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]:

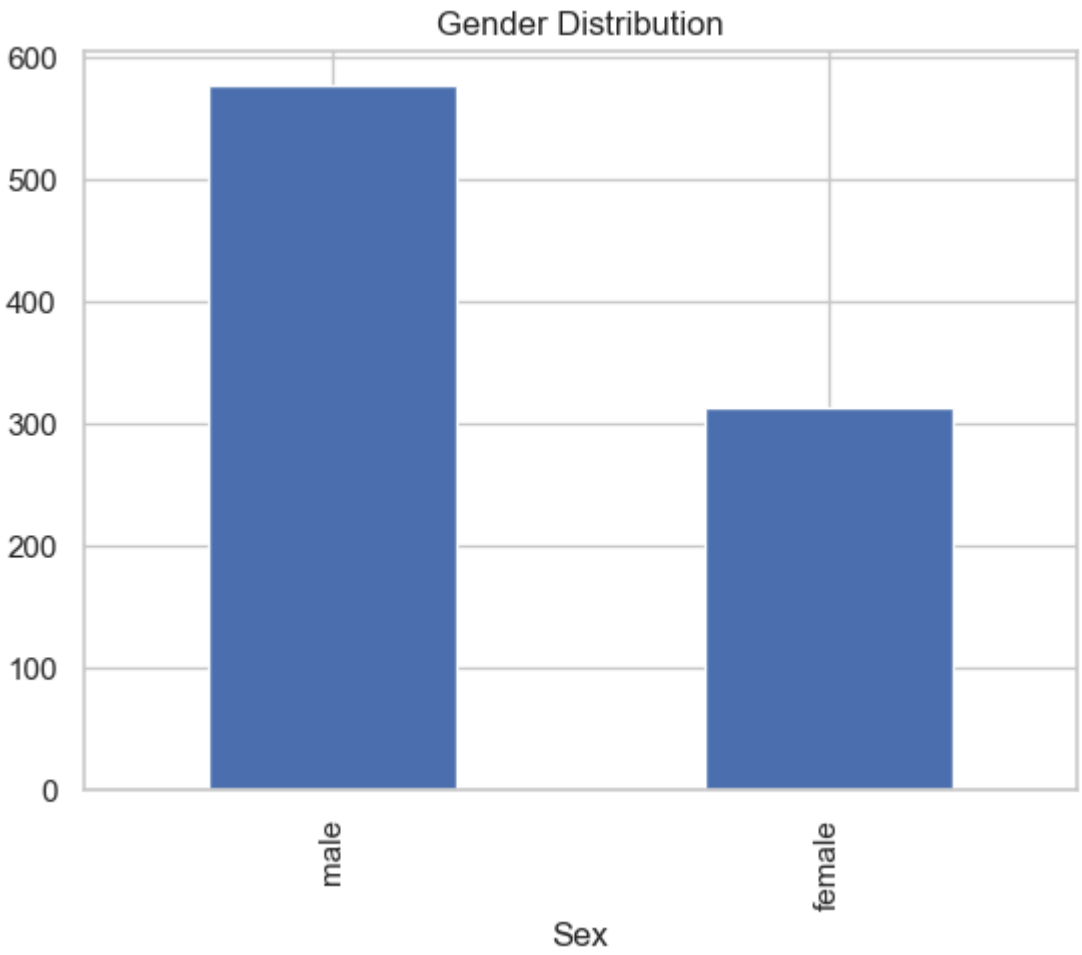
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques 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

