In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_style("whitegrid")
```

In [2]:
```
df = pd.read_csv("train.csv")
df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	I
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0

In [3]:
```
df.info()
```

```
<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
```

In [4]:
```
df.describe()
```

Out[4]:

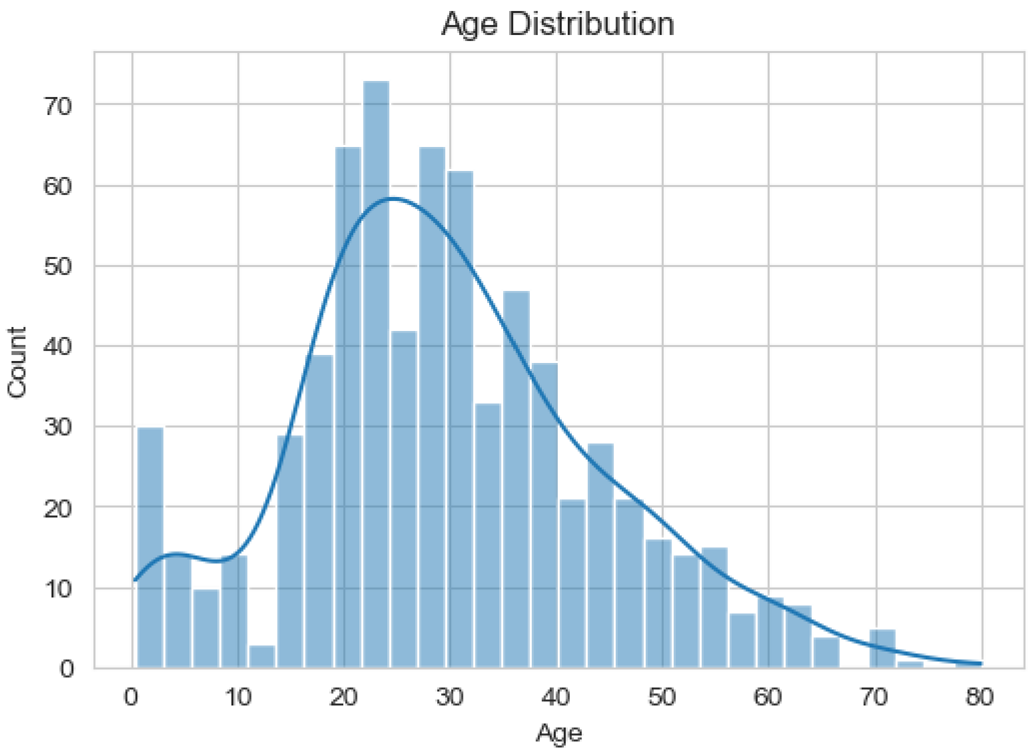
	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.00(
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.20
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.69
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.00(
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.91(
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.45
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.00(
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.32

In [5]:
```
df.isnull().sum()
```

Out[5]:
```
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 [6]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

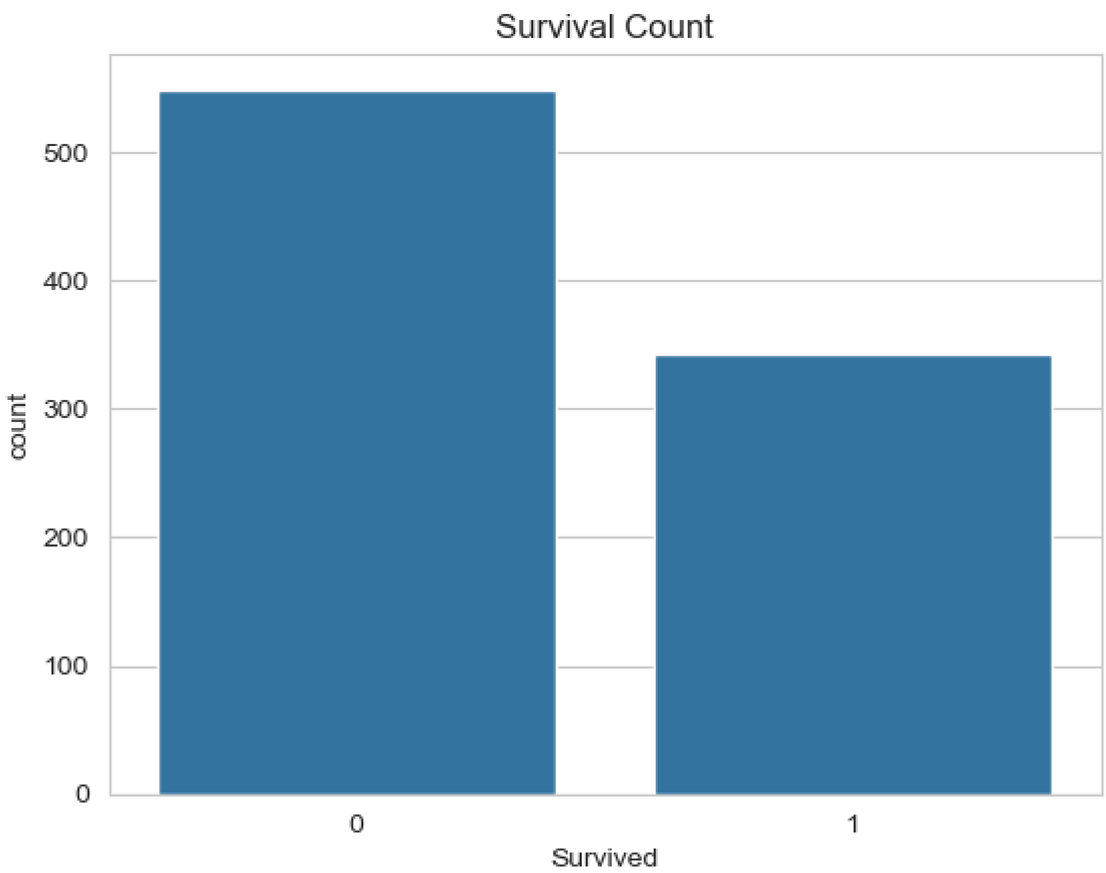
```
plt.figure(figsize=(6,4))
sns.histplot(df['Age'], bins=30, kde=True)
plt.title("Age Distribution")
plt.show()
```



Observation:

- The age distribution of passengers is slightly right-skewed.
- Most passengers fall in the age range of 20–40 years, indicating a predominantly young adult population onboard.
- Very few passengers were elderly (above 60 years), and a smaller proportion were children.
- This suggests that the Titanic mainly carried young and middle-aged adults, which may influence survival-related patterns.

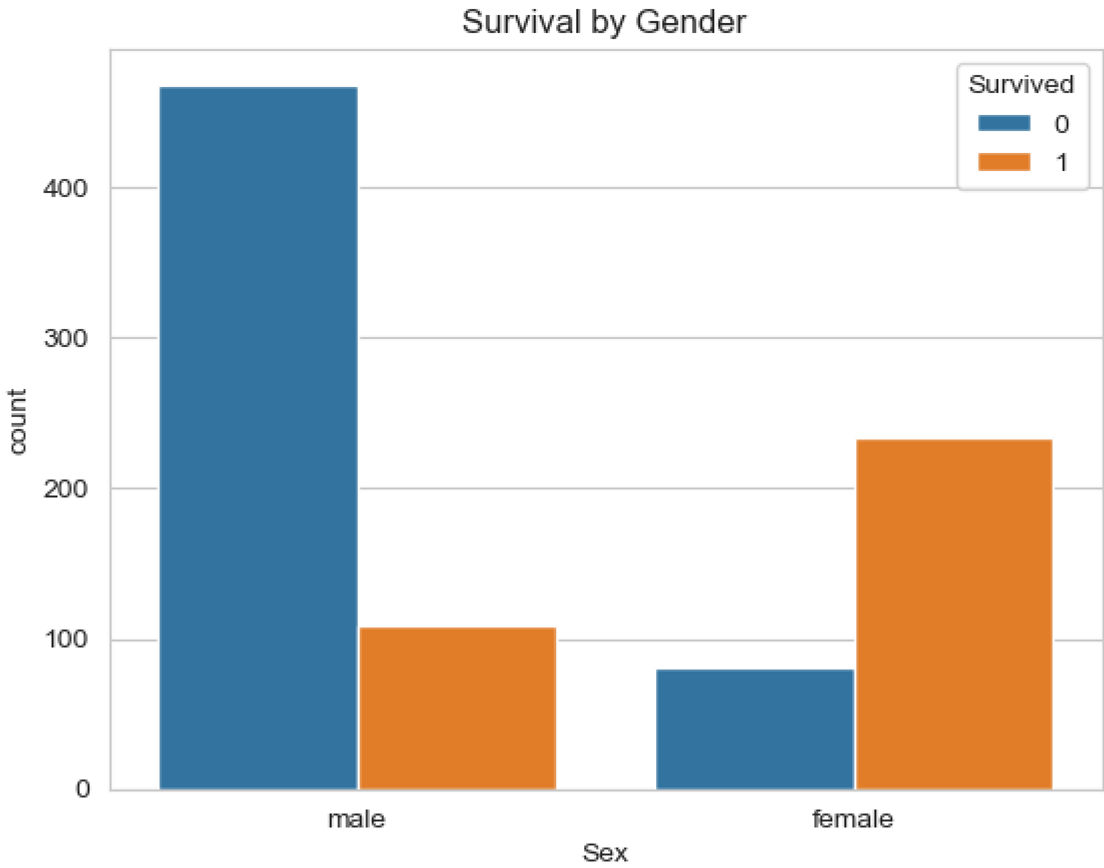
In [7]:
```
sns.countplot(x='Survived', data=df)
plt.title("Survival Count")
plt.show()
```



Survival Count Observation:

- The number of passengers who did not survive is higher than those who survived.
- This indicates that the overall survival rate on the Titanic was low.
- The dataset is imbalanced with respect to the target variable (Survived), which is important to consider during analysis and modeling.

