Assignment No 9 : Data Visualization 2

Import required Libraries

In [1]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

Read csv into dataframe

In [2]:
```python
df = pd.read_csv('train.csv')
```

In [3]:
```python
df
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	
...	...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	

891 rows × 12 columns

Data preprocessing

In [4]:
```python
df.shape
```

Out[4]: (891, 12)

In [5]:
```python
df.columns
```

Out[5]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [6]:
```python
df.dtypes
```

Out[6]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
```

```
Embarked         object
dtype: object
```

In [7]:
```python
df.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 [8]:
```python
df.isnull()
```

Out[8]:

	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 [9]:
```python
df.isnull().sum()
```

Out[9]:
```
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 [10]:
```python
df.describe()
```

Out[10]:

	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

Data Visualization

In [12]:
```python
sns.countplot(df['Sex'])
```

Out[12]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x8345203508>
```
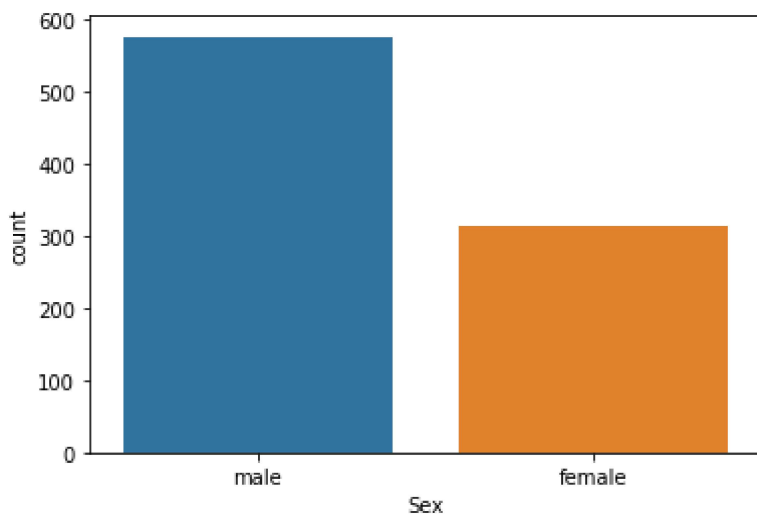

In [13]:
```python
sns.countplot(df['Survived'])
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x8345288b88>
```
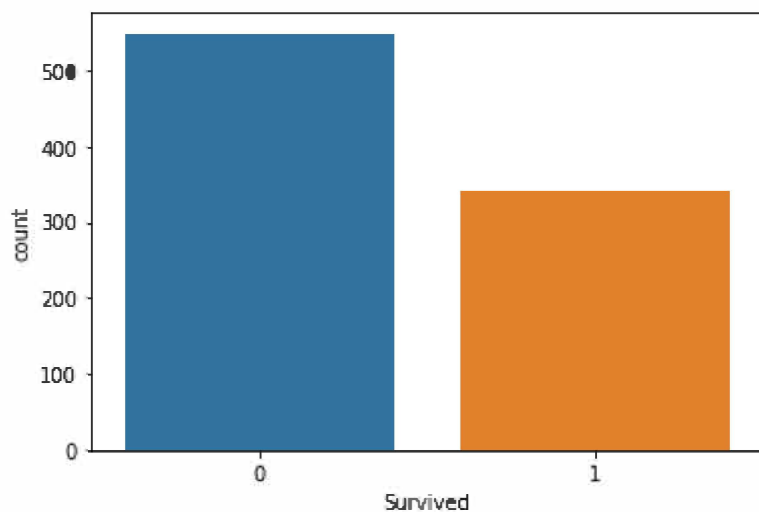
```
In [15]:   df['Sex'].value_counts().plot(kind = 'pie', autopct = '%.2f')
```
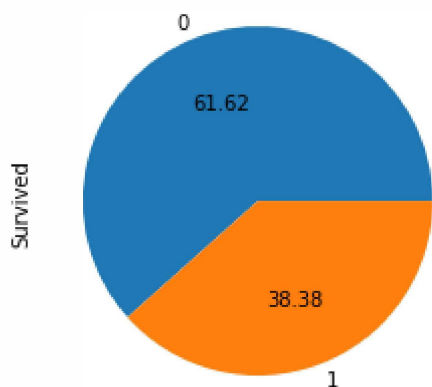
Out[15]: <matplotlib.axes._subplots.AxesSubplot at 0x8345593d08>
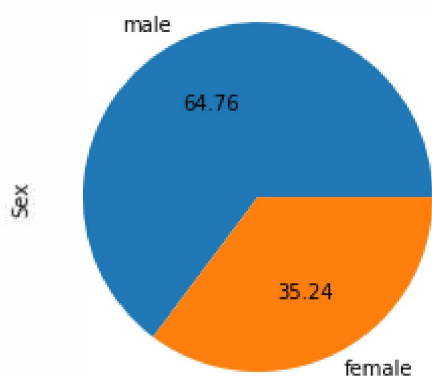