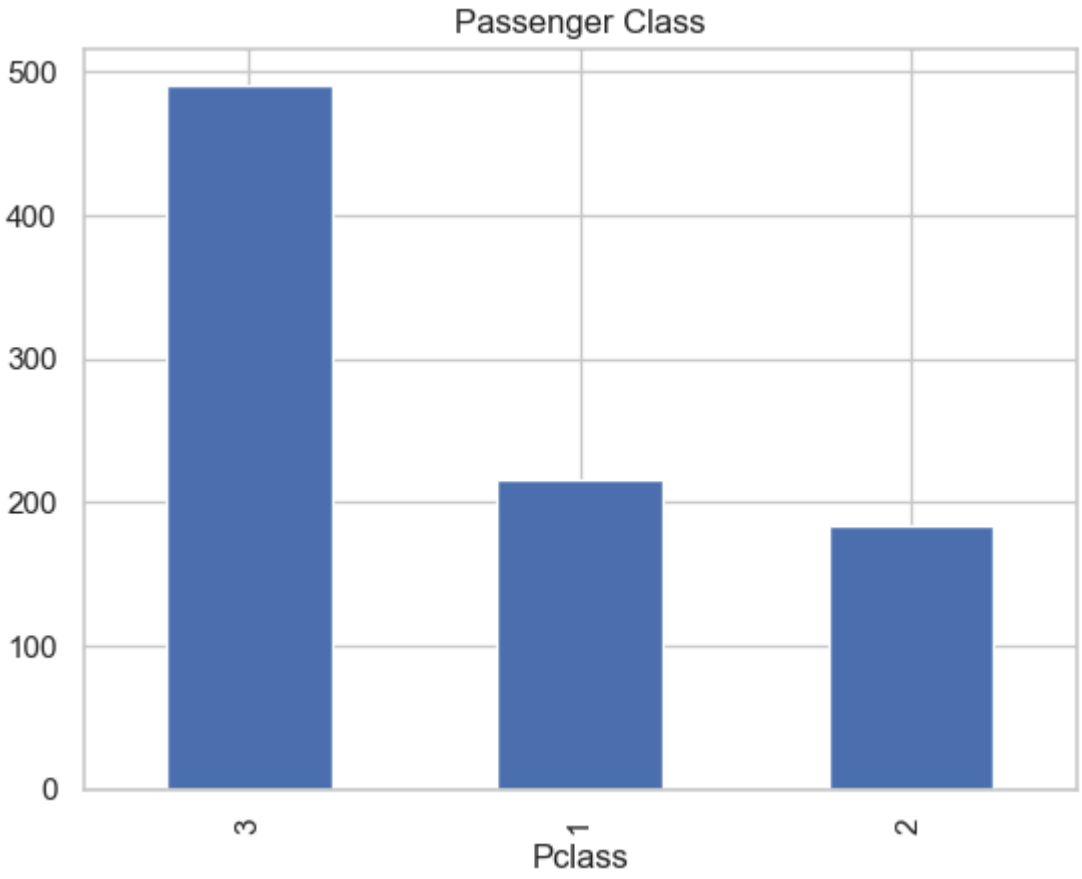
In [3]:
```
df.info()
df.describe()
df.isnull().sum()
df.nunique()
```

```
<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    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
Embarked         3
dtype: int64
```

