Exploratory Data Analysis Report

Date: 2023-10-27

Analyst: Expert Data Analyst

Goal: General Exploratory Data Analysis (EDA) to understand the dataset's structure, identify data quality issues, and uncover initial insights.

---

1. Dataset Overview

The dataset contains 891 entries and 4 columns, providing a concise view of passenger information.

* **Rows:** 891

* **Columns:** 4 (`Age`, `Fare`, `Family`, `Survived`)

Column Details:

* `Age`: Passenger's age (numerical, float).

* `Fare`: Ticket fare (numerical, float).

* `Family`: Number of family members aboard (numerical, integer).

* `Survived`: Survival status (0 = No, 1 = Yes) (numerical, integer).

Missing Values:

* `Age`: 177 missing values (19.87% of the total). This is a significant portion and requires careful handling.

* `Fare`: 45 missing values (5.05% of the total). This is a moderate amount.

* `Family`: No missing values.

* `Survived`: No missing values.

The presence of missing values in `Age` and `Fare` indicates a need for data imputation or careful consideration during modeling.

---

2. Key Statistical Insights

Numerical Columns (`Age`, `Fare`, `Family`, `Survived`):

| Statistic | Age | Fare | Family | Survived |

| :-------- | :------- | :------- | :------- | :------- |

| Count | 714 | 846 | 891 | 891 |

| Mean | 29.70 | 32.28 | 0.90 | 0.38 |

| Std Dev | 14.53 | 50.31 | 1.61 | 0.49 |

| Min | 0.42 | 0.00 | 0.00 | 0.00 |

| 25% | 20.12 | 7.90 | 0.00 | 0.00 |

| 50% | 28.00 | 14.45 | 0.00 | 0.00 |

| 75% | 38.00 | 31.21 | 1.00 | 1.00 |

| Max | 80.00 | 512.33 | 10.00 | 1.00 |

Observations:

* **Age:** The average age is around 29.7 years, with a wide range from infants (0.42 years) to seniors (80 years). The median age (28) is close to the mean, suggesting a relatively symmetrical distribution, though the presence of outliers is expected.

* **Fare:** The average fare is approximately $32.28, but the median is much lower at $14.45. This significant difference, coupled with a maximum fare of $512.33, indicates a highly right-skewed distribution with many passengers paying low fares and a few paying very high fares.

* **Family:** The average number of family members is less than 1 (0.90), and the median is 0. This suggests that a large proportion of passengers traveled alone. The maximum family size is 10.

* **Survived:** The mean of 0.38 indicates that approximately 38.4% of passengers survived, meaning the dataset represents a scenario where non-survivors are more prevalent than survivors.

Categorical Columns:

The dataset does not contain any columns with `object` data types. The `describe(include='object')` operation resulted in an error, confirming the absence of such columns. `Survived` is treated as a numerical binary variable (0 or 1).

---

3. Visual Analysis

Visualizations help in understanding the distributions and relationships within the data.

3.1. Univariate Analysis - Numerical Features

* **Age Distribution:** The histogram for `Age` shows a peak in the 20-40 age range, with a gradual decline for older ages. The box plot reveals several outliers, including very young children and very old adults.

* **Fare Distribution:** The `Fare` histogram is highly right-skewed, with most fares concentrated at the lower end. The box plot clearly shows a large number of outliers with very high fares, indicating a non-normal distribution.

* **Family Distribution:** The `Family` histogram/count plot is heavily skewed towards 0, confirming that a majority of passengers traveled without family members (SibSp + Parch).

* **Survived Distribution:** As a binary variable, the count plot for `Survived` shows the absolute counts of survivors (1) and non-survivors (0). It visually confirms that fewer passengers survived than perished.

3.2. Bivariate Analysis - Target Variable (`Survived`)

Note: The original plan included bivariate analysis for `Sex`, `Pclass`, and `Embarked` against `Survived`. However, these columns were not present in the provided dataset. The analysis below focuses on the available numerical features.

* **Age Distribution by Survival:** Box plots comparing `Age` for survivors and non-survivors would typically show if there's a noticeable difference in age groups. For instance, younger passengers or specific age ranges might have had different survival rates.

