ML-6

After getting the data ,The first thing you should do is

Asking basic questions

1) How big is the data?

Ans. shape

2) How does the data look like ?

Ans. head() / [sample()]→ this will give you random 5 rows or arguments you give

3)what is the data type of cols?

Ans. info()

4)Are there any missing values?

Ans. isnull()

5)How does the data look mathematically?

Ans. describe()

6) Are there duplicate values?

Ans. duplicated()

7)How is the correlation between cols?

Ans. corr() → arranges values from -1 to 1 and it tells the correlation between the cols.

-1 means that it is inversely proportional .

EDA[Exploratory Data Analysis]

1) Univariate Analysis

looks at single variable

understand the data distribution and identity anu outliers.

Method of representation

Numeric Variable

Histogram

violin plot

Box plot

distplot

Categorical data

bar graph

count plot

pie chart

2) Bivariate Analysis

looks at two variables

identify the relation between the two variables and also see if any pattern exists

method of representation

Numeric Variables

scatter plot

joint plot

Categorical Data

count plot

heat map

cluster map

Numeric and categorical

bar plot

box plot

violin plot

distplot

3)Multivariate Analysis

looks at multiple variables.

can help us finding the relationship between several variables and also find any complex patterns if exists

Method of representation

stacked bar plot

pair plot

heatmap

histogram

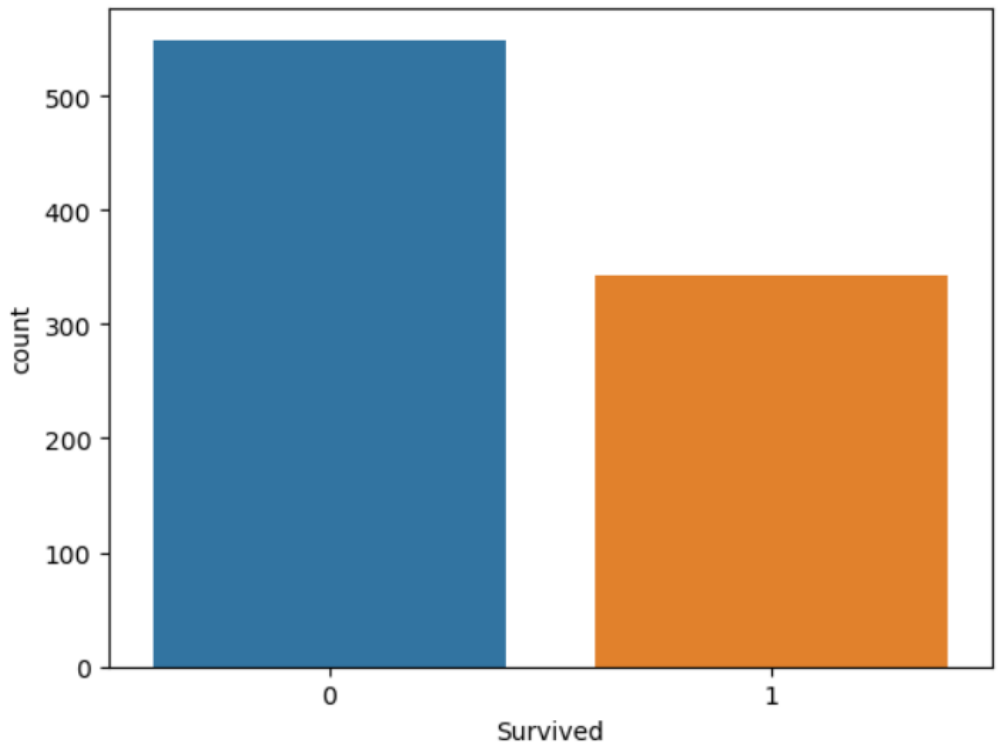
▼ EDA in univariate

```
df=pd.read_csv("titanic.csv")
```

Countplot

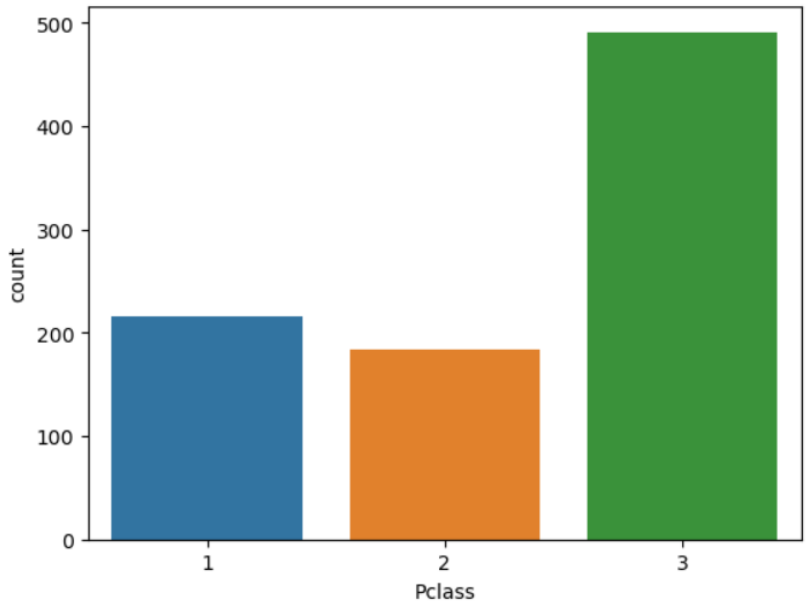
```
In [27]: sns.countplot(x='Survived' , data=df)
```

```
Out[27]: <Axes: xlabel='Survived', ylabel='count'>
```