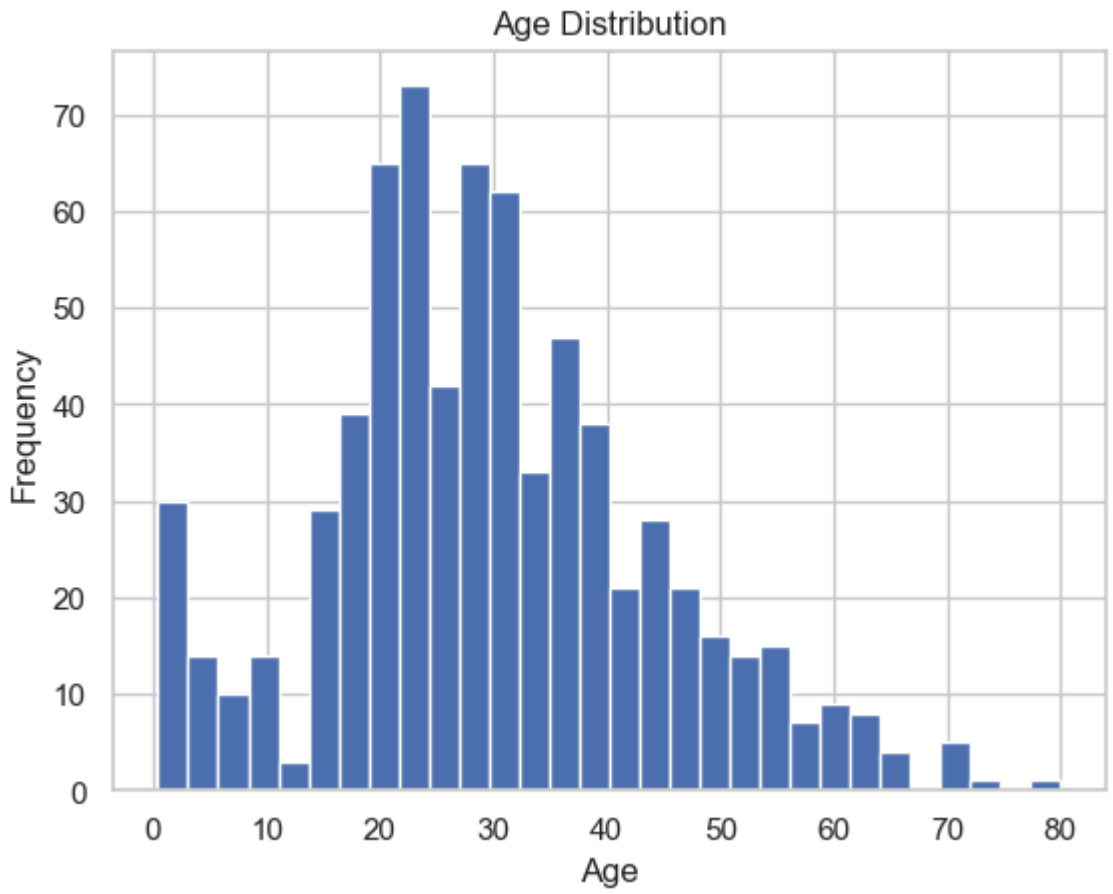
In [4]:
```
df['Sex'].value_counts().plot(kind='bar', title='Gender Distribution')
plt.show()

df['Pclass'].value_counts().plot(kind='bar', title='Passenger Class')
plt.show()
```



