In [3]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
```

In [4]:
```python
dt1 = pd.read_excel("titanic.xlsx")
dt1.head(3)
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S

In [5]:
```python
dt1.tail(3)
```

Out[5]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

Analysing the Data

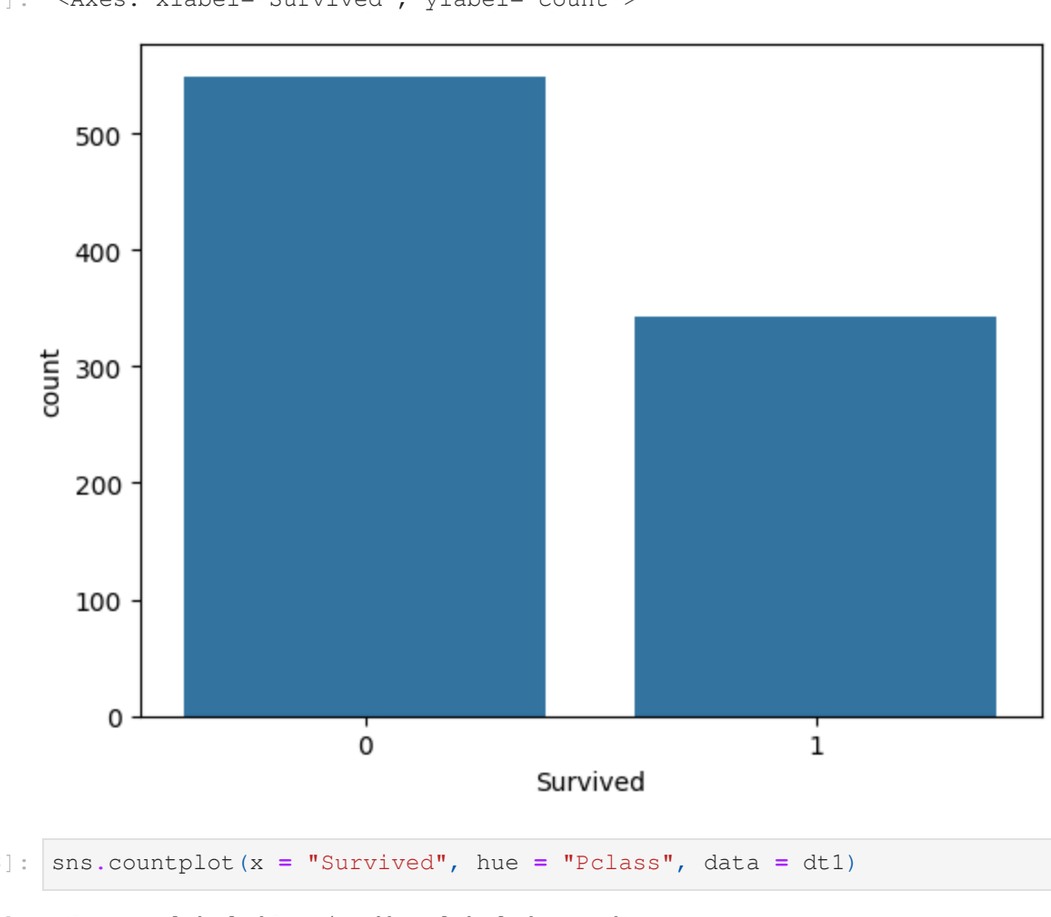
In [6]:
```python
print("# passenger in the original data:" +str(len(dt1.index)))
```

```
# passenger in the original data:891
```