```
In [16]:   df['Survived'].value_counts().plot(kind = 'pie', autopct = '%.2f')
```
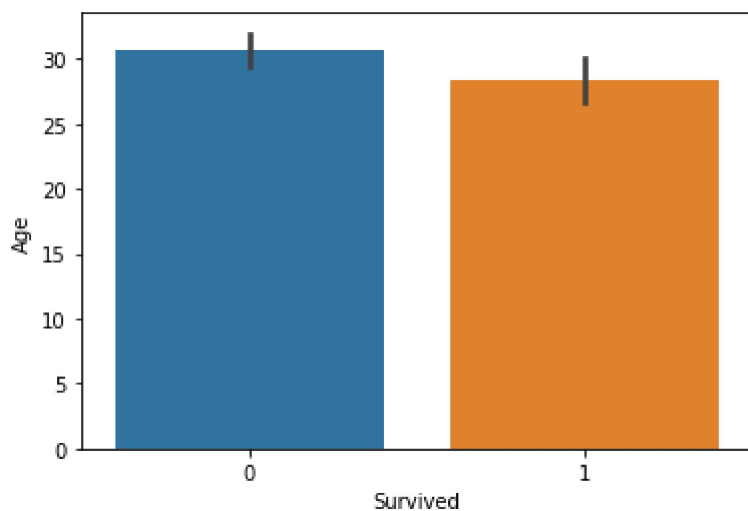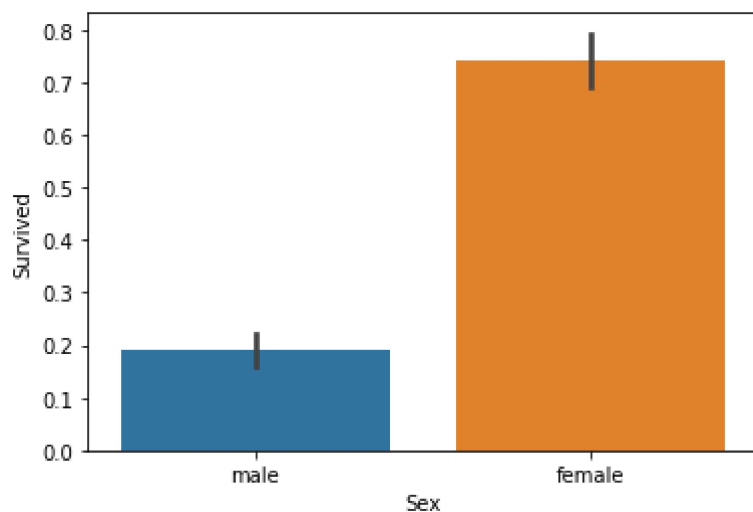
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x83455b8ec8>


```
In [17]:   sns.barplot(df['Survived'], df['Age'])
```
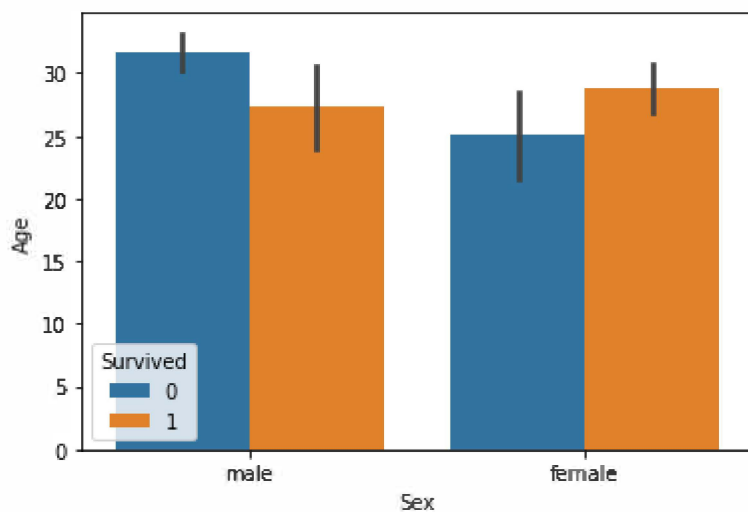
Out[17]:   <matplotlib.axes._subplots.AxesSubplot at 0x8345bd2788>

In [19]:
```python
sns.barplot(df['Sex'], df['Survived'])
```

Out[19]:   <matplotlib.axes._subplots.AxesSubplot at 0x8345bcda08>

In [20]:
```python
sns.barplot(df['Sex'],df['Age'], hue = df['Survived'])
```
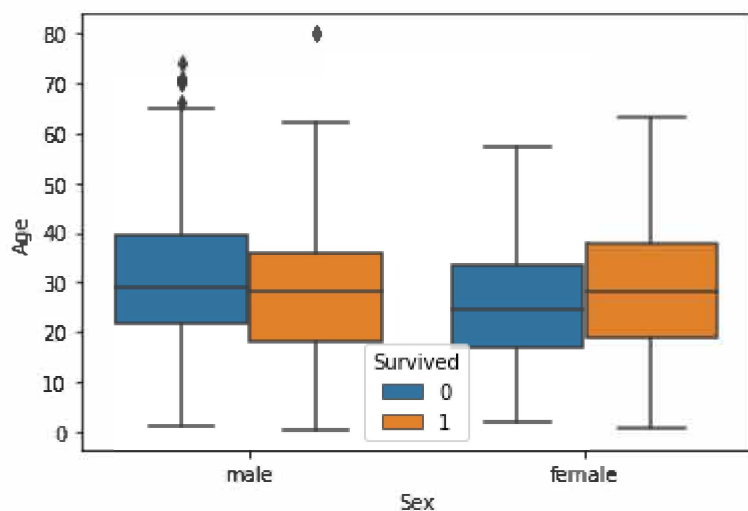
Out[20]:   <matplotlib.axes._subplots.AxesSubplot at 0x8345cf55c8>

In [21]:
```python
sns.boxplot(df['Sex'], df['Age'])
```

Out[21]: `<matplotlib.axes._subplots.AxesSubplot at 0x8345d66c48>`
