

Assignment No - 08 (Group A)

Problem statement:

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.

Pre-requisite

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

Objective

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

Software and Hardware requirements:-

1. **Operating system:** Linux- Ubuntu 16.04 to 17.10, or Windows 7 to 10,
2. **RAM-** 2GB RAM (4GB preferable)
3. **IDE :-** Anaconda Jupiter Notebook / pycharm / Visual Studio

Theory-

Data Visualisation plays a very important role in Data mining. Various data scientists spent their time exploring data through visualisation. To accelerate this process we need to have a well-documentation of all the plots. Even plenty of resources can't be transformed into valuable goods without planning and architecture

1. Seaborn Library Basics

Seaborn is a Python data visualisation library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

For the installation of Seaborn, you may run any of the following in your command line.

pip install seaborn

conda install seaborn

To import seaborn you can run the following command.

```
import seaborn as sns
```

2. Know your data

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()
```

3. Finding patterns of data.

Patterns of data can be find out with the help of different types of plots

Types of plots are:

- A. Distribution Plots
 - a. Dist-Plot
 - b. Joint Plot
 - d. Rug Plot

A. Distribution Plots

These plots help us to visualise the distribution of data. We can use these plots to understand the mean, median, range, variance, deviation, etc of the data.

a. Distplot

Dist plot gives us the histogram of the selected continuous variable.

It is an example of a univariate analysis.

We can change the number of bins i.e. number of vertical bars in a histogram

```
import seaborn as sns

sns.distplot(x = dataset['age'], bins = 10)
```

b. Joint Plot

It is the combination of the distplot of two variables.

It is an example of bivariate analysis.

We additionally obtain a scatter plot between the variables to reflect their linear relationship. We can customise the scatter plot into a hexagonal plot, where, the more the colour intensity, the more will be the number of observations.

```
import seaborn as sns
sns.jointplot(x = dataset['age'], y = dataset['fare'], kind = 'scatter')
```

c. The Rug Plot

b. The rugplot() is used to draw small bars along the x-axis for each point in the dataset. To plot a rug plot, you need to pass the name of the column. Let's plot a rug plot for fare.

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

B. Categorical Plots

- a. Bar Plot
- b. Count Plot
- c. Box Plot
- d. Violin Plot

b. 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)
```

b. 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)
```

c The Box Plot

The box plot is used to display the distribution of the categorical data in the form of quartiles. The centre 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)
```

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 data point. The `violinplot()` 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.

Let's plot a violin plot that displays the distribution for the age with respect to each gender.

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

Conclusion:

Seaborn is an advanced data visualisation library built on top of Matplotlib library.

In this assignment, we looked at how we can draw distributional and categorical plots using the Seaborn library.

We have seen how to plot matrix plots in Seaborn. We also saw how to change plot styles and use grid functions to manipulate subplots.