Creating Plots to check the Relationship between Variables

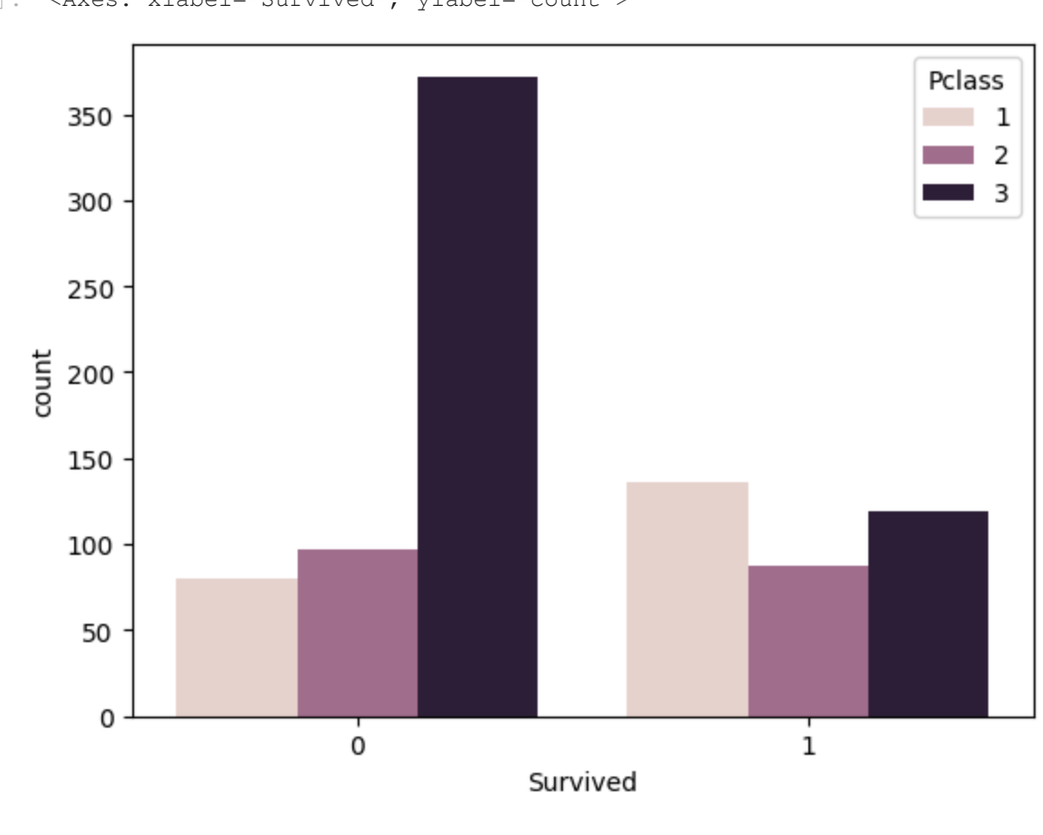
In [7]:
```python
sns.countplot(x = "Survived", data = dt1)
```

Out[7]: <Axes: xlabel='Survived', ylabel='count'>



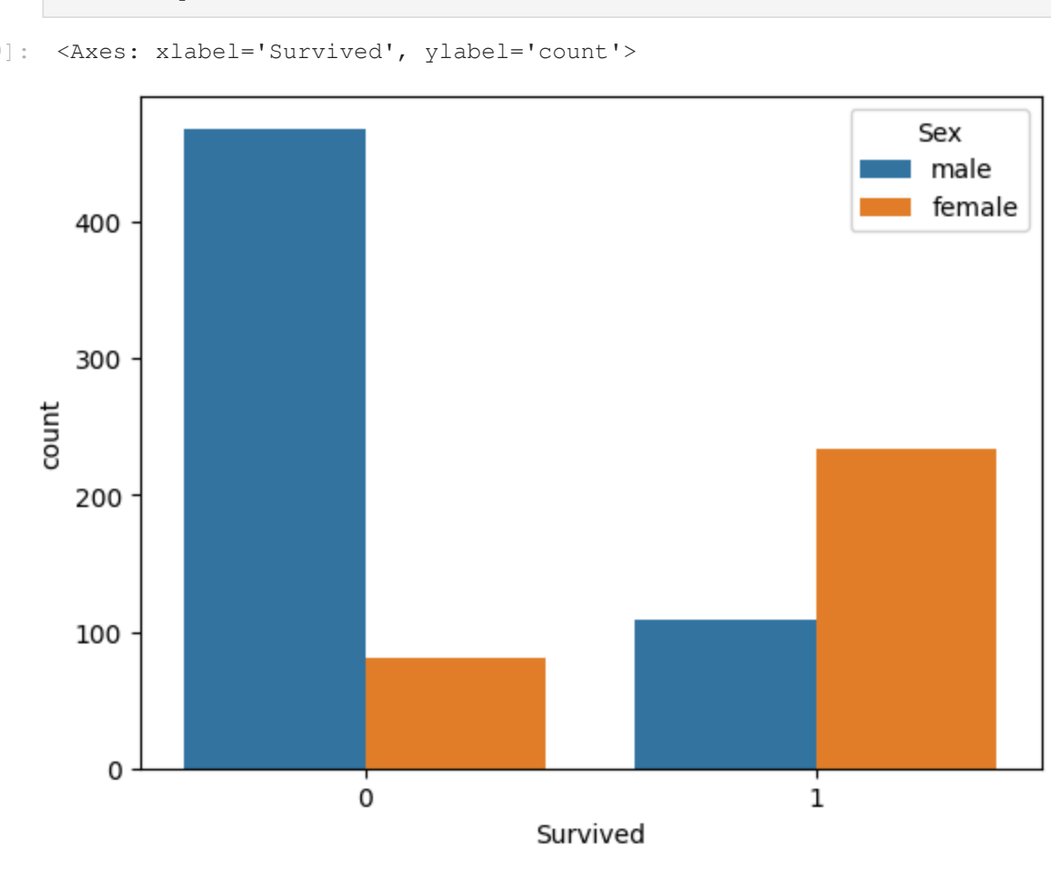
In [8]:
```python
sns.countplot(x = "Survived", hue = "Pclass", data = dt1)
```

