```
In [2]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
```

```
In [3]: df=pd.read_csv(r"D:\Stranger Things\New folder\you_ex_data seience\YOUTUBE\EDA_y
```
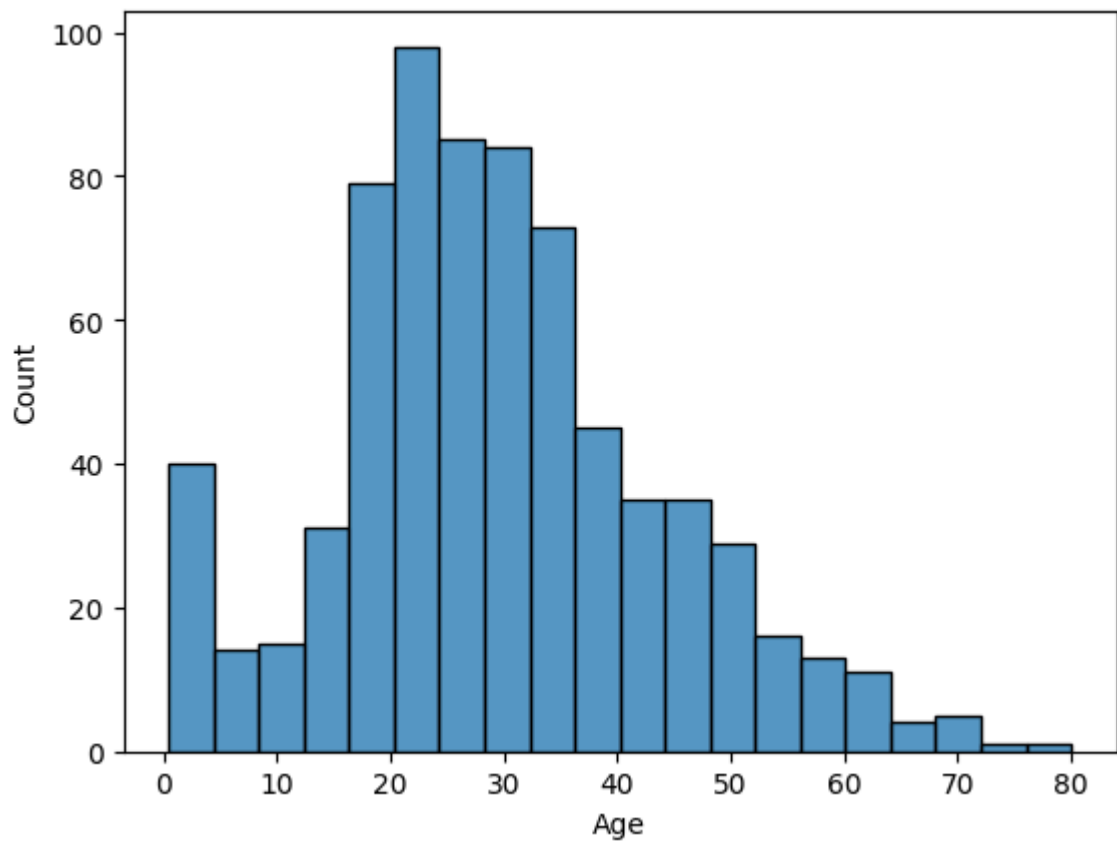
```
In [4]: df.head()
```

Out[4]:

	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 [5]: df['Age'].describe()
```

```
Out[5]: count    714.000000
        mean      29.699118
        std       14.526497
        min        0.420000
        25%       20.125000
        50%       28.000000
        75%       38.000000
        max       80.000000
        Name: Age, dtype: float64
```
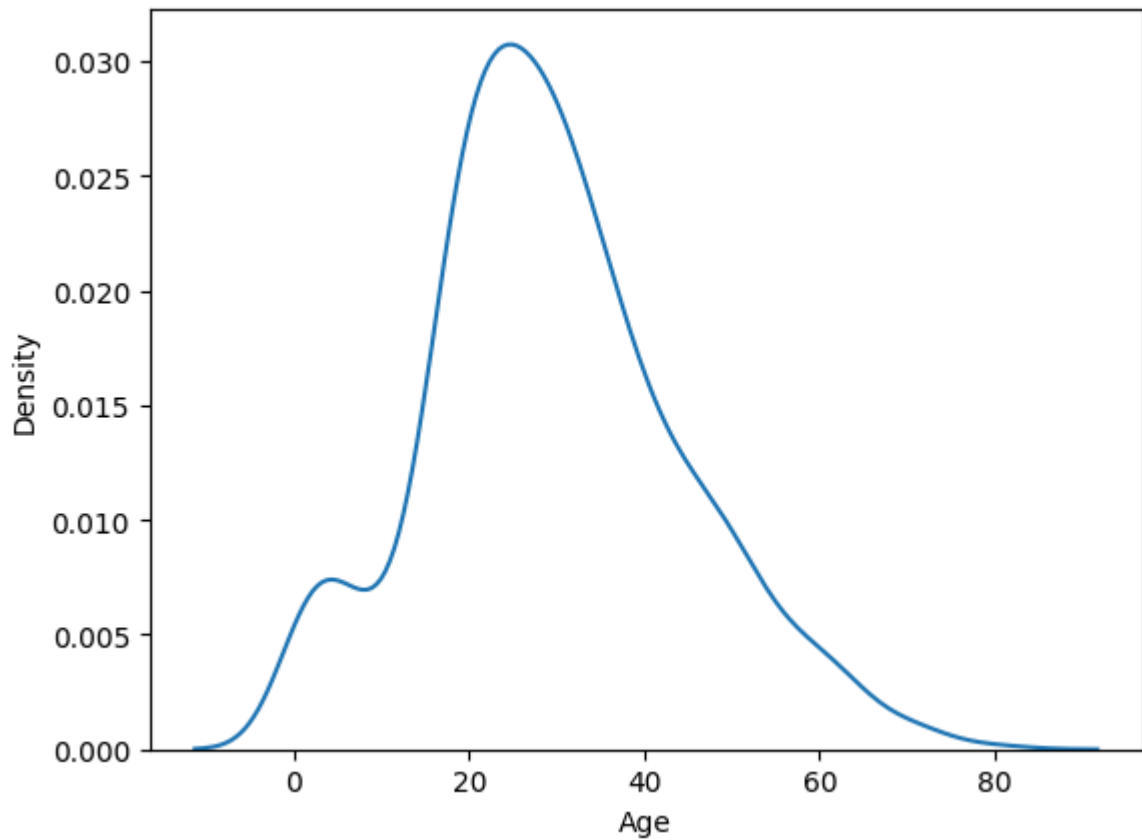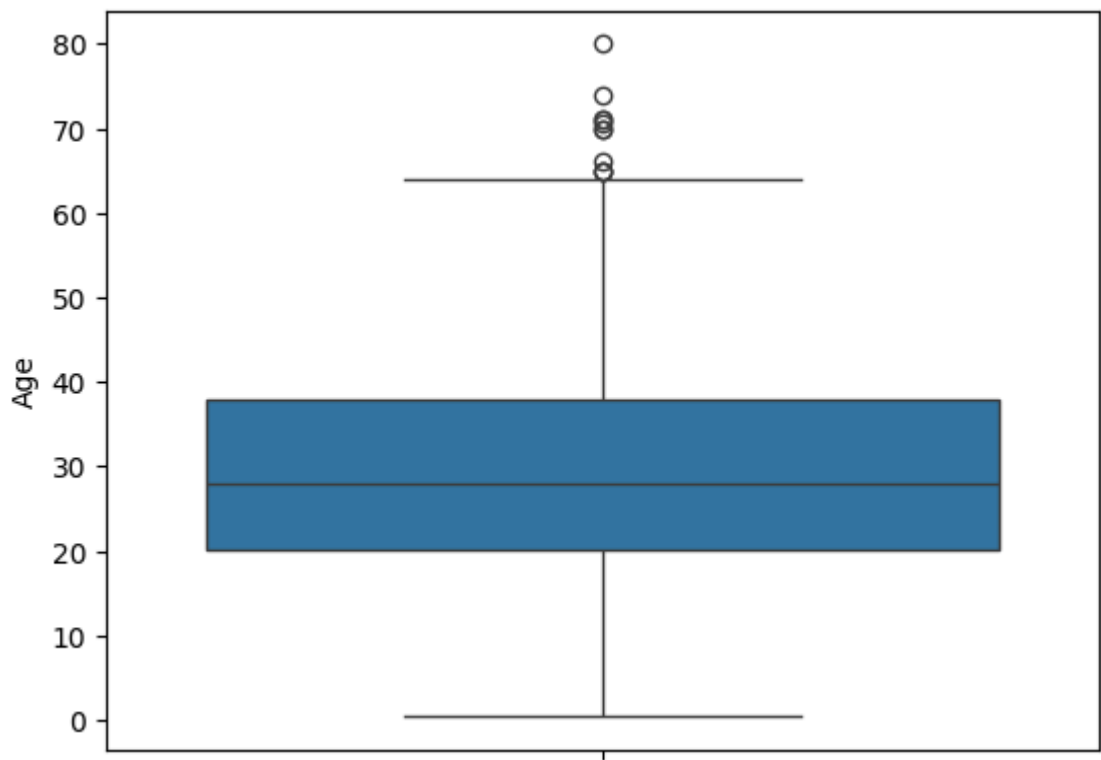
```
In [6]: sns.histplot(df['Age'])
```