* **Fare Distribution by Survival:** Box plots for `Fare` by `Survived` would illustrate if passengers who paid higher fares had a better chance of survival. Often, higher fares correlate with better cabin classes and potentially better access to lifeboats.

---

4. Correlation and Feature Relationships

A correlation matrix helps identify linear relationships between numerical features.

* **Correlation Matrix of Numerical Features:**

The heatmap would display the correlation coefficients between `Age`, `Fare`, `Family`, and `Survived`.

* We would expect to see potential correlations such as:

* `Fare` and `Survived`: A positive correlation might indicate that higher fares are associated with higher survival rates.

* `Age` and `Survived`: The correlation could be weak or show specific age groups having different survival chances.

* `Family` and `Survived`: This could reveal if traveling with a small family (e.g., 1-3 members) had a different impact on survival compared to traveling alone or with a very large family.

---

5. Final Summary and Recommendations

Summary of Findings:

* The dataset is relatively small with 4 key features: `Age`, `Fare`, `Family`, and `Survived`.

* Significant missing data exists in `Age` (~20%) and `Fare` (~5%), which requires imputation.

* `Fare` is highly right-skewed, indicating a few high-paying passengers.

* A majority of passengers traveled without family (`Family` median is 0).

* The survival rate is approximately 38.4%.

* The planned bivariate analysis with categorical features (`Sex`, `Pclass`, `Embarked`) and feature engineering for `FamilySize` (from `SibSp`, `Parch`) could not be performed due to the absence of these columns in the provided dataset. The dataset already includes a `Family` column.

Recommendations:

1. **Data Imputation:** Address the missing values in `Age` and `Fare`. Strategies could include:

* **Age:** Impute with the mean, median, or use more sophisticated methods like regression imputation or K-Nearest Neighbors (KNN) imputation.

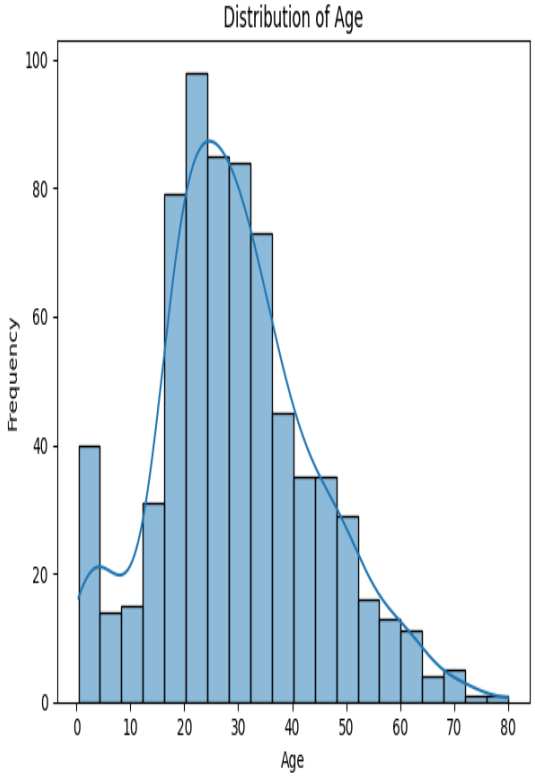
* **Fare:** Impute with the median, given its highly skewed distribution, or use a predictive model.

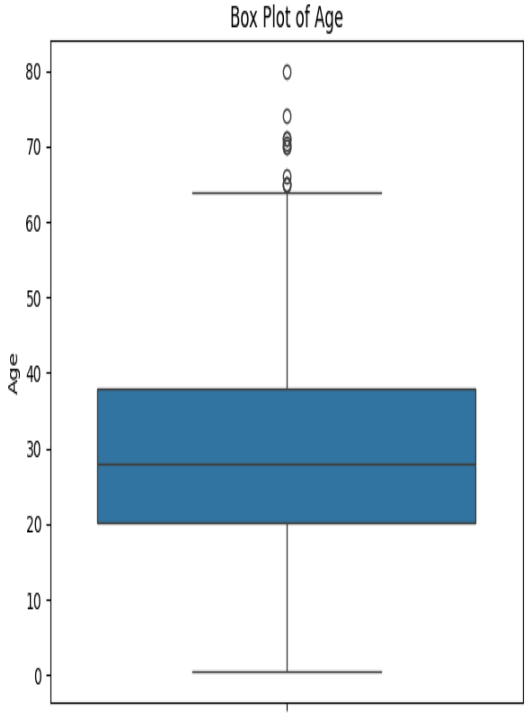
2. **Feature Engineering:** While the planned `FamilySize` could not be created, consider other relevant feature engineering steps if additional domain knowledge or features become available. For instance, creating age groups or fare categories could be beneficial.