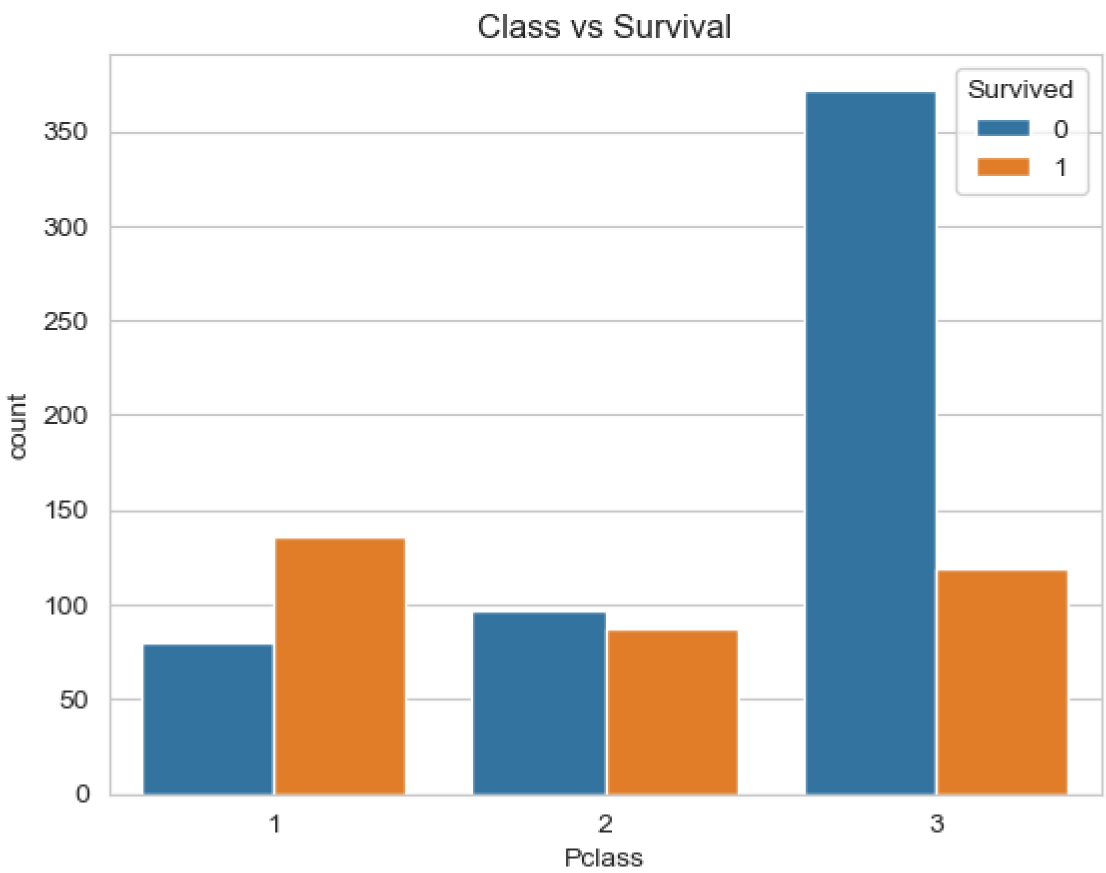
In [8]:
```
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival by Gender")
plt.show()
```



Survival by Gender Observation:

- Female passengers had a significantly higher survival rate compared to male passengers.
- A large proportion of male passengers did not survive, whereas most female passengers survived.
- This indicates that gender played a crucial role in survival, likely due to evacuation priorities during the disaster.

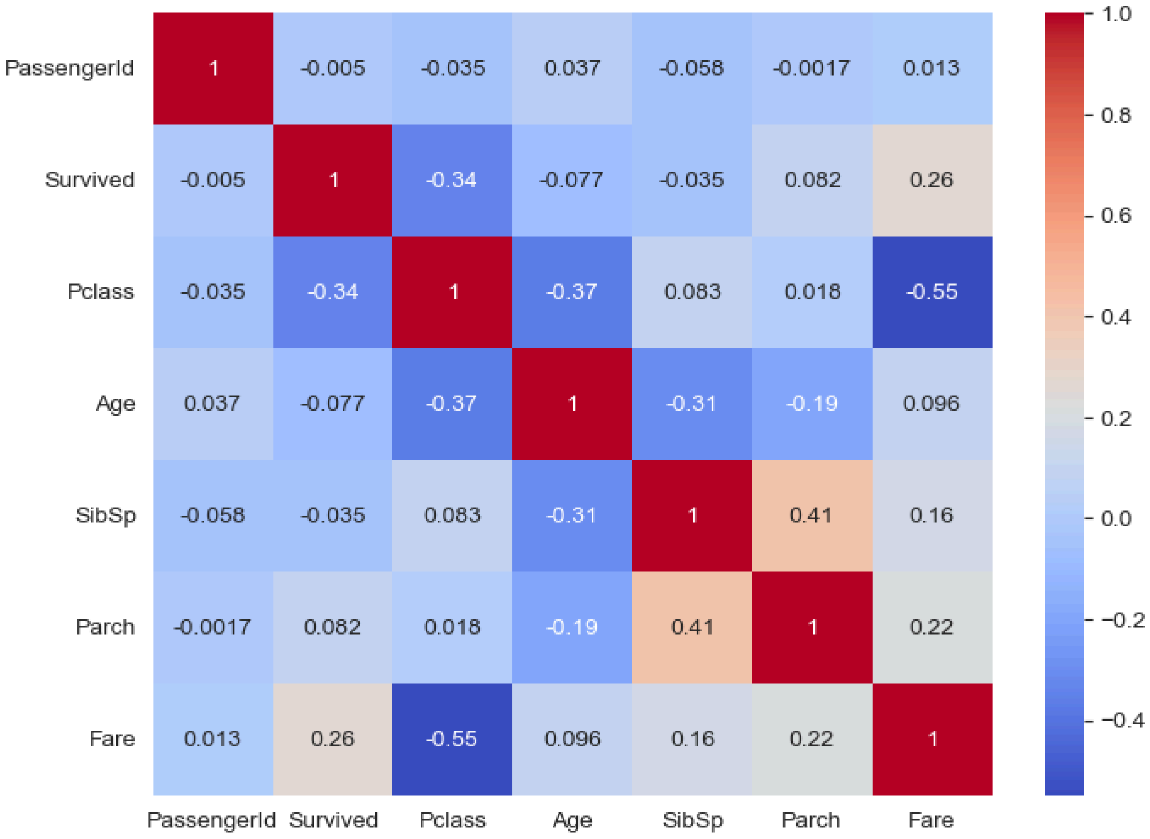
In [9]:
```
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Class vs Survival")
plt.show()
```



Survival by Passenger Class Observation:

- Passengers in first class (Pclass = 1) had the highest survival rate.
- Survival rate decreases as passenger class moves from first to third class.
- A large number of third-class passengers did not survive compared to first- and second-class passengers.
- This indicates that socio-economic status and access to resources played a significant role in survival.

In [10]:
```
plt.figure(figsize=(8,6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap="coolwarm")
plt.show()
```



Correlation Heatmap Observation:

- Survival shows a moderate positive correlation with Fare, indicating that passengers who paid higher fares were more likely to survive.
- Survival has a negative correlation with Pclass, suggesting that lower-class passengers had lower survival chances.
- Pclass and Fare show a strong negative correlation, meaning higher-class passengers generally paid higher fares.
- Other features such as Age, SibSp, and Parch show weak correlations with survival.

Insight:

- Fare and passenger class are important predictors of survival.
- No strong multicollinearity is observed among most features, except between Pclass and Fare.

🔍 Summary of Findings

- The majority of passengers were young adults aged between 20–40 years.
- Overall survival rate was low, with more passengers not surviving.
- Female passengers and first-class passengers had significantly higher survival rates.
- Higher fares were associated with better survival outcomes.
- Passenger class and fare emerged as key factors influencing survival.

```
In [ ]:
```