In [3]:
```
import pandas as pd

df = pd.read_csv('C:\\Users\\Kurien\\Downloads\\EDA_Titanic_Task5\\train.csv')  # R
df.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500

In [4]:
```
# Basic structure
df.info()

# Statistical summary
df.describe()

# Checking null values
df.isnull().sum()
```

```
<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[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
```

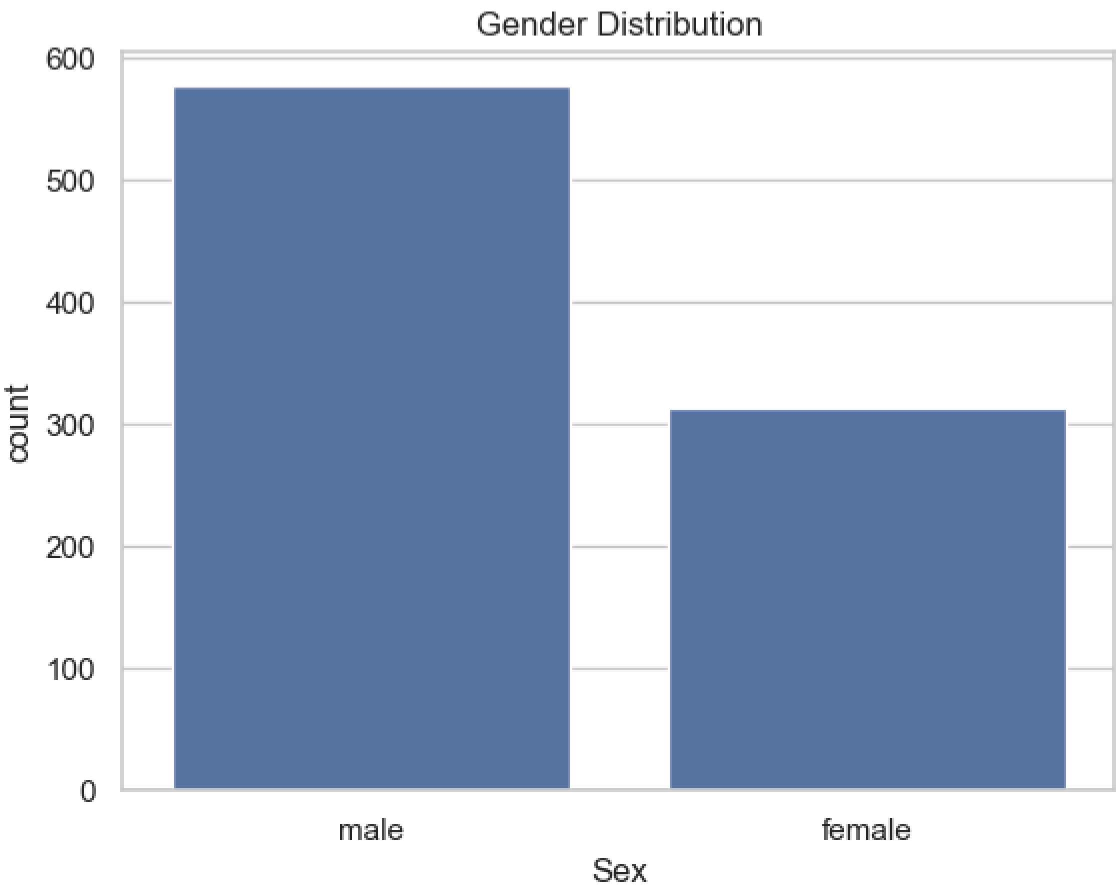
In [15]:

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Style setup
sns.set(style='whitegrid')
```

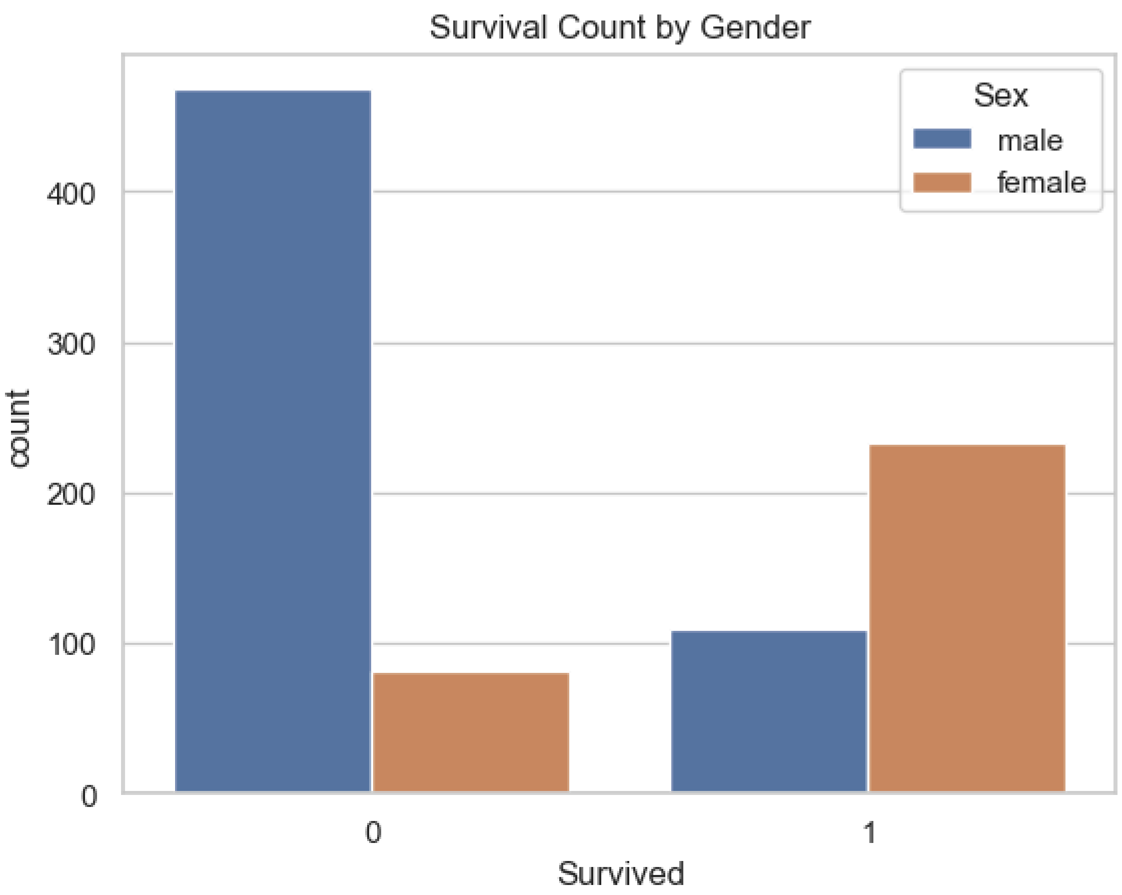
In [9]:

```python
sns.countplot(x='Sex', data=df)
plt.title('Gender Distribution')
plt.show()
```



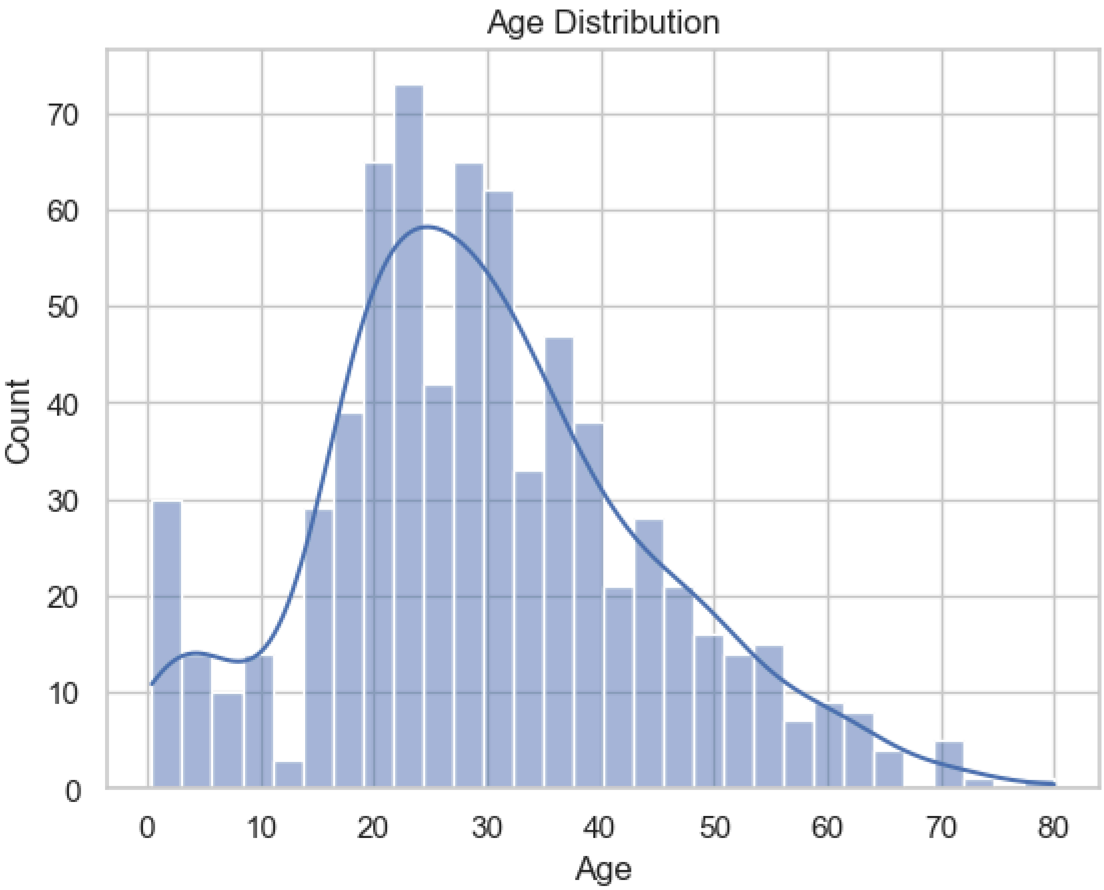
In [10]:

```
sns.countplot(x='Survived', hue='Sex', data=df)
plt.title('Survival Count by Gender')
plt.show()
```