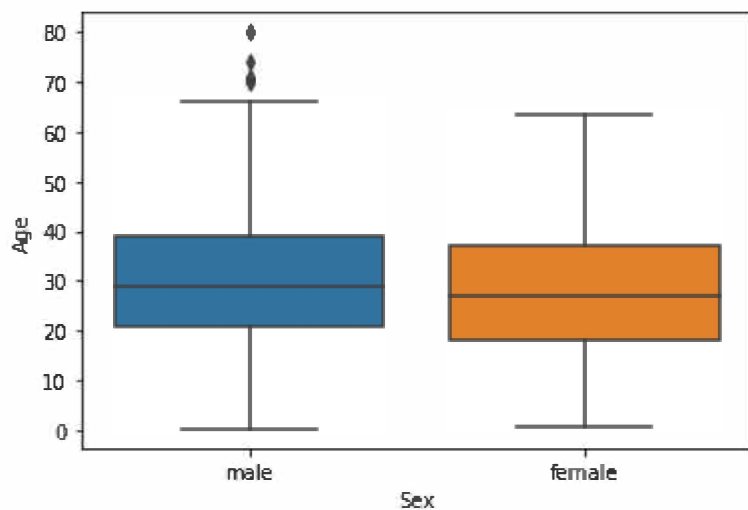
In [4]:
```python
sns.boxplot(df['Sex'], df['Age'], df['Survived'])
```

Out[4]: `<matplotlib.axes._subplots.AxesSubplot at 0x44b8ecf788>`

In [24]:
```python
pd.crosstab(df['Sex'], df['Survived'])
```

Out[24]:

Survived	0	1
Sex		
female	81	233
male	468	109

In [25]:
```python
pd.crosstab(df['Age'], df['Survived'])
```

Out[25]:

Survived	0	1
Age		
0.42	0	1
0.67	0	1
0.75	0	2
0.83	0	2
0.92	0	1
...	...	...
70.00	2	0
70.50	1	0
71.00	2	0
74.00	1	0
80.00	0	1

88 rows × 2 columns

In [26]:
```python
sns.heatmap(pd.crosstab(df['Sex'], df['Survived']))
```

Out[26]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x8345eaa848>
```
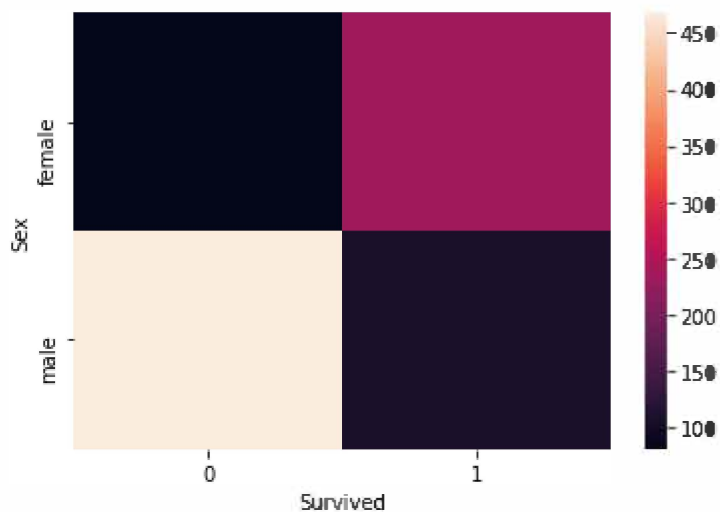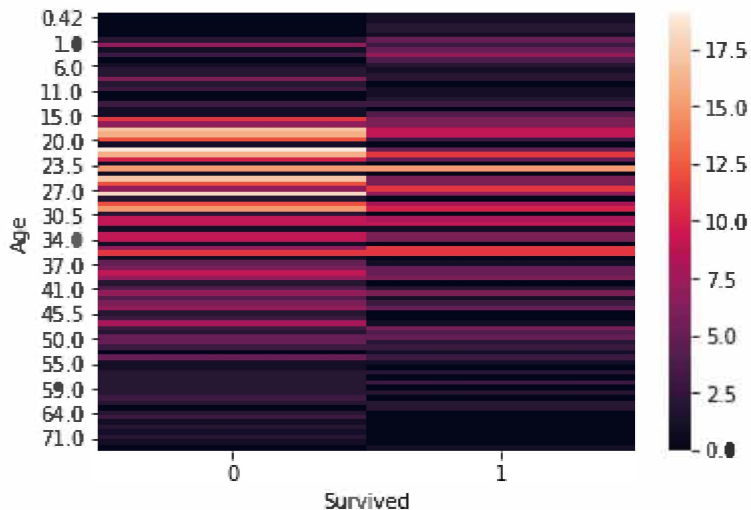
In [27]:
```python
sns.heatmap(pd.crosstab(df['Age'], df['Survived']))
```
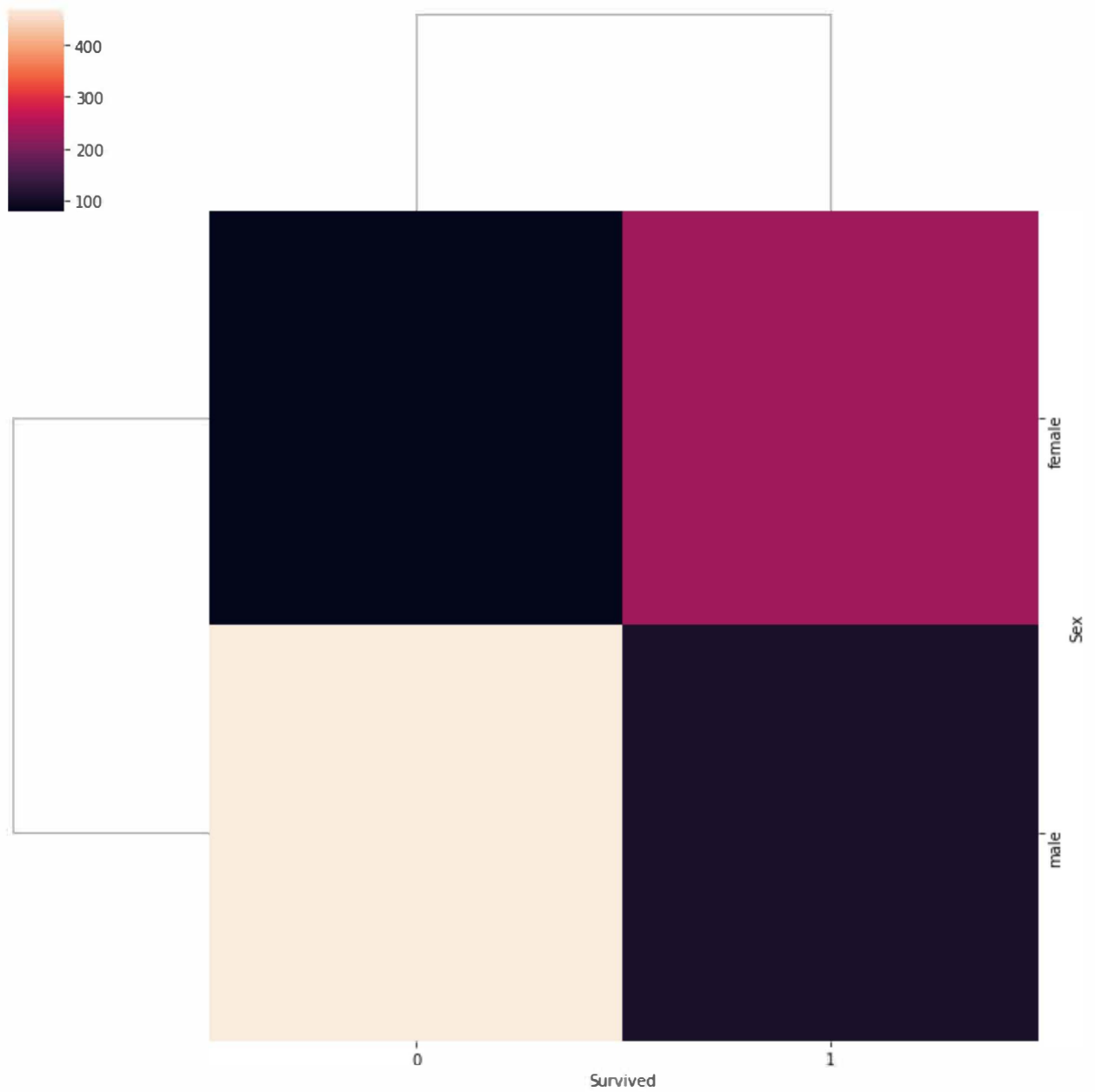
Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x8346973308>

In [28]:
```python
sns.clustermap(pd.crosstab(df['Sex'], df['Survived']))b
```

Out[28]: <seaborn.matrix.ClusterGrid at 0x8345c8bd48>

```
In [29]:   sns.clustermap(pd.crosstab(df['Age'], df['Survived']))
```

Out[29]:   <seaborn.matrix.ClusterGrid at 0x83469736c8>