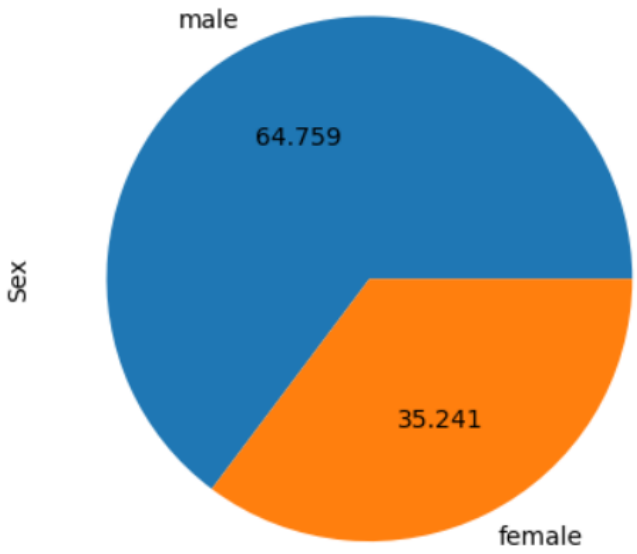
In [30]:
```
sns.countplot(x='Pclass' , data=df)
df['Pclass'].value_counts()
```

Out[30]:
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```



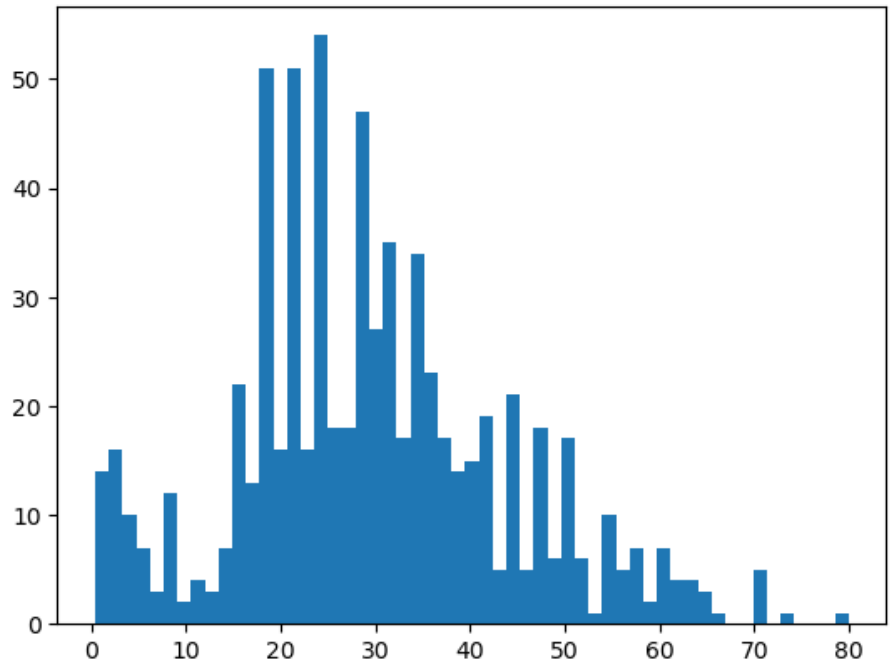
Pie chart

```
In [44]: df['Sex'].value_counts().plot(kind='pie', autopct="%.3f")
Out[44]: <Axes: ylabel='Sex'>
```



Histogram

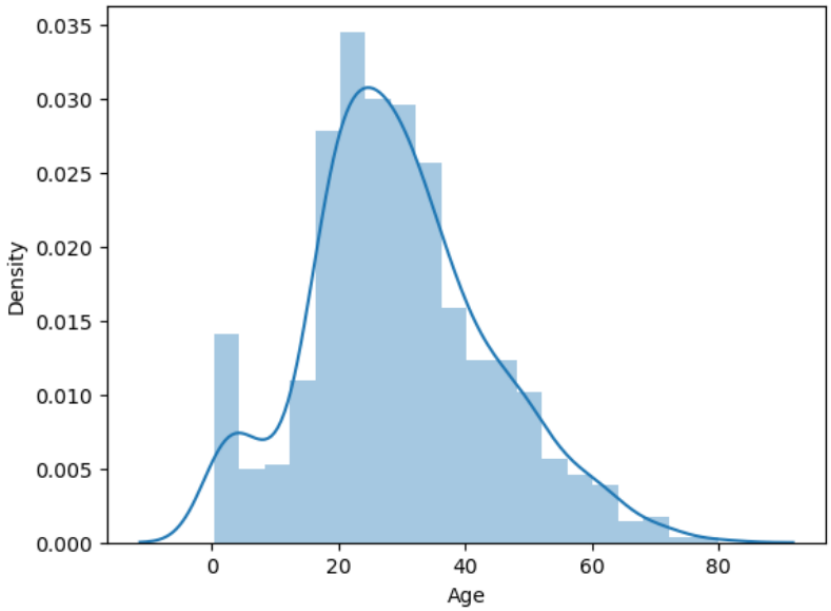
```
In [55]: plt.hist(df['Age'],bins=55)

Out[55]: (array([14., 16., 10.,  7.,  3., 12.,  2.,  4.,  3.,  7., 22., 13., 51.,
                 16., 51., 16., 54., 18., 18., 47., 27., 35., 17., 34., 23., 17.,
                 14., 15., 19.,  5., 21.,  5., 18.,  6., 17.,  6.,  1., 10.,  5.,
                  7.,  2.,  7.,  4.,  4.,  3.,  1.,  0.,  0.,  5.,  0.,  1.,  0.,
                  0.,  0.,  1.]),
          array([ 0.42      ,  1.86690909,  3.31381818,  4.76072727,  6.20763636,
                  7.65454545,  9.10145455, 10.54836364, 11.99527273, 13.44218182,
                 14.88909091, 16.336     , 17.78290909, 19.22981818, 20.67672727,
                 22.12363636, 23.57054545, 25.01745455, 26.46436364, 27.91127273,
                 29.35818182, 30.80509091, 32.252     , 33.69890909, 35.14581818,
                 36.59272727, 38.03963636, 39.48654545, 40.93345455, 42.38036364,
                 43.82727273, 45.27418182, 46.72109091, 48.168     , 49.61490909,
                 51.06181818, 52.50872727, 53.95563636, 55.40254545, 56.84945455,
                 58.29636364, 59.74327273, 61.19018182, 62.63709091, 64.084     ,
                 65.53090909, 66.97781818, 68.42472727, 69.87163636, 71.31854545,
                 72.76545455, 74.21236364, 75.65927273, 77.10618182, 78.55309091,
                 80.        ]),
          <BarContainer object of 55 artists>)
```



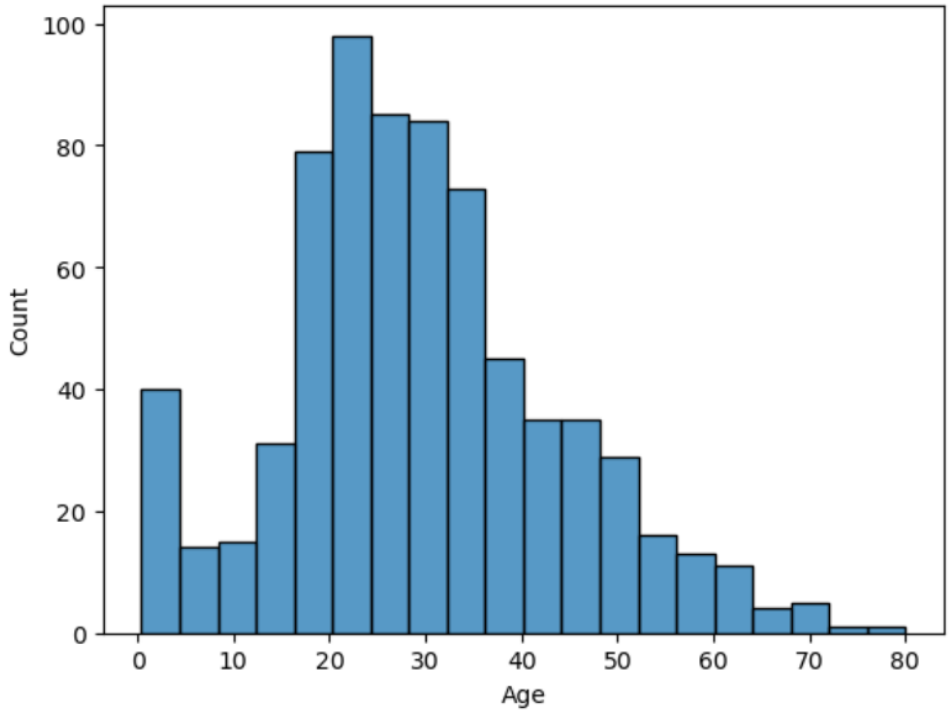
Ditplot /Histplot

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

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

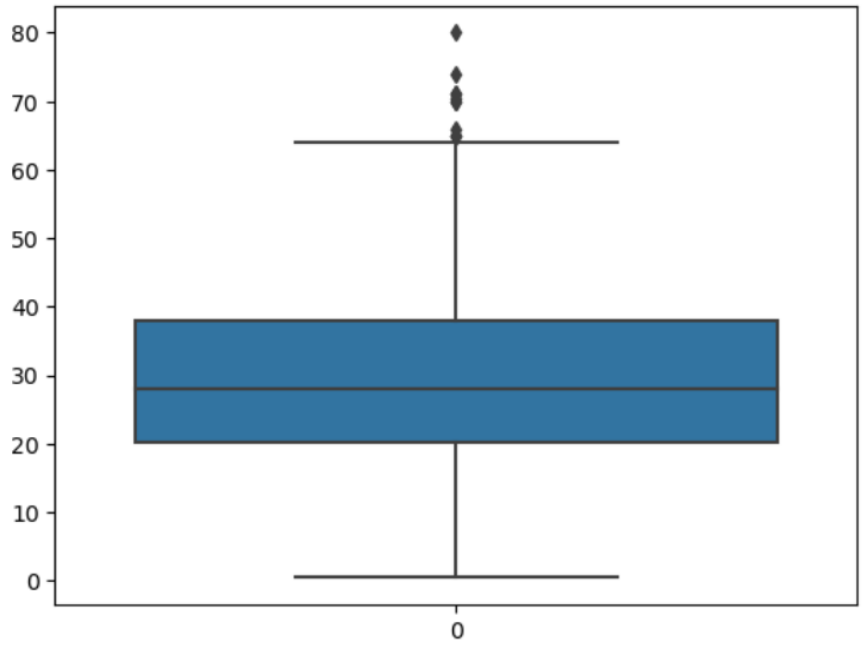


```
In [59]: sns.histplot(df['Age'])
Out[59]: <Axes: xlabel='Age', ylabel='Count'>
```



Boxplot

```
In [63]: sns.boxplot(df['Age'])
Out[63]: <Axes: >
```

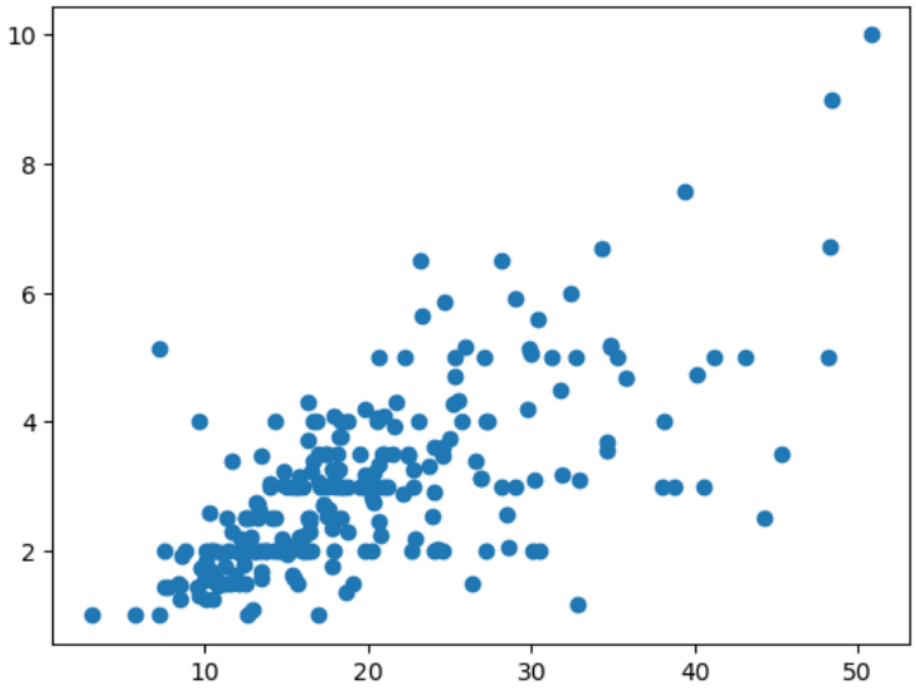


EDA in bivariate

Scatter plot

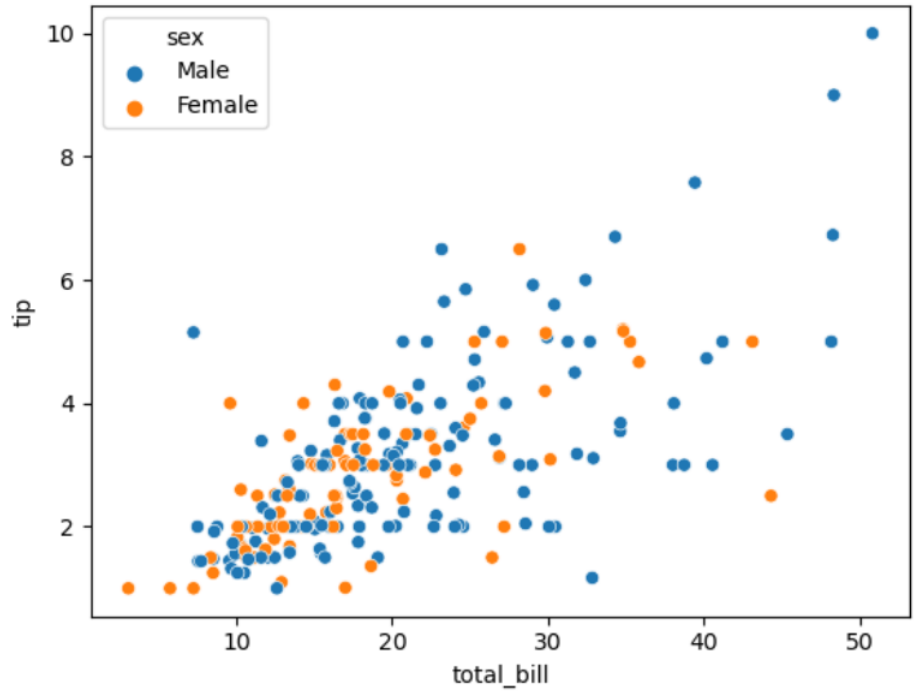
In [98]:
```
plt.scatter(x=bill['total_bill'],y=bill['tip'])
```

Out[98]:
```
<matplotlib.collections.PathCollection at 0x1602d28e2d0>
```



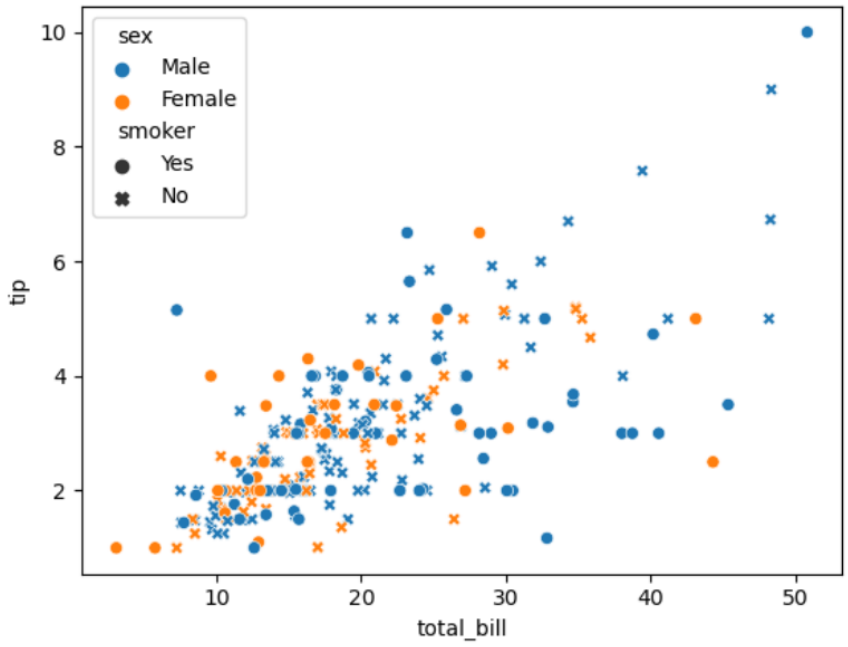
In [103]:
```
sns.scatterplot(x=bill['total_bill'],y=bill['tip'],hue=bill['sex'])
```

Out[103]: <Axes: xlabel='total_bill', ylabel='tip'>



In [106]:
```
sns.scatterplot(x=bill['total_bill'],y=bill['tip'],hue=bill['sex'],style=bill['smoker'])
```

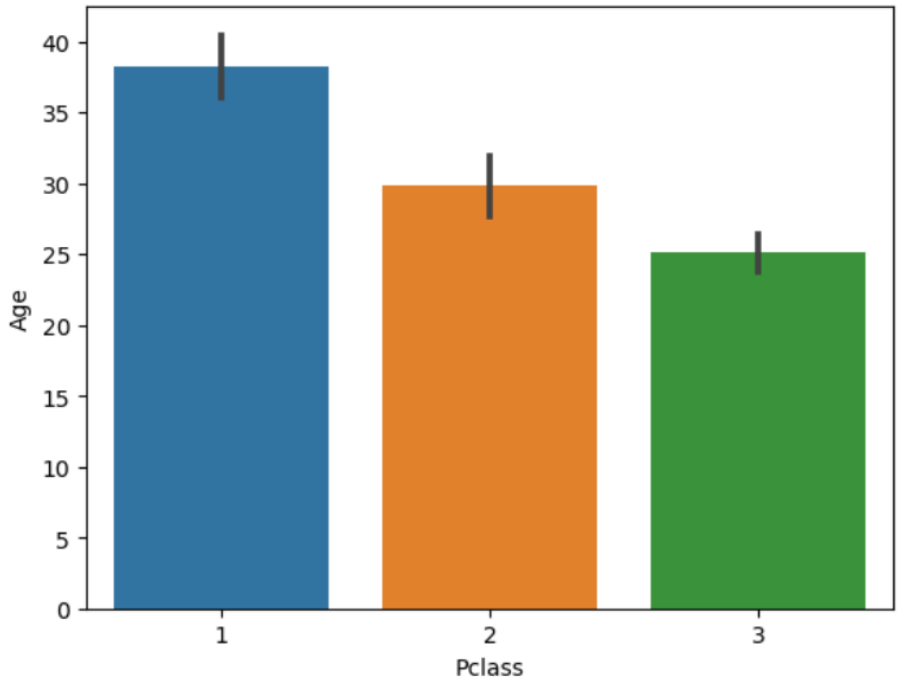
Out[106]: <Axes: xlabel='total_bill', ylabel='tip'>



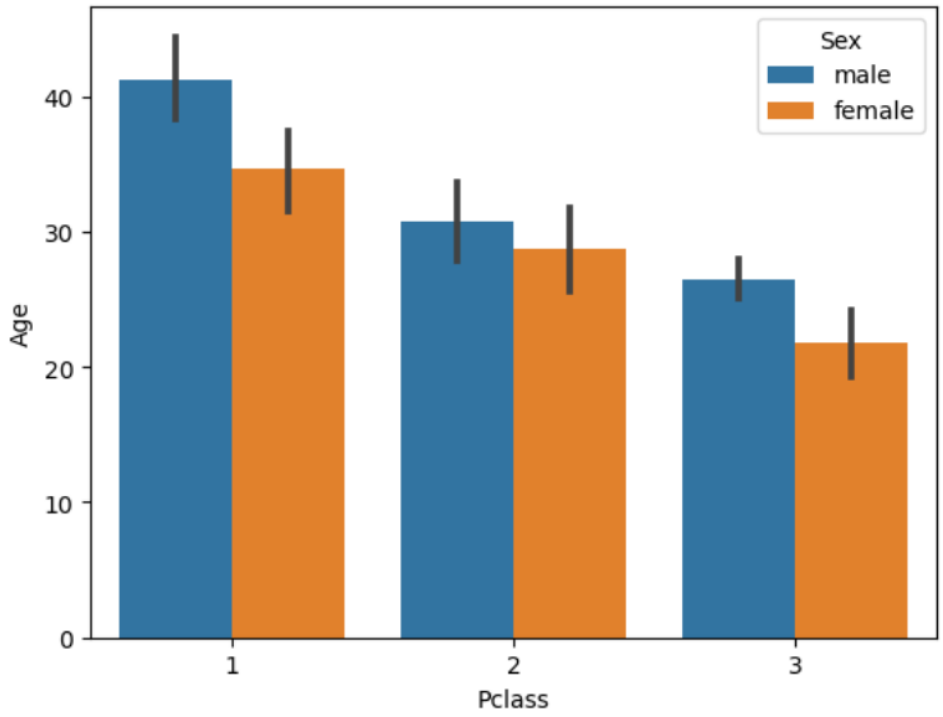
Bar plot

```
In [109]: sns.barplot(x=df['Pclass'],y=df['Age'])
```

```
Out[109]: <Axes: xlabel='Pclass', ylabel='Age'>
```



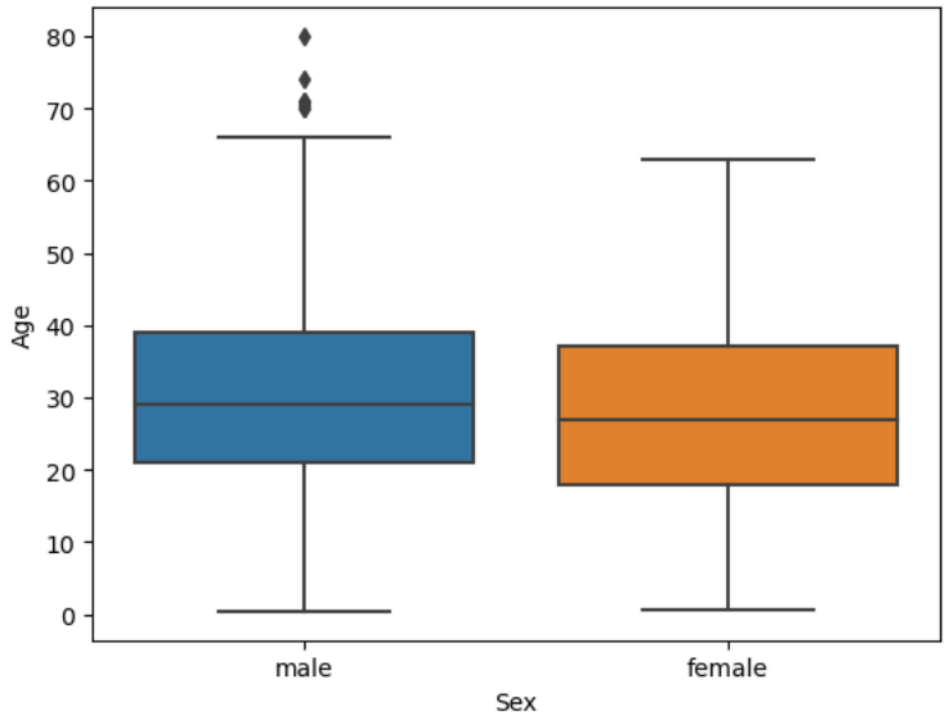
```
In [112]: sns.barplot(x=df['Pclass'],y=df['Age'],hue=df['Sex'])
Out[112]: <Axes: xlabel='Pclass', ylabel='Age'>
```



