Group A
Assignment No: 8

Title of the Assignment: Data Visualization I

1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information
about the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see
if we can find any patterns in the data.
2. Write a code to check how the price of the ticket (column name: 'fare') for each passenger
is distributed by plotting a histogram.

Objective of the Assignment: Students should be able to perform the data Visualization
operation using Python on any open source dataset

Prerequisite:

1. Basic of Python Programming
2. Seaborn Library, Concept of Data Visualization.

Introduction

In the previous article, we looked at how Python's Matplotlib library can be used for data visualization. In this article we will look at Seaborn which is another extremely useful library for data visualization in Python. The Seaborn library is built on top of Matplotlib and offers many advanced data visualization capabilities.

Though, the Seaborn library can be used to draw a variety of charts such as matrix plots, grid plots, regression plots etc., in this article we will see how the Seaborn library can be used to draw distributional and categorial plots. In the second part of the series, we will see how to draw regression plots, matrix plots, and grid plots.

Downloading the Seaborn Library :

The seaborn library can be downloaded in a couple of ways. If you are using pip installer for Python libraries, you can execute the following command to download the library:

pip install seaborn

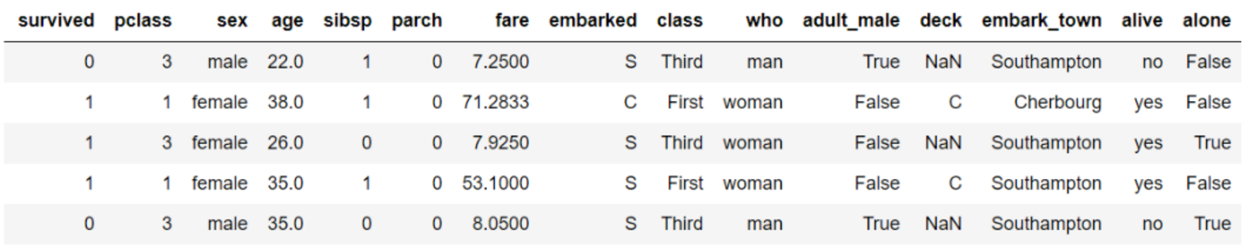
The Dataset :

The dataset that we are going to use to draw our plots will be the Titanic dataset, which is downloaded by default with the Seaborn library. All you have to do is use the load_dataset function and pass it the name of the dataset.

Let's see what the Titanic dataset looks like. Execute the following script:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

dataset = sns.load_dataset('titanic')

dataset.head()
```

The script above loads the Titanic dataset and displays the first five rows of the dataset using the head function. The output looks like this:



survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

Distributional Plots :

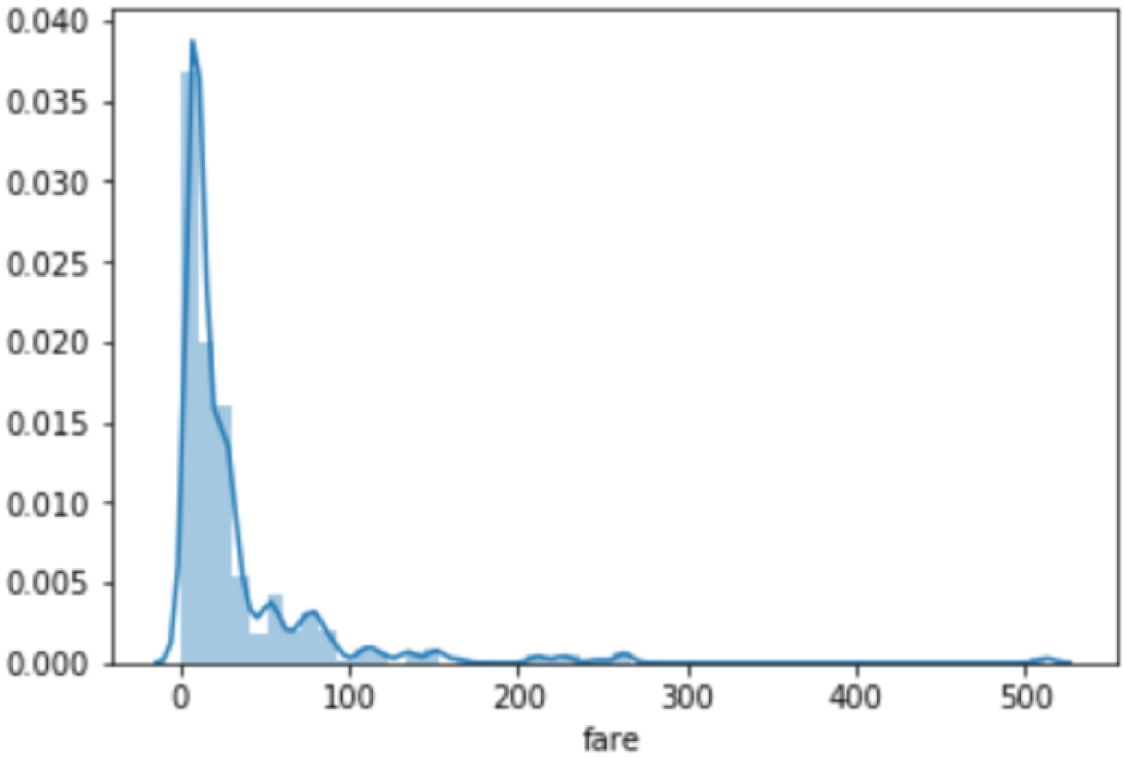
Distributional plots, as the name suggests are type of plots that show the statistical distribution of data. In this section we will see some of the most commonly used distribution plots in Seaborn.

The Dist Plot :

The distplot() shows the histogram distribution of data for a single column. The column name is passed as a parameter to the distplot() function. Let's see how the price of the ticket for each passenger is distributed. Execute the following script:

```
sns.distplot(dataset['fare'])
```

Output:



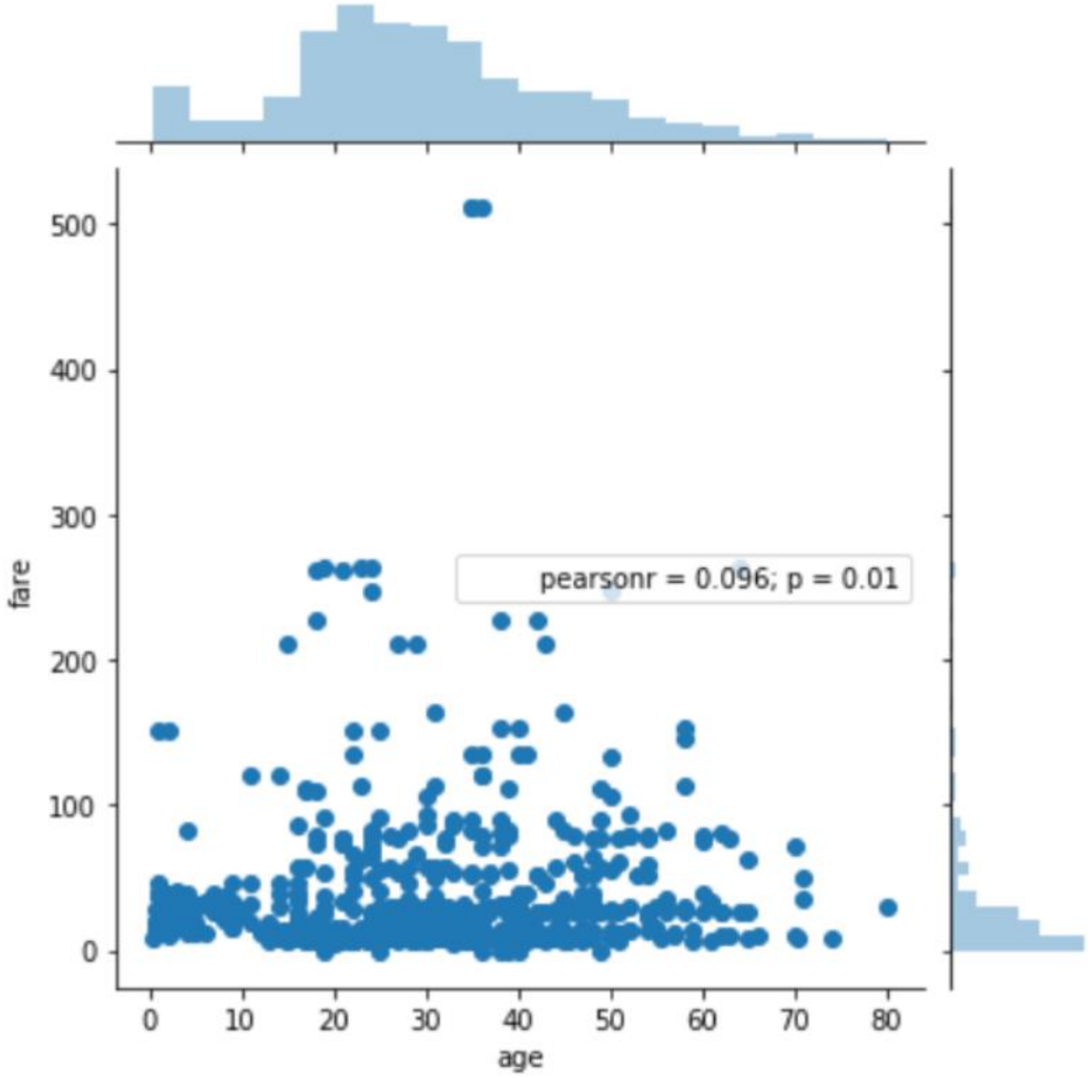
The Joint Plot

The jointplot()is used to display the mutual distribution of each column. You need to pass three parameters to jointplot. The first parameter is the column name for which you want to display the distribution of data on x-axis. The second parameter is the column name for which you want to display the distribution of data on y-axis. Finally, the third parameter is the name of the data frame.

Let's plot a joint plot of age and fare columns to see if we can find any relationship between the two.

```
sns.jointplot(x='age', y='fare', data=dataset)
```

Output:

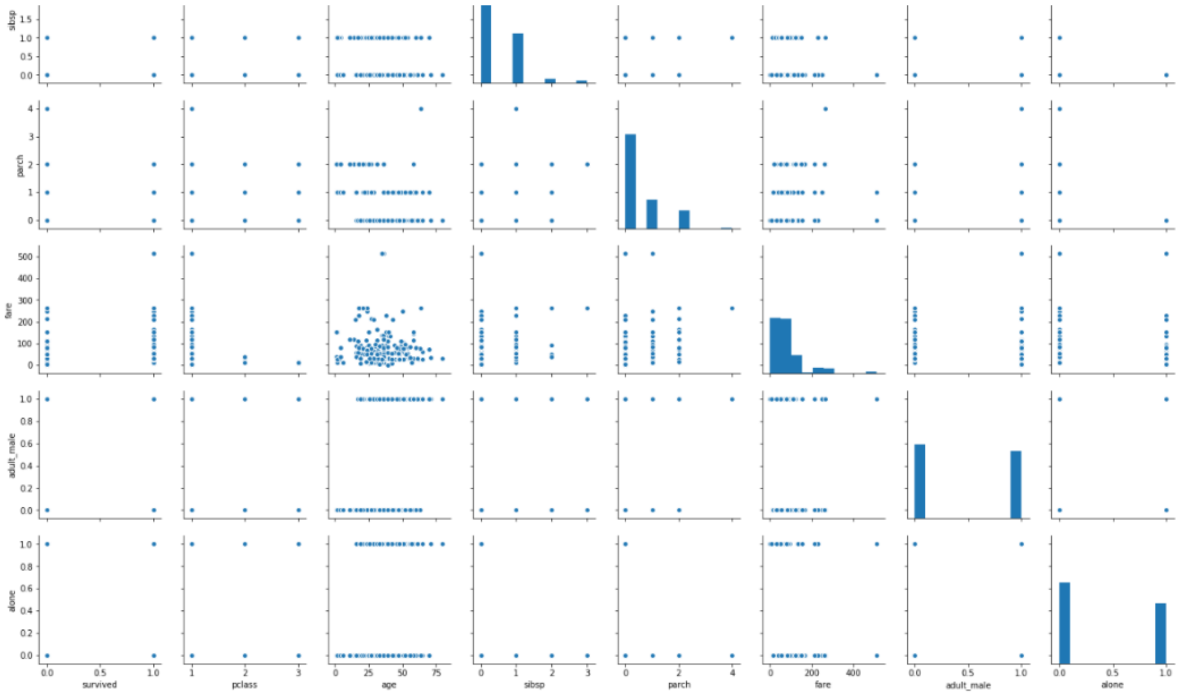


The Pair Plot

The paitplot() is a type of distribution plot that basically plots a joint plot for all the possible combination of numeric and Boolean columns in your dataset. You only need to pass the name of your dataset as the parameter to the pairplot() function as shown below:

```
sns.pairplot(dataset)
```

A snapshot of the portion of the output is shown below:



Categorical Plots

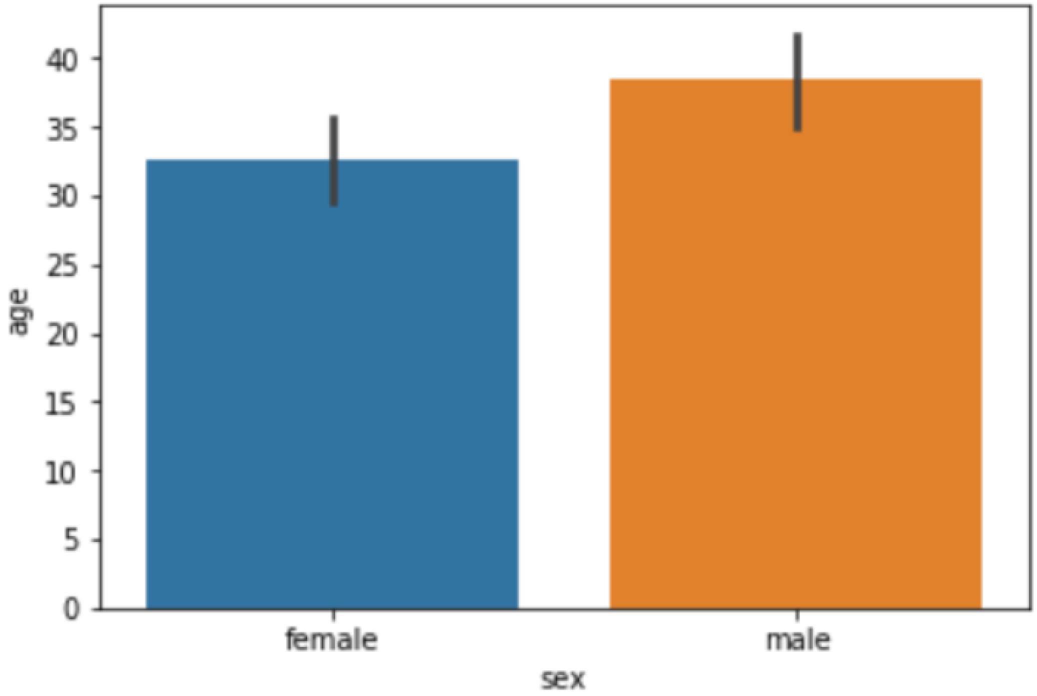
Categorical plots, as the name suggests are normally used to plot categorical data. The categorical plots plot the values in the categorical column against another categorical column or a numeric column. Let's see some of the most commonly used categorical data.

The Bar Plot

The barplot() is used to display the mean value for each value in a categorical column, against a numeric column. The first parameter is the categorical column, the second parameter is the numeric column while the third parameter is the dataset. For instance, if you want to know the mean value of the age of the male and female passengers, you can use the bar plot as follows.

```
sns.barplot(x='sex', y='age', data=dataset)
```

OUTPUT:-

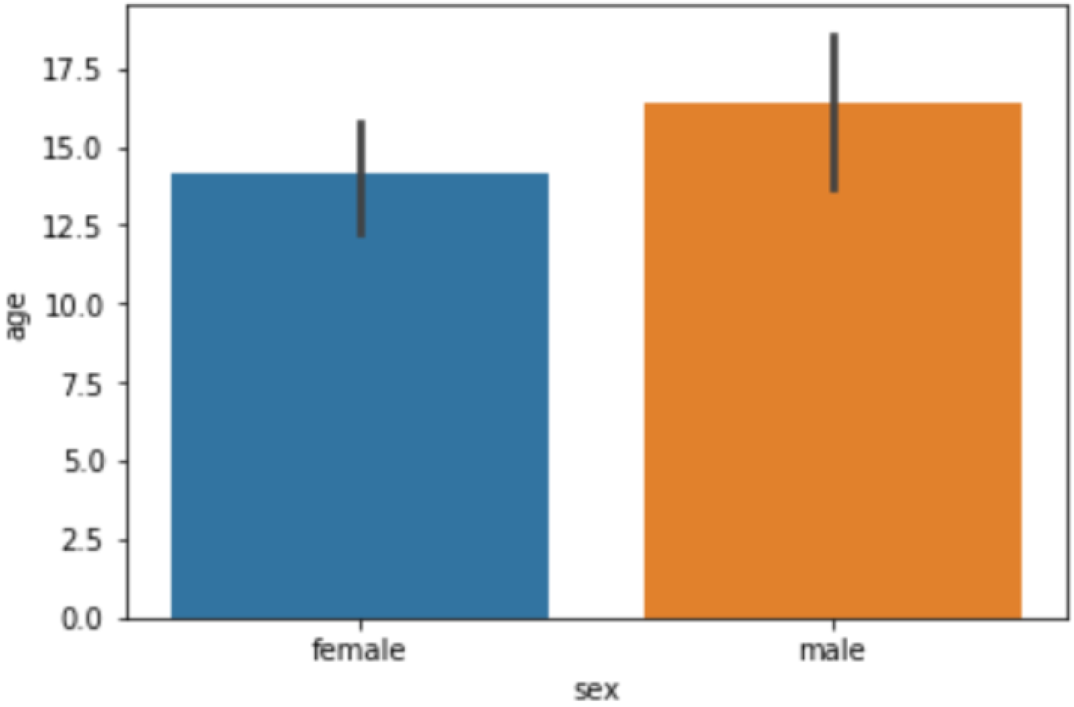


From the output, you can clearly see that the average age of male passengers is just less than 40 while the average age of female passengers is around 33.

In addition to finding the average, the bar plot can also be used to calculate other aggregate values for each category. To do so, you need to pass the aggregate function to the estimator. For instance, you can calculate the standard deviation for the age of each gender as follows:

```
import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

sns.barplot(x='sex', y='age', data=dataset, estimator=np.std)
```

output:-



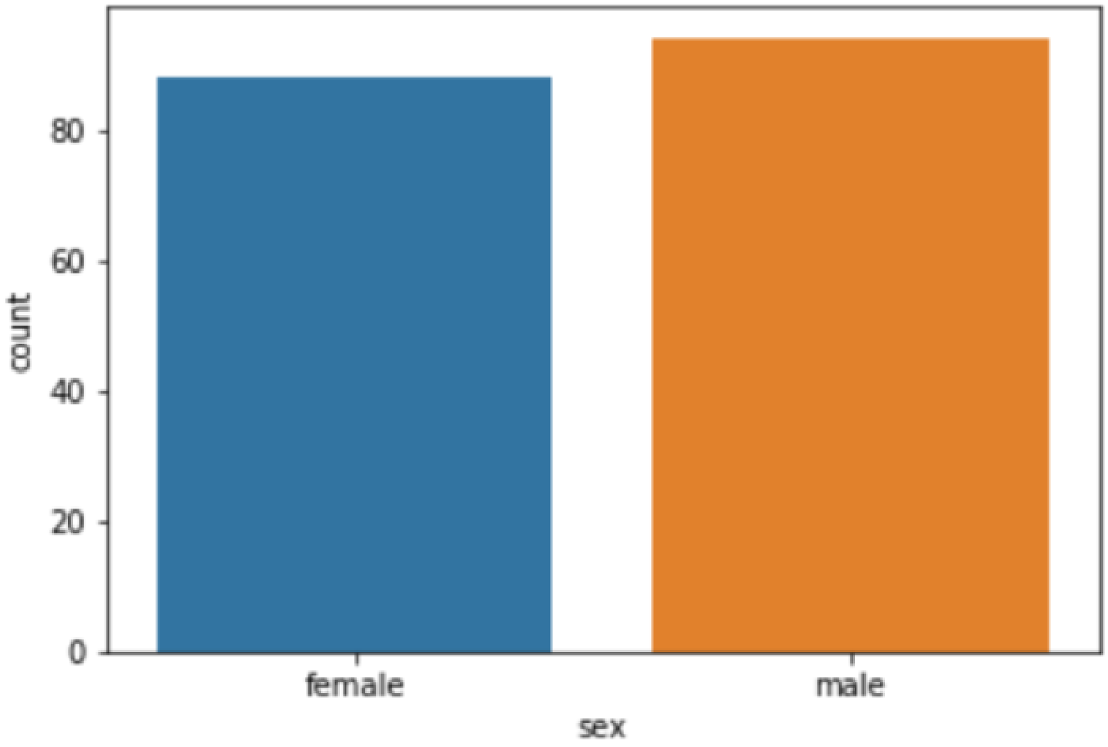
The Count Plot

The count plot is similar to the bar plot, however it displays the count of the categories in a specific column. For instance, if we want to count the number of males and women passenger we can do so using count plot as follows:

```
sns.countplot(x='sex', data=dataset)
```

The output shows the count as follows:

Output:



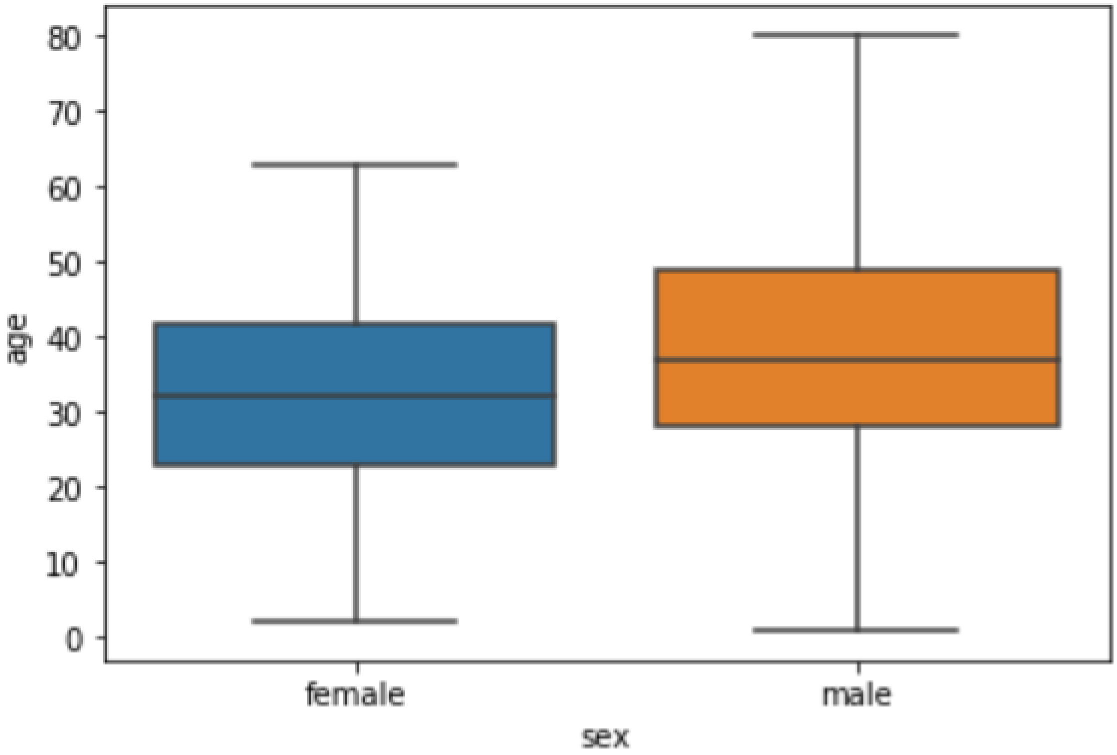
The Box Plot

The box plot is used to display the distribution of the categorical data in the form of quartiles. The center of the box shows the median value. The value from the lower whisker to the bottom of the box shows the first quartile. From the bottom of the box to the middle of the box lies the second quartile. From the middle of the box to the top of the box lies the third quartile and finally from the top of the box to the top whisker lies the last quartile.

Now let's plot a box plot that displays the distribution for the age with respect to each gender. You need to pass the categorical column as the first parameter (which is sex in our case) and the numeric column (age in our case) as the second parameter. Finally, the dataset is passed as the third parameter, take a look at the following script:

```
sns.boxplot(x='sex', y='age', data=dataset)
```

output :-



Conclusion

Seaborn is an advanced data visualization library built on top of Matplotlib library. In this article, we looked at how we can draw distributional and categorical plots using Seaborn library. This is Part 1 of the series of article on Seaborn. In the second article of the series, we will see how we play around with grid functionalities in Seaborn and how we can draw Matrix and Regression plots in Seaborn.