Applied Statistics for Data Scientists with R

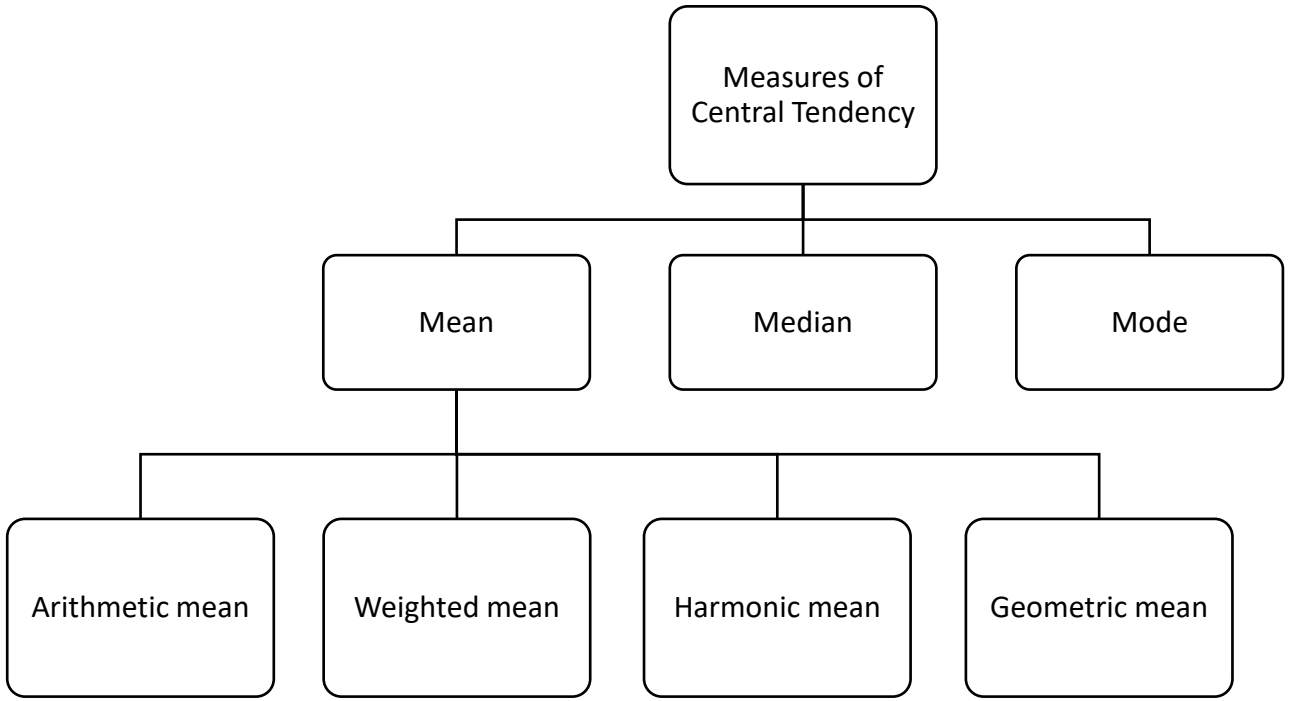
Class 08: Descriptive Statistics and EDA

Learning Objective

1. Common measures of central tendency
2. Common measures of dispersion
3. Sample vs population
4. Degrees of freedom
5. Using dplyr and other packages to get summary statistics

What is central tendency?

- Tendency for a set of values to gather around the middle of the set.



Mean

- Requires values of all points under consideration (sample)
- Data should be numeric

Type	Formula	Usage
Arithmetic mean	$\frac{x_1 + x_2 + \cdots + x_n}{n}$	Used in most cases
Geometric mean	$\sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$	Average growth, Average ratio or percent change
Harmonic mean	$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$	Average speed

N.B. A special case of arithmetic mean is the weighted mean.

Weighted Mean

- The weights of each observation are considered

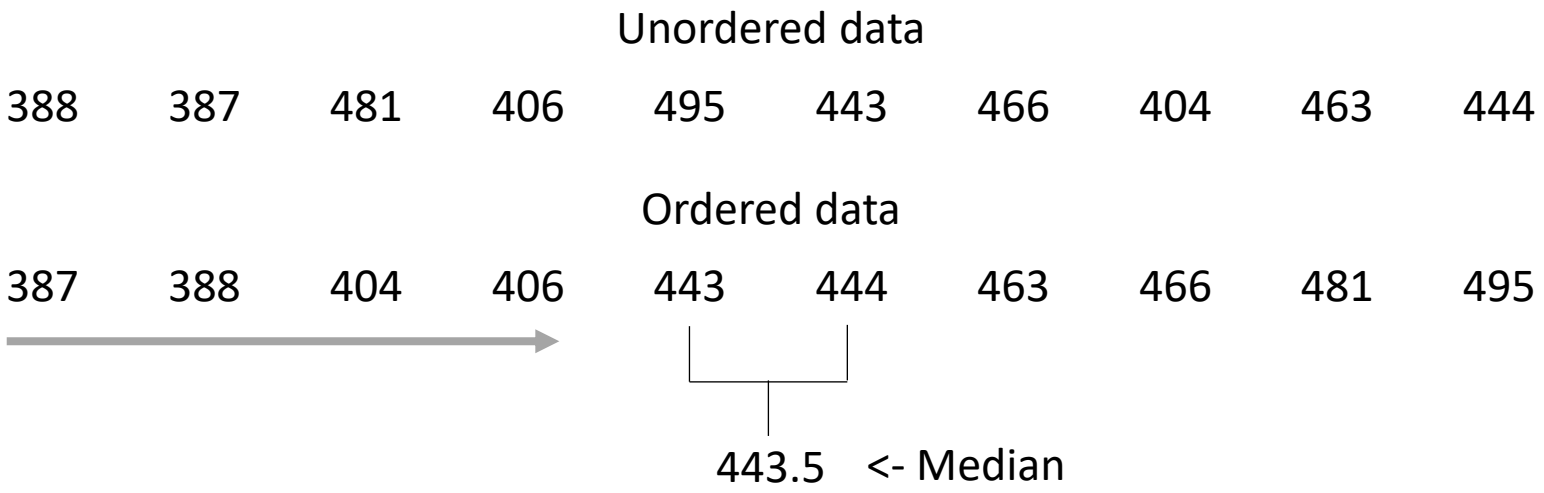
$$\bar{x}_{weighted} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n}$$

In usual case, weights are assumed the same.
Hence, you can consider, $w_1 = w_2 = \cdots = w_n = w$

$$\begin{aligned} &\frac{wx_1 + wx_2 + \cdots + wx_n}{w + w + \cdots + w} \\ &= \frac{w(x_1 + x_2 + \cdots + x_n)}{w \times n} \\ &= \frac{w \sum_{i=1}^{n} x_i}{w \times n} \\ &= \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \end{aligned}$$

Median

- Order the values from highest to lowest or lowest to highest
- Find the middle value
- If there are two middle values, then average of those middle value is the median
- Data should be numeric (ratio and interval scale)



Mode

- Mode is the most frequent value in a set of values
- Both numeric and categorical data are suitable

Which measure of central tendency to use?

Around 90 countries have launched over 13401 satellites, so on average, each country has 148 satellites. However, median is around 7 and mode is 1.

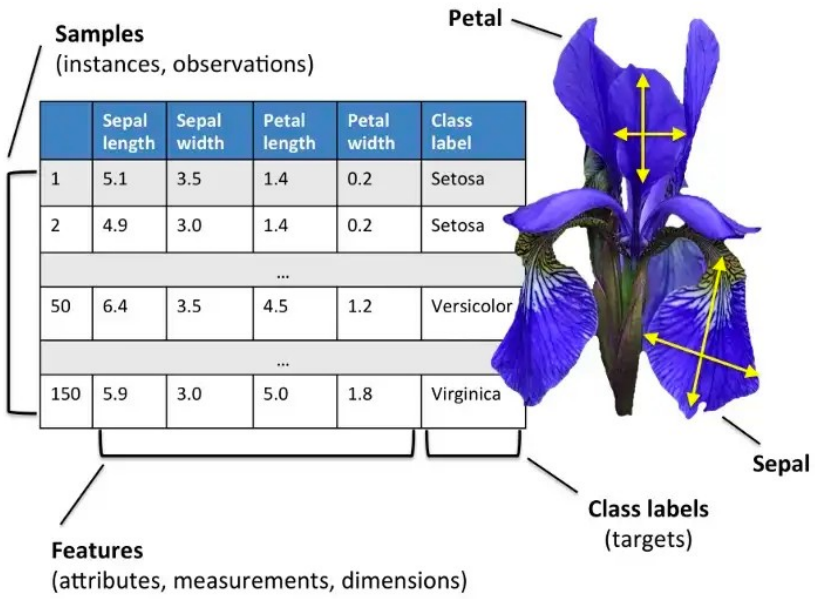
Source: Satellites by Country or Organization

Measures of Dispersion

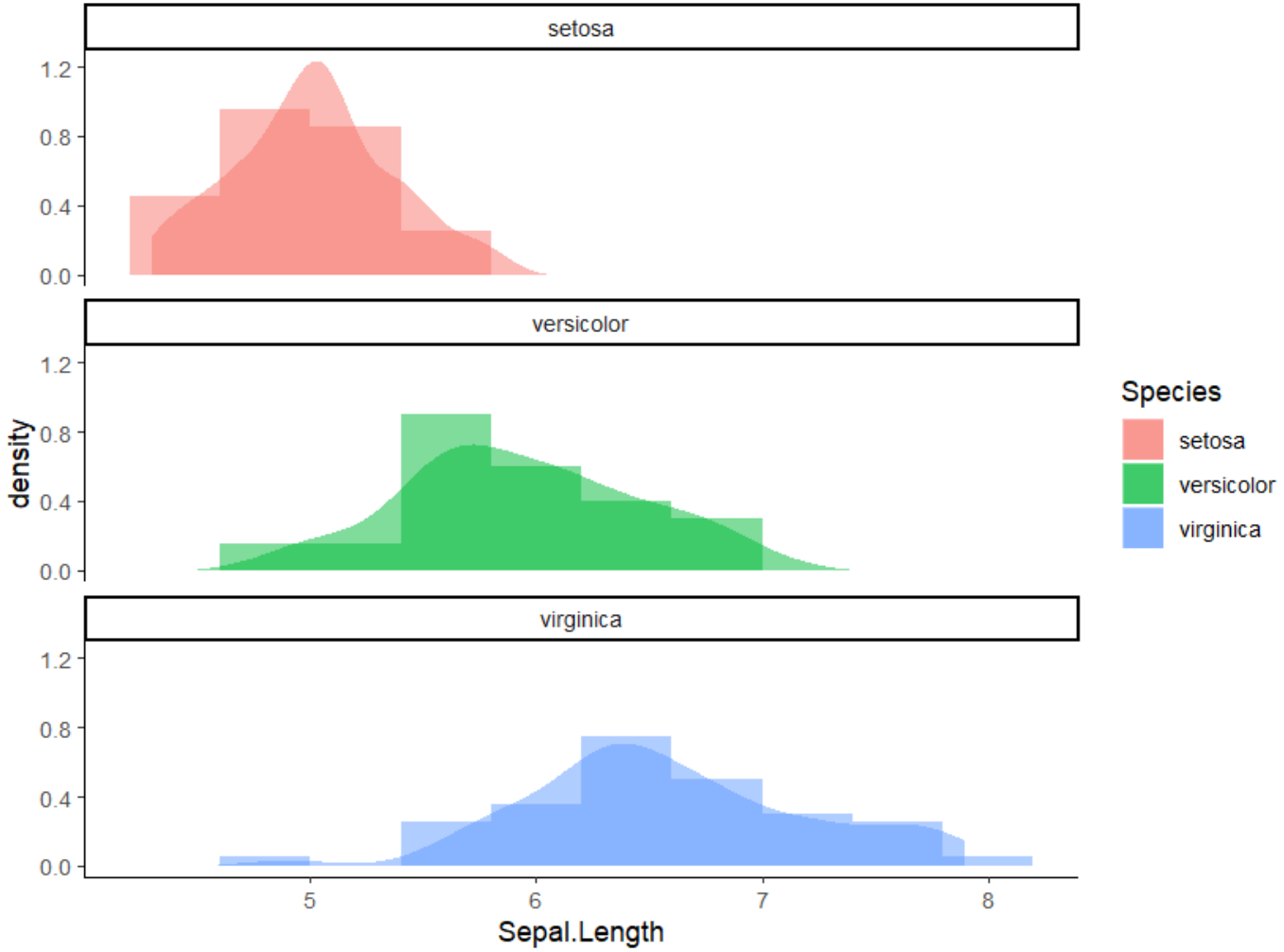
- Quantifies how much the data varies around the average value
- Can be assessed graphically or statistically

(continues)

Measures of Dispersion



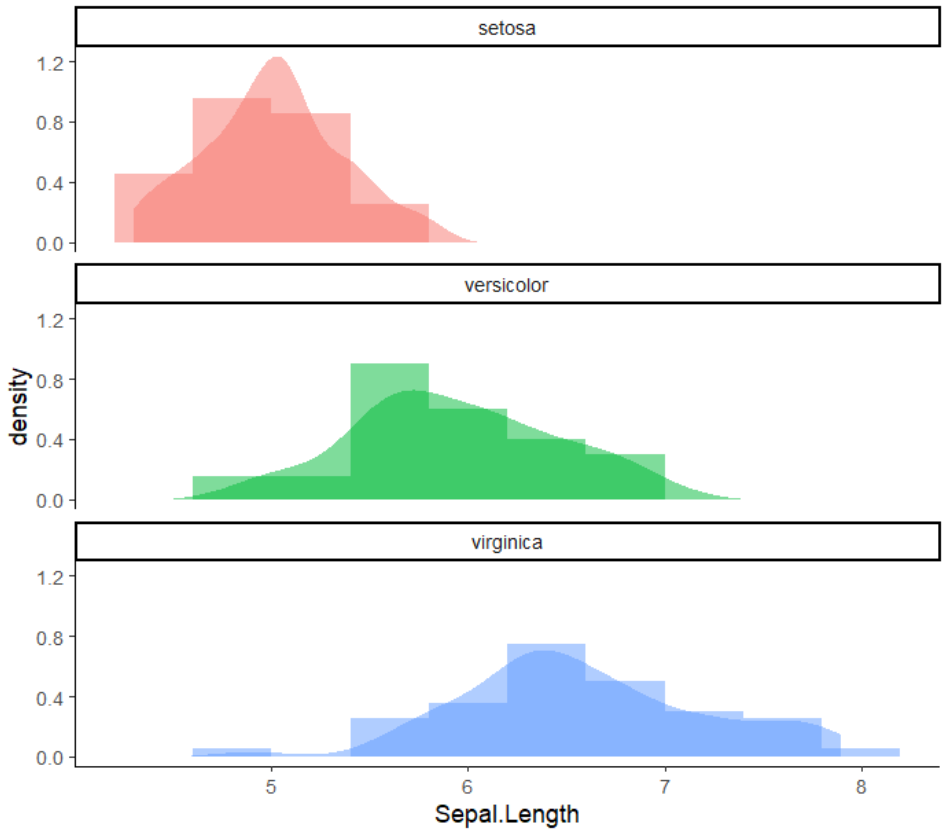
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Measures of Dispersion

Measure	Formula	Usage
Range	Max – Min	Measures the total spread of data; affected by outliers.
Interquartile range	Q3 – Q1	Measures the spread of the middle 50% of the data; less affected by outliers.
Variance	$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	Quantifies how much the data varies around the mean
Standard deviation	$\sqrt{Variance}$	Indicates the average deviation from the mean; easier to interpret compared to variance.
Coefficient of variation	$\frac{SD}{Mean}$	Useful for comparing variability between datasets with different units or means.

Measures of Dispersion



Species	Minimum	Maximum	Range	Variance	SD	IQR	CV
setosa	4.3	5.8	1.5	0.124	0.352	0.400	0.070
versicolor	4.9	7.0	2.1	0.266	0.516	0.700	0.087
virginica	4.9	7.9	3.0	0.404	0.636	0.675	0.097

Sample and Population

Consider a study, where the aim is to determine proportion of ‘Dieback’ or ‘Aga mora’ disease infected Mango trees in Chapai Nawabganj district

Population	Sample
All the Mango trees in Chapai Nawabganj district	Randomly selected 300 trees from all 5400 trees
Population size, N = 5400	Sample size, n = 300

Parameter and Statistic

- You call something parameter, when it is calculated using the population.
  - It is the true value
  - Usually assumed unknown

- You call something statistic, when it is calculated using the sample.
  - Any function of sample observations is called statistic
  - But when a statistic is used to estimate a parameter, it is called estimator

$\mathrm{f}(x) = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ is an estimator of population variance,

but $f(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2$ is not.

Degrees of Freedom

- The maximum number of logically independent values, which are values that have the freedom to vary in the sample.

Sample variance, $s^2 = \frac{\sum_{i=1}^{n}(x_i - mean)^2}{n-1}$

Population variance, $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - mean)^2}{N}$

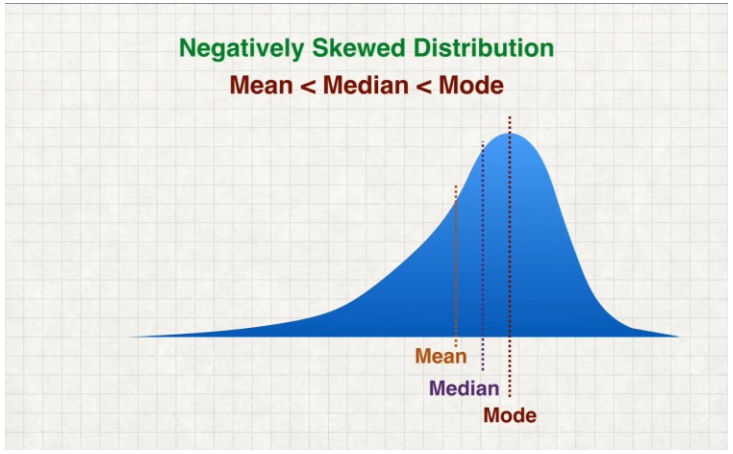
- Each parameter you estimate "uses up" one degree of freedom

Degrees of Freedom

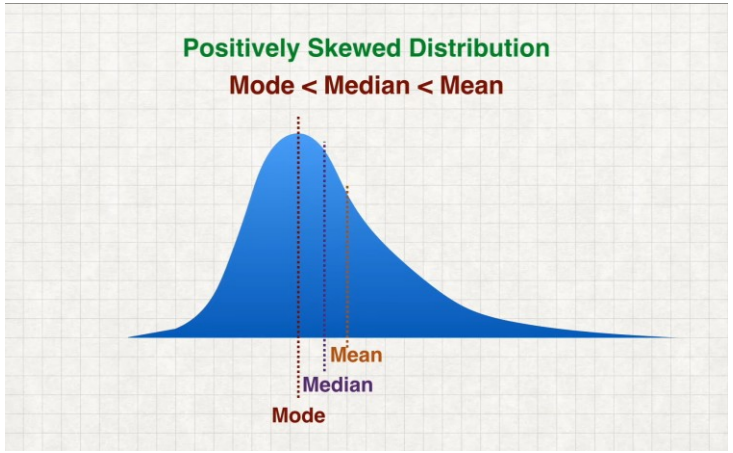
Scenario	Degrees of Freedom	Reason
Variance of sample	n-1	Loses 1 degrees of freedom because 1 parameter (mean) is estimated from the sample
Simple linear regression	n-2	When calculating variance of the residuals two degrees of freedom is lost; one for estimating the intercept parameter and another for estimating the slope parameter.

Shape of Distribution of Data

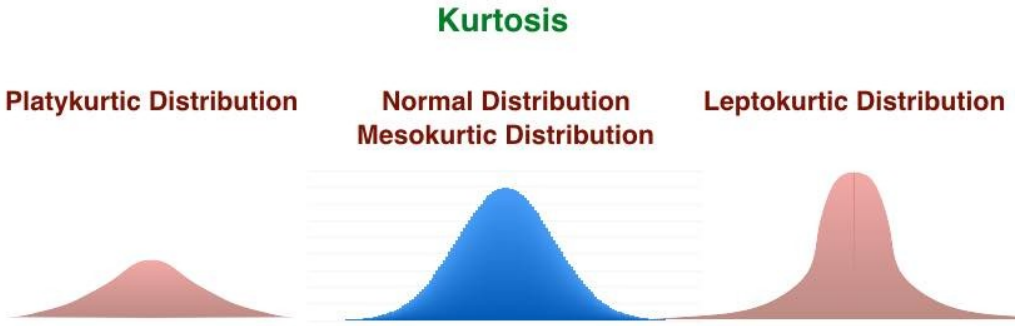
- Skewness measures the symmetry of the distribution
- Kurtosis measures the flatness of a distribution.



Skewness < 0



Skewness > 0



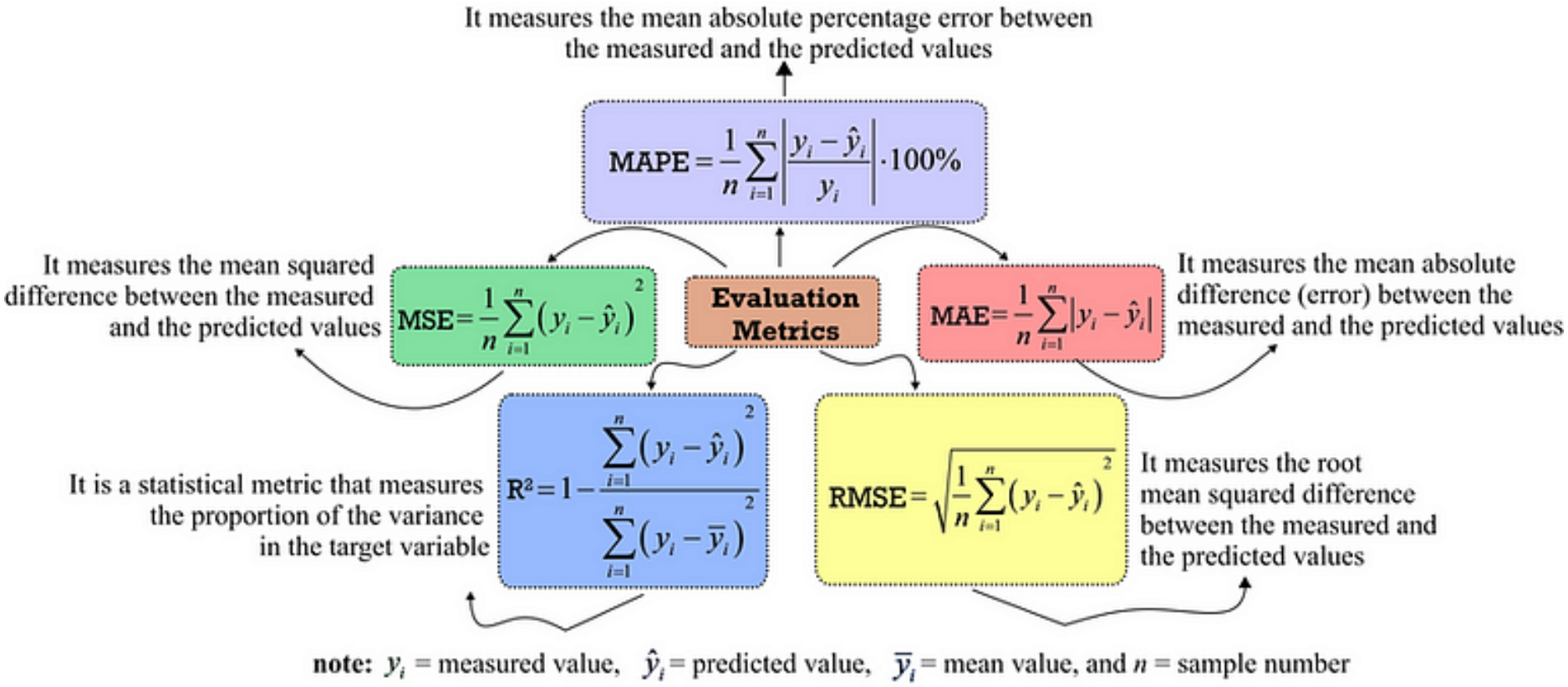
Kurtosis < 0

Kurtosis = 0

Skewness = 0

Kurtosis > 0

Model Evaluation Matrices



Measures of Dispersion in Modelling

Measure	Usage	Example
Mean Squared Error (MSE)	For a cost prediction model, MSE = 17689 $Taka^2$ means the average squared error is 17689 $Taka^2$.	Penalizes larger errors heavily; useful for comparing models.
Root Mean Squared Error (RMSE)	For a cost prediction model, RMSE = 133 Taka means the average error in predictions is about 133 Taka.	Easier to interpret as it's in the original units of the data.
Mean Absolute Error / Deviation (MAE / MAD)	For a cost prediction model, MAE = 133 Taka means the average error in predictions is about 133 Taka.	Simple and intuitive, giving an average error without penalizing large errors more heavily.
Mean Absolute Percentage Error	For a predictive model, MAPE = 4.5% means predictions are off by an average of 4.5% of the actual values.	Useful when percentage-based error is needed; sensitive to small actual values in the denominator.