In [11]:

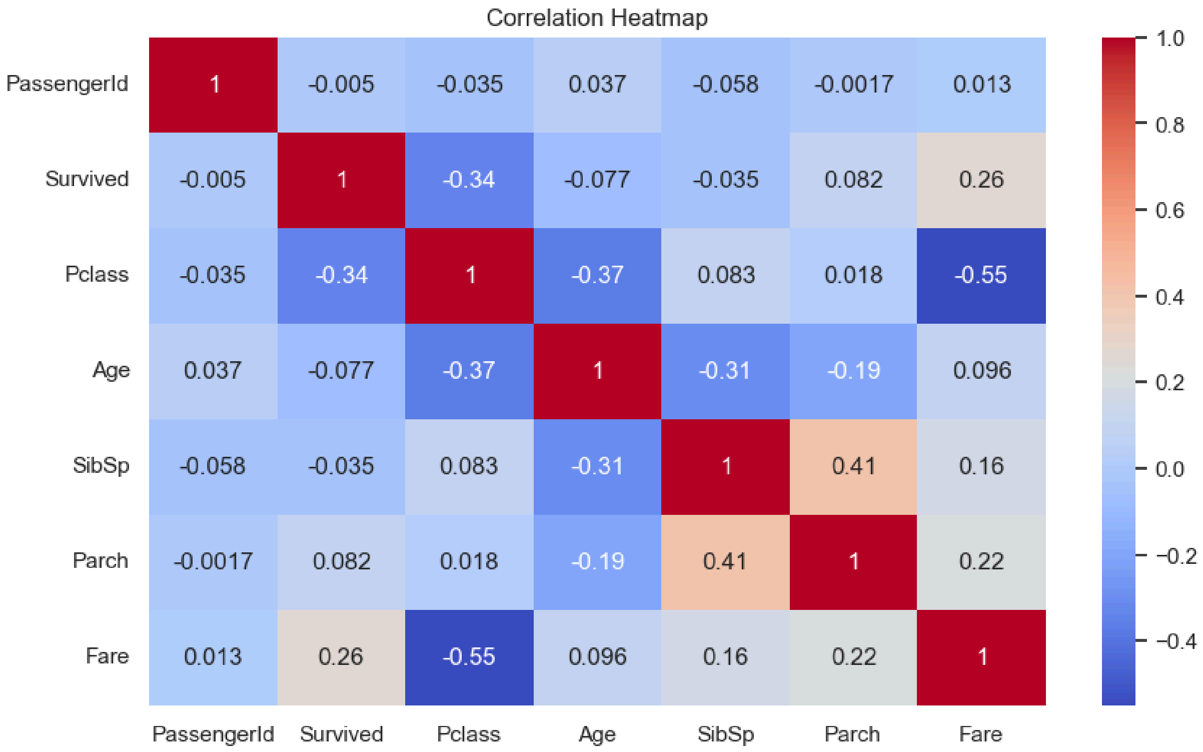
```
sns.histplot(df['Age'].dropna(), kde=True, bins=30)
plt.title('Age Distribution')
plt.show()
```



In [13]:

