Scatter and KDE Plots

In this lesson, scatter plots and KDE plots are discussed.

We'll cover the following

- Scatter plot
- KDE (kernel density estimation) plots

Scatter plot

A scatter plot represents the values of data as points in a *cartesian plane*. It displays the data points based on the cartesian coordinates on an XY-plane. It is preferably used when a pair of dependent and independent variables need to be represented or visualized.

The same example from the regression plot lesson will be used to display a scatter plot.

```
import numpy as np
import seaborn as sns

df = sns.load_dataset('tips')

sns1 = sns.scatterplot(x = 'total_bill', y = 'tip', data = df)
```

Just like in the *regression* lesson, the `total_bill` is an independent variable, and `tip` is the dependent variable. The `total_bill` is on the x-axis, and the `tip` is along the y-axis. A circle is placed on the graph where the values of these two planes coincide.

For more functionalities and information on scatter plots, refer here.

KDE (kernel density estimation) plots

This function calculates and plots the probability density of a given dataset, which means that it estimates the value of a random variable. Let’s understand this concept from an example.

```
import numpy as np
import seaborn as sns

x = np.random.randn(100) # Generating random data

sns1 = sns.kdeplot(x, color = 'red') # ploting KDE plot
```

The above function plots the probability density of the dataset declared at **line 4**. The probability density is different from probability. It represents the probability of a single point in a range of values as the area under the curve. As can be seen in the output, the resultant plot is in the shape of a curve. The y-axis of a **KDE** plot represents the height of the curve, and sometimes it might be larger than *one*. However, by multiplying it with the width of the area under the curve the final probability value will be between [0,1] (inclusive). To put it simply, this plot informs us of where the majority of the population of data lies.

In the above example, `x` is a random variable with 100 random values. The `kdeplot` function estimates where how many points of data lie. The x-axis represents the range of values from the random variable, and the y-axis represents the *probability density* values for the corresponding range value.

According to the above output, the *probability density* value for **0** is the highest, meaning most of the values are close to **0**. Detailed information on *KDE plots* can be obtained here.

Next, some challenges await to test your newly acquired visualization skills.