Exploratory Data Analysis: Observations for Each Visual

The Approach I used To Solve The Problem:

Step 1: Initial Data Loading and Inspection

- **Action: Loaded the train.csv dataset into a pandas DataFrame.**
- **Information Gathered:**
    - **Checked basic dataset information (.info()) to understand data types and non-null counts for each column.**
    - **Obtained descriptive statistics (.describe()) for numerical columns (mean, std, min, max, quartiles).**
    - **Analyzed value counts (.value_counts()) for key categorical columns like 'Survived', 'Pclass', 'Sex', and 'Embarked' to understand their distribution.**
    - **Identified the number of missing values (.isnull().sum()) in each column, noting significant missing data in 'Age', 'Cabin', and 'Embarked'.**

Step 2: Missing Value Handling

- **Action:**
    - **Imputed missing 'Age' values with the *median* age to handle numerical missing data robustly.**
    - **Imputed missing 'Embarked' values with the *mode* (most frequent port) as it's a categorical feature.**
    - **Dropped the 'Cabin' column entirely due to its high percentage of missing values (over 70%), which would make reliable imputation or direct use challenging for a basic EDA.**

Step 3: Visual Exploration - Histograms of Numerical Features

- **Action: Generated histograms for 'Age', 'Fare', 'SibSp', and 'Parch'.**
- **Observations:**
    - **'Age' showed a slightly right-skewed distribution, peaking around 20-30 years, with a notable number of children.**
    - **'Fare' was highly right-skewed, indicating most passengers paid low fares, with a few paying very high fares.**
    - **'SibSp' and 'Parch' distributions revealed that the majority of passengers traveled alone or with very few family members.**

Step 4: Visual Exploration - Boxplots for Numerical Features by Survival

- **Action: Created boxplots comparing 'Age' and 'Fare' distributions against 'Survived' status.**
- **Observations:**

    - **Age vs. Survival: The median age for non-survivors was slightly higher, with a slightly wider age range for those who perished.**
    - **Fare vs. Survival: Survivors generally paid significantly higher fares, indicating a strong correlation between fare (and thus class) and survival chances.**

Step 5: Visual Exploration - Countplots for Categorical Features

- **Action: Produced countplots for 'Survived', 'Pclass', 'Sex', and 'Embarked'.**
- **Observations:**
    - **Overall Survival: More passengers did not survive than survived.**
    - **Passenger Class: 3rd class had the most passengers, followed by 1st and 2nd.**
    - **Gender: There were more male passengers than female passengers.**
    - **Port of Embarkation: Southampton ('S') was the most common embarkation port.**

Step 6: Visual Exploration - Countplots for Categorical Features vs. Survived

- **Action: Generated countplots to show survival rates across 'Pclass', 'Sex', and 'Embarked'.**
- **Observations:**
    - **Survival by Pclass: 1st class had the highest survival rate, 3rd class the lowest.**
    - **Survival by Sex: Females had a significantly higher survival rate than males (strongest factor).**
    - **Survival by Embarked: Passengers from Cherbourg ('C') showed a relatively higher survival rate.**

Step 7: Visual Exploration - Pairplot of Numerical Features by Survival

- **Action: Created a pairplot for selected numerical features ('Age', 'Fare', 'SibSp', 'Parch', 'Survived') to visualize their pairwise relationships and distributions, colored by survival status.**
- **Observations: This plot visually confirmed that higher fares correlate with survival and highlighted that larger family sizes (higher SibSp/Parch) were associated with lower survival rates.**

Step 8: Visual Exploration - Correlation Heatmap of Numerical Features

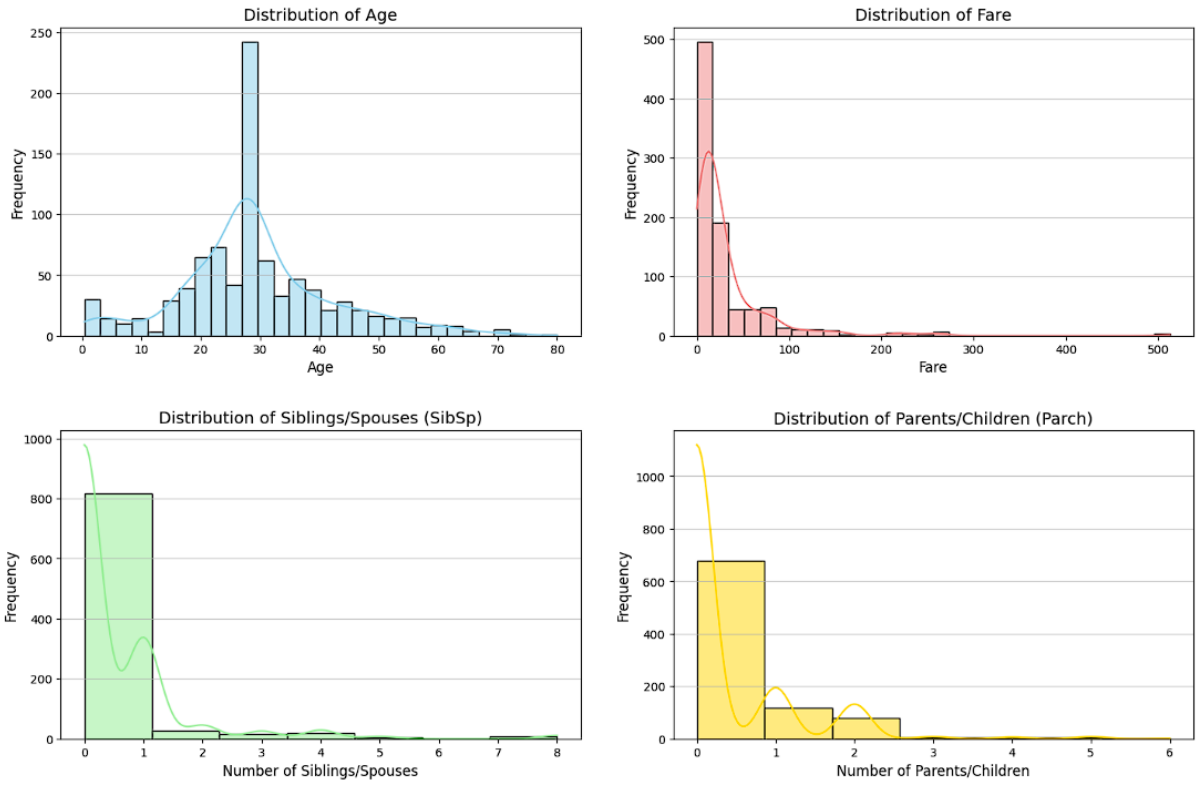
- **Action: Generated a heatmap showing the correlation matrix of all numerical features.**
- **Observations:**
    - **Fare had a positive correlation with Survived, and Pclass had a negative correlation (meaning 1st class correlates with survival).**
    - **SibSp and Parch were positively correlated with each other.**
    - **Age showed a very weak correlation with Survived.**

Step 9: Visual Exploration - Scatterplots

- **Action: Created scatterplots for 'Age vs. Fare', 'Age vs. Pclass', and 'SibSp vs. Parch', colored by 'Survived' status.**
- **Observations:**
    - **Age vs. Fare by Survival: Reinforced that high-fare passengers predominantly survived.**
    - **Age vs. Pclass by Survival: Clearly showed the age distribution within each class and how survival varied across classes.**
    - **SibSp vs. Parch by Survival: Suggested that very large families had lower survival rates.**

Findings:

1. **Histograms of Numerical Features**

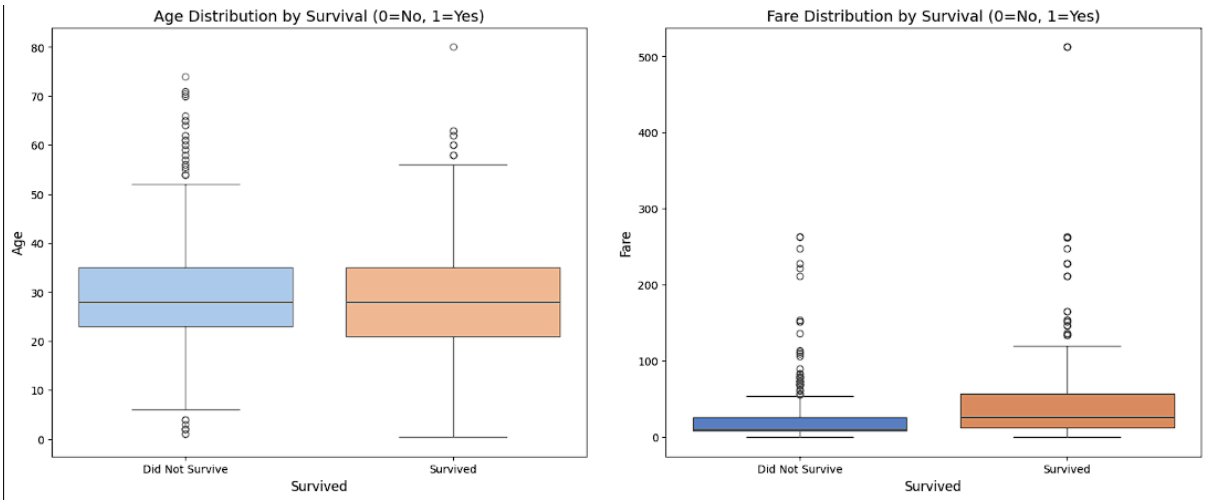


- **Distribution of Age:**
    - **Observation:** The age distribution, after median imputation for missing values, appears to be slightly right-skewed, with a prominent peak around 20-30 years. This indicates that the majority of passengers were young adults. There's also a noticeable number of very young passengers (children) and a decreasing frequency as age increases beyond 30.
- **Distribution of Fare:**
    - **Observation:** The fare distribution is highly right-skewed, with a large concentration of passengers paying low fares (below $50). A few passengers paid significantly

higher fares, suggesting a wide range of economic backgrounds and the presence of luxury cabins. This skewness is common in real-world pricing data.

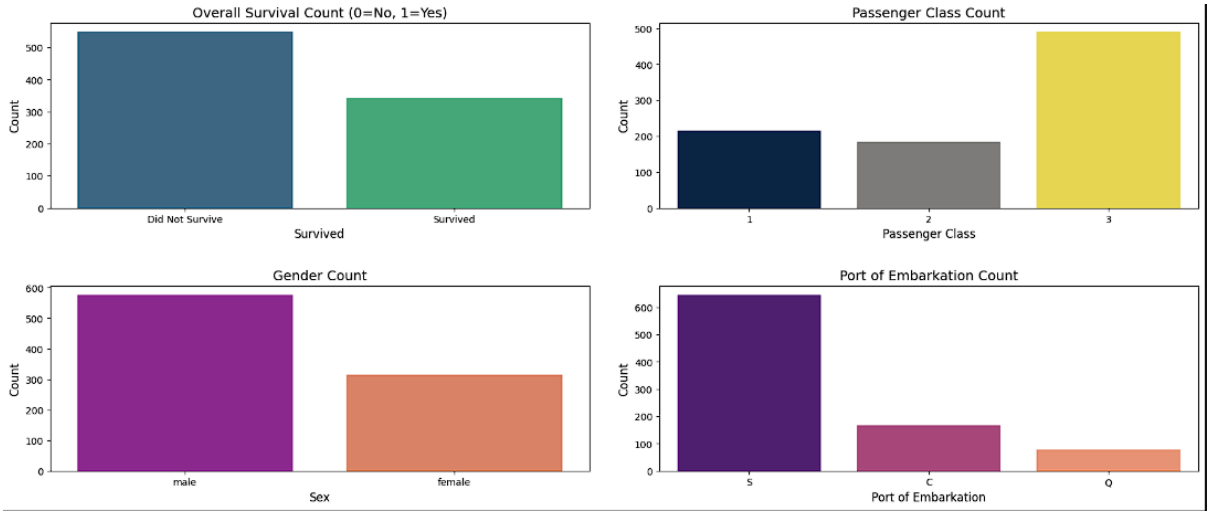
- **Distribution of SibSp (Siblings/Spouses Aboard):**
    - **Observation:** The vast majority of passengers (~600) traveled without siblings or spouses (SibSp = 0). A smaller proportion had 1 sibling/spouse, and very few had 2 or more. This suggests that many passengers were traveling alone or with a single family member.
- **Distribution of Parch (Parents/Children Aboard):**
    - **Observation:** Similar to SibSp, most passengers (~700) traveled without parents or children (Parch = 0). Very few had 1 or 2 parents/children, and even fewer had larger families on board. This further reinforces the idea that many passengers were traveling independently or with very small immediate families.

2. Boxplots for Numerical Features by Survival



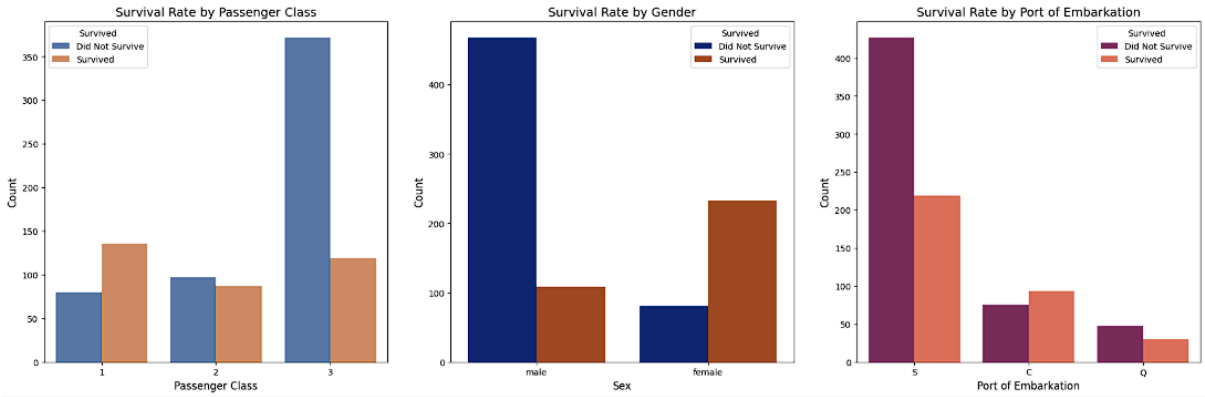
- **Age Distribution by Survival:**
    - **Observation:** The median age for those who did not survive (0) is slightly higher than for those who survived (1), although the difference is not very large. The interquartile range (IQR) for non-survivors appears slightly wider, suggesting a more varied age group among those who perished. Notably, there are some outliers in both groups, especially older individuals who did not survive.
- **Fare Distribution by Survival:**
    - **Observation:** Passengers who survived (1) generally paid significantly higher fares compared to those who did not survive (0). The median fare for survivors is substantially greater, and the distribution for survivors is much wider, indicating that high-fare passengers (likely from higher classes) had a better chance of survival. This plot shows a strong relationship between fare and survival.

3. Countplots for Categorical Features



- **Overall Survival Count (0=No, 1=Yes):**
    - **Observation:** There were more passengers who did not survive (549) than those who survived (342). This visually confirms the initial statistic that the overall survival rate was less than 50%.
- **Passenger Class Count:**
    - **Observation:** The 3rd class had the highest number of passengers (491), followed by 1st class (216) and 2nd class (184). This reflects the class demographics on the Titanic, with the largest proportion of passengers in the most affordable class.
- **Gender Count:**
    - **Observation:** There were significantly more male passengers (577) than female passengers (314) on board.
- **Port of Embarkation Count:**
    - **Observation:** Southampton ('S') was the most common port of embarkation, with a large majority of passengers boarding there. Cherbourg ('C') was the second most common, followed by Queenstown ('Q').

4. Countplots for Categorical Features vs. Survived



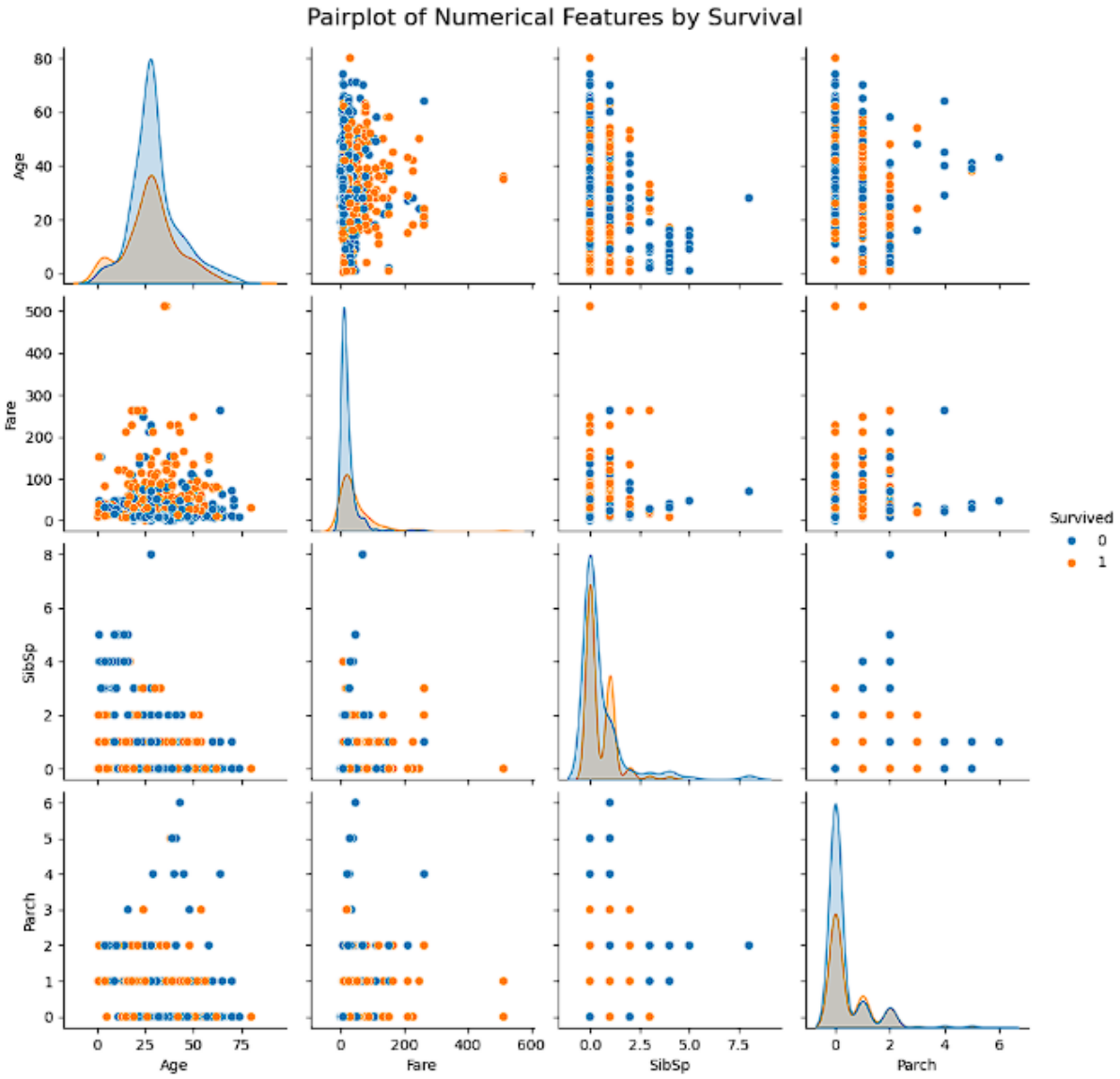
- **Survival Rate by Passenger Class:**

- **Observation:** A stark difference in survival rates is evident across passenger classes. 1st class passengers had the highest survival rate, followed by 2nd class. 3rd class passengers had the lowest survival rate, with the majority of non-survivors coming from this class. This suggests that class was a major determinant of survival.

- **Survival Rate by Gender:**
  - **Observation:** Female passengers had a significantly higher survival rate than male passengers. This is one of the most prominent observations, strongly supporting the "women and children first" evacuation protocol.

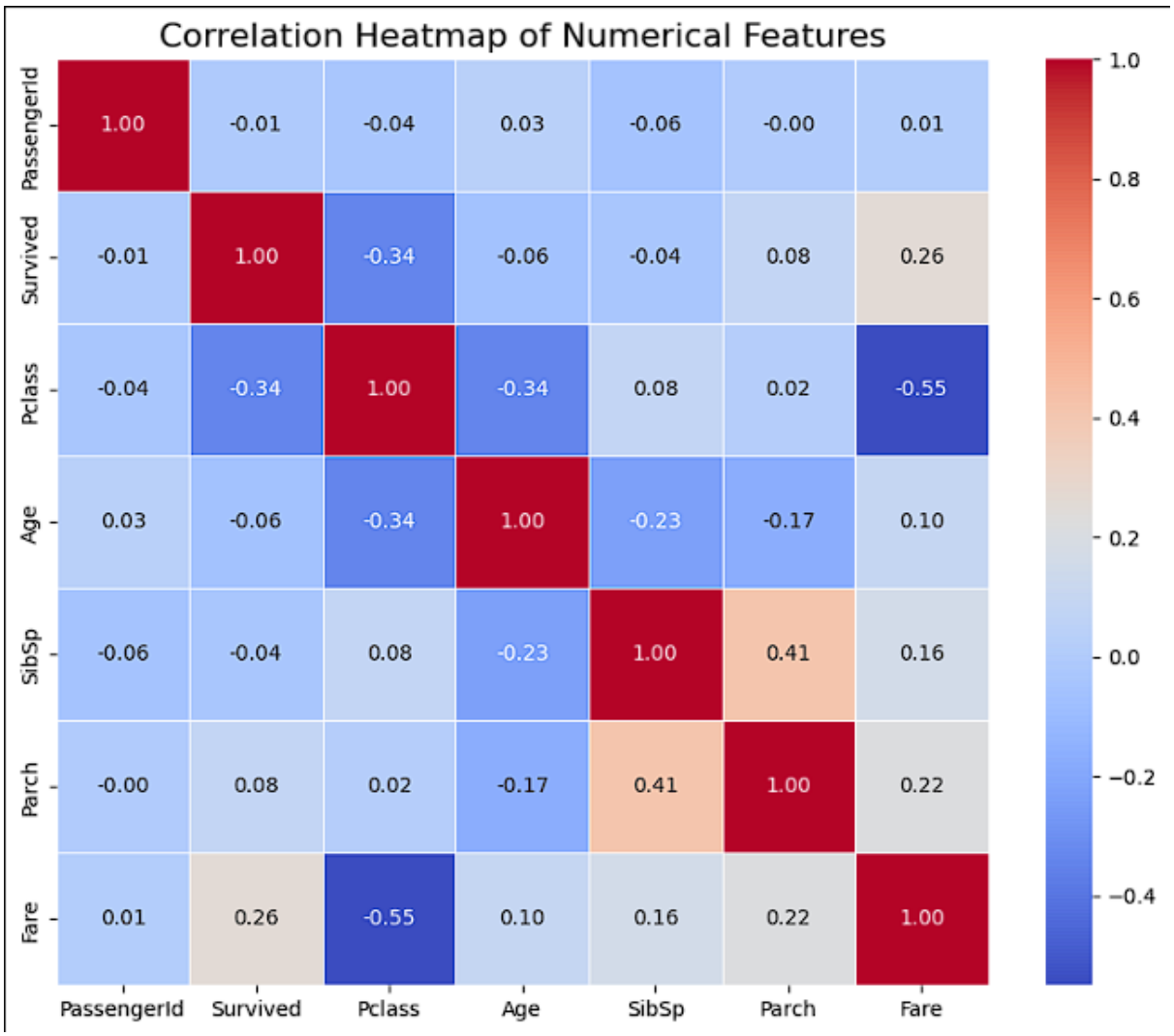
- **Survival Rate by Port of Embarkation:**
  - **Observation:** Passengers who embarked from Cherbourg ('C') appear to have a higher survival rate relative to their numbers compared to those from Southampton ('S') and Queenstown ('Q'). This might be correlated with the class composition of passengers from each port.

5. Pairplot of Numerical Features by Survival



- **Observation:** The pairplot provides a holistic view of relationships between numerical features, colored by survival status.
    - **Age vs. Fare:** Survivors (orange) tend to cluster in the higher fare ranges, regardless of age, and there's a denser cluster of non-survivors (blue) at lower fares.
    - **Age vs. SibSp/Parch:** For lower SibSp and Parch values (i.e., smaller families or individuals), survival is more mixed. However, for larger family sizes (higher SibSp or Parch), the blue points (non-survivors) dominate, suggesting that traveling in large groups was a disadvantage for survival.
    - The diagonal plots show the individual distributions, reaffirming observations from the histograms, with colors indicating survival groups.

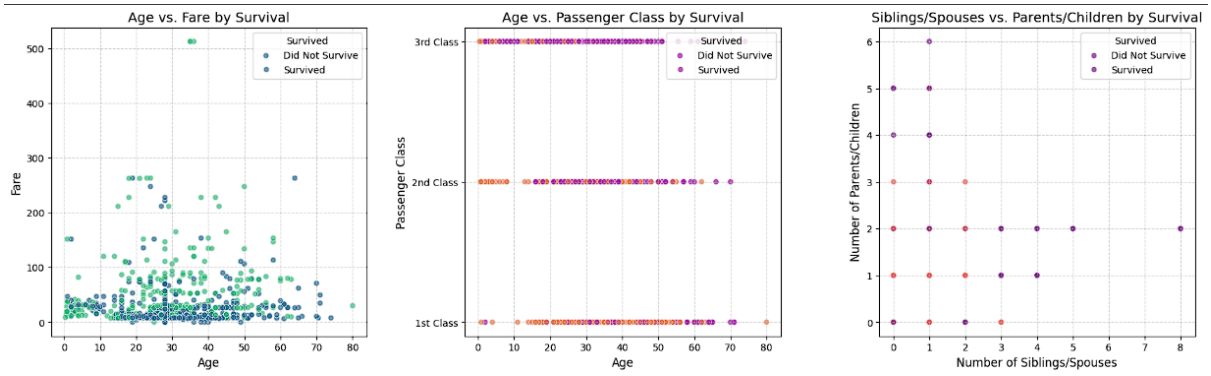
6. Correlation Heatmap of Numerical Features



- **Observation:**
    - Survived shows a moderate positive correlation with Fare (approx. 0.26), indicating that higher fares (and likely higher class) are associated with higher survival rates.

- Survived shows a moderate negative correlation with Pclass (approx. -0.34), meaning that a lower Pclass value (i.e., 1st class) is positively correlated with survival.
- SibSp and Parch are positively correlated (approx. 0.41), which is expected as they both represent family members.
- Age has a very weak negative correlation with Survived (approx. -0.06), suggesting that age alone (after median imputation) is not a strong linear predictor of survival.

7. Scatterplots



- **Age vs. Fare by Survival:**
  - **Observation:** There's a clear visual separation by fare: passengers who paid higher fares are predominantly survivors, regardless of age. Non-survivors are heavily concentrated at lower fare values. This reiterates the strong influence of fare/class on survival.
- **Age vs. Passenger Class by Survival:**
  - **Observation:** This plot distinctly shows the distribution of ages within each passenger class. 1st class has a wide age range of survivors, while 3rd class has a denser cluster of non-survivors across various ages. Very young children (under ~5-10) in 3rd class show some survival, but the overall trend for 3rd class is high non-survival.
- **Siblings/Spouses vs. Parents/Children by Survival:**
  - **Observation:** Passengers with no siblings/spouses and no parents/children (SibSp=0, Parch=0) form the largest cluster, and survival within this group is mixed. As SibSp or Parch increases (larger family sizes), the concentration of non-survivors (blue points) tends to become more dominant, particularly for families with 3 or more members in either category. This suggests that the presence of too many dependents or family members could have hindered escape.

Summary of Findings:

The comprehensive Exploratory Data Analysis of the Titanic dataset clearly indicates that **gender (Sex) and passenger class (Pclass), directly influenced by the Fare paid, were the most significant**

factors determining survival. Females and first-class passengers had a disproportionately higher chance of survival. While age itself wasn't a strong standalone predictor, very young children showed better survival, and the ability to escape with large families appeared to be a disadvantage. The port of embarkation also showed some correlation, likely due to underlying demographic differences of passengers from each port. These findings lay a solid foundation for further feature engineering and predictive modeling.