Out[8]: <Axes: xlabel='Survived', ylabel='count'>



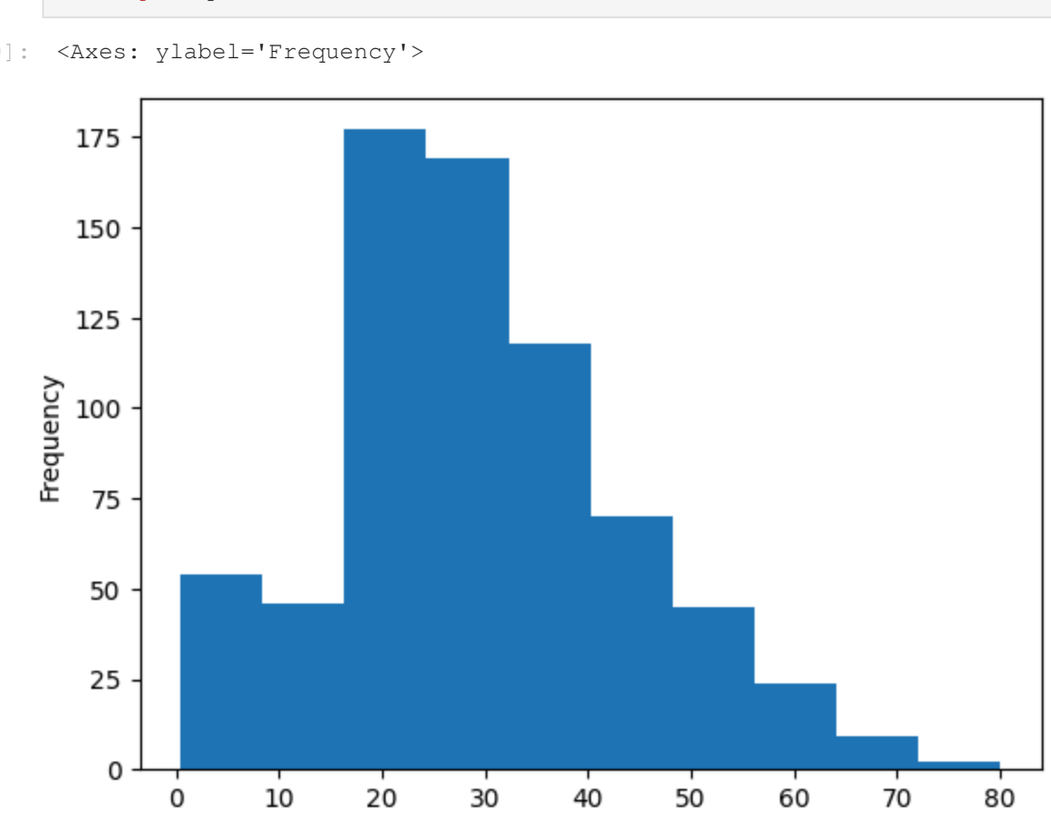
In [9]:
```python
sns.countplot(x = "Survived", hue = "Sex", data = dt1)
```

Out[9]: <Axes: xlabel='Survived', ylabel='count'>



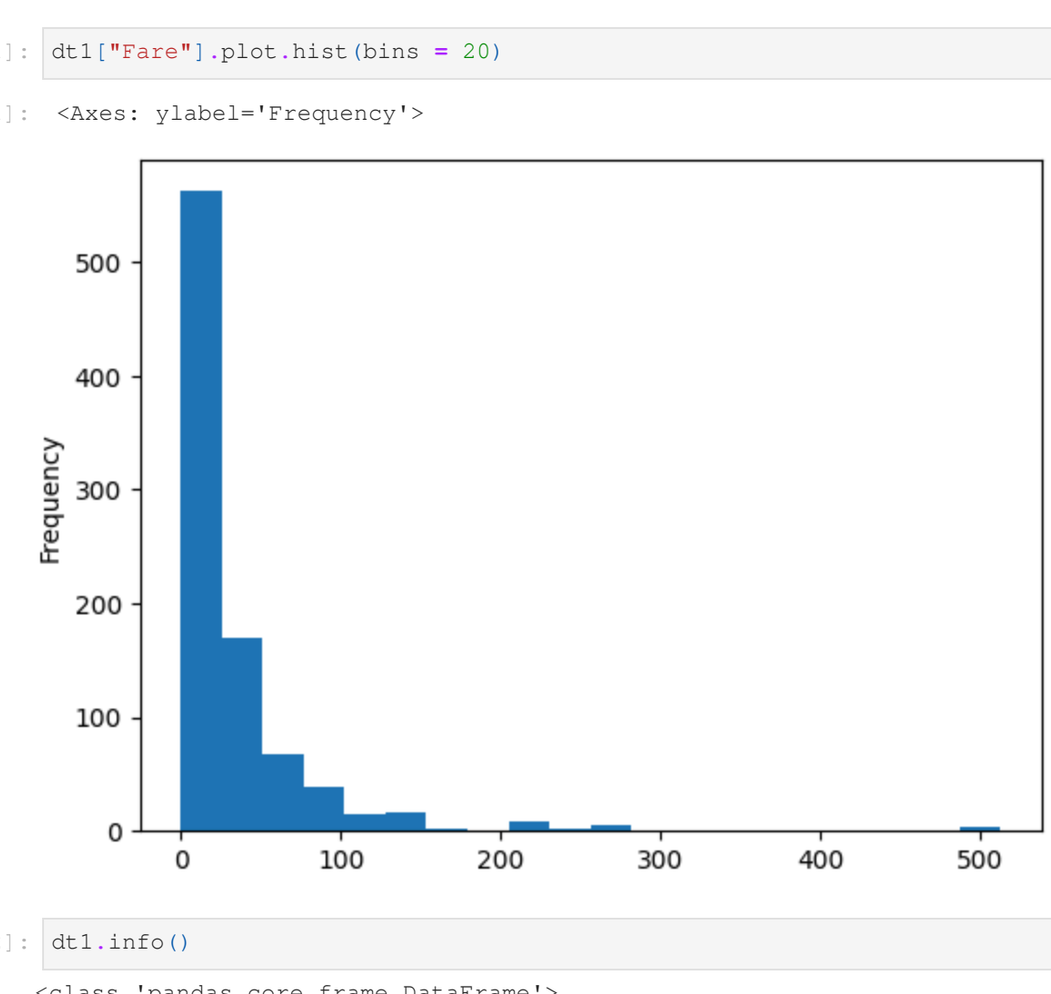
In [10]:
```python
dt1["Age"].plot.hist()
```

Out[10]: <Axes: ylabel='Frequency'>



In [11]:
```python
dt1["Fare"].plot.hist(bins = 20)
```

Out[11]: <Axes: ylabel='Frequency'>

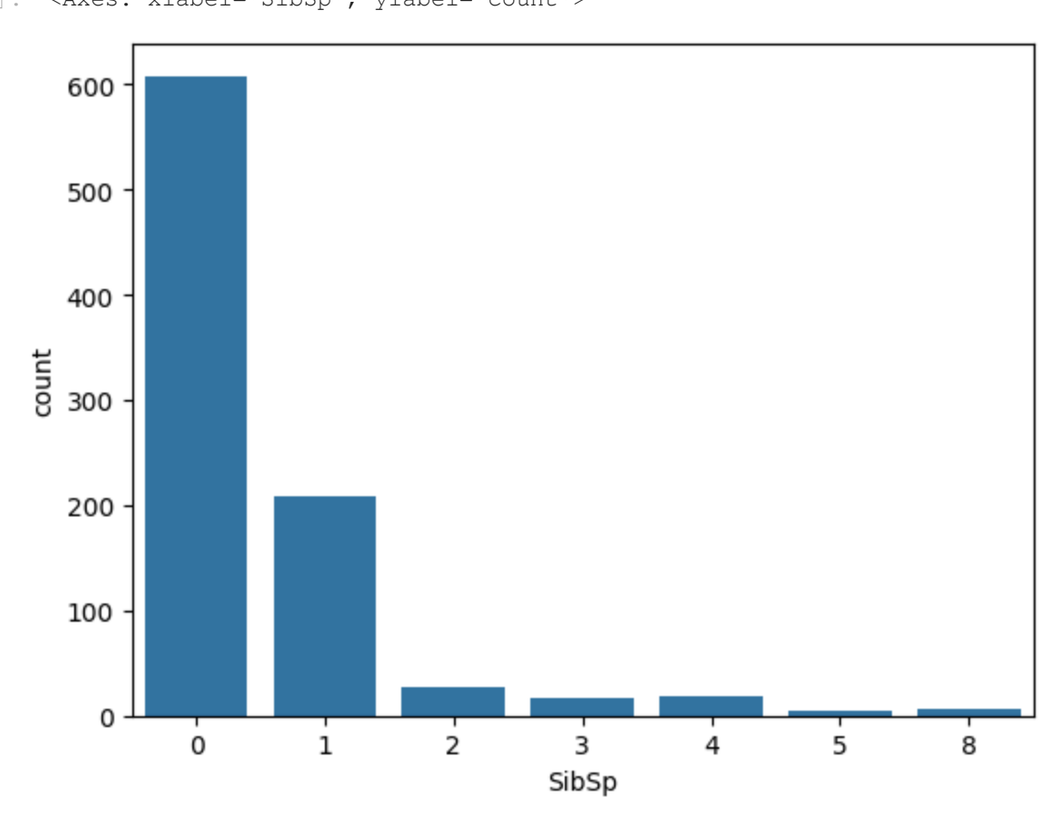


In [12]:
```python
dt1.info()
```

```
<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 [13]:
```python
sns.countplot(x = "SibSp", data = dt1)
```

Out[13]: <Axes: xlabel='SibSp', ylabel='count'>



Data Wrangling or Data Mugging

In [14]:
```python
dt1.isnull()
```

Out[14]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

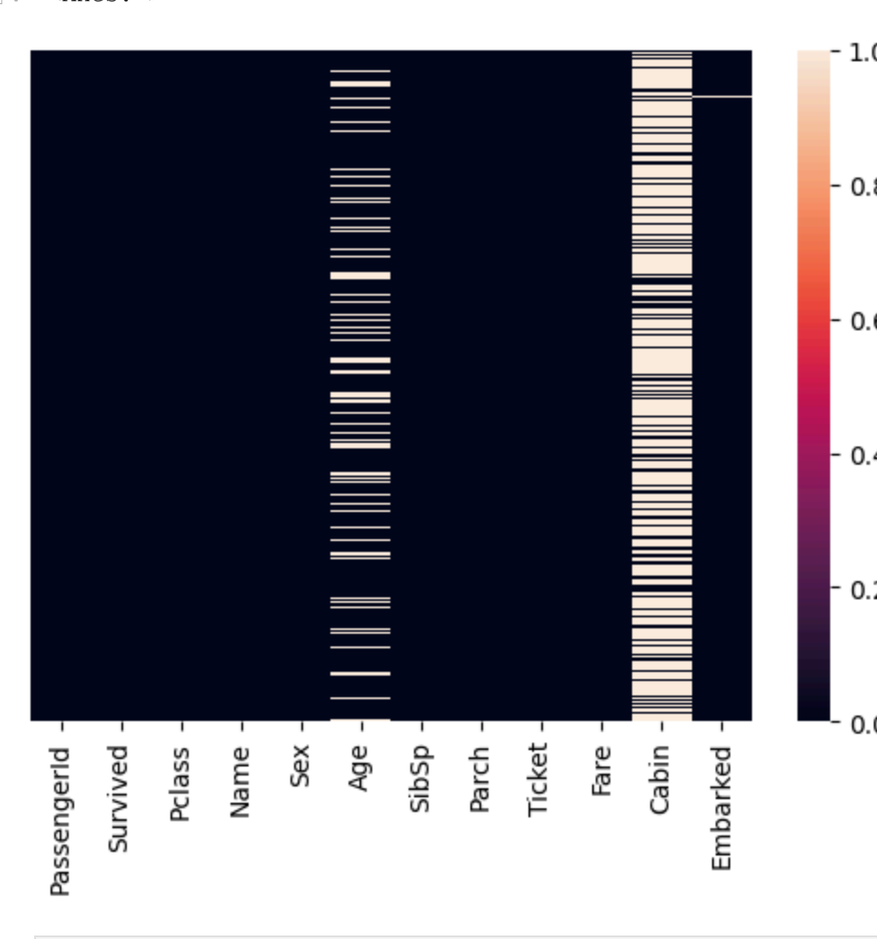
891 rows × 12 columns

In [15]:
```python
dt1.isnull().sum()
```

Out[15]:
```
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
```

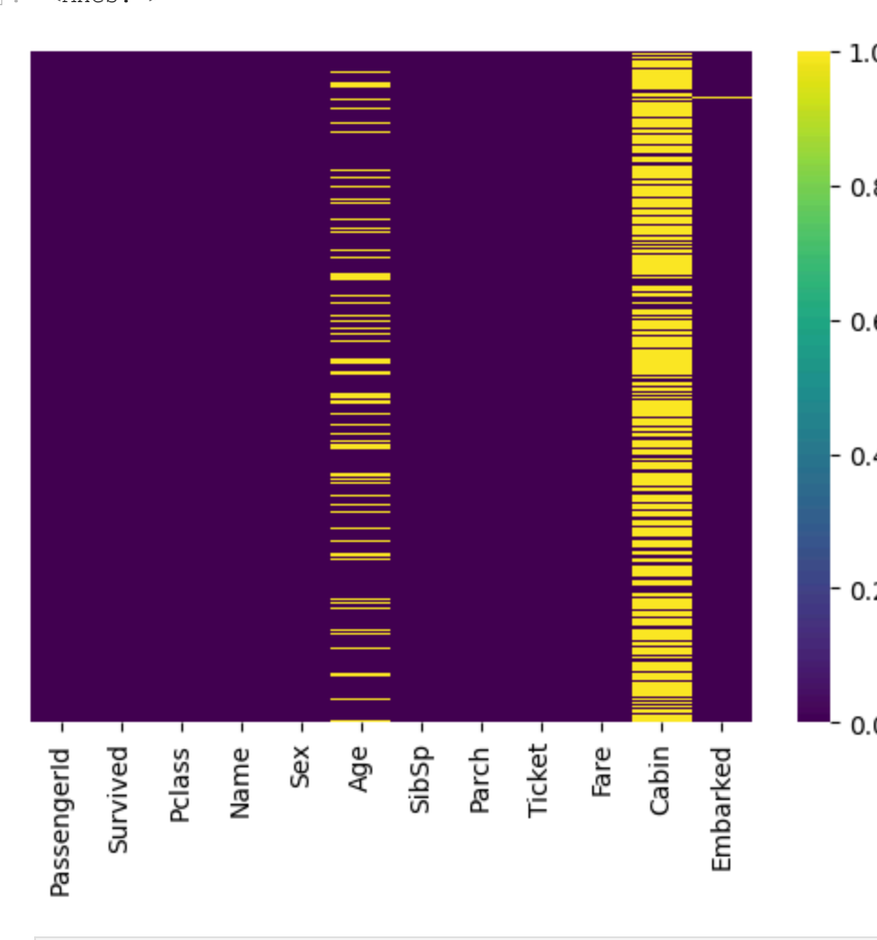
In [16]:
```python
sns.heatmap(dt1.isnull(),yticklabels=False)
```

Out[16]: <Axes: >



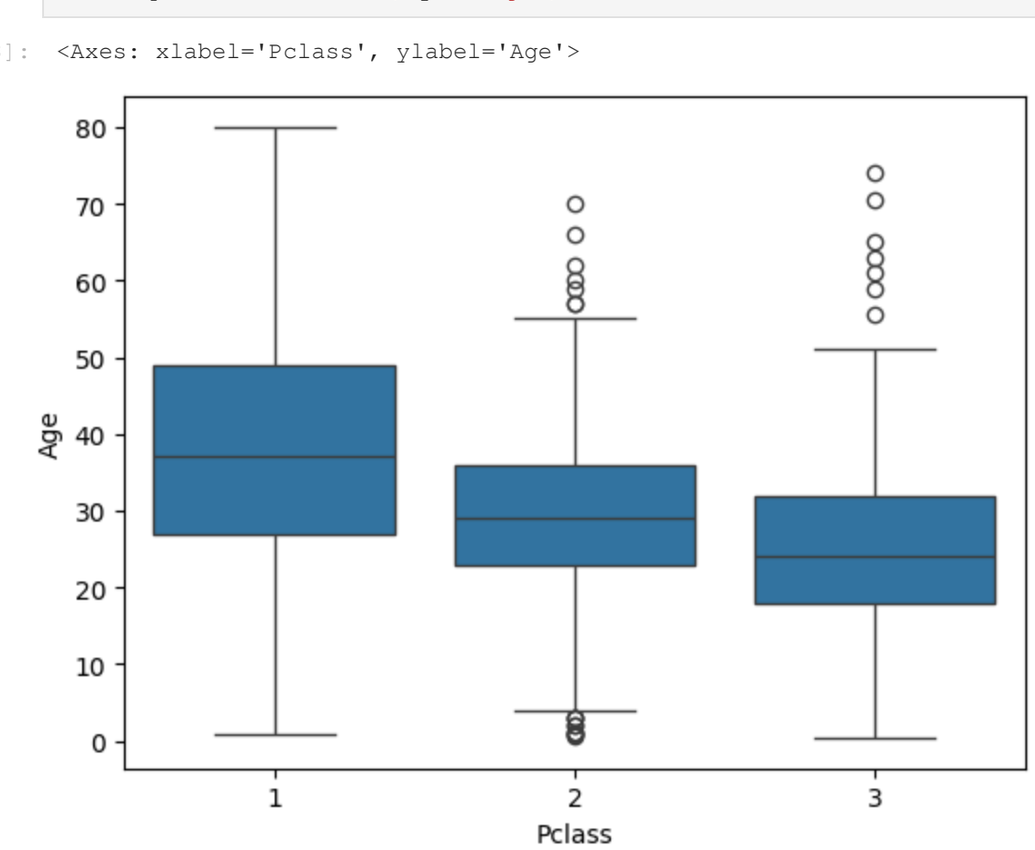
In [17]:
```python
sns.heatmap(dt1.isnull(),yticklabels=False, cmap="viridis")
```

Out[17]: <Axes: >



In [18]:
```python
sns.boxplot(x = "Pclass", y = "Age", data = dt1)
```

Out[18]: <Axes: xlabel='Pclass', ylabel='Age'>



In [19]:
```python
dt1.columns
```

Out[19]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [20]:
```python
dt1.Survived.value_counts()
```

Out[20]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

In [21]:
```python
dt1.Embarked.value_counts()
```

Out[21]:
```
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

In [22]:
```python
dt1.Pclass.value_counts()
```

Out[22]:
```
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

In [23]:
```python
dt1.Sex.value_counts()
```

Out[23]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [24]:
```python
dt2 = dt1.copy()
```

In [25]:
```python
dt2.head(2)
```

Out[25]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C

In [26]:
```python
dt2.drop("Cabin",axis = 1, inplace = True)
```

In [27]:
```python
dt2.head(3)
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S

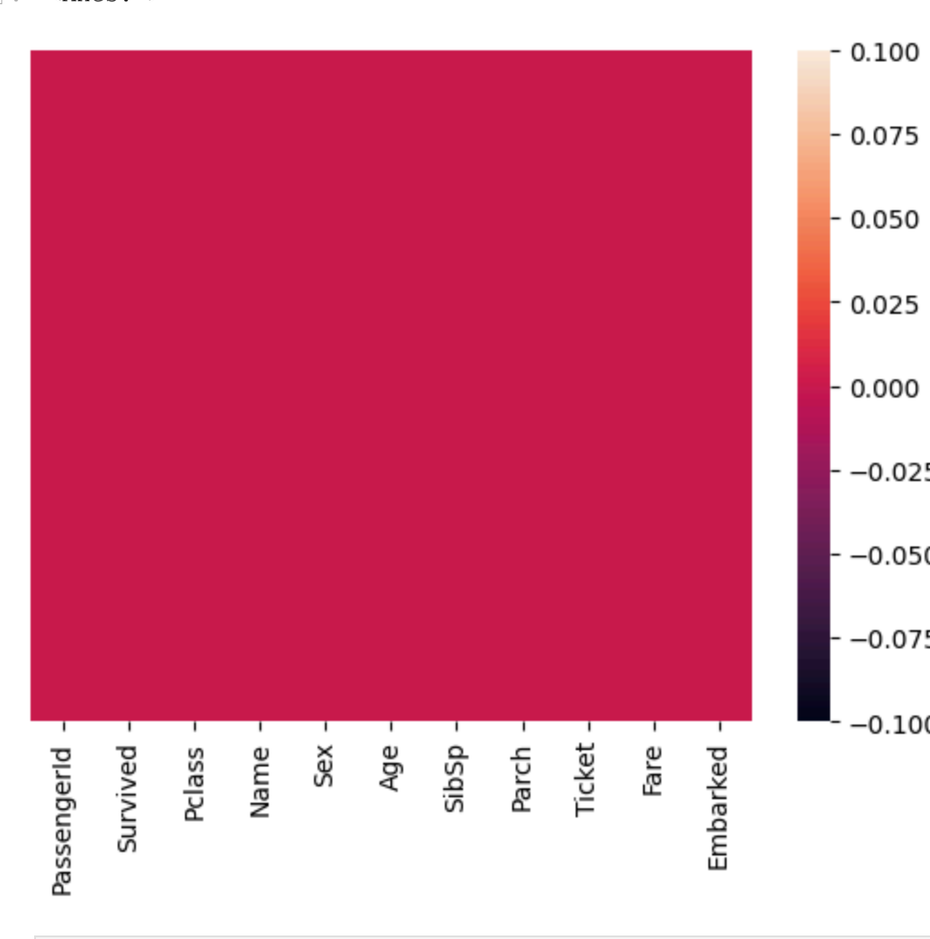
In [28]:
```python
dt2.dropna(inplace=True)
```

In [29]:
```python
dt2.isnull().sum()
```

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

In [30]:
```python
sns.heatmap(dt2.isnull(),yticklabels=False)
```

Out[30]: <Axes: >



In [31]:
```python
sex = pd.get_dummies(dt2["Sex"],drop_first=True)
```

In [32]:
```python
sex
```

Out[32]:

	male
0	True
1	False
2	False
3	False
4	True
...	...
885	False
886	True
887	False
889	True
890	True

712 rows × 1 columns

In [33]:
```python
embark = pd.get_dummies(dt2["Embarked"],drop_first=True)
```

In [34]:
```python
embark
```

Out[34]:

	Q	S
0	False	True
1	False	False
2	False	True
3	False	True
4	False	True
...	...	...
885	True	False
886	False	True
887	False	True
889	False	False
890	True	False

712 rows × 2 columns

In [35]:
```python
pls = pd.get_dummies(dt2["Pclass"],drop_first=True)
```

In [36]:
```python
pls
```

Out[36]:

	2	3
0	False	True
1	False	False
2	False	True
3	False	False
4	False	True
...	...	...
885	False	True
886	True	False
887	False	False
889	False	False
890	False	True

712 rows × 2 columns

In [37]:
```python
dt2.head(3)
```

Out[37]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S

In [38]:
```python
dt2 = pd.concat([dt2,sex,embark,pls],axis = 1)
```

In [39]:
```python
dt2.head(3)
```

Out[39]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	male	Q	S	2	3
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S	True	False	True	False	True
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C	False	False	False	False	False
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S	False	False	True	False	True

In [40]:
```python
dt2.drop("Embarked",axis = 1, inplace = True)
```

In [41]:
```python
dt2.drop("Sex",axis = 1, inplace = True)
```

In [42]:
```python
dt2.drop("Pclass",axis = 1, inplace = True)
```

In [43]:
```python
dt2.drop("PassengerId",axis = 1, inplace = True)
```

In [44]:
```python
dt2.drop("Ticket",axis = 1, inplace = True)
```

In [45]:
```python
dt2.drop("Fare",axis = 1, inplace = True)
```

In [46]:
```python
dt2.head(3)
```

Out[46]:

	Survived	Name	Age	SibSp	Parch	male	Q	S	2	3
0	0	Braund, Mr. Owen Harris	22.0	1	0	True	False	True	False	True
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	1	0	False	False	False	False	False
2	1	Heikkinen, Miss. Laina	26.0	0	0	False	False	True	False	True

In [47]:
```python
dt2.drop("SibSp",axis = 1, inplace = True)
```

In [49]:
```python
dt2.drop("Parch",axis = 1, inplace = True)
```

In [50]:
```python
dt2.head(3)
```

Out[50]:

	Survived	Name	Age	male	Q	S	2	3
0	0	Braund, Mr. Owen Harris	22.0	True	False	True	False	True
1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	False	False	False	False	False
2	1	Heikkinen, Miss. Laina	26.0	False	False	True	False	True

In [51]:
```python
dt2.drop("Name",axis = 1, inplace = True)
```

In [52]:
```python
dt2
```

Out[52]:

	Survived	Age	male	Q	S	2	3
0	0	22.0	True	False	True	False	True
1	1	38.0	False	False	False	False	False
2	1	26.0	False	False	True	False	True
3	1	35.0	False	False	True	False	False
4	0	35.0	True	False	True	False	True
...	...	...	...	...	...	...	...
885	0	39.0	False	True	False	False	True
886	0	27.0	True	False	True	True	False
887	1	19.0	False	False	True	False	False
889	1	26.0	True	False	False	False	False
890	0	32.0	True	True	False	False	True

712 rows × 7 columns