**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 informationabout the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see if we can find any patterns in the data.

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 opensource dataset

**Prerequisite:**

1. Basic of Python Programming
2. Concept of Data Visualization, Distributional Plotting Histogram

**Contents for Theory:**

1. Introduction of Seaborn Library
2. Downloading the Seaborn Library
3. Loading Dataset
4. Plotting Histogram Using the Dist Plot

1. **Introduction of Seaborn Library:**

The Seaborn library is built on top of Matplotlib and offers many advanced data visualization capabilities.

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.

1. **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 |  |

Alternatively, if you are using the Anaconda distribution of Python, you can use execute the following command to download the seaborn library:

|  |  |
| --- | --- |
| conda install seaborn |  |

1. **Loading 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:



The dataset contains 891 rows and 15 columns and contains information about the passengers who boarded the unfortunate Titanic ship. The original task is to predict whether or not the passenger survived depending upon different features such as their age, ticket, cabin they boarded, the class of the ticket, etc. We will use the Seaborn library to see if we can find any patterns in the data.

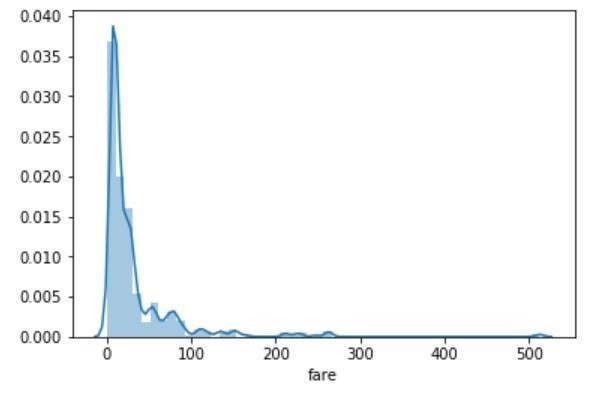
**4. Ploting Histogram Using the Dist Plot**

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']) |  |

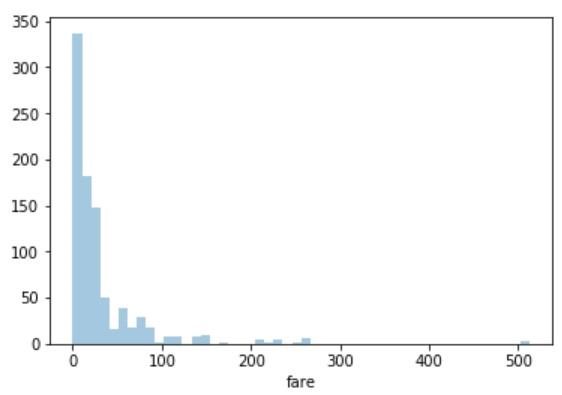
**Output:**



You can see that most of the tickets have been solved between 0-50 dollars. The line that you see represents the kernel density estimation. You can remove this line by passing False as the parameter for the kde attribute as shown below:

|  |  |
| --- | --- |
| sns.distplot(dataset['fare'], kde=False) |  |

**Output:**



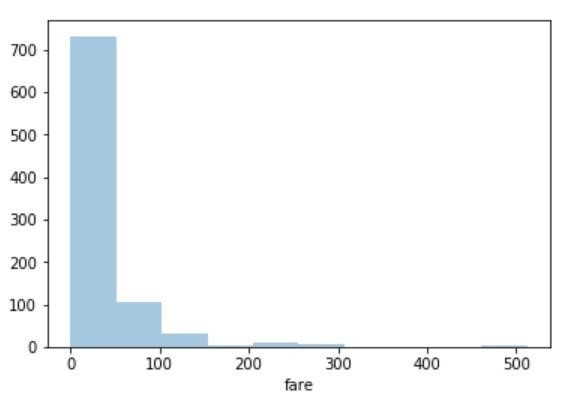
Now you can see there is no line for the kernel density estimation on the plot.

You can also pass the value for the bins parameter in order to see more or less details in the graph. Take a look at the following script:

|  |  |
| --- | --- |
| sns.distplot(dataset['fare'], kde=False, bins=10) |  |

Here we set the number of bins to 10. In the output, you will see data distributed in 10 bins as shown below:

**Output:**



You can clearly see that for more than 700 passengers, the ticket price is between 0 and

50.

**Source Code:**

|  |  |  |
| --- | --- | --- |
|  | import pandas as pandas  import numpy as numpy  import matplotlib.pyplot as pyplot  import seaborn as sns  dataset = sns.load\_dataset('titanic')  dataset.head()  sns.distplot(dataset['fare'],kde=False,bins=10)  pyplot.title("Fare")  pyplot.show() |  |
|  |

**Conclusion:**

Seaborn is an advanced data visualization library built on top of Matplotlib library. In this Assignment, we looked at how we can draw distributional plots using Seaborn library.

**Questions:**

1. **What is Data visualization?**
2. **What is Histogram?**
3. **What is the difference between a bar plot and a histogram?**