Titanic EDA

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
data = pd.read_csv('./titanic.csv')
data.head(10)
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708

In [3]:

```
data.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]:

```
data.describe()
```

Out[4]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [5]:

```
data.shape
```

Out[5]:

```
(891, 12)
```

In [6]:

```
data.isnull().mean()*100
```

Out[6]:

```
PassengerId     0.000000
Survived        0.000000
Pclass          0.000000
Name            0.000000
Sex             0.000000
Age            19.865320
SibSp           0.000000
Parch           0.000000
Ticket          0.000000
Fare            0.000000
Cabin          77.104377
Embarked        0.224467
dtype: float64
```

In [7]:

```
data.drop(columns='Cabin', inplace=True)
```

In [8]:

```
data.drop_duplicates(inplace=True)
```

In [9]:

```
data.dropna(inplace=True)
```

In [10]:

```
data.shape
```

Out[10]:

```
(712, 11)
```

EDA

In [11]:

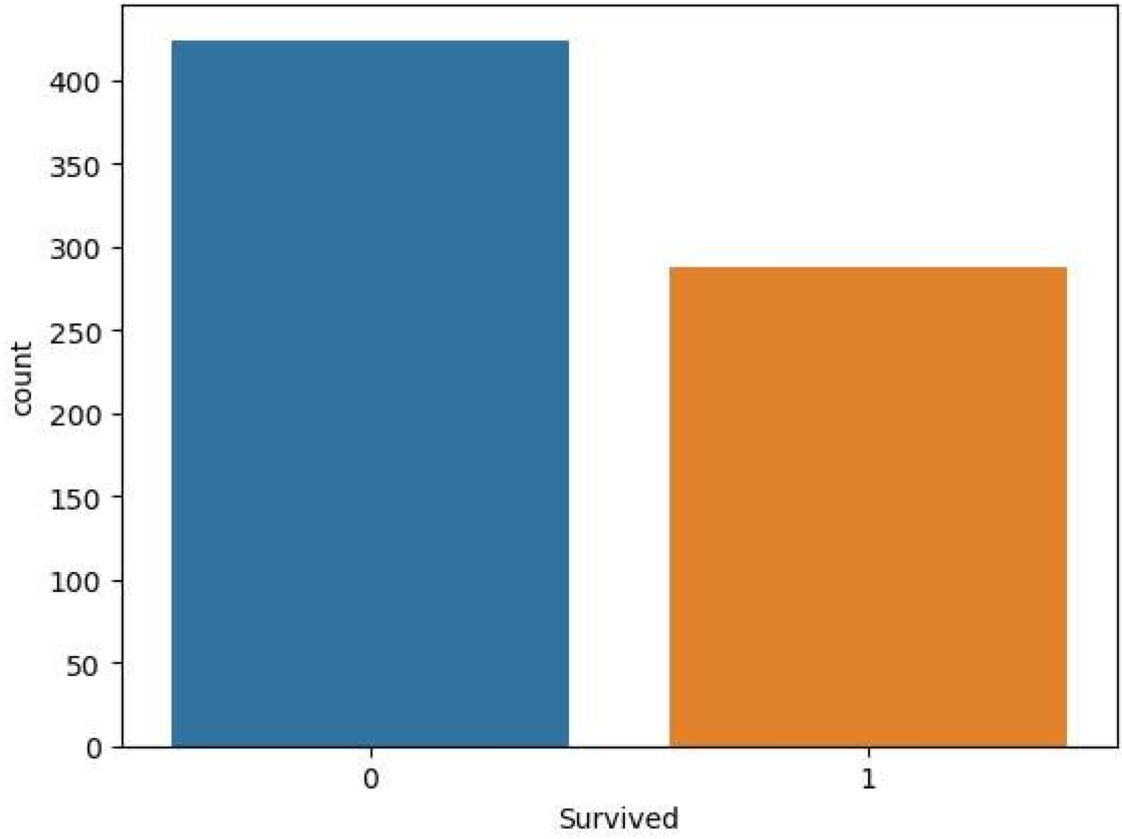
```
data.columns
```

Out[11]:

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

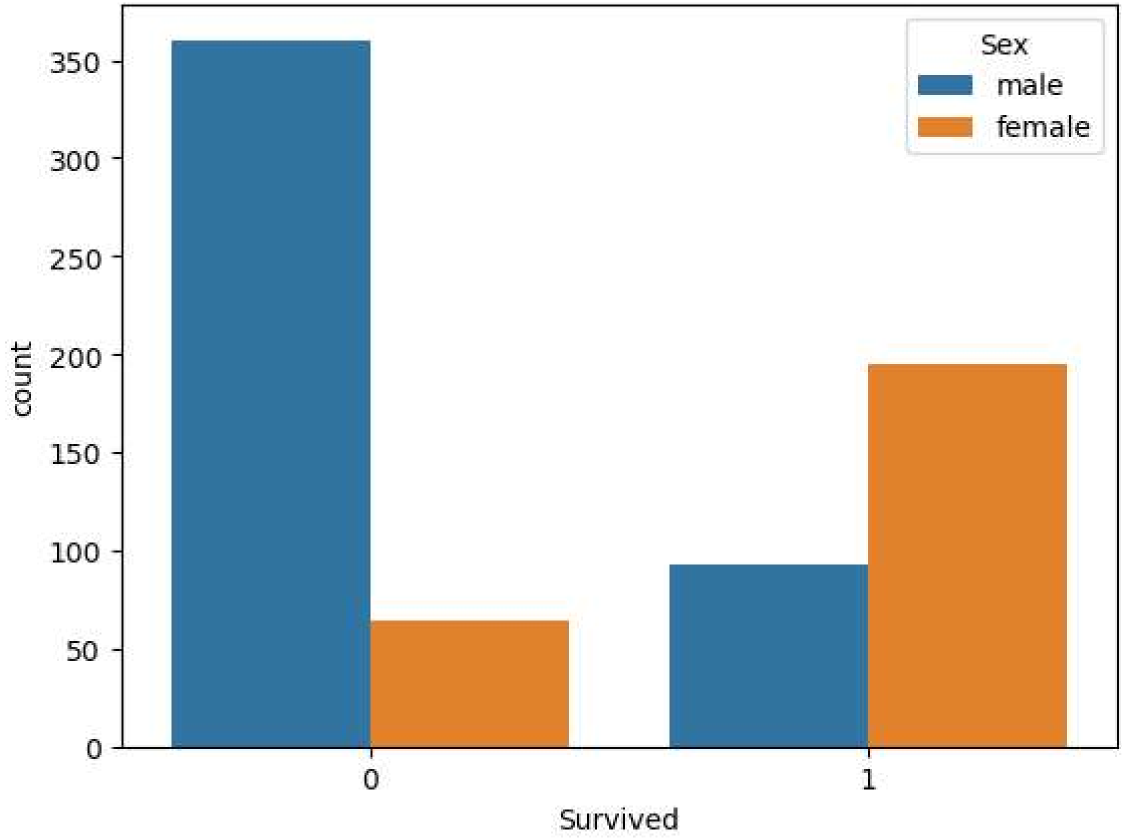
In [12]:

```
sns.countplot(x='Survived', data=data);
```



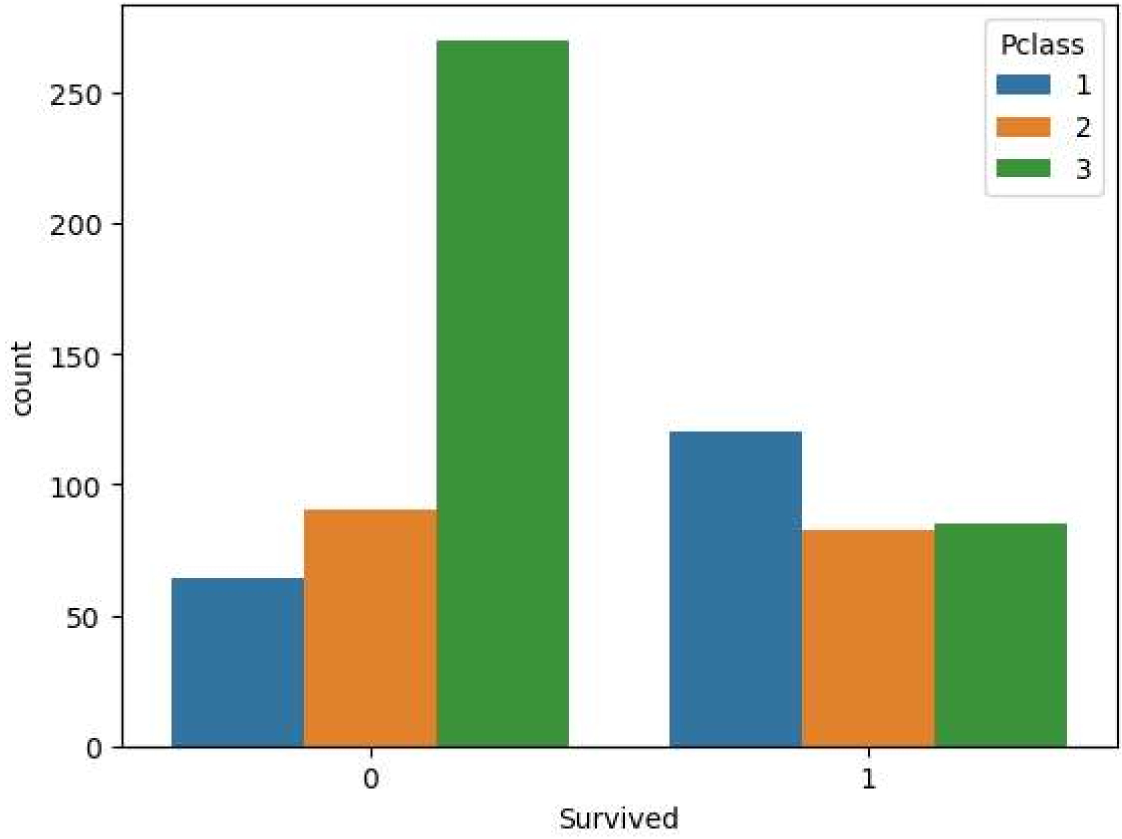
In [13]:

```
sns.countplot(x='Survived', data=data, hue='Sex');
```



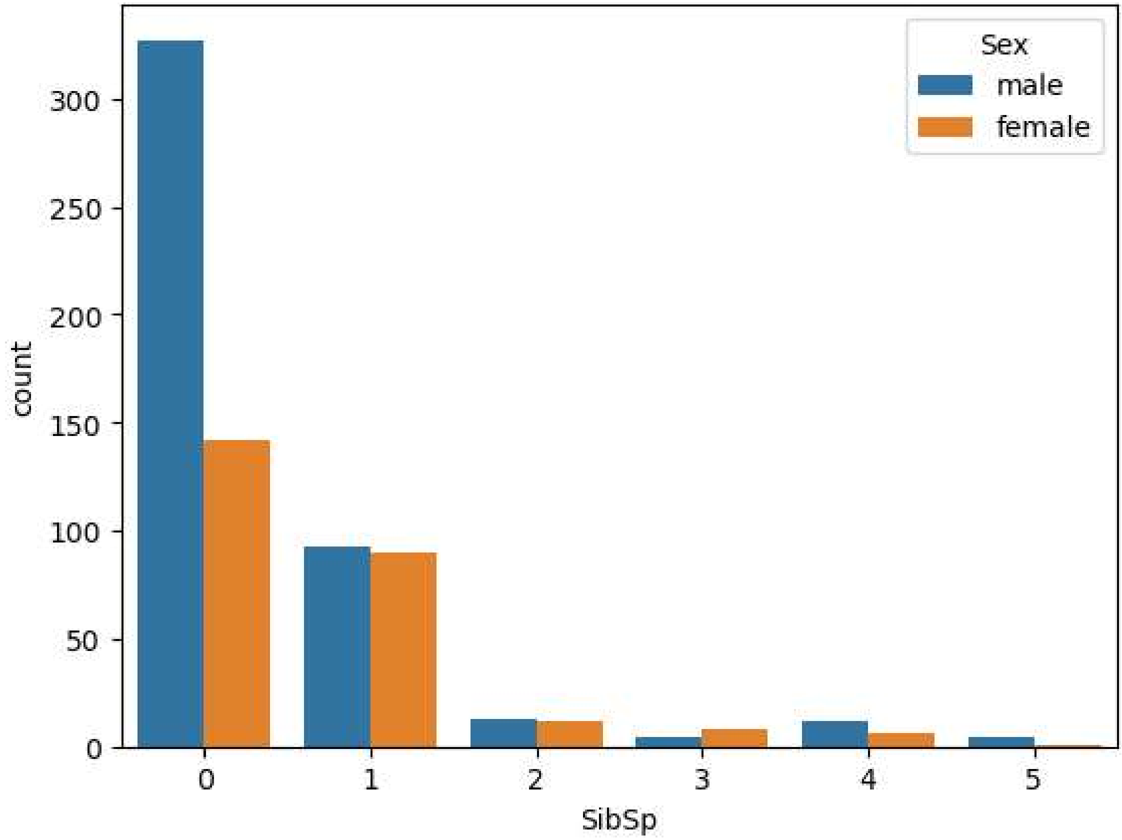
In [14]:

```
sns.countplot(x='Survived', data=data, hue='Pclass');
```



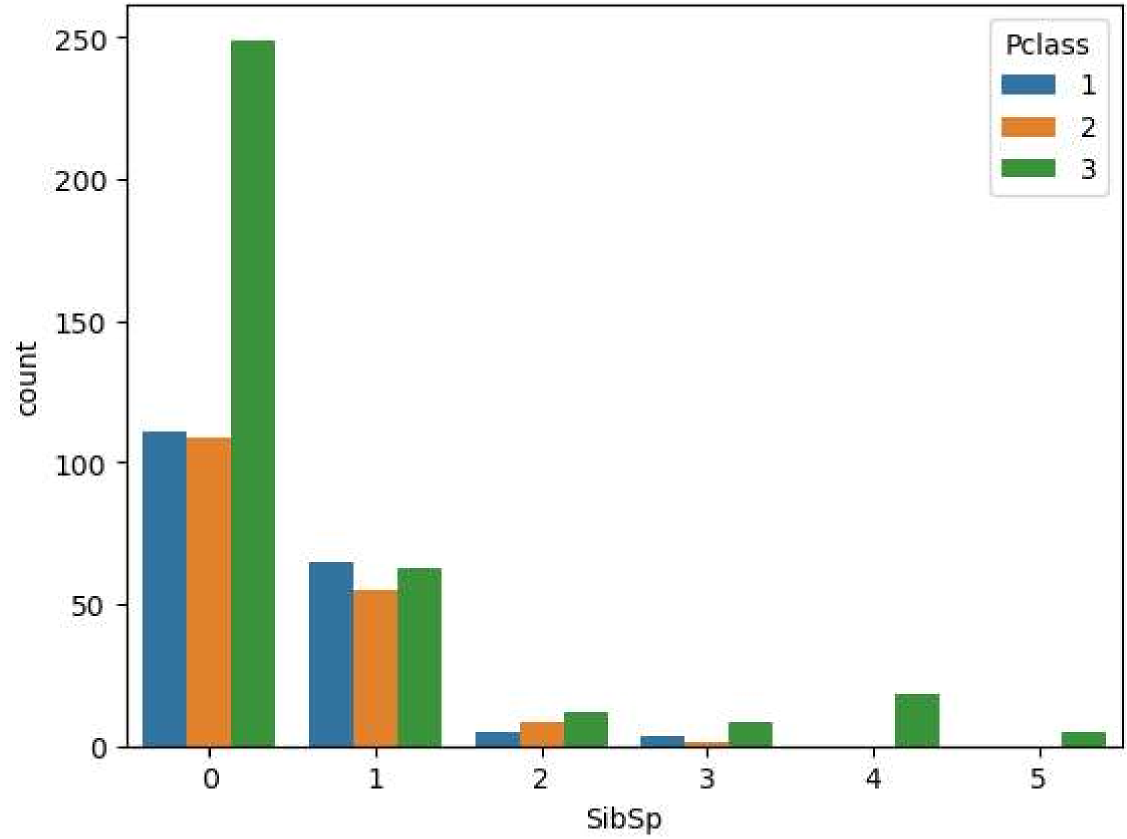
In [15]:

```
sns.countplot(x='SibSp', hue='Sex', data=data);
```



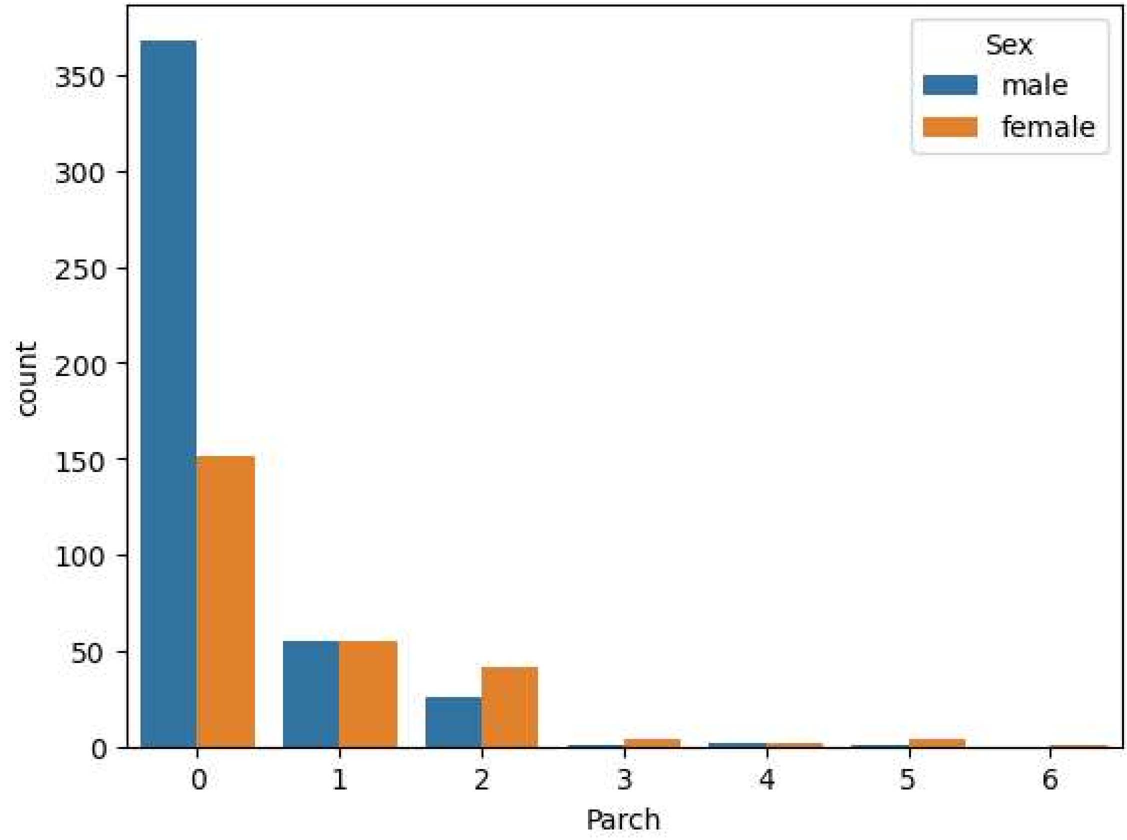
In [16]:

```
sns.countplot(x='SibSp', hue='Pclass', data=data);
```



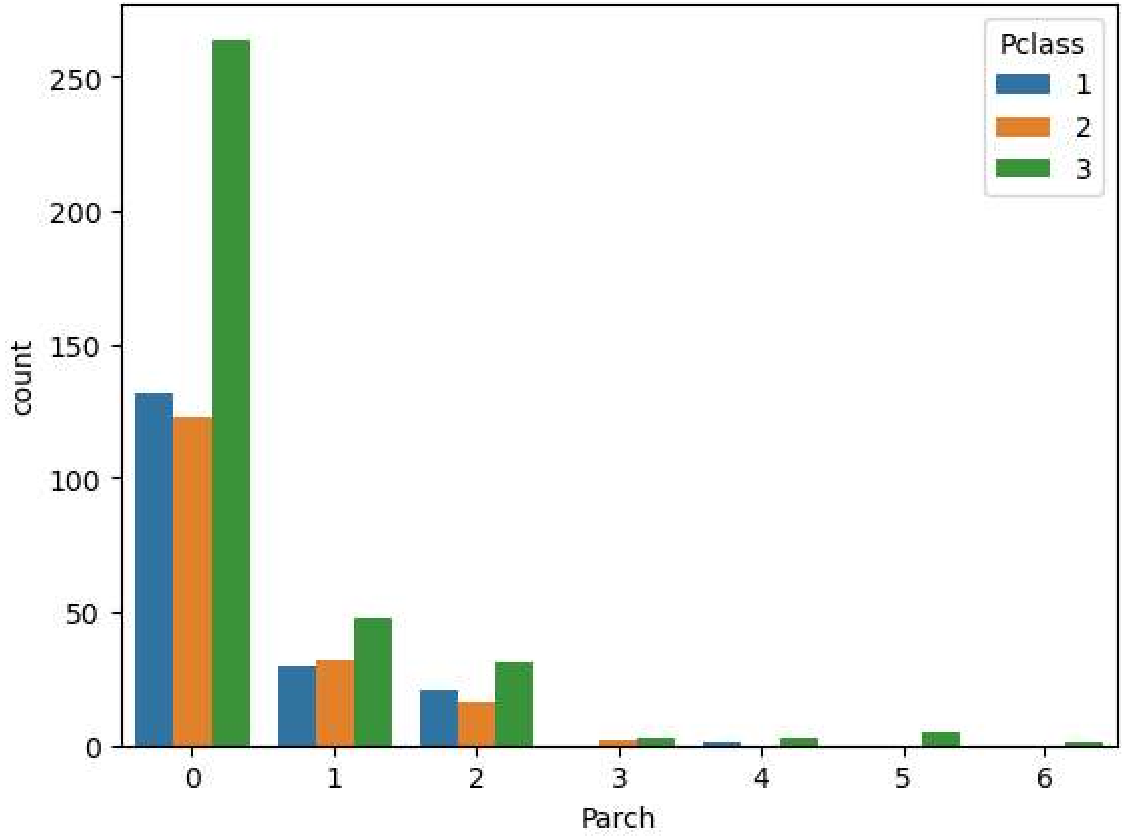
In [17]:

