

EDA-Skewness and Kurtosis

Skewness and Kurtosis: Understanding the Shape of Your Data

In Exploratory Data Analysis (EDA), two important measures are used to describe the shape of your data distribution: **Skewness** and **Kurtosis**. These metrics help you understand whether your data is symmetric, heavy-tailed, or has other characteristics.

1. Skewness: Measuring Asymmetry

- **Definition:** Skewness measures the asymmetry of a distribution. It indicates how much the data points are deviated from the normal distribution.
- **Types:**
 - **Positive Skewness:** Data is skewed to the right, with more extreme values on the higher side.
 - **Negative Skewness (or Left Skewness):** Data is skewed to the left, with more extreme values on the lower side.
 - **Zero Skewness:** Data is symmetric around the mean.

Example:

Suppose we have a dataset of house prices in a city. The distribution of prices might be:

Price	Frequency
200K	10
300K	20
400K	30
500K	40

In this example, the prices are skewed to the right (positive skewness), as there are more extreme values on the higher side (e.g., \$500K).

2. Kurtosis: Measuring Tails

- **Definition:** Kurtosis measures the heaviness or "tailed-ness" of a distribution.
- **Types:**
 - **Leptokurtic:** Data has heavy tails, indicating more extreme values than expected in a normal distribution.
 - **Platykurtic (or Light-Tailed):** Data has relatively light tails, with fewer extreme values.

Example:

Suppose we have a dataset of incomes. The distribution might be leptokurtic, as there are more people earning high salaries and fewer people earning very low or moderate salaries.

To calculate Skewness and Kurtosis in Python, you can use the following libraries:

```
import pandas as pd
from scipy import stats

# Load your dataset into a Pandas DataFrame
df = pd.read_csv('your_data.csv')

# Calculate Skewness
skewness = df['column_name'].skew()
print(f'Skewness: {skewness:.2f}')

# Calculate Kurtosis
kurtosis = df['column_name'].kurtosis()
print(f'Kurtosis: {kurtosis:.2f}')
```

By understanding and visualizing Skewness and Kurtosis, you can better interpret the characteristics of your data distribution. This can help you:

- Identify outliers or unusual patterns
- Choose appropriate statistical tests for your analysis
- Communicate insights effectively to stakeholders