Experiment No. 9

Aim: Data Visualization II:

1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each gender along with the information about whether they survived or not. (Column names : 'sex' and 'age').

2. Write observations on the inference from the above statistics.

Theory: Exploratory Data Analysis

There are various techniques to understand your data, And the basic need is you should have the knowledge of Numpy for mathematical operations and Pandas for data manipulation. We are using Titanic dataset. For demonstrating some of the techniques we will also use an inbuilt dataset of seaborn as tips data which explains the tips each waiter gets from different customers.

Import libraries and loading Data

```
import numpy as np
import pandas pd
import matplotlib.pyplot as plt
import seaborn as sns
from seaborn import load_dataset
#titanic dataset
data = pd.read_csv("titanic_train.csv")
#tips dataset
tips = load_dataset("tips")
```

Univariate Analysis

Univariate analysis is the simplest form of analysis where we explore a single variable. Univariate analysis is performed to describe the data in a better way. we perform Univariate analysis of Numerical and categorical variables differently because plotting uses different plots.

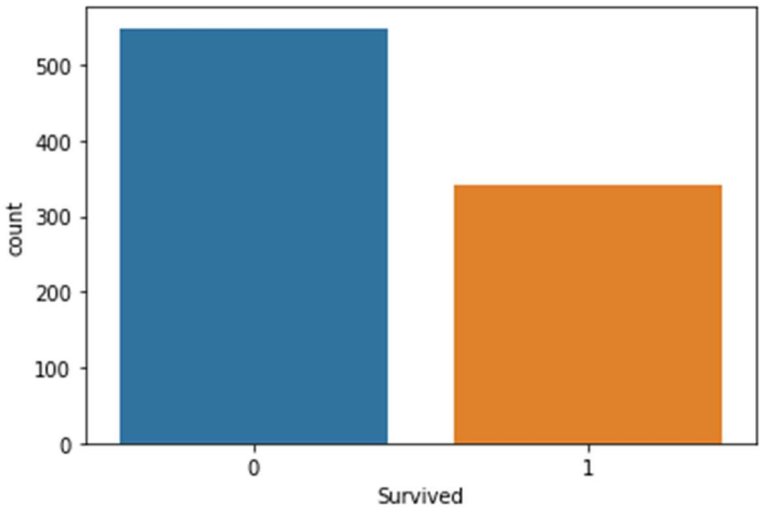
Categorical Data:

A variable that has text-based information is referred to as categorical variables. Now following are various plots which we can use for visualizing Categorical data.

1) CountPlot:

Countplot is basically a count of frequency plot in form of a bar graph. It plots the count of each category in a separate bar. When we use the pandas' value counts function on any column. It is the same visual form of the value counts function. In our data-target variable is survived and it is categorical so plot a countplot of this.

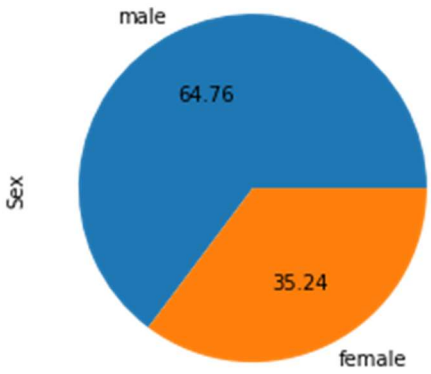
```
sns.countplot(data['Survived'])
plt.show()
```



2) Pie Chart:

The pie chart is also the same as the countplot, only gives us additional information about the percentage presence of each category in data means which category is getting how much weightage in data. Now we check about the Sex column, what is a percentage of Male and Female members traveling.

```
data['Sex'].value_counts().plot(kind="pie", autopct="%.2f")
plt.show()
```



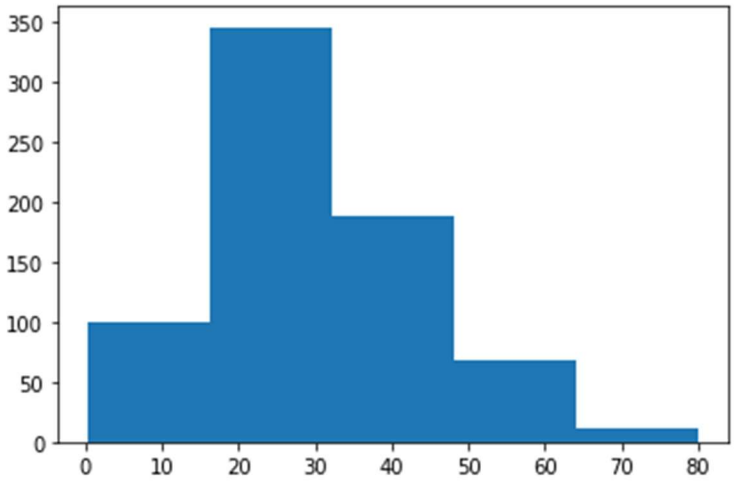
Numerical Data:

Analyzing Numerical data is important because understanding the distribution of variables helps to further process the data. Most of the time, we will find much inconsistency with numerical data so we have to explore numerical variables.

1) Histogram:

A histogram is a value distribution plot of numerical columns. It basically creates bins in various ranges in values and plots it where we can visualize how values are distributed. We can have a look where more values lie like in positive, negative, or at the center(mean). Let's have a look at the Age column.

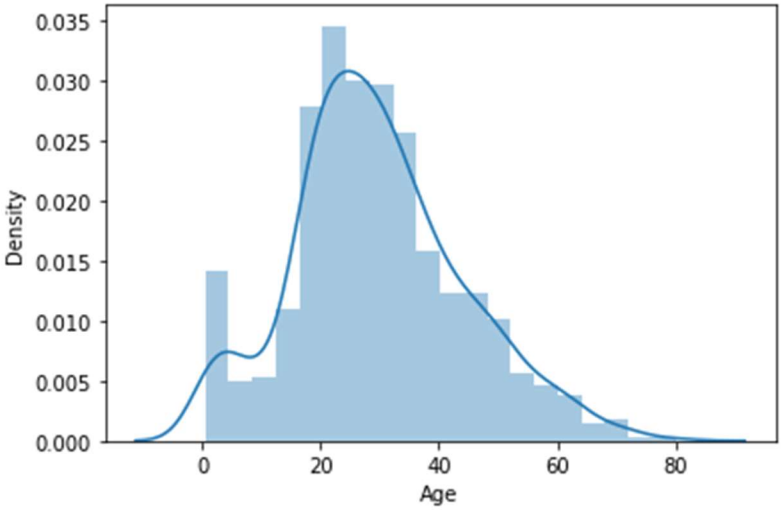
```
plt.hist(data['Age'], bins=5)
plt.show()
```



2) Distplot:

Distplot is also known as the second Histogram because it is a slight improvement version of the Histogram. Distplot gives us a KDE(Kernel Density Estimation) over histogram which explains PDF(Probability Density Function) which means what is the probability of each value occurring in this column.

```
sns.distplot(data['Age'])
plt.show()
```



3) Boxplot:

Boxplot is a very interesting plot that basically plots a 5 number summary. to get 5 number summary some terms we need to describe.

- Median – Middle value in series after sorting
- Percentile – Gives any number which is number of values present before this percentile like for example 50 under 25th percentile so it explains total of 50 values are there below 25th percentile
- Minimum and Maximum – These are not minimum and maximum values, rather they describe the lower and upper boundary of standard deviation which is calculated using Interquartile range(IQR).

```
IQR = Q3 - Q1
Lower_boundary = Q1 - 1.5 * IQR
Upper_bounday = Q3 +  1.5 * IQR
```

Here Q1 and Q3 is 1st quantile (25th percentile) and 3rd Quantile(75th percentile).

Bivariate/ Multivariate Analysis:

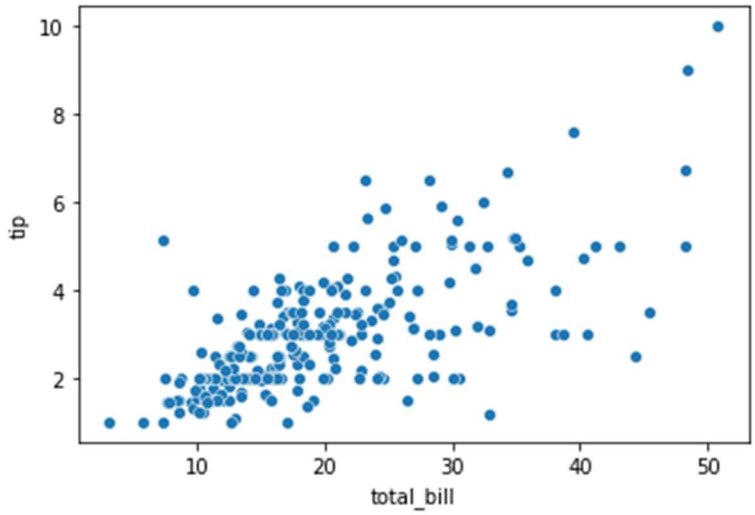
We have study about various plots to explore single categorical and numerical data. Bivariate Analysis is used when we have to explore the relationship between 2 different variables and we have to do this because, in the end, our main task is to explore the relationship between variables to build a powerful model. And when we analyze more than 2 variables together then it is known as Multivariate Analysis. we will work on different plots for Bivariate as well on Multivariate Analysis.

Explore the plots when both the variable is numerical.

1) Scatter Plot:

To plot the relationship between two numerical variables scatter plot is a simple plot to do. Let us see the relationship between the total bill and tip provided using a scatter plot.

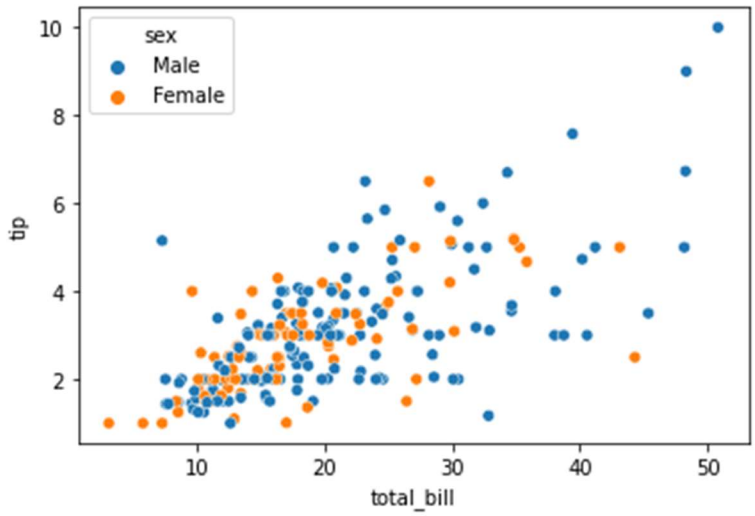
```
sns.scatterplot(tips["total_bill"], tips["tip"])
```



Multivariate analysis with scatter plot:

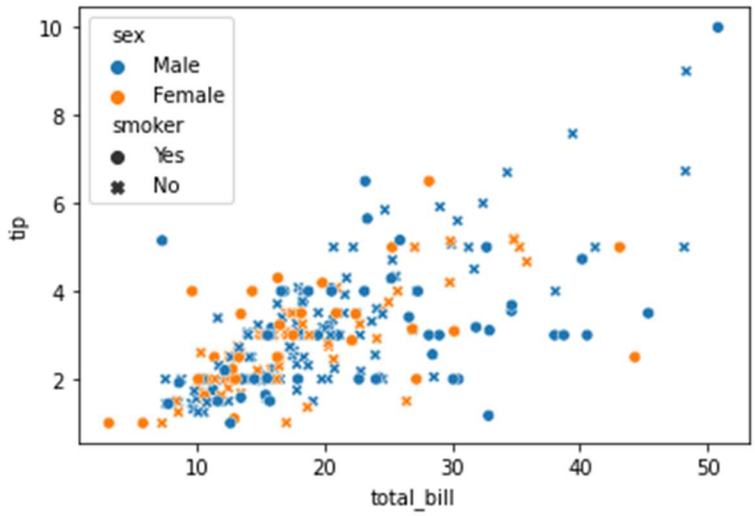
We can also plot 3 variable or 4 variable relationships with scatter plot. suppose we want to find the separate ratio of male and female with total bill and tip provided.

```
sns.scatterplot(tips["total_bill"], tips["tip"], hue=tips["sex"])
plt.show()
```



We can also see 4 variable multivariate analyses with scatter plots using style argument. Suppose along with gender we also want to know whether the customer was a smoker or not so we can do this.

```
sns.scatterplot(tips["total_bill"], tips["tip"], hue=tips["sex"],
style=tips['smoker'])
plt.show()
```



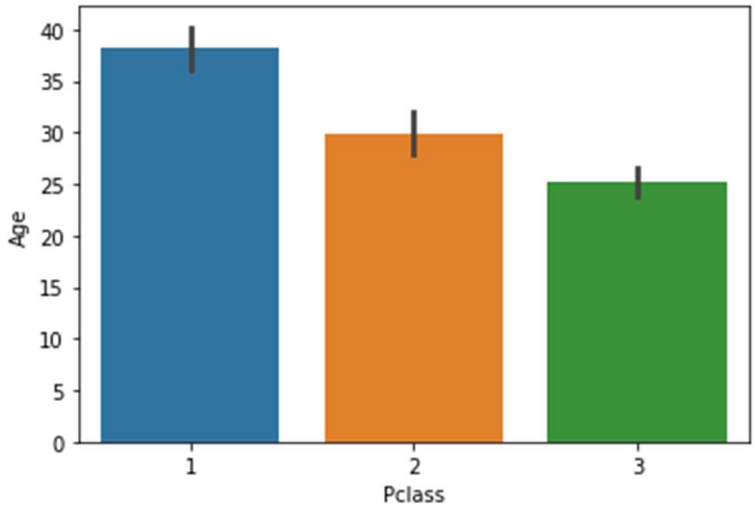
Numerical and Categorical:

If one variable is numerical and one is categorical then there are various plots that we can use for Bivariate and Multivariate analysis.

1) Bar Plot:

Bar plot is a simple plot which we can use to plot categorical variable on the x-axis and numerical variable on y-axis and explore the relationship between both variables. The blacktip on top of each bar shows the confidence Interval. let us explore P-Class with age.

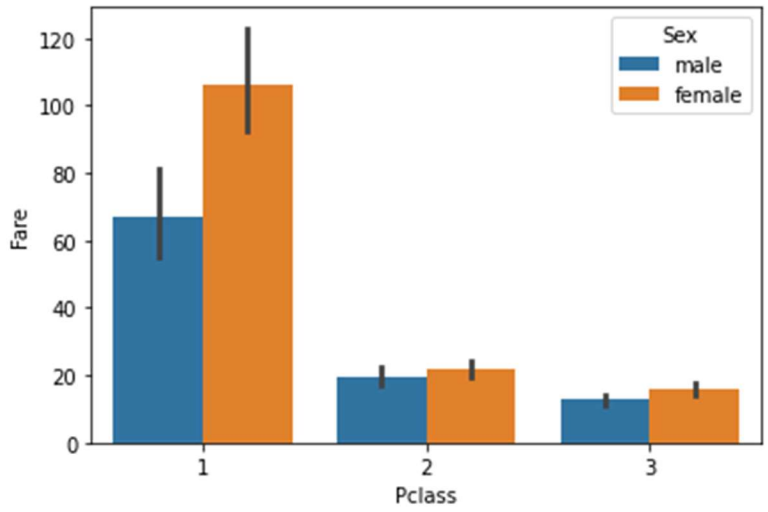
```
sns.barplot(data['Pclass'], data['Age'])
plt.show()
```



Multivariate analysis using Bar plot:

Hue's argument is very useful which helps to analyze more than 2 variables. Now along with the above relationship we want to see with gender.

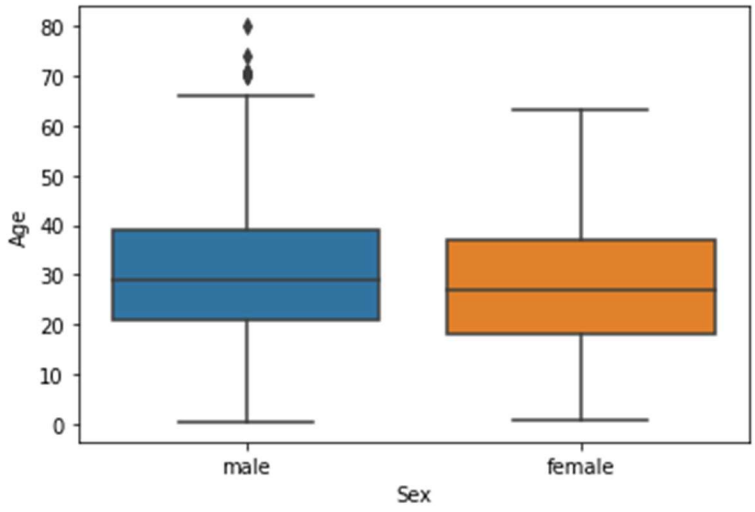
```
sns.barplot(data['Pclass'], data['Fare'], hue = data["Sex"])
plt.show()
```



2) Boxplot:

We have already study about boxplots in the Univariate analysis above. we can draw a separate boxplot for both the variable. let us explore gender with age using a boxplot.

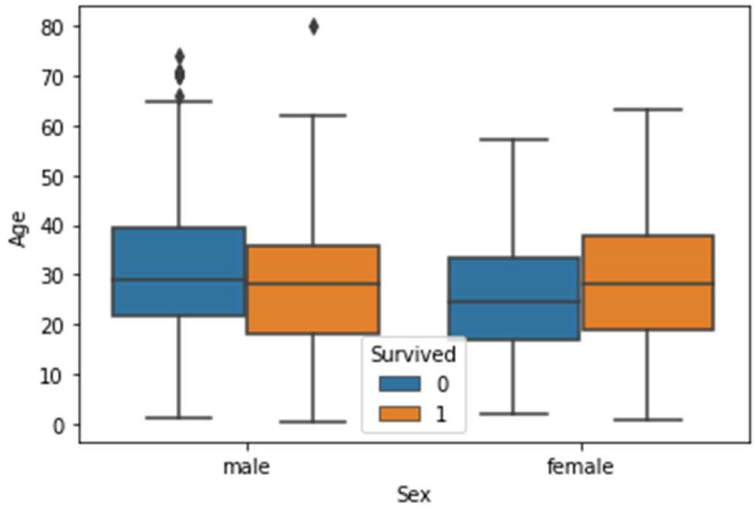
```
sns.boxplot(data['Sex'], data["Age"])
```



Multivariate analysis with boxplot:

Along with age and gender let's see who has survived and who has not.

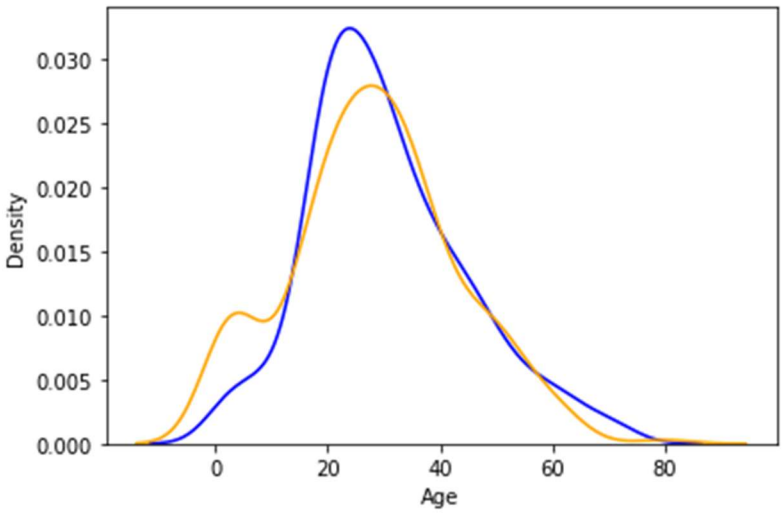
```
sns.boxplot(data['Sex'], data["Age"], data["Survived"])
plt.show()
```



3) Distplot:

Distplot explains the PDF function using kernel density estimation. Distplot does not have a hue parameter but we can create it. Suppose we want to see the probability of people with an age range that of survival probability and find out whose survival probability is high to the age range of death ratio.

```
sns.distplot(data[data['Survived'] == 0]['Age'], hist=False, color="blue")
sns.distplot(data[data['Survived'] == 1]['Age'], hist=False, color="orange")
plt.show()
```



In above graph, the blue one shows the probability of dying and the orange plot shows the survival probability. If we observe it we can see that children's survival probability is higher than death and which is the opposite in the case of aged peoples. This small analysis tells sometimes some big things about data and it helps while preparing data stories.

Categorical and Categorical:

Now, we will work on categorical and categorical columns.

1) Heatmap:

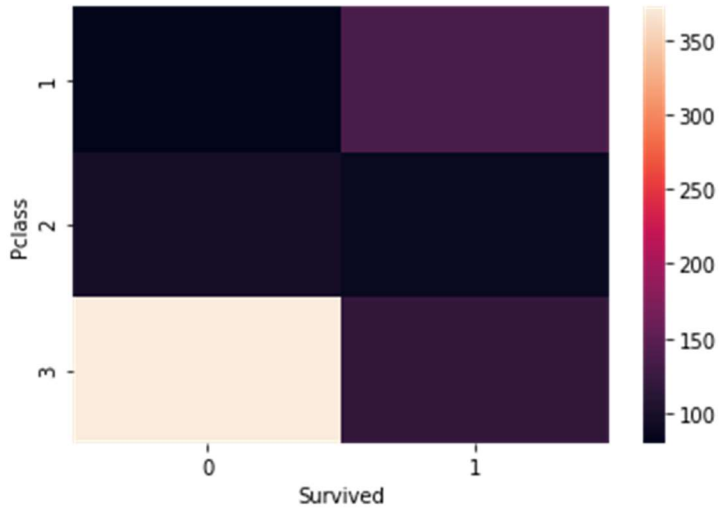
If you have ever used a crosstab function of pandas then Heatmap is a similar visual representation of that only. It basically shows that how much presence of one category concerning another category is present in the dataset. let me show first with crosstab and then with heatmap.

```
pd.crosstab(data['Pclass'], data['Survived'])
```

Survived	0	1
Pclass		
1	80	136
2	97	87
3	372	119

Now with heatmap, we have to find how many people survived and died.

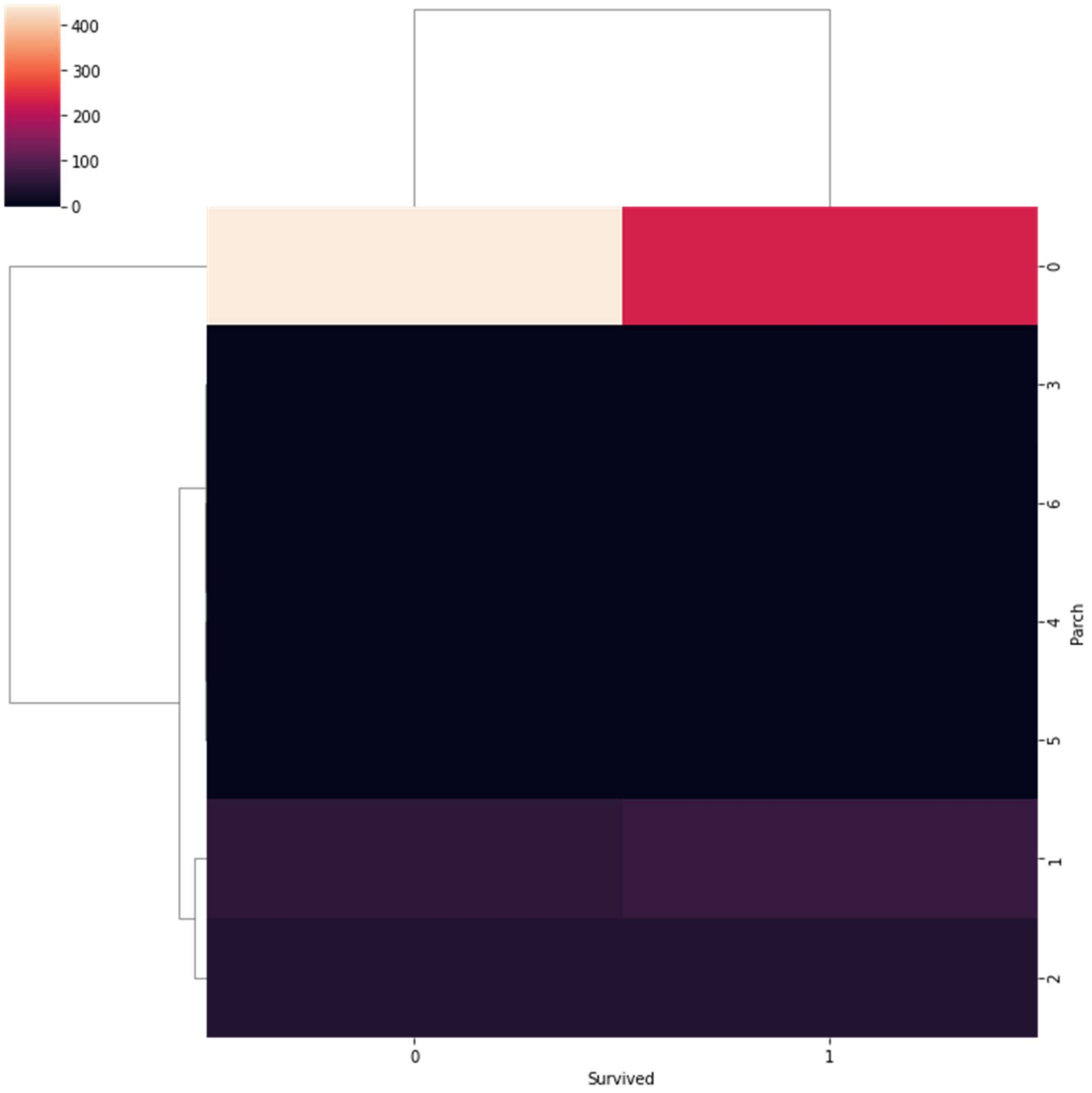
```
sns.heatmap(pd.crosstab(data['Pclass'], data['Survived']))
```



2) Cluster map:

We can also use a cluster map to understand the relationship between two categorical variables. A cluster map basically plots a dendrogram that shows the categories of similar behavior together.

```
sns.clustermap(pd.crosstab(data['Parch'], data['Survived']))
plt.show()
```


400
300
200
100
0
0
3
6
4
5
1
2
Parch
0
1
Survived