```
sns.countplot(x='Parch', hue='Sex', data=data);
```



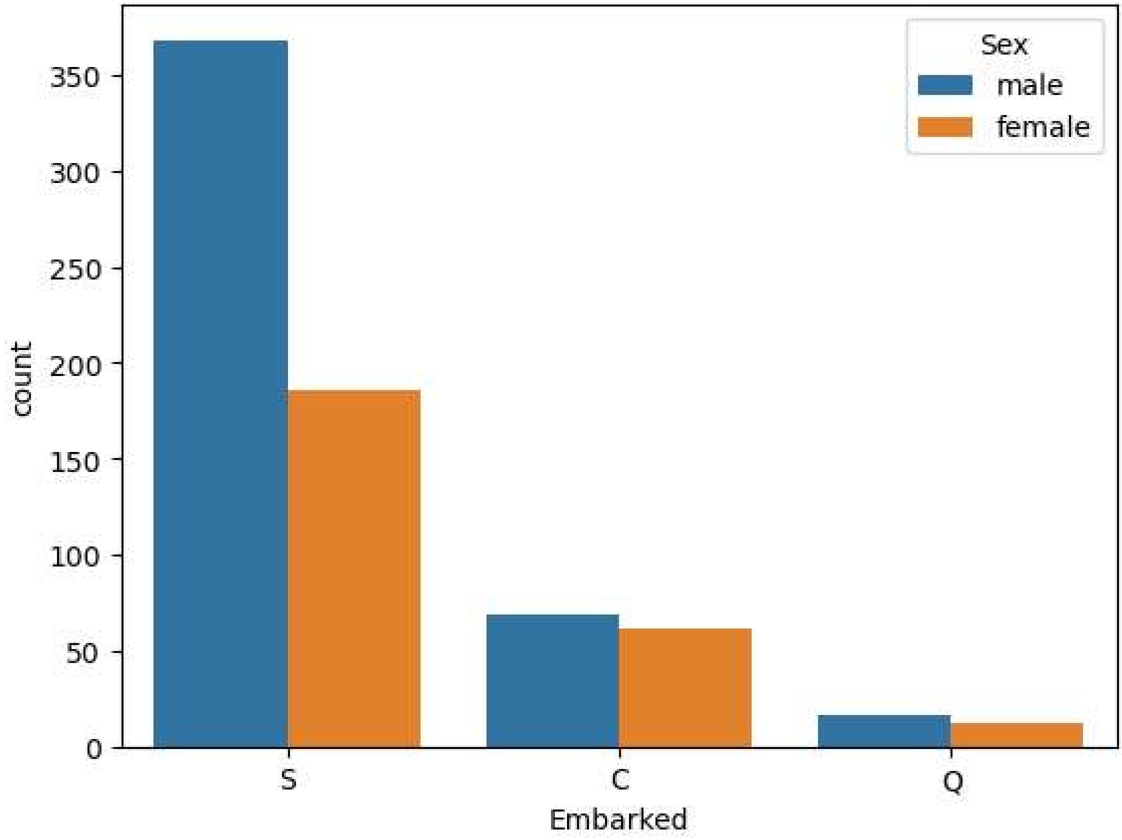
In [18]:

```
sns.countplot(x='Parch', hue='Pclass', data=data);
```



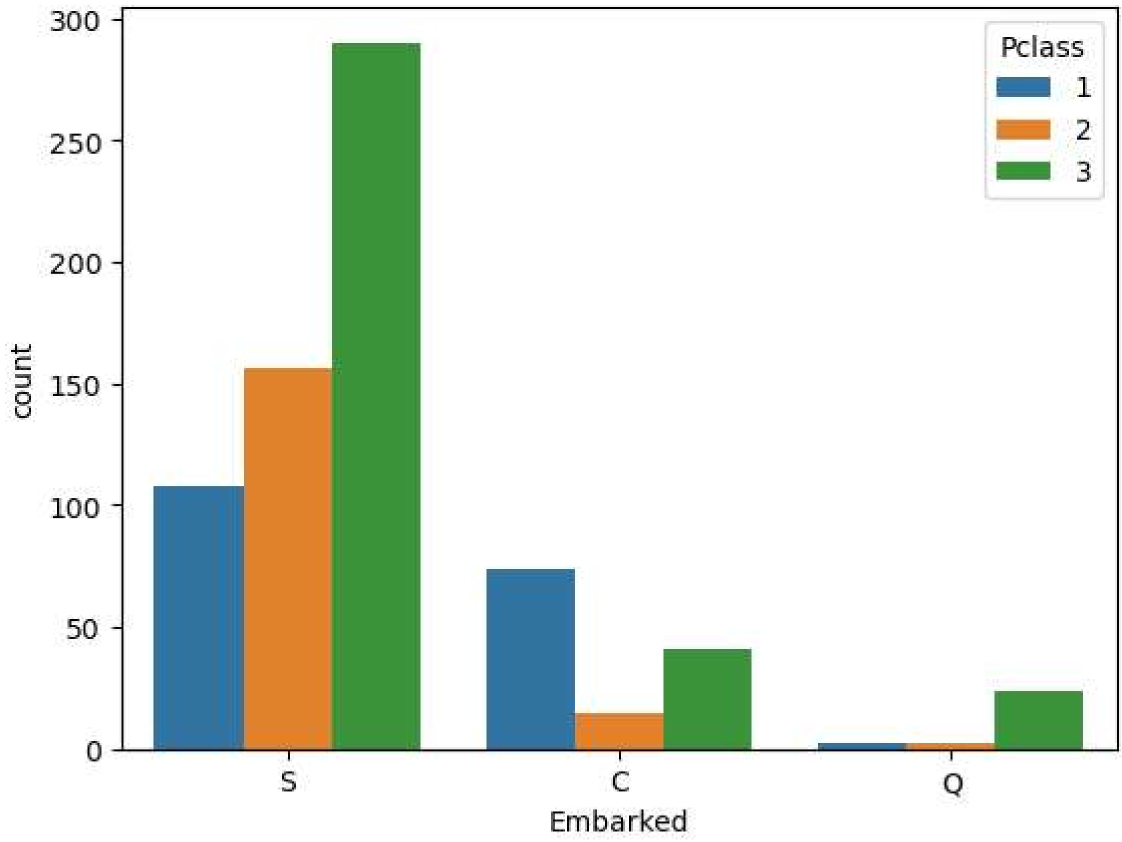
In [19]:

```
sns.countplot(x='Embarked', data=data, hue='Sex');
```



In [20]:

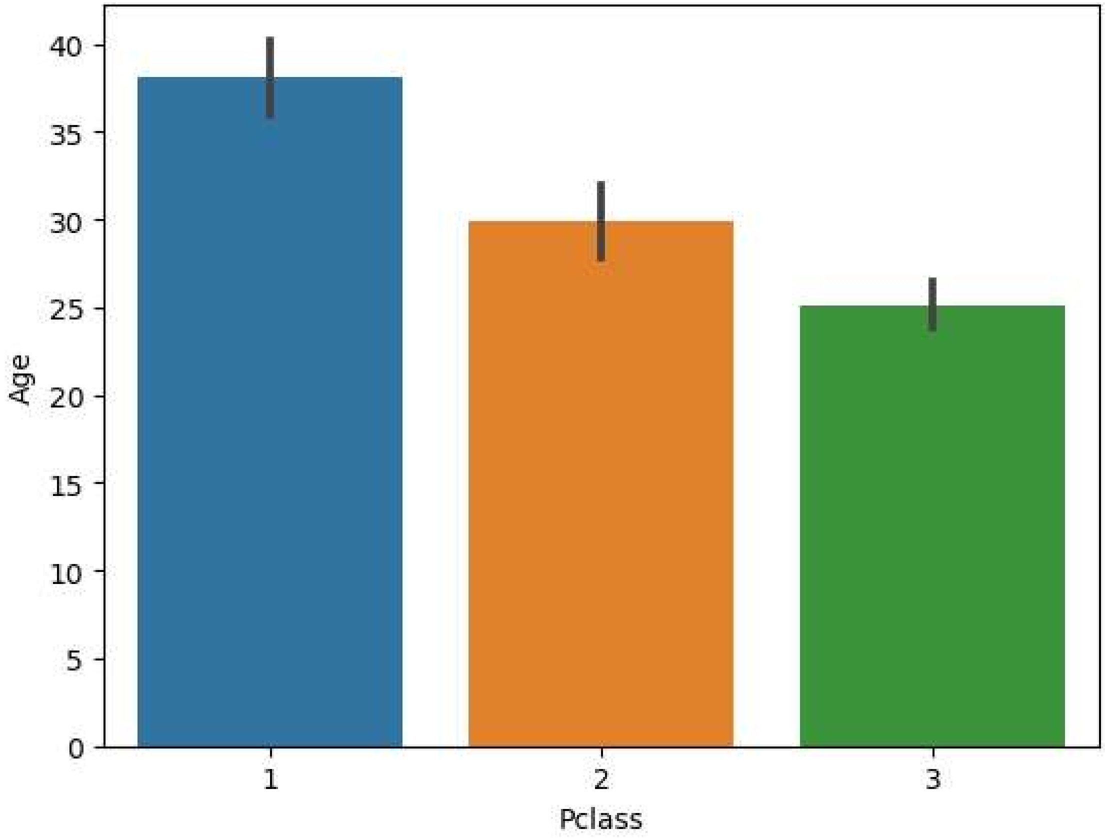
```
sns.countplot(x='Embarked', data=data, hue='Pclass');
```



Bar chart

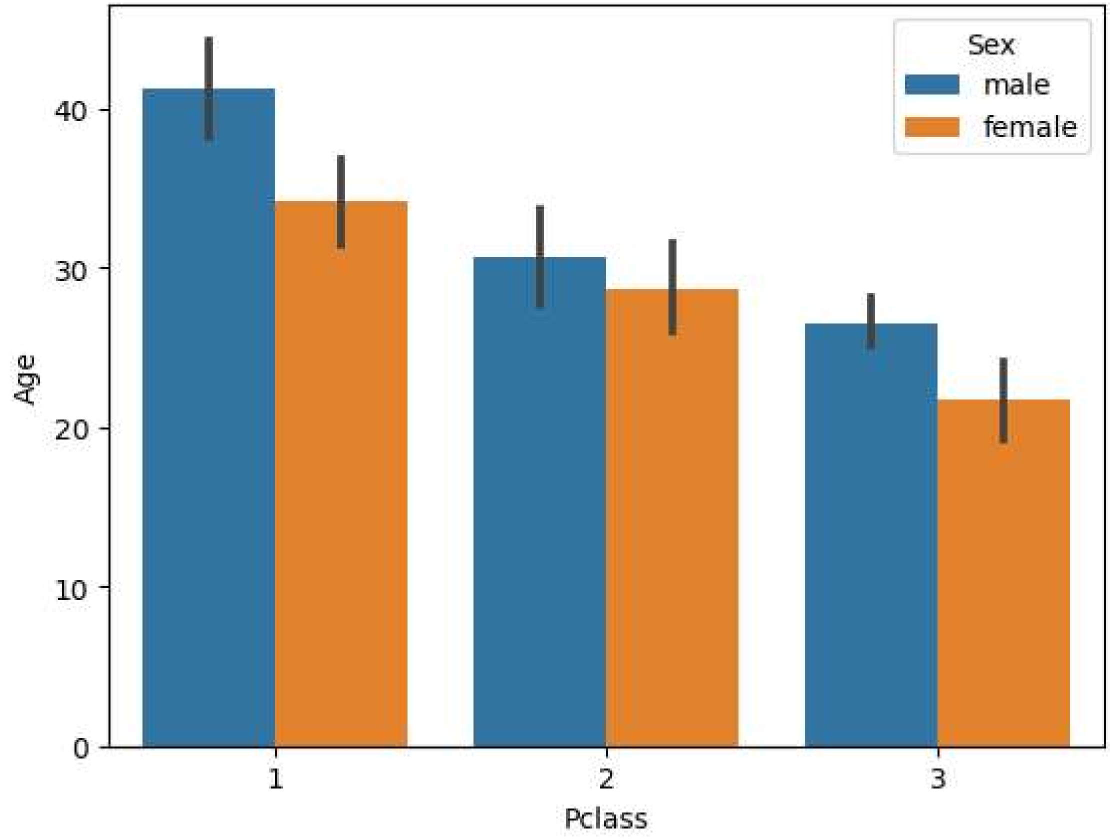
In [21]:

```
sns.barplot(x='Pclass', y='Age',data=data);
```



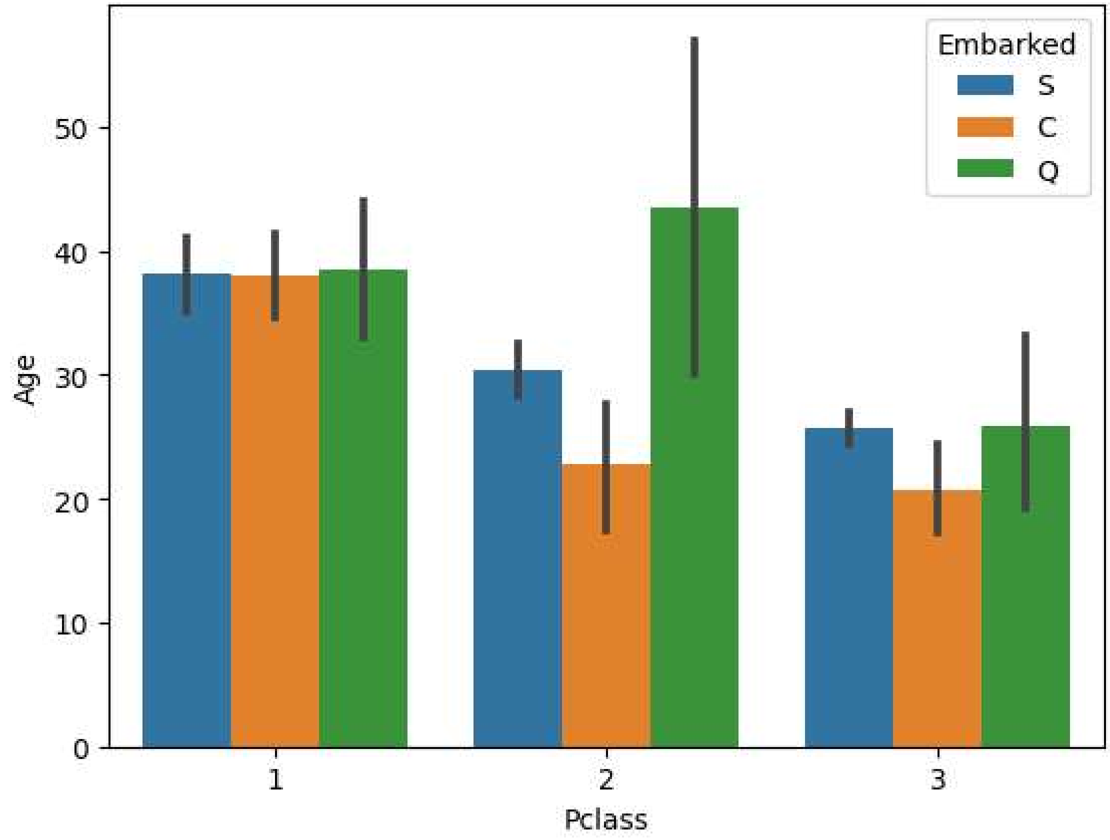
In [22]:

