Unit 1 – Point 1.4: Data Types, Measures of Central Tendency, and Measures of Dispersion

Overview

- **Unit 1 – Point 1.4 Overview:** Introduction to core statistical concepts relevant for data analytics: data types, measures of central tendency, and measures of dispersion.
- **Data Types:** Classification of data into qualitative and quantitative categories with further subtypes.
- **Measures of Central Tendency:** Mean, median, and mode as indicators of where the center of data lies.
- **Measures of Dispersion:** Range, interquartile range, variance, and standard deviation to quantify data spread.

Photo by Luke Chesser on Unsplash

Data Types

Qualitative & Quantitative

- **Qualitative:** Non-numeric, descriptive data such as colors, names, or categories.
- **Quantitative:** Numeric data, further classified into interval and ratio scales.
- **Structured vs. Unstructured:** Structured data stored in rows/columns vs. unstructured text, images, or videos.
- **Examples:** Nominal (gender), Ordinal (rank), Interval (Celsius temperature), Ratio (weight).

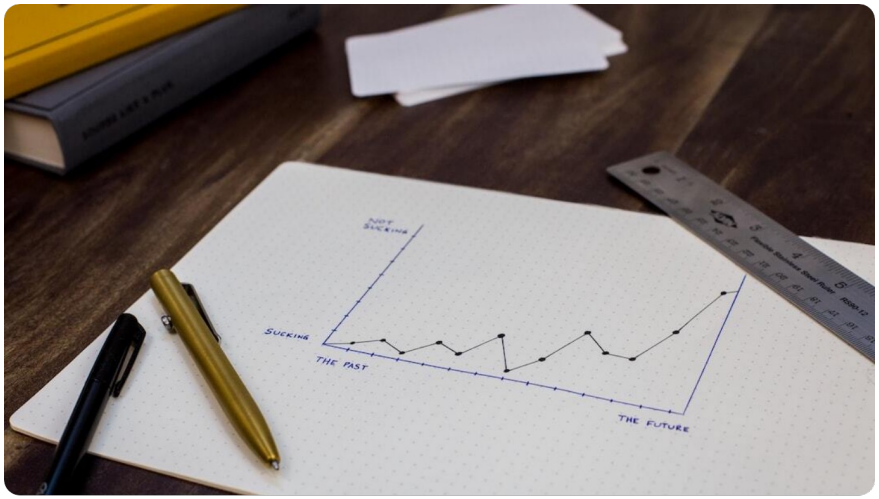
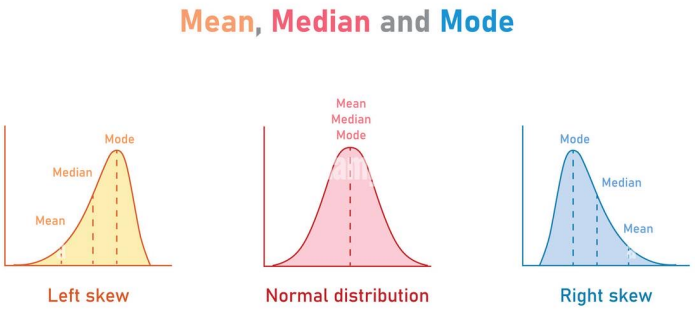


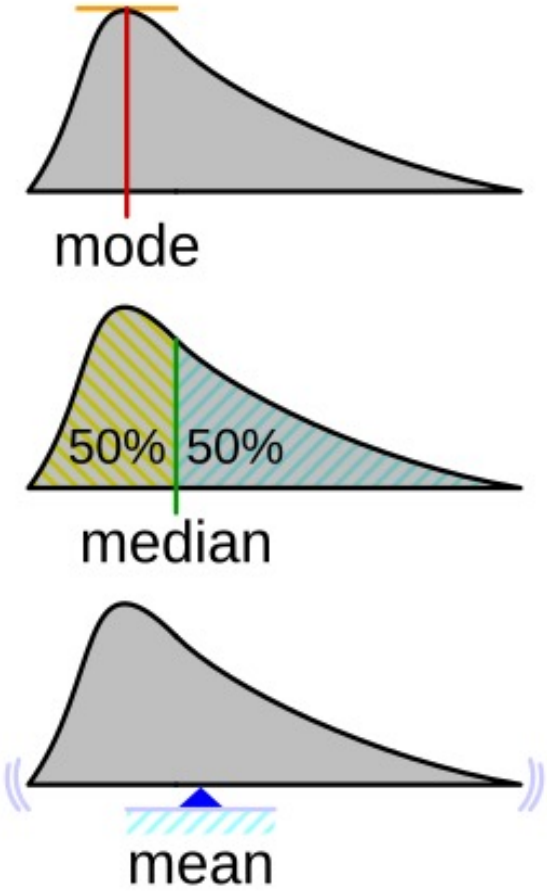
Photo by Isaac Smith on Unsplash

Introduction to Measures of Central Tendency

Finding the Center of Data

- **Shows the Center of Data:** Central tendency summarizes the middle point of a dataset.
- **Common Measures:** Mean, median, and mode are the primary statistics used.
- **Purpose:** Identifies a representative value to simplify large datasets.

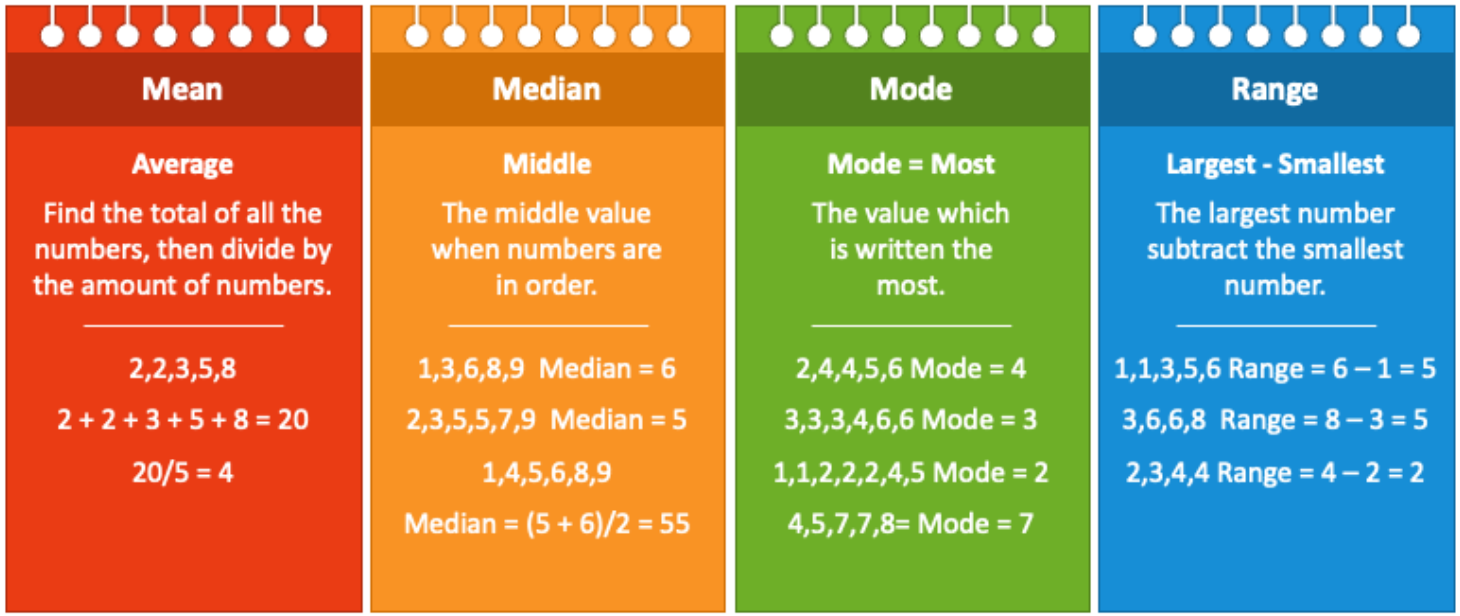




Introduction to Measures of Central Tendency

Finding the Center of Data

MEAN MEDIAN MODE RANGE



Mean: Concept & Formula

Arithmetic Average

- **Definition:** The mean is the sum of all data values divided by the number of values.
- **Formula:** $\bar{x} = \frac{\sum xi}{n}$
- **Best Use:** Works best for balanced, numeric datasets without extreme outliers.

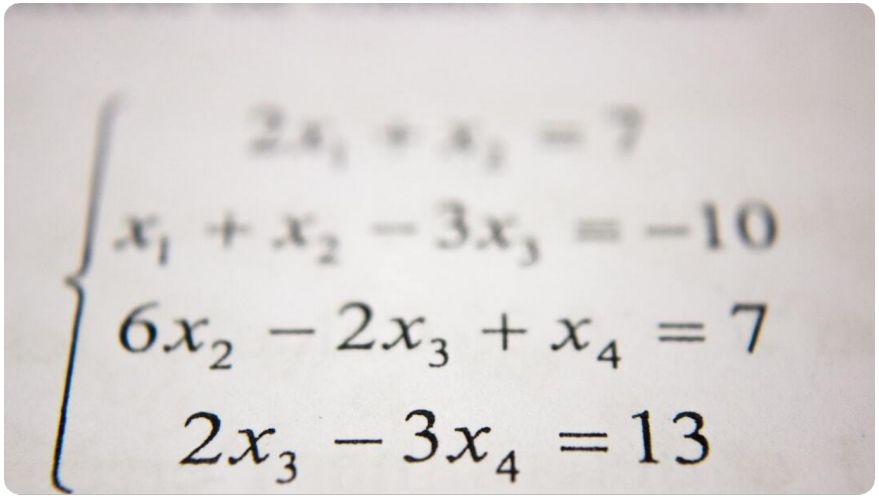


Photo by Antoine Dautry on Unsplash

Mean: Graphical Visualization

The Balance Point

- **Reference Line:** The mean can be represented as a horizontal line across data plots.
- **Balance Point:** Acts as the equilibrium where data values balance out.
- **Symmetry vs. Skew:** In symmetric distributions, the mean is central; in skewed data, it shifts toward the tail.

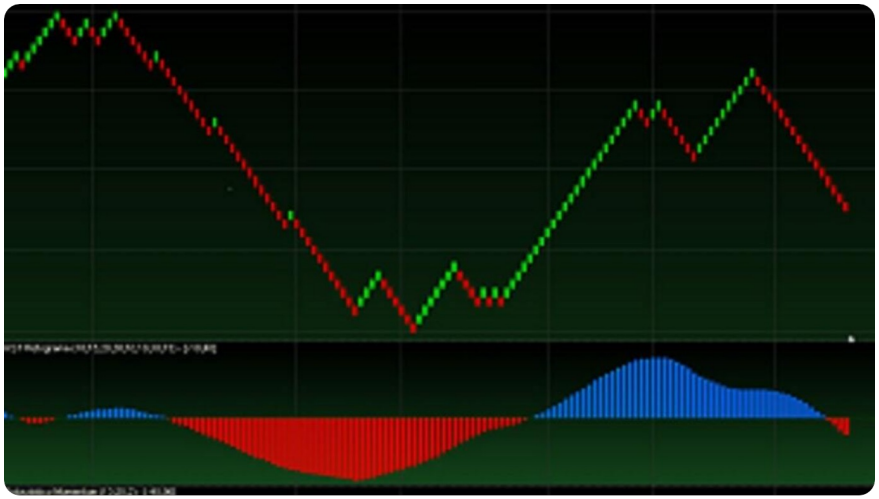
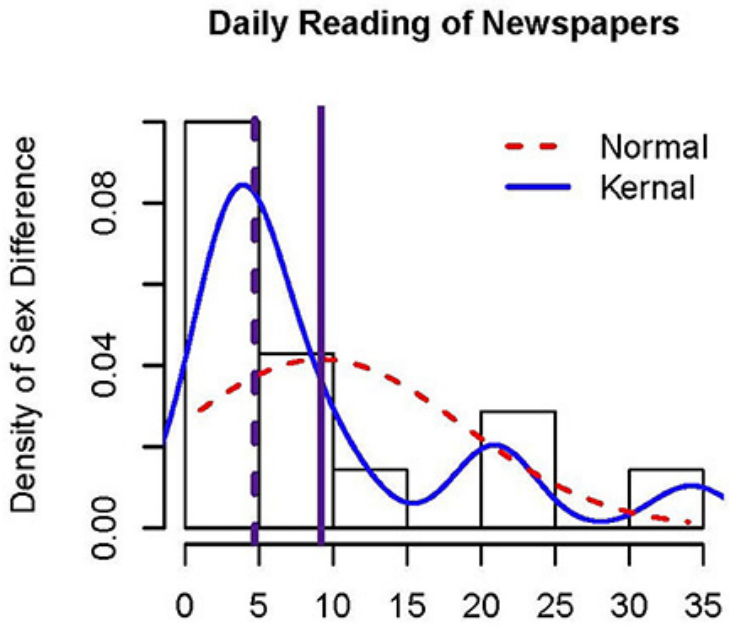
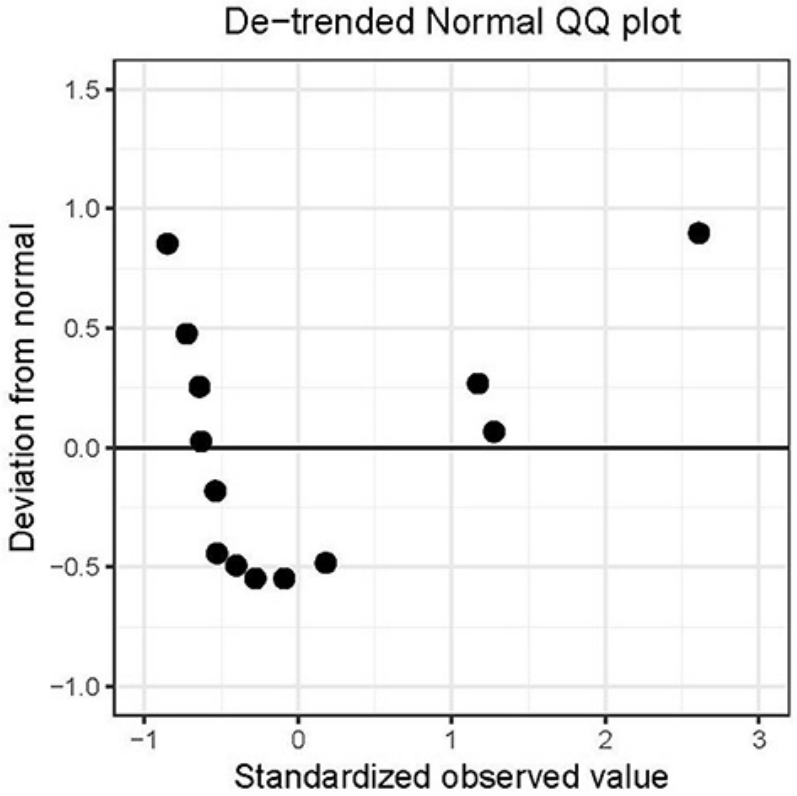


Photo by Jianu tr on Unsplash

Mean: Graphical Visualization

The Reference Line

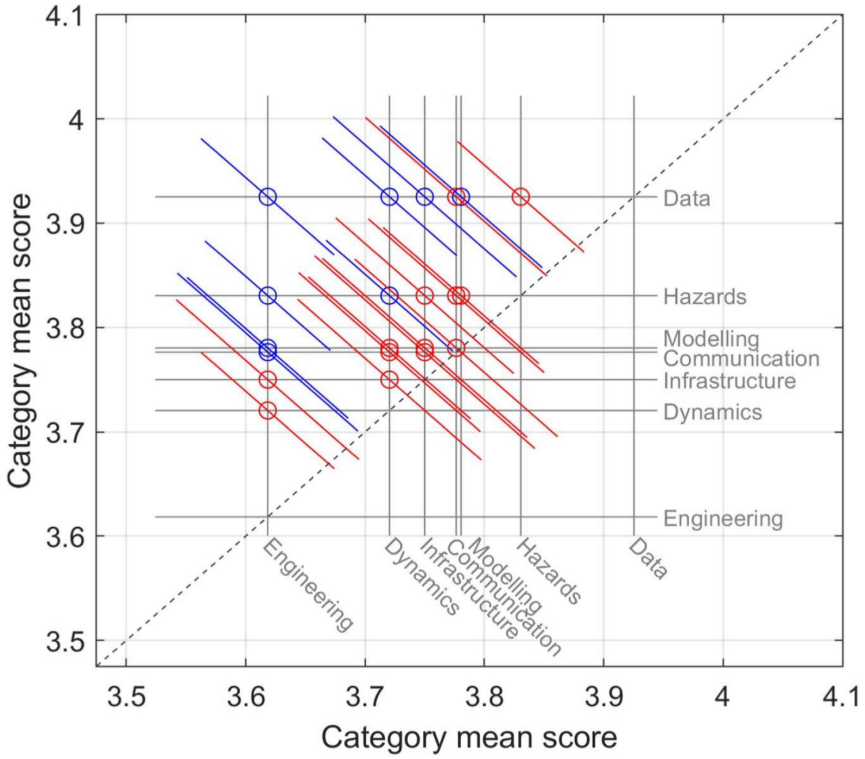