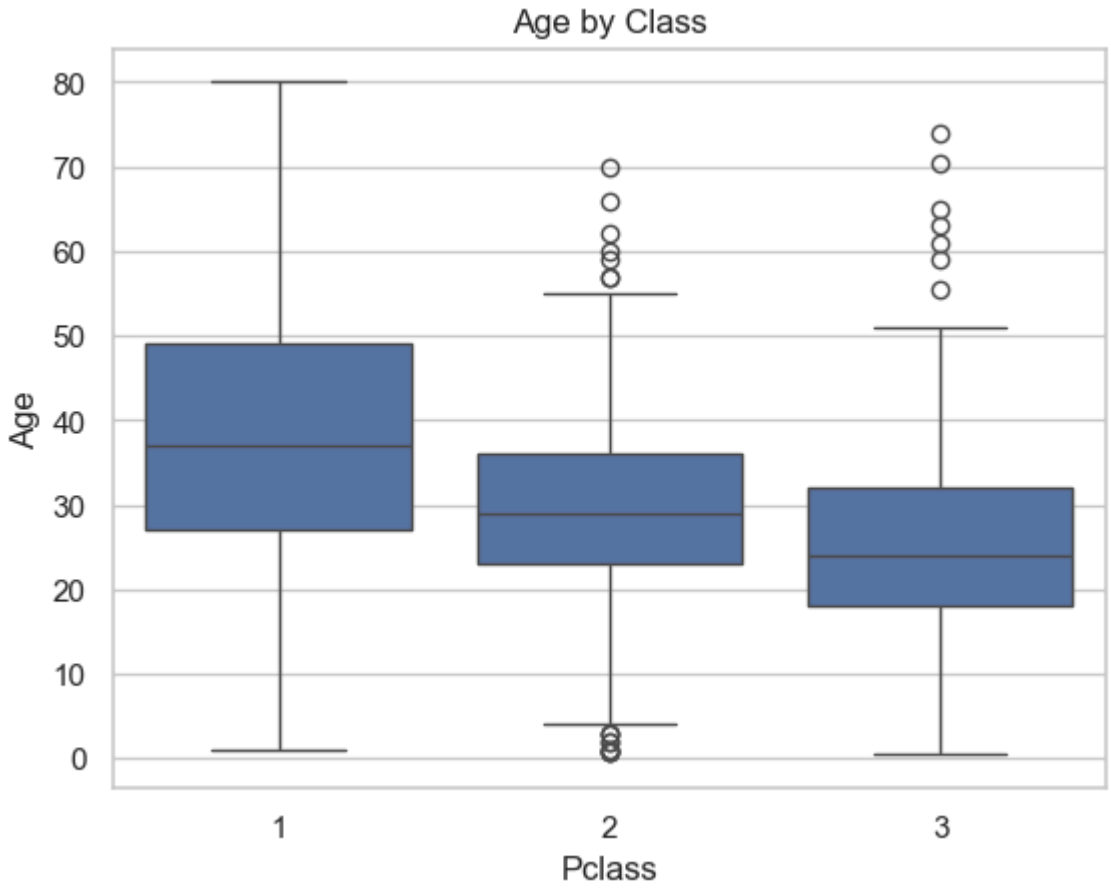


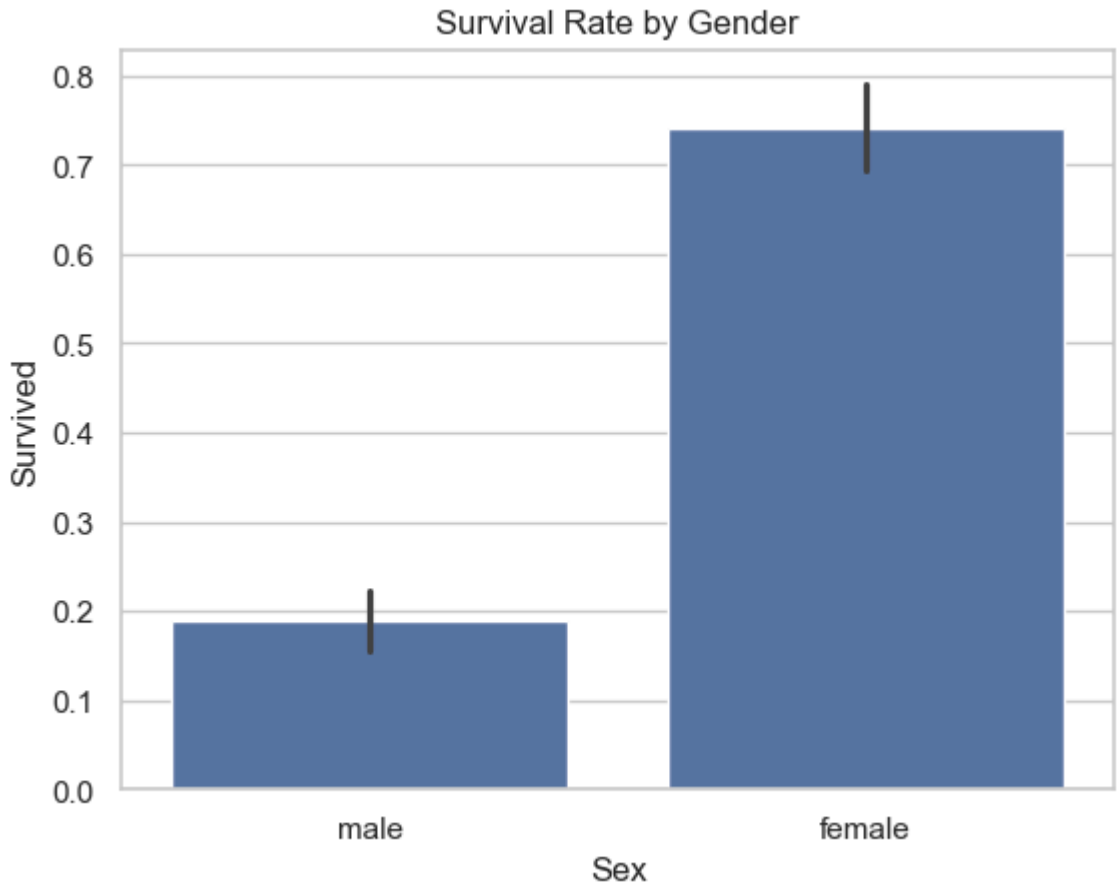
In [5]:
```
df['Age'].hist(bins=30)
plt.title('Age Distribution')
plt.xlabel('Age')
```

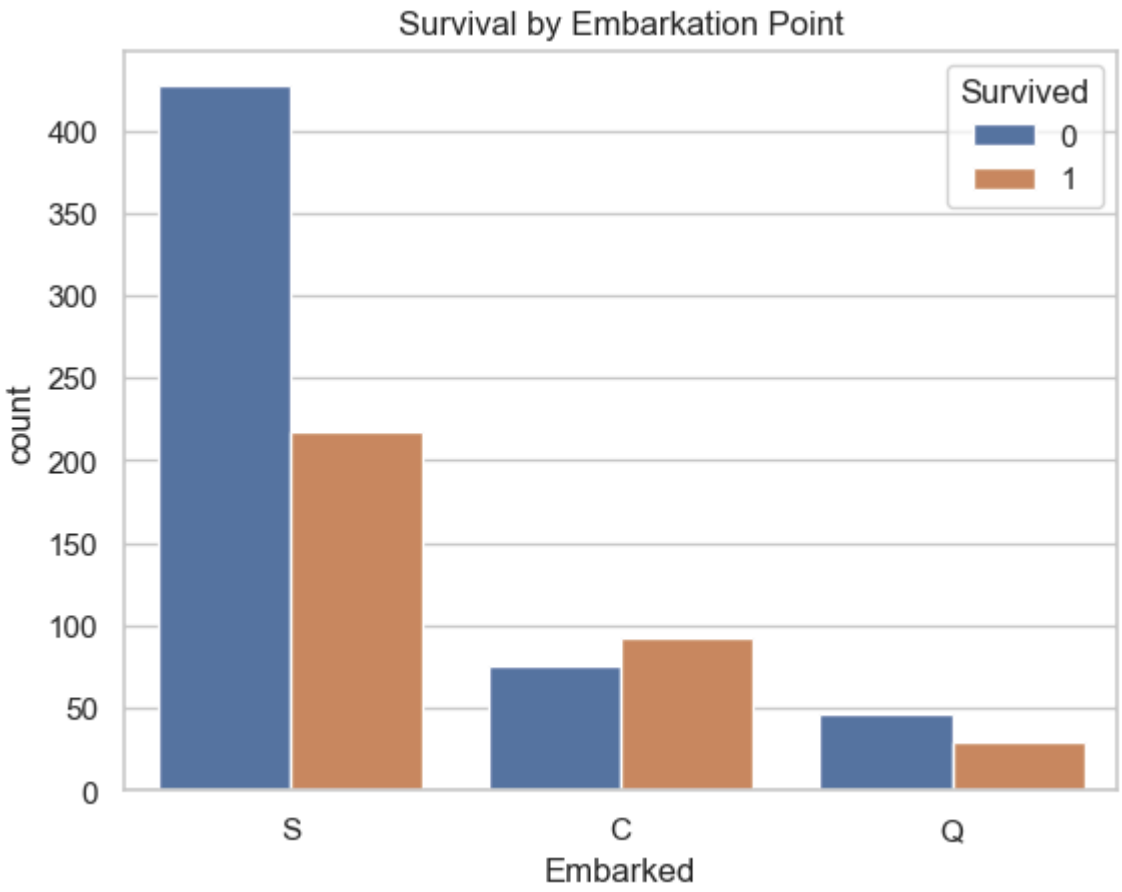
```
plt.ylabel('Frequency')
plt.show()
```



