EDA of titanic dataset

Describe()

The describe() function provides a summary of the statistical properties of the dataset. For each numeric column, it includes the following key metrics:

1. **Count**: The number of non-null values in the column (e.g., 891 passengers).
2. **Mean**: The average value of the column (e.g., the average age of passengers is approximately 29.7 years).
3. **Standard Deviation (Std)**: A measure of the variability or spread of the data (e.g., the standard deviation for Fare is 49.69, indicating a wide range of fares).
4. **Minimum (Min)**: The lowest value in the column (e.g., the youngest passenger was 0.42 years old).
5. **25th Percentile (25%)**: The value below which 25% of the data falls (e.g., 25% of passengers are aged 20.13 years or younger).
6. **50th Percentile (50%)**: The median or middle value of the data (e.g., the median age is 28 years).
7. **75th Percentile (75%)**: The value below which 75% of the data falls (e.g., 75% of passengers are aged 38 years or younger).
8. **Maximum (Max)**: The highest value in the column (e.g., the oldest passenger was 80 years old).

For the "Survived" column, the mean value of 0.38 indicates that about 38% of passengers survived. The data also provides insights into the distribution of other features such as class (Pclass), age, the number of siblings/spouses aboard (SibSp), the number of parents/children aboard (Parch), and the fare paid for the ticket.

Info()

The info() function provides a concise summary of the dataset, including the following key details:

1. **RangeIndex**: The dataset consists of 891 entries (from index 0 to 890).
2. **Columns**: The dataset has 12 columns in total, each representing a different feature of the passengers.
3. **Non-Null Count**: Indicates the number of non-null (non-missing) values in each column. For example, the "Age" column has 714 non-null values, while the "Cabin" column has only 204 non-null values.
4. **Data Type (Dtype)**: The data types of each column are shown:

- **int64**: Integer values (e.g., PassengerId, Survived, Pclass, SibSp, Parch).
- **float64**: Floating-point values (e.g., Age, Fare).
- **object**: Object (string) values (e.g., Name, Sex, Ticket, Cabin, Embarked).

5. **Memory Usage**: The dataset occupies 83.7 KB of memory.

Value counts()

A short description of the value counts for the columns

1. **Sex:**
    - 577 male passengers
    - 314 female passengers

This indicates a higher number of male passengers compared to female passengers in the dataset.

2. **Survived:**
    - 549 passengers did not survive (Survived = 0)
    - 342 passengers survived (Survived = 1)

This shows that a larger proportion of passengers did not survive the incident, with only about 38% surviving.

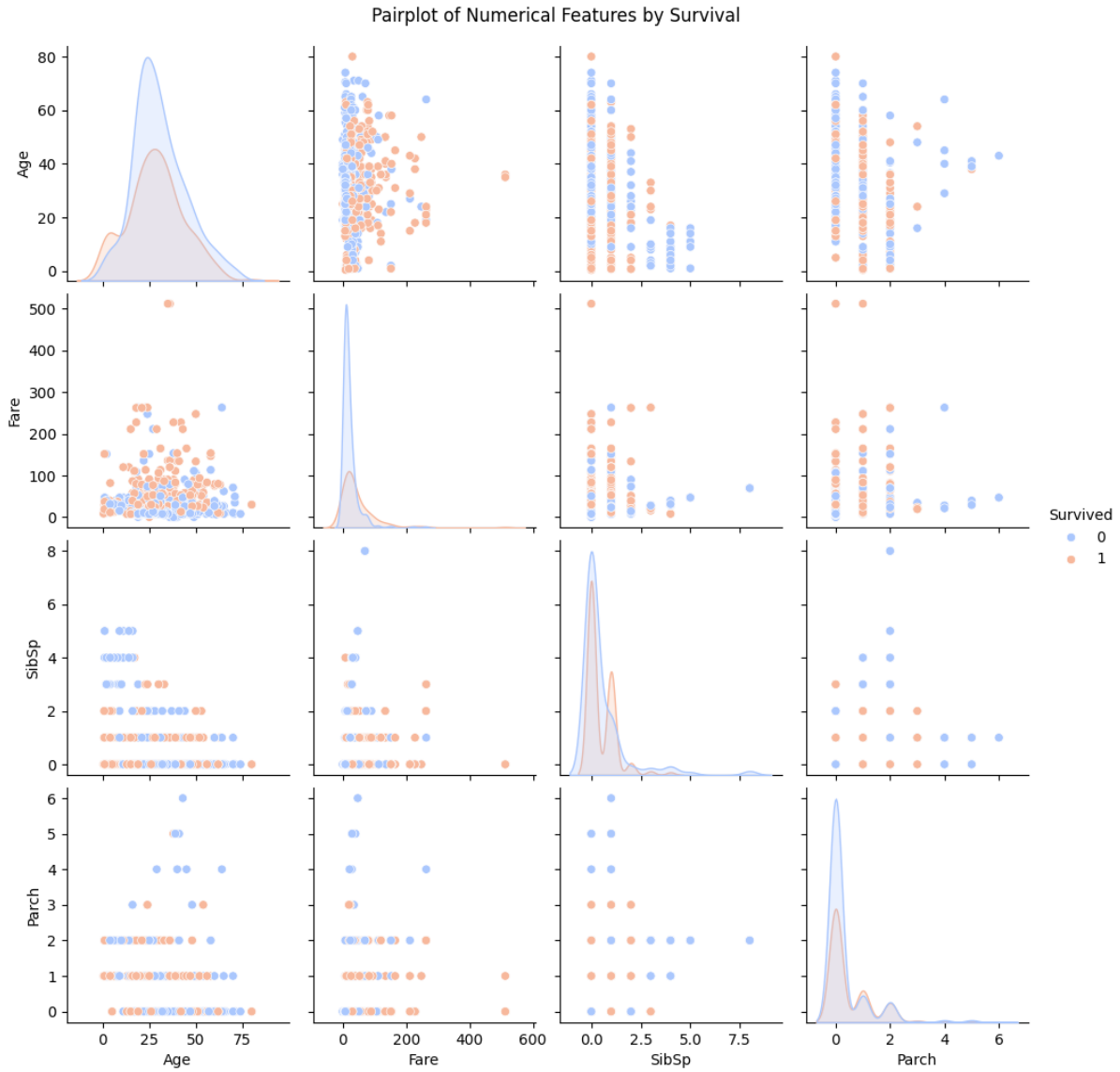
3. **Pclass (Passenger Class):**
    - 491 passengers were in third class (Pclass = 3)
    - 216 passengers were in first class (Pclass = 1)
    - 184 passengers were in second class (Pclass = 2)

The majority of passengers were in third class, followed by first class, and the least number were in second class.

4. **Embarked (Embarkation Point):**
    - 644 passengers embarked from Southampton (S = S)
    - 168 passengers embarked from Cherbourg (C = C)
    - 77 passengers embarked from Queenstown (Q = Q)

The majority of passengers boarded the ship in Southampton, followed by Cherbourg, and the fewest from Queenstown.

Pairplot of numerical features by survival



Insights from Pairplot Analysis

The pairplot provides a visual representation of numerical relationships and survival trends in the Titanic dataset. Below are key insights derived from the pairwise comparisons:

1. Fare vs. Survival

- Passengers who paid **higher fares** had a **greater probability of survival**, indicating that **cabin class played a significant role** in survival chances.
- Lower-fare passengers (closer to zero) show a **higher proportion of non-survivors**, suggesting that **third-class passengers faced more difficulties during evacuation**.

- Some outliers exist where **lower-fare passengers survived**, possibly due to **early access to lifeboats or strategic positioning**.

2. Age vs. Survival

- There is **no strong correlation** between age and survival, meaning passengers of **all ages had mixed survival outcomes**.
- However, infants and young children show a **slight survival advantage**, likely due to **prioritization during evacuation**.
- Elderly passengers had **a relatively lower survival rate**, possibly because of **mobility constraints during the chaotic rescue process**.

3. SibSp (Siblings/Spouses) vs. Survival

- Passengers traveling **alone** had **lower survival rates**, indicating that **having family onboard may have helped survival chances**.
- Those traveling with **1-2 relatives** show a **higher survival trend**, possibly due to **assistance and coordination during escape**.
- Families with **larger groups (3+ members)** saw **lower survival rates**, likely due to **difficulties in staying together and securing lifeboat space**.

4. Parch (Parents/Children) vs. Survival

- Individuals traveling with **at least one parent or child** had a **moderate survival advantage**, reinforcing the likelihood that **families were prioritized in rescue operations**.
- Lone travelers had **mixed survival trends**, depending on their **age and ticket class**.

5. General Pattern Across the Pairplot

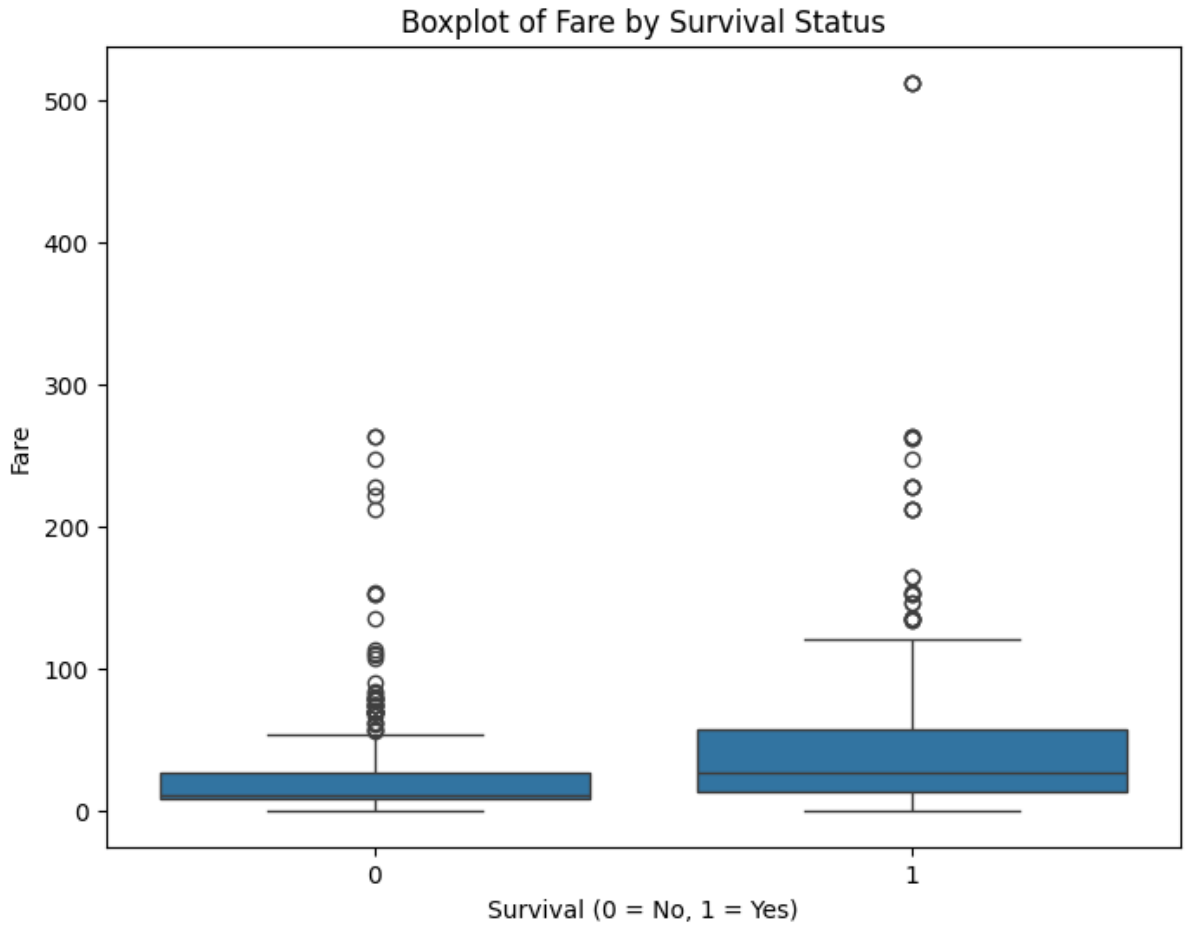
- **Fare had the strongest correlation with survival**, supporting the idea that **economic class influenced rescue priority**.
- **Smaller family groups had better survival rates** compared to larger ones.
- **Age shows no strong correlation**, though infants received a slight advantage.
- The data suggests **first-class passengers had better access to lifeboats**, significantly improving survival chances.

Conclusion:

This pairplot reveals important survival patterns, showing how economic status, age, and family size influenced a passenger's chance of survival. The visual analysis

suggests that **ticket class and group size were key factors**, while **age played a secondary role** in determining outcomes.

BoxPlot



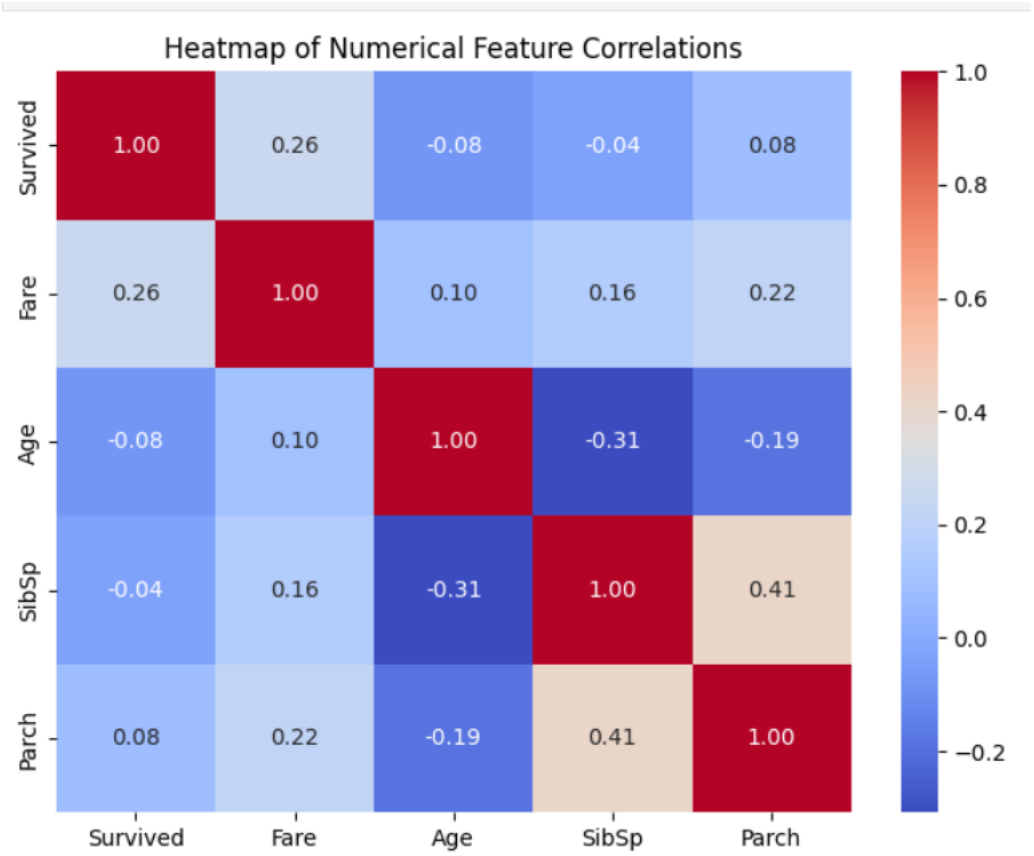
Here's an easy breakdown of insights from the boxplot comparing **Fare and Survival**:

1. **Higher Fare → Higher Survival**
    - Passengers who paid **higher fares** had a **better chance of survival**.
    - This suggests **first-class passengers were prioritized for lifeboats**.
2. **Lower Fare → More Non-Survivors**
    - Many passengers with **low fares** did not survive.
    - Third-class travelers had **less access to lifeboats**, affecting their survival rate.
3. **Outliers in Fare**
    - Some passengers paid **very high fares** (above 500).

- These could be **wealthy individuals in luxury cabins**, possibly having better rescue chan

Conclusion: This boxplot strongly indicates that **fare played a role in survival**, likely because **higher-class passengers had better access to lifeboats and assistance**. The economic divide is clearly reflected in the survival statistics.

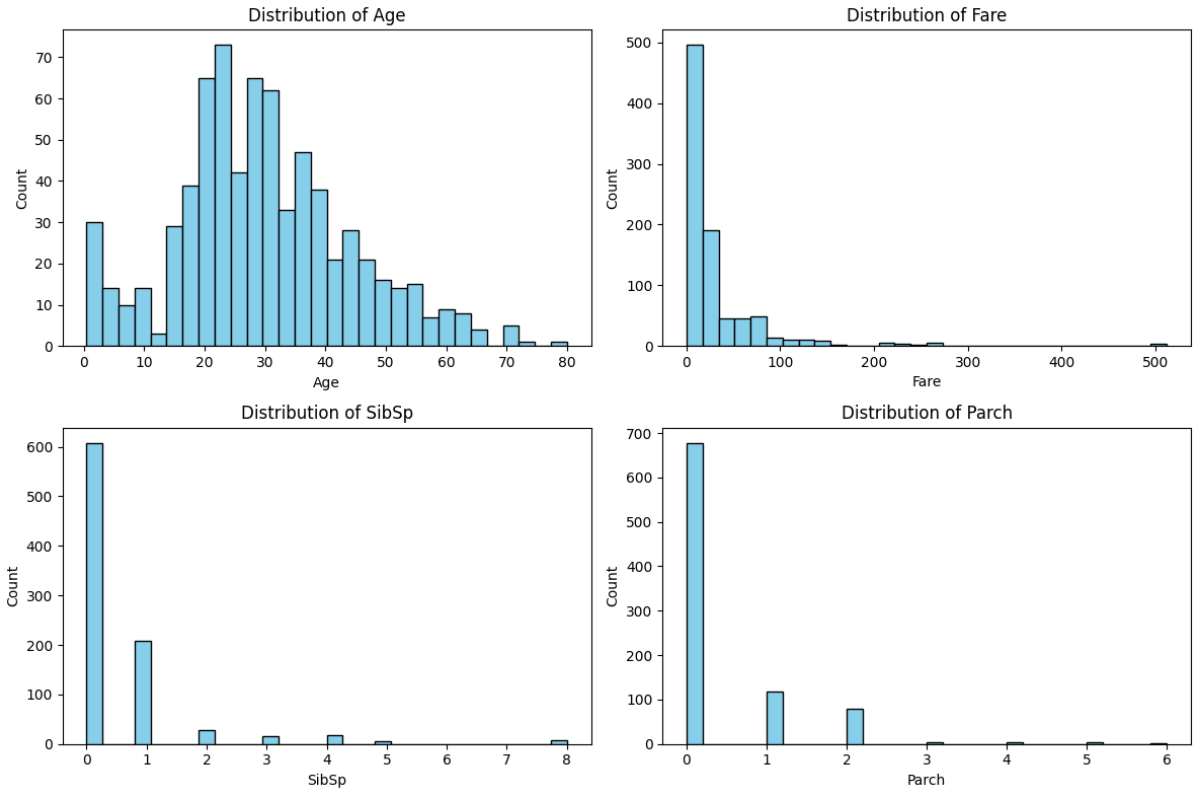
Heatmap



Here's a concise summary of the heatmap insights:

- **Fare & Survival:** Higher fare passengers had better survival chances, likely due to first-class priority in lifeboat access.
- **Age & Survival:** Weak correlation—infants had a slight advantage, but overall age wasn't a major factor.
- **Family Size (SibSp & Parch):** Small families had a better chance of survival, while larger groups struggled.
- **Fare & Family:** Wealthier passengers often traveled with family, influencing their accommodation level.
- **Age & Travel Companions:** Younger passengers were more likely to travel with parents, while older ones often traveled alone.
- **Overall Conclusion:** Fare had the strongest impact on survival, highlighting class-based rescue advantages.

Histograms



Histograms show the **distribution of numerical features** in the dataset, helping us understand the range and frequency of values. Below are the key observations:

1. Age Distribution

- Most passengers were between **20-40 years old**, peaking around age **25**.
- Very few passengers were **above 60**, indicating a younger population onboard.
- Infants and children (0-10 years) form a small but noticeable group.

2. Fare Distribution

- The **majority of fares** are between **0-50**, indicating that most passengers were in **lower-class tickets**.
- A few passengers paid **exceptionally high fares (above 200)**—these are likely **first-class travelers**.

- The distribution is **right-skewed**, meaning most passengers paid lower fares, but a few paid extremely high amounts.

3. SibSp (Siblings/Spouses) Distribution

- Most passengers were **traveling alone (SibSp = 0)**.
- Few people traveled with **1-2 siblings/spouses**, showing that small family units were common.
- Larger family groups (SibSp = 3 or more) were rare, indicating fewer large families onboard.

4. Parch (Parents/Children) Distribution

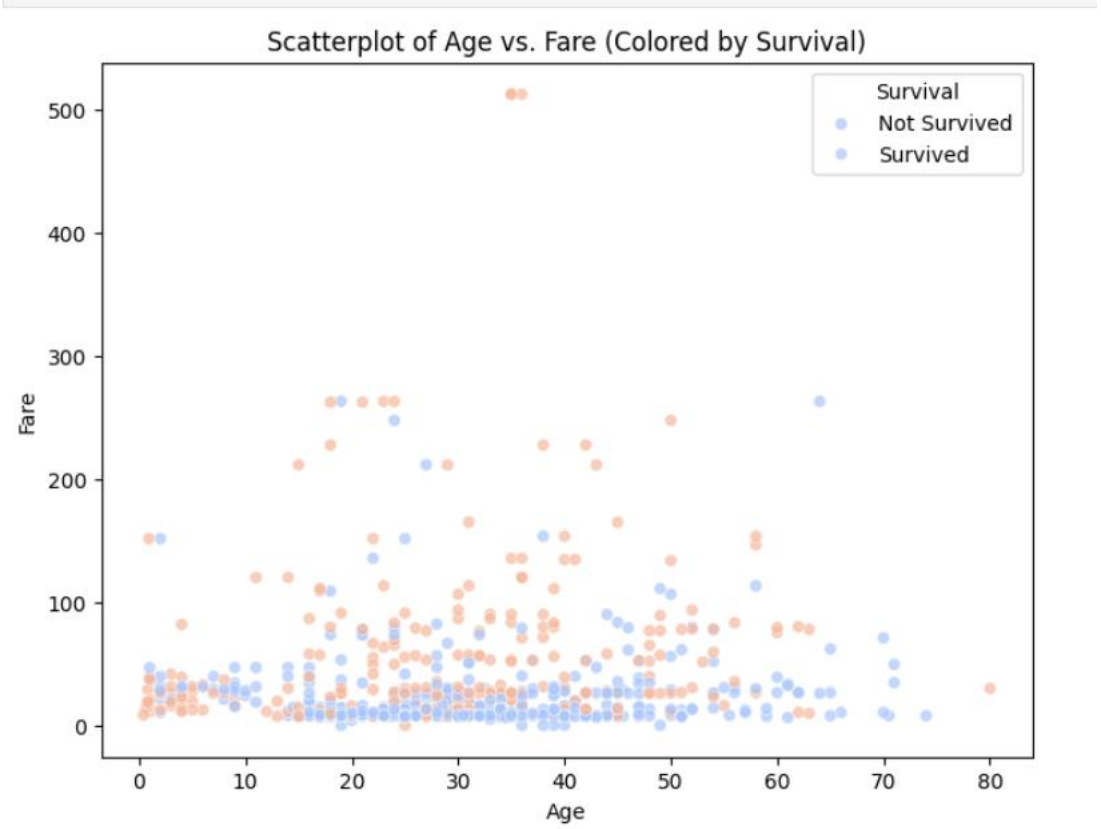
- Most passengers had **zero parents or children aboard**, meaning they traveled alone or with other relatives.
- Some passengers had **1-2 parents or children**, suggesting **family travel was less common**.
- Very few passengers had **3 or more family members onboard**, reinforcing the trend of small family groups.

Conclusion

- The **Age and Fare distributions** reveal a **young demographic with mostly low-cost tickets**.
- **SibSp and Parch distributions** show that **most passengers traveled alone or in small family groups**.
- The **right-skewed Fare distribution** confirms a **large economic gap**, with a few passengers paying very high fares.

This analysis helps in understanding **passenger demographics, travel groups, and economic class distribution** aboard the Titanic.

Scatterplot



This scatterplot visualizes the **relationship between Age and Fare**, with data points colored based on **survival status**. Below are the key insights:

1. Overall Distribution

- The x-axis represents **Age**, ranging from **0 to 80 years**.
- The y-axis represents **Fare**, ranging from **0 to 500**.
- The majority of fares are **below 100**, with more passengers **aged between 20-50 years**.

2. Survival-Based Coloring

- **Orange dots = Non-survivors** (Passengers who did not survive).
- **Blue dots = Survivors** (Passengers who survived).
- The color pattern helps in understanding **which age and fare groups had higher survival chances**.

3. Trends in Fare and Survival

- **Higher fare passengers (above 200) mostly survived**, indicating **first-class passengers had better survival chances**.
- **Lower fare passengers (below 50) show more non-survivors**, suggesting **third-class travelers faced more difficulties in evacuation**.

4. Age and Survival Patterns

- **Younger passengers (ages 0-10)** have more **survivors**, indicating **children were prioritized in evacuation**.
- **Older passengers (ages above 60) are fewer**, and survival rates are **mixed** among them.

5. Observing Outliers

- Some passengers **paid very high fares (above 300)**, and most of them **survived**, showing a **clear economic influence** on survival.
- A few **low-fare passengers survived**, possibly due to **early rescue or strategic location near lifeboats**.

Conclusion

- This scatterplot highlights that **Fare played a significant role in survival**—higher-paying passengers had better chances. **Age had a moderate influence**, with **younger children benefiting from priority rescue**. However, **not all high-fare passengers survived**, meaning other factors also affected survival chances.

Summary of my findings:

1. **Fare & Survival** → Higher fare passengers had better survival chances, indicating first-class priority in lifeboat access.
2. **Age & Survival** → Weak correlation—infants had a slight survival advantage, but age wasn't a major deciding factor.
3. **Family Size (SibSp & Parch) & Survival** → Small families had better chances, while larger groups faced difficulties.
4. **Heatmap Insights** → Fare had the strongest positive correlation with survival, reinforcing class-based rescue advantages.
5. **Scatterplot Insights** → Younger passengers and high-fare travelers had higher survival rates, but exceptions exist.
6. **Histogram Insights** → Most passengers were young (20-40 years), paid low fares, and traveled alone or in small groups.