Navigating the Titanic: An In-Depth Exploration of Survival Factors

Description about the Dataset and summary of its attributes

The Titanic dataset contains information about the passengers and crew members who were on board the Titanic when it sank in 1912. The dataset contains of 1306 entries with 11 columns.

- **PassengerId**: A unique identifier for each passenger [categorical(nominal)]
- **Survived**: Whether or not the passenger survived the disaster (1 = Survived, 0 = Died) [categorical(nominal)]
- **Pclass**: The passenger's class (1 = first class, 2 = second class, 3 = third class) [categorical(ordinal)]
- **Name**: The passenger's name
- **Sex**: The passenger's gender (male or female) [categorical(nominal)]
- **Age**: The passenger's age [numeric(continuous)]
- **SibSp**: The number of siblings and spouses the passenger was traveling with. [numeric(continuous)]
- **Parch**: The number of parents and children the passenger was traveling with. [numeric(continuous)]
- **Ticket**: The passenger's ticket number [categorical(nominal)]
- **Fare**: The price the passenger paid for their ticket [numeric(discrete)]
- **Embarked**: The port where the passenger boarded the Titanic (Southampton, Cherbourg, or Queenstown) [categorical(ordinal)]

Initial Plan for Data Exploration

The initial data exploration focused on understanding key features such as survival rates, age distribution, gender distribution, passenger class distribution, and fare distribution. Visualizations, including countplot, pairplot, scatter plots were employed to gain insights.

Data cleaning and Feature engineering

Handling Missing Values:

- For the ‘Age’ column, 263 missing values were replaced with the median age.

- The ‘Fare’ column had one missing value, which was filled in with an appropriate value.
- The ‘Cabin’ column, with 1014 missing values, was dropped from the dataset.

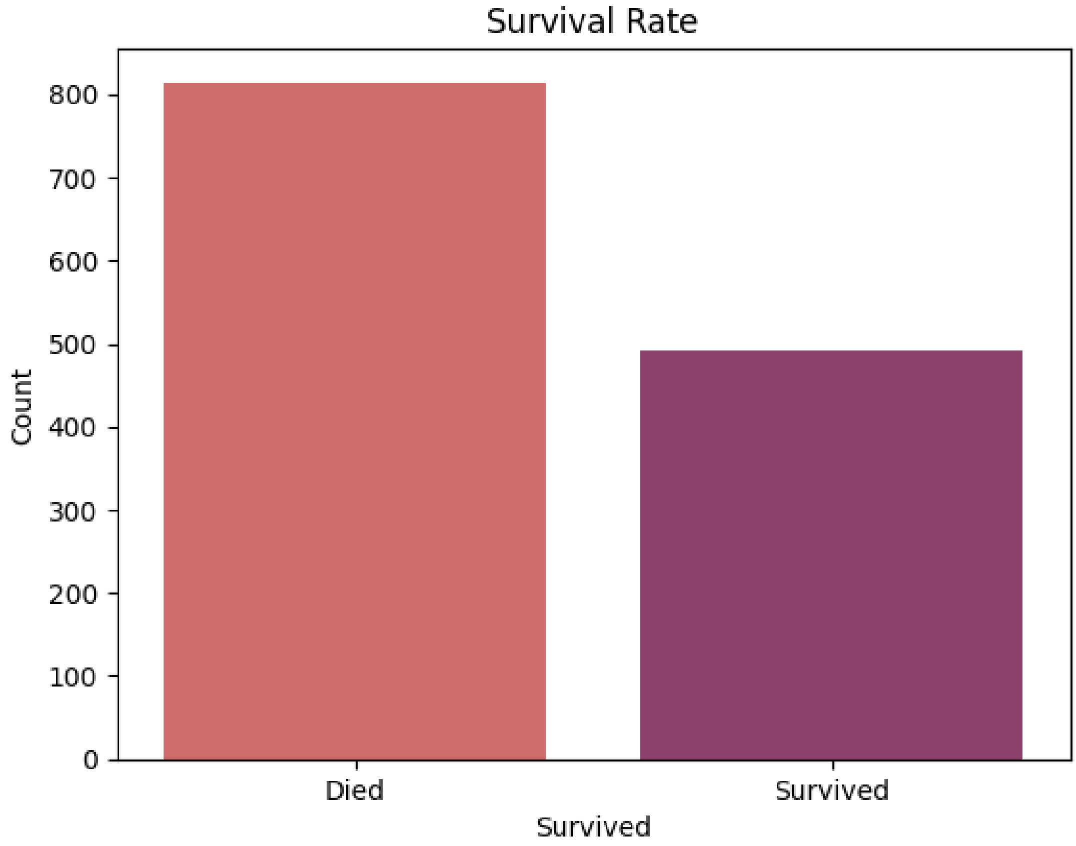
Handling Categorical Data:

- The 'Embarked' column labels ('S', 'C', 'Q') were replaced with meaningful port names ('Southampton', 'Cherbourg', 'Queenstown').
- The 'Survived' column values (0 and 1) were replaced with descriptive labels ('Died' and 'Survived').
- The 'Pclass' column values (1, 2, 3) were replaced with more interpretable class names ('First Class', 'Second Class', 'Third Class').

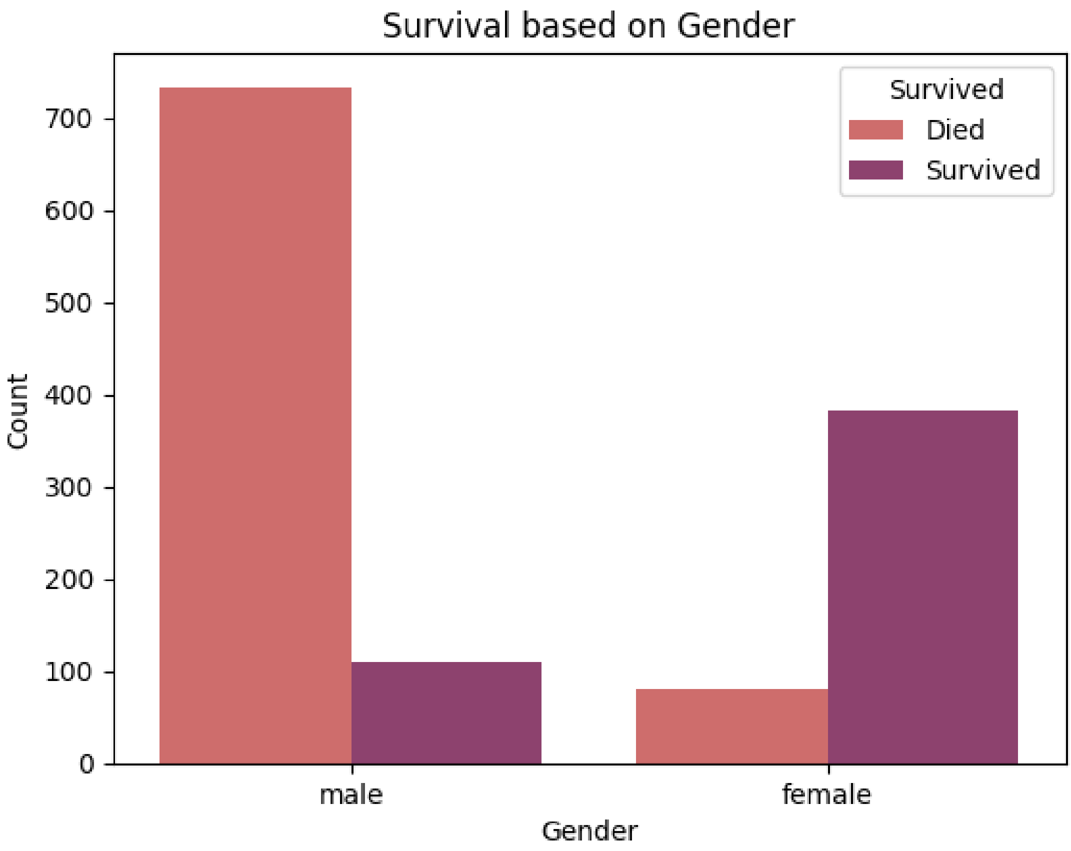
Family Size Feature:

- Created a new feature, 'FamilySize,' by summing the 'SibSp' and 'Parch' columns and adding 1 for the individual.

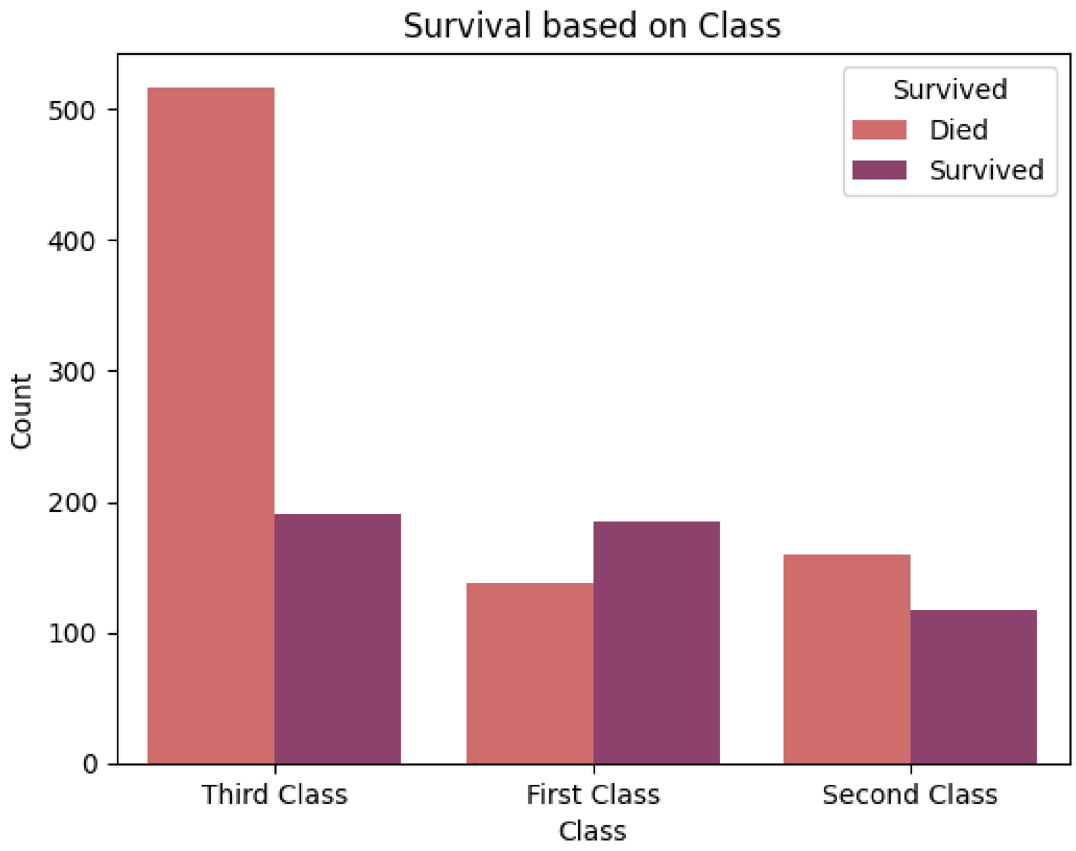
Key Findings and Insights



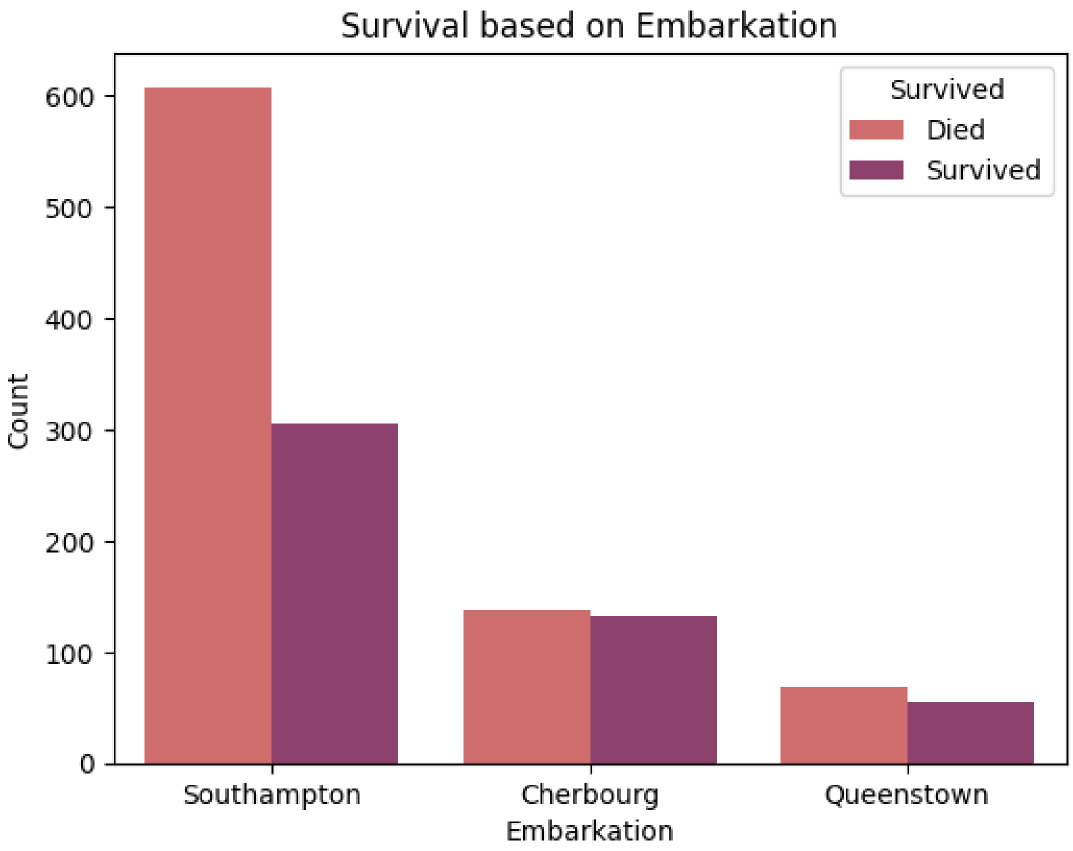
- The overall survival rate appears to be lower than the non-survival rate, indicating that the majority of passengers did not survive the Titanic disaster.



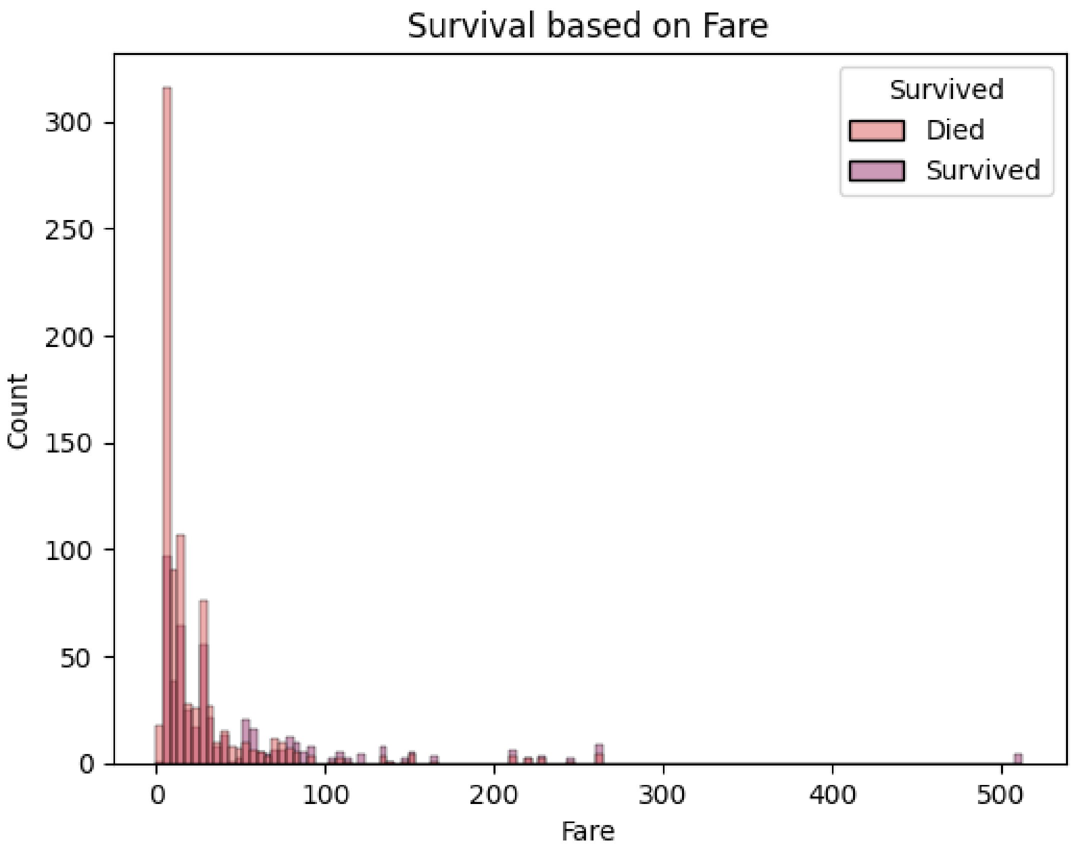
- Females has a notably higher survival rate compared to males. This aligns with the historical account of prioritizing women and children during the evacuation.



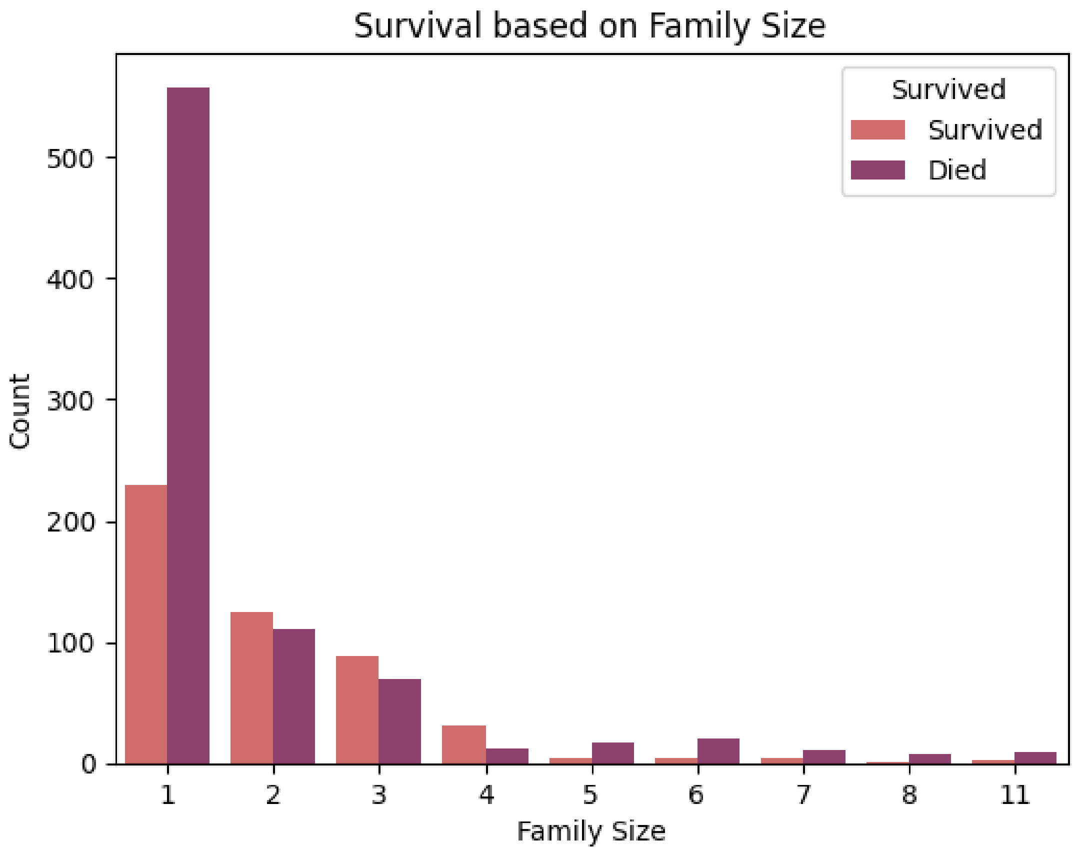
- Passengers in the first class had a higher chance of survival compared to those in the second and third classes. This suggests a potential correlation between social class and survival.



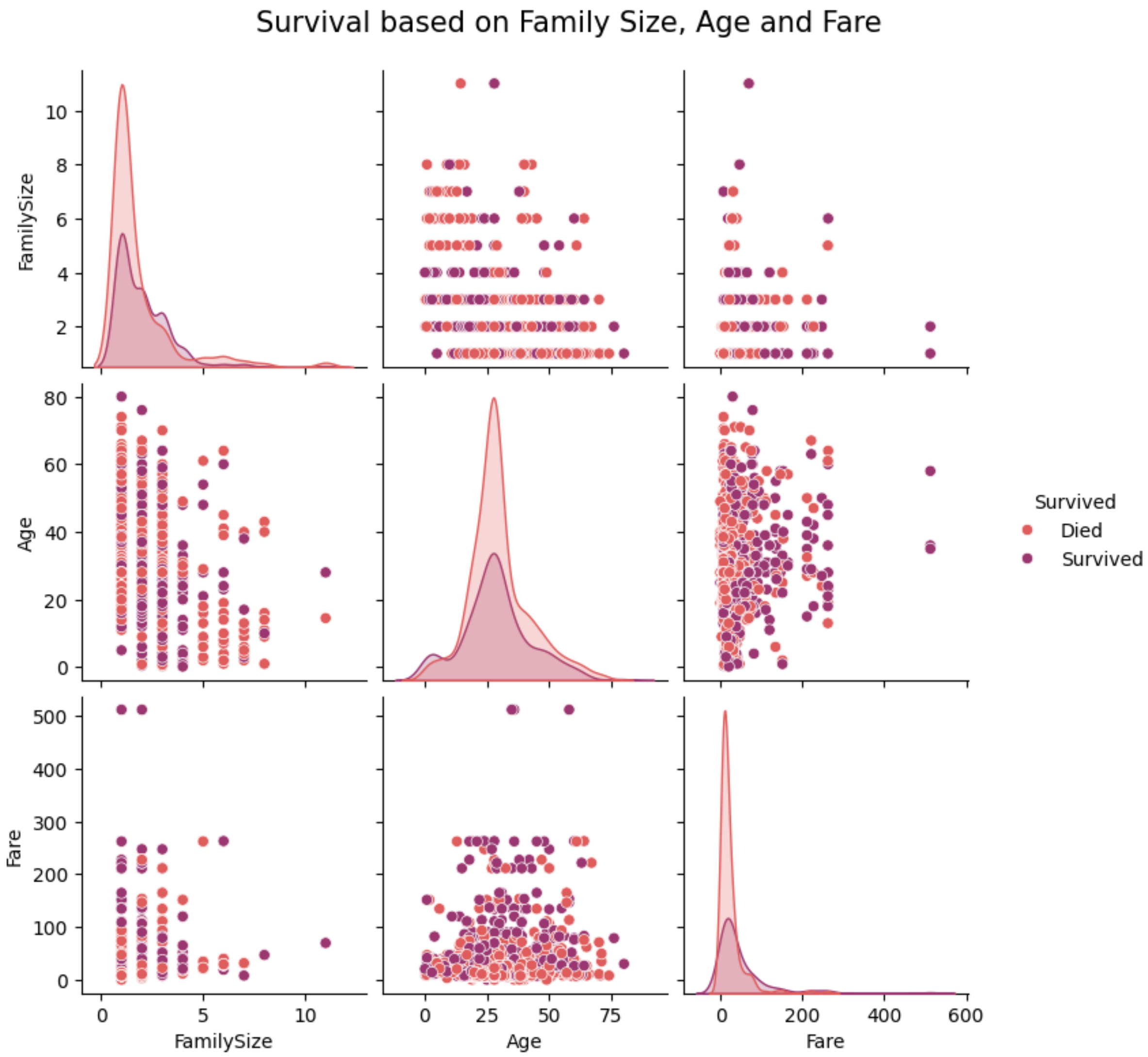
- Passengers who boarded from Southampton("S") has a lower survival rate compared to those from Cherbourg("C") and Queenstown("Q"). This could be due to various factors, such as the distribution of classes or other unexplored variables.



- Passengers who paid lower fares generally had a lower survival rate, while those who paid higher fares had a relatively higher survival rate.
- This suggests that individuals who paid higher fares might have had preferential access to lifeboats or better accommodations, potentially impacting their chances of survival



- Passengers with smaller family sizes (e.g., alone or with a small family) had a slightly higher survival rate compared to those with larger family sizes. This suggests that individuals or smaller groups may have had better chances of evacuation.



- Family Size vs Age: The scatter plots for ‘FamilySize’ against ‘Age’ indicate that passengers traveling alone or with a smaller family tended to be of varying ages. Larger families seem to include a wider range of ages, but no distinct pattern is evident.
- Family Size vs Fare: The scatter plot for ‘FamilySize’ against ‘Fare’ show that passengers with smaller family sizes generally paid a range of fares, while larger families predominantly paid lower fares. This suggests potential cost-saving choices for larger groups.
- Age vs Fare: The scatter plots for ‘Age’ against ‘Fare’ exhibit a diverse distribution of ages across different fare ranges. There isn’t a clear correlation between age and fare, indicating that passengers of various ages paid similar fares.

Insights Summary:

- Gender and class were influential factors in survival, with females and first-class passengers having higher survival rates.
- Embarkation port and family size may have some correlation with survival but require further investigation.

Formulating Hypotheses

1. Hypothesis 1:
   Null Hypothesis: No significant difference in survival rates between genders.
   Alternative Hypothesis: Females have a significantly higher survival rate.
2. Hypothesis 2:
   Null Hypothesis: Age is not correlated with the likelihood of survival.
   Alternative Hypothesis: Younger passengers are more likely to survive.
3. Hypothesis 3:
   Null Hypothesis: Passenger class does not affect the changes of survival.
   Alternative Hypothesis: Higher-class passengers have a higher survival rate.

Formal Significance Test:

Hypothesis:

Null Hypothesis: No significant difference in survival rates between genders.

Alternative Hypothesis: Females have a significantly higher survival rate.

A Chi-Square test of independence was performed on 'Survived' and 'Sex', resulting in a chi-square statistic of 614.17 and a p-value of 1.39e-135. With a p-value well below 0.05, we reject the null hypothesis, concluding that females exhibit a significantly higher survival rate than males. This aligns with historical records and establishes gender as a crucial factor influencing survival on the Titanic.

Suggestions for Next Steps

- Build a machine learning model to predict survival rate based on the different attributes in the dataset.

- Use the machine learning model to identify passengers who were at a higher risk of dying in the disaster.
- Investigate the factors that contributed to the sinking of the Titanic.

Quality of Dataset and Additional Data

The Titanic dataset is of high quality. Despite the missing values. The data is good and there were no duplicated values. Additionally, the data is well-organized and easy to understand.

I do not have any specific requests for additional data at this time. However, it would be interesting to have access to data about the Titanic's crew members. This data could provide additional insights into the factors that affected survival on the Titanic.