Page No.
Date

Name:- Rohini Janardan Devkar
Roll no:- 23272
PRN no:- 72030818G
Class :- TE2
DSBDA pr-8.

Practical No.8.
Data Visualization-I.

Aim:- 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 to price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.

Theory:-

* **Data Visualization:-**

Data Visualization ~~of data~~ is a field in data analysis that deals with visual representation of data. It graphically plots data and is an effective way to communicate inferences from data.

With pictures, maps and graphs, the human mind has an easier time processing and understanding any given data.

Python offers several plotting libraries, namely Matplotlib, Seaborn and many other such data visualization packages with different features for creating informative, customized and appealing plots to present data in the most simple and effective way.

Page No.
Date

* Python libraries :-

Seaborn :-

When you read the official documentation on Seaborn, it is defined as the data visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics. Putting it simply, seaborn is an extension of Matplotlib with advanced features.

Matplotlib :-

This is undoubtedly my favourite and a quintessential python library. You can create stories with the data visualized with Matplotlib. Another library from the Scipy stack, Matplotlib plots 2D figures.

* Benefits of Data Visualization :-

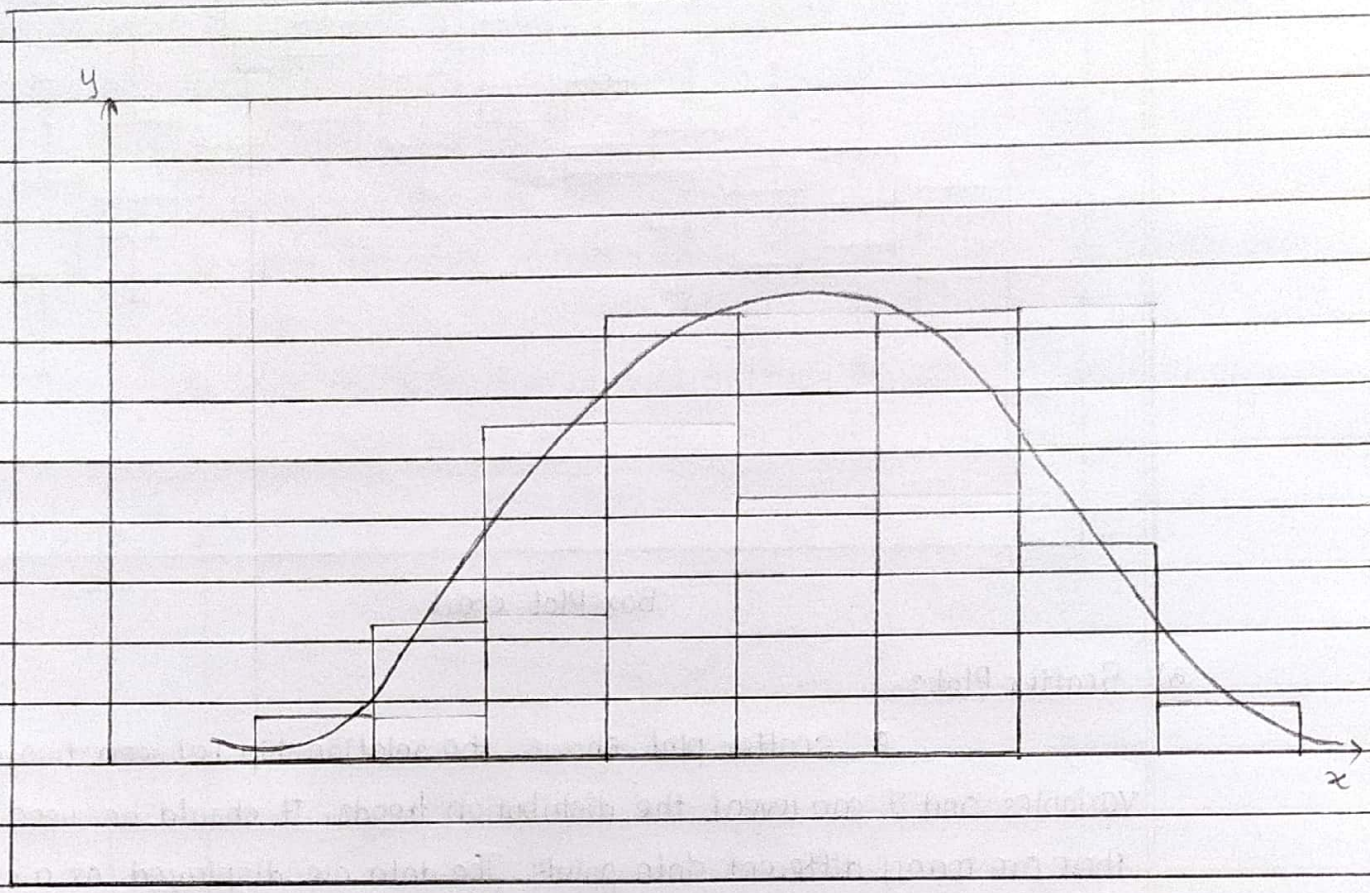
1) It promotes improved absorption of business information.
2) With the help of data visualization, decision-makers can easily understand how the data is being interpreted to determine business variations.
3) A large amount of data is handled and is visualized to establish patterns in the data. Many meaningful insights and the evidence behind the data can be used to establish a business goal.
4) Visualizing the data helps managers to achieve growth and use the ~~the~~ new pattern trends found in business strategies.

* Types of Graphs :-

1) Histogram :-

The histogram is a representation of the numerical data, not accurate but an estimate. The histogram represents the frequency of occurrence of specific

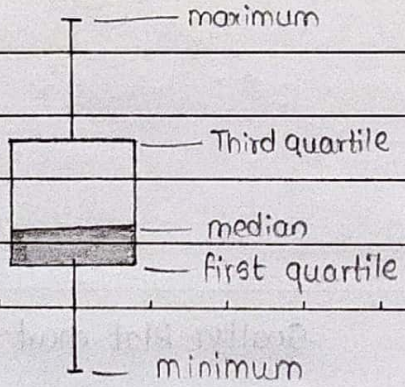
phenomena which lie within a specific range of values and arranged in consecutive and fixed intervals. A histogram graph is a popular graphing tool that provides a visual representation of data distribution.

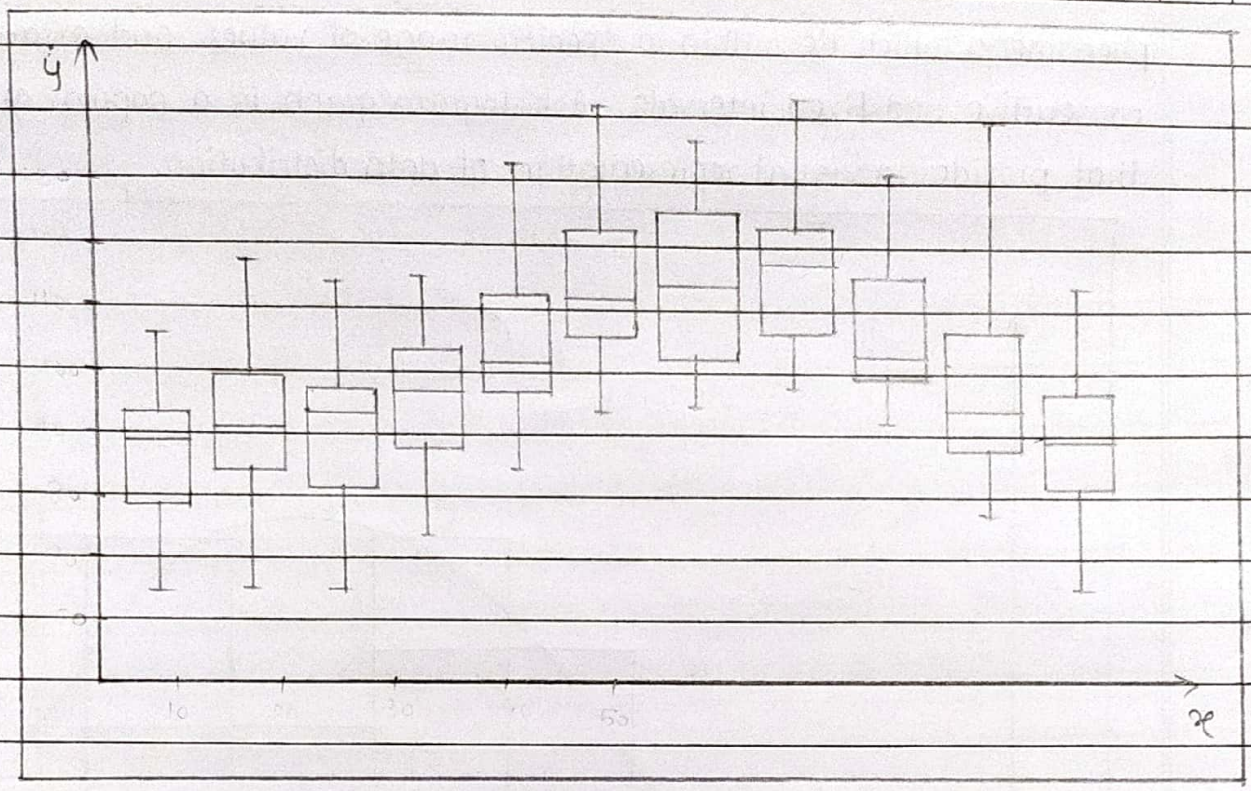


Histogram

2) Box Plot Chart :-

A box plot chart is a graphical representation of statistical data based of the minimum, first quartile, median, third quartile and maximum. The term "box plot" comes from the fact that the graph looks like a rectangle with lines extending from the top and bottom. Because of the extending lines, this type of graph is sometimes called a box-and-whisker plot.

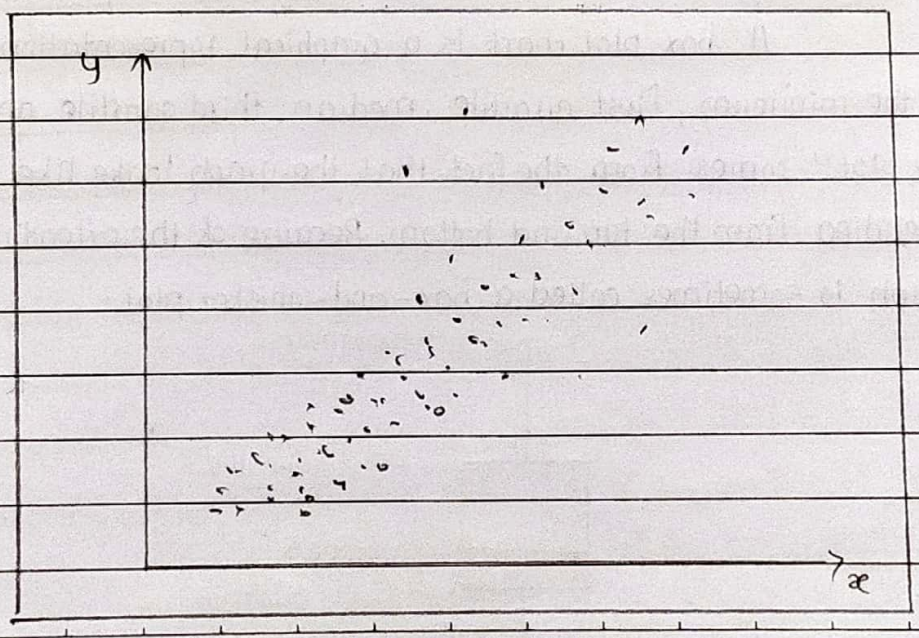




Box Plot chart

3) Scatter Plot:-

A scatter plot shows the relationship between two different variables and it can reveal the distribution trends. It should be used when there are many different data points. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of other variable determining the position on the vertical axis.



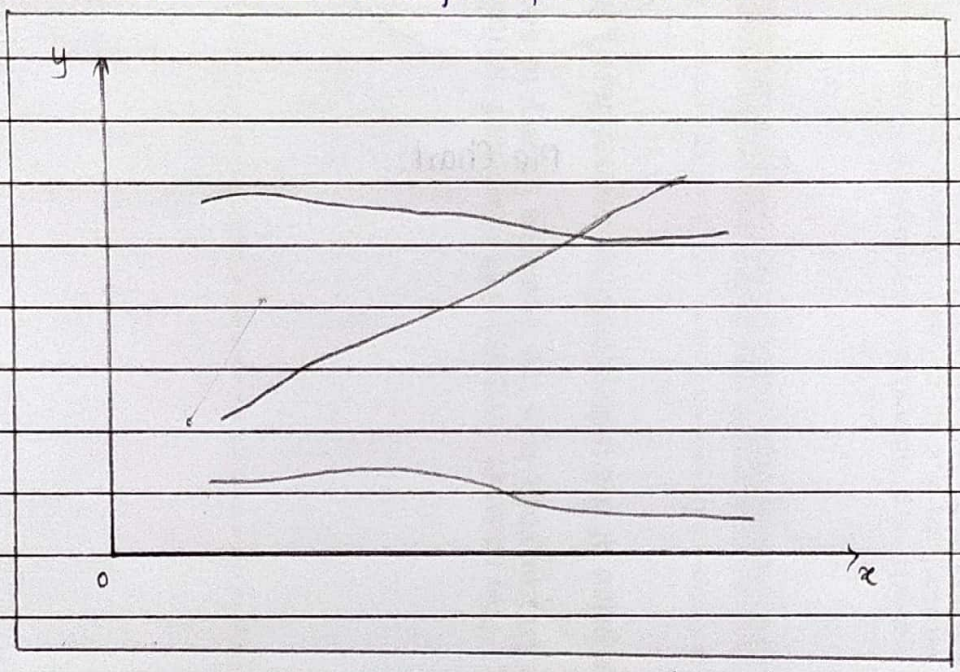
Scatter Plot chart

4) Line graph :-

A line chart graphically displays data that changes continuously over time. Each line graph consists of points that connect data to show a trend. Line graphs have x-axis and a y-axis. In the most cases, time is distributed on the horizontal axis.

- Uses of line graphs :-

1) When you want to show trends
2) When you want to make predictions based on a data history over time.
3) When comparing two or more different variables, situations, and information over a given period of time.



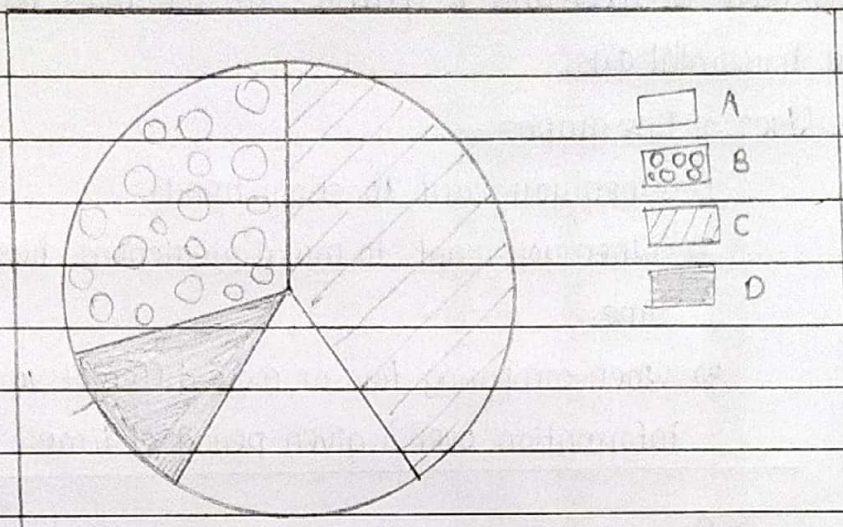
Line charts track several variables at once.

Line graph

5) Pie Charts :-

When it comes to statistical types of graphs and charts, the pie chart has a crucial place and meaning. It displays data and statistics in an easy-to-understand 'pie-slice' format and illustrates numerical ~~portion~~ proportion. The larger a slice is the bigger portion of the total quantity it represents.

When you want to create ~~represent~~ and represent the composition of something, it uses a pie chart. To show percentage or proportional data.



Pie Chart.

Data Science And Big Data Analytics Practical - 8

==========

Name:- Rohini Devkar

Roll no:- 23272

Prn no:- 72030818G

Class :- TE-2 (COMPUTER)

==========

Problem Statement:-

Data Visualization I

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.

Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.

==========

In [1]:
```
pip install seaborn
```

```
Requirement already satisfied: seaborn in c:\users\lenovo\anaconda3\lib\site-packages (0.11.2)
Requirement already satisfied: matplotlib>=2.2 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (3.4.3)
Requirement already satisfied: pandas>=0.23 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (1.3.4)
Requirement already satisfied: scipy>=1.0 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (1.7.1)
Requirement already satisfied: numpy>=1.15 in c:\users\lenovo\anaconda3\lib\site-packages (from seaborn) (1.20.3)
Requirement already satisfied: cycler>=0.10 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->seabor
n) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->se
aborn) (1.3.1)
```

```
Requirement already satisfied: python-dateutil>=2.7 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2-
>seaborn) (2.8.2)
Requirement already satisfied: pillow>=6.2.0 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->seabor
n) (8.4.0)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\lenovo\anaconda3\lib\site-packages (from matplotlib>=2.2->sea
born) (3.0.4)
Requirement already satisfied: six in c:\users\lenovo\anaconda3\lib\site-packages (from cycler>=0.10->matplotlib>=2.2->se
aborn) (1.16.0)
Requirement already satisfied: pytz>=2017.3 in c:\users\lenovo\anaconda3\lib\site-packages (from pandas>=0.23->seaborn)
(2021.3)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```
conda install seaborn
```

```
Collecting package metadata (current_repodata.json): ...working... done
Solving environment: ...working... done

## Package Plan ##

  environment location: C:\Users\Lenovo\anaconda3

  added / updated specs:
    - seaborn


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    conda-4.11.0               |   py39haa95532_0        14.4 MB
    ------------------------------------------------------------
                                           Total:        14.4 MB

The following packages will be UPDATED:

  conda                              4.10.3-py39haa95532_0 --> 4.11.0-py39haa95532_0



Downloading and Extracting Packages

conda-4.11.0         | 14.4 MB   |            |   0%
conda-4.11.0         | 14.4 MB   |            |   0%
conda-4.11.0         | 14.4 MB   | 3          |   4%
```

```
conda-4.11.0          | 14.4 MB   | 8           |   9%
conda-4.11.0          | 14.4 MB   | #4          |  14%
conda-4.11.0          | 14.4 MB   | ##          |  21%
conda-4.11.0          | 14.4 MB   | ##5         |  26%
conda-4.11.0          | 14.4 MB   | ###1        |  31%
conda-4.11.0          | 14.4 MB   | ###6        |  36%
conda-4.11.0          | 14.4 MB   | ####1       |  42%
conda-4.11.0          | 14.4 MB   | ####7       |  48%
conda-4.11.0          | 14.4 MB   | #####3      |  54%
conda-4.11.0          | 14.4 MB   | ######1     |  61%
conda-4.11.0          | 14.4 MB   | ######8     |  68%
conda-4.11.0          | 14.4 MB   | #######4    |  75%
conda-4.11.0          | 14.4 MB   | ########1   |  81%
conda-4.11.0          | 14.4 MB   | ########9   |  89%
conda-4.11.0          | 14.4 MB   | #########5  |  96%
conda-4.11.0          | 14.4 MB   | ########## | 100%
Preparing transaction: ...working... done
Verifying transaction: ...working... done
Executing transaction: ...working... done

Note: you may need to restart the kernel to use updated packages.
```

In [3]:

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

dataset = sns.load_dataset('titanic')

dataset.head()
```

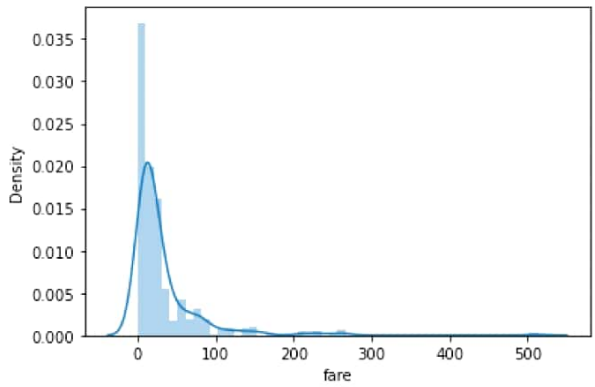
Out[3]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

In [4]:
```
sns.distplot(dataset['fare'])
```

```
C:\Users\Lenovo\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated func
tion and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function wit
h similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[4]:
```
<AxesSubplot:xlabel='fare', ylabel='Density'>
```



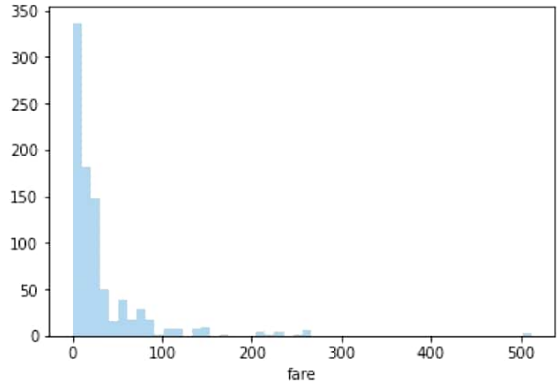
You can see that most of the tickets have been solved between 0-50 dollars.

The line that you see represents the kernel density estimation.

In [5]:
```
sns.distplot(dataset['fare'], kde=False)
```

```
C:\Users\Lenovo\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot` is a deprecated func
tion and will be removed in a future version. Please adapt your code to use either `displot` (a figure-level function wit
h similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

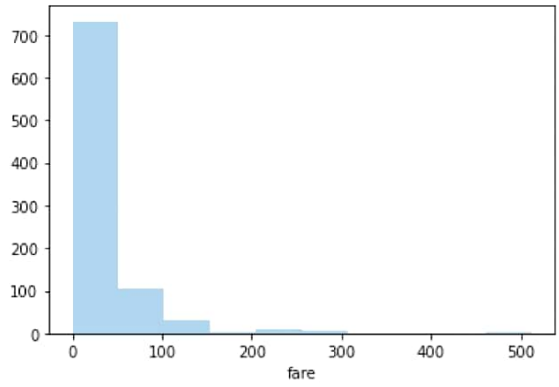
Out[5]:
```
<AxesSubplot:xlabel='fare'>
```



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

In [6]:
```
sns.distplot(dataset['fare'], kde=False, bins=10)
```

Out[6]:
```
<AxesSubplot:xlabel='fare'>
```



You can clearly see that for more than 700 passengers, the ticket price is between 0 and 50.

Page No.
Date

* Conclusion :-

Seaborn is an advanced data visualization library built on top of matplotlib library. In this practical, we looked at how we can draw histogram, distributional and categorical plots using Seaborn library. We implemented the the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.