Titanic Dataset - Exploratory Data Analysis (EDA) Report

DataSet

survived	pclass	sex	age	sibsp	parch	fare	embarked	deck	
0	0	3	male	22.0	1	0	7.2500	S	NaN
1	1	1	female	38.0	1	0	71.2833	C	C
2	1	3	female	26.0	0	0	7.9250	S	NaN
3	1	1	female	35.0	1	0	53.1000	S	C
4	0	3	male	35.0	0	0	8.0500	S	NaN
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S	NaN
887	1	1	female	19.0	0	0	30.0000	S	B
888	0	3	female	NaN	1	2	23.4500	S	NaN
889	1	1	male	26.0	0	0	30.0000	C	C
890	0	3	male	32.0	0	0	7.7500	Q	NaN

891 rows × 9 columns

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

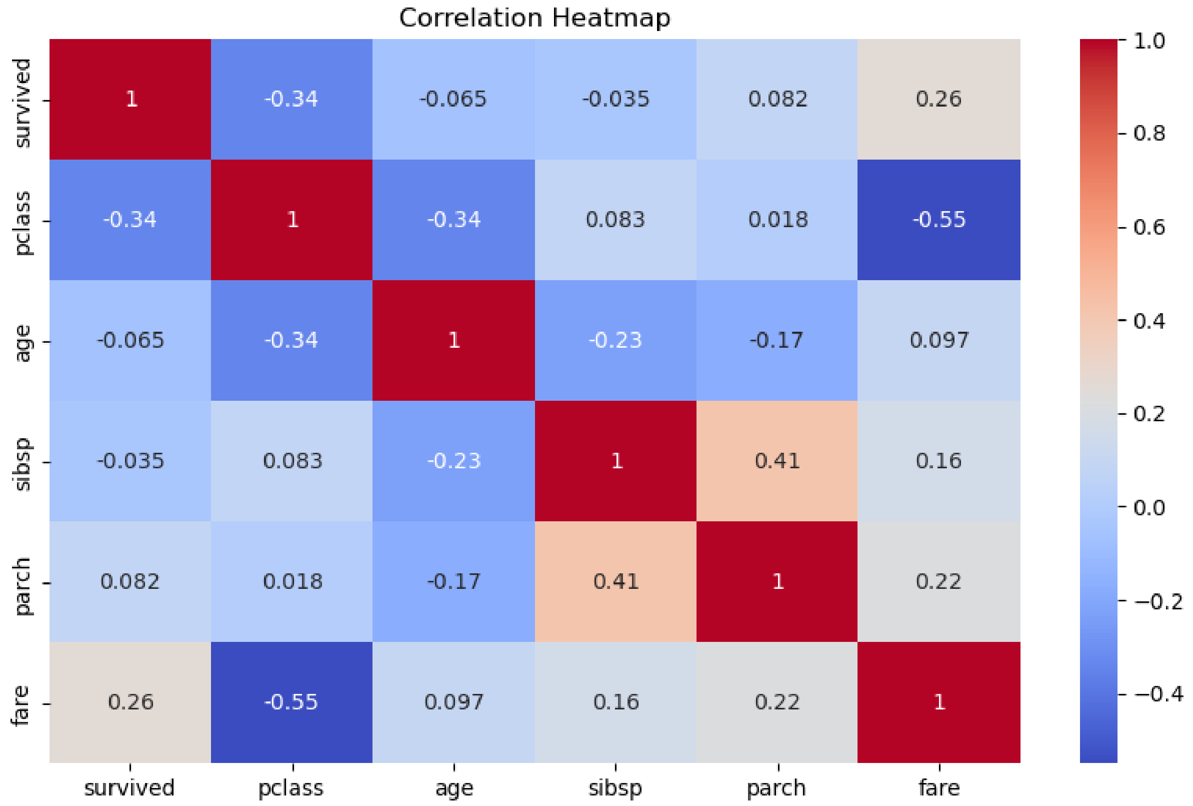
parch: The dataset defines family relations in this way...
Parent = mother, father = 1
Child = daughter, son, stepdaughter, stepson = 2
Some children travelled only with a nanny, therefore parch=0 for them.

Statistical summary

survived	pclass	age	sibsp	parch	fare	
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.361582	0.523008	0.381594	32.204208
std	0.486592	0.836071	13.019697	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	22.000000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

1. Survival Rate:
   - Only 38.4% of passengers survived.
   - The distribution is skewed toward non-survival (median = 0.0).
2. Passenger Class (Pclass):
   - Most passengers were in 3rd class (75% ≤ 3).
   - The average class value is 2.31, indicating a majority in lower classes.
3. Age:
   - The average age was 29.36 years, with most passengers between 22 and 35 years old (IQR).
   - The youngest was 0.42 years, and the oldest was 80 years.
   - A high standard deviation (13.02) suggests a wide spread in ages.
4. Siblings/Spouses Aboard (SibSp):
   - Median = 0, mean = 0.52 → Most people were traveling without siblings/spouses.
   - Maximum value of 8 indicates a few large families/groups.
5. Parents/Children Aboard (Parch):
   - Similar to SibSp, the average is low (0.38) and 75% had no parents/children aboard.
   - Maximum of 6 suggests some were traveling with large families.
6. Fare:
   - The average fare was ₹32.20, but with a high standard deviation (49.69) and a maximum fare of ₹512, indicating many outliers.

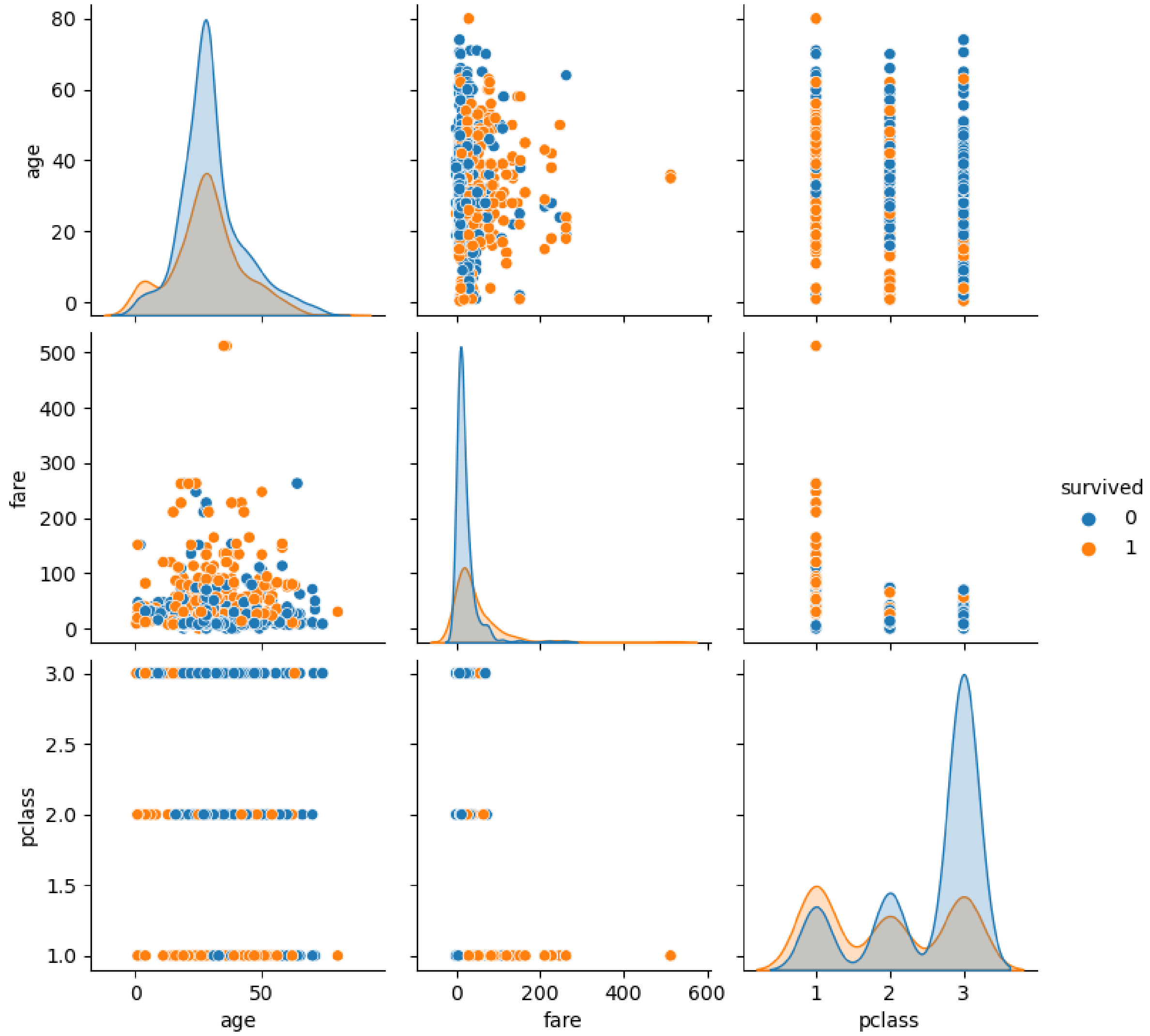
Correlation heatmap



Conclusion:

- **Survival is negatively correlated with pclass (-0.34)** — lower class (higher number) passengers had lower survival chances.
- **Fare** is **positively correlated** with **survival (0.26)** — passengers who paid more had better chances of survival.
- **SibSp and Parch** (family aboard) are moderately positively correlated (0.41) — expected since both relate to family.
- Other variables show weak or no correlation with survival.

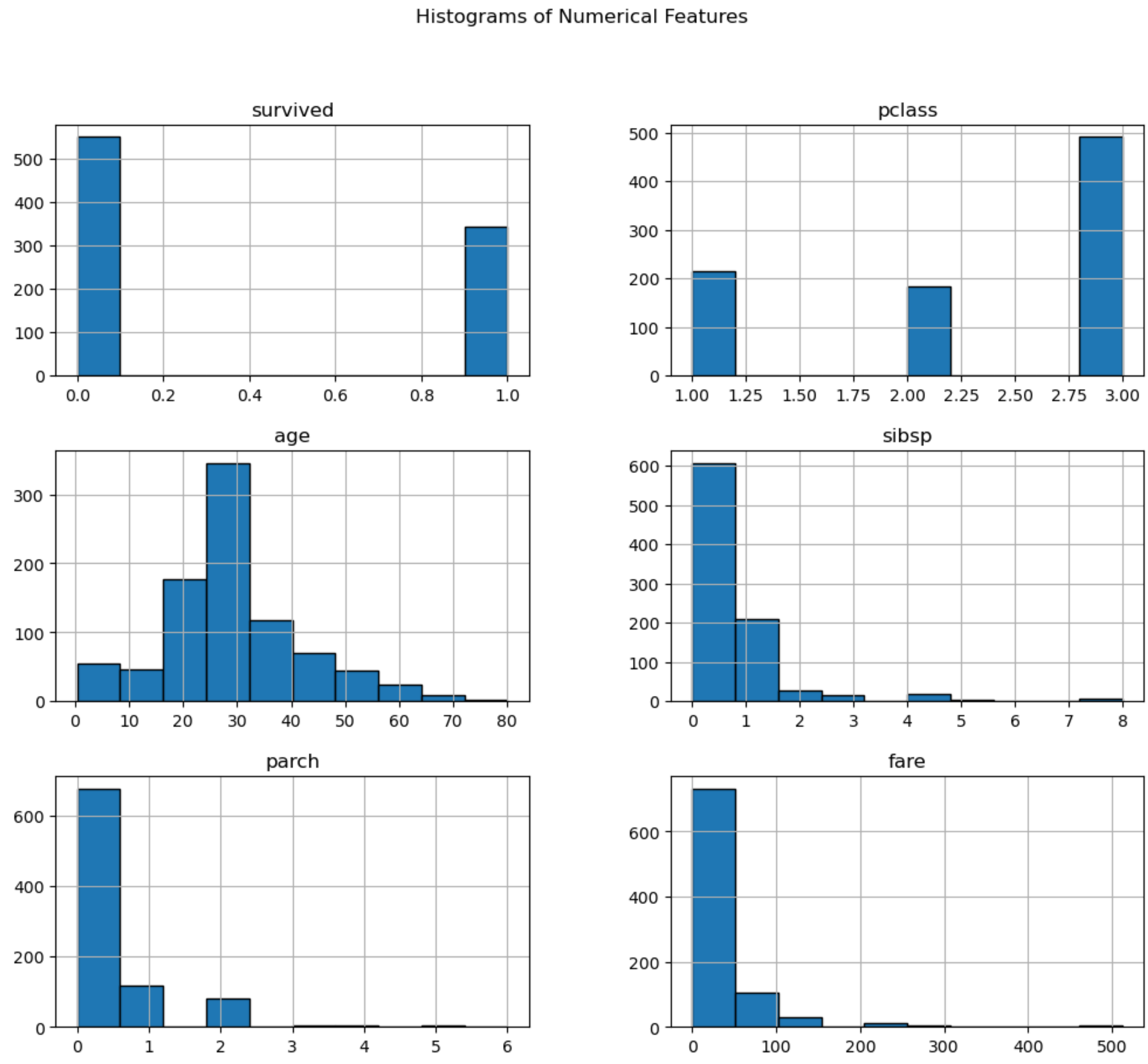
Pairplot



Conclusion:

- **Survivors (orange)** tend to cluster around **lower Pclass (1)** and **higher fare**.
- Survivors are slightly younger on average.
- There is high fare variance within 1st class, and survival is much higher in that class.

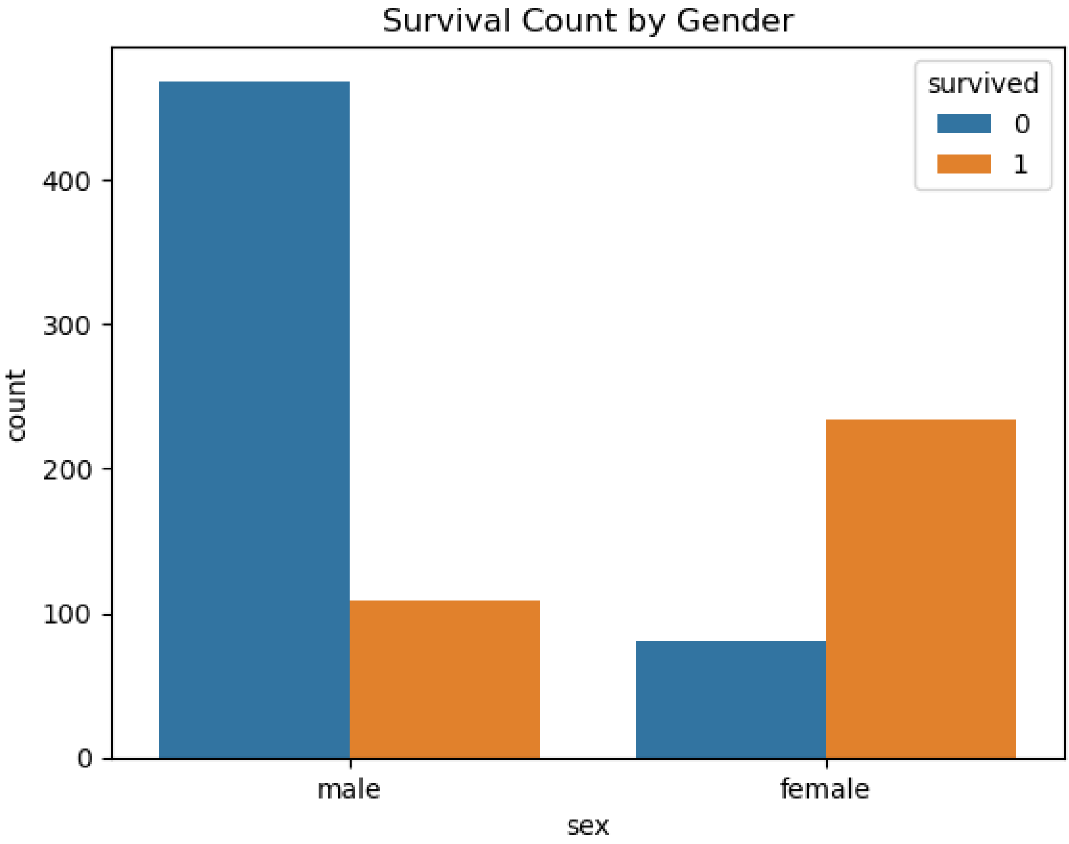
Histograms



Conclusion:

- Survived: Imbalanced – more people died (0) than survived (1).
- Pclass: Most passengers were in **3rd class**.
- Age: Right-skewed — most were in the **20–40 age range**, with fewer elderly passengers.
- SibSp & Parch: Most passengers had **0 siblings/spouse or parents/children** aboard.
- Fare: Highly **right-skewed** — majority paid below 100.

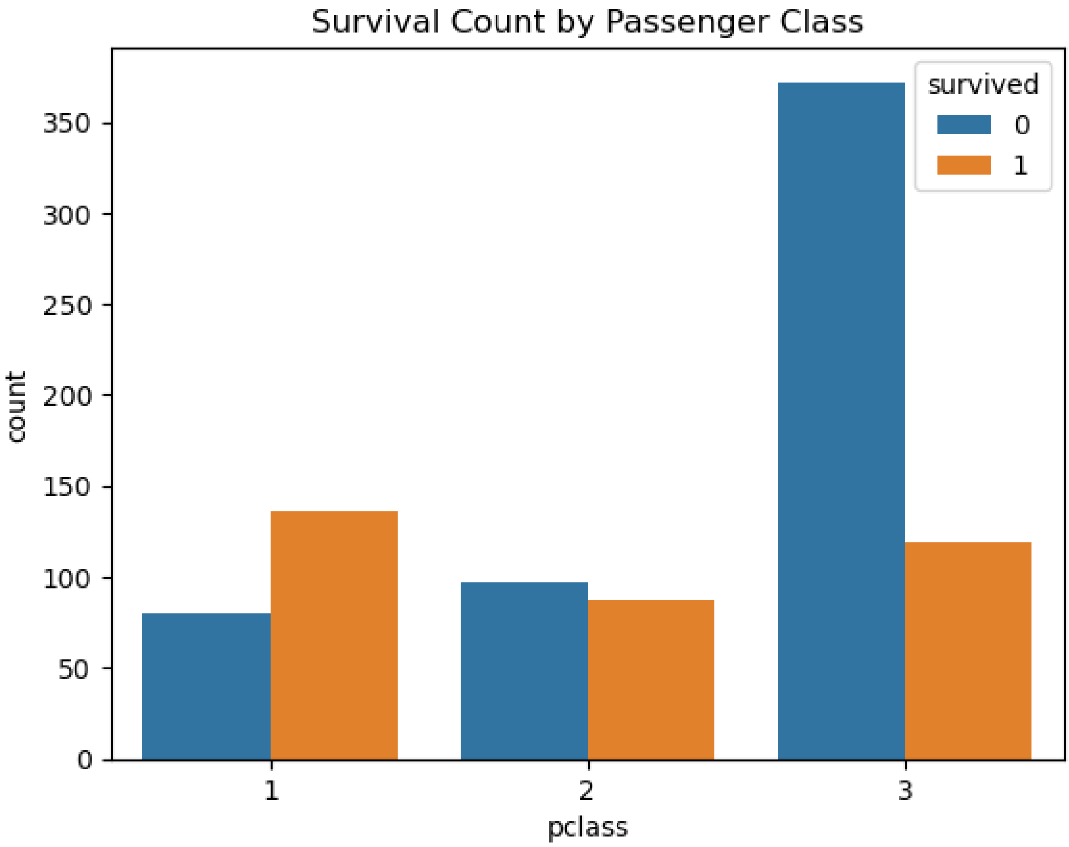
Bar plot of survival by sex



Conclusion:

- **Females had a significantly higher survival rate** than males.
- **Most males did not survive**, highlighting the “women and children first” policy during evacuation.

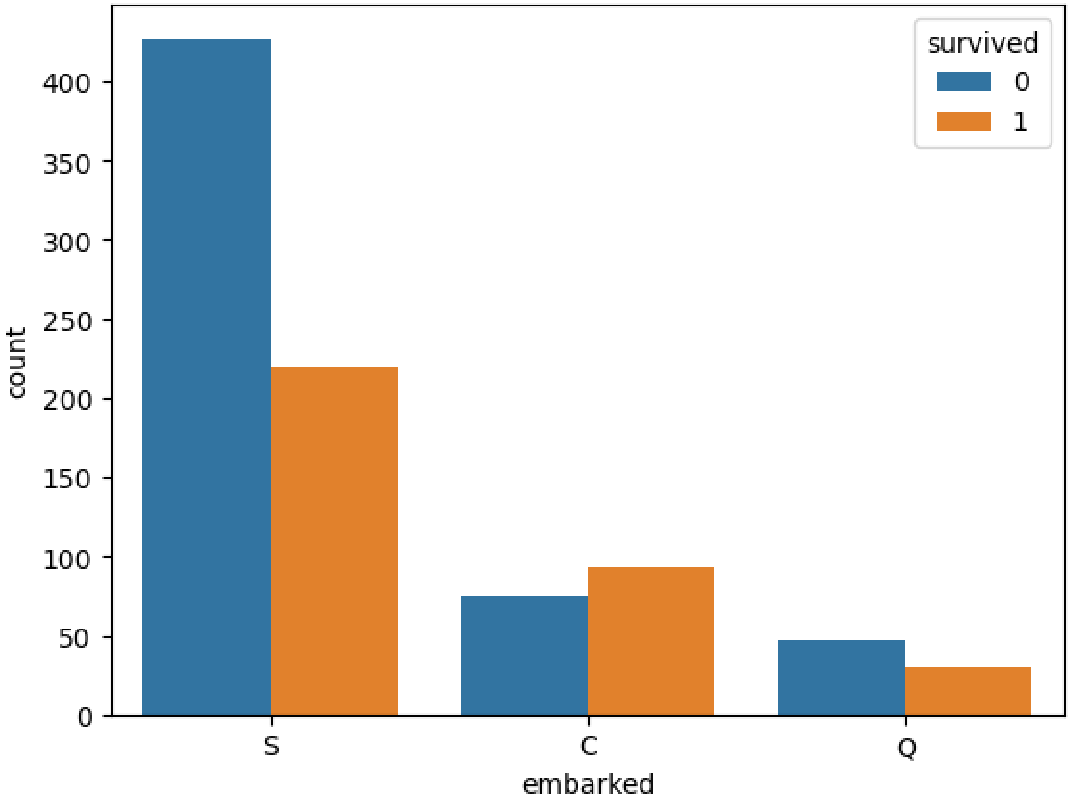
Bar plot of survival by Pclass



Conclusion:

- **1st class** had the **highest survival rate** — more passengers survived than died.
- **2nd class** had nearly equal survival and death counts.
- **3rd class** had **very low survival** — the majority of 3rd class passengers died.

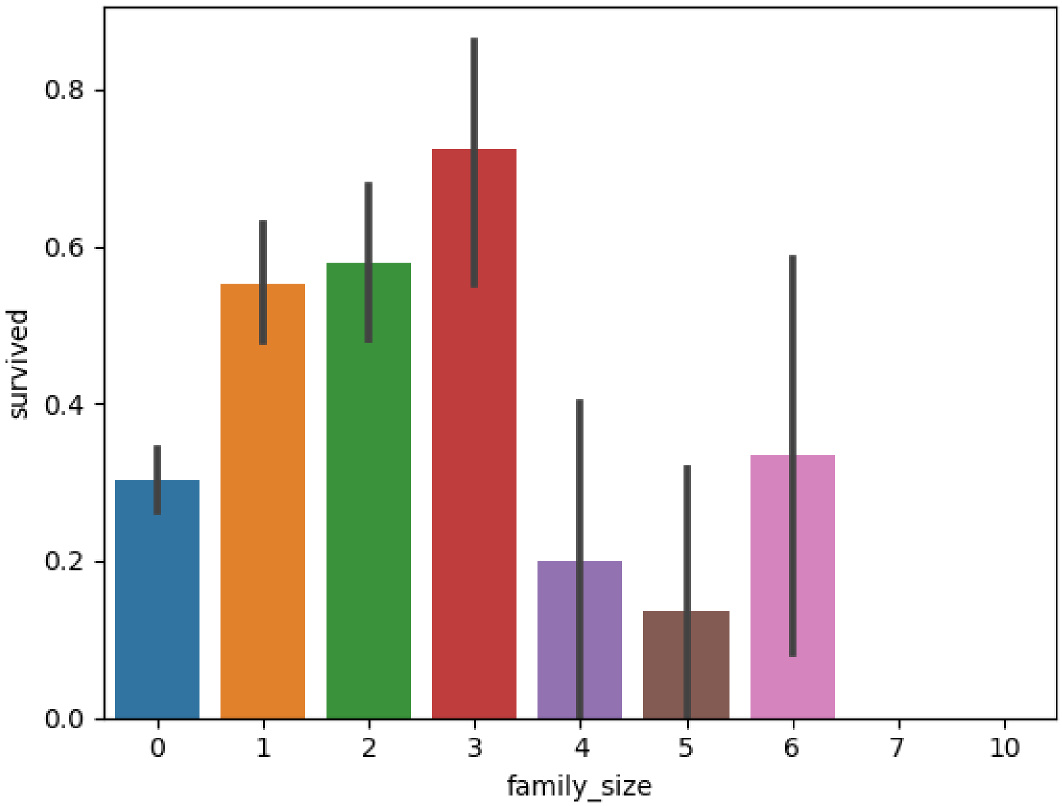
Survival Count by Embarked Port



Conclusion:

- Most passengers boarded at **Southampton (S)**, followed by **Cherbourg (C)** and **Queenstown (Q)**.
- **Highest survival rate** appears among those who embarked at **Cherbourg (C)**.
- Passengers from **Southampton (S)** had the **highest death count**, possibly because many 3rd class passengers boarded there.

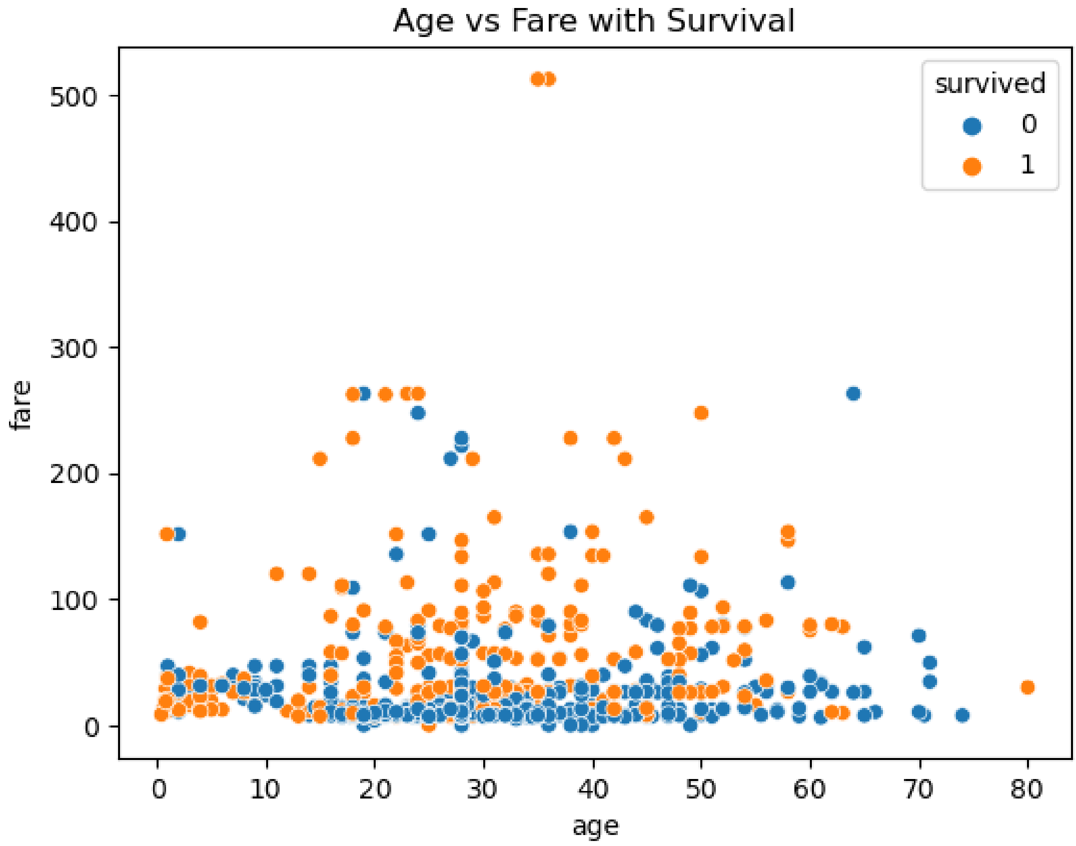
Survival Rate by Family Size



Conclusion:

- **Passengers with 1 to 3 family members aboard** had the **highest survival rates** (up to ~75%).
- **Solo travelers (family_size = 0)** had **lower survival**, around 30%.
- **Larger families (4 or more)** had significantly **lower survival**, possibly due to difficulty evacuating together.

Scatter plot: Age vs Fare



Conclusion:

- **Survivors (orange)** tend to be:
  - Spread across all age groups
  - More frequent among passengers who paid higher fares
- **Non-survivors (blue)** are more concentrated in the **low-fare range**
- There is **no strong linear pattern** between age and fare, but survival is clearly more likely in **higher fare brackets**