```
Out[6]: <Axes: xlabel='Age', ylabel='Count'>
```

`sns.kdeplot(df['Age'])`

`<Axes: xlabel='Age', ylabel='Density'>`

`sns.boxplot(df['Age'])`

`<Axes: ylabel='Age'>`

```
In [10]: df['Age'].isnull().sum()
```

Out[10]: 177

```
In [11]: df['Fare'].describe()
```

Out[11]:
```
count    891.000000
mean      32.204208
std       49.693429
min        0.000000
25%        7.910400
50%       14.454200
75%       31.000000
max      512.329200
Name: Fare, dtype: float64
```

```
In [12]: sns.histplot(df['Fare'])
```

Out[12]: <Axes: xlabel='Fare', ylabel='Count'>
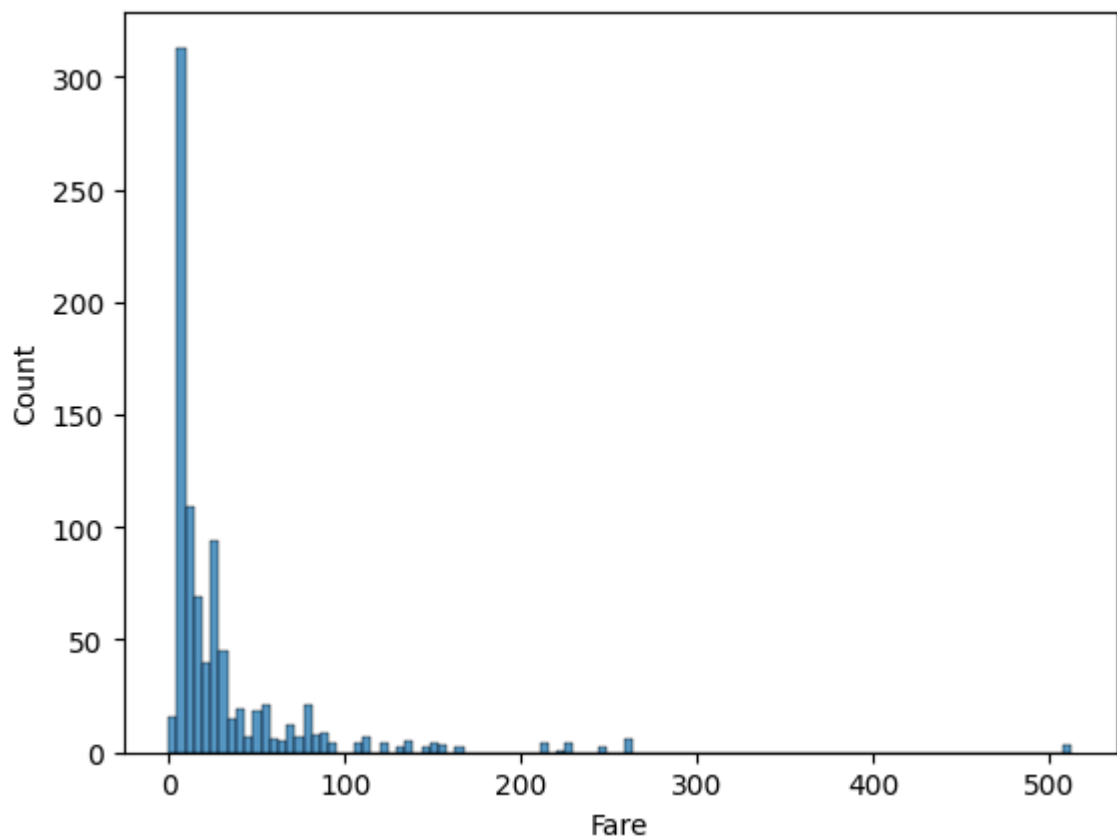

```
In [13]: sns.kdeplot(df['Fare'])
```

Out[13]: <Axes: xlabel='Fare', ylabel='Density'>

```
In [14]:  sns.boxplot(df['Fare'])
```

```
Out[14]:  <Axes: ylabel='Fare'>
```
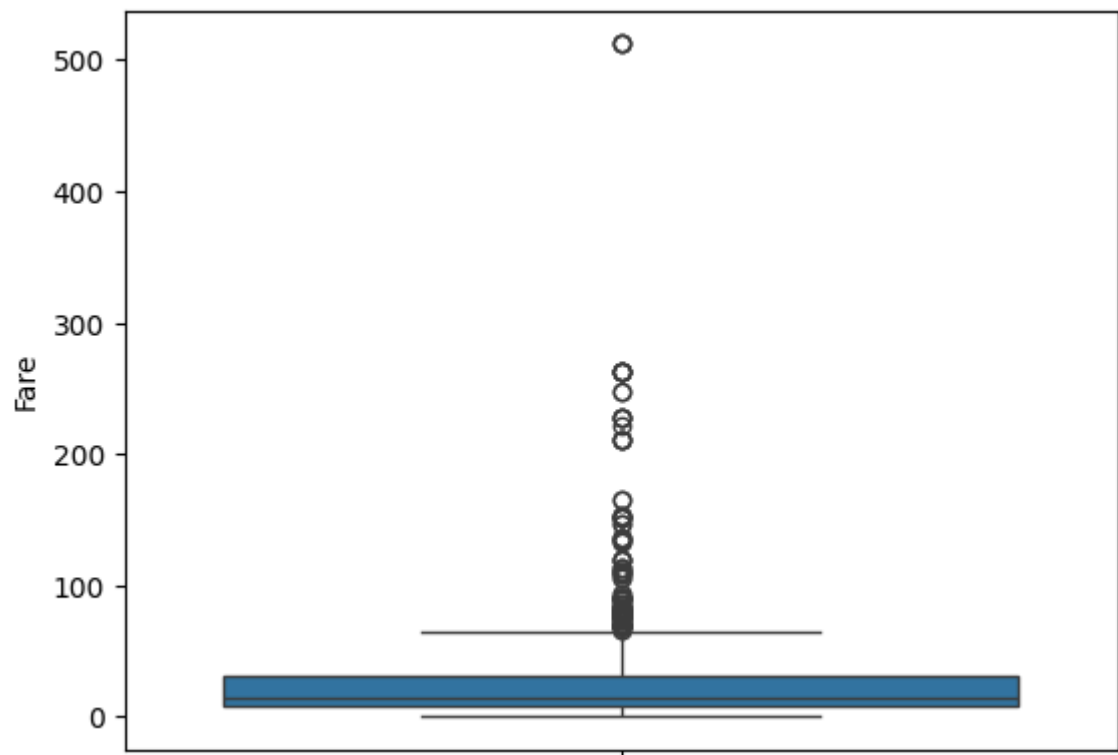


```
In [15]:  df[df['Fare']>250]
```

Out[15]:

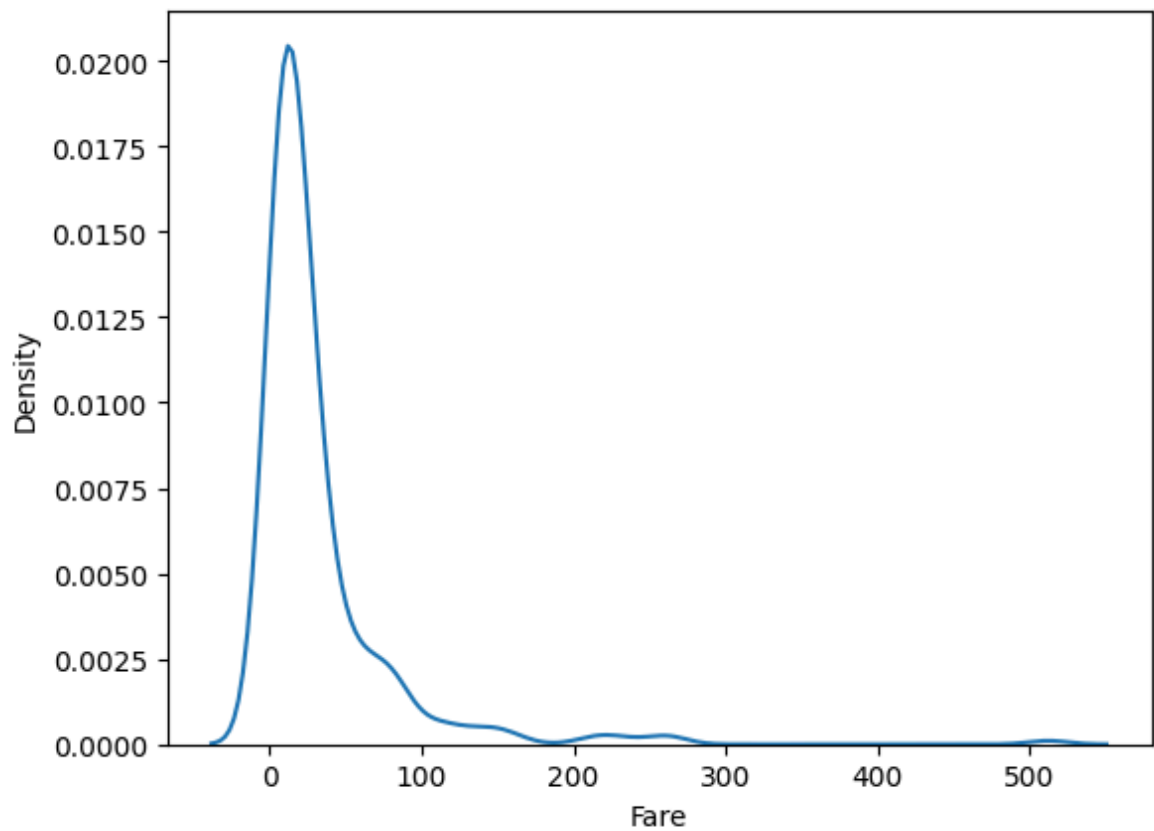|	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
27	28	0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0
88	89	1	1	Fortune, Miss. Mabel Helen	female	23.0	3	2	19950	263.0
258	259	1	1	Ward, Miss. Anna	female	35.0	0	0	PC 17755	512.3
311	312	1	1	Ryerson, Miss. Emily Borie	female	18.0	2	2	PC 17608	262.3
341	342	1	1	Fortune, Miss. Alice Elizabeth	female	24.0	3	2	19950	263.0
438	439	0	1	Fortune, Mr. Mark	male	64.0	1	4	19950	263.0
679	680	1	1	Cardeza, Mr. Thomas Drake Martinez	male	36.0	0	1	PC 17755	512.3
737	738	1	1	Lesurer, Mr. Gustave J	male	35.0	0	0	PC 17755	512.3
742	743	1	1	Ryerson, Miss. Susan Parker "Suzette"	female	21.0	2	2	PC 17608	262.3

In [16]: `df['Fare'].isnull().sum()`

Out[16]: 0

In [17]: `df['Survived'].unique()`

Out[17]: array([0, 1], dtype=int64)

In [18]: `df['Survived'].value_counts()`

```
Out[18]:  Survived
          0    549
          1    342
          Name: count, dtype: int64
```

```
In [19]:  sns.countplot(data=df,x='Survived')
```

```
Out[19]:  <Axes: xlabel='Survived', ylabel='count'>
```
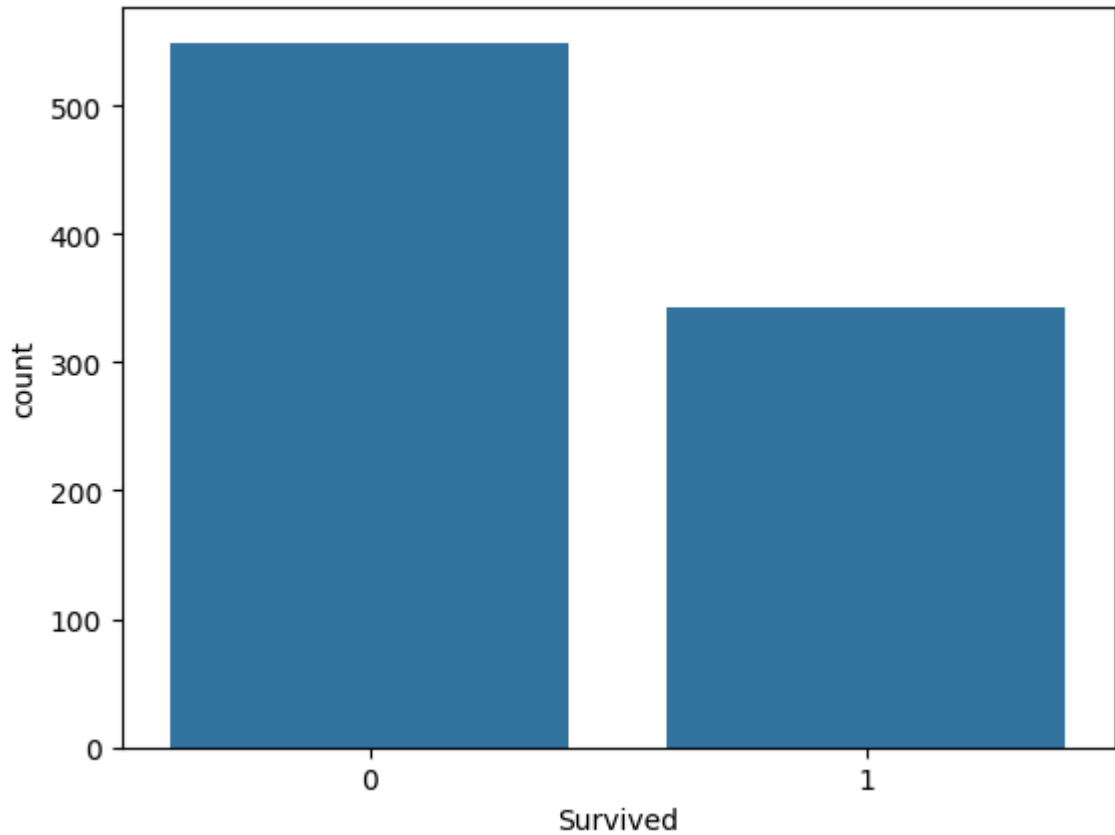


