Task 5 - Exploratory Data Analysis (EDA)

Basic Analysis

📋 Dataset Column Explanation (Row-wise Sample)

Column	Meaning & Example
Passenger Id	Unique ID for each passenger. Example: 1 = First entry in the dataset.
Survived	Whether the passenger survived: 0 = No, 1 = Yes.Row 1: Did not survive. Row 2: Survived.
Pclass	Passenger class (proxy for socio-economic status): 1 = 1st (upper), 2 = 2nd (middle), 3 = 3rd (lower). Most in this sample are 3.
Name	Full name including title. Example: "Braund, Mr. Owen Harris".Can be used to extract titles like Mr., Mrs., Miss, which help estimate gender or social status.
Sex	Gender of the passenger: male / female. Useful for survival analysis (females had higher survival rate).
Age	Age in years. Some values are missing (NaN). Row 1: 22.0, Row 2: 38.0, Row 3: 26.0. Helps understand age distribution and survival of children.
SibSp	Number of siblings/spouses aboard. Row 1: 1 → had 1 sibling or spouse with them.0 means alone.

Parch Number of parents/children aboard. Row 1: 0 → no parent/child with them.This and SibSp can be combined to find family size.

Ticket Ticket number. Can be alphanumeric. Example: A/5 21171.May reveal booking patterns but not directly useful unless grouped.

Fare Price paid for the ticket (in £). Row 1: £7.25, Row 2: £71.28.Strongly correlates with Pclass.

Cabin Cabin number (many missing - NaN). Row 2: C85. High missing rate but sometimes reveals deck.

Embarked Port of embarkation: C = Cherbourg, Q = Queenstown, S = Southampton. Most people boarded at S.

🔍 Observations from the First 5 Rows:

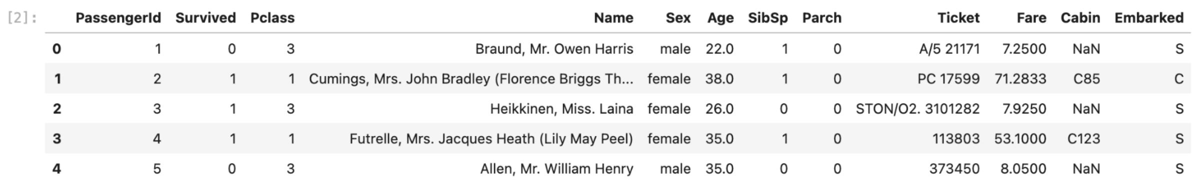
- Row 1 (Mr. Braund): Male, 22 years old, 3rd class, didn’t survive, paid very little, no cabin listed, boarded from Southampton.

- Row 2 (Mrs. Cumings): Female, 38, 1st class, survived, paid a high fare, had a private cabin (C85), embarked at Cherbourg.

- Row 3 (Miss Heikkinen): Young female, 26, 3rd class, survived, no cabin assigned, likely traveled alone.

- Row 4 (Mrs. Futrelle): Female, 35, 1st class, survived, has a cabin, had a spouse on board.

- Row 5 (Mr. Allen): Male, 35, 3rd class, didn’t survive, paid a low fare.



	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

Dataset Structure and Quality Summary

Dataset Overview

- Total Entries (Rows): 891 passengers
- Total Features (Columns): 12
- DataFrame Type: pandas.core.frame.DataFrame
- Memory Usage: ~83.7 KB

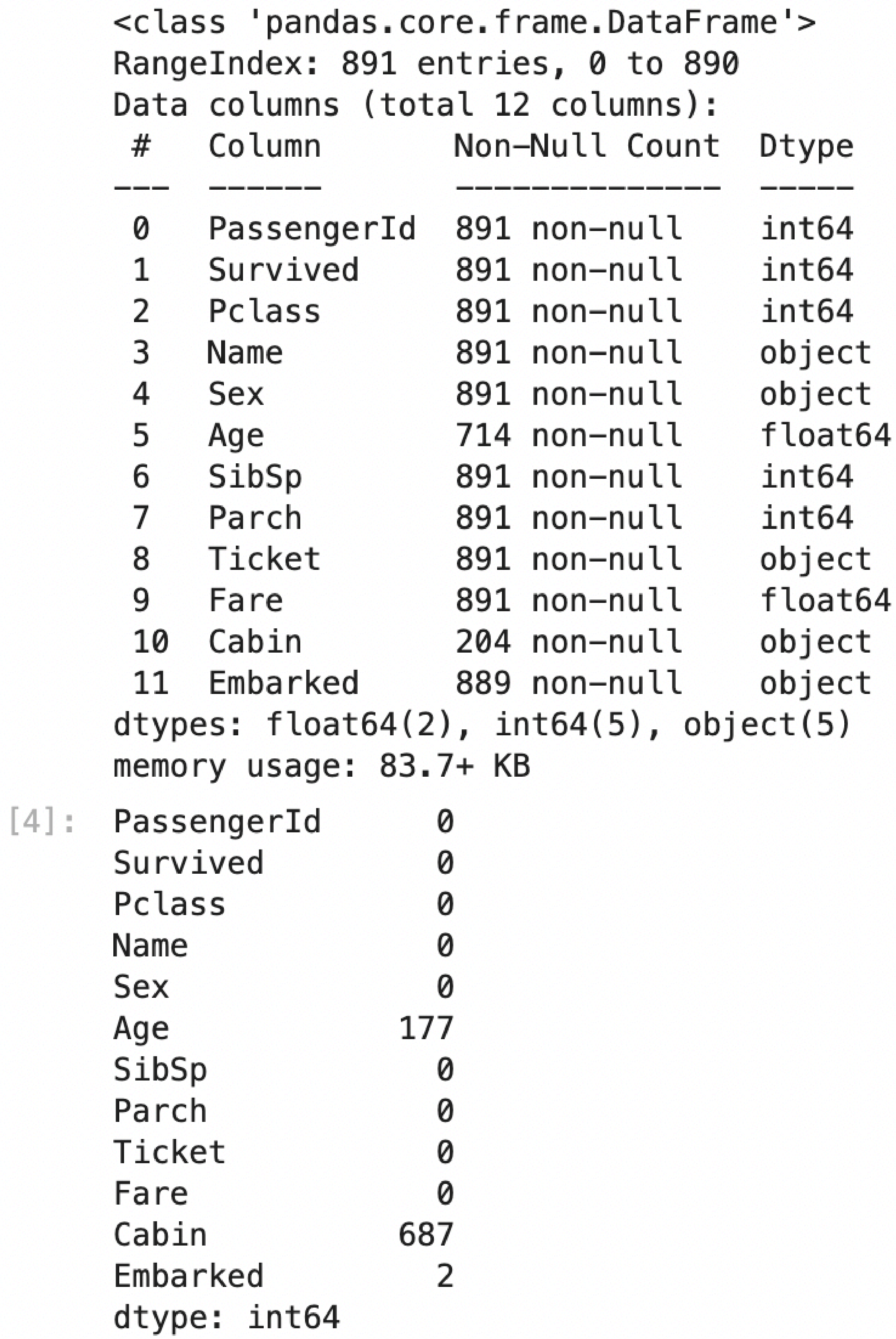
Column-wise Summary

Column Name	Non-Null Count	Data Type	Description
PassengerId	891 / 891	int64	Unique ID assigned to each passenger
Survived	891 / 891	int64	Survival status (0 = No, 1 = Yes)
Pclass	891 / 891	int64	Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	891 / 891	object	Full name, including title
Sex	891 / 891	object	Gender (male/female)

Age	714 / 891	float64	Age in years (177 missing)
SibSp	891 / 891	int64	Number of siblings/spouses aboard
Parch	891 / 891	int64	Number of parents/children aboard
Ticket	891 / 891	object	Ticket number
Fare	891 / 891	float64	Passenger fare
Cabin	204 / 891	object	Cabin number (687 missing)
Embarked	889 / 891	object	Port of embarkation (2 missing)

Missing Data Summary

Column	Missing Values	% Missing	Remarks
Age	177	19.9%	Impute with median or by group
Cabin	687	77.1%	Consider dropping or extracting deck letter
Embarked	2	0.2%	Fill with most frequent value (S)

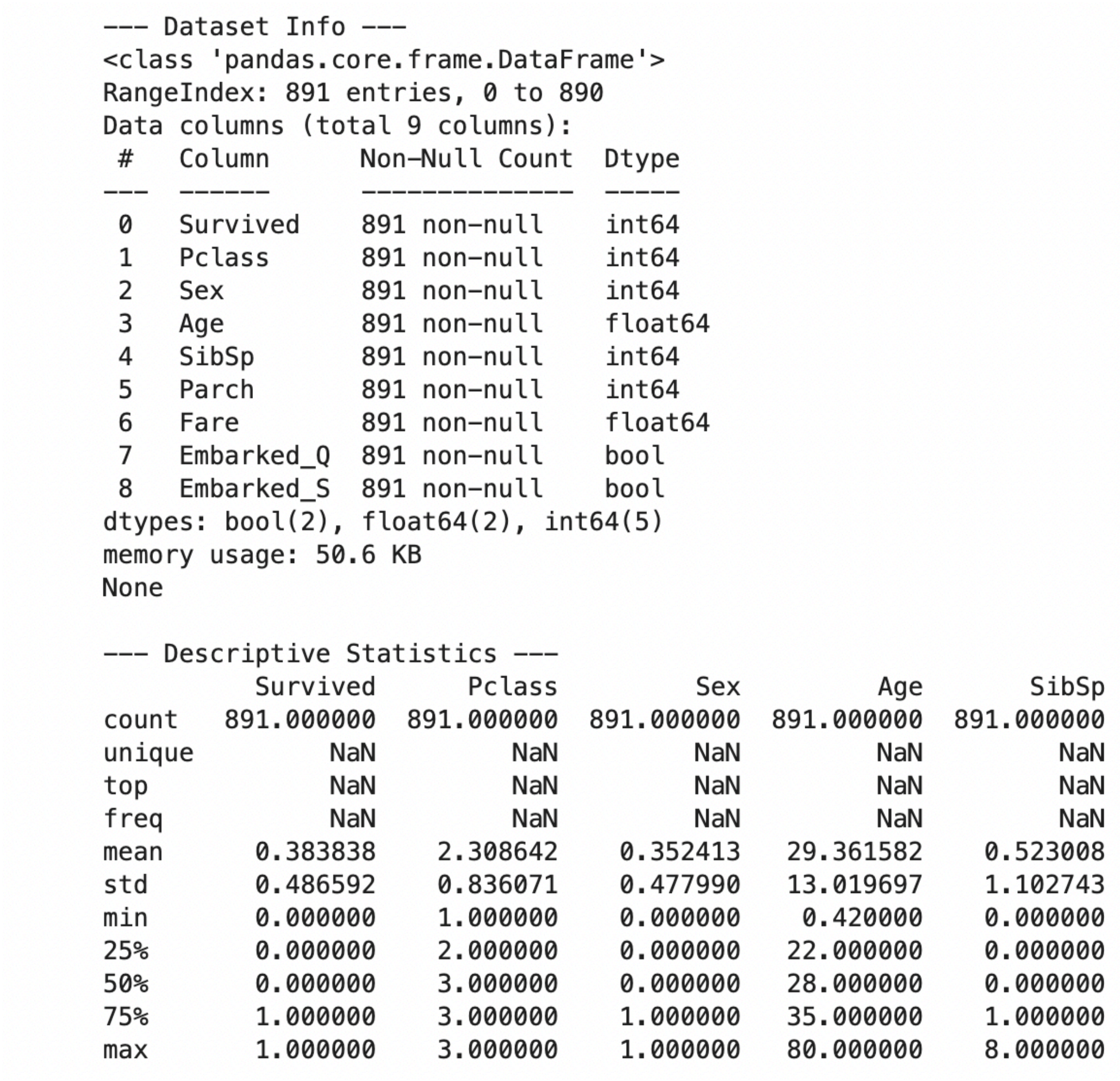


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

```
[4]:  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
```

Insights & Interpretation

- Survival Rate: Only 38.4% of passengers survived.
- Pclass: Majority passengers were in 3rd class (median = 3).
- Sex Encoding: ~35% are female (Sex = 0), and ~65% are male (Sex = 1).
- Age Distribution:
  - Average age is 29.36 years, range spans infants (0.42) to elderly (80).
  - 50% of passengers were between 22 and 35 years old.
- Fare Spread:
  - Skewed distribution; many paid below ₹30, but some outliers went up to ₹512.
- Embarkation:
  - Embarked_S (Southampton): 646 True → 72.5% embarked from S.
  - Embarked_Q (Queenstown): 77 True → 8.6%.
  - Remaining (~19%) embarked from Cherbourg (Embarked_C).



```
--- Dataset Info ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Survived    891 non-null    int64  
 1   Pclass      891 non-null    int64  
 2   Sex         891 non-null    int64  
 3   Age         891 non-null    float64
 4   SibSp       891 non-null    int64  
 5   Parch       891 non-null    int64  
 6   Fare        891 non-null    float64
 7   Embarked_Q  891 non-null    bool   
 8   Embarked_S  891 non-null    bool   
dtypes: bool(2), float64(2), int64(5)
memory usage: 50.6 KB
None

--- Descriptive Statistics ---
          Survived      Pclass         Sex         Age       SibSp
count   891.000000  891.000000  891.000000  891.000000  891.000000
unique         NaN         NaN         NaN         NaN         NaN
top            NaN         NaN         NaN         NaN         NaN
freq           NaN         NaN         NaN         NaN         NaN
mean      0.383838    2.308642    0.352413   29.361582    0.523008
std       0.486592    0.836071    0.477990   13.019697    1.102743
min       0.000000    1.000000    0.000000    0.420000    0.000000
25%       0.000000    2.000000    0.000000   22.000000    0.000000
50%       0.000000    3.000000    0.000000   28.000000    0.000000
75%       1.000000    3.000000    1.000000   35.000000    1.000000
max       1.000000    3.000000    1.000000   80.000000    8.000000
```

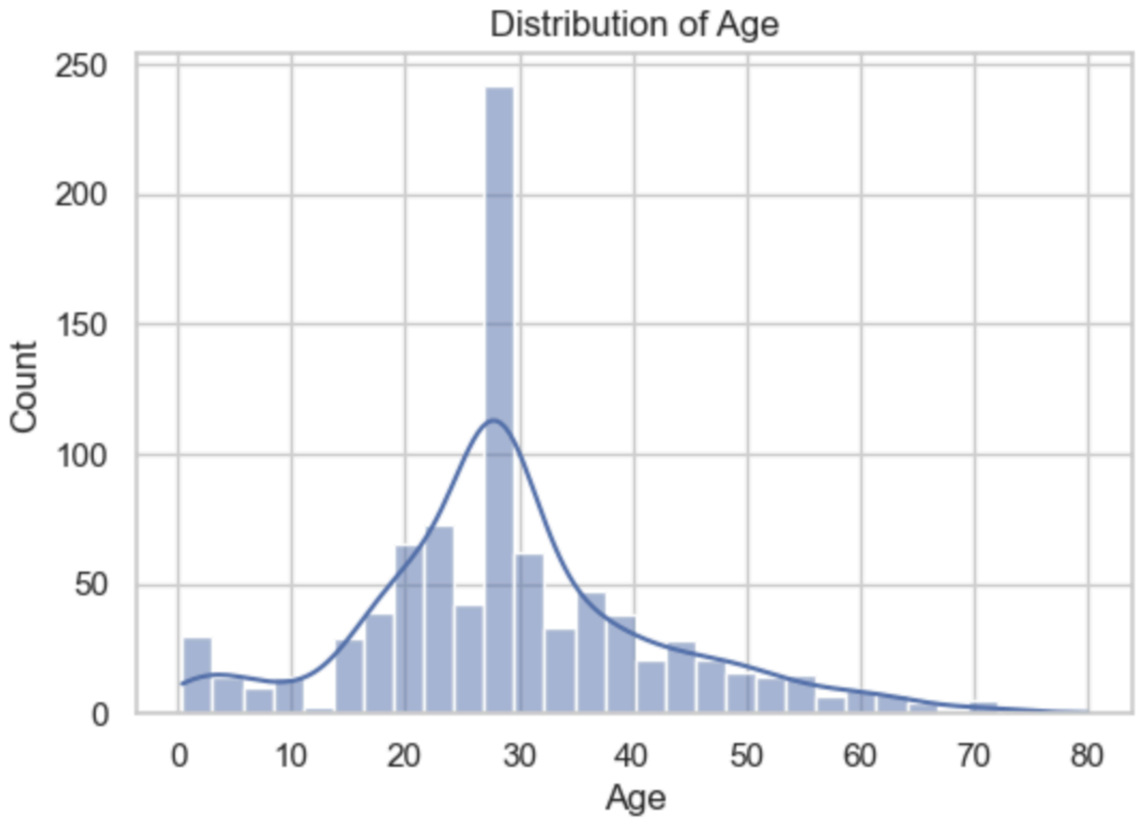
```
           Parch        Fare Embarked_Q Embarked_S
count   891.000000  891.000000        891        891
unique         NaN         NaN          2          2
top            NaN         NaN      False       True
freq           NaN         NaN        814        646
mean      0.381594   32.204208        NaN        NaN
std       0.806057   49.693429        NaN        NaN
min       0.000000    0.000000        NaN        NaN
25%       0.000000    7.910400        NaN        NaN
50%       0.000000   14.454200        NaN        NaN
75%       0.000000   31.000000        NaN        NaN
max       6.000000  512.329200        NaN        NaN

--- Value Counts ---
```

[6]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked_Q	Embarked_S
0	0	3	0	22.0	1	0	7.2500	False	True
1	1	1	1	38.0	1	0	71.2833	False	False
2	1	3	1	26.0	0	0	7.9250	False	True
3	1	1	1	35.0	1	0	53.1000	False	True
4	0	3	0	35.0	0	0	8.0500	False	True

Data Visualisation



Age Distribution Summary

The histogram and KDE (Kernel Density Estimate) above display the distribution of passengers' ages on the Titanic.

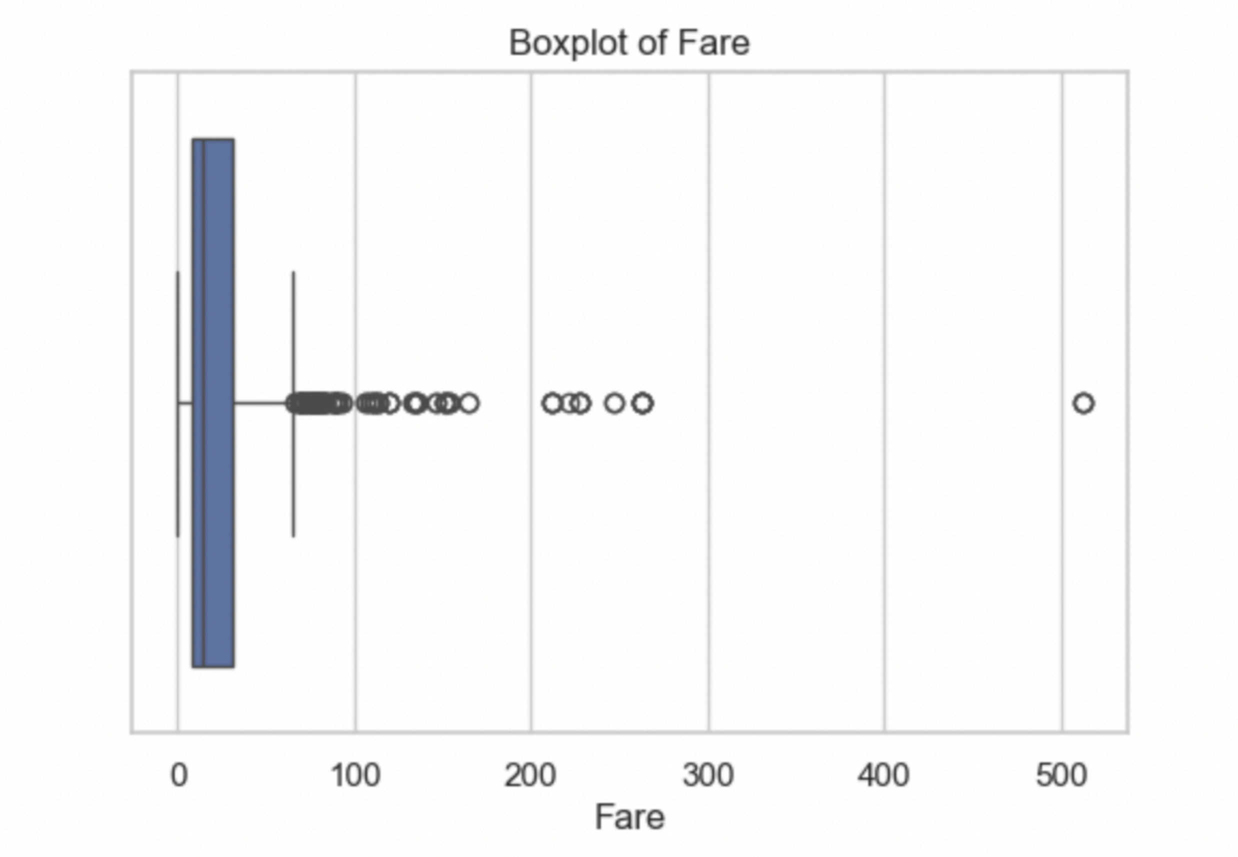
Key Observations:

- Most Common Age Group:
  A significant number of passengers were around 28–30 years old, as seen from the sharp peak near age 29. This spike is likely due to imputation, where missing age values were filled with the median age.

- General Shape:
  The distribution is right-skewed, meaning there are more younger

passengers than older ones, but a long tail extends toward the elderly (up to age 80).

- Passenger Age Spread:
  - Infants as young as 0.42 years were onboard.
  - The oldest passenger was 80 years old.
  - Majority of passengers were between 20 and 40 years.
- Bimodal Hint:
  While the main peak is around 29, there's also a smaller hump in the child age group (0–10), suggesting a moderate presence of families with young children.



Fare Distribution Summary

The boxplot above visualizes the distribution of passenger fares paid on the Titanic.

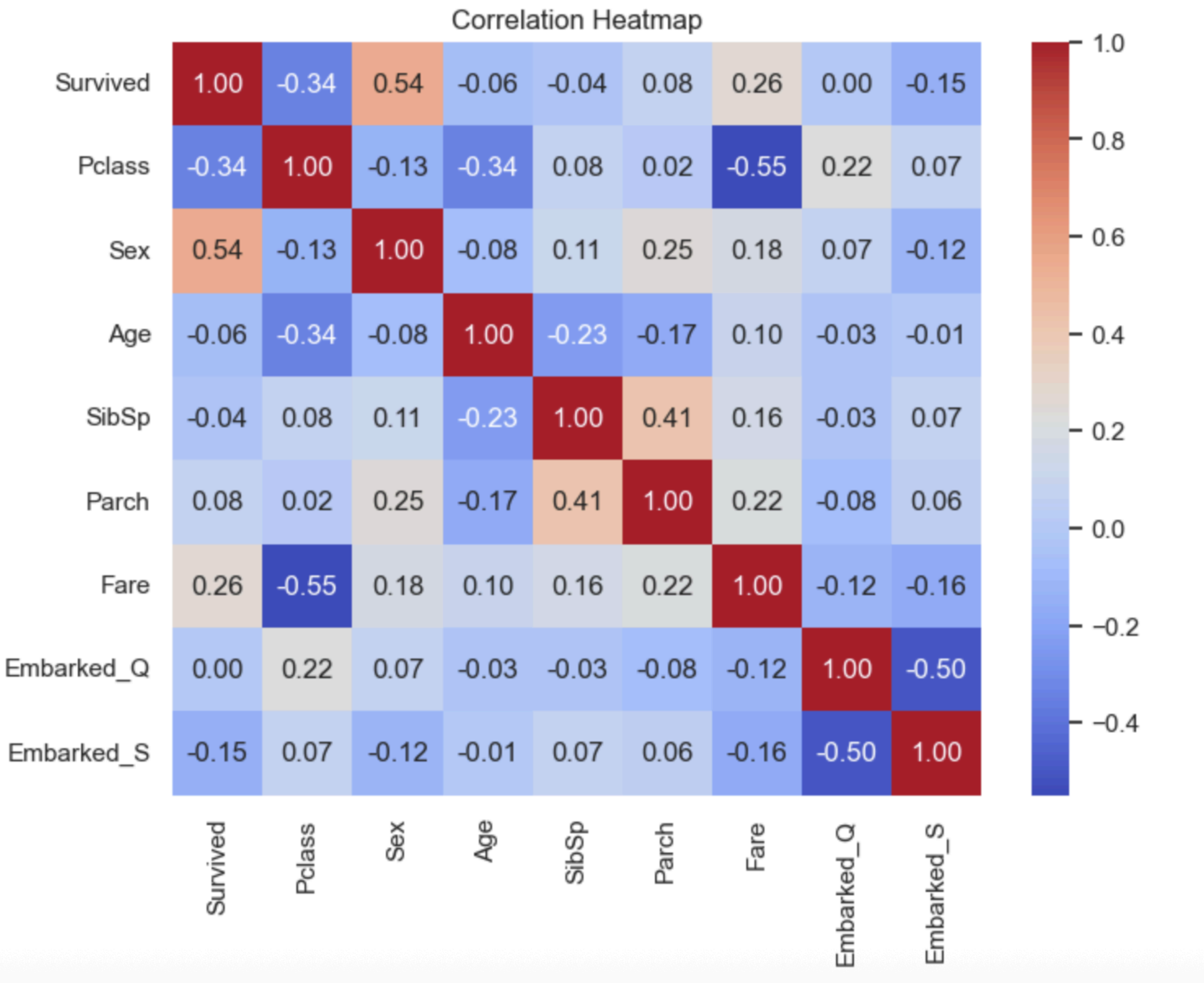
Key Observations:

- Right-Skewed Distribution:
  The majority of fares are concentrated on the lower end, indicating that most passengers paid under $100.

- Median Fare:
  The black line inside the box shows the median fare, which lies well below 50, confirming low average ticket prices.

- Outliers:
 A large number of outliers exist beyond the upper whisker, with some fares exceeding $500. These outliers are likely first-class passengers or those traveling in luxury cabins.

- Interquartile Range (IQR):
 The middle 50% of fare values (between Q1 and Q3) are tightly packed, suggesting low fare variability for most passengers.



Correlation Heatmap

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked_Q	Embarked_S
Survived	1.00	-0.34	0.54	-0.06	-0.04	0.08	0.26	0.00	-0.15
Pclass	-0.34	1.00	-0.13	-0.34	0.08	0.02	-0.55	0.22	0.07
Sex	0.54	-0.13	1.00	-0.08	0.11	0.25	0.18	0.07	-0.12
Age	-0.06	-0.34	-0.08	1.00	-0.23	-0.17	0.10	-0.03	-0.01
SibSp	-0.04	0.08	0.11	-0.23	1.00	0.41	0.16	-0.03	0.07
Parch	0.08	0.02	0.25	-0.17	0.41	1.00	0.22	-0.08	0.06
Fare	0.26	-0.55	0.18	0.10	0.16	0.22	1.00	-0.12	-0.16
Embarked_Q	0.00	0.22	0.07	-0.03	-0.03	-0.08	-0.12	1.00	-0.50
Embarked_S	-0.15	0.07	-0.12	-0.01	0.07	0.06	-0.16	-0.50	1.00

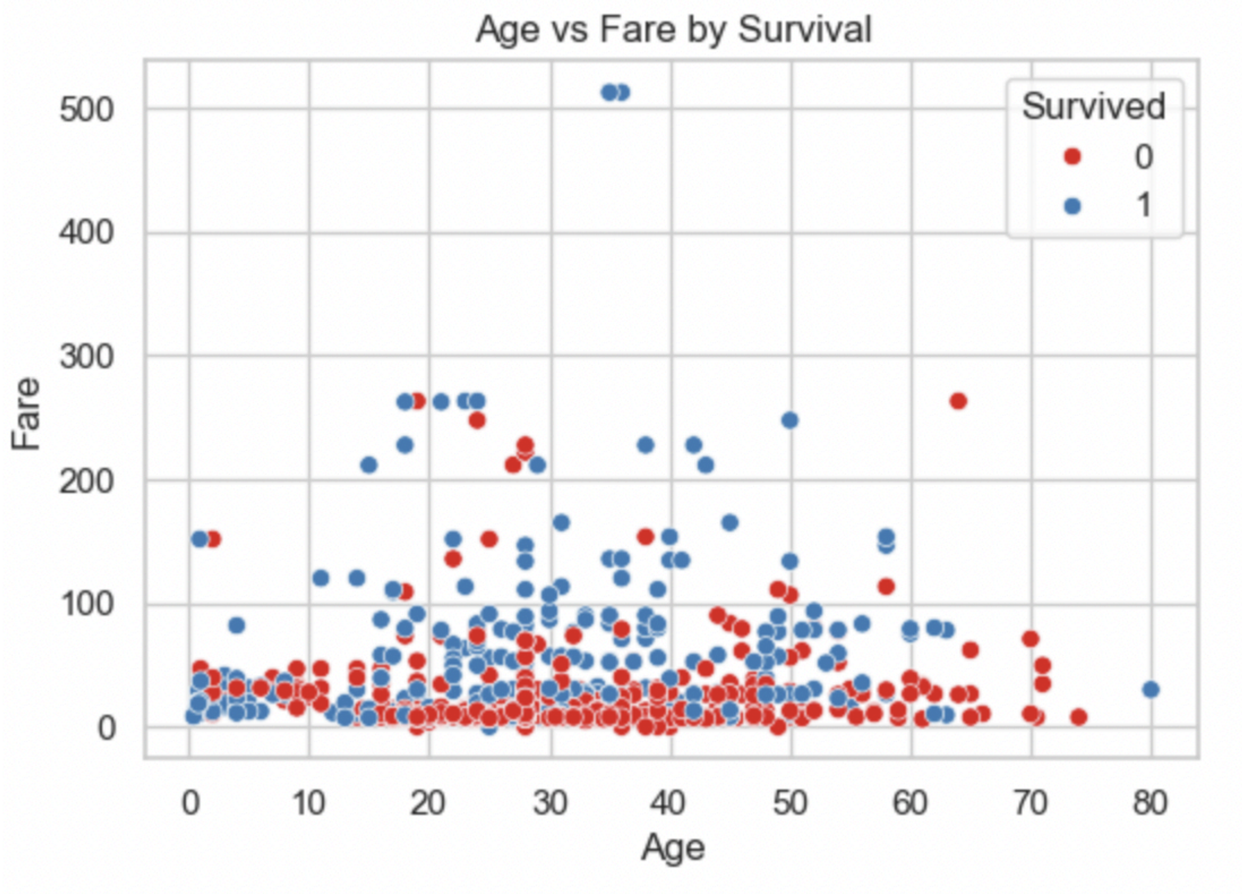
Correlation Heatmap Summary

This heatmap visualizes Pearson correlation coefficients between different features in the Titanic dataset. Values range from -1 (perfect negative) to +1 (perfect positive) correlation.

Key Insights:

- Survival Influences:
  - Sex has the strongest positive correlation with Survived (0.54) → Females had higher survival rates.
  - Pclass has a negative correlation with Survived (-0.34) → Lower-class passengers were less likely to survive.
  - Fare shows a mild positive correlation with survival (0.26) → Higher-paying passengers had better chances.
- Strong Feature Relationships:
  - SibSp and Parch are positively correlated (0.41) → Larger families onboard.
  - Fare and Pclass have a strong negative correlation (-0.55) → Higher class means higher fare.
  - Embarked_Q and Embarked_S are strongly negatively correlated (-0.50) due to one-hot encoding (mutually exclusive categories).
- Weak/No Correlation:

- ○ Age has almost no correlation with Survived (-0.06), indicating age alone was not a major survival factor.
- ○ Embarkation points show little direct impact on survival.



Key Observations:

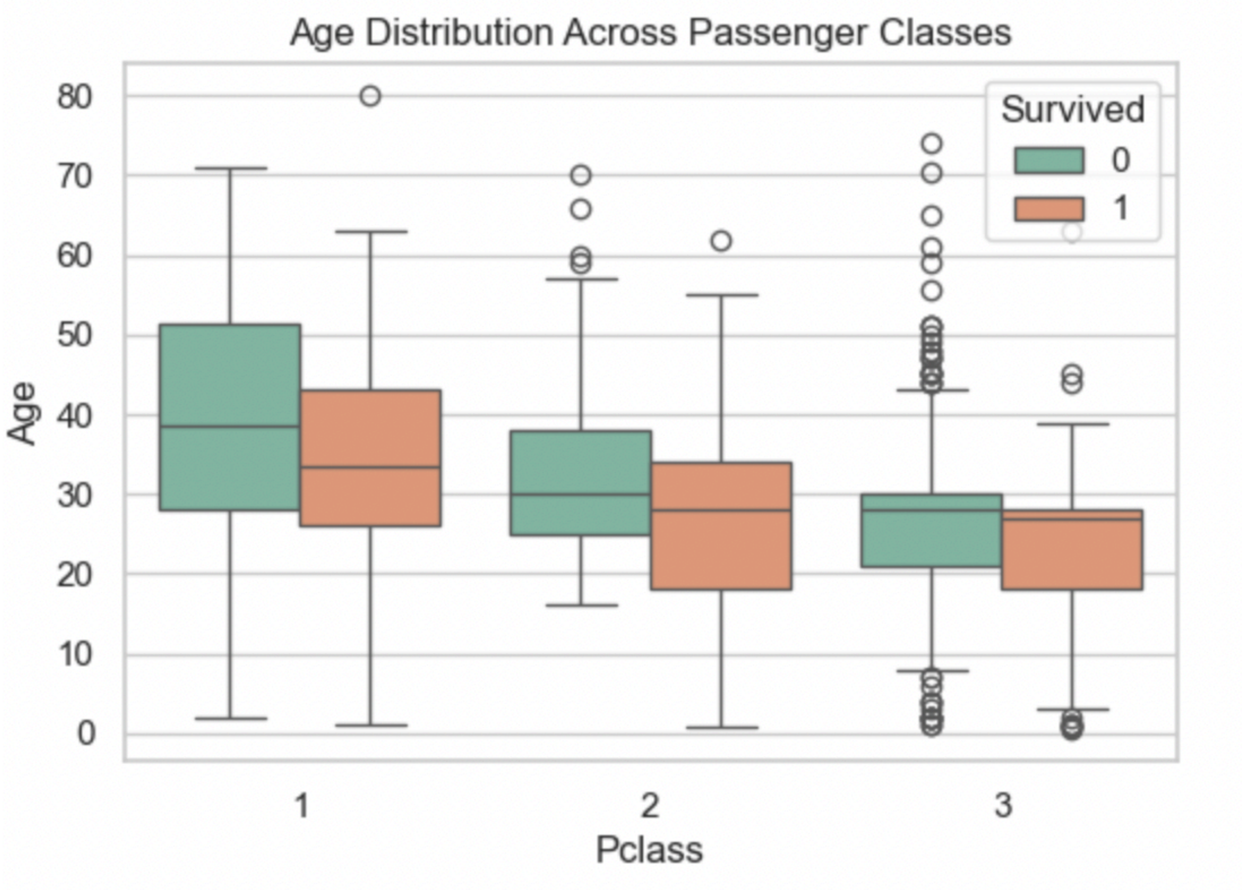
- Survivors (blue) generally appear across a wider fare range, including many who paid high fares (>100), especially between ages 20–50.

- Non-survivors (red) are more densely clustered in the lower fare range (0–50), suggesting many non-survivors paid lower fares.

- High-fare passengers (especially those paying >200) were more likely to survive, indicating that higher class passengers may have had better survival chances.

- Age distribution is fairly spread for both groups, but survival appears somewhat more frequent among younger adults and children, especially those who paid more.

- There is no strong linear correlation between age and fare, but fare seems more predictive of survival than age.



Titanic passenger classes (Pclass) broken down by survival status:

- Green (0) = Did not survive
- Orange (1) = Survived

Key Insights:

Pclass 1 (First Class):

- Generally older passengers, with median age around 38–40 for both survivors and non-survivors.
- Survivors have slightly younger median ages than non-survivors.
- Wide age range, with several passengers over 60.
- More even distribution between survivors and non-survivors.

Pclass 2 (Second Class):

- Median ages are slightly lower than first class (around 30).
- Survivors tend to be younger than non-survivors.
- Fewer outliers compared to other classes.

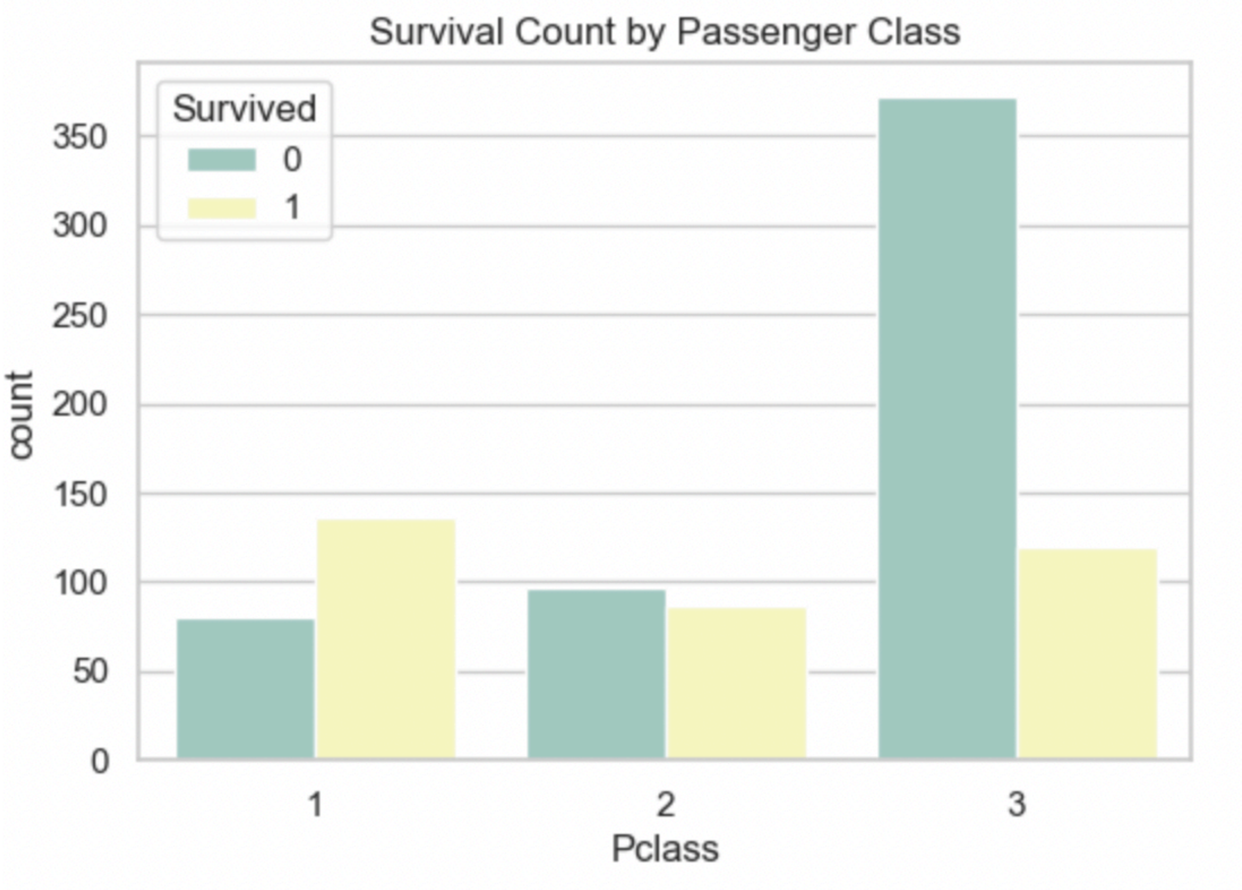
Pclass 3 (Third Class):

- Youngest group overall; median age around 20–25.

- Survivors are notably younger, suggesting younger third-class passengers had better survival chances.
- Many outliers (especially among survivors), indicating some older individuals did survive despite being in the lowest class.

Overall Observations:

- Survivors were generally younger, especially in 2nd and 3rd classes.
- First-class passengers were older on average, and had a higher chance of survival across all age groups.
- The age-survival relationship becomes more evident in lower classes, where survival seems to favor younger individuals.



Key Observations:

First Class (Pclass 1):

- More survivors than non-survivors, indicating the highest survival rate among all classes.

- Suggests first-class passengers were prioritized or had better access to lifeboats.

Second Class (Pclass 2):

- Survival and non-survival counts are almost equal, suggesting a moderate survival rate.

- Neither group significantly dominates.

Third Class (Pclass 3):

- Dramatically more non-survivors than survivors.
- Largest total number of passengers, but the lowest survival rate, highlighting class-based survival disparity.

Overall Insight:

There is a clear correlation between class and survival:

- Higher-class passengers (especially 1st class) had a significantly better chance of survival.
- 3rd class passengers had the highest number of fatalities, reinforcing the historical accounts of unequal access to lifeboats and safety during the Titanic disaster.

Final Summary

Top 3 Business Insights

1. Passenger Class Strongly Influences Survival
   Survival rates were highest in 1st class and lowest in 3rd class, confirming that socio-economic status played a crucial role during evacuation. First-class passengers likely had better cabin locations and earlier access to lifeboats, while 3rd class passengers faced the greatest

risk.

2. Gender is a Strong Predictor of Survival
   Female passengers had a much higher survival rate (correlation = 0.54). This reinforces the historical “women and children first” evacuation protocol. Gender should be a key feature in any predictive survival model.

3. Higher Fare = Higher Survival Odds
   Passengers who paid more (especially > ₹100) were far more likely to survive. This reflects a strong association between ticket price, class, and survival, suggesting that privilege and access heavily influenced life-or-death outcomes.

Underperforming / High-Risk Segments

- Third-Class Passengers (Pclass = 3):
  This group had the largest number of fatalities. Despite making up the majority of passengers, they were disproportionately underrepresented among survivors.

- Male Passengers:
  The majority of non-survivors were male, across all classes. Even among 1st class passengers, females had a clear survival advantage.

- Low Fare / Large Families:
  Passengers who paid low fares, often traveling with many family members, had lower survival rates. They were concentrated in 3rd class and likely had cabins located deeper in the ship.

Data Quality Insights

- Missing Age Data (19.9%):
  May bias age-related insights. Median imputation appears to have artificially increased the frequency around age 29.

- Cabin Information (77% Missing):
  Limits ability to analyze the impact of deck location on survival. If available, deck could provide strong spatial insights.

- Embarked Port:
  Majority embarked from Southampton (72.5%), but port of embarkation had little effect on survival.

Visual Summary Takeaways

- Age Distribution:
  Right-skewed; most passengers were young adults (20–40). Survivors in 3rd class skewed younger, suggesting youth aided survival in lower classes.

- Fare Distribution:
  Highly right-skewed. Outliers paying ₹300+ were typically survivors from 1st class.

- Correlation Heatmap:

  - Sex and Survived: Strongest positive correlation

    - Pclass and Fare: Strong negative correlation, reinforcing class-price stratification

    - Age has minimal direct correlation with survival

- Survival by Class:

    - 1st Class: Most likely to survive

    - 2nd Class: Survival ~50%

    - 3rd Class: Majority perished

Recommendations

1. Prioritize Socio-Demographic Factors in Predictive Models
   Focus on Sex, Pclass, and Fare as primary predictors of survival in any machine learning or logistic regression model.

2. Segmented Analysis for Deeper Insights
   Analyze survival within subgroups (e.g., females in 3rd class) to uncover nuanced trends.

3. Consider Feature Engineering
   Create derived features like:

    - FamilySize = SibSp + Parch + 1

    - Title from Name (e.g., Mr., Mrs., Miss)

4. Impute and Clean Data Cautiously
   Use group-wise imputation for Age (e.g., by Sex and Pclass) and drop or categorize cabin values by deck if possible.