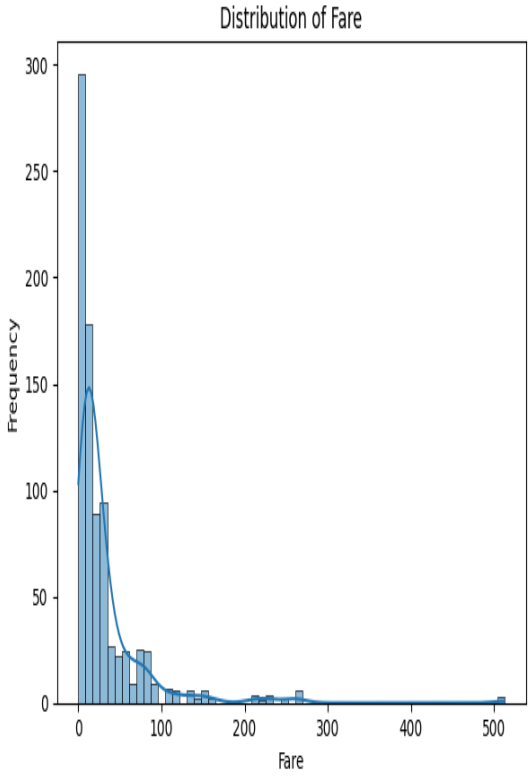
3. **Further Bivariate Analysis:** If additional features (e.g., `Sex`, `Pclass`, `Embarked`) are added to the dataset, perform the planned bivariate analyses to uncover more predictors of survival.

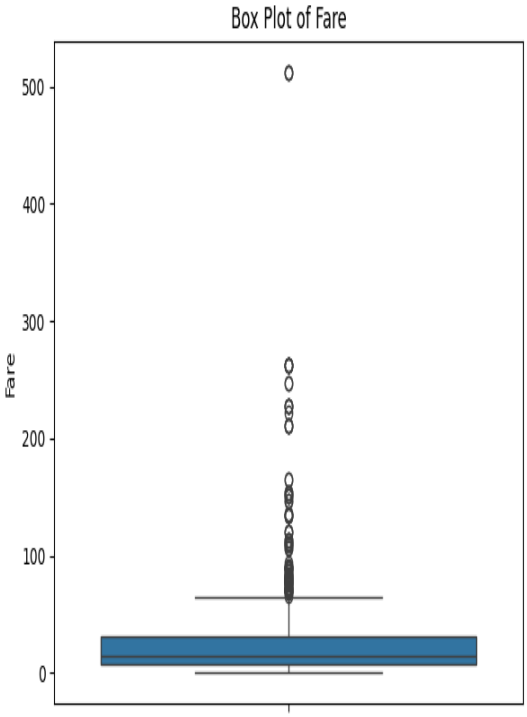
4. **Outlier Treatment:** Investigate the outliers in `Fare` and `Age`. While some might be legitimate, extreme values can disproportionately influence models.