```
In [20]:  df['Survived'].value_counts().plot(kind='pie', autopct='%1.1f%%')
```

```
Out[20]:  <Axes: ylabel='count'>
```
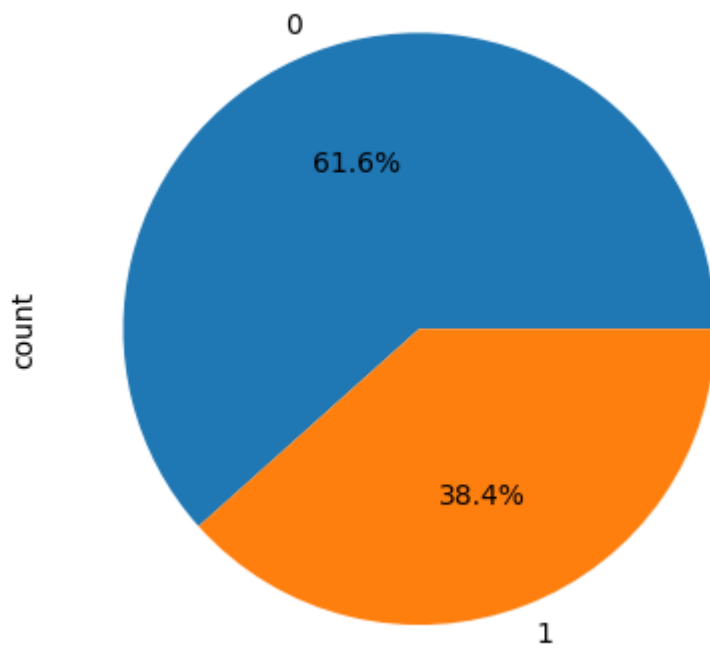
```
In [21]:  df['Survived'].isnull().sum()
```

Out[21]: 0

```
In [22]:  df['Pclass'].value_counts()
```

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

```
In [23]:  sns.countplot(data=df,x='Pclass')
```

Out[23]: <Axes: xlabel='Pclass', ylabel='count'>
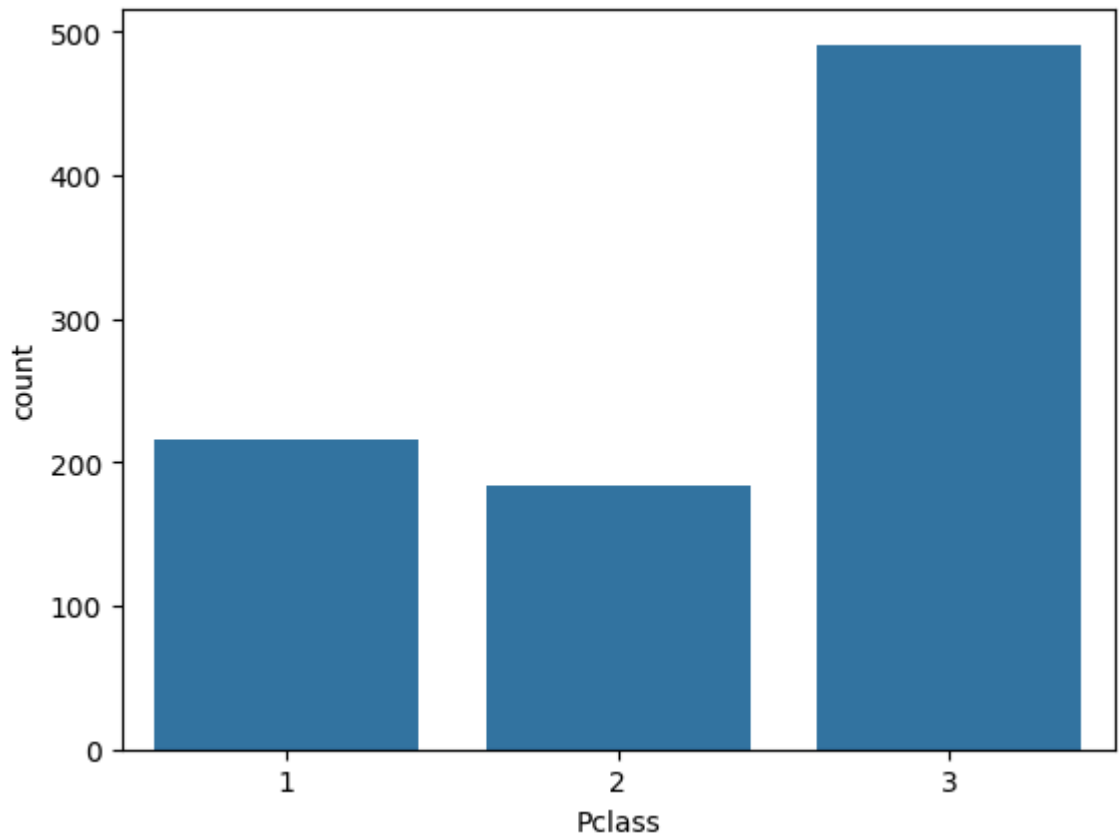
```
In [24]: df['Pclass'].value_counts().plot(kind='pie',autopct='%1.1f%%')
```

Out[24]: <Axes: ylabel='count'>
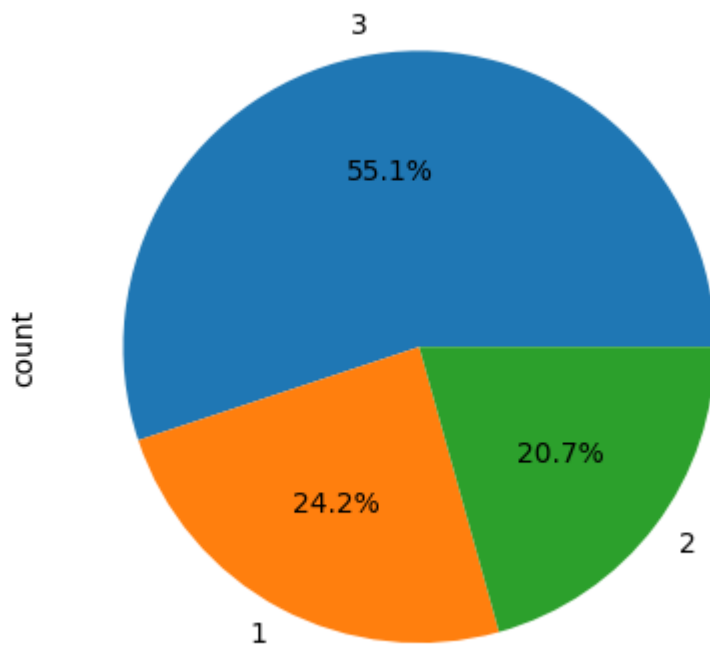

```
In [25]: df['Pclass'].isnull().sum()
```

Out[25]: 0

```
In [26]: df['Sex'].value_counts()
```

Out[26]:   Sex
            male   577
            female   314
            Name: count, dtype: int64

In [27]:   sns.countplot(data=df,x='Sex')

Out[27]:   <Axes: xlabel='Sex', ylabel='count'>
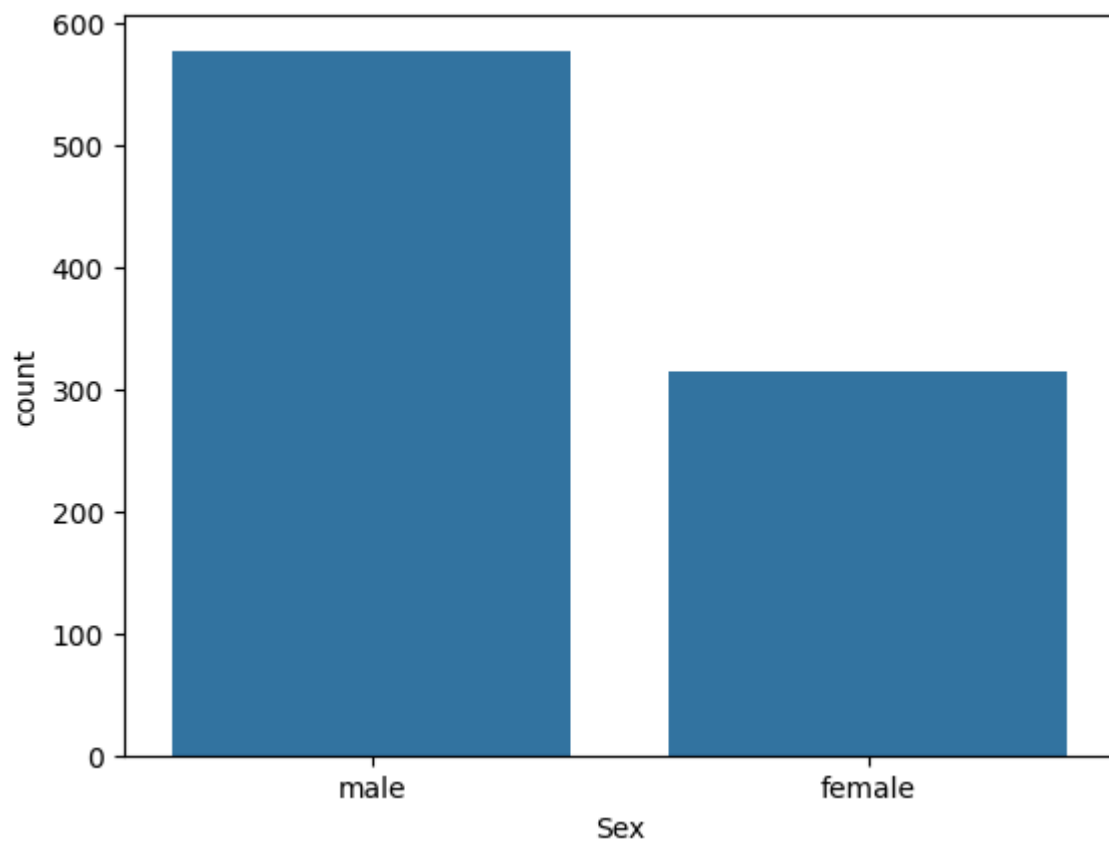
In [28]:   df['Sex'].value_counts().plot(kind='pie',autopct='%1.1f%%')

Out[28]:   <Axes: ylabel='count'>
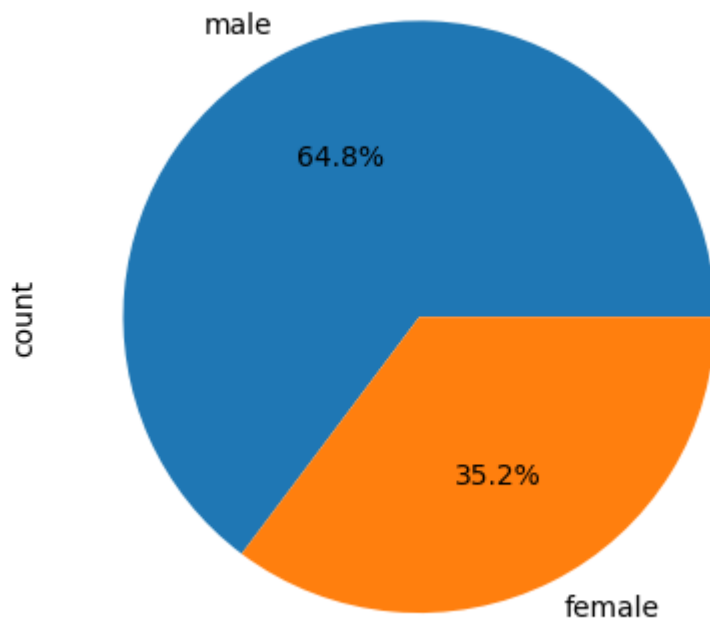
```
In [29]: df['Sex'].isnull().sum()

Out[29]: 0

In [30]: df['SibSp'].unique()

Out[30]: array([1, 0, 3, 4, 2, 5, 8], dtype=int64)

In [31]: df['SibSp'].value_counts()

Out[31]: SibSp
         0    608
         1    209
         2     28
         4     18
         3     16
         8      7
         5      5
         Name: count, dtype: int64

In [32]: sns.countplot(data=df,x='SibSp')

Out[32]: <Axes: xlabel='SibSp', ylabel='count'>
```
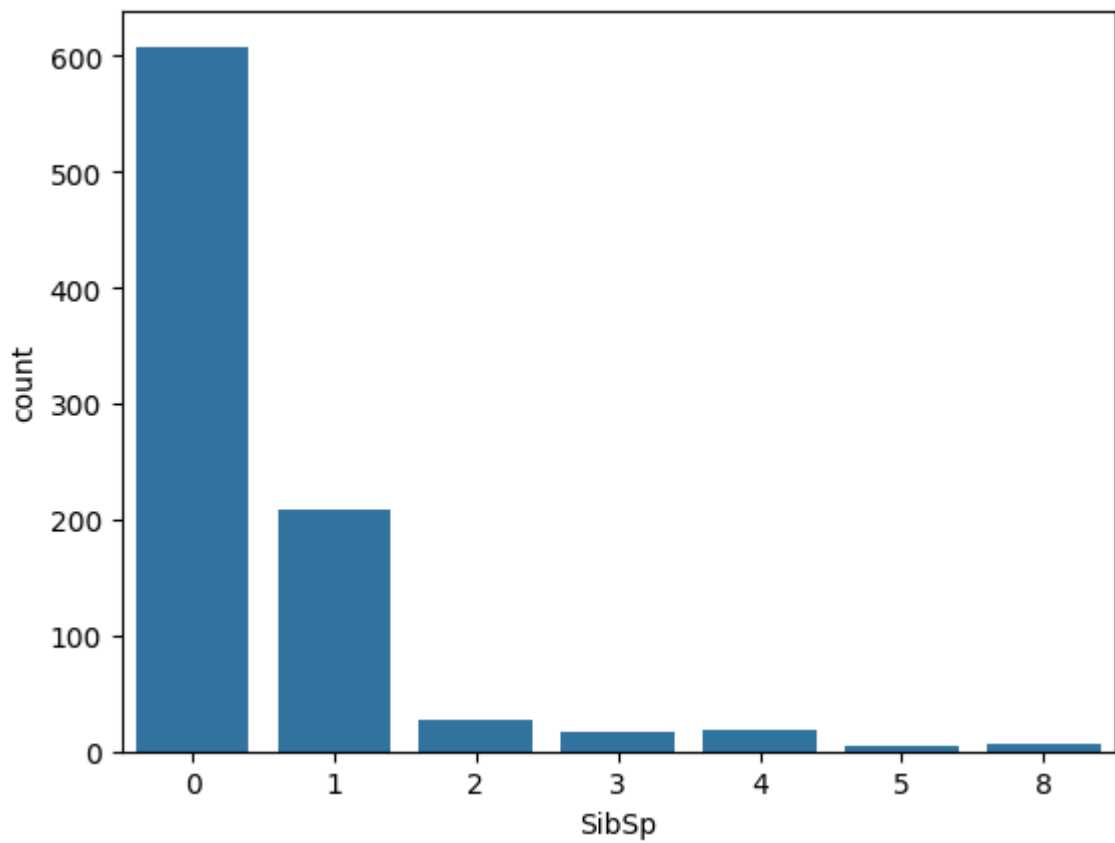
```
In [33]: df['SibSp'].value_counts().plot(kind='pie',autopct='%1.1f%%')
```

Out[33]: <Axes: ylabel='count'>
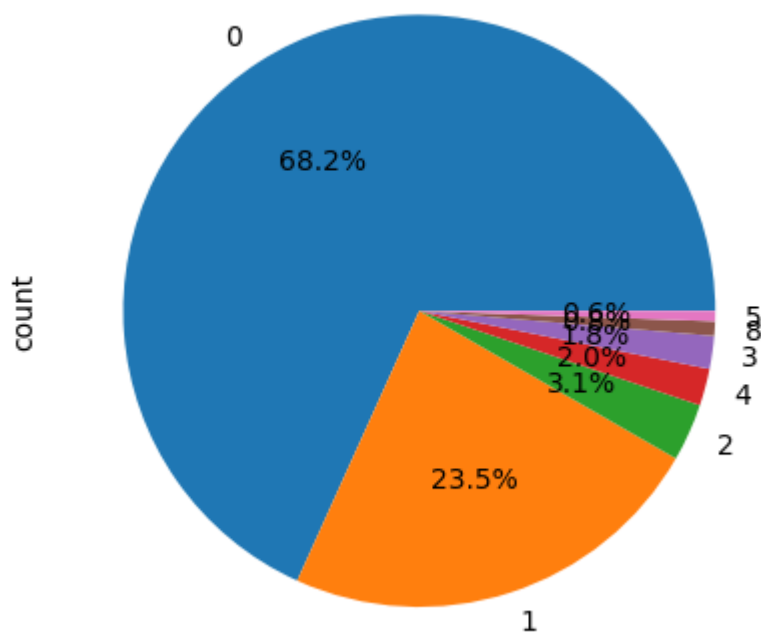

```
In [34]: df['SibSp'].isnull().sum()
```

Out[34]: 0

```
In [35]: df["Parch"].value_counts()
```

Out[35]:
```
Parch
0    678
1    118
2     80
5      5
3      5
4      4
6      1
Name: count, dtype: int64
```

In [36]:
```python
df['Embarked'].value_counts()
```

Out[36]:
```
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

In [37]:
```python
sns.countplot(data=df,x='Embarked')
```

Out[37]: `<Axes: xlabel='Embarked', ylabel='count'>`
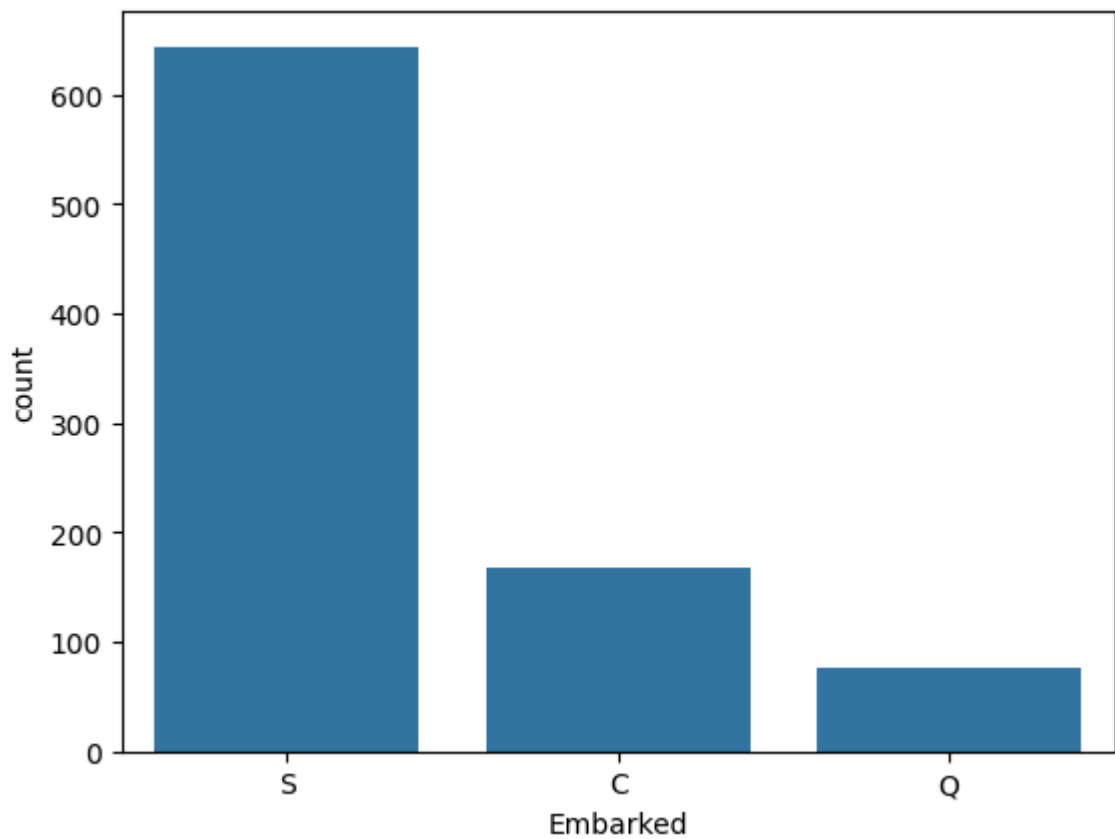
In [38]:
```python
df['Embarked'].value_counts().plot(kind='pie',autopct='%1.1f%%')
```

Out[38]: `<Axes: ylabel='count'>`
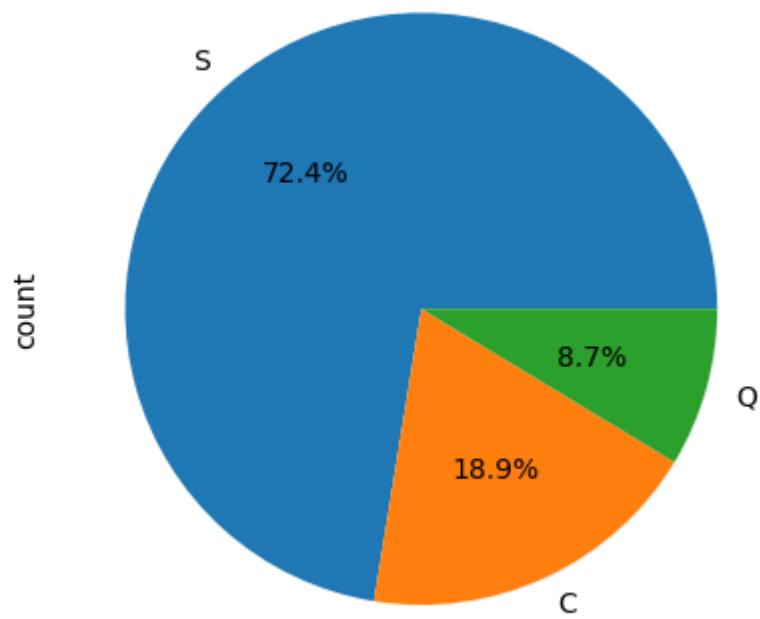
```
In [39]: df['Embarked'].isnull().sum()
```

Out[39]: 2

```
In [40]: df
```

Out[40]:

	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
...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376

891 rows × 12 columns

In [47]:
```python
pd.crosstab(df['Survived'],df['Pclass'],normalize='columns')*100
```

Pclass	1	2	3
Survived			
0	37.037037	52.717391	75.763747
1	62.962963	47.282609	24.236253

```python
sns.heatmap(pd.crosstab(df['Survived'],df['Pclass'],normalize='columns')*100)
```

