Experiment - 08 (Group A)

Title of the assignment - Data Visualization - I.

Problem statement - 1. Use the in built 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.

Objective of the assignment - Students should be able to perform the data visualization using python on any open source dataset.

Prerequisite: 1. Basic of Python.
2. Seaborn library, Concept of data visualization.

THEORY :

1) Seaborn :

1. 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.

2. Though the seaborn library can be used to draw a variety of charts such as matrix plots, grid plots

regression plots, etc. Seaborn library can be used to draw distributional and categorical plots.
To draw regression plots, matrix plots and grid plots, seaborn library need to download.

- Downloading Seaborn library.

pip install seaborn.

- or for anaconda — conda install seaborn.

2] The dataset:

The titanic dataset is in built.

```
import pandas as pd.
import numpy as np.
import matplotlib as plt.
import seaborn as sns.

dataset = sns.load_dataset ('titanic').
```

- Perform EDA before visualization.

3] Here are all the various different visualization than seaborn has to offer —

(a) Distributional Plots -

The distplot shows the histogram distribution of data for a single column. The column name is passed as a parameter to the distplot() function.

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

The output will be done in form of graph columns.

(b). The Joint Plot. -

The joint plot is used to display mutual distribution of each column. There are three parameters to joint plot.

- The first parameter is the column name which display the distribution of data on x-axis.
- The second parameter is the column name which display the distribution of data on y-axis.
- Finally third parameter is the name of the dataframe.

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

(C) Pair plot.

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

sns.pairplot(dataset).

(d). The rug plot :-

The rug plot is used to draw small bars along x-axis for each point in dataset. To plot a rug plot, pass the name of column plot a rug plot for fare.

sns.rugplot(dataset['fare']).

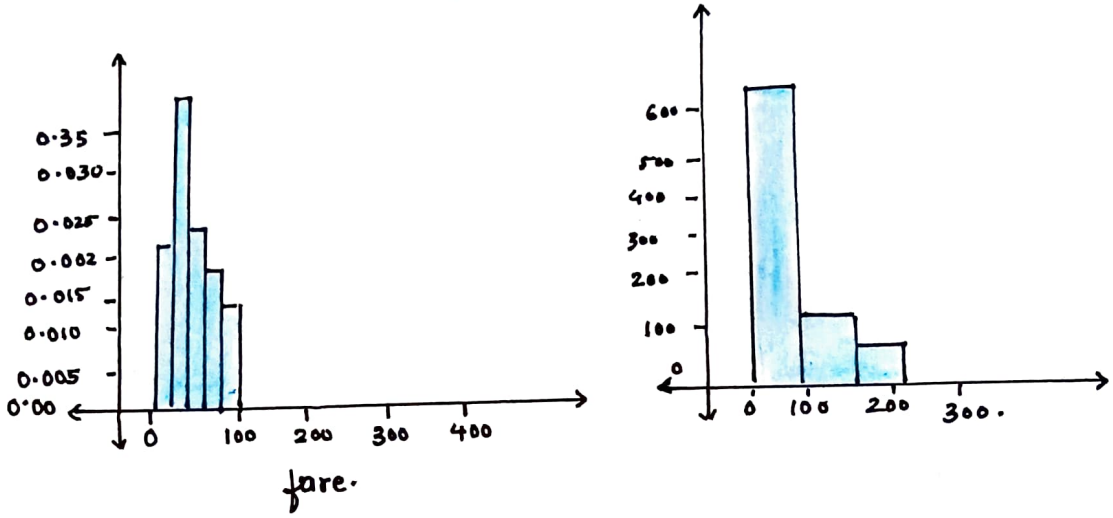
* Categorical plot:

- Categorical plot 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. Most commonly used categorical data is as follows –

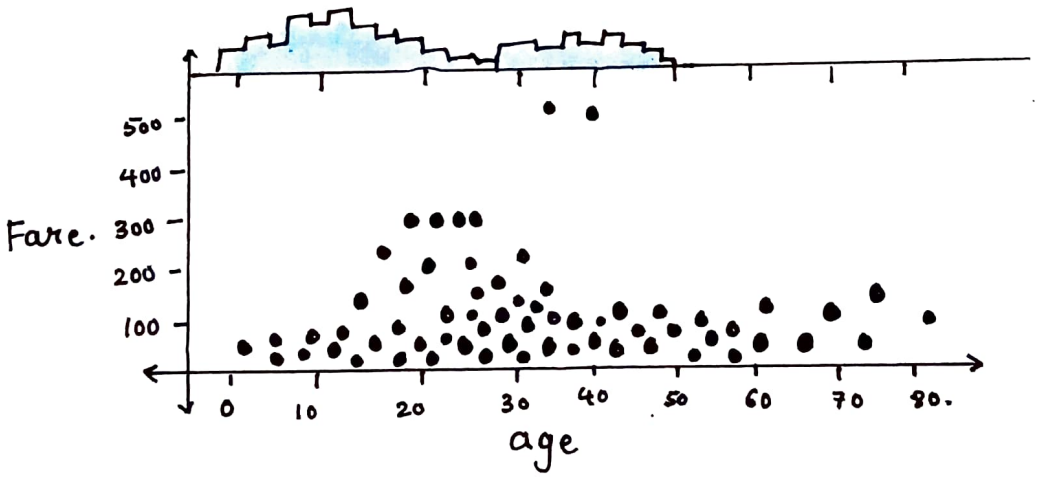
(a) The Bar plot:

- The bar plot, as the name suggests are normally used to plot categorical data.
- The bar plot is used to display the mean value for each in categorical column.
- The first parameter is the categorical column, the second parameter is the numeric column while third parameter is the dataset.

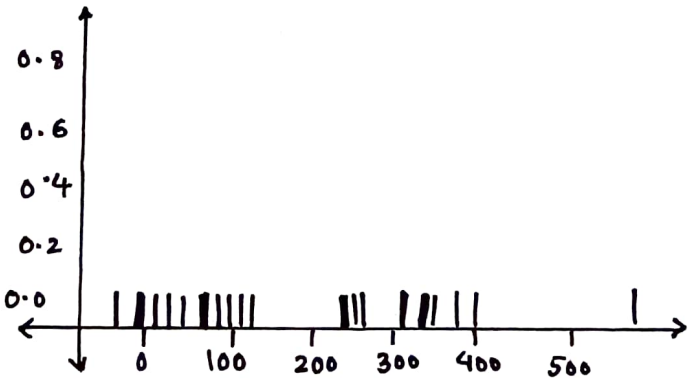
Distribution plot.



JOINT PLOT.



RUG PLOT.



To find the average bar plot can also be used to calculate other aggregate values for each category.

- Pass the aggregate function to the estimator.

To 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 ).
```

(b) Count plot:

The count plot is similar to the bar plot, It displays the count number of males and women passenger, use count plot as follows:-

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

(c) Box plot.

- The box plot is used to display the distribution of the categorical data in the form of quartiles.
- The center of box shows the median value.
- The value from 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 to the middle of the box lies the second quartile. From the middle of the box to the top of the box. lies third quartile. Then finally on top lies fourth quartile.

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

(d). The violin plot. –

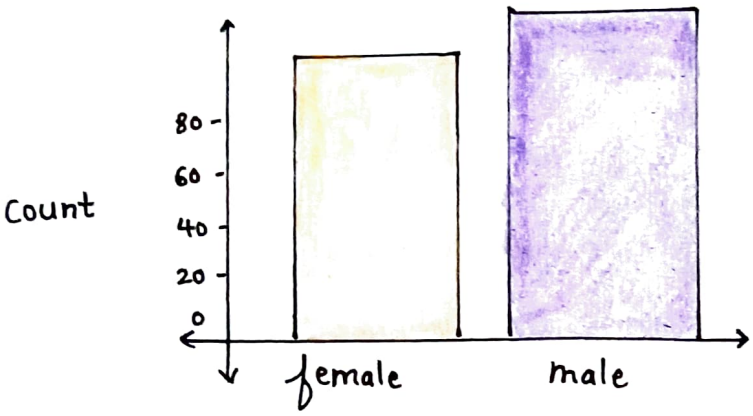
- The violin plot is similar to the box plot, however the violin plot allows us to display all the components that actually correspond to the datapoint. The violinplot() function is used to plot the violin plot.
- Like box plot, the first parameter is categorical column, the second parameter is numeric column. while third is the dataset.

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

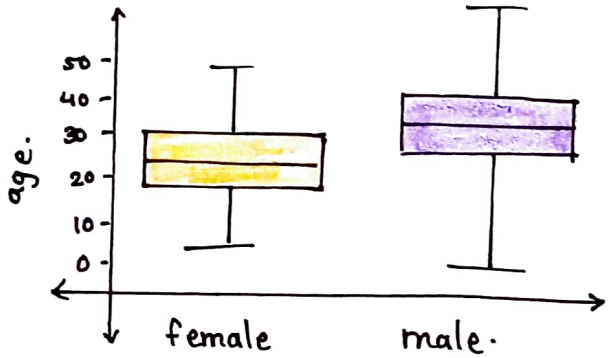
(e). The strip plot –

- The strip plot draws a scatter plot where one of the variables is categorical. We have seen scatter plots in the joint plot and the pair plot sections where we had two numeric variables.
- The strip plot is different in a way that

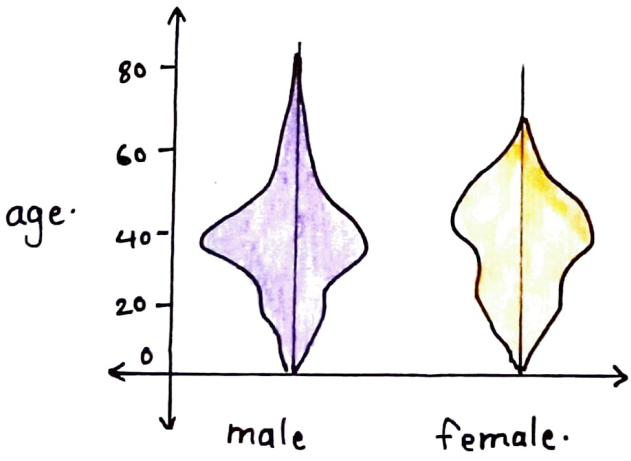
BAR PLOT.



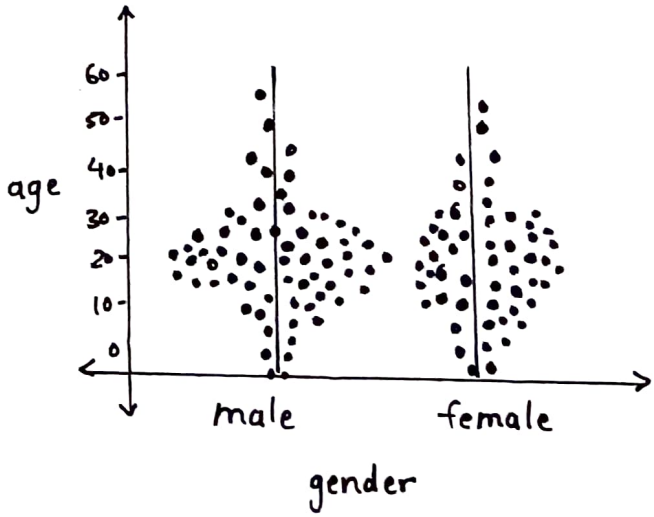
Box PLOT



VIOLIN PLOT



SWARM PLOT



one of variables is categorical in this case, and for each category in the categorical variable, you will see scatter plot with respect to the numeric column.

- The stripplot() function is used to plot the violin plot, like the box plot, the first parameter is the categorical column, the second parameter is the numeric column while the third parameter is the dataset. Look at following script:

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

(f). The swarm plot -

- The swarm plot is a combination of the strip and violin plots. In the swarm plot, the points are adjusted in such a way that they don't overlap.
- Lets plot a swarm plot for the distribution of age against gender
- The swarmplot() function is used to plot swarm plot.
- Like box plot() the first parameter is catagorical column, the second parameter is the numeric column while third parameter is the dataset.
- Ex -

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

(g). Combining swarm plots and violin plots -

- Swarm plots are not recommended if you have a huge dataset since they do not scale well because they have to plot each data point.
- If you really like swarm plots, a better way is to combine two plots, For instance, to combine a violin plot with swarm plot, you need to execute the following script.

```
sns.violinplot(x='sex', y='age', data='dataset')
sns.swarmplot(x='sex', y="age", data='ds', color='black').
```

* **Conclusion:** Seaborn is an advanced data visualization library built on top of matplotlib library.
In this assignment we have explored distributional data and categorical plots using seaborn library.

* * * * * * *