```
sns.barplot(x='Pclass', y='Age', hue='Sex', data=data);
```



In [23]:

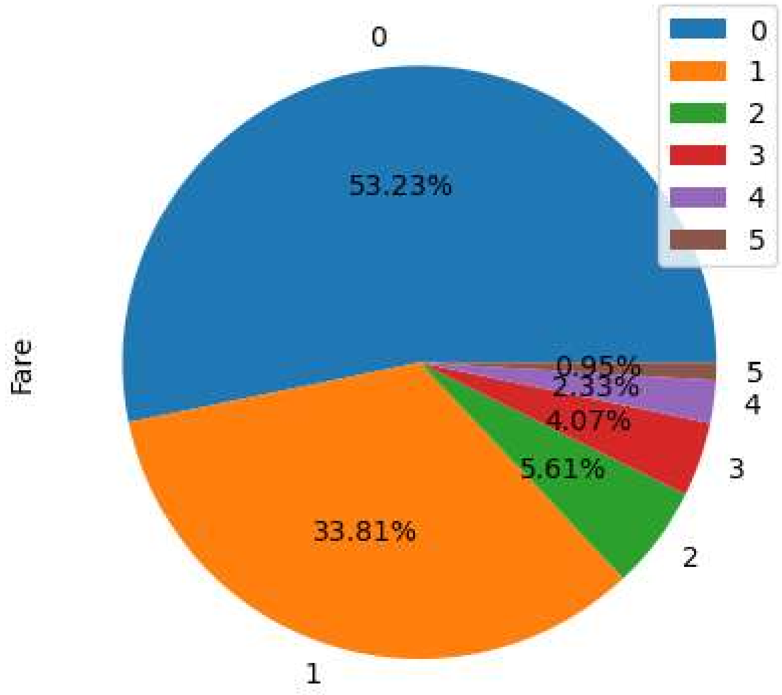
```
sns.barplot(x='Pclass', y='Age', hue='Embarked', data=data);
```



Pie chart

In [24]:

```
SibSp_Fare = data.groupby(['SibSp'])['Fare'].sum()
SibSp_Fare.plot.pie(legend =True, autopct = '%1.2f%%');
```

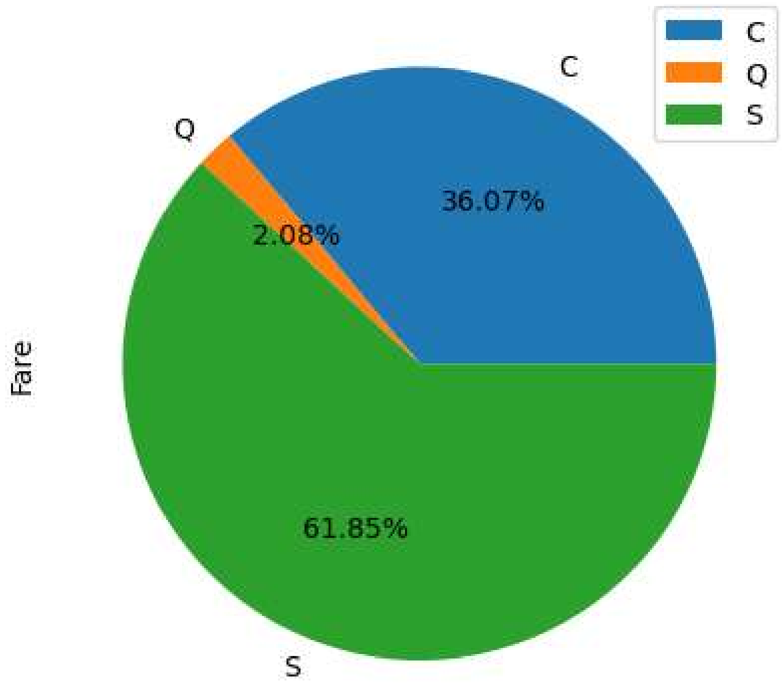


In [25]:

```
Embarked_fare = data.groupby(['Embarked'])['Fare'].sum()
```

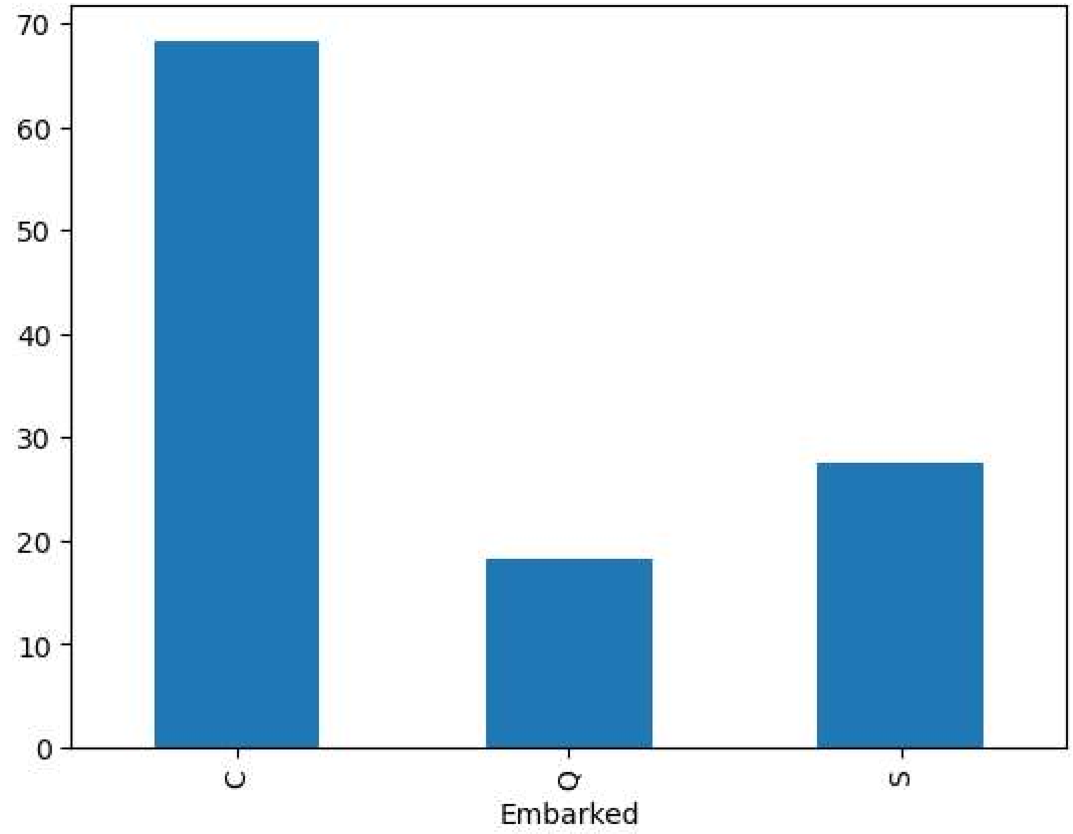
In [26]:

```
Embarked_fare.plot.pie(legend=True, autopct = '%1.2f%%');
```



In [27]:

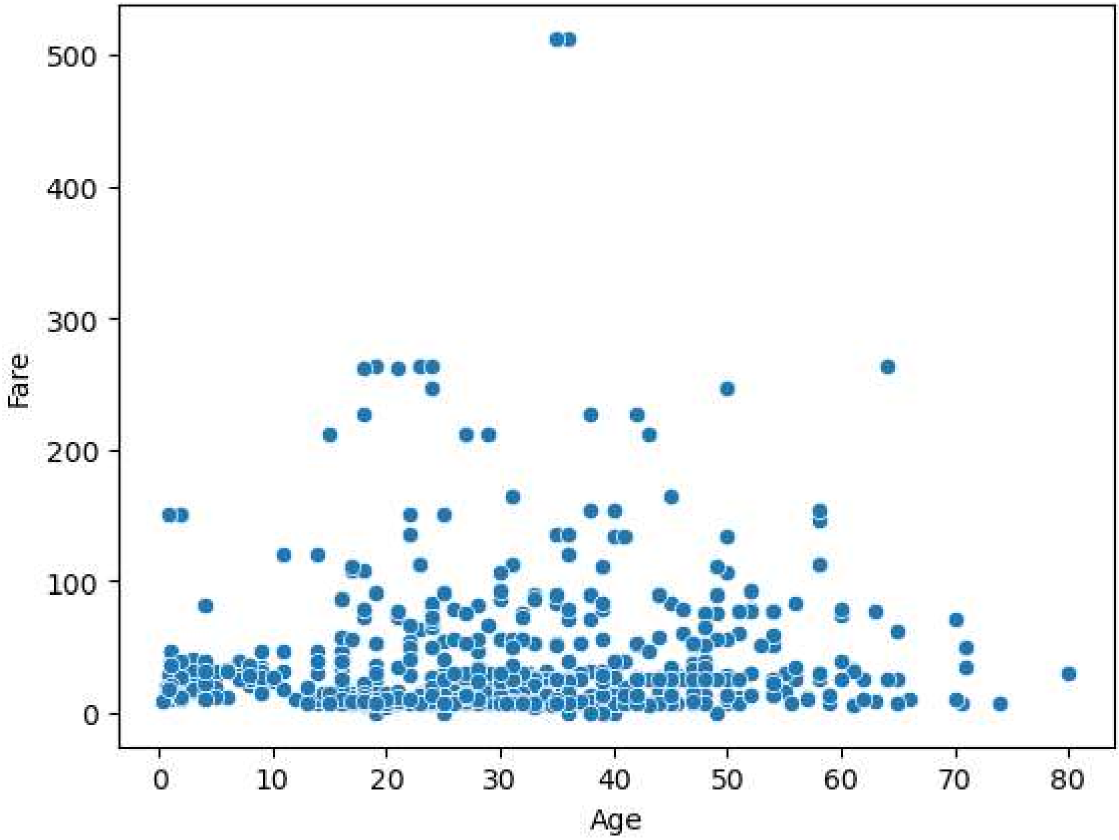
```
Embarked_fare_mean = data.groupby(['Embarked'])['Fare'].mean()
Embarked_fare_mean.plot.bar();
```



Scatter plot

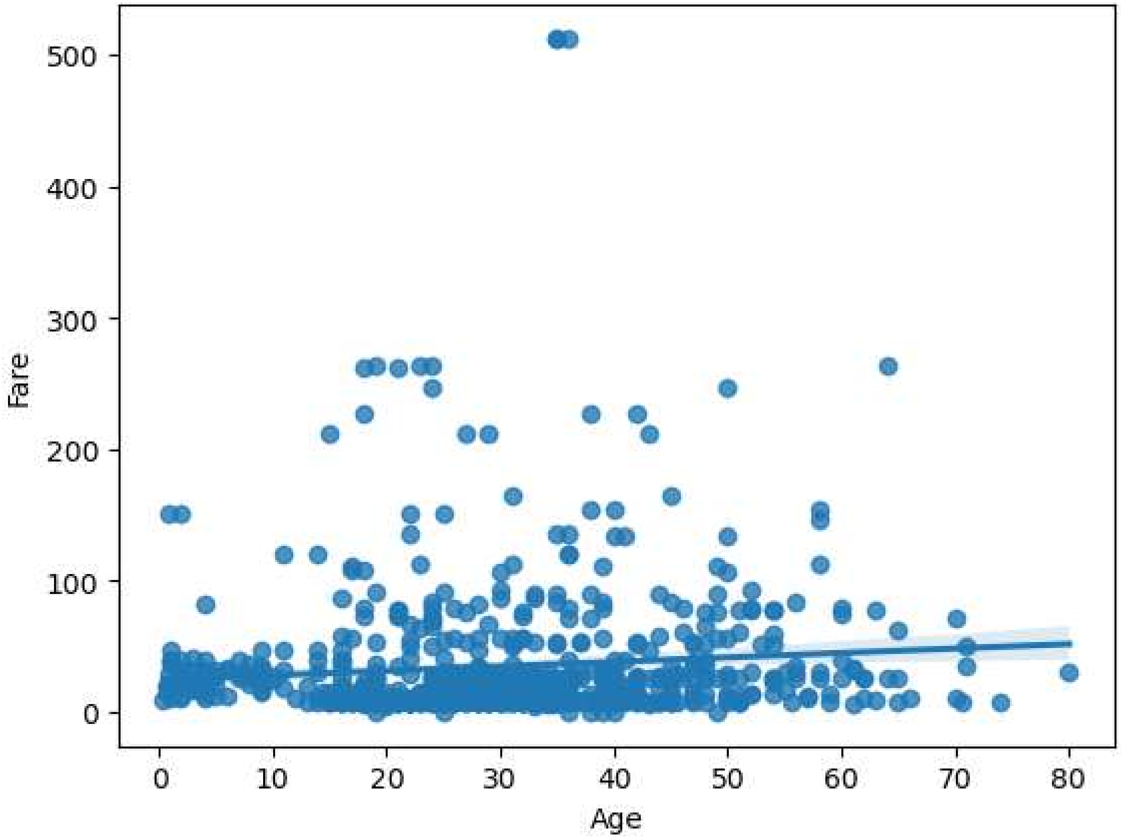
In [28]:

```
sns.scatterplot(x='Age', y='Fare', data=data);
```



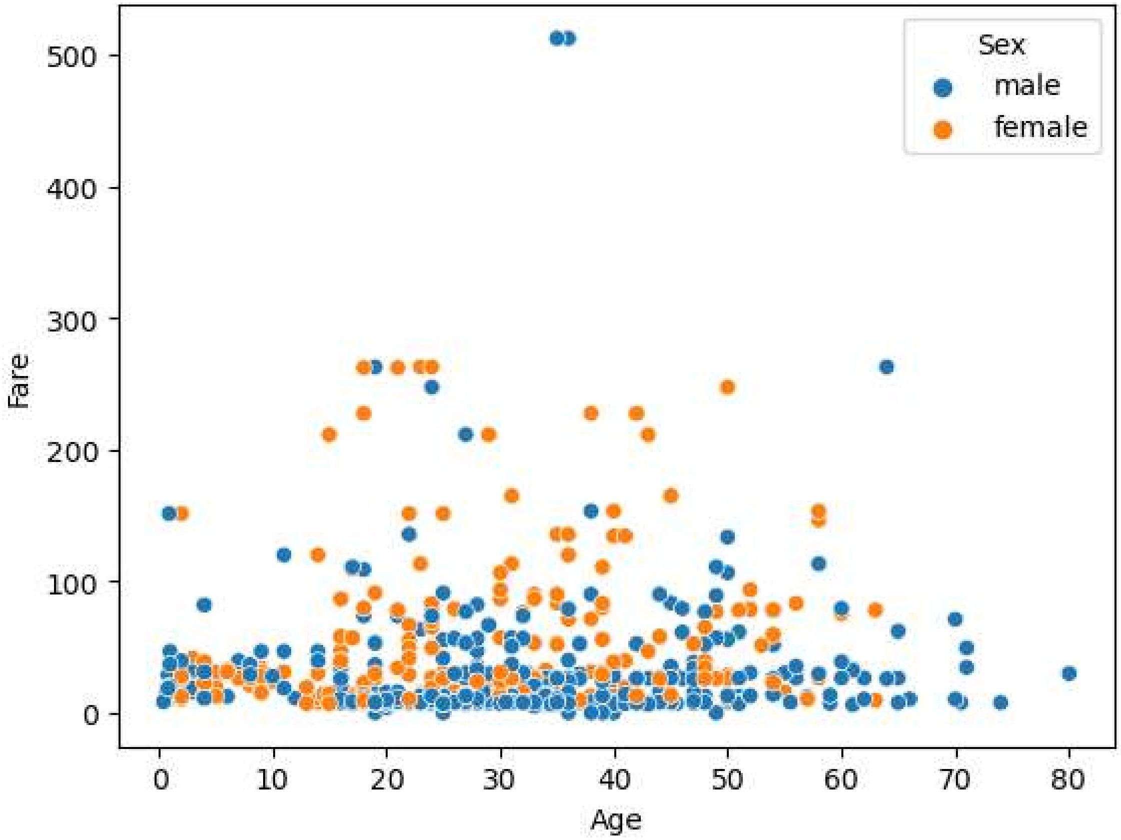
In [29]:

```
sns.regplot(x='Age', y='Fare', data=data);
```



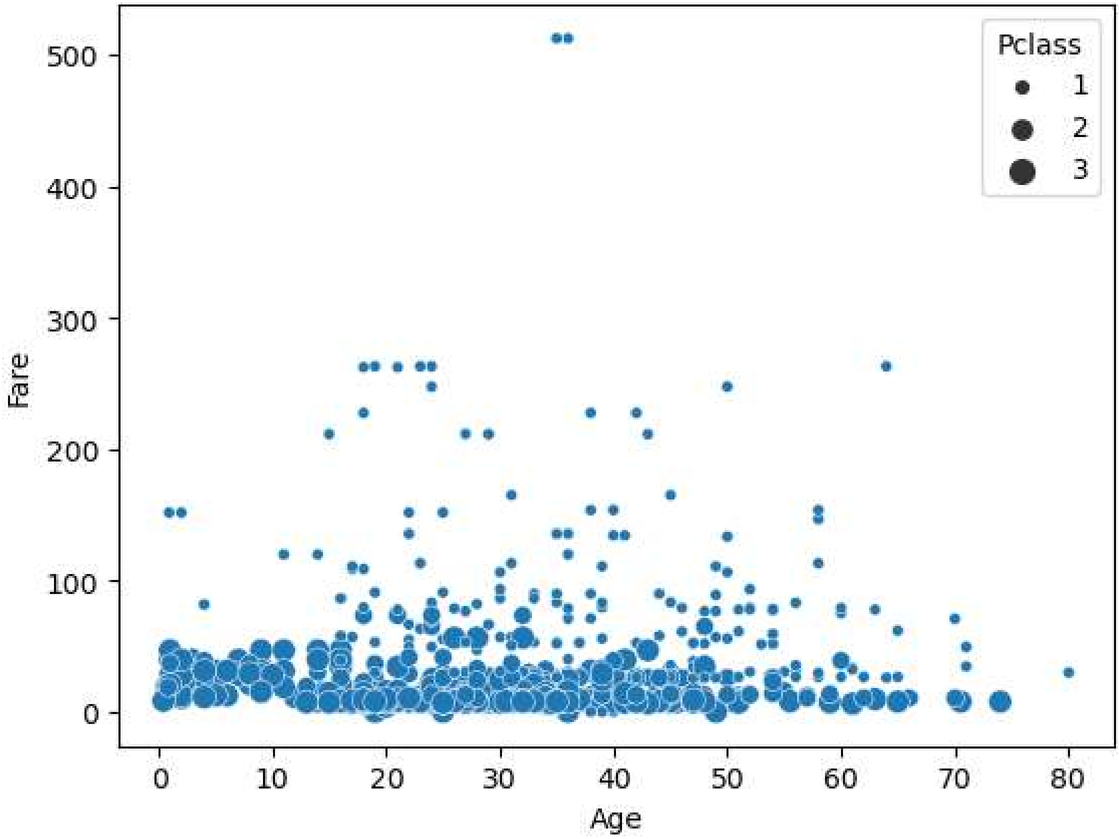
In [30]:

```
sns.scatterplot(x='Age', y='Fare',hue= 'Sex',data=data);
```



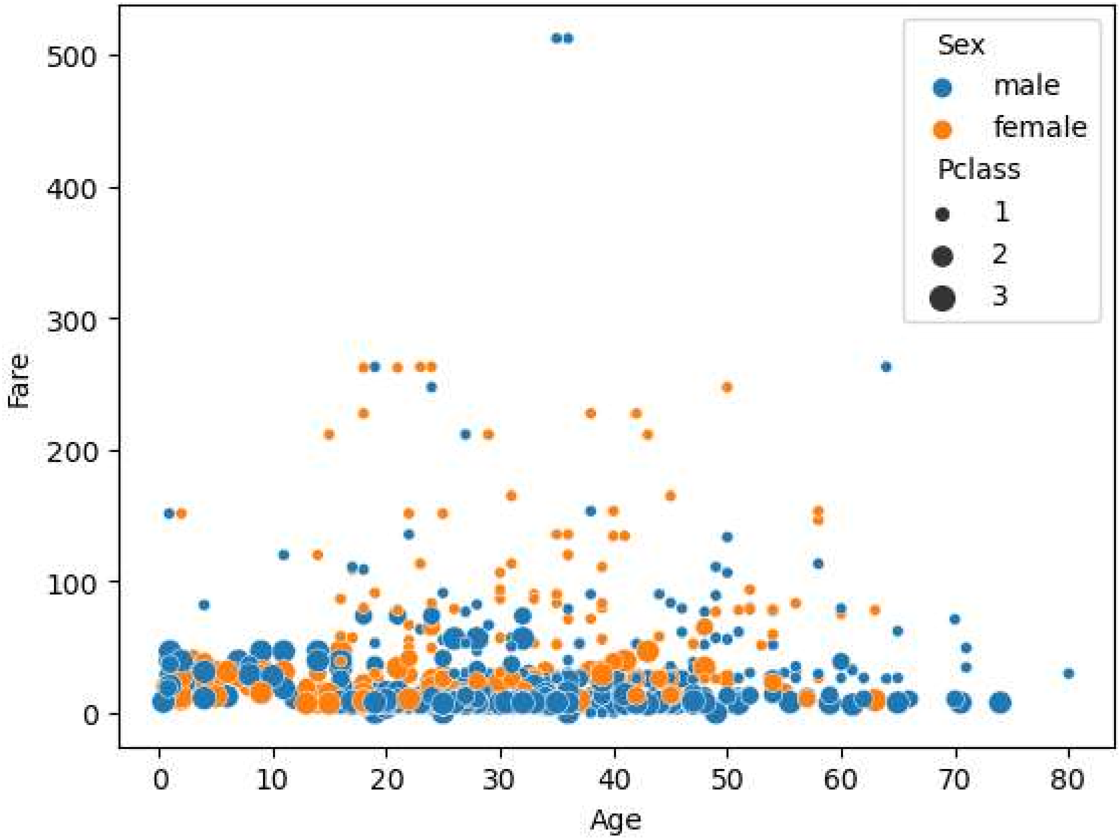
In [31]:

```
sns.scatterplot(x='Age', y='Fare',size='Pclass',data=data);
```



In [32]:

```
sns.scatterplot(x='Age', y='Fare',hue= 'Sex',size='Pclass',data=data);
```


Sex
● male
● female
Pclass
● 1
● 2
● 3
Fare
Age