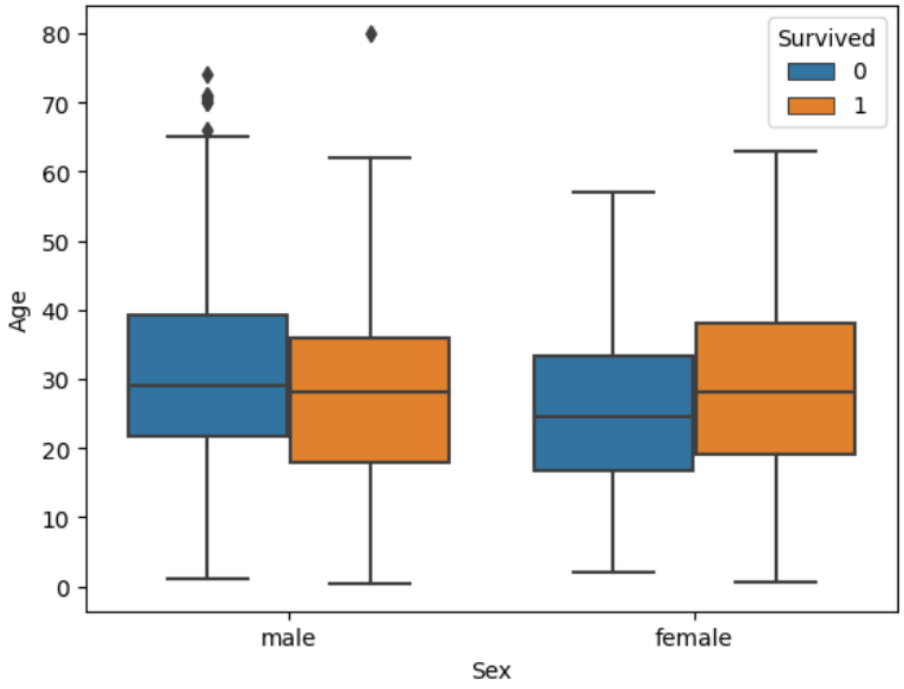
Box plot

```
In [119]: sns.boxplot(x=df['Sex'],y=df['Age'])
```

```
Out[119]: <Axes: xlabel='Sex', ylabel='Age'>
```

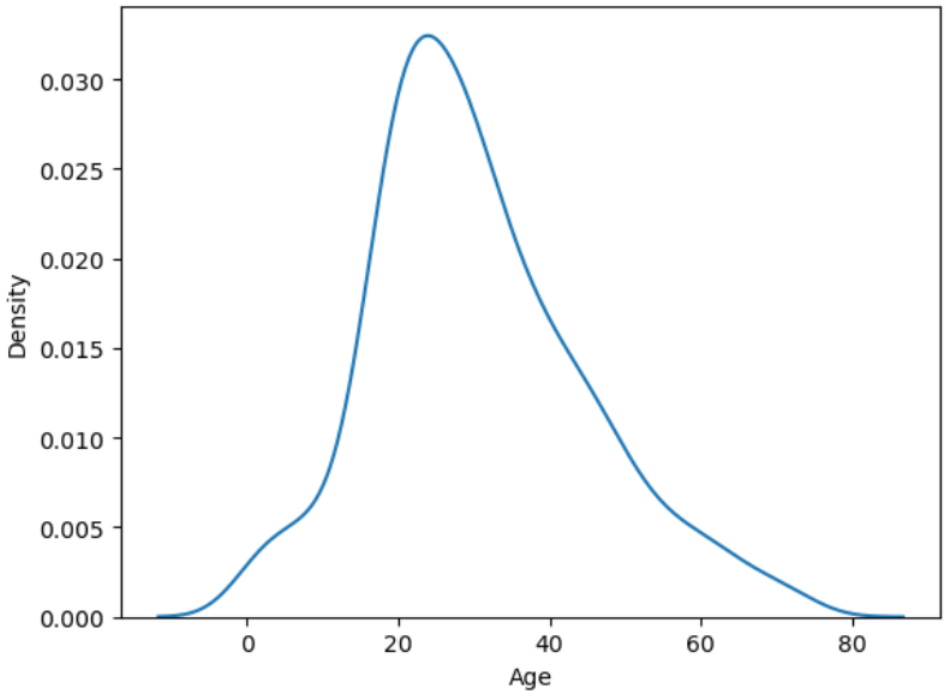


```
In [117]: sns.boxplot(x=df['Sex'],y=df['Age'],hue=df['Survived'])
Out[117]: <Axes: xlabel='Sex', ylabel='Age'>
```



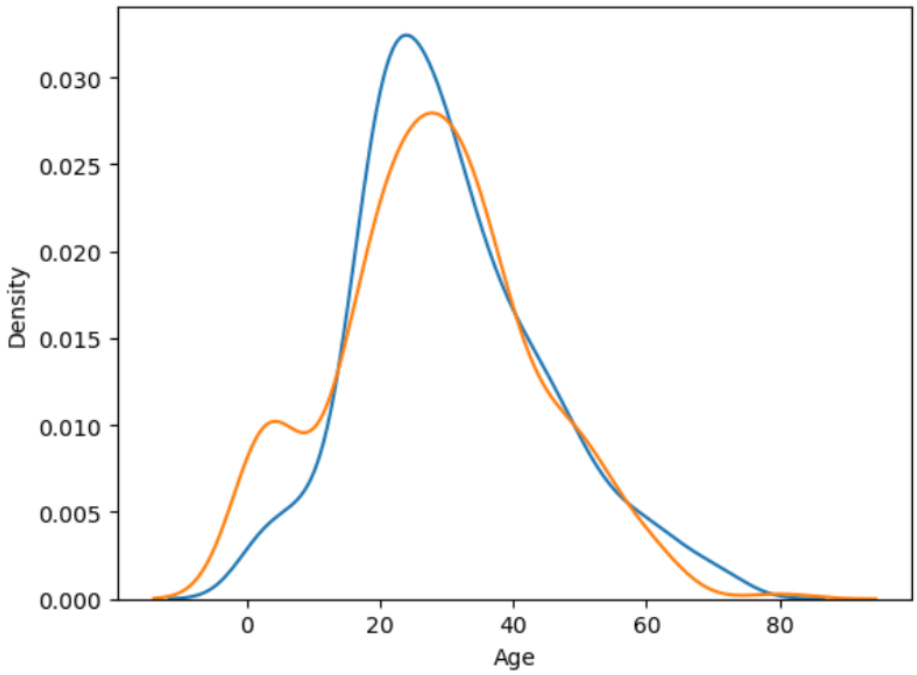
Distplot/Histplot

```
sns.distplot(df[df['Survived']==0]['Age'],hist=False)
```




```
sns.distplot(df[df['Survived']==0]['Age'],hist=False)
sns.distplot(df[df['Survived']==1]['Age'],hist=False)
```

Out[131]: <Axes: xlabel='Age', ylabel='Density'>

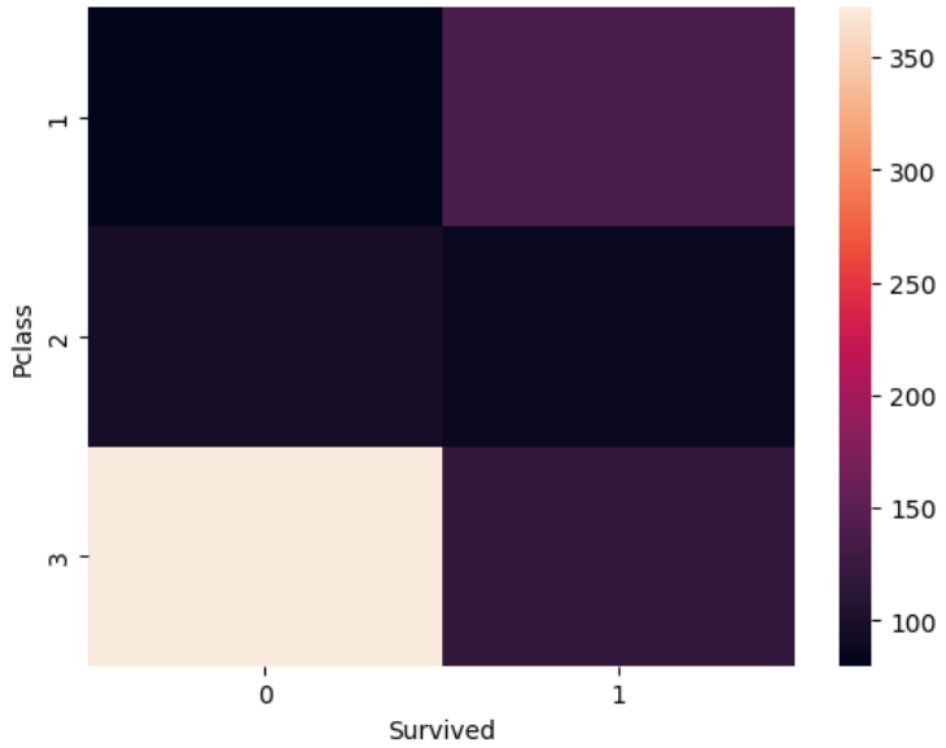


Heatplot

```
In [138]: a=pd.crosstab(df['Pclass'],df['Survived'])
```

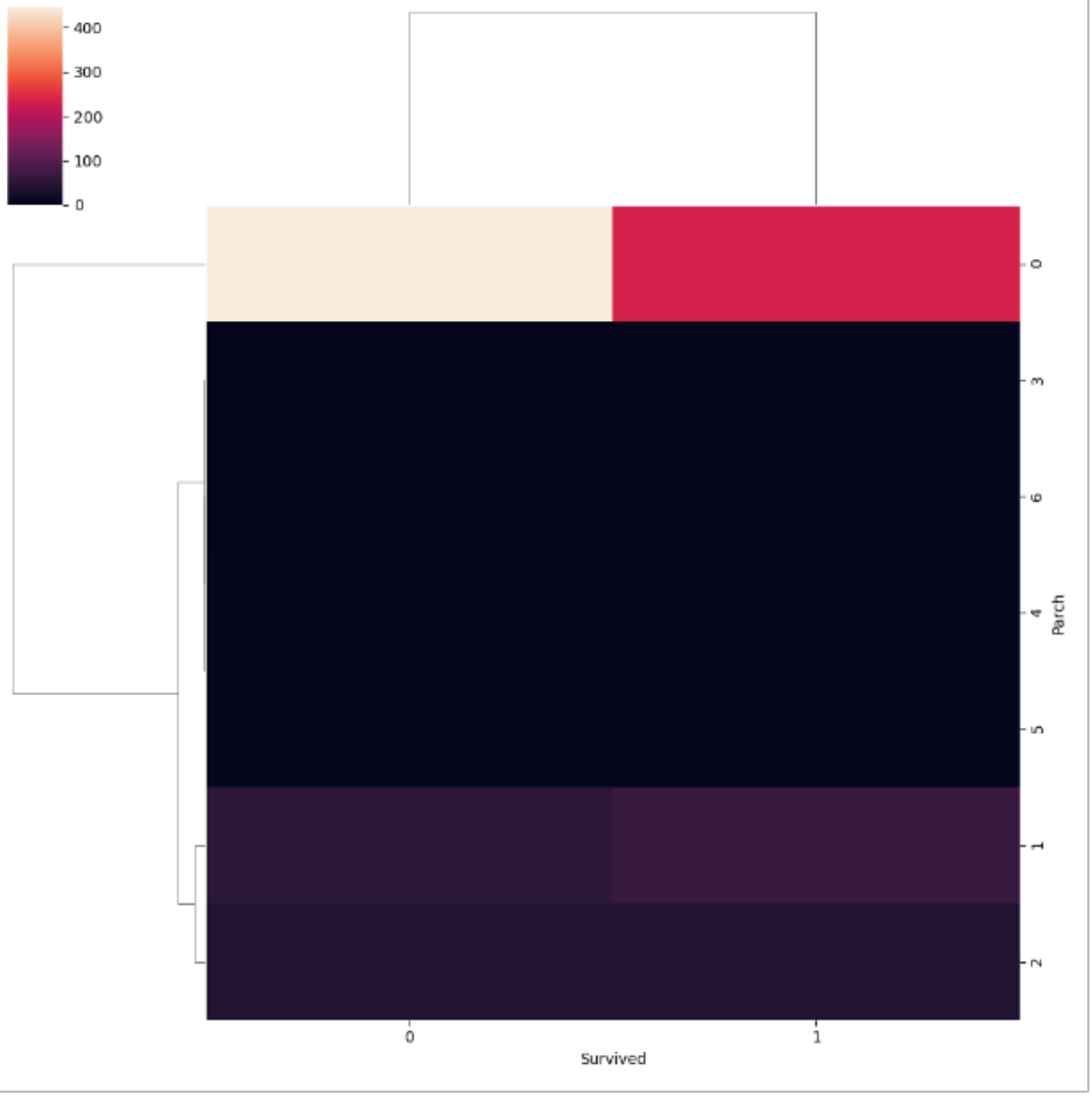
```
In [139]: sns.heatmap(a)
```

```
Out[139]: <Axes: xlabel='Survived', ylabel='Pclass'>
```



clustermap

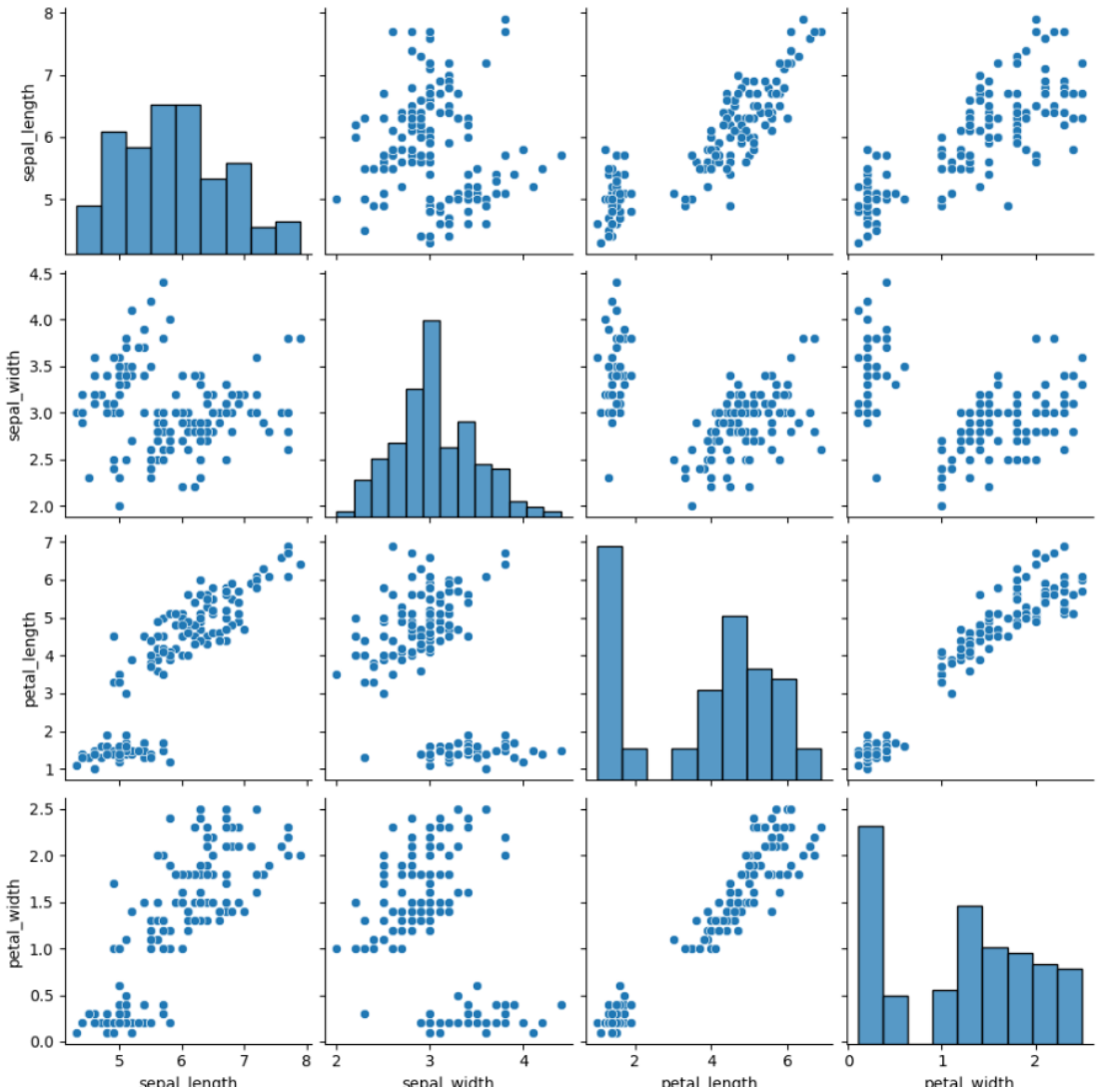
```
In [140]: a=pd.crosstab(df['Parch'],df['Survived'])

In [142]: sns.clustermap(a)

Out[142]: <seaborn.matrix.ClusterGrid at 0x16030bbce90>
```



Pair plot

```
In [149]: sns.pairplot(irisflwr)
```

```
Out[149]: <seaborn.axisgrid.PairGrid at 0x16030bdc090>
```



```
In [151]: sns.pairplot(irisflwr,hue='species')
```

```
Out[151]: <seaborn.axisgrid.PairGrid at 0x16033cf09d0>
```