In [26]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid")
```

In [27]:
```
df = pd.read_csv("train.csv")
```

In [28]:
```
df.head()
df.tail()
df.shape
df.info()
df.describe(include="all")
df.isnull().sum()
```

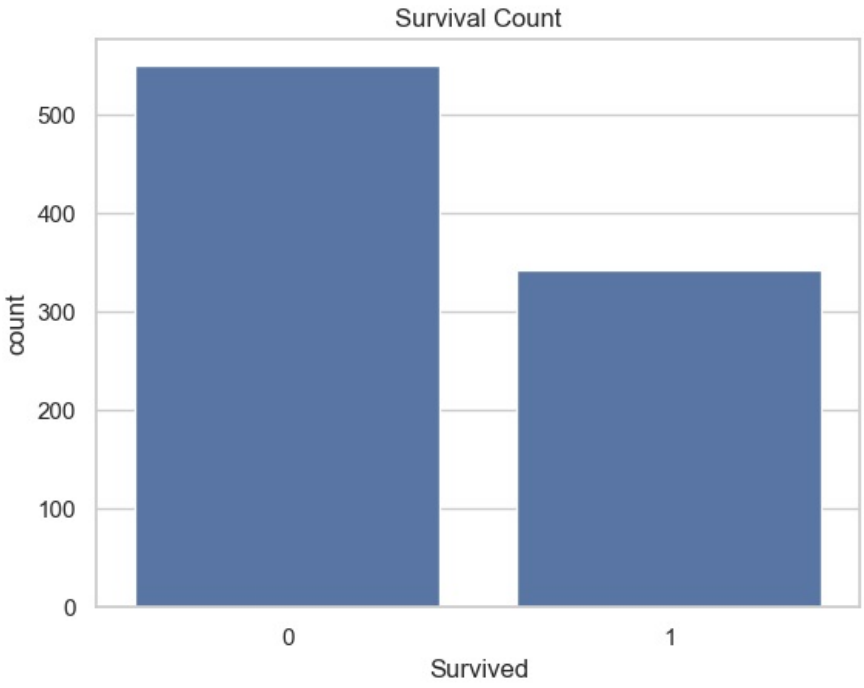
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    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[28]:
```
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
```

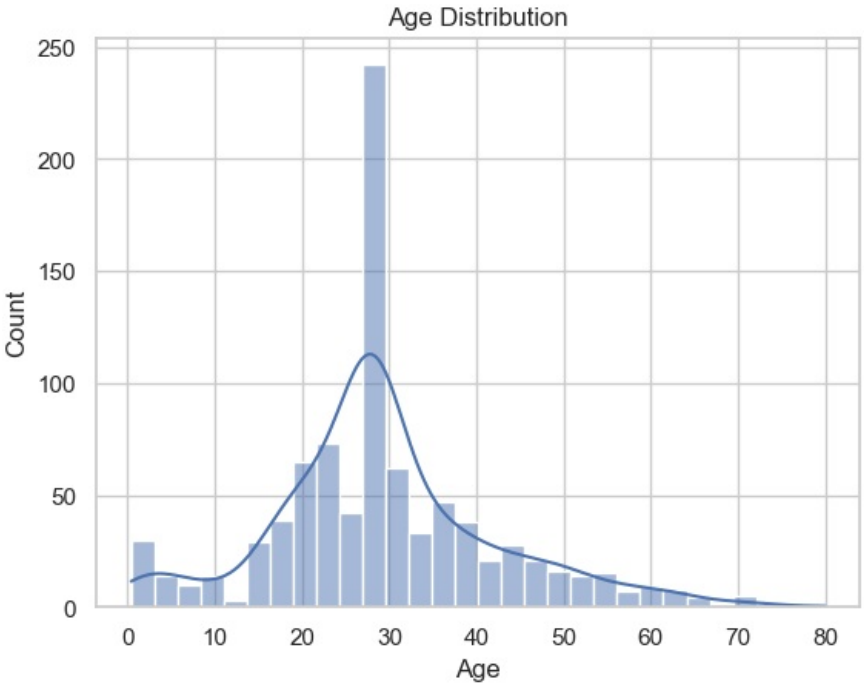
In [29]:
```
df['Age'] = df['Age'].fillna(df['Age'].median())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
df['Deck'] = df['Cabin'].astype(str).str[0]
df['Deck'] = df['Deck'].replace('n', np.nan)   # n means 'nan'
df.isnull().sum()
```

Out[29]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         0
Deck           687
dtype: int64
```

