In [1]:

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

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
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/ O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450

In [3]:

```
df.info()
df.describe()
df.columns
df.shape
```

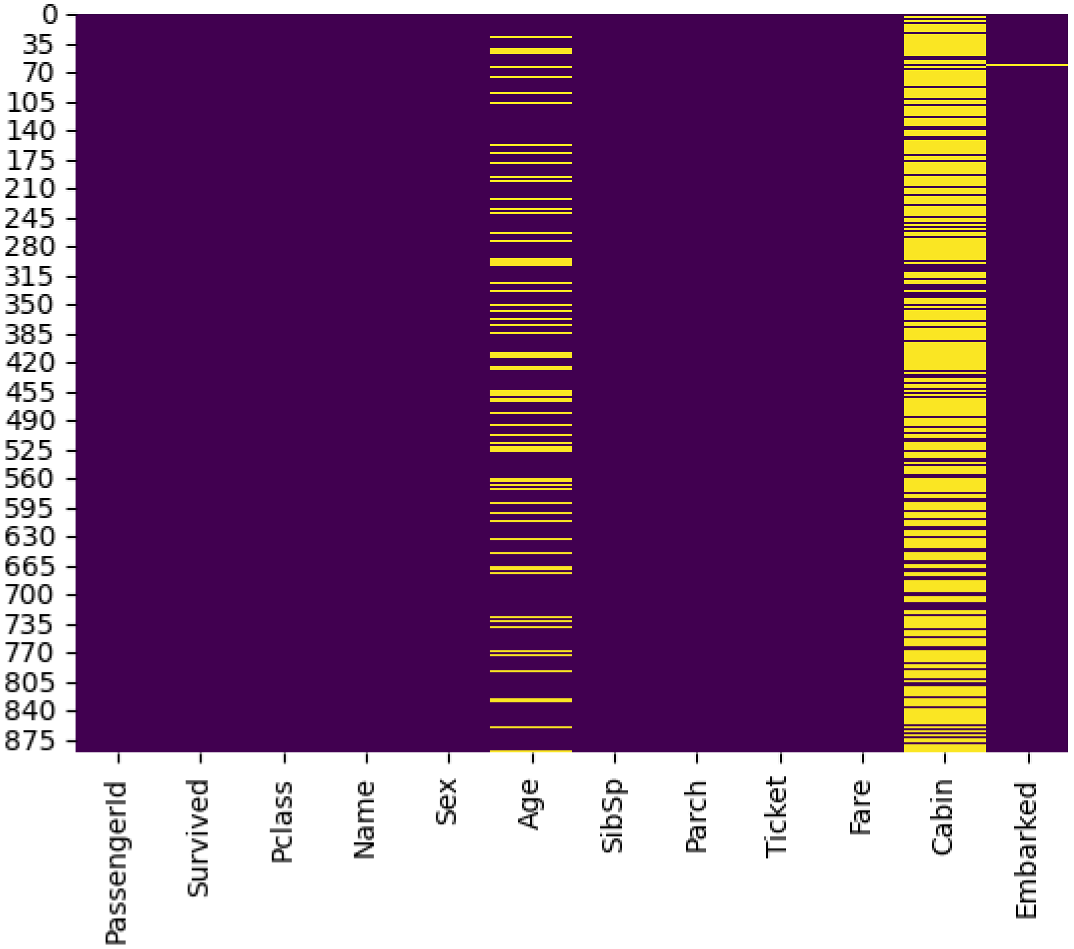
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null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

In [4]:

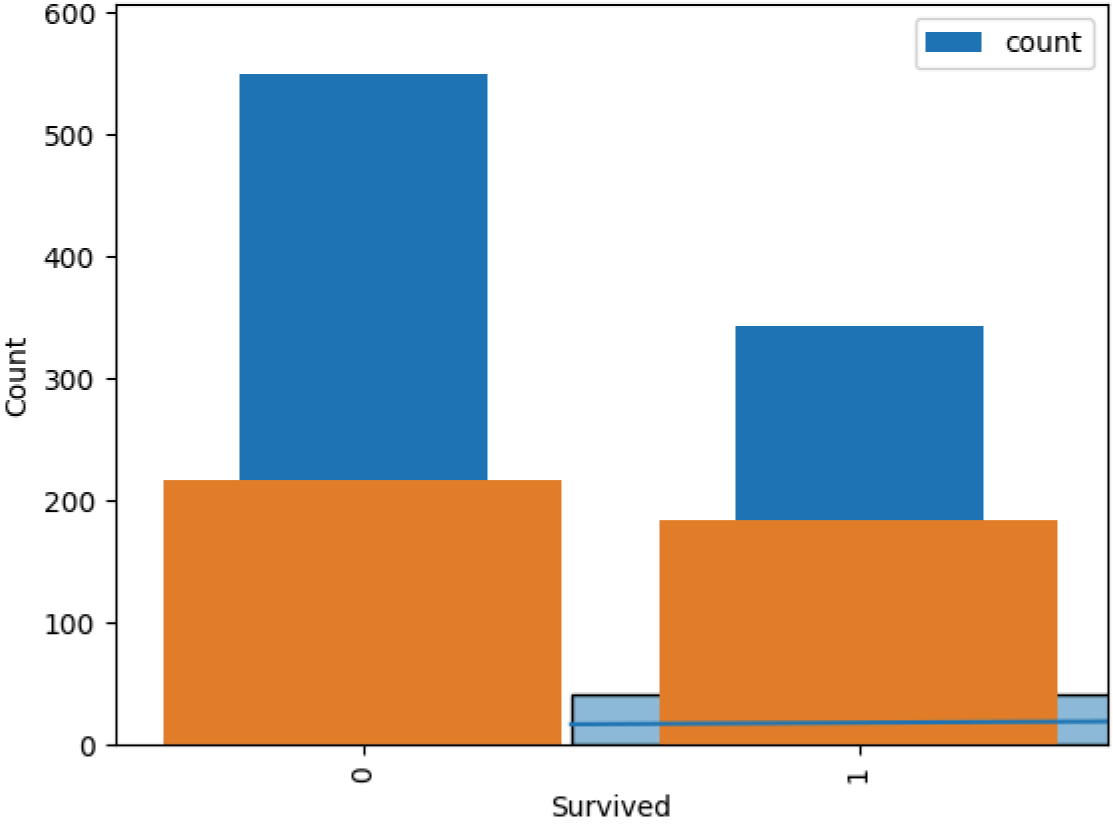
```python
df.isnull().sum()
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
```

```
Out[4]: <Axes: >
```



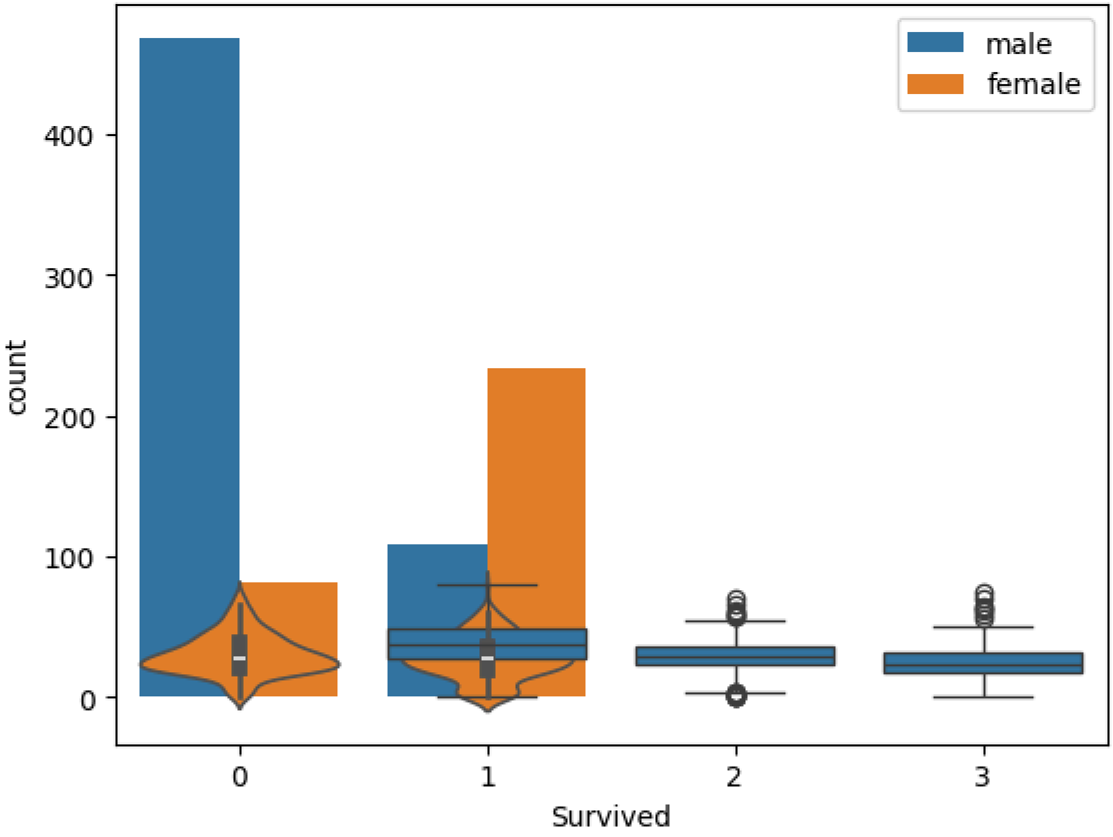
```
In [5]: df['Survived'].value_counts().plot(kind='bar')
        sns.histplot(df['Age'].dropna(), kde=True)
        sns.countplot(x='Pclass', data=df)
        sns.countplot(x='Sex', data=df)
```

```
Out[5]: <Axes: xlabel='Survived', ylabel='Count'>
```



```
In [6]: sns.countplot(x='Survived', hue='Sex', data=df)
        sns.boxplot(x='Pclass', y='Age', data=df)
        sns.violinplot(x='Survived', y='Age', data=df)
        sns.barplot(x='Pclass', y='Survived', data=df)
```

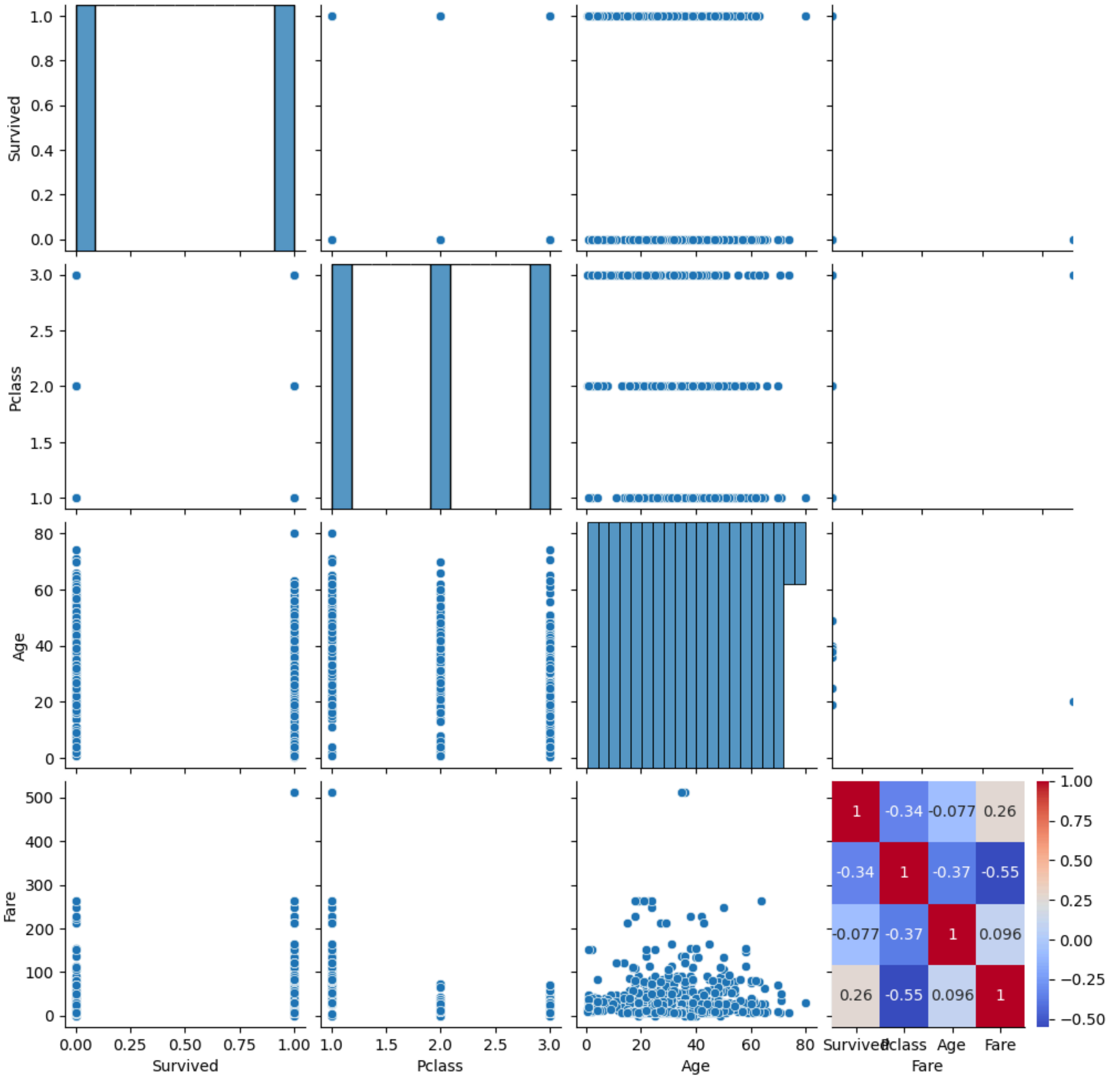
```
Out[6]: <Axes: xlabel='Survived', ylabel='count'>
```



```
In [9]: # Fix: Use only numeric columns
        sns.pairplot(df[['Survived', 'Pclass', 'Age', 'Fare']])

        # Heatmap for correlation (only numeric columns)
        corr_matrix = df[['Survived', 'Pclass', 'Age', 'Fare']].corr()
        sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')

Out[9]: <Axes: >
```



In [10]:

```
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop('Cabin', axis=1, inplace=True)  # Too many missing values
```

Total passengers and survival rate

Survival rate by gender

Survival rate by passenger class

Age and fare distribution

Any anomalies or patterns