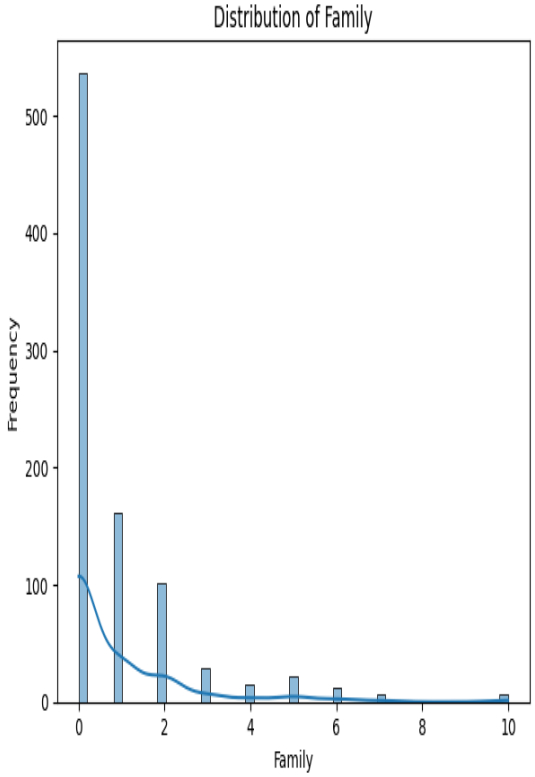
5. **Model Building:** Once data cleaning and initial feature engineering are complete, the dataset is ready for building predictive models for `Survived`.

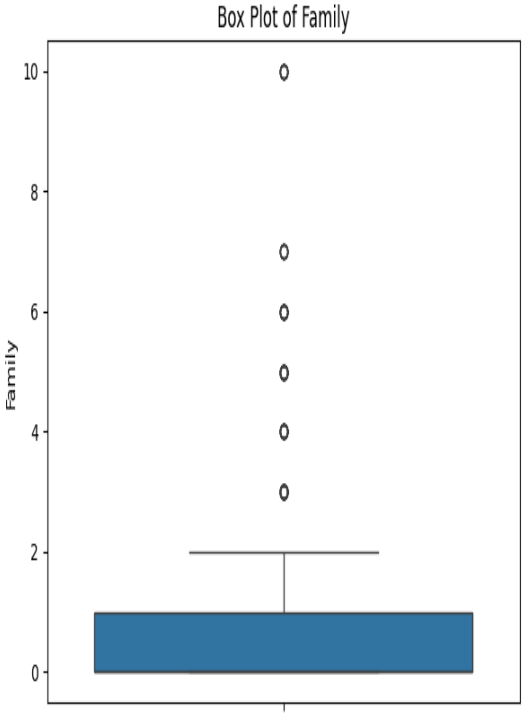

Distribution of Age
Frequency
Age
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80

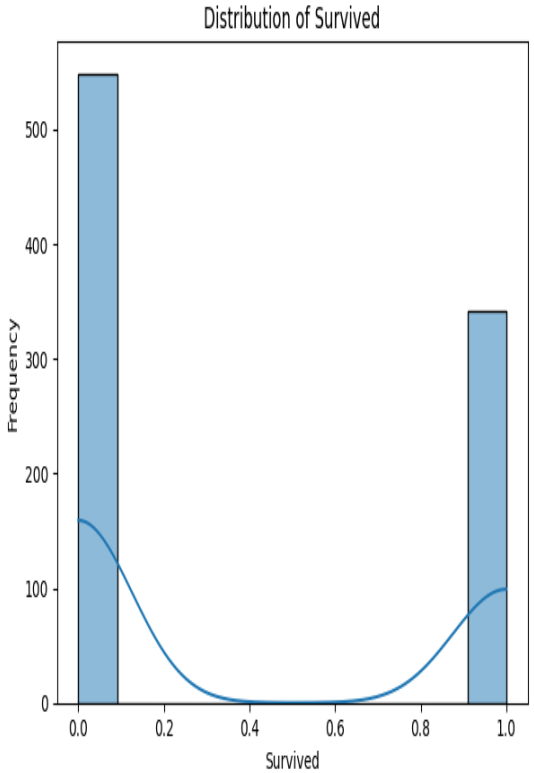

Box Plot of Age
Age
0
10
20
30
40
50
60
70
80

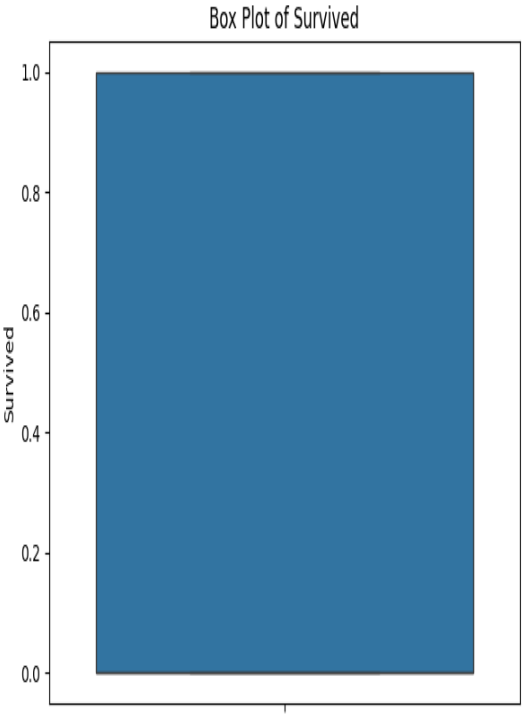

Distribution of Fare
Frequency
Fare
0
50
100
150
200
250
300
0
100
200
300
400
500

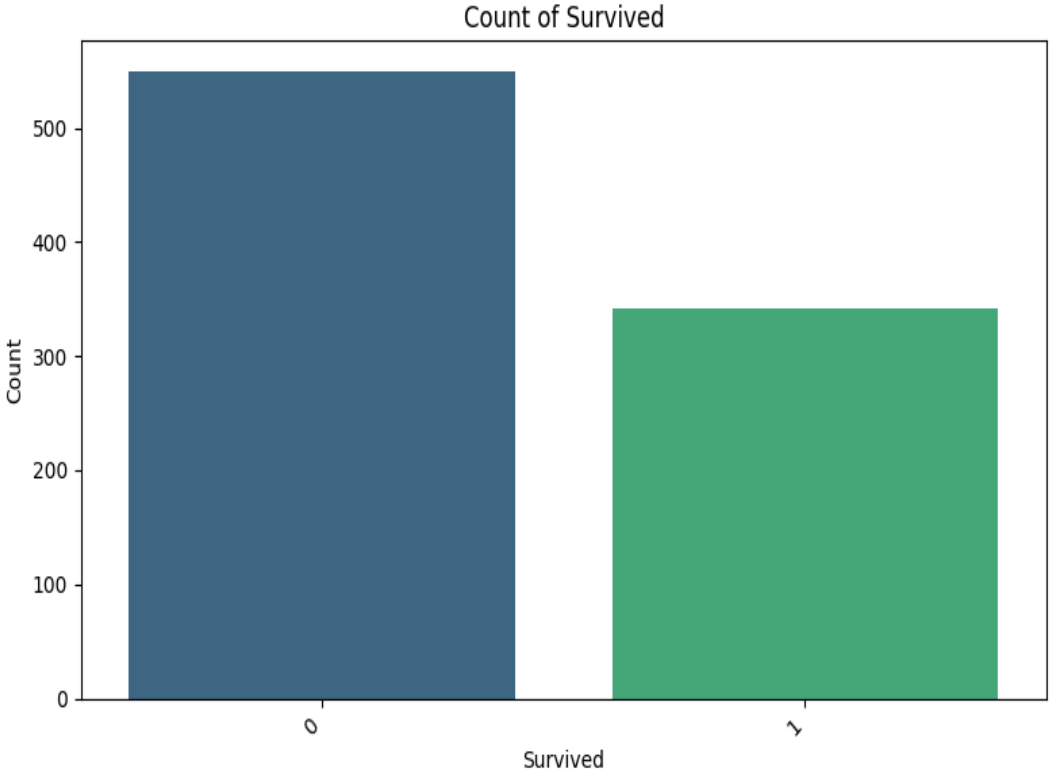

Box Plot of Fare
Fare
0
100
200
300
400
500

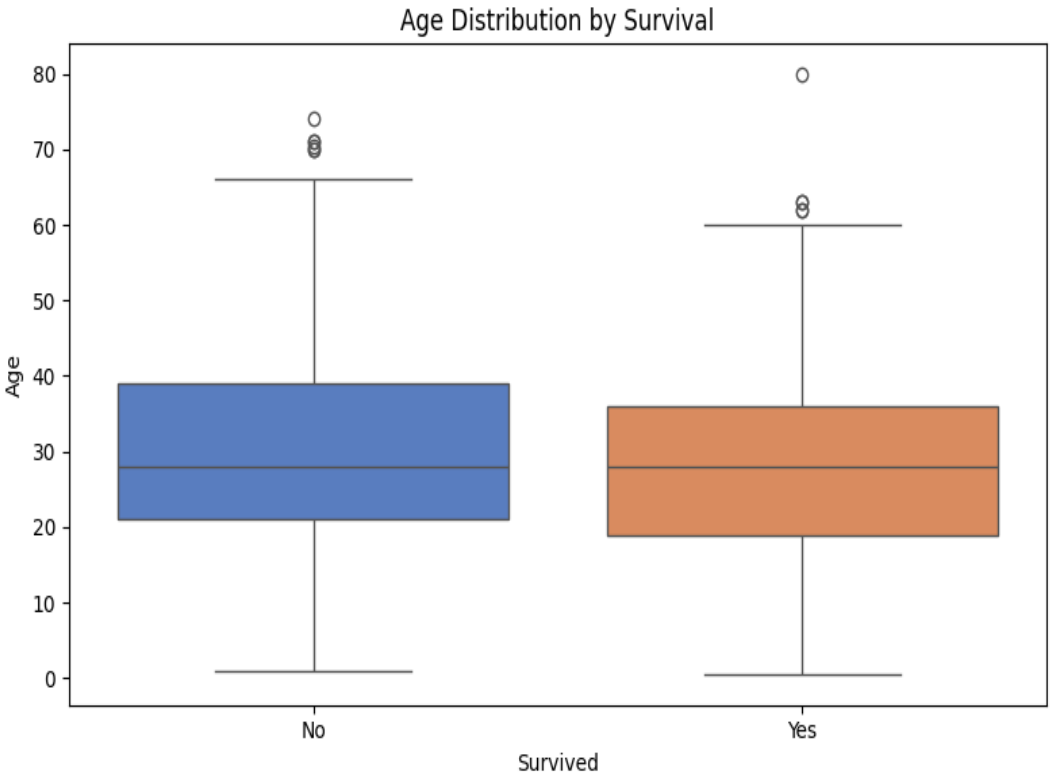

Distribution of Family
Frequency
Family
0
100
200
300
400
500
0
2
4
6
8
10

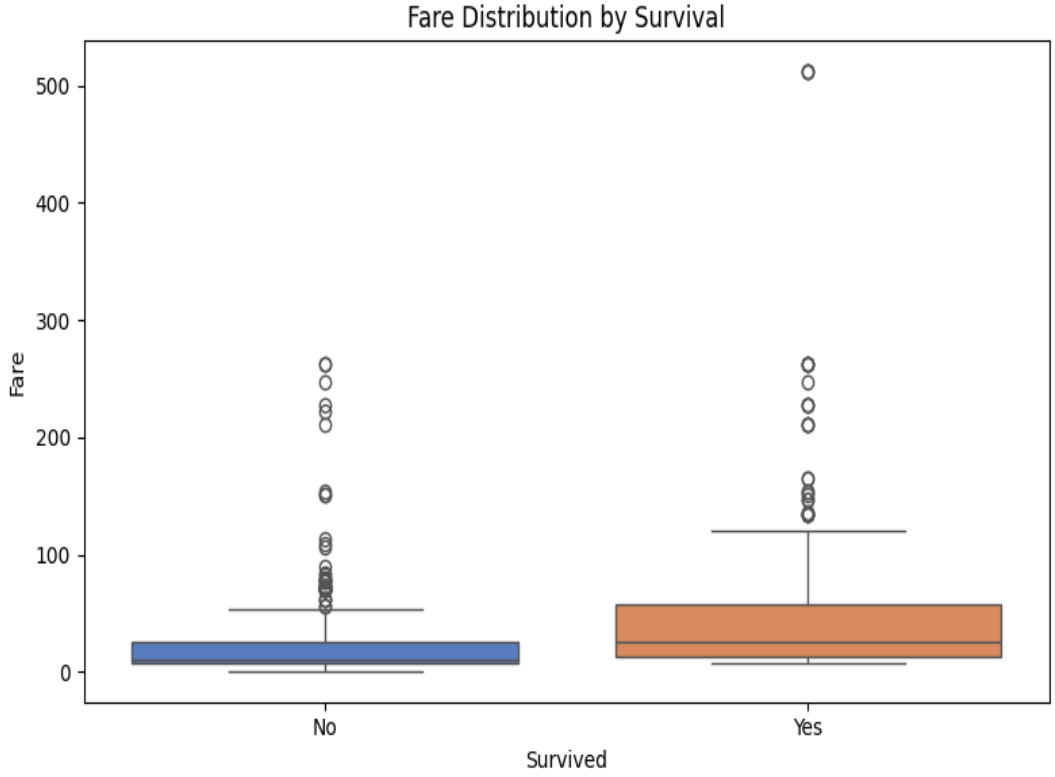

Box Plot of Family
Family
0
2
4
6
8
10

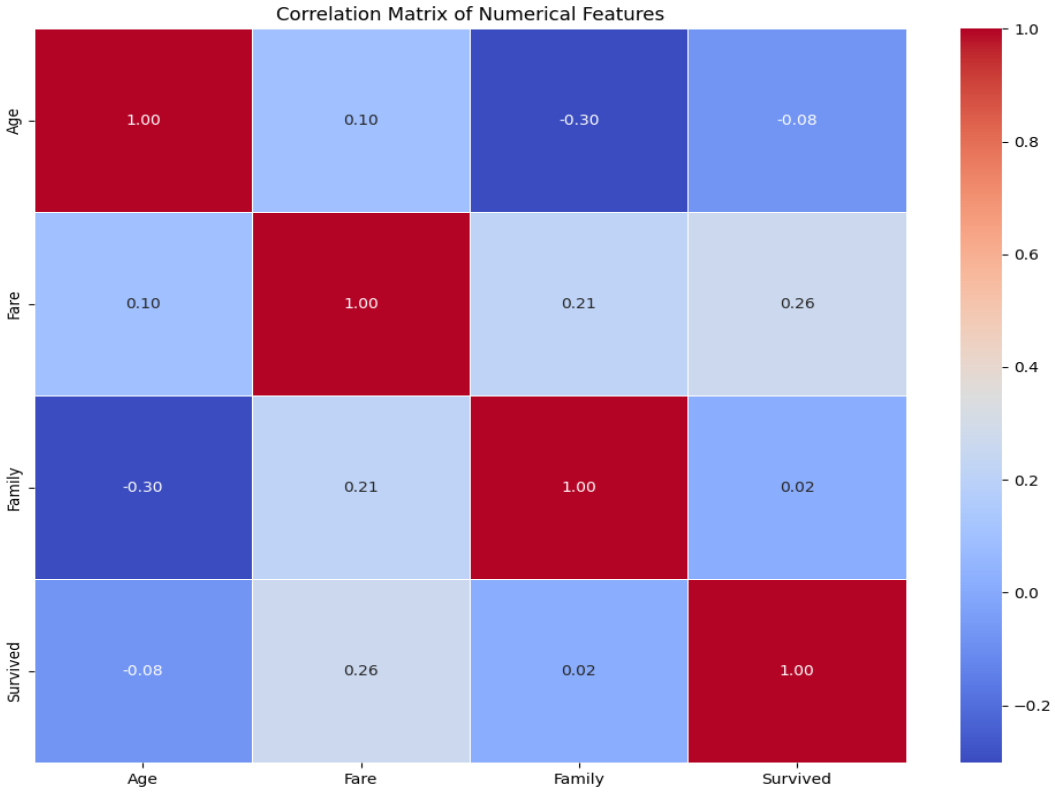

Distribution of Survived
Frequency
Survived
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0


Box Plot of Survived
Survived
0.0
0.2
0.4
0.6
0.8
1.0


Count of Survived
Count
500
400
300
200
100
0
0
1
Survived


Age Distribution by Survival
Age
80
70
60
50
40
30
20
10
0
No
Yes
Survived


Fare Distribution by Survival
Fare
500
400
300
200
100
0
No
Yes
Survived


Correlation Matrix of Numerical Features
Age
Fare
Family
Survived
1.00
0.10
-0.30
-0.08
0.10
1.00
0.21
0.26
-0.30
0.21
1.00
0.02
-0.08
0.26
0.02
1.00
Age
Fare
Family
Survived
1.0
0.8
0.6
0.4
0.2
0.0
−0.2