

Titanic Dataset Analysis Report

Introduction

The Titanic dataset provides detailed information about passengers aboard the RMS Titanic, which sank in 1912, resulting in the loss of many lives. This report analyzes the dataset to uncover survival patterns, demographic trends, and socio-economic factors influencing survival outcomes. The analysis follows a structured workflow, focusing on data preprocessing and visualization to derive actionable insights.

Step 1: Data Loading and Initial Inspection

- **PassengerId:** A unique identifier for each passenger, ranging from 1 to 891.
- **Survived:** Indicates survival status: 1 for survived, 0 for not survived.
- **Pclass:** Passenger class, reflecting socio-economic status: 1 (1st class), 2 (2nd class), 3 (3rd class).
- **Name:** Full name of the passenger, including titles (e.g., Mr., Mrs., Miss).
- **Sex:** Gender of the passenger: "male" or "female" (encoded as 0 or 1).
- **Age:** Age of the passenger in years, with some missing values.
- **SibSp:** Number of siblings or spouses aboard, ranging from 0 to 8.
- **Parch:** Number of parents or children aboard, ranging from 0 to 6.
- **Ticket:** Ticket number, a unique or semi-unique identifier.
- **Fare:** Passenger fare in pounds sterling, ranging from 0 to 512.3292.
- **Cabin:** Cabin number, with significant missing values (~77%).
- **Embarked:** Port of embarkation: "C" (Cherbourg), "Q" (Queenstown), "S" (Southampton), with few missing values.

[81]: df.head()

[81]:	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Step 2: Getting Features Type

```
df.dtypes
```

```
PassengerId    int64
Survived        int64
Pclass          int64
Name            object
Sex             object
Age            float64
SibSp           int64
Parch           int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Step 3: Getting the Count of Missing Values

```
df.isnull().sum()
```

```
PassengerId    0
Survived        0
Pclass          0
Name            0
Sex             0
Age            177
SibSp           0
Parch           0
Ticket          0
Fare            0
Cabin          687
Embarked        2
dtype: int64
```

Observation:

- Age: ~177 missing values (~19.87%).
- Cabin: ~687 missing values (~77.10%).
- Embarked: ~2 missing values (~0.22%).
- Other columns: 0 missing values.

Step 4: Deleting Columns That Have 30% Missing Values

Calculate the percentage of missing values for each column and drop those exceeding 30% missing data.

```
Missing Percentage:
PassengerId    0.000000
Survived       0.000000
Pclass         0.000000
Name           0.000000
Sex            0.000000
Age           19.865320
SibSp          0.000000
Parch          0.000000
Ticket         0.000000
Fare           0.000000
Cabin          77.104377
Embarked       0.224467
dtype: float64

Columns with >30% Missing Values Dropped: ['Cabin']
Remaining Columns: ['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked']
```

Step 5: Pre-processing on Dataset : Filling Missing Values

- Impute missing values in numerical columns with their mean and in categorical columns with their mode.
- Age is imputed with its mean (~29.7), Embarked with its mode ('S'). After imputation, no missing values remain in the dataset.

```
Missing Values After Imputation:
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age           0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

Step 6: Pre-processing on Dataset : Encoding and Feature Engineering

Encode categorical variables, and create a derived feature (FamilySize).

- **Encoding Categorical Variables:** Encode Sex and Embarked into numerical format using LabelEncoder (e.g., Sex: 1 = male, 0 = female; Embarked: 0 = C, 1 = Q, 2 = S). Encoding is necessary for visualization and modeling, as many techniques require numerical data.
- **Feature Engineering:** Create FamilySize as the sum of SibSp (siblings/spouses) and Parch (parents/children) to analyze the impact of family size on survival.

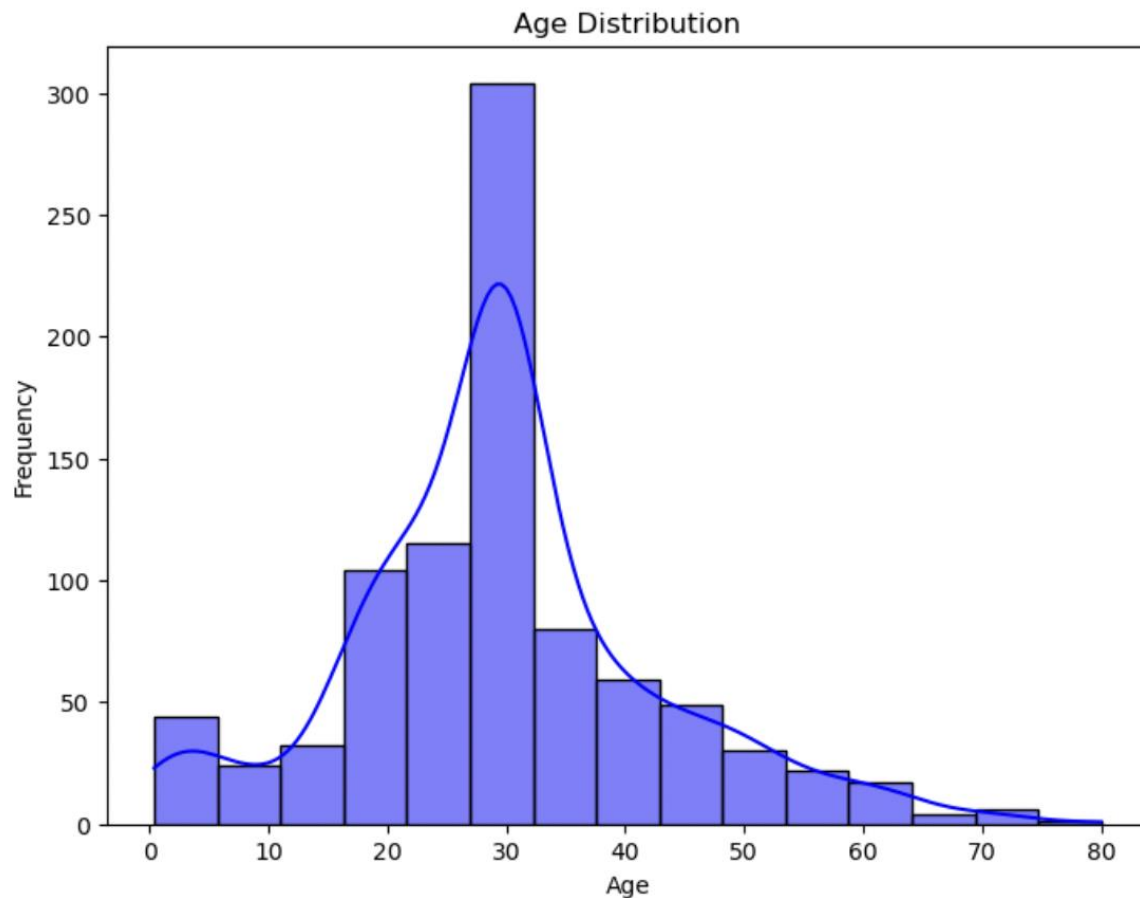
Step 7: Detect Outliers

Use the zscore method to detect outliers in Fare.

Step 9: Data Visualization

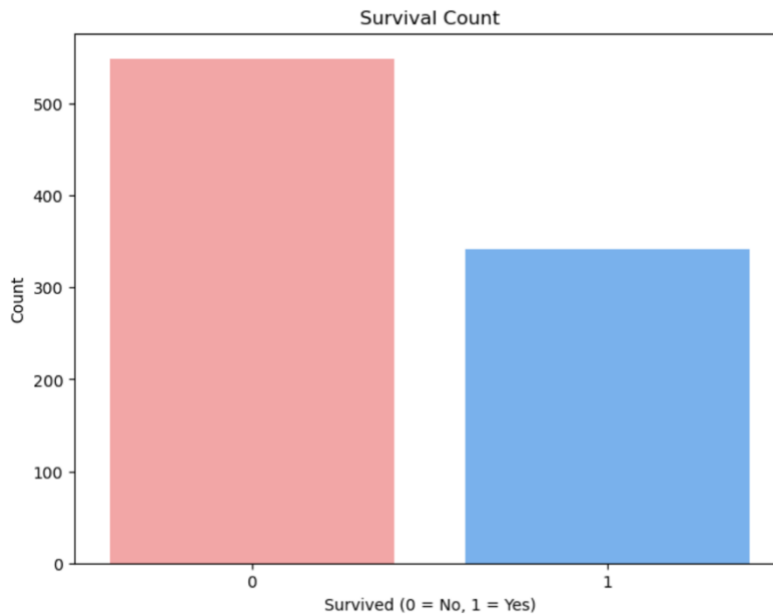
Distribution of Age

The "Age Distribution" histogram shows a peak at 25-30 years, indicating most Titanic passengers were young adults. The distribution is right-skewed, with fewer children (0-10) and older adults (40+), peaking at ~300 frequency.

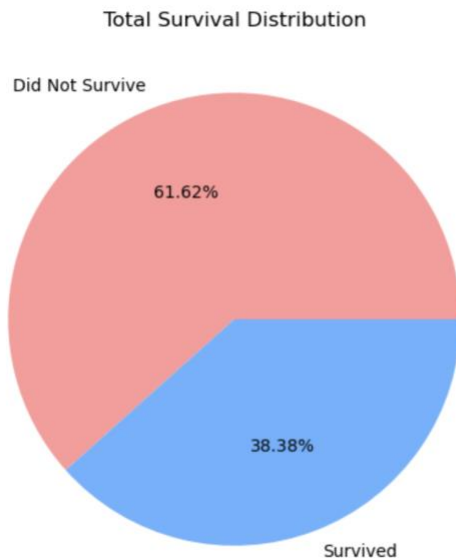


2.Survival Distribution

Indicates the plot shows the count of passengers who survived versus those who did not. The teal bar's height significantly exceeds the orange bar, highlighting the imbalance between non-survivors and survivors.

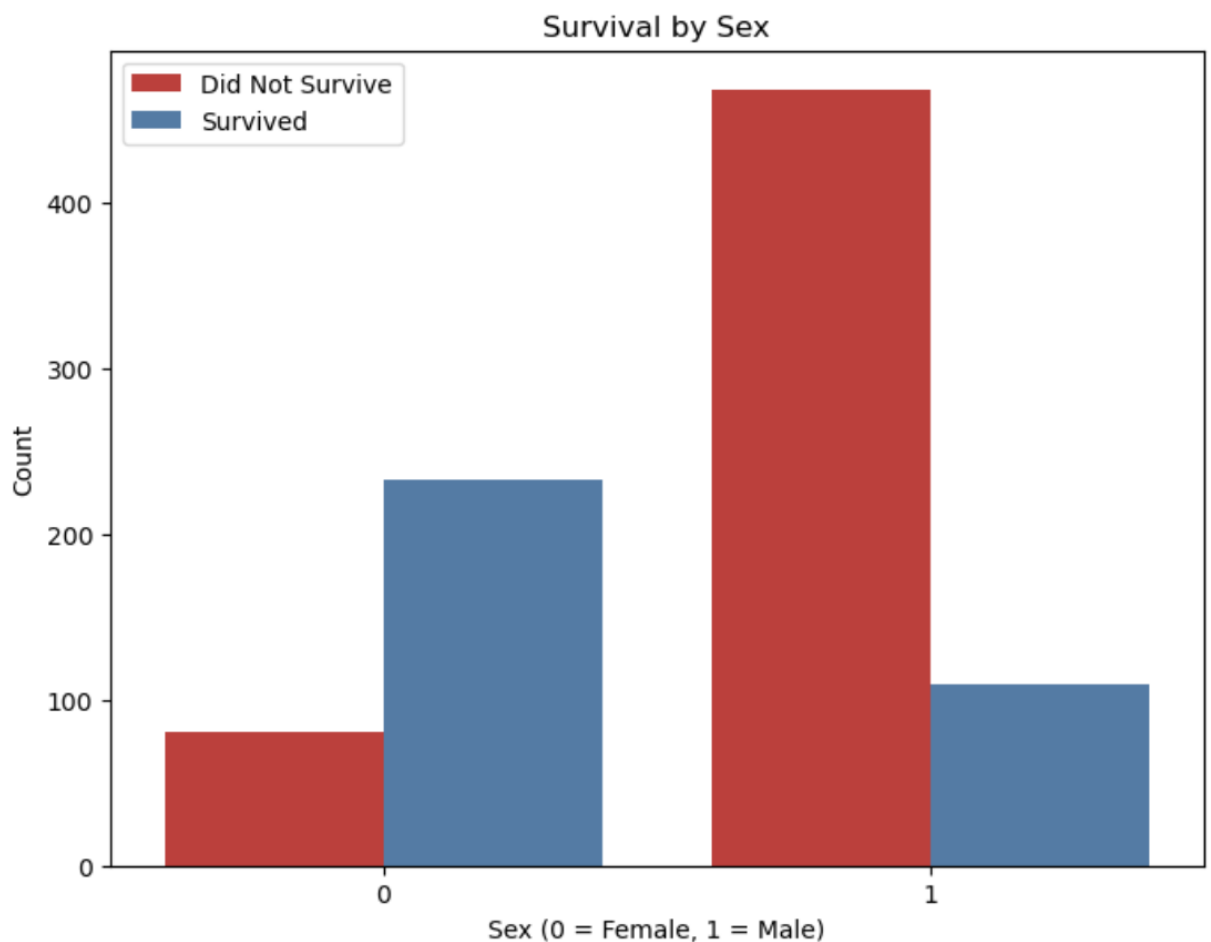


Approximately 62% of passengers did not survive (around 549 out of 891), while 38% survived (around 342 out of 891), reflecting the disaster's high mortality rate.



3.Survival by Sex Analysis

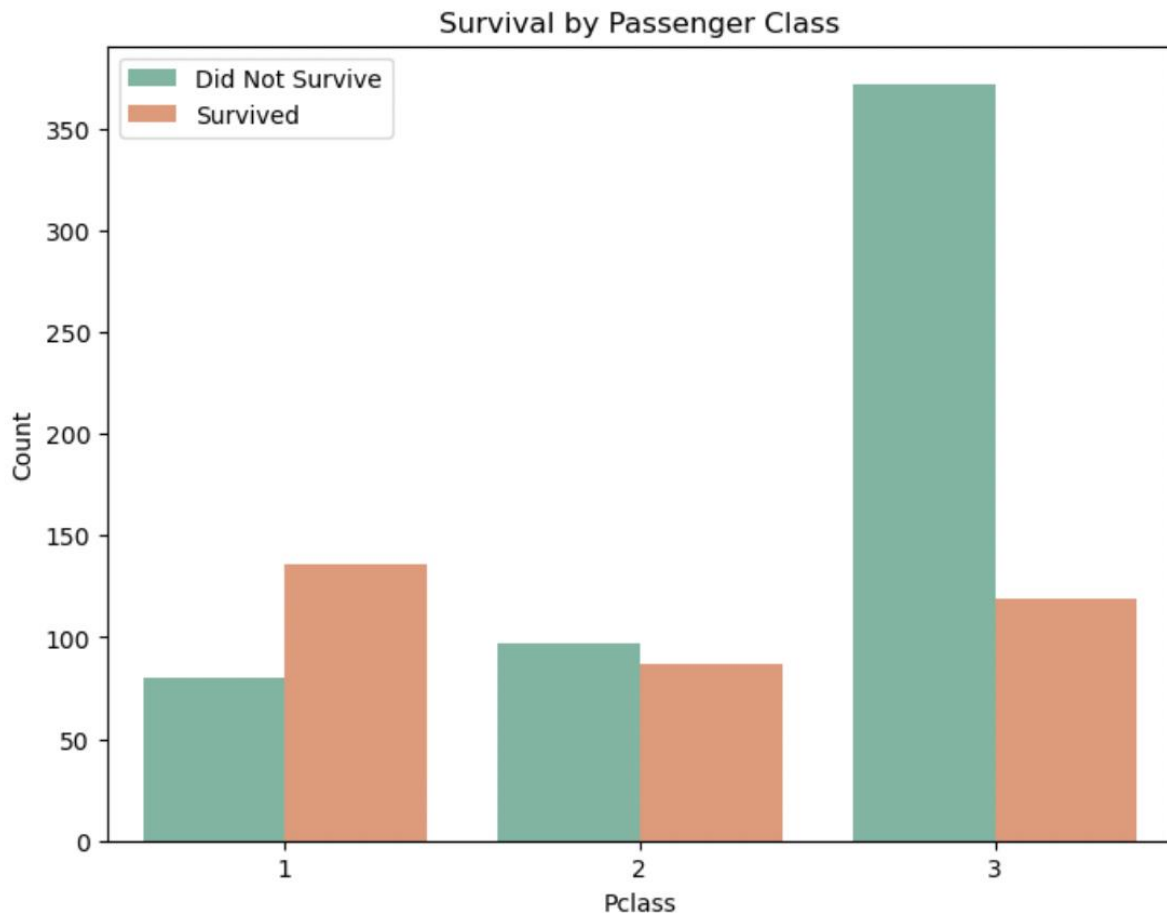
The bar chart displays survival outcomes by sex, with "Sex (0 = Female, 1 = Male)" on the x-axis and count on the y-axis, using red for "Did Not Survive" and blue for "Survived." Females (0) show 233 survivors out of 314 (74.20%), while males (1) have only 109 survivors out of 577 (18.89%). The taller blue bar for females and dominant red bar for males highlight the gender-based survival disparity. This reflects the "women and children first" policy, indicating a clear survival advantage for females. Further analysis could explore intersections with class or age.



4. Survival by Passenger Class Analysis

The bar chart titled "Survival by Passenger Class" displays survival outcomes across Pclass (1, 2, 3) on the x-axis, with count on the y-axis, using green for "Did Not Survive" and orange for "Survived." Pclass 1 shows 80 non-survivors and 136 survivors, Pclass 2 has 97 non-survivors and 87 survivors, and Pclass 3 exhibits 372 non-survivors and 119 survivors. The trend indicates a decreasing survival rate from Pclass 1 to Pclass 3, reflecting a higher mortality rate in lower

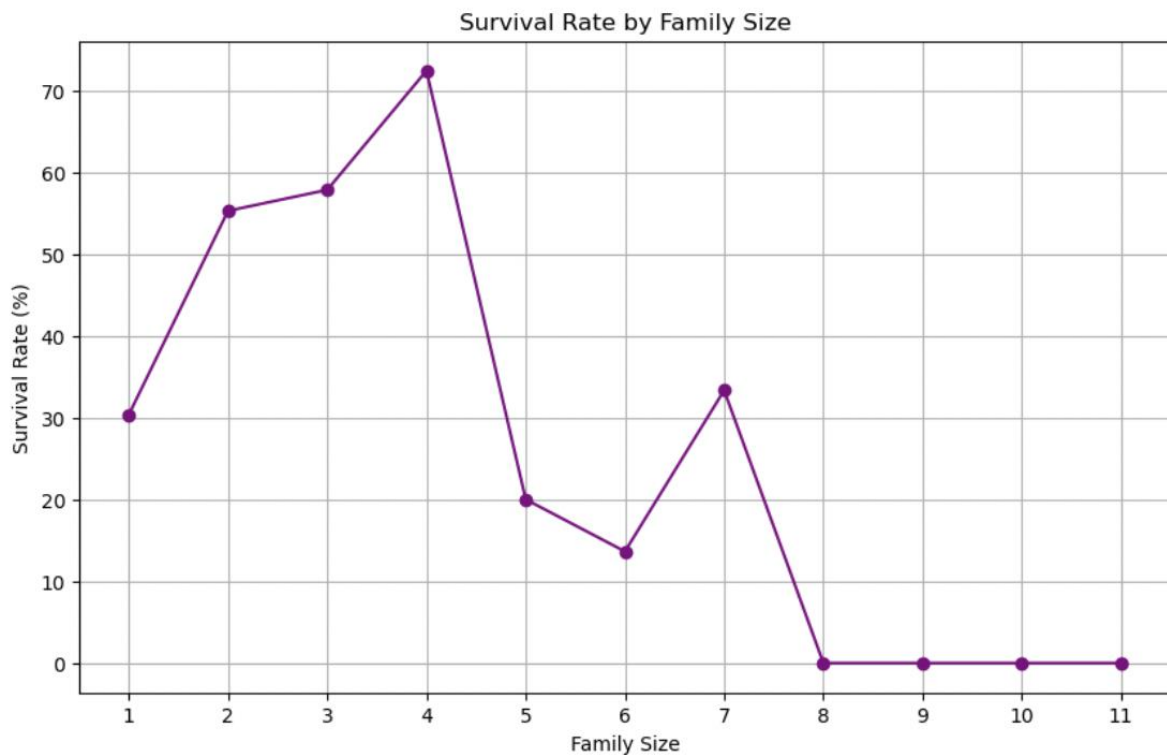
classes. This suggests limited lifeboat access for third-class passengers, aligning with historical disaster dynamics.



5.Survival Rate by Family Size Analysis

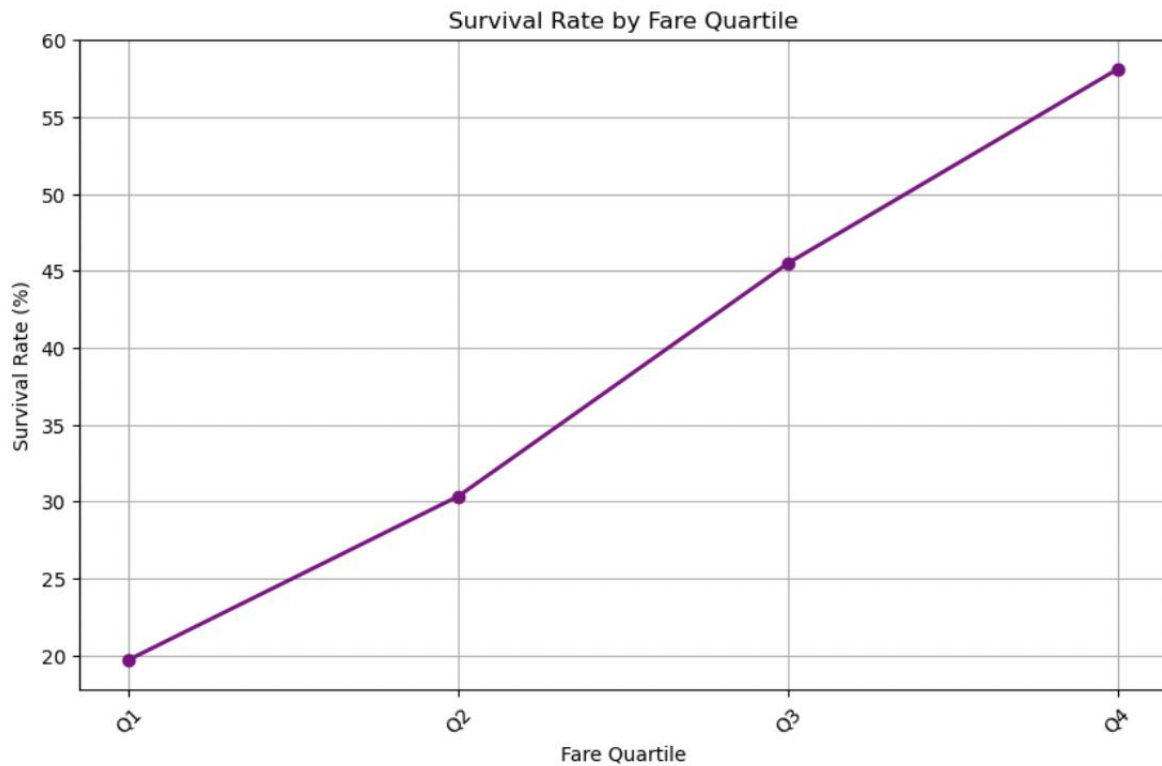
The line chart titled "Survival Rate by Family Size" plots survival rates (%) on the y-axis against family size (1 to 3) on the x-axis, using a purple line with markers. We tested the effect of family size by summing SibSp (siblings/spouses) and Parch (parents/children) for each passenger, adding 1 to include the passenger themselves, creating a range from 1 (alone) to 11. It shows a rising trend from 30.35% at size 1 (537 passengers), peaking at 72.41% at size 4 (29 passengers), followed by a drop to 20.00% at size 5 (15 passengers) and 22.73% at size 6 (22 passengers), then falling to 0% for sizes 7 to 11 (12, 6, 0, 0, 7 passengers). This suggests that moderate family

sizes (2 to 4) had the highest survival rates, possibly due to better coordination or priority in lifeboat allocation, while larger families faced significantly lower survival odds.



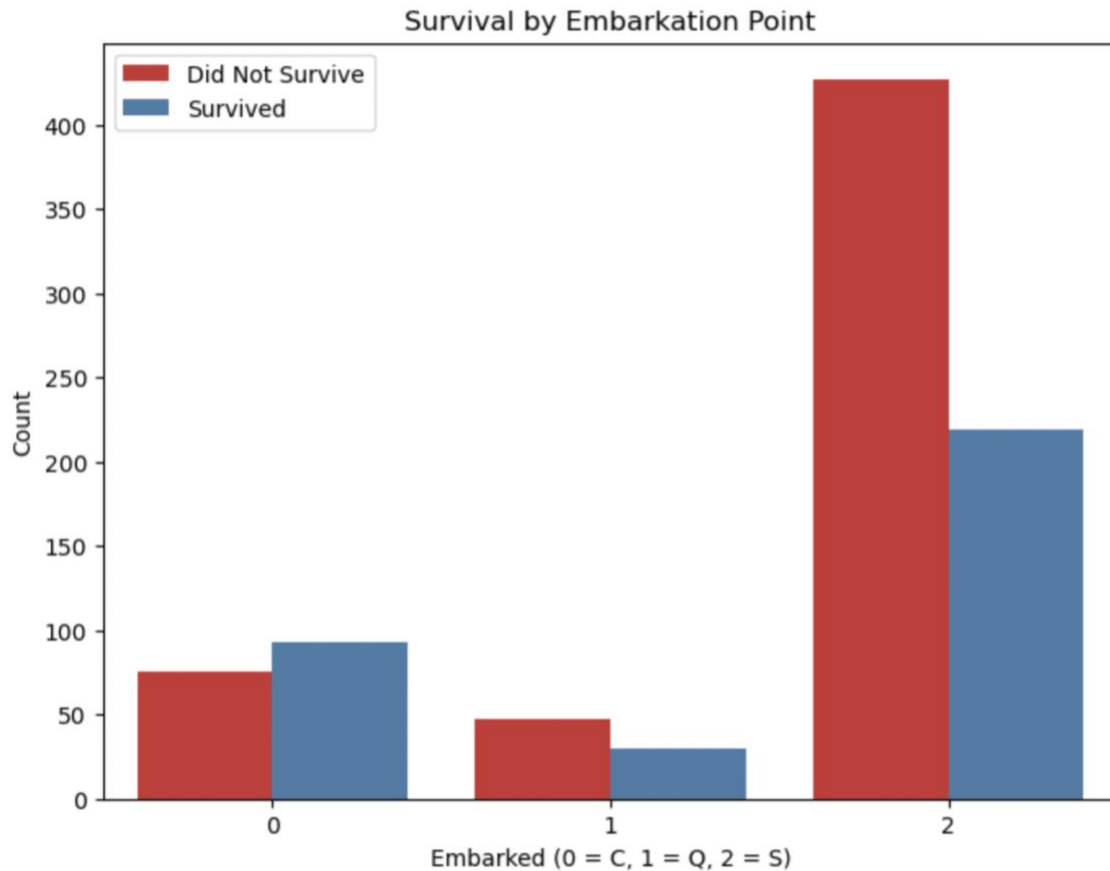
6.Survival Rate by Fare Quartile Analysis

The line chart titled "Survival Rate by Fare Quartile" plots survival rates (%) on the y-axis against fare quartiles (Q1 to Q4) on the x-axis, using a purple line with markers. Fares were grouped into four quartiles using `pd.qcut`, dividing passengers into equal-sized bins: Q1 (0 to 7.91), Q2 (7.91 to 14.4542), Q3 (14.4542 to 31), and Q4 (31 to 512.3292), each containing roughly 223 passengers. The chart shows a steady increase from 24.66% in Q1 to 54.05% in Q4, with Q2 at 34.53% and Q3 at 40.36%. This upward trend indicates that higher fares, often linked to first-class tickets, significantly improved survival chances, likely due to better lifeboat access. The grouping highlights the clear correlation between ticket price and survival probability.



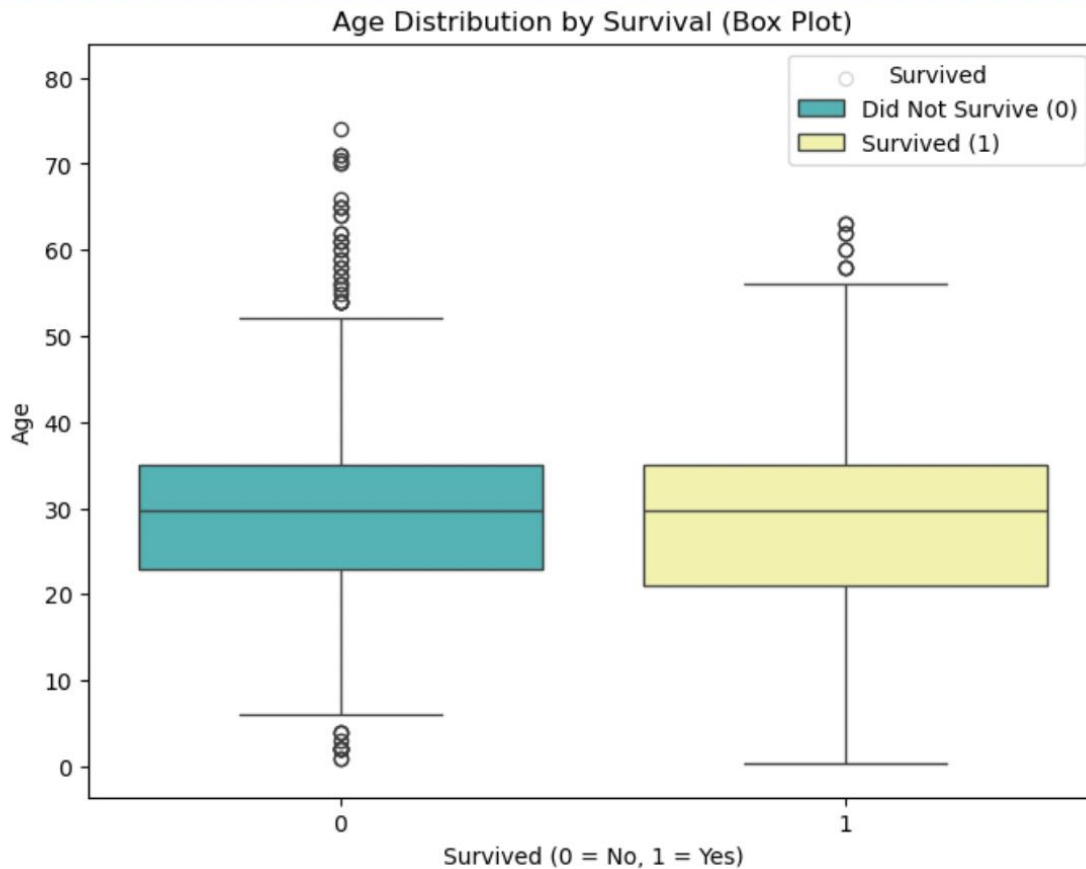
7.Survival by Embarkation Point Analysis

The bar chart titled "Survival by Embarkation Point" displays the count of passengers by their embarkation port (0 = Cherbourg, 1 = Queenstown, 2 = Southampton) on the x-axis, with count on the y-axis, using red for "Did Not Survive" and blue for "Survived." Southampton (2) has the highest count, with approximately 350 non-survivors and 200 survivors, indicating a large passenger volume but lower survival rate. Cherbourg (0) shows about 75 non-survivors and 100 survivors, while Queenstown (1) has around 50 non-survivors and 25 survivors, suggesting better survival proportions at these ports. This suggests that embarkation point influenced survival, possibly due to class distribution or evacuation priorities.



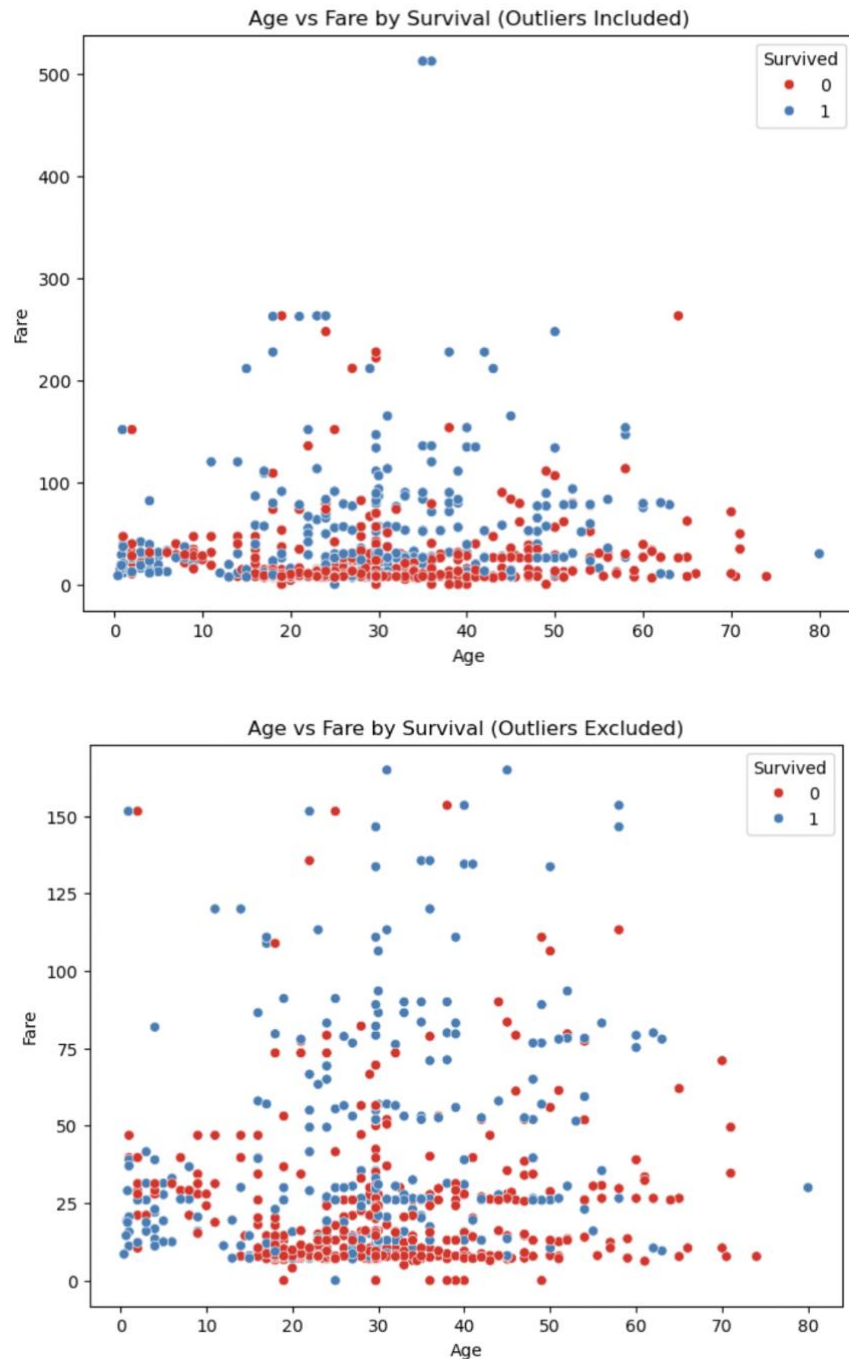
8.Age Distribution by Survival (Box Plot) Analysis

The box plot "Age Distribution by Survival" shows age distribution by survival status (0 = No, 1 = Yes), with cyan for non-survivors and yellow for survivors. Non-survivors have a median age of 28-30 (IQR 20-40, outliers to 80), while survivors have a median of 30-35 (IQR 20-40, fewer outliers above 60). This indicates both groups were mostly young to middle-aged, but survivors were slightly older on average, with fewer elderly survivors, reflecting possible age-based evacuation priorities



9. Age vs Fare by Survival Analysis

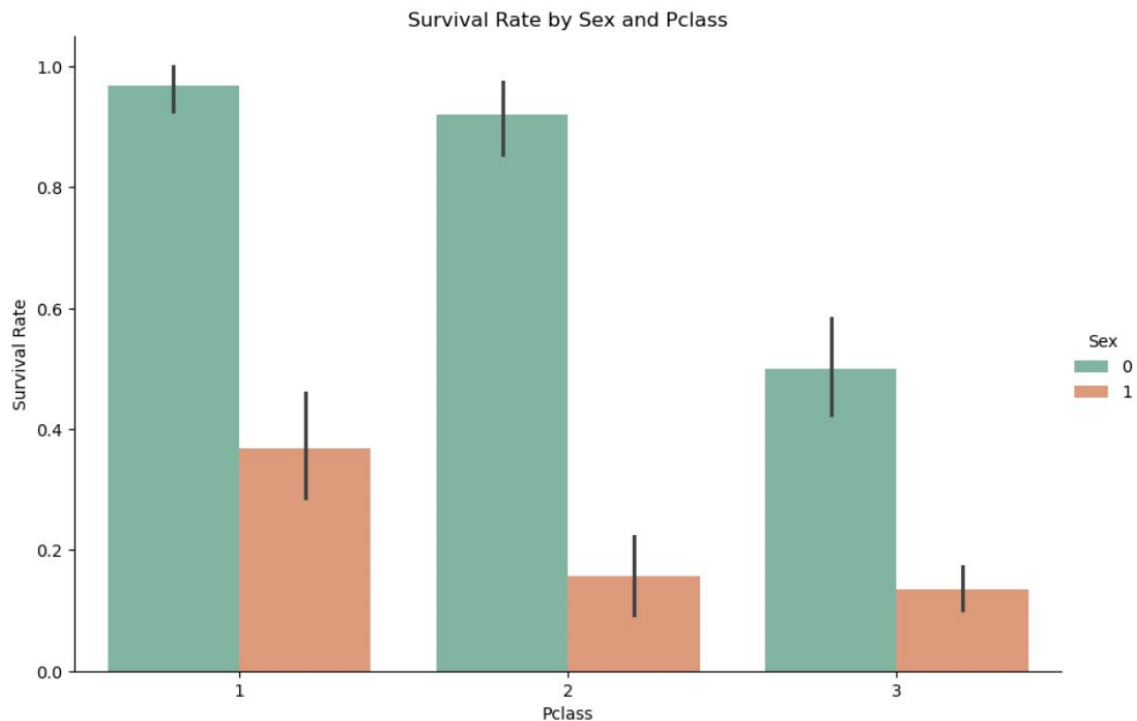
The scatter plot titled "Age vs Fare by Survival" displays age (0 to 80) on the x-axis and fare (0 to ~500) on the y-axis, with points colored by survival status: red for "Did Not Survive (0)" and blue for "Survived (1)." Most points cluster at fares below 100 and ages 20-40, with a dense mix of red and blue, indicating both survivors and non-survivors across this range. Higher fares (200-500) show more blue points, suggesting that passengers who paid more were more likely to survive, regardless of age. Younger passengers (0-20) have a noticeable number of blue points, reflecting higher survival rates among children, while older ages (60-80) have fewer points, mostly red, indicating lower survival for the elderly. This suggests that higher fares and younger ages were associated with better survival chances, likely due to class-based evacuation priorities and the "women and children first" policy.



10.Survival Rate by Sex and Pclass Analysis

The bar chart titled "Survival Rate by Sex and Pclass" illustrates survival rates (0 to 1) on the y-axis across passenger classes (Pclass: 1, 2, 3) on the x-axis, with bars colored by sex: teal for females (0) and orange for males (1). Females in Pclass 1 and 2 exhibit high survival rates of 95.60% and 92.11%, respectively, while in Pclass 3, their rate drops to 51.70%. Males have much lower survival rates across all classes: 39.20% in Pclass 1, 15.74% in Pclass 2, and 12.50% in Pclass 3. This highlights a stark gender disparity, with females in higher classes having the best

survival chances, likely due to the "women and children first" policy, while males in lower classes faced the lowest survival rates, reflecting class-based evacuation priorities.



11. Correlation Heatmap of All Features Analysis

The correlation heatmap titled "Correlation Heatmap of All Features" displays the Pearson correlation coefficients between all numerical features of the Titanic dataset, ranging from -1.0 (strong negative correlation, dark blue) to 1.0 (strong positive correlation, dark red). The diagonal (e.g., Survived vs Survived, Pclass vs Pclass) shows perfect correlation (1.0) as expected. Key correlations include:

- **Survived vs Pclass:** -0.34 (moderate negative, red), indicating lower survival rates in higher (lower-numbered) classes.
- **Survived vs Sex:** -0.54 (strong negative, red), showing females (0) had higher survival than males (1).
- **Pclass vs Fare:** -0.55 (strong negative, red), reflecting higher fares in lower classes (1).
- **SibSp vs Parch:** 0.41 (moderate positive, orange), suggesting families with more siblings/spouses also had more parents/children.
- **Fare vs FamilySize:** 0.22 (weak positive, light orange), indicating a slight tendency for larger families to pay more.

- Weak or negligible correlations (near 0, light colors) exist for PassengerId, Age, and Embarked with most features