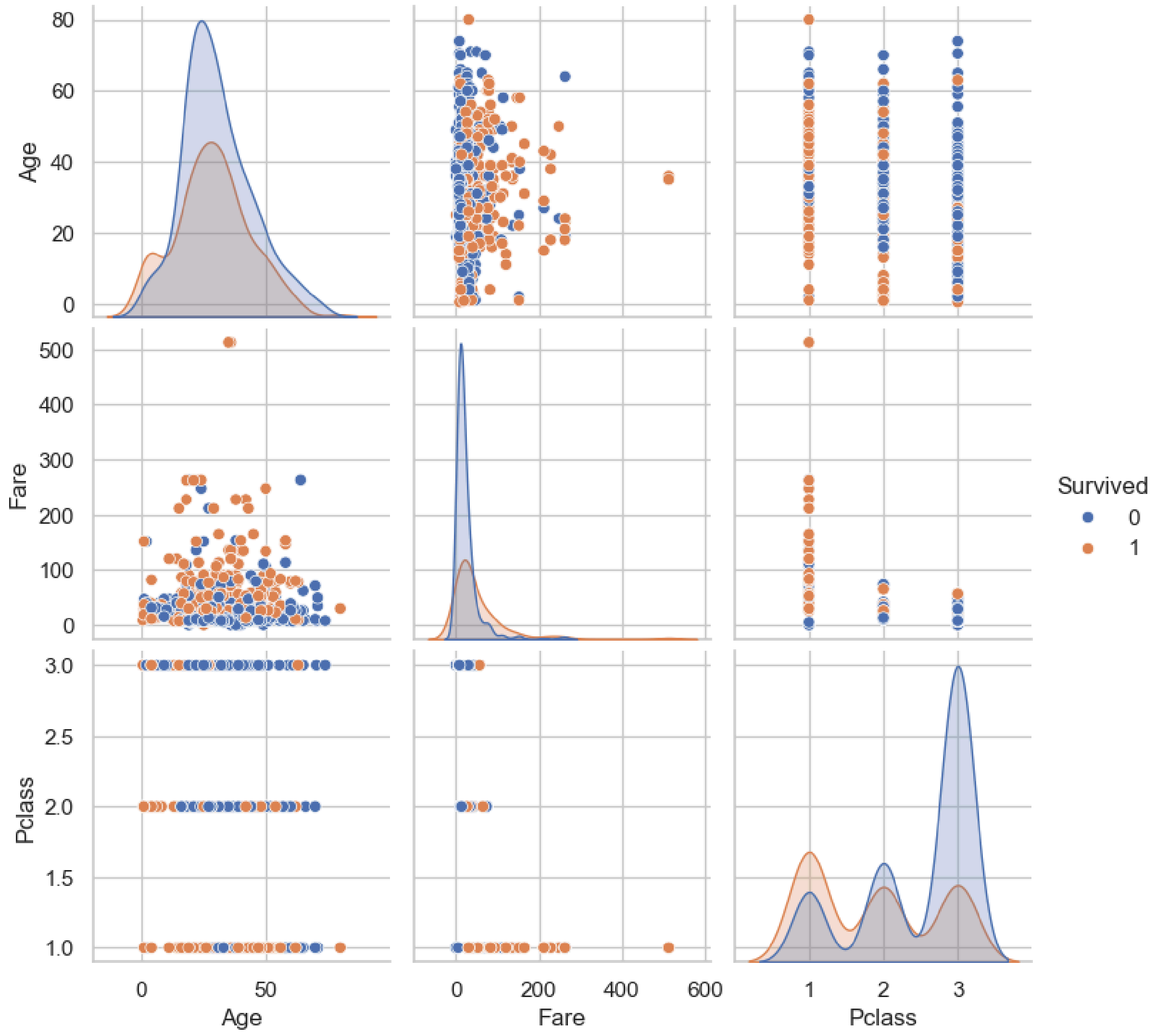
```python
# Select only the numeric columns from the DataFrame
numeric_df = df.select_dtypes(include=['float64', 'int64'])

# Plot the heatmap with the numeric data
plt.figure(figsize=(10,6))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



```
In [14]: sns.pairplot(df.dropna(subset=['Age', 'Fare']), vars=['Age', 'Fare', 'Pclass'], hue
```

```
Out[14]: <seaborn.axisgrid.PairGrid at 0x25c0e0302f0>
```



📊 Gender Distribution

- There are **more males** on board than females.
- Male passengers outnumber female passengers by a significant margin.

📊 Survival Count by Gender

- **Females had a higher survival rate** compared to males.
- Most male passengers did not survive, while a larger portion of female passengers did.

📊 Age Distribution

- The **majority of passengers were younger adults and children**.
- There is a noticeable peak around ages **20–30**.
- Fewer elderly passengers were present, and very few were above **60**.

📊 Correlation Heatmap

- **Survival** is **positively correlated with Fare** (higher fare, higher survival chances).
- **Pclass** (ticket class) is **negatively correlated with Survival** (lower class, lower survival chances).
- **Age and Survival** have a weak negative correlation (younger passengers slightly more likely to survive).

In [ ]: