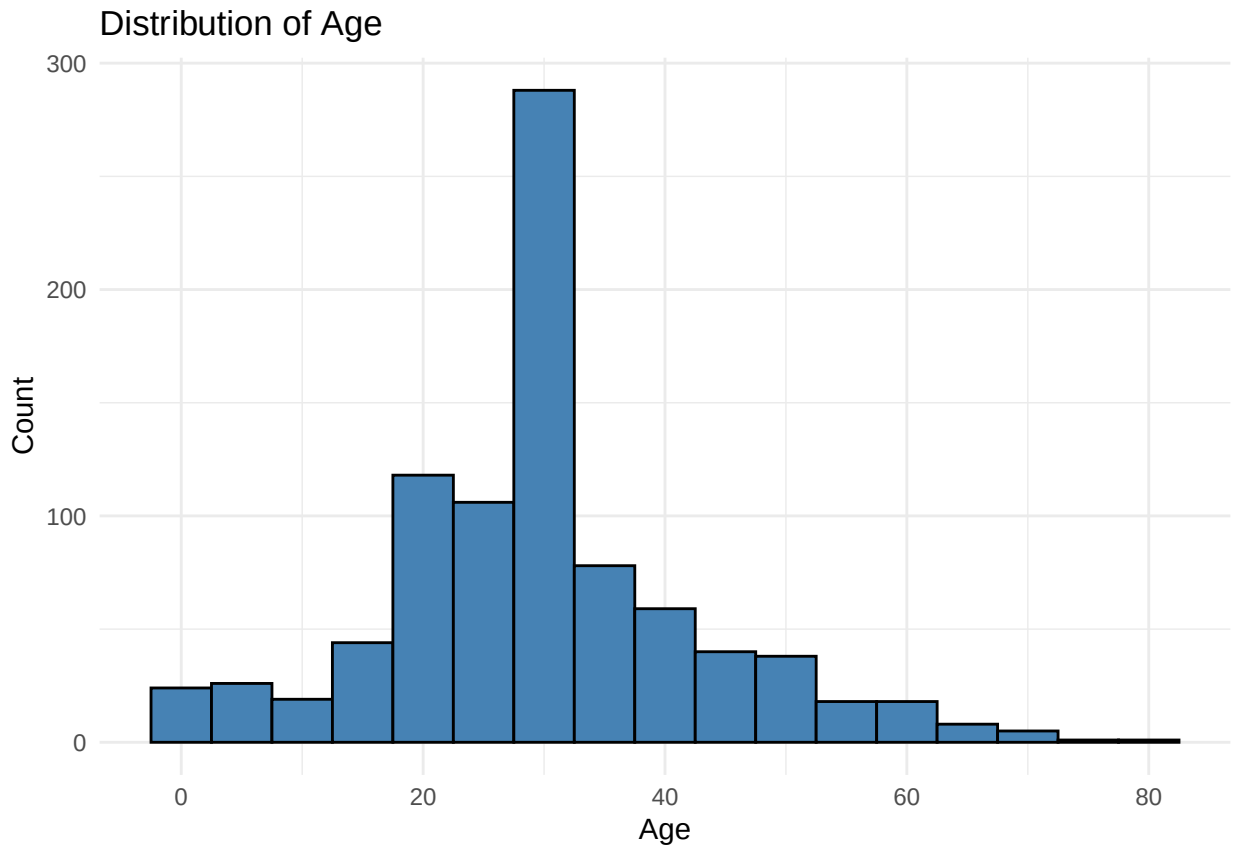
Cleaning Strategy:
1.Handling Missing Values:
- Age: Replace missing values with the median.
- Embarked: Replace missing values with the most frequently occurring value.
- Cabin: Remove the column if more than 50% of its values are missing.
2. Validating Data Consistency:
- Remove or correct unrealistic values in Age (e.g., less than 0 or greater than 120).
- Ensure only valid categories exist in Embarked (C, Q, S).
3. Standardizing Text Fields:
- Convert the Sex column to lowercase to ensure consistency.

2. Data Exploration and Visualizations

2.1. Distribution of Single Variables

2.1.1. Age Distribution

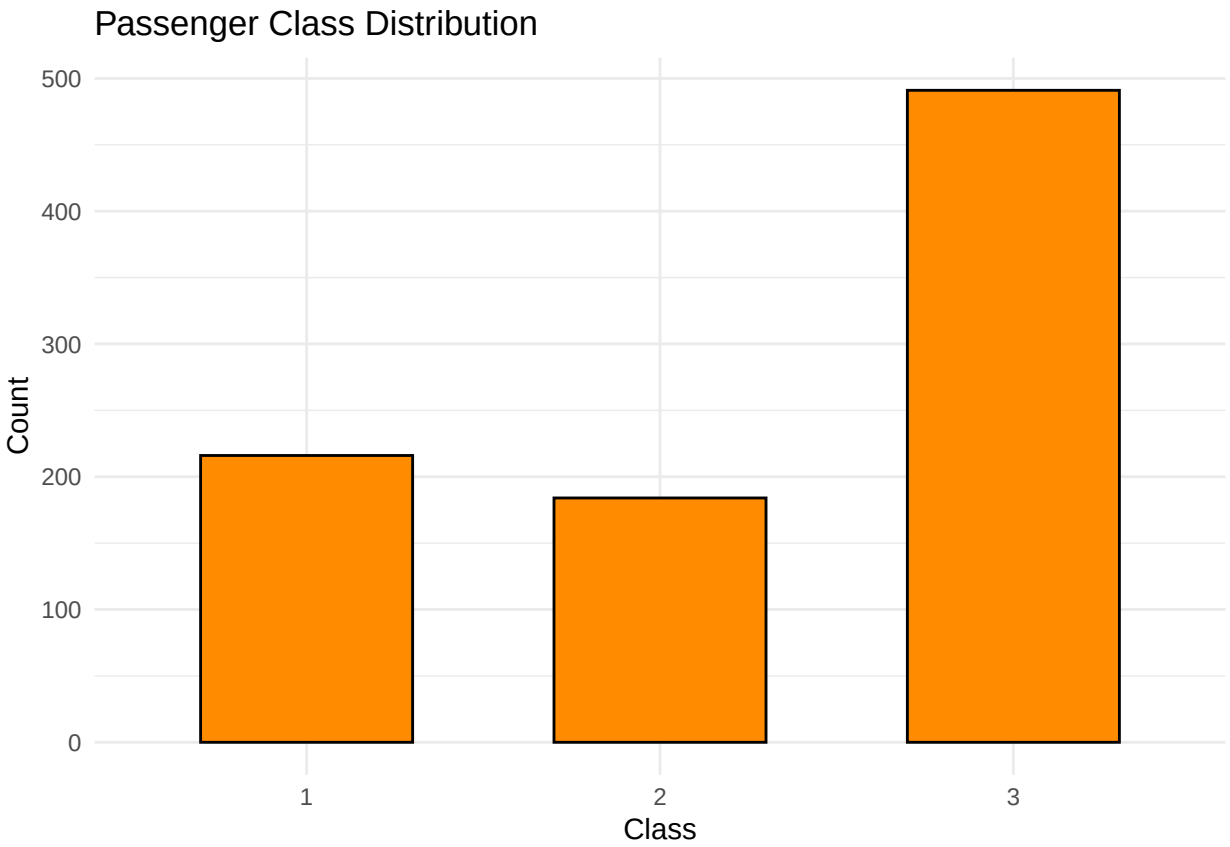
A histogram provides an overview of the age distribution among passengers.



This histogram displays the distribution of ages in the dataset, revealing that the majority of individuals fall within the age range of 20 to 30 years. This is evident from the tallest bar at the center of the chart, indicating that this age group is the most common in the population. The distribution is right-skewed, with a gradual decline in frequency as age increases. There are relatively few individuals above the age of 60, and the data spans a wide range of ages, from newborns (around 0 years) to approximately 80 years old. The overall pattern suggests a population dominated by younger individuals, with smaller representations of both children and elderly participants.

2.1.2. Passenger Class Distribution

A bar chart shows the frequency of passengers in each class.

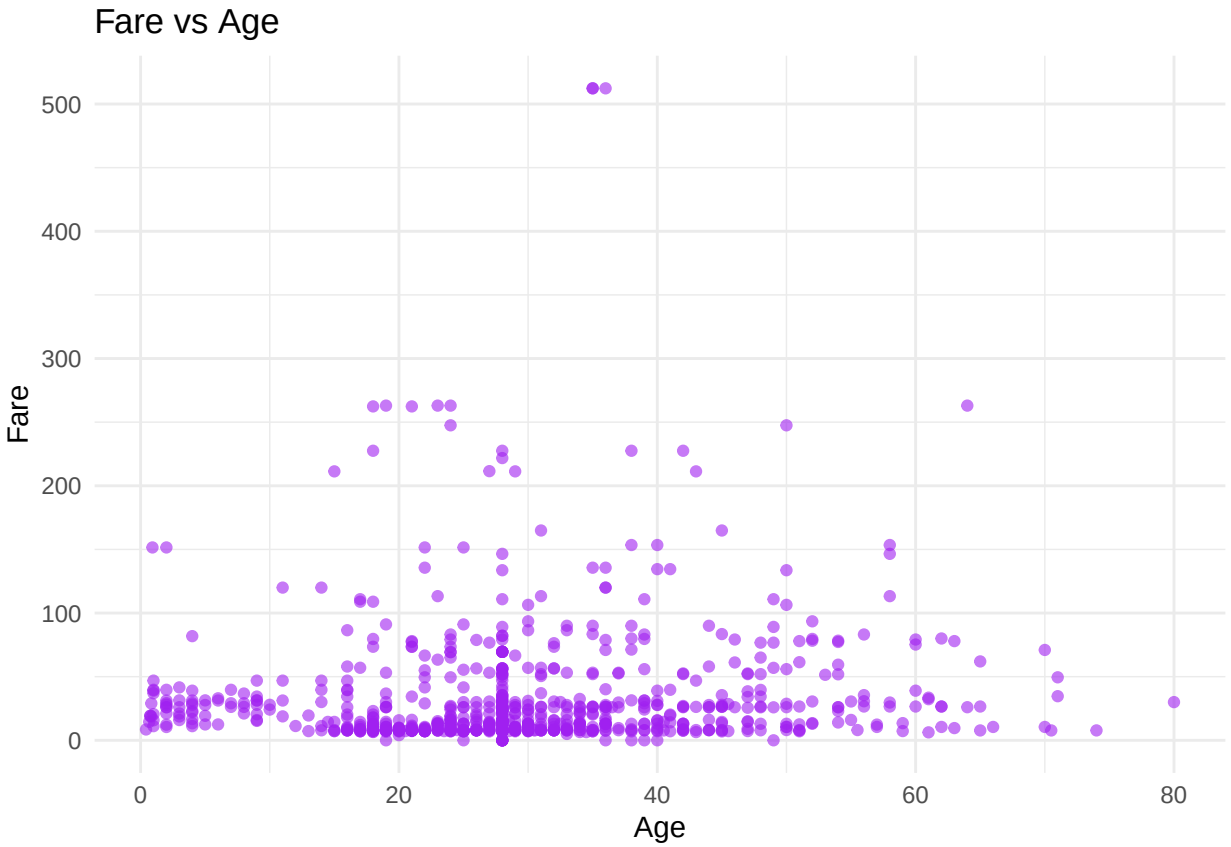


This bar chart illustrates the distribution of passengers across three classes. It is clear that the third class has the largest number of passengers, with a count exceeding 500, making it significantly more populous than the other classes. The first and second classes, on the other hand, have relatively similar counts, each hovering around 200 passengers. This distribution suggests a strong skew towards the third class, which could indicate a larger representation of economically lower-tier passengers.

2.2. Relationship Between Two Variables

2.2.1. Fare vs. Age

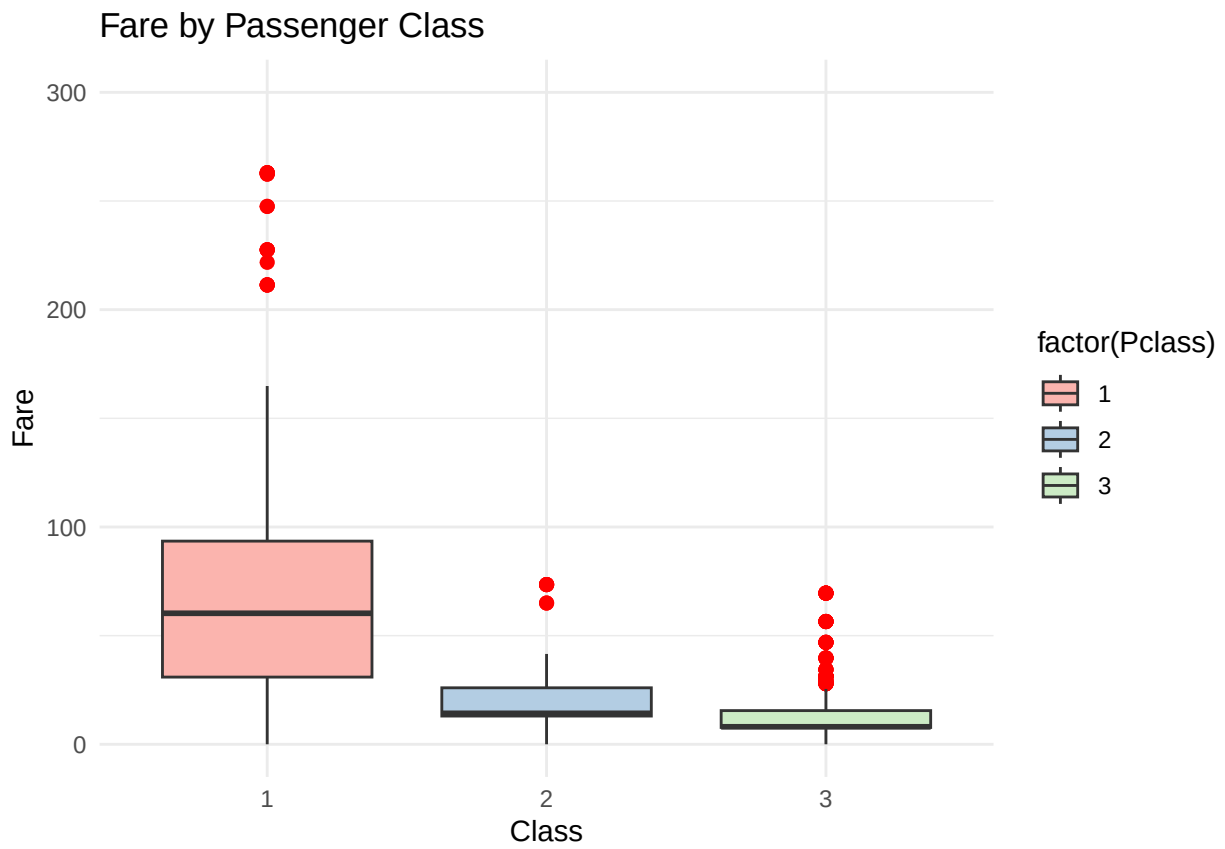
A scatter plot highlights the relationship between passenger age and fare paid.



This scatter plot illustrates the relationship between age and fare. The majority of data points are clustered near the bottom of the chart, indicating that most passengers paid lower fares. Additionally, the distribution of fares does not seem to have a strong correlation with age, as passengers of all age groups are represented across various fare levels. There are, however, a few outliers where very high fares, exceeding 500, were paid by a small number of individuals. These high fares are exceptions and stand out clearly in the plot. Younger individuals appear more evenly distributed across fare ranges, while older passengers seem concentrated in the lower fare categories.

2.2.2. Fare by Passenger Class

A boxplot shows the distribution of fares across different passenger classes.

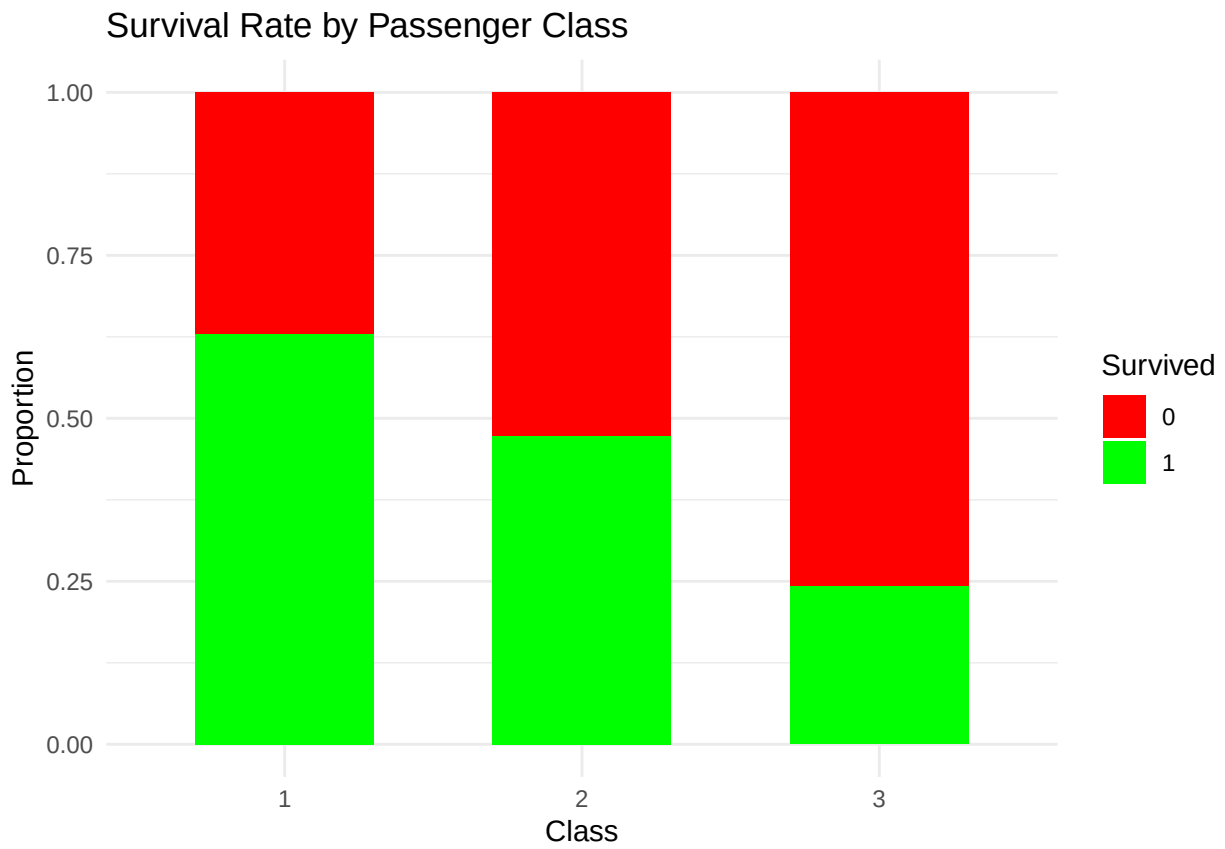


This boxplot visualizes the distribution of fares across the three passenger classes. First-class passengers paid significantly higher fares, as indicated by their wider interquartile range (IQR) and higher median fare. While most first-class fares fall below 150, several outliers exceed 200, indicating some passengers paid exceptionally high fares. Second-class fares are more modest, with a much smaller IQR and a median fare considerably lower than that of the first class. Although there are a few outliers for second-class fares, they do not exceed the 100 range. Third-class passengers paid the lowest fares, with a narrow IQR close to zero. The majority of third-class fares are tightly clustered near the lower range, although a few outliers slightly exceed 50. Overall, the chart emphasizes the socioeconomic divide among the three classes, with first-class passengers enjoying the most expensive and diverse fare options, while third-class passengers paid the least with minimal variation. This distribution likely reflects the level of amenities and accommodations provided to each class.

2.3. Survival Analysis

2.3.1. Survival Rate by Passenger Class

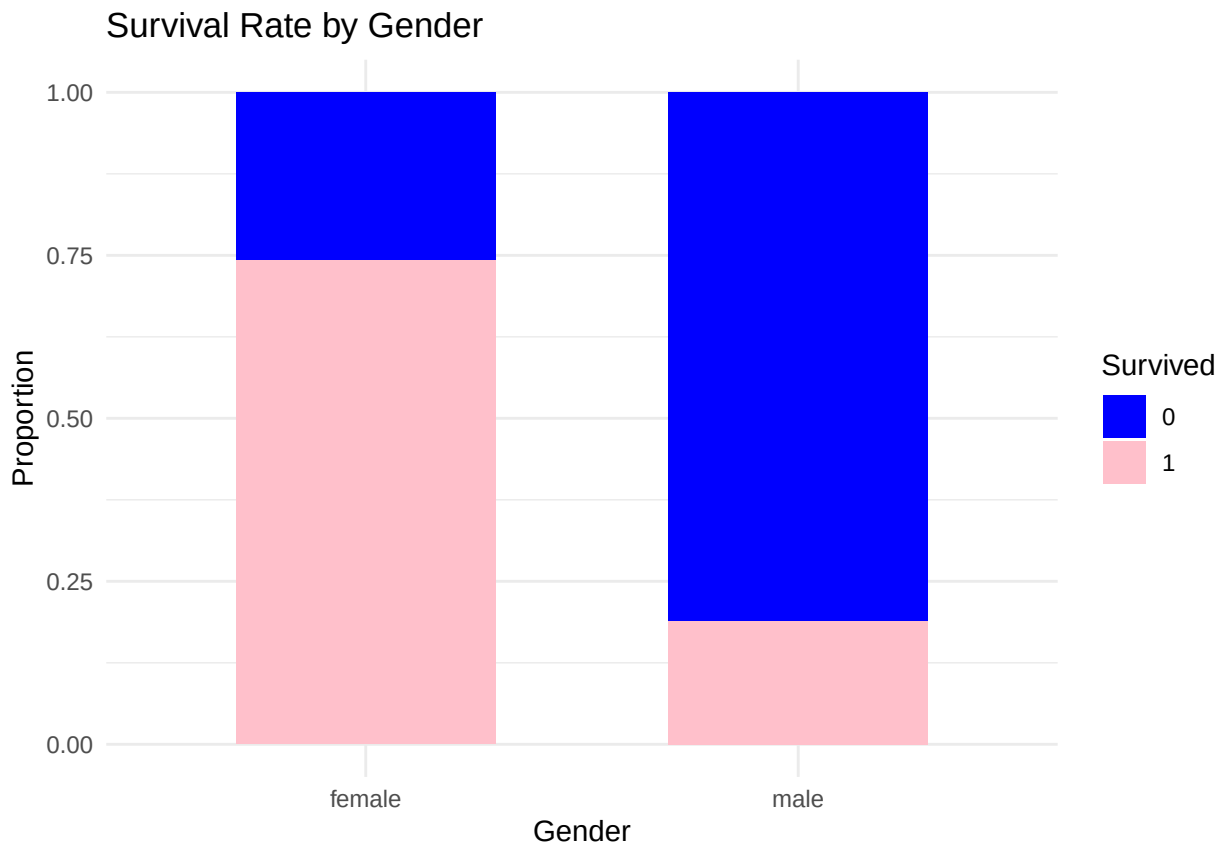
A stacked bar chart shows survival rates across passenger classes.



This stacked bar chart illustrates the survival rates across the three passenger classes, expressed as proportions. The green segment represents passengers who survived, while the red segment represents those who did not survive. The chart clearly shows a strong relationship between survival rate and passenger class. First-class passengers had the highest survival rate, with more than half surviving. Second-class passengers also show a relatively balanced survival rate, though slightly fewer survived compared to the first class. In stark contrast, third-class passengers experienced the lowest survival rate, with a majority not surviving.

2.3.2. Survival Rate by Gender

A bar chart compares survival rates for males and females.

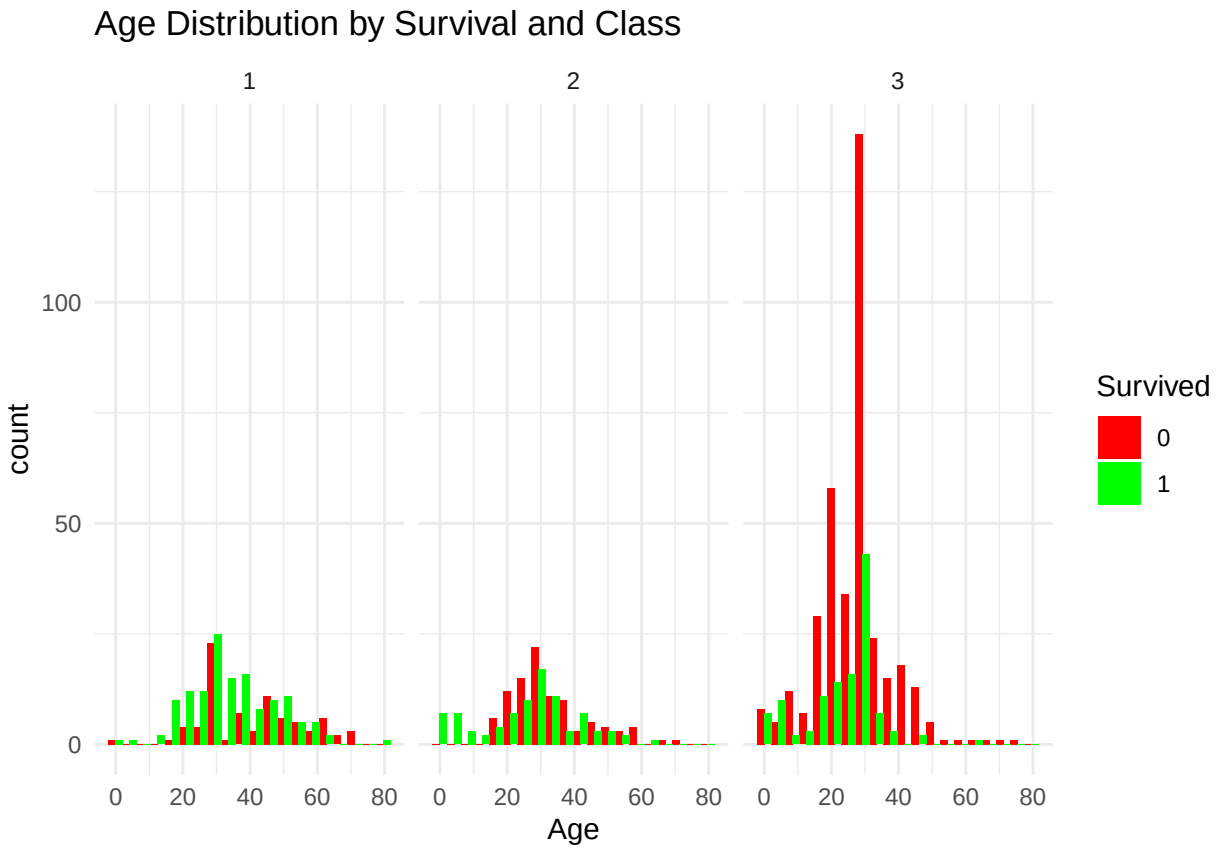


This stacked bar chart represents the survival rates by gender, expressed as proportions. The pink segment corresponds to passengers who survived, while the blue segment represents those who did not survive. The chart clearly indicates that survival rates were significantly influenced by gender. Among females, a much larger proportion survived. In contrast, the blue segment is relatively smaller, showing fewer fatalities among women. This reflects a survival rate that exceeds 75% for females. For males, the trend is reversed. The blue segment dominates, illustrating that the majority of male passengers did not survive. The pink segment is much smaller, indicating a survival rate below 25% for males. This visualization underscores the prioritization of women during rescue efforts, reflecting societal norms or evacuation protocols where women were given priority access to lifeboats or other safety measures.

2.4. Multivariate Analysis

2.4.1. Age Distribution by Survival and Passenger Class

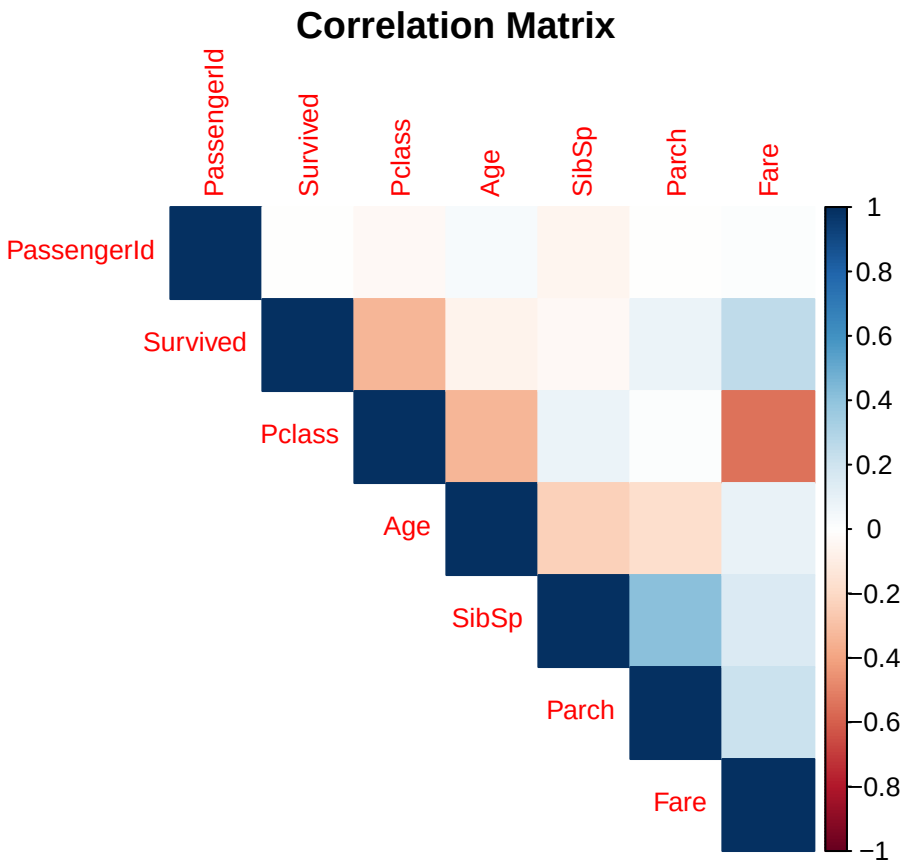
A faceted histogram compares the age distribution for survivors and non-survivors across passenger classes.



This faceted bar chart illustrates the age distribution of passengers by survival status across the three passenger classes. Each facet corresponds to a specific passenger class, with red bars representing passengers who did not survive and green bars representing those who survived. In the first class, survival rates appear relatively high across all age groups, with green bars frequently outnumbering the red bars. This pattern suggests that first-class passengers had better chances of survival, regardless of age. In the second class, the survival rates are more balanced, but survival is still evident across a range of age groups. However, younger individuals and those in middle age seem to have had slightly better survival rates compared to the elderly. The third class presents a stark contrast. The red bars dominate most age groups, particularly among younger adults and children, indicating a much lower survival rate for third-class passengers. However, a small number of survivors (green bars) can still be seen, primarily among the youngest passengers.

2.4.2. Correlation Heatmap

A heatmap shows the correlations among numeric variables.



This correlation matrix visualizes the relationships between the numeric variables in the dataset. The color gradient, ranging from deep blue to red, represents the strength and direction of the correlation. Positive correlations are shown in shades of red, while negative correlations are in shades of blue, with the intensity of the color indicating the strength of the relationship. The most notable observations include the following: - Pclass and Fare: A strong negative correlation exists between Pclass and Fare, indicating that passengers in higher classes (lower numerical value for Pclass) tended to pay higher fares. - FamilySize and SibSp/Parch: There is a strong positive correlation between FamilySize, SibSp (siblings/spouses), and Parch (parents/children). This relationship is expected since FamilySize is derived from these variables. - Survived and Pclass: There is a negative correlation between Survived and Pclass, suggesting that passengers in higher classes (lower numerical value for Pclass) had a better chance of survival. - Fare and Survived: A positive correlation is observed between Fare and Survived, indicating that passengers who paid higher fares were more likely to survive. - Other variables: Variables like Age show weaker correlations with other features, suggesting limited direct influence on the relationships shown here.

3. Machine Learning Modeling

The goal is to model the "Survived" variable using multiple machine learning approaches.

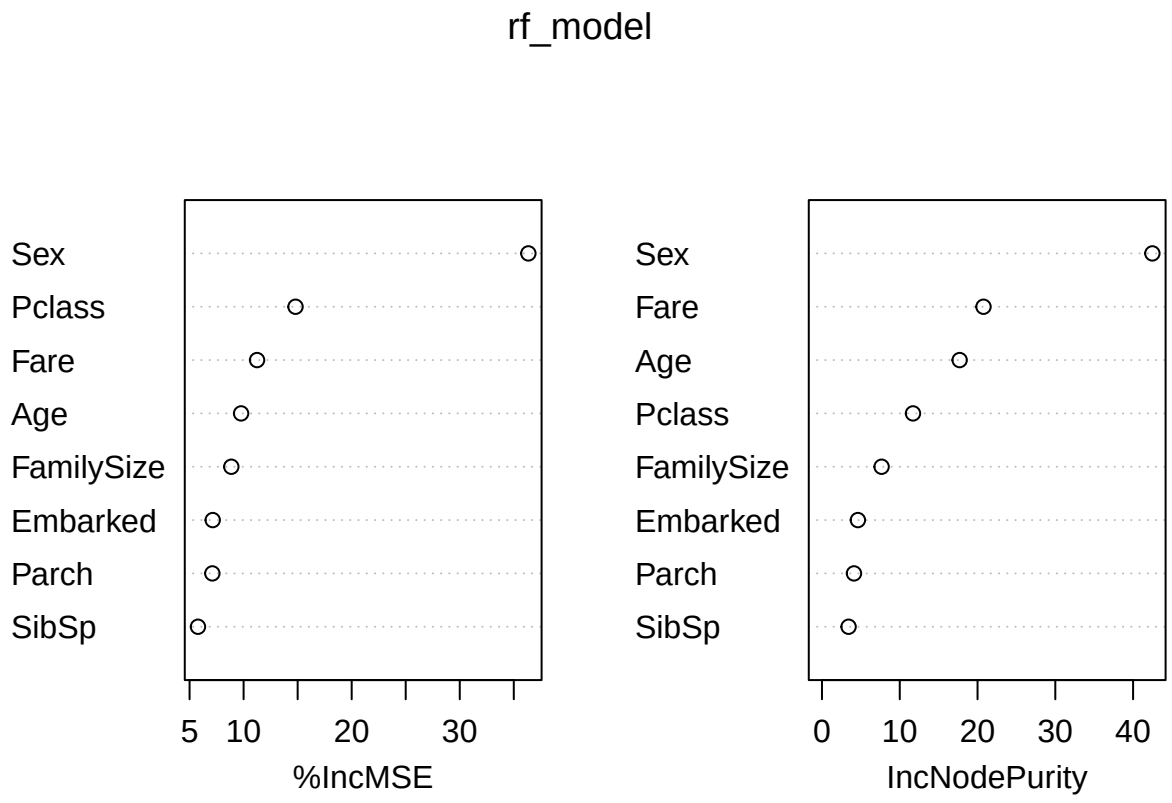
3.1. Feature selection

The goal is to perform feature selection and identify the most impactful variables using several methods, including correlation analysis and feature importance from machine learning models.

3.1.1. Feature Importance Using Random Forest

Random Forest can calculate feature importance based on how much each variable improves the model's performance.

```
##              %IncMSE IncNodePurity
## Pclass     14.803971     11.701306
## Sex        36.350335     42.479908
## Age         9.767267     17.699838
## SibSp       5.778081      3.419830
## Parch       7.107529      4.113789
## Fare       11.244549     20.763804
## Embarked    7.140601      4.615281
## FamilySize  8.861615      7.663652
```



This chart represents the feature importance from a random forest model (rf_model), as measured by the IncNodePurity metric. The IncNodePurity (increase in node purity) quantifies how much each variable contributes to reducing the impurity in the model's decision trees. The variable Sex has the highest IncNodePurity, indicating that it is the most important predictor in the random forest model. This means that knowing the value of Sex provides the most significant improvement in the model's ability to predict the outcome. Fare is the second most important variable, followed by Age. These features also contribute significantly to reducing impurity, though less than Sex. Pclass and FamilySize have lower IncNodePurity values, indicating that these variables are less influential in the model's predictions compared to the top features. This importance ranking suggests that demographic features (Sex and Age) and socioeconomic factors (Fare) are key drivers in the prediction task. Variables like Pclass and FamilySize, while still relevant, have a comparatively smaller impact.

3.1.2. Recursive Feature Elimination (RFE)

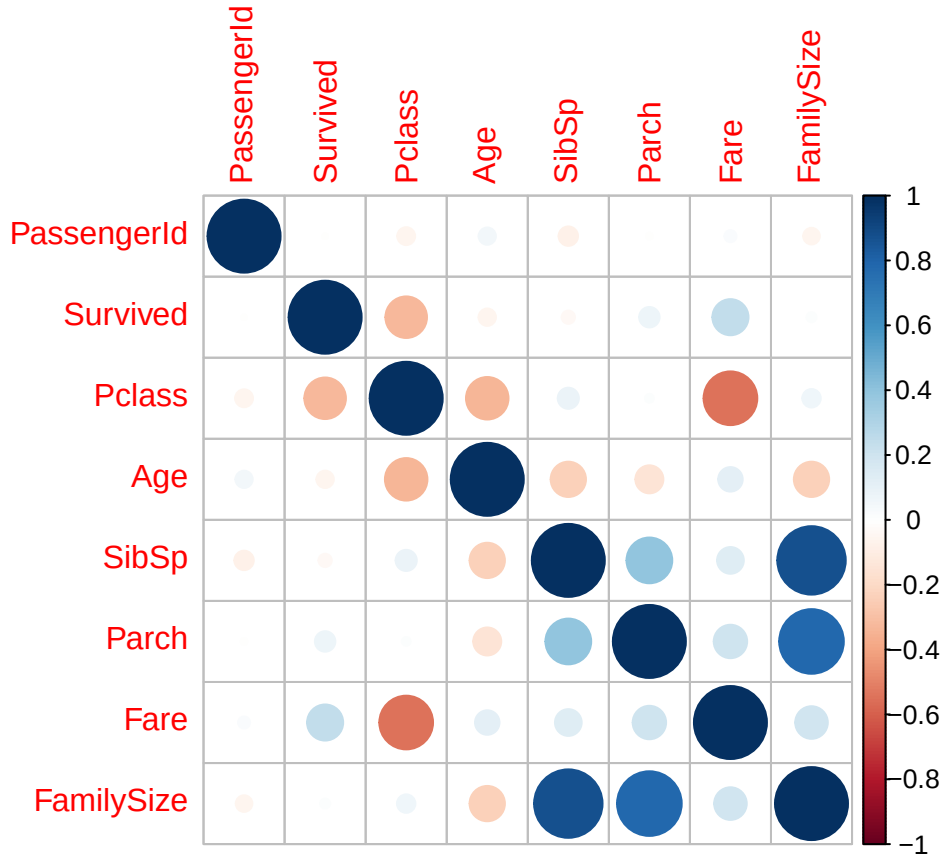
RFE systematically selects features by recursively removing the least important ones and assessing model performance.

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold)
##
## Resampling performance over subset size:
##
##  Variables   RMSE Rsquared    MAE  RMSESD RsquaredSD   MAESD Selected
##          1 0.4060   0.3104 0.3309 0.03562     0.1108 0.02760
##          2 0.3867   0.3779 0.3230 0.03640     0.1204 0.03472
##          3 0.3800   0.4026 0.3172 0.03930     0.1290 0.03503
##          4 0.3712   0.4354 0.3097 0.03904     0.1296 0.03423
##          5 0.3682   0.4543 0.3124 0.04228     0.1459 0.03603
##          6 0.3581   0.4561 0.2647 0.04975     0.1452 0.04468
##          7 0.3598   0.4541 0.2729 0.04707     0.1406 0.04100
##          8 0.3580   0.4618 0.2756 0.04727     0.1425 0.03898        *
##
## The top 5 variables (out of 8):
##    Sex, Pclass, Fare, Age, FamilySize
```

Based on the analysis the top 5 variables are: Sex, Pclass, Fare, Age and FamilySize.

3.1.3. Correlation Analysis for Numerical Variables

Identify highly correlated features to remove redundancy.



This plot reveals how variables are interrelated, providing insights into potential predictors of survival. For

instance, the negative correlation between Pclass and Survived suggests that socioeconomic status (reflected in class) was a significant factor in survival. Similarly, the strong link between FamilySize, SibSp, and Parch highlights how these variables are structurally related (FamilySize was developed by summing SibSp and Parch variables). This correlation plot helps identify key features that may be important for modeling and highlights variables with minimal influence, which might not contribute significantly to predictive models. Let me know if you'd like further analysis or help with visualizations!

3.1.4. Stepwise Regression

Stepwise regression is another method to select predictors based on statistical significance.

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Embarked,
##     family = binomial, data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5592  -0.5950  -0.4139   0.6299   2.4620
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.382048   0.555340   9.691  < 2e-16 ***
## Pclass      -1.099542   0.139099  -7.905 2.68e-15 ***
## Sexmale     -2.758035   0.219061 -12.590  < 2e-16 ***
## Age         -0.037787   0.008847  -4.271 1.94e-05 ***
## SibSp       -0.364705   0.118964  -3.066  0.00217 **
## EmbarkedQ    0.070335   0.416233   0.169  0.86581
## EmbarkedS   -0.606082   0.254981  -2.377  0.01746 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 948.95  on 712  degrees of freedom
## Residual deviance: 626.38  on 706  degrees of freedom
## AIC: 640.38
##
## Number of Fisher Scoring iterations: 5
```

All these methods suggested to reduce number of predictors to 5 which are: Sex, Pclass, Fare, Age and FamilySize.

4. Models

4.1. Logistic Regression

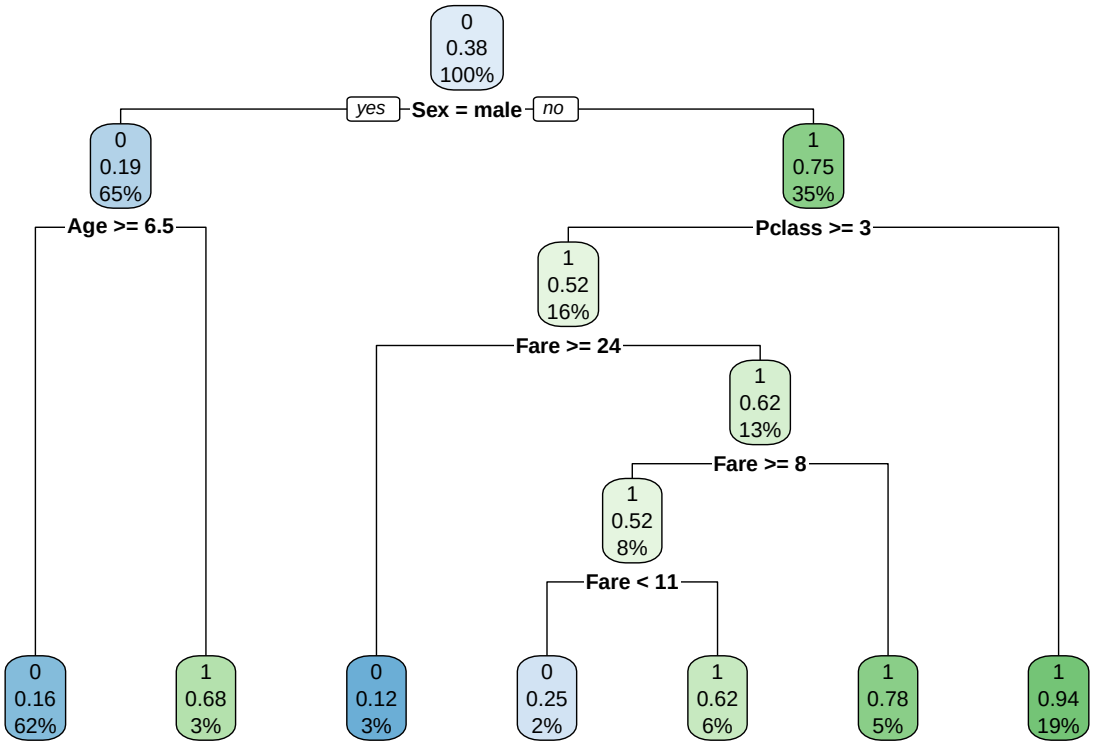
A statistical model for binary classification, estimating the probability of survival based on predictors.

```
##           Reference
## Prediction  0  1
##          0 96 24
##          1 13 45
```

4.2.Decision Tree Classifier

A tree-based model that recursively splits the data into subsets based on feature values.

```
##           Reference
## Prediction  0  1
##          0 94 20
##          1 15 49
```



4.3. Random Forest Classifier

An ensemble method using multiple decision trees to improve accuracy and robustness.

```
##           Reference
## Prediction  0  1
##          0 97 15
##          1 12 54
```

4.4. Support Vector Machines (SVM)

Support Vector Machines (SVM) are supervised learning algorithms used for classification and regression tasks. They work by finding a hyperplane that best separates data points into different classes while maximizing the margin between the classes. SVMs can handle linear and non-linear relationships using kernel functions to project data into higher dimensions, making them powerful for complex datasets.

```
##           Reference
## Prediction  0  1
```

```
##          0 92 23
##          1 17 46
```

4.5. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, non-parametric supervised learning algorithm used for classification and regression. It works by identifying the K closest data points (neighbors) to a given input and assigning a label or value based on the majority class (for classification) or the average (for regression) of the neighbors. It relies on distance metrics like Euclidean distance to measure similarity.

```
##           Reference
## Prediction  0  1
##          0 93 24
##          1 16 45
```

5. Results

```
##                          Model Accuracy Precision Recall F1_Score ROC_AUC
## 1          Logistic Regression     0.79      0.65   0.78     0.71    0.77
## 2      Decision Tree Classifier     0.80      0.71   0.77     0.71    0.79
## 3      Random Forest Classifier     0.85      0.78   0.82     0.71    0.84
## 4 Support Vector Machines (SVM)     0.78      0.67   0.73     0.71    0.76
## 5     K-Nearest Neighbors (KNN)     0.78      0.65   0.74     0.71    0.75
```

Best Model: The Random Forest Classifier demonstrates the strongest performance across all metrics, making it the most reliable choice for this dataset.
Good Alternative: The Decision Tree Classifier also performs well, with balanced precision and recall, though slightly less effective than Random Forest.
Middle Performers: Logistic Regression and SVM perform similarly, with moderate metrics across the board.
Weakest Model: While KNN is not the weakest in accuracy, its slightly lower ROC-AUC and dependence on feature scaling make it less competitive compared to Random Forest or Decision Tree.