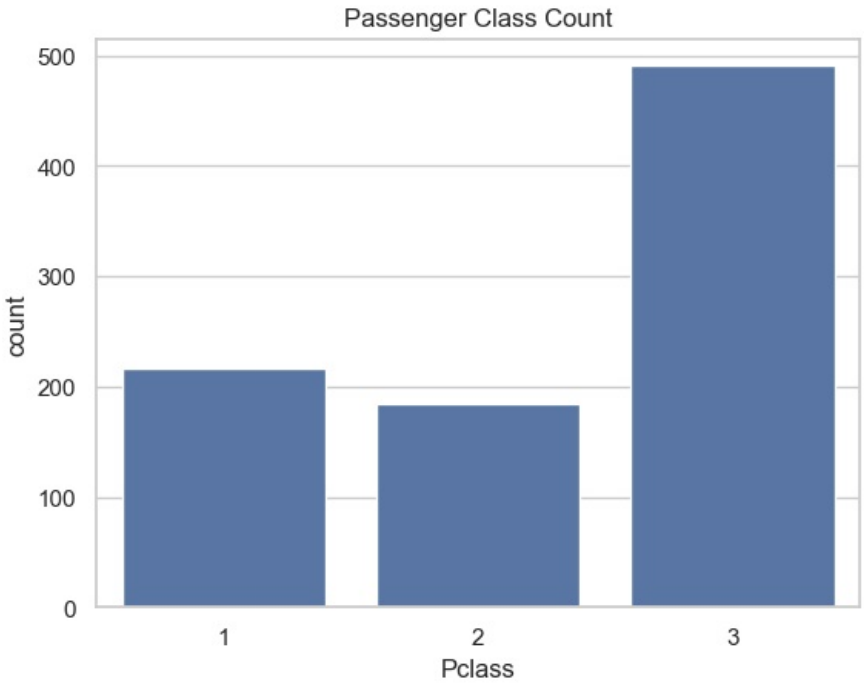
In [30]:
```
sns.countplot(x='Survived', data=df)
plt.title("Survival Count")
plt.show()
```



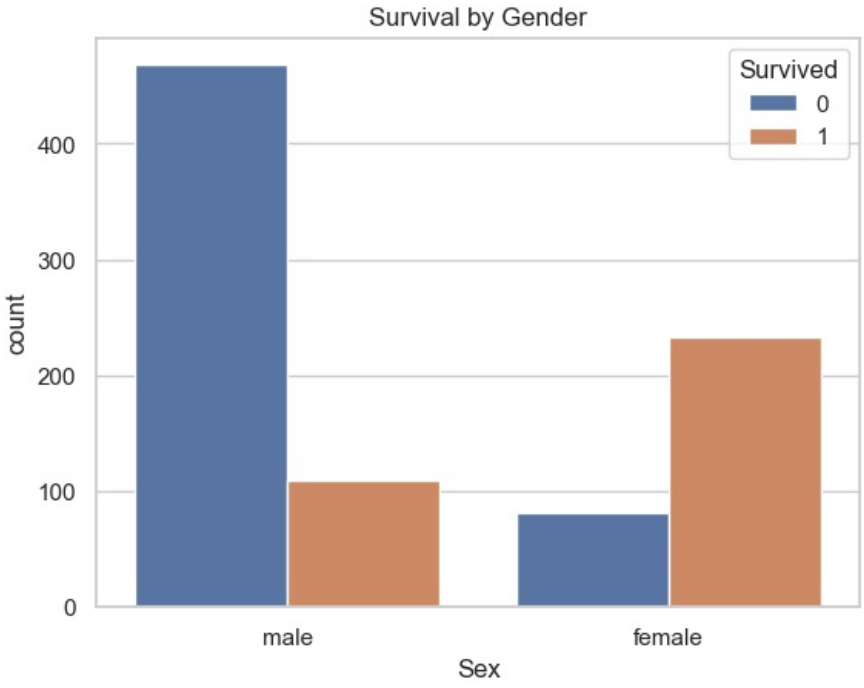
In [31]:
```
sns.histplot(df['Age'], kde=True)
plt.title("Age Distribution")
plt.show()
```



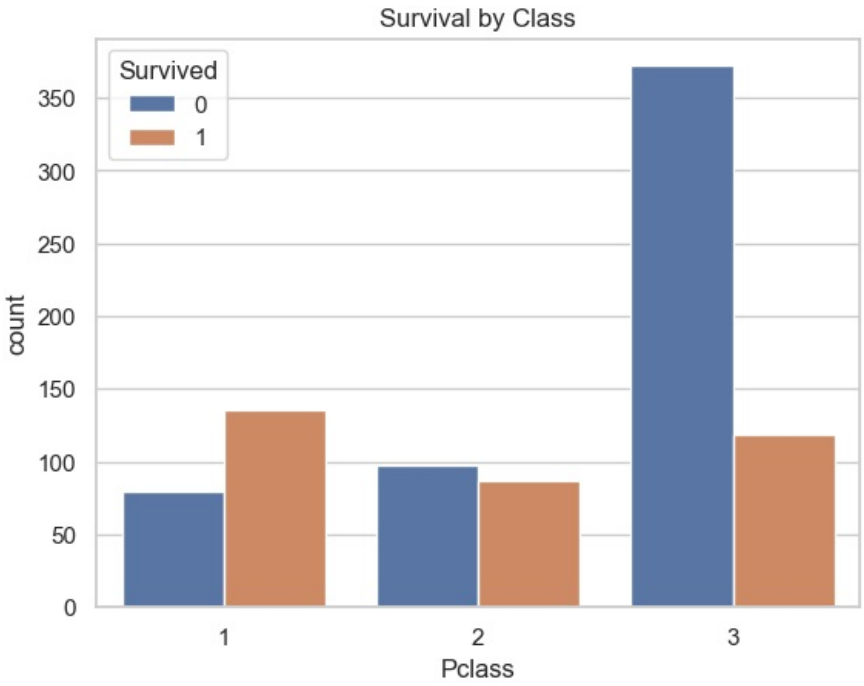
In [32]:
```
sns.countplot(x='Pclass', data=df)
plt.title("Passenger Class Count")
plt.show()
```



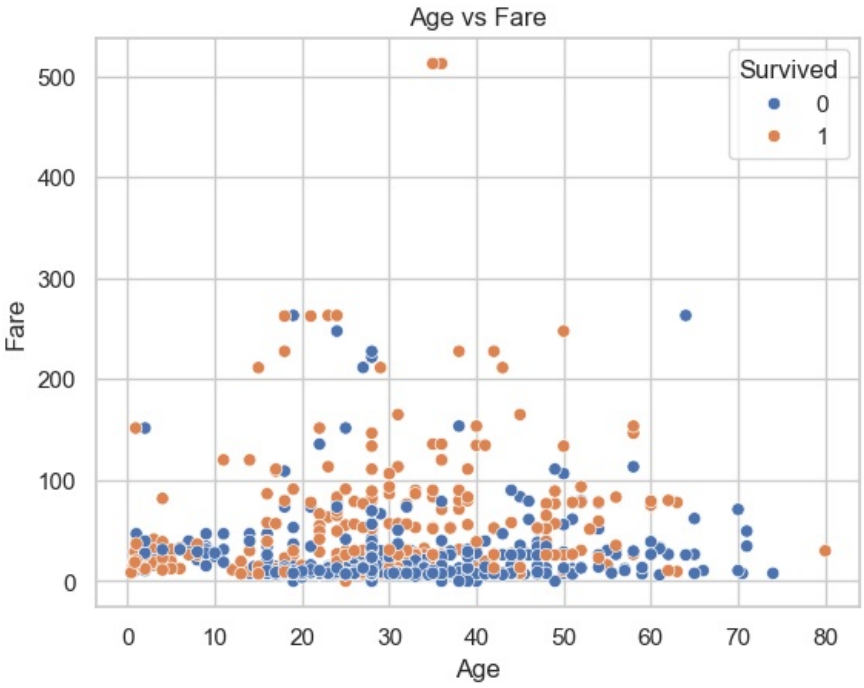
In [33]:
```
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival by Gender")
plt.show()
```



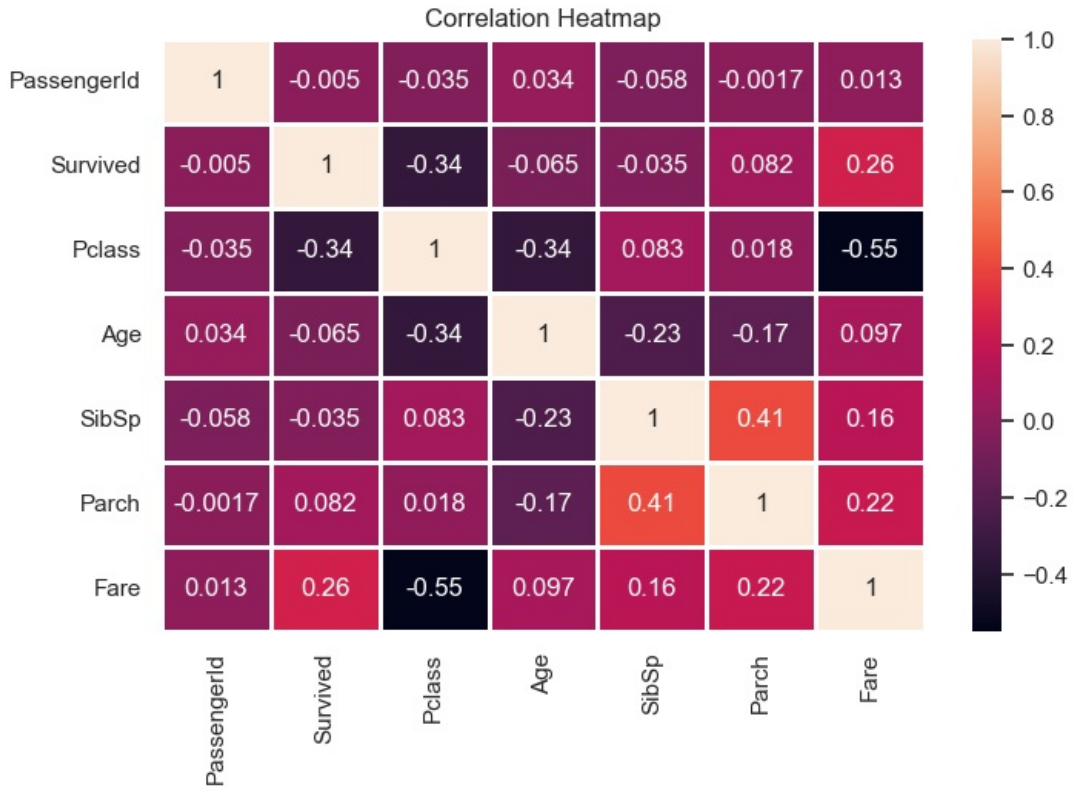
In [34]:
```
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Survival by Class")
plt.show()
```



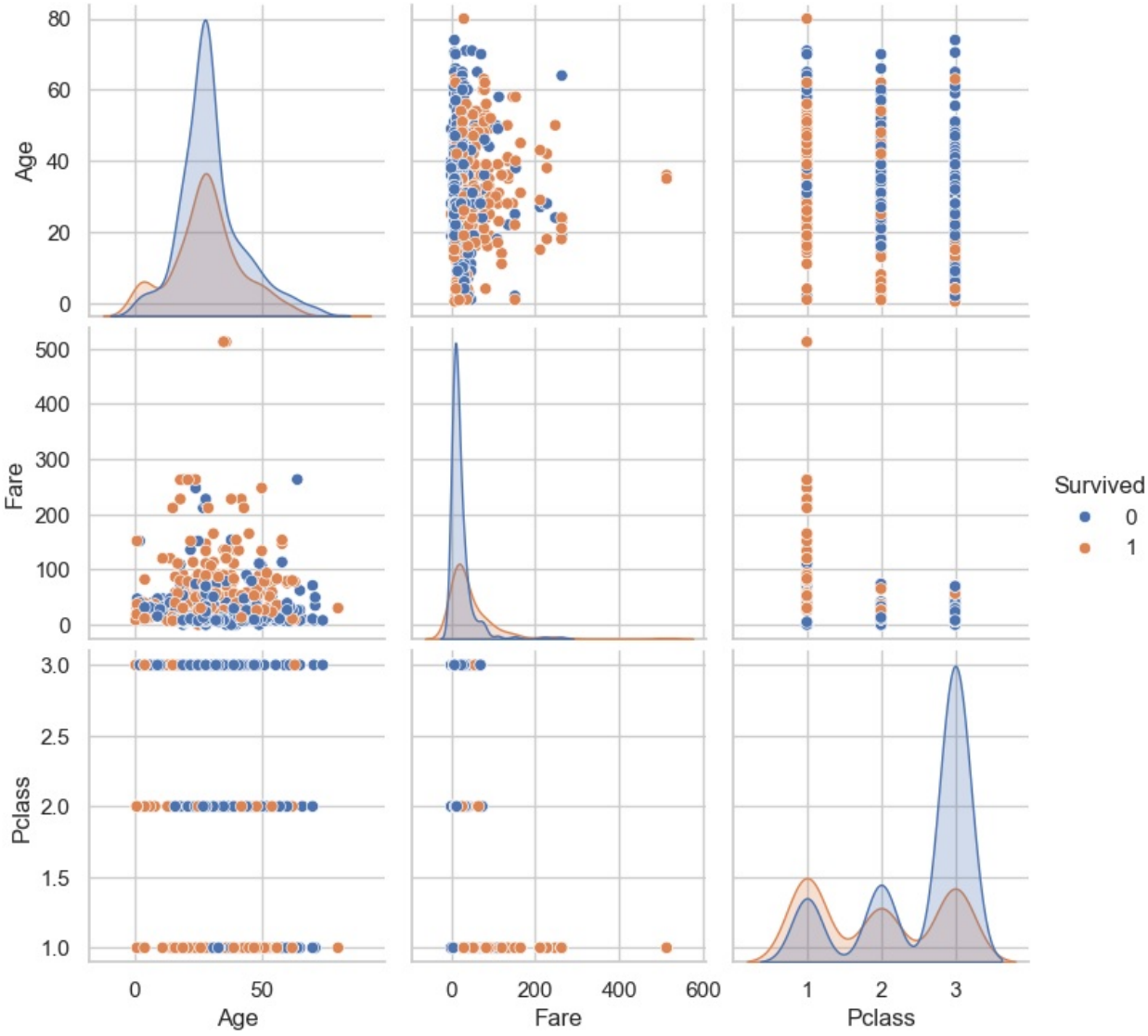
In [35]:
```
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title("Age vs Fare")
plt.show()
```



In [36]:
```
plt.figure(figsize=(8,5))
sns.heatmap(df.corr(numeric_only=True), annot=True, linewidths=1)
plt.title("Correlation Heatmap")
plt.show()
```



In [37]:
```
sns.pairplot(df[['Survived','Age','Fare','Pclass']], hue='Survived')
plt.show()
```



In [38]:
```
df.to_csv('train_analysis.csv',index=False)
```

In [ ]: