Univariate Analysis for Numerical Variables

1. What is Univariate Analysis?

Univariate analysis is the process of analyzing a single variable. It helps understand the distribution, central tendency, and spread of data. It is mainly used for:

- Identifying the shape of data distribution
- Detecting outliers
- Understanding data spread

2. Plots for Numerical Data

2.1 Histogram (Distribution Plot)

A histogram is used to visualize the frequency of numerical values. It shows how values are distributed and whether the data is skewed. The KDE (Kernel Density Estimation) curve provides a smooth estimate of the distribution.

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.histplot(df["age"], bins=20, kde=True)
plt.title("Age Distribution")
plt.show()
```

2.2 Box Plot (Outlier Detection)

A box plot helps in detecting outliers and understanding data spread. The box represents the interquartile range (IQR), while dots outside the whiskers indicate outliers.

```
sns.boxplot(y=df["age"])
plt.title("Boxplot of Age")
plt.show()
```

2.3 Violin Plot (Density & Distribution)

A violin plot is a combination of a box plot and a KDE plot, showing both distribution and outliers.

```
sns.violinplot(y=df["age"])
plt.title("Violin Plot of Age")
plt.show()
```

2.4 KDE Plot (Smooth Density Estimate)

A KDE plot provides a smooth density estimation, helping to identify peaks and spread.

```
sns.kdeplot(df["age"], shade=True)
plt.title("KDE Plot of Age")
plt.show()
```

2.5 Strip Plot (Raw Value Distribution)

A strip plot shows individual data points and helps observe clustering.

```
sns.stripplot(y=df["age"], jitter=True)
plt.title("Strip Plot of Age")
plt.show()
```

3. Conclusion

Different plots help in understanding numerical variables in a dataset. Histograms and KDE plots show distributions, box and violin plots help detect outliers, and strip plots provide insights into individual data points. Choose the right plot based on the analysis you need.