Descriptive Statistic

Course:

INFO-6145 Data Science and Machine Learning

Developed by:
Mohammad Noorchenarboo

September 12, 2024

Contents

Current Section

Descriptive Statistics Overview

Descriptive statistics are used to summarize and describe the key characteristics of a dataset. The main metrics include:

- Mean
- Median
- Mode
- Range
- Standard Deviation
- Variance
- Interquartile Range (IQR)
- Skewness
- Kurtosis

Each metric provides valuable insights into the data's central tendency, spread, and distribution characteristics.

Mean

Mean: The average value of a dataset, calculated by summing all values and dividing by the number of observations.

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example

For the data set: 10, 20, 30, 40, 50

$$\text{Mean} = \frac{10+20+30+40+50}{5} = 30$$

Median

Median: The middle value in a sorted dataset. If the dataset has an odd number of observations, the median is the middle value. If the dataset has an even number, it is the average of the two middle values.

$$\text{Median} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even} \end{cases}$$

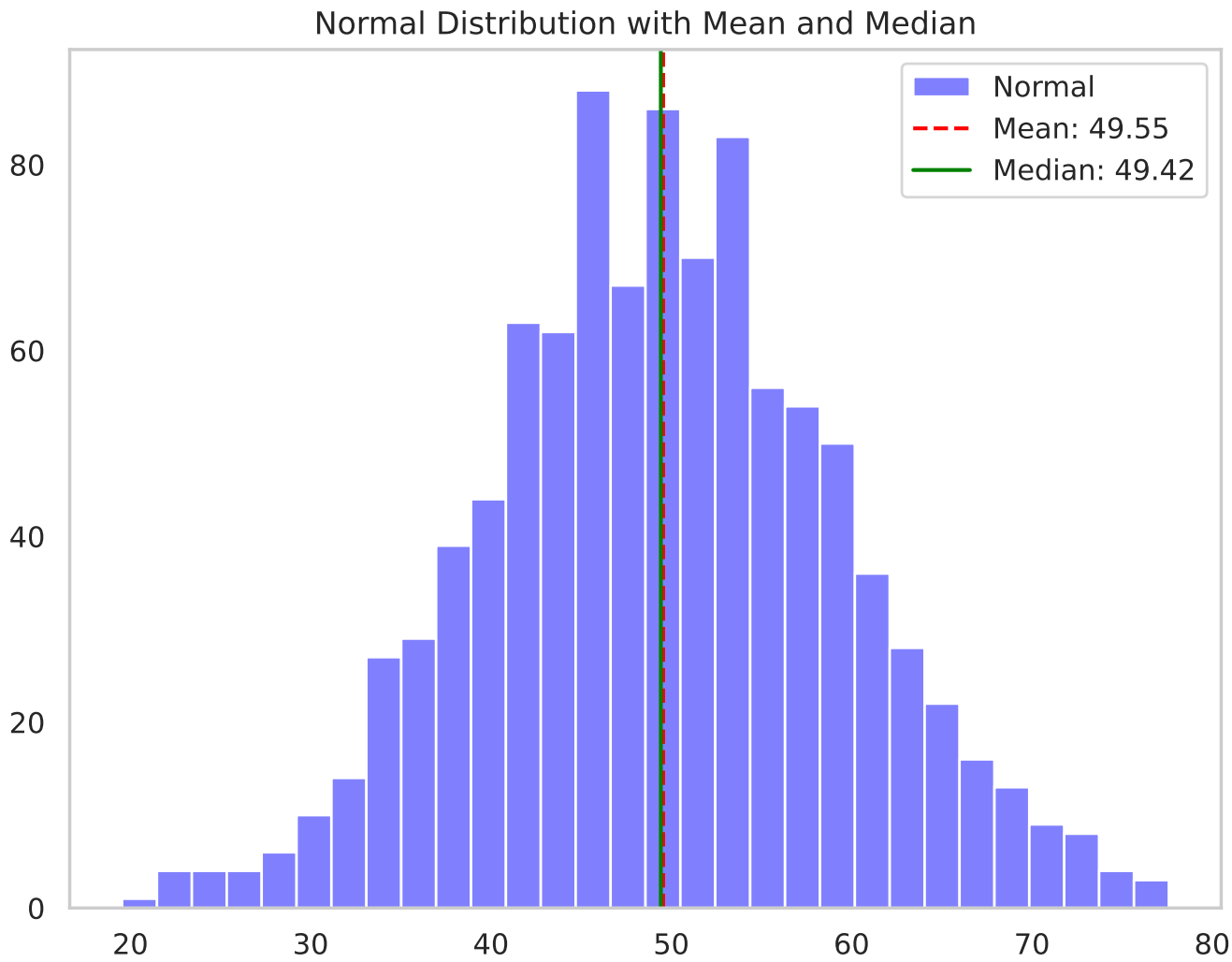
Example

For the dataset: 10, 20, 30, 40, 50

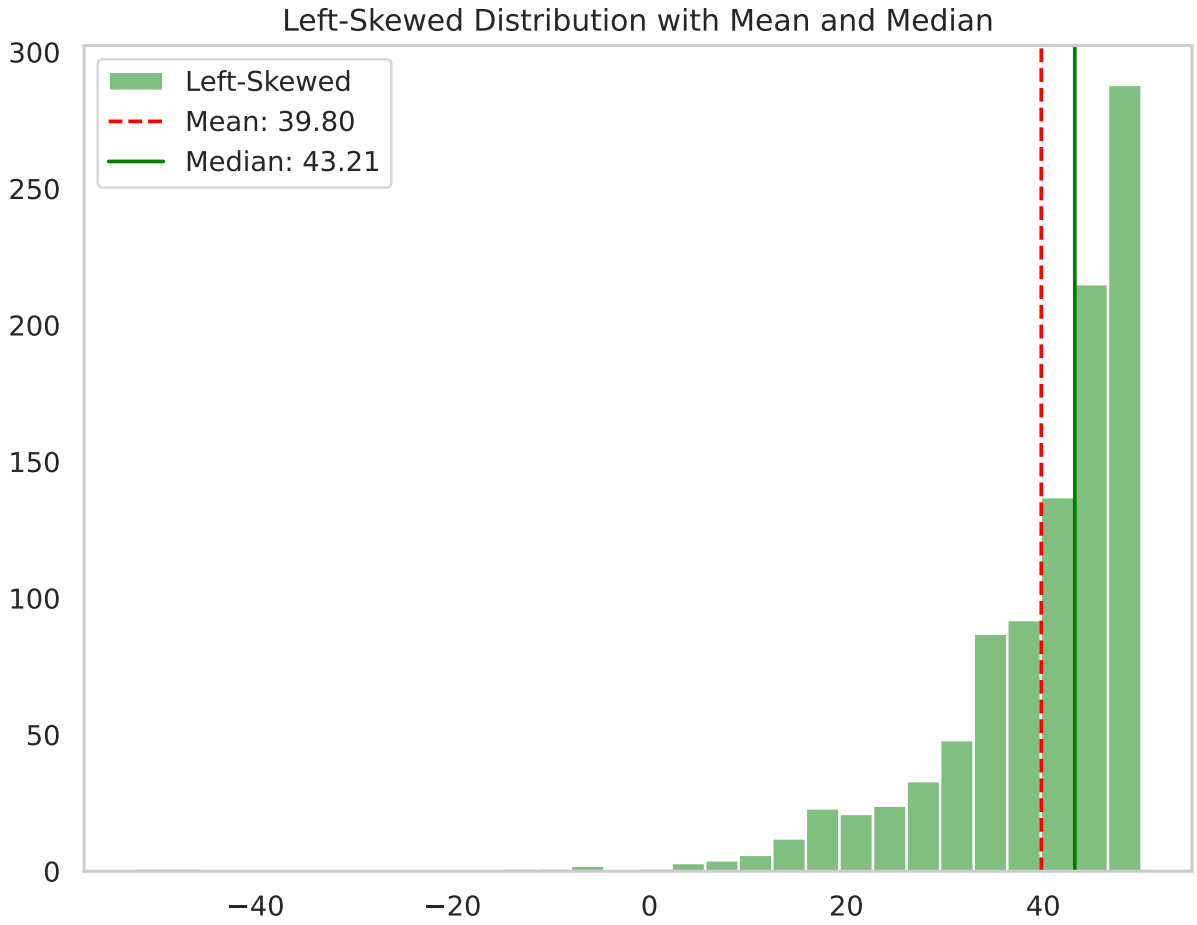
$$\text{Median} = 30$$

Normal Distribution

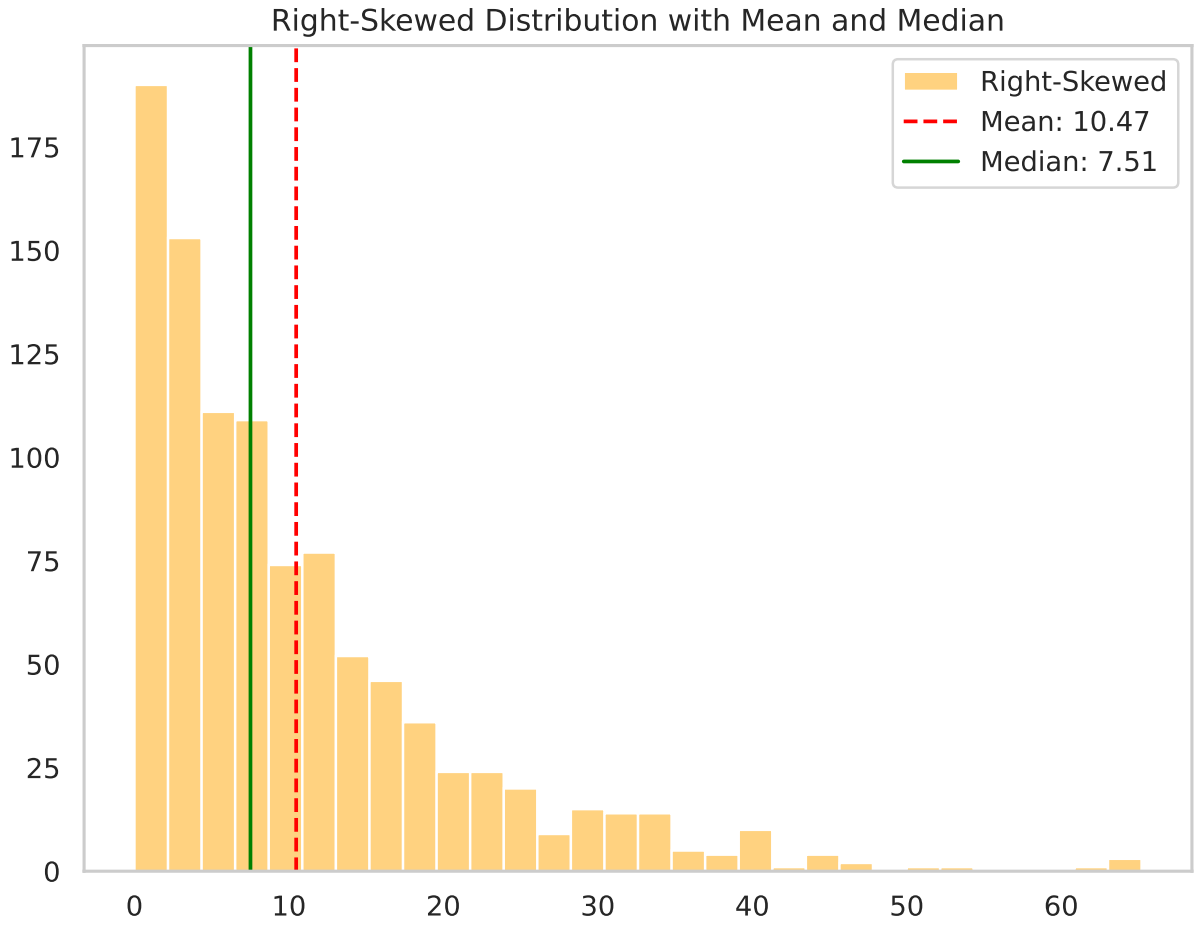
Normal Distribution



Left skewed distribution



Right skewed distribution



Mode

Mode: The value that appears most frequently in a dataset.

$$\text{Mode} = \text{Most Frequent Value}$$

Example

For the dataset: 5, 5, 10, 15, 20, 5, 20

$$\text{Mode} = 5$$

Range

Range: The difference between the maximum and minimum values in a dataset.

$$\text{Range} = \text{Max} - \text{Min}$$

Example

For the dataset: 10, 20, 30, 40, 50

$$\text{Range} = 50 - 10 = 40$$

Standard Deviation

Standard Deviation: Measures the dispersion of data points around the mean. It shows how spread out the values are.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Example

For the dataset: 10, 20, 30, 40

$$\sigma = \sqrt{\frac{(10-30)^2 + (20-30)^2 + (30-30)^2 + (40-30)^2}{4}} = 11.18$$

Variance

Variance: Measures how far each data point in the set is from the mean, calculated as the average of the squared differences from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Example

For the dataset: 10, 20, 30, 40, 50

$$\sigma^2 = \frac{(10-30)^2 + (20-30)^2 + (30-30)^2 + (40-30)^2 + (50-30)^2}{5} = 200$$

Interquartile Range (IQR)

IQR: The range between the first quartile (Q1) and third quartile (Q3), which represents the middle 50

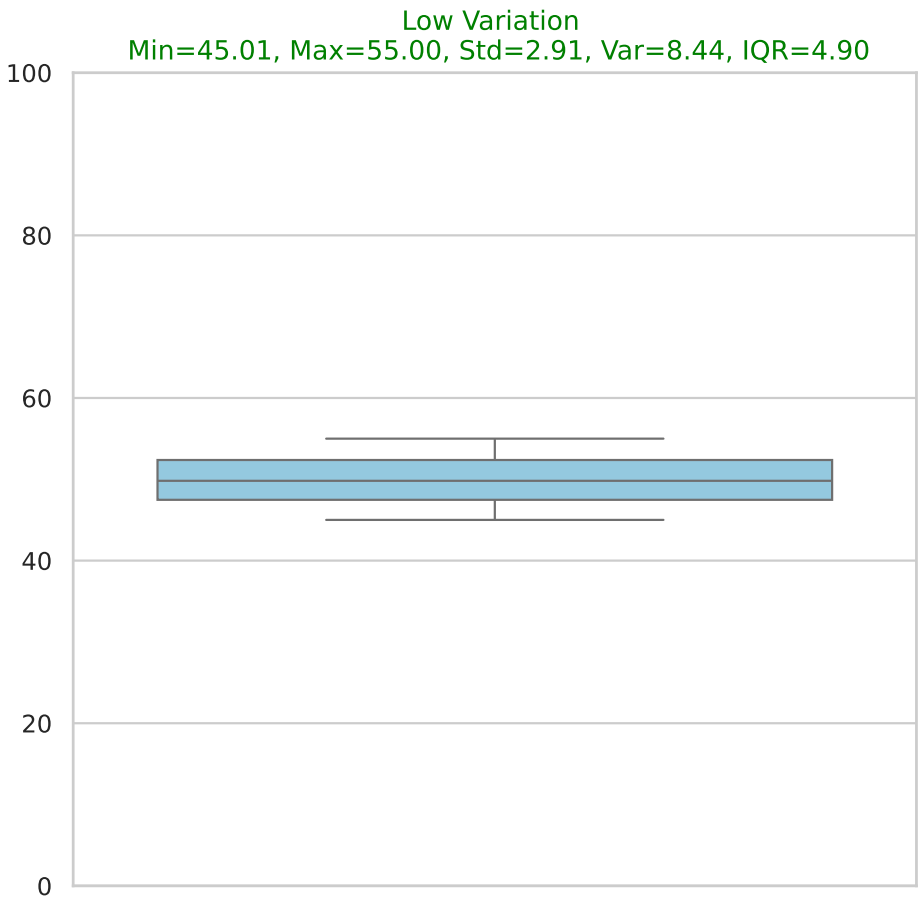
$$\text{IQR} = Q3 - Q1$$

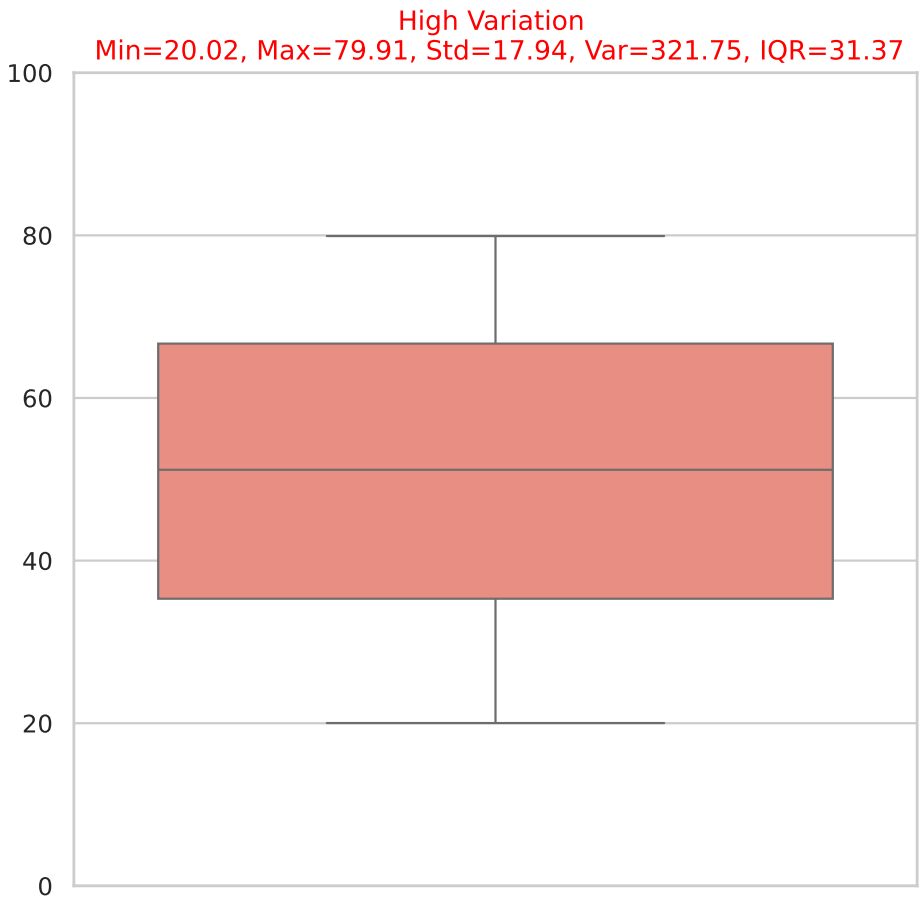
Example

For the dataset: 5, 10, 15, 20, 25, 30, 35, 40, 45

$$Q1 = 15,\ Q3 = 35,\ \text{IQR} = 35 - 15 = 20$$

Min, Max, SD, IQR





Skewness

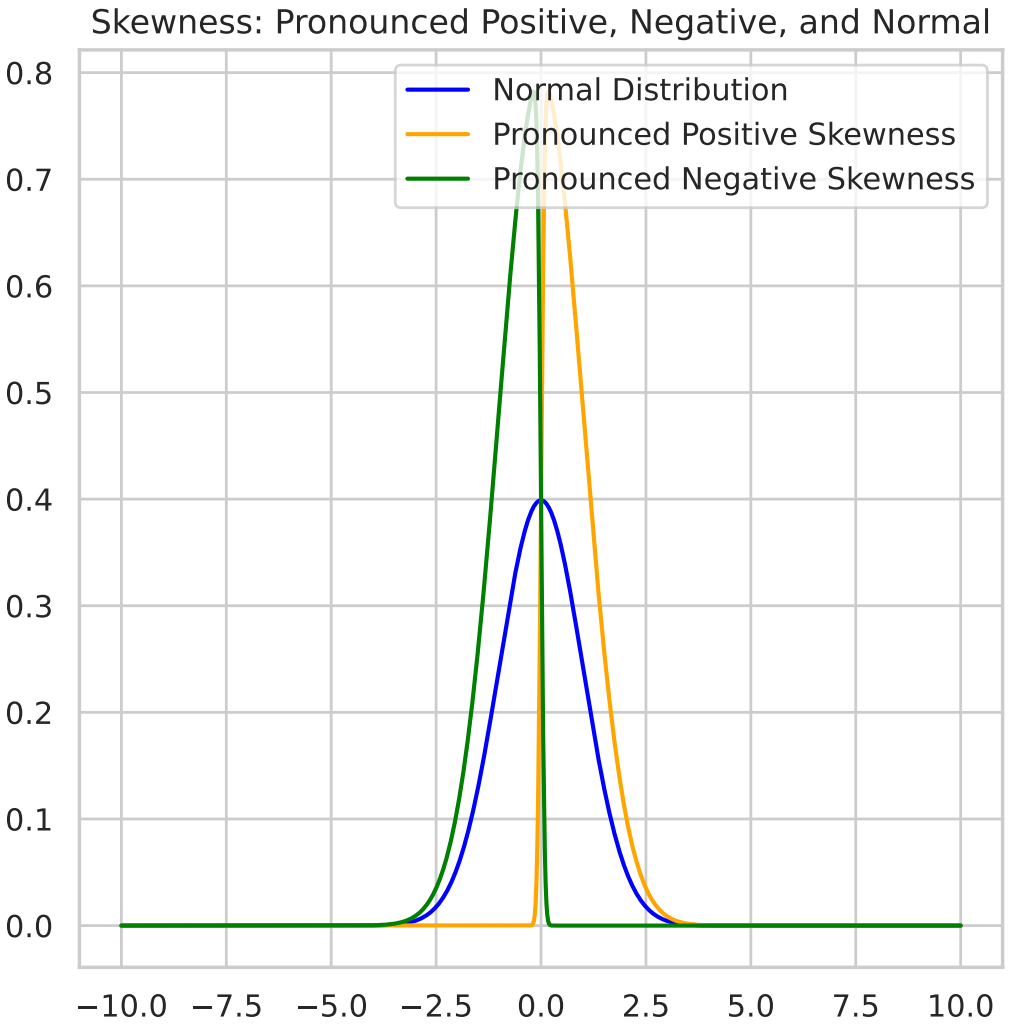
Skewness: A measure of the asymmetry of the probability distribution of a real-valued random variable.

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \mu}{\sigma} \right)^3$$

Example

A right-skewed distribution has more values on the lower end, while a left-skewed distribution has more on the higher end.

Skewness


Skewness: Pronounced Positive, Negative, and Normal
Normal Distribution
Pronounced Positive Skewness
Pronounced Negative Skewness
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Kurtosis

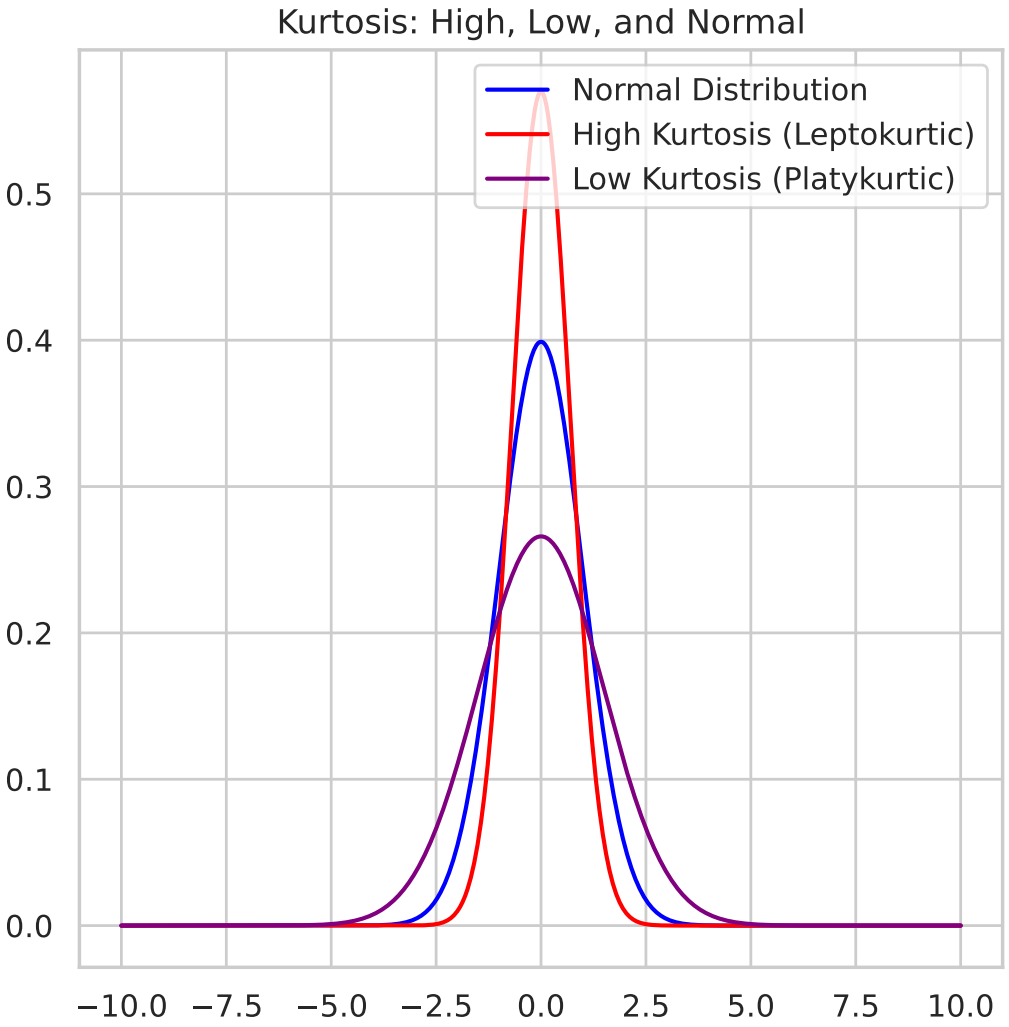
Kurtosis: Describes the "tailedness" of the data distribution. High kurtosis means heavy tails; low kurtosis means light tails.

$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \mu}{\sigma} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Example

A distribution with high kurtosis has extreme values (outliers) compared to a normal distribution.

Kurtosis


Kurtosis: High, Low, and Normal
Normal Distribution
High Kurtosis (Leptokurtic)
Low Kurtosis (Platykurtic)
0.5
0.4
0.3
0.2
0.1
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Current Section

1. Descriptive Statistics

2. Correlation

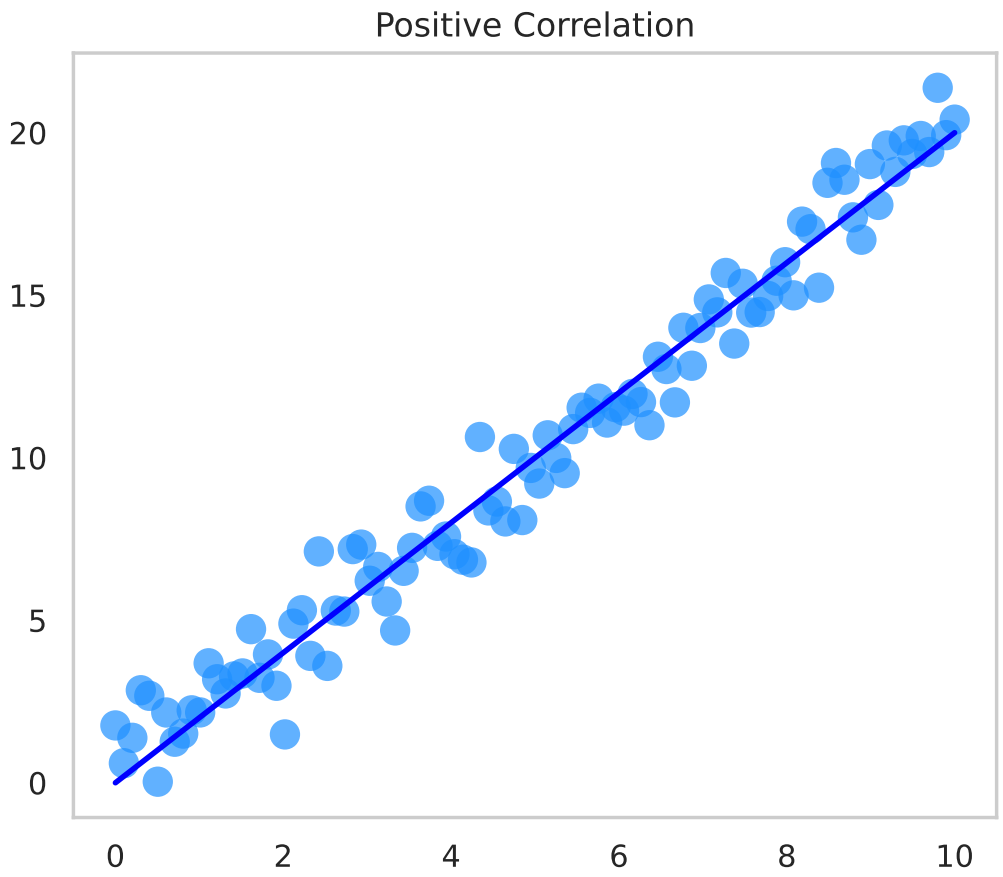
Correlation

Correlation: Measures the strength and direction of the relationship between two continuous numerical variables.

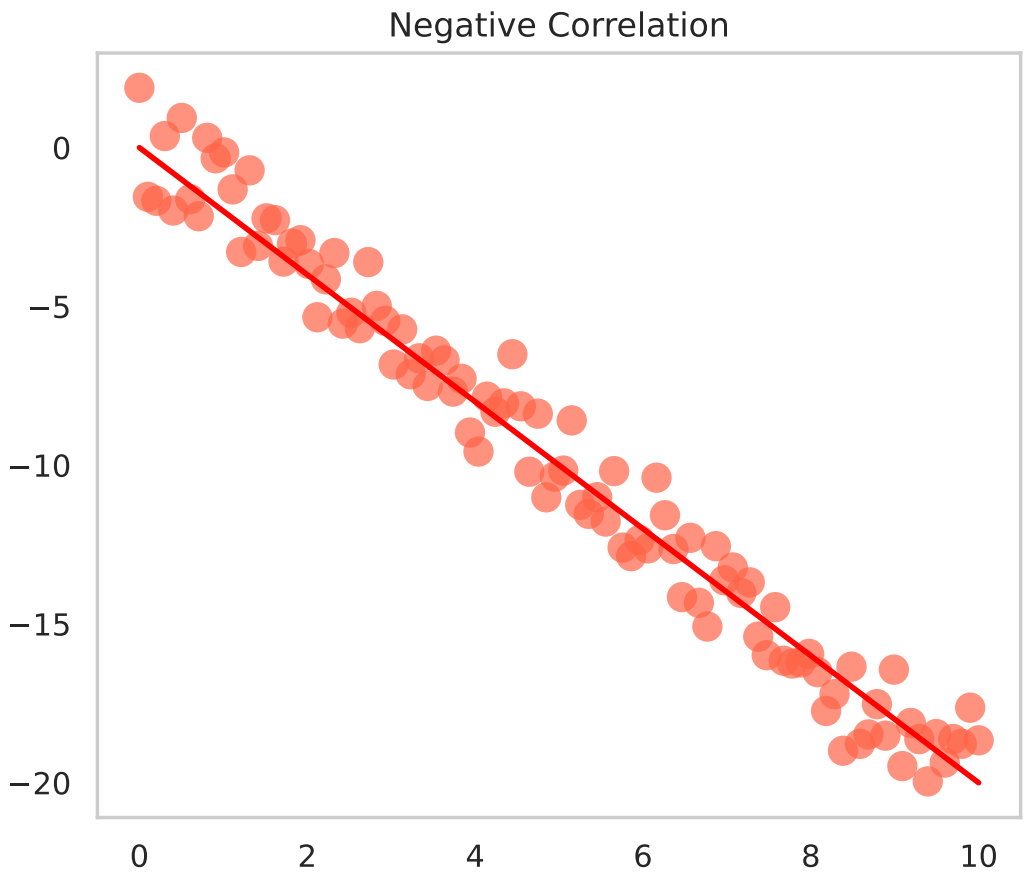
- Positive Correlation: When one variable increases, the other also increases.
- Negative Correlation: When one variable increases, the other decreases.
- No Correlation: Changes in one variable do not predict changes in the other; no relationship exists.

Useful for understanding how two continuous variables are related.

Positive Correlation



Negative Correlation



No Correlation