In [1]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# To show plots inline
%matplotlib inline
```

In [2]:

```python
df = pd.read_csv("train.csv")
df.head()
```

Out[2]:

	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

In [3]:

```python
df.shape  # rows & columns
df.info() # column types & null values
df.describe() # statistics
df.isnull().sum() # missing values count
```

```
<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
```

Out[3]:

```
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
```

In [4]:

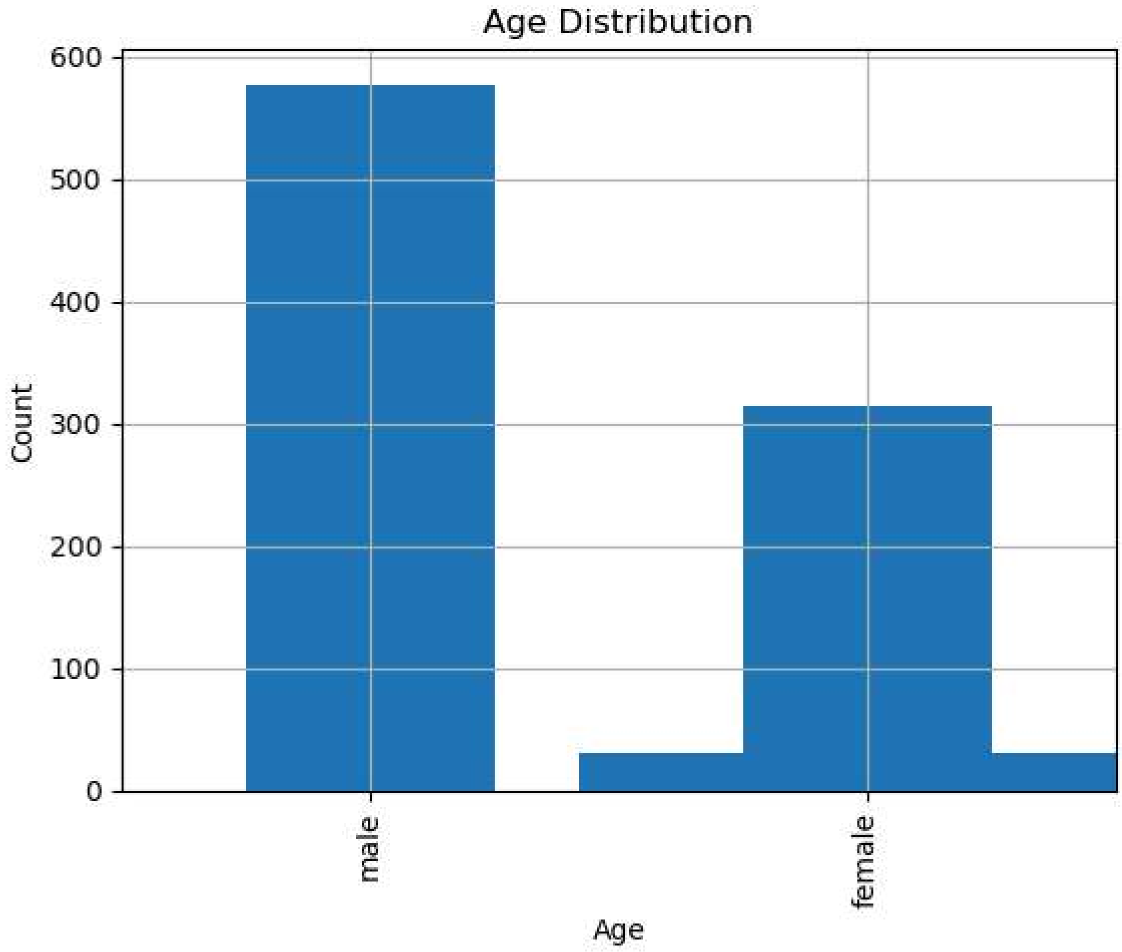
```
#Univariate Analysis
# Categorical variable
df['Sex'].value_counts().plot(kind='bar', title='Gender Count')

# Numerical variable
df['Age'].hist(bins=30)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')
```

Out[4]:

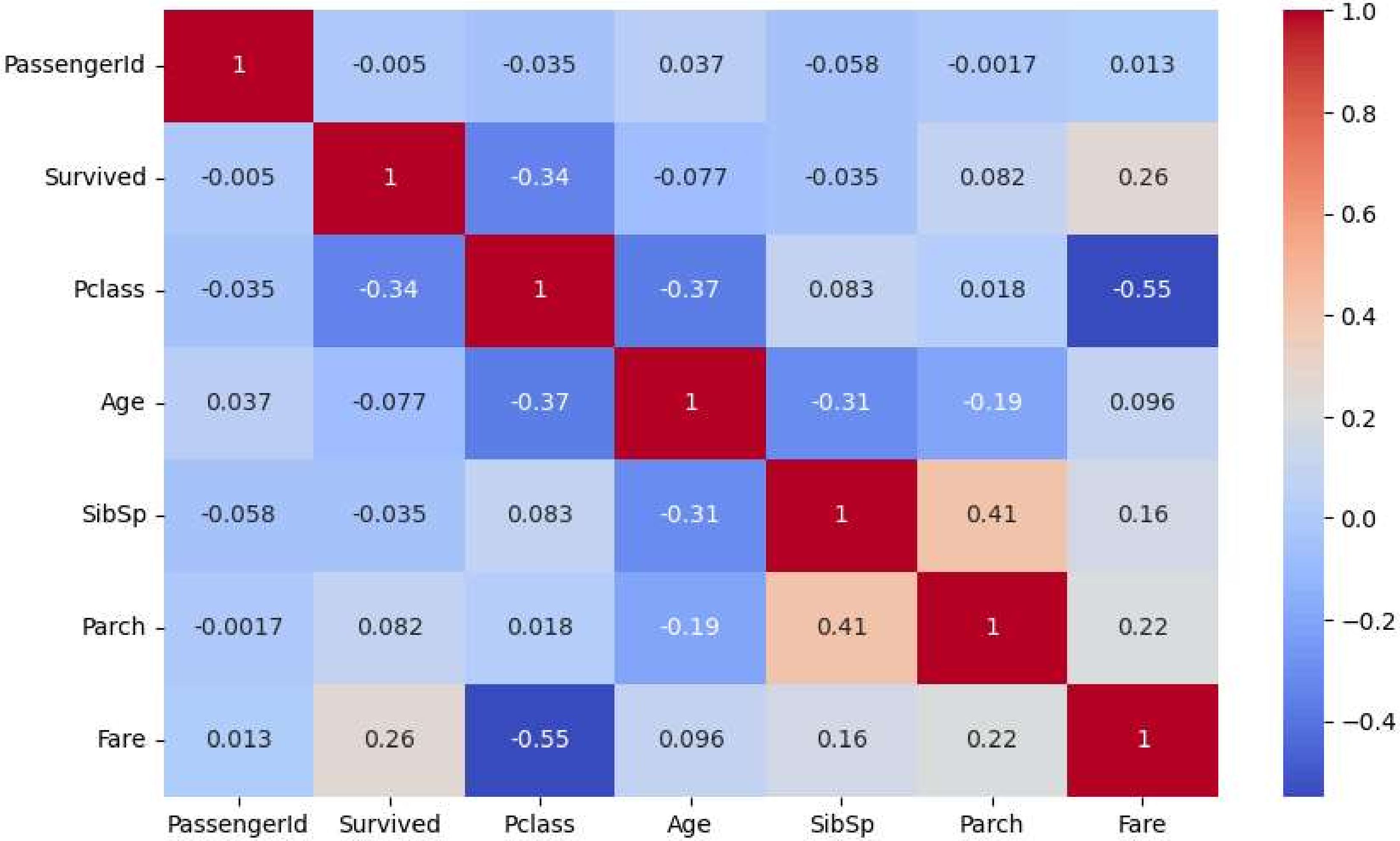
```
Text(0, 0.5, 'Count')
```

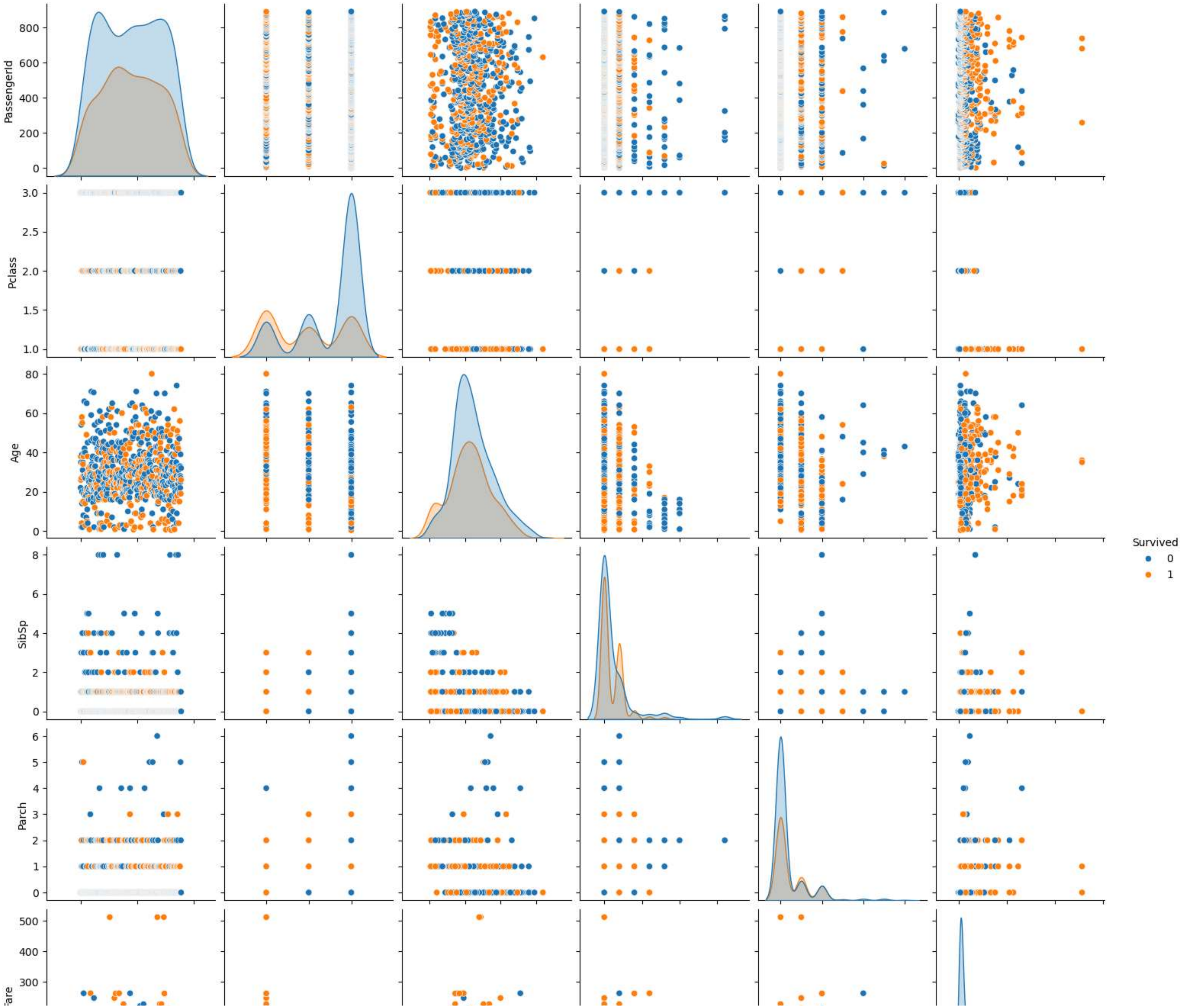
In [7]:

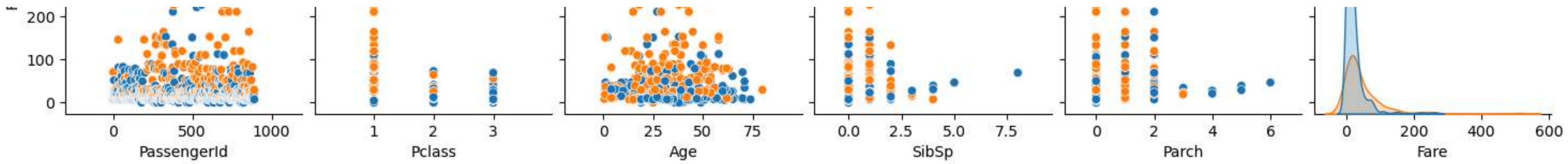
```
#multivariate analysis
# Correlation heatmap
plt.figure(figsize=(10,6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')


# Pairplot
sns.pairplot(df, hue='Survived')
plt.show()
```

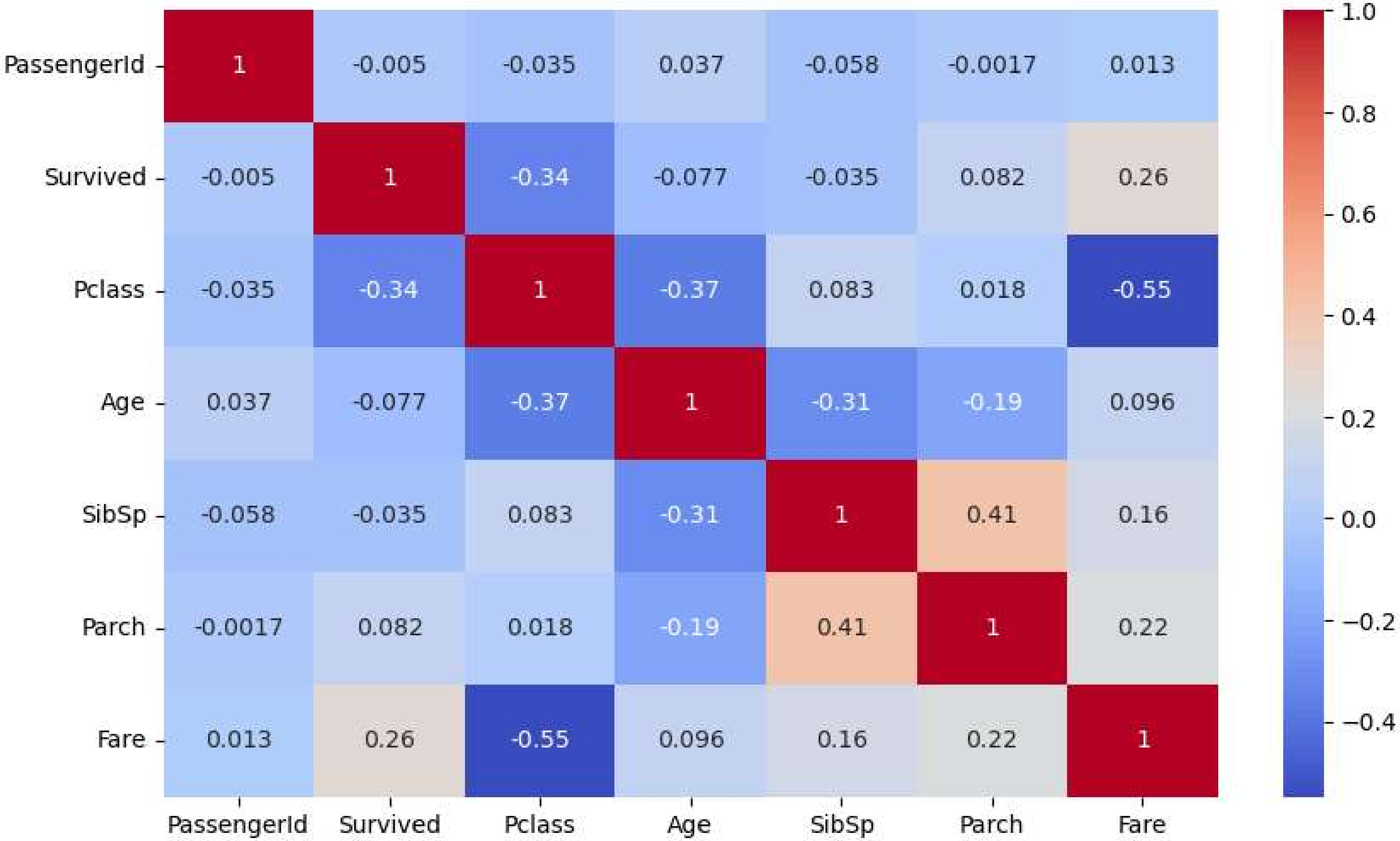

Age Distribution
600
500
400
300
200
100
0
Count
male
female
Age

```
<Figure size 1000x600 with 0 Axes>
```

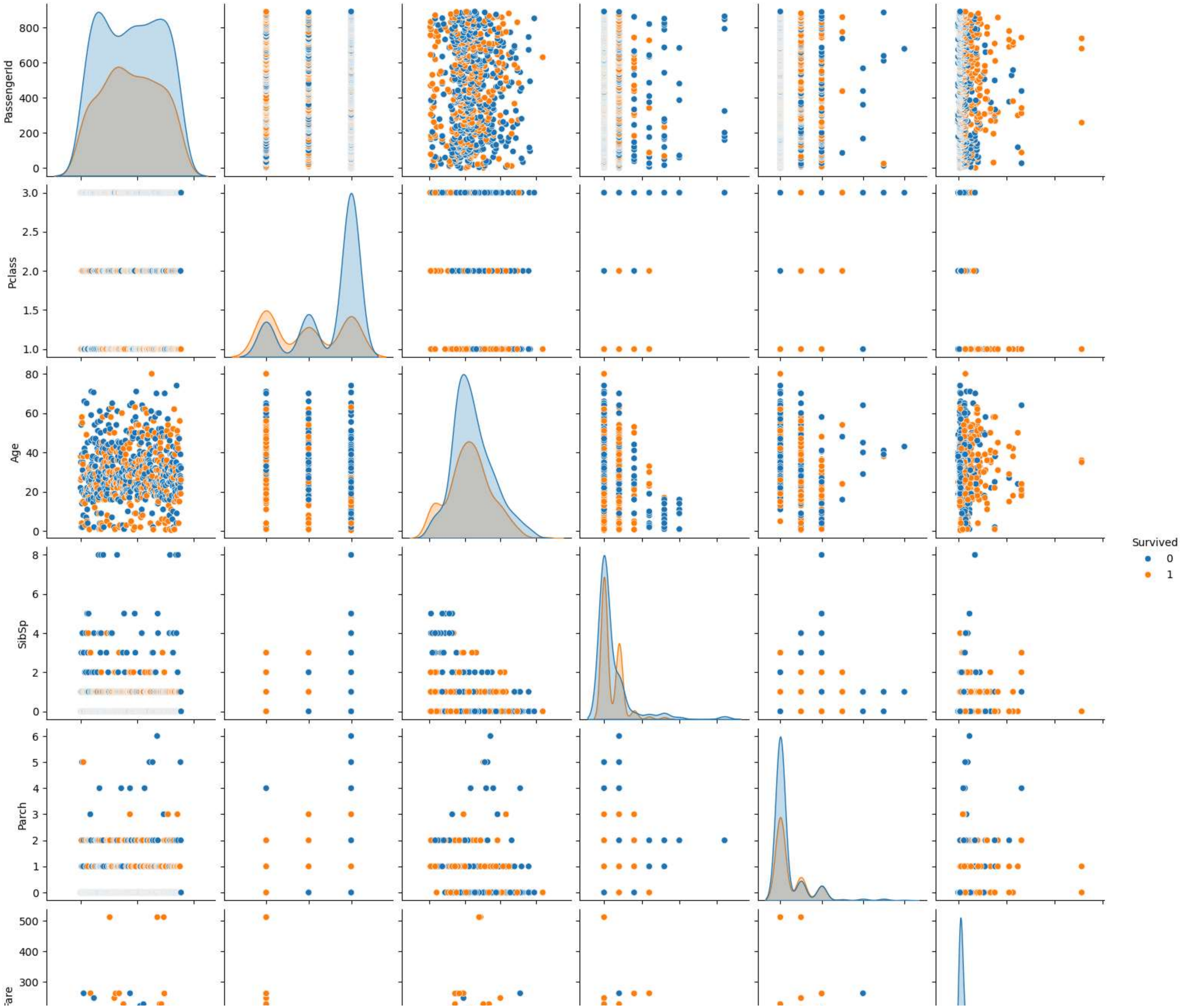

PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1 -0.005 -0.035 0.037 -0.058 -0.0017 0.013
-0.005 1 -0.34 -0.077 -0.035 0.082 0.26
-0.035 -0.34 1 -0.37 0.083 0.018 -0.55
0.037 -0.077 -0.37 1 -0.31 -0.19 0.096
-0.058 -0.035 0.083 -0.31 1 0.41 0.16
-0.0017 0.082 0.018 -0.19 0.41 1 0.22
0.013 0.26 -0.55 0.096 0.16 0.22 1
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

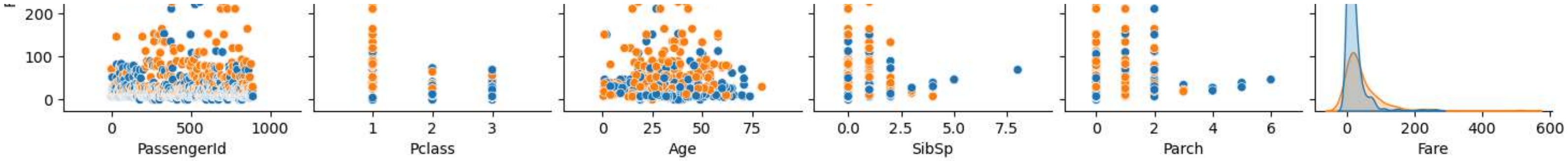

PassengerId
800
600
400
200
0
Pclass
3.0
2.5
2.0
1.5
1.0
Age
80
60
40
20
0
SibSp
8
6
4
2
0
Parch
6
5
4
3
2
1
0
Fare
500
400
300
Survived
0
1





	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1	-0.005	-0.035	0.037	-0.058	-0.0017	0.013
Survived	-0.005	1	-0.34	-0.077	-0.035	0.082	0.26
Pclass	-0.035	-0.34	1	-0.37	0.083	0.018	-0.55
Age	0.037	-0.077	-0.37	1	-0.31	-0.19	0.096
SibSp	-0.058	-0.035	0.083	-0.31	1	0.41	0.16
Parch	-0.0017	0.082	0.018	-0.19	0.41	1	0.22
Fare	0.013	0.26	-0.55	0.096	0.16	0.22	1


PassengerId
800
600
400
200
0
Pclass
3.0
2.5
2.0
1.5
1.0
Age
80
60
40
20
0
SibSp
8
6
4
2
0
Parch
6
5
4
3
2
1
0
Fare
500
400
300
Survived
0
1



📝 Observations & Insights

- **Gender vs Survival:** Females had a higher survival rate compared to males.
- **Age Distribution:** Most passengers were between 20–40 years old.
- **Survival by Class:** Passengers in 1st class had a higher chance of survival than those in 3rd class.
- **Heatmap:** Fare shows a moderate positive correlation with survival, while Pclass has a negative correlation.
- **Embarked vs Survival:** Passengers boarding from port 'C' had a slightly higher survival rate.

In [9]:

```
#handle missing data
df.fillna({'Age': df['Age'].median(),
           'Embarked': df['Embarked'].mode()[0]}, inplace=True)
```

In [ ]: