DSBDAL

Assignment No: 08

* Aim:- Data visualization II

1) use the inbuilt dataset titanic ! The dataset contains 891 rows and contains information about the passangers 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 for each passanger is distributed by plotting a histogram.

* Theory:-

seaborn is a library mostly used for statistical plotting in python. It is built on top of matplotlib and provides beautiful default styles & color palettes to make statistical plots more attractive.

*Different categories of plot in seaborn.

1) Relational plots:-

This plot is used to understand the relation between two variables.

2) categorical plots -

This plot deals With categorical variables and how they can be visualized.

3) Distribution plots:-
This plot is used for examining univariate and bivariate distribution.

4) Regression plots:-
The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyzes.

5) matrix plots:-
A matrix plot is an array of scatter plots.

6) multi-plot -grids:-
It is useful approach is to draw multiple instances of the same plot an different subsets of the datasets.

* The datasets consists of 891 rows & 12 columns.
1) passenger Id
2) survived
3) pclass
4) Name
5) sex
6) age

7) sibsp
8) parch
9) Ticket
10) fare
11) cabin
12) Embarked

The seaborn library is buit on top of matplotlib and offers many advanced data visualization capabilities.

* features:-
The titanic dataset has roughly the following types of features:

- categorical / Nominal:
variables that can be divided into multiple categories but having no order or priority.
E.g. Embarked (C = cherbourg; Q = queen stown; S = southampton)

- Binary:
A subtype of categorical features, where the variable has two categories.
E.g. Sex (male / female)

- ordinal:-
They are similar to categorical feature but they have an order (ie. can be started).
E.g. pclass (1,2,3).

- continous:-
They can take up any value beth the minimum & maximum values in a column.
E.g. Age, fare

- count:-
They represent the count of a variable
E.g. - sibsp, parch.

- useless:-
They don't contribute to the final outcome of an ML model.

- distribution plots:-
distribution plots, as the name suggests are type of plot that shows the statical distribution of data.
The distplot() shows the histogram distribution of data for a single column.
We can see that most of the tickets have been solved between the 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)

sns.distplot(dataset['fare'], kde = false, bins =10)
```

Here we set the no. of bins to 10. In the o/p, you will see data distributed in 10 bins as shown below.

* Histrogram:-
  - Histrograms are visualization tools that represents the distribution of set of continuous data.
  - In a histogram, the data is divided into a set of intervals or bin & the count of data points that fall into each bin corresponding to the height of the bar above that bin.

Syntax:-

```
seaborn.histoplot(data x, y, hue,
```

stat, bins, bin width, discrete, kde, log-scale).

* parameters :-

- data = input data in the form of pandaforms or Numpy array.
- x, y (optional) = key of the data to be positioned on the x & y axes respectively.
- hue (optional) = semantic data key which is mapped to determine the color.
- stat (optional) = count, frequency, etc.
- Return :- This method returns the matplotlib axes with plot drawn on it.