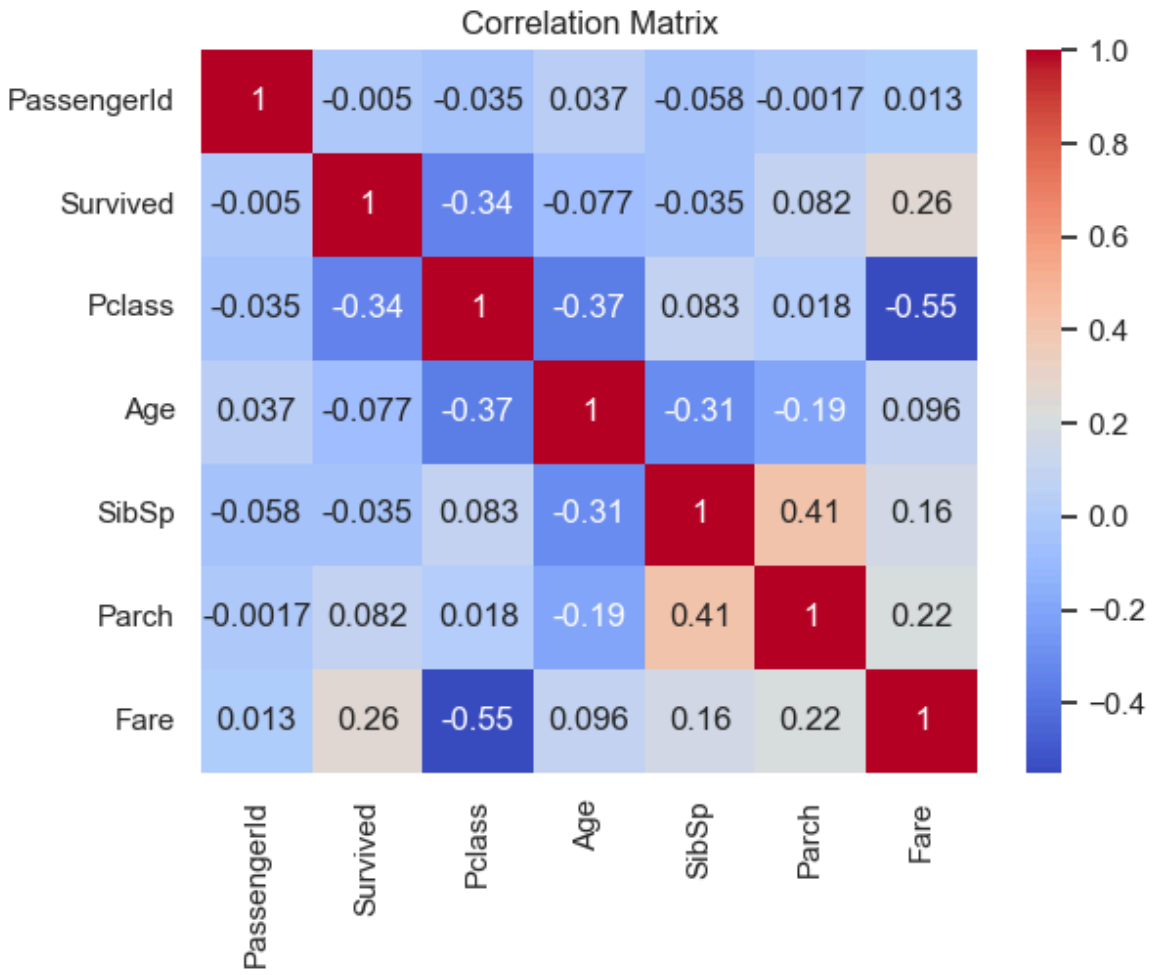
In [6]:
```
sns.boxplot(x='Pclass', y='Age', data=df)
plt.title('Age by Class')
plt.show()

sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Gender')
plt.show()

sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title('Survival by Embarkation Point')
plt.show()
```


Age by Class
Age
80
70
60
50
40
30
20
10
0
1
2
3
Pclass


Survival Rate by Gender
Survived
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
male
female
Sex



In [9]:
```
# Select only numeric columns for correlation
numeric_df = df.select_dtypes(include=['number'])
corr = numeric_df.corr()

# Plot heatmap
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



In [10]:
```
# Summary

- Most survivors were women and children.
- Passengers in 1st class had higher survival rates.
- Age and Fare are somewhat correlated.
- Missing values in Age and Cabin require attention before modeling.
```

```
  Cell In[10], line 4
    - Passengers in 1st class had higher survival rates.
                    ^
SyntaxError: invalid decimal literal
```

In [ ]: