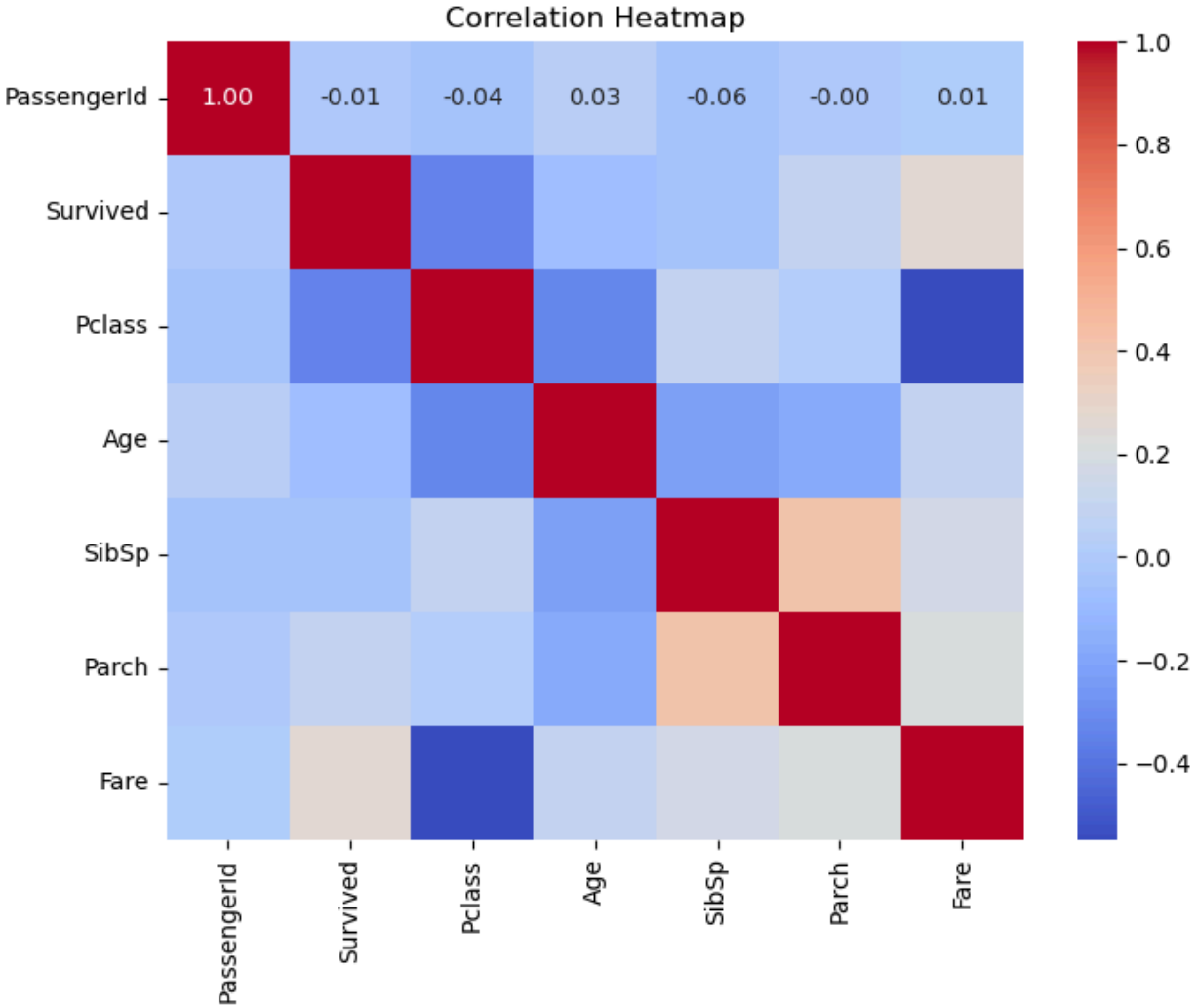
In [27]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sn

df = pd.read_csv("C:/Users/HP/Documents/DATA SCIENCE/Data analyst internship/t
print(df)

## a.basic checks
print(df.info()) # structure, data types, nulls
print(df.describe())
print(df.dtypes)
print('_____________________________________')

df.fillna(df.mean(numeric_only=True), inplace=True)

for col in df.select_dtypes(include=['object','category']).columns:
        df[col].fillna(df[col].mode()[0], inplace=True)        ## mode()[0]- m
print(df.isnull().sum())   ## to check the result
print('_____________________________________')

# Value counts for categorical features
print(df['Sex'].value_counts())
print(df['Pclass'].value_counts())
```

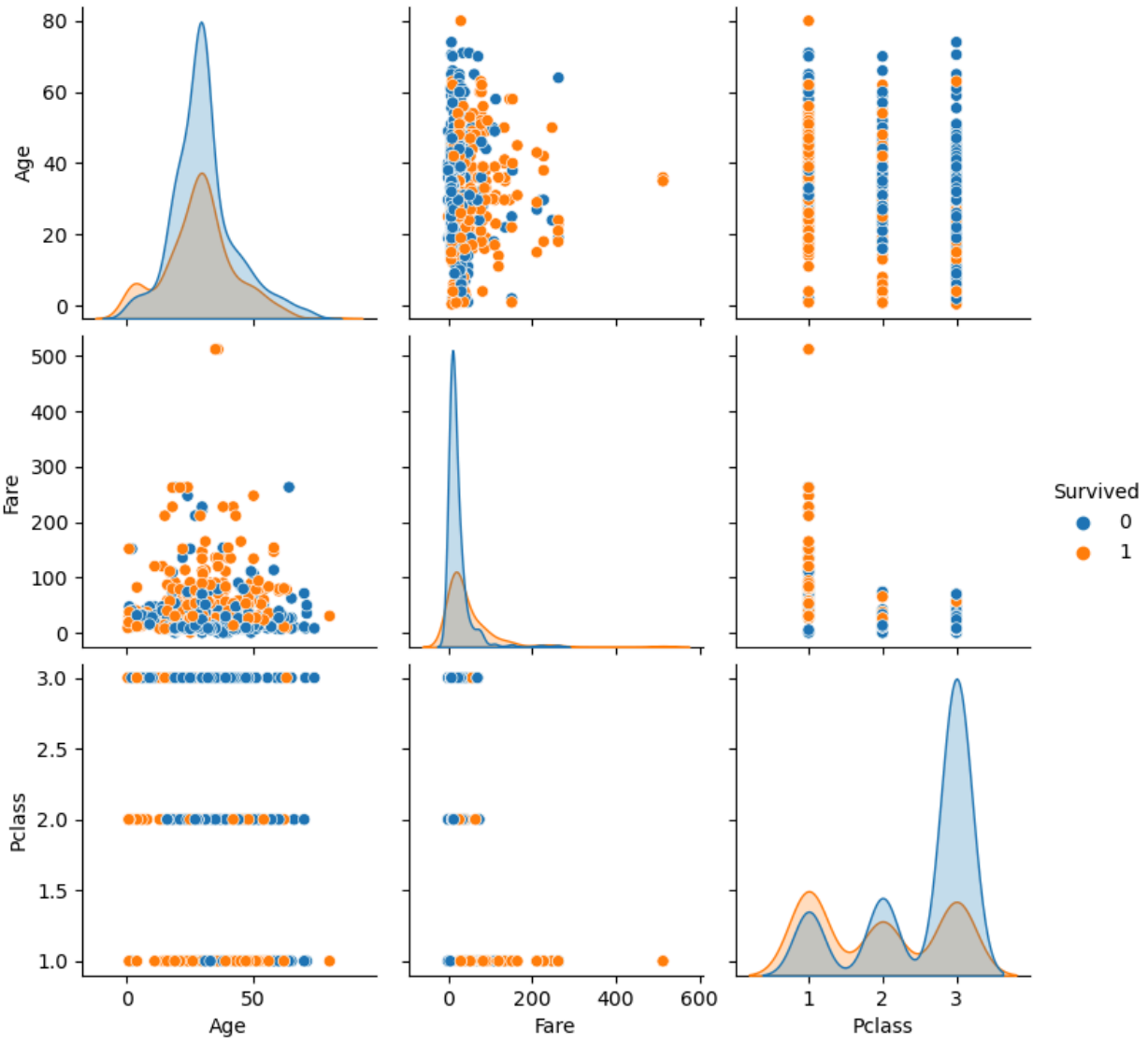
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0
..                                                ...     ...   ...    ...
886                            Montvila, Rev. Juozas    male  27.0      0
887                     Graham, Miss. Margaret Edith  female  19.0      0
888         Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                            Behr, Mr. Karl Howell    male  26.0      0
890                              Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
<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
None
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
________________________________________
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
________________________________________
male      577
```

```
female    314
Name: Sex, dtype: int64
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

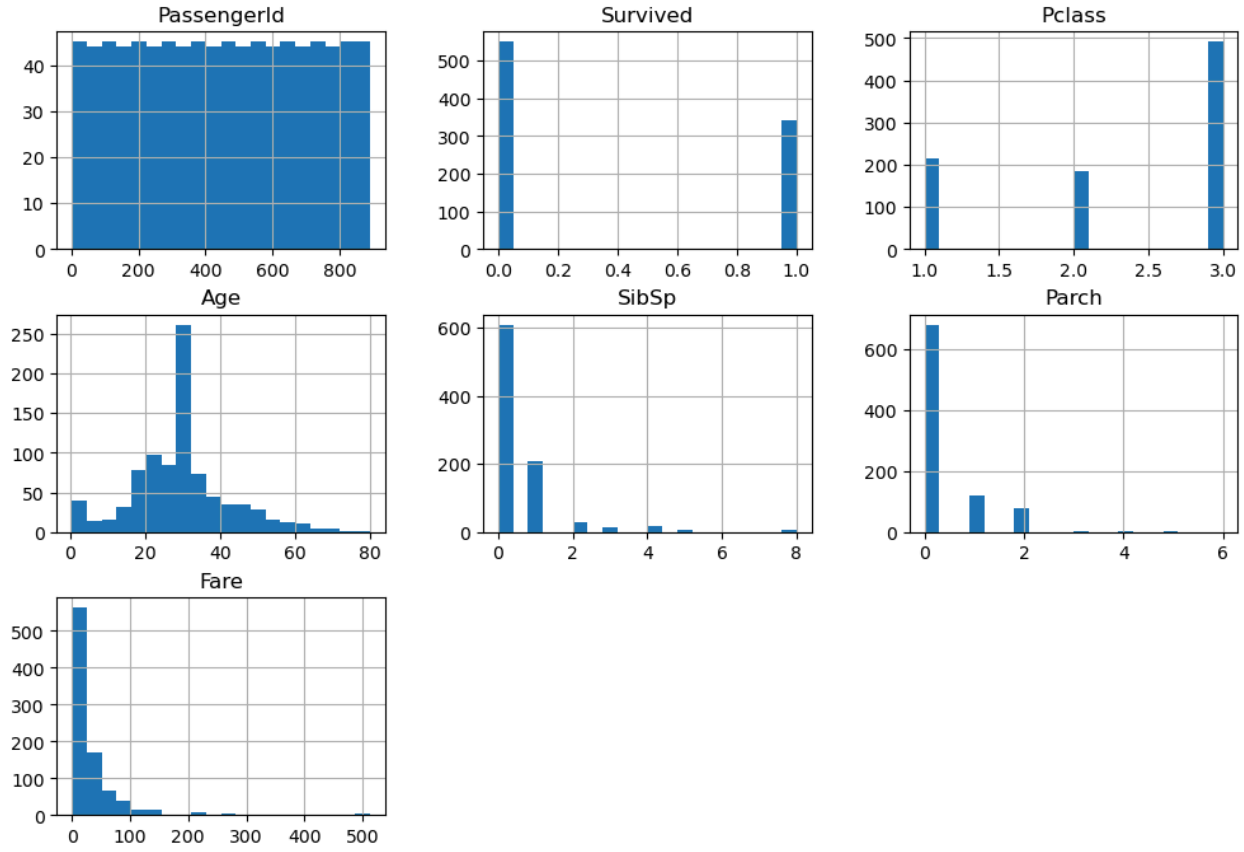
In [26]:
```python
# b.heatmap and pairplot (correlation analysis)
# heatmap
corr = df.corr(numeric_only=True)
plt.figure(figsize=(8,6))
sn.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()

# Pairplot
sn.pairplot(df[['Age', 'Fare', 'Survived', 'Pclass']], hue='Survived')
plt.show()
```

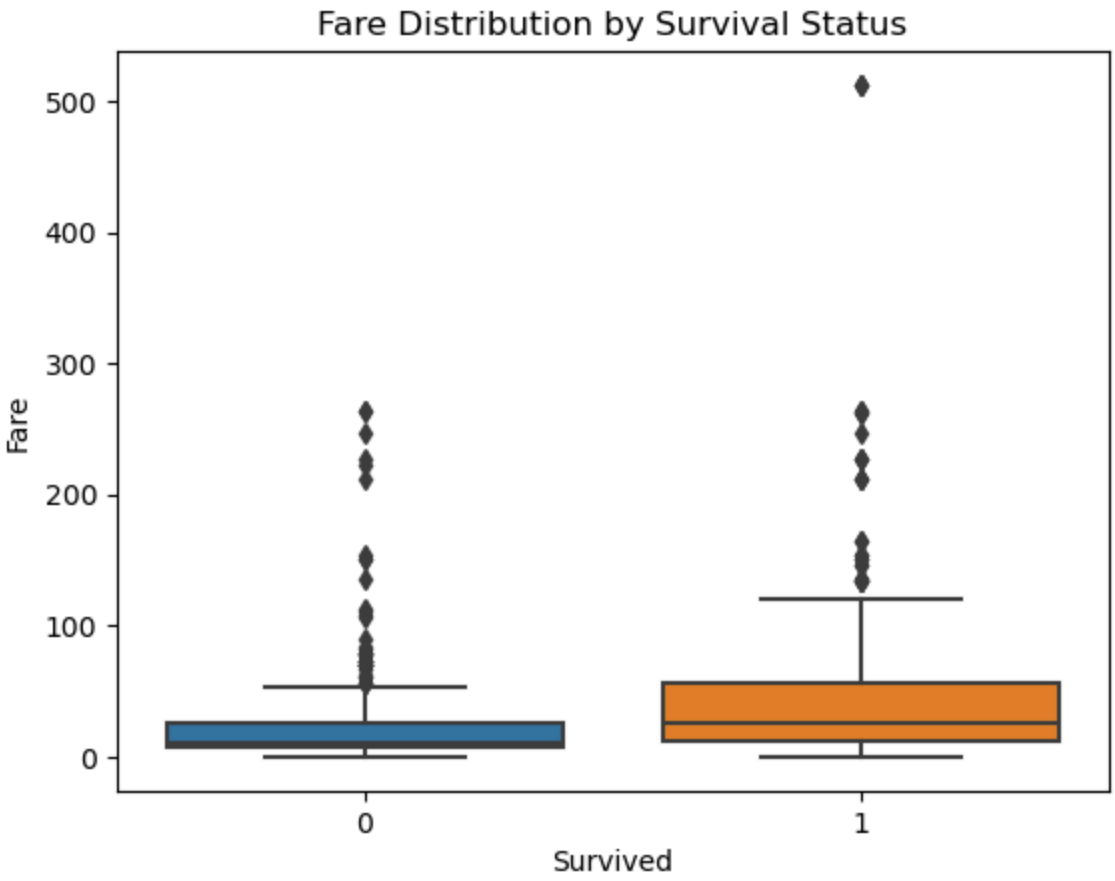


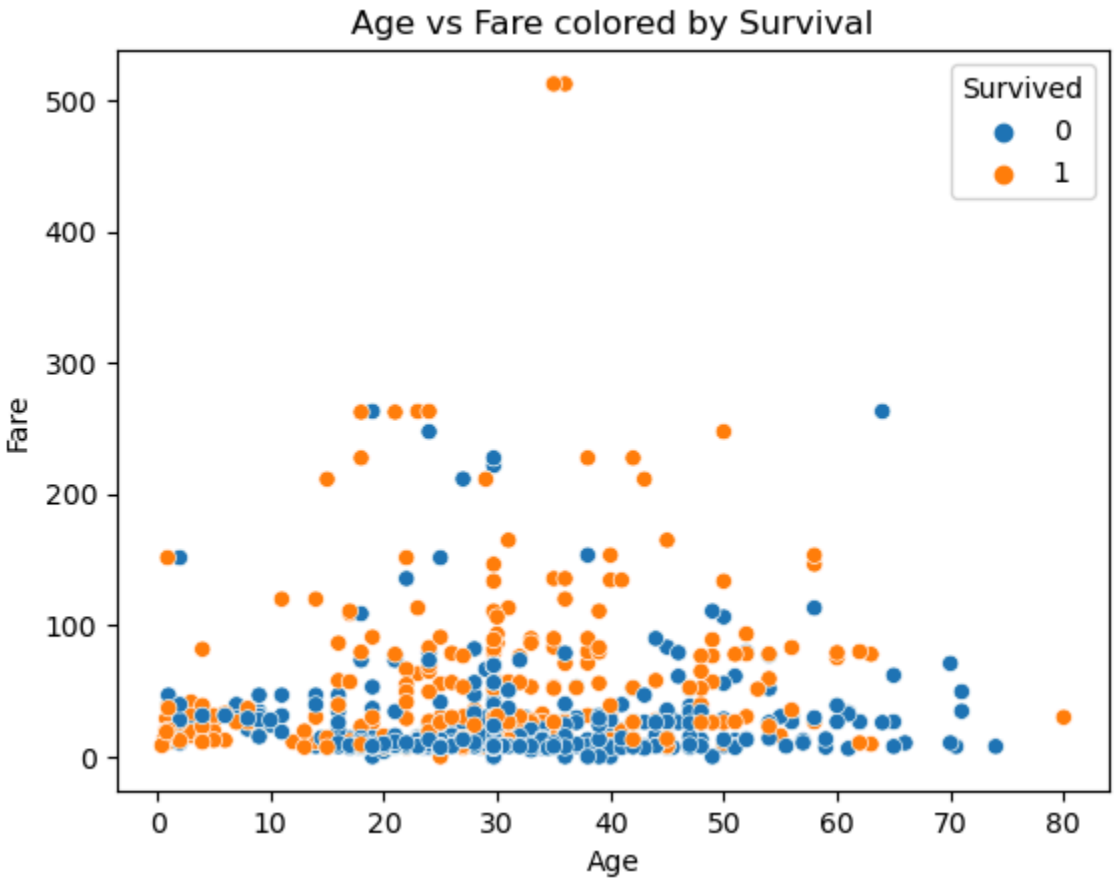


In [19]:

```
# d Histograms for numerical variables (univariate analysis)
df.hist(bins=20, figsize=(12, 8))
plt.suptitle("Histograms of Numerical Features")
plt.show()

# Boxplot: Fare vs Survived (bivariate analysis)
sn.boxplot(data=df, x='Survived', y='Fare')
plt.title("Fare Distribution by Survival Status")
plt.show()

# Scatterplot: Age vs Fare
sn.scatterplot(data=df, x='Age', y='Fare', hue='Survived')
plt.title("Age vs Fare colored by Survival")
plt.show()
```


Histograms of Numerical Features
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare


Fare Distribution by Survival Status
Fare
Survived
0
1
0
100
200
300
400
500


Age vs Fare colored by Survival
Survived
0
1
Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80

In [ ]:
```
# c. Identify relationships and trends

# (univariate analysis)
plt.figure(figsize=(6,4))
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()

# (bivariate anlaysis)
plt.figure(figsize=(6,4))
sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Gender')
plt.show()

plt.figure(figsize=(6,4))
sns.barplot(x='Pclass', y='Survived', data=df)
plt.title('Survival Rate by Passenger Class')
plt.show()

# Key Relationships & Trends

# - Gender & Survival: Female passengers had a much higher survival rate than
# - Class & Survival: First-class passengers survived more often than second c
# - Age & Survival: Younger passengers showed slightly higher survival chances
# - Fare & Survival: Higher fares were linked to higher survival probability,
# - Correlations:
# - Passenger Class ↔ Survival: Negative correlation (-0.338) – lower class =
# - Fare ↔ Survival: Positive correlation (0.257) – higher fare = higher survi
```

In [ ]:
```
# e) Observations
# - Higher survival rate for females than males.
# - First-class passengers had the highest survival rate.
# - Age distribution shows most passengers were 20–40 years old.
# - Fare is right-skewed – some passengers paid extremely high fares.
# - Survival is positively correlated with Fare and negatively with Pclass.
```

In [ ]:
```
# f) Summary of Findings
# 1. Class and Gender were strong determinants of survival.
# 2. High fares and lower classes show different distributions in survival.
# 3. Missing values found in 'Age' and 'Cabin' need imputation or removal.
# 4. Fare distribution is skewed – consider log transformation before modeling
```