`<Axes: xlabel='Pclass', ylabel='Survived'>`
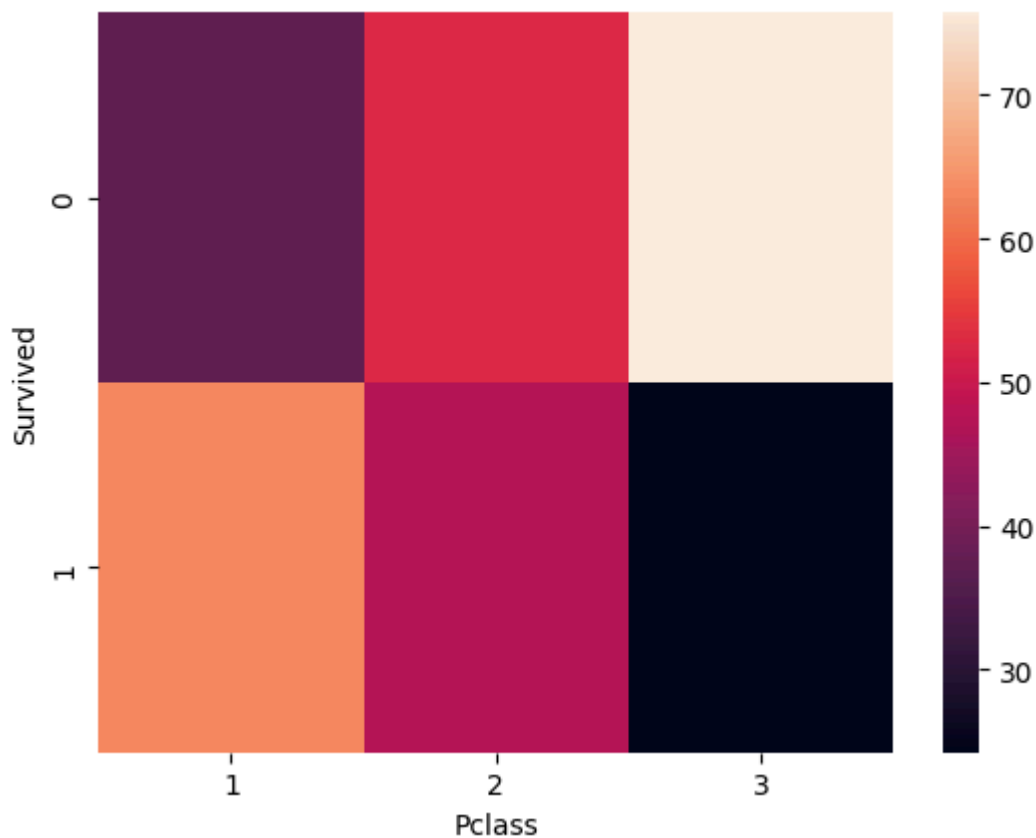
```python
pd.crosstab(df['Survived'],df['Sex'],normalize='columns')*100
```

Sex	female	male
Survived		
0	25.796178	81.109185
1	74.203822	18.890815

```python
pd.crosstab(df['Survived'],df['Embarked'],normalize='columns')*100
```

Embarked	C	Q	S
Survived			
0	44.642857	61.038961	66.304348
1	55.357143	38.961039	33.695652

```
In [54]:  pd.crosstab(df['Sex'],df['Embarked'],normalize='columns')*100
```

Out[54]:

Embarked	C	Q	S
Sex			
female	43.452381	46.753247	31.521739
male	56.547619	53.246753	68.478261

```
In [55]:  pd.crosstab(df['Pclass'],df['Embarked'],normalize='columns')*100
```

Out[55]:

Embarked	C	Q	S
Pclass			
1	50.595238	2.597403	19.720497
2	10.119048	3.896104	25.465839
3	39.285714	93.506494	54.813665

```
In [58]:  sns.boxplot(x='Survived', y='Age', data=df,hue='Survived')
```

Out[58]: <Axes: xlabel='Survived', ylabel='Age'>
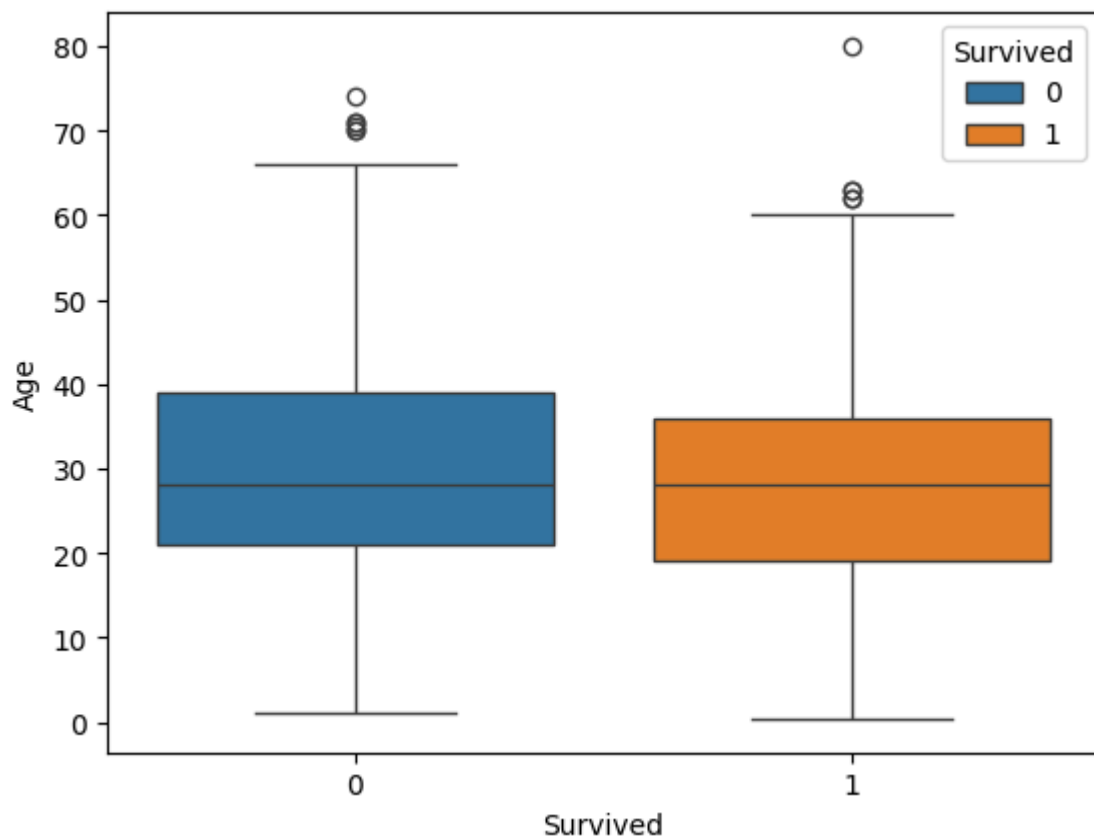

```
In [62]:  df.columns
```

Out[62]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
               'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
              dtype='object')
```

```
In [63]: df['SibSp'].value_counts()
```

```
Out[63]: SibSp
         0    608
         1    209
         2     28
         4     18
         3     16
         8      7
         5      5
         Name: count, dtype: int64
```

```
In [64]: df[df['SibSp']==8]
```

Out[64]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Far
159	160	0	3	Sage, Master. Thomas Henry	male	NaN	8	2	CA. 2343	69.5
180	181	0	3	Sage, Miss. Constance Gladys	female	NaN	8	2	CA. 2343	69.5
201	202	0	3	Sage, Mr. Frederick	male	NaN	8	2	CA. 2343	69.5
324	325	0	3	Sage, Mr. George John Jr	male	NaN	8	2	CA. 2343	69.5
792	793	0	3	Sage, Miss. Stella Anna	female	NaN	8	2	CA. 2343	69.5
846	847	0	3	Sage, Mr. Douglas Bullen	male	NaN	8	2	CA. 2343	69.5
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5

```
In [69]: df[df['Ticket']=='CA. 2343']
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Far
159	160	0	3	Sage, Master. Thomas Henry	male	NaN	8	2	CA. 2343	69.5
180	181	0	3	Sage, Miss. Constance Gladys	female	NaN	8	2	CA. 2343	69.5
201	202	0	3	Sage, Mr. Frederick	male	NaN	8	2	CA. 2343	69.5
324	325	0	3	Sage, Mr. George John Jr	male	NaN	8	2	CA. 2343	69.5
792	793	0	3	Sage, Miss. Stella Anna	female	NaN	8	2	CA. 2343	69.5
846	847	0	3	Sage, Mr. Douglas Bullen	male	NaN	8	2	CA. 2343	69.5
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5

In [65]:
```python
df1=pd.read_csv(r"D:\Stranger Things\New folder\you_ex_data seience\YOUTUBE\EDA_
```

In [73]:
```python
df=pd.concat([df, df1])
```

In [74]:
```python
df[df['Ticket']=='CA. 2343']
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Far
159	160	0.0	3	Sage, Master. Thomas Henry	male	NaN	8	2	CA. 2343	69.5
180	181	0.0	3	Sage, Miss. Constance Gladys	female	NaN	8	2	CA. 2343	69.5
201	202	0.0	3	Sage, Mr. Frederick	male	NaN	8	2	CA. 2343	69.5
324	325	0.0	3	Sage, Mr. George John Jr	male	NaN	8	2	CA. 2343	69.5
792	793	0.0	3	Sage, Miss. Stella Anna	female	NaN	8	2	CA. 2343	69.5
846	847	0.0	3	Sage, Mr. Douglas Bullen	male	NaN	8	2	CA. 2343	69.5
863	864	0.0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5
188	1080	NaN	3	Sage, Miss. Ada	female	NaN	8	2	CA. 2343	69.5
342	1234	NaN	3	Sage, Mr. John George	male	NaN	1	9	CA. 2343	69.5
360	1252	NaN	3	Sage, Master. William Henry	male	14.5	8	2	CA. 2343	69.5
365	1257	NaN	3	Sage, Mrs. John (Annie Bullen)	female	NaN	1	9	CA. 2343	69.5

In [82]: `df['indi_fare']=df['Fare']/(df['SibSp']+df['Parch']+1)`

In [83]: `df.sample(10)`

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
515	516	0.0	1	Walker, Mr. William Anderson	male	47.0	0	0	36967	34
309	1201	NaN	3	Hansen, Mrs. Claus Peter (Jennie L Howard)	female	45.0	1	0	350026	14
65	957	NaN	2	Corey, Mrs. Percy C (Mary Phyllis Elizabeth Mi...	female	NaN	0	0	F.C.C. 13534	2
346	347	1.0	2	Smith, Miss. Marion Elsie	female	40.0	0	0	31418	13
258	1150	NaN	2	Bentham, Miss. Lilian W	female	19.0	0	0	28404	13
697	698	1.0	3	Mullens, Miss. Katherine "Katie"	female	NaN	0	0	35852	1
219	220	0.0	2	Harris, Mr. Walter	male	30.0	0	0	W/C 14208	1(
703	704	0.0	3	Gallagher, Mr. Martin	male	25.0	0	0	36864	1
68	69	1.0	3	Andersson, Miss. Erna Alexandra	female	17.0	4	2	3101281	1
580	581	1.0	2	Christy, Miss. Julie Rachel	female	25.0	1	1	237789	3(

In [84]: df

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
413	1305	NaN	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236
414	1306	NaN	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758
415	1307	NaN	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262
416	1308	NaN	3	Ware, Mr. Frederick	male	NaN	0	0	359309
417	1309	NaN	3	Peter, Master. Michael J	male	NaN	1	1	2668

1309 rows × 13 columns

```
In [85]: df['Family_si']=df['SibSp']+df['Parch']+1
```

```
In [88]: df.sample(6)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
733	734	0.0	2	Berriman, Mr. William John	male	23.0	0	0	28425	13.0
698	699	0.0	1	Thayer, Mr. John Borland	male	49.0	1	1	17421	110.8
64	65	0.0	1	Stewart, Mr. Albert A	male	NaN	0	0	PC 17605	27.7
710	711	1.0	1	Mayne, Mlle. Berthe Antonine ("Mrs de Villiers")	female	24.0	0	0	PC 17482	49.5
56	57	1.0	2	Rugg, Miss. Emily	female	21.0	0	0	C.A. 31026	10.5
270	1162	NaN	1	McCaffry, Mr. Thomas Francis	male	46.0	0	0	13050	75.2

In [95]:
```python
df["Surname"]=df['Name'].str.split(',').str.get(0)
```

In [113...
```python
df['Title']=df['Name'].str.split(',').str.get(1).str.split(' ').str.get(1)
```

In [115...
```python
df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.0	1	0	PC 17599
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
413	1305	NaN	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236
414	1306	NaN	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758
415	1307	NaN	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262
416	1308	NaN	3	Ware, Mr. Frederick	male	NaN	0	0	359309
417	1309	NaN	3	Peter, Master. Michael J	male	NaN	1	1	2668

1309 rows × 16 columns

```python
df['Title'].value_counts()
```

```
Out[105...   Title
             Mr.           757
             Miss.         260
             Mrs.          197
             Master.        61
             Rev.            8
             Dr.             8
             Col.            4
             Mlle.           2
             Major.          2
             Ms.             2
             Lady.           1
             Sir.            1
             Mme.            1
             Don.            1
             Capt.           1
             the             1
             Jonkheer.       1
             Dona.           1
             Name: count, dtype: int64
```

In [117...
```python
df["Title"] = df['Title'].str.replace(r'(Col\.|Major\.|Don\.|Capt\.|the|Jonkheer
```

In [118...
```python
df['Title'].value_counts()
```

```
Out[118...   Title
             Mr.           757
             Miss.         260
             Mrs.          197
             Master.        61
             other          10
             Rev.            8
             Dr.             8
             Ms.             2
             Mlle.           2
             Mme.            1
             Lady.           1
             Sir.            1
             Dona.           1
             Name: count, dtype: int64
```

In [123...
```python
pd.crosstab(df['Survived'],df['Title'],normalize='columns')*100
```

Out[123...

Title	Dr.	Lady.	Master.	Miss.	Mlle.	Mme.	Mr.	Mrs.	Ms.	R(
Survived										
0.0	57.142857	0.0	42.5	30.21978	0.0	0.0	84.332689	20.8	0.0	100
1.0	42.857143	100.0	57.5	69.78022	100.0	100.0	15.667311	79.2	100.0	(

In [124...
```python
df['Cabin'].isnull().sum()
```

Out[124...   1014

In [126...
```python
df['Cabin'].fillna('O',inplace=True)
```

```
In [127…    df['Cabin'].value_counts()
```

```
Out[127…    Cabin
            O                    1014
            C23 C25 C27             6
            B57 B59 B63 B66         5
            G6                      5
            F33                     4
                           ...
            A14                     1
            E63                     1
            E12                     1
            E38                     1
            C105                    1
            Name: count, Length: 187, dtype: int64
```

```
In [131…    df['deck']=df['Cabin'].str[0]
```

```
In [133…    df['deck'].value_counts()
```

```
Out[133…    deck
            O    1014
            C      94
            B      65
            D      46
            E      41
            A      22
            F      21
            G       5
            T       1
            Name: count, dtype: int64
```

```
In [134…    df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450
...	...	...	...	...	...	...	...	...	...
413	1305	NaN	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236
414	1306	NaN	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758
415	1307	NaN	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262
416	1308	NaN	3	Ware, Mr. Frederick	male	NaN	0	0	359309
417	1309	NaN	3	Peter, Master. Michael J	male	NaN	1	1	2668

1309 rows × 17 columns

In [137...
```python
pd.crosstab(df['deck'],df['Pclass'])
```

deck	Pclass 1	2	3
A	22	0	0
B	65	0	0
C	94	0	0
D	40	6	0
E	34	4	3
F	0	13	8
G	0	0	5
O	67	254	693
T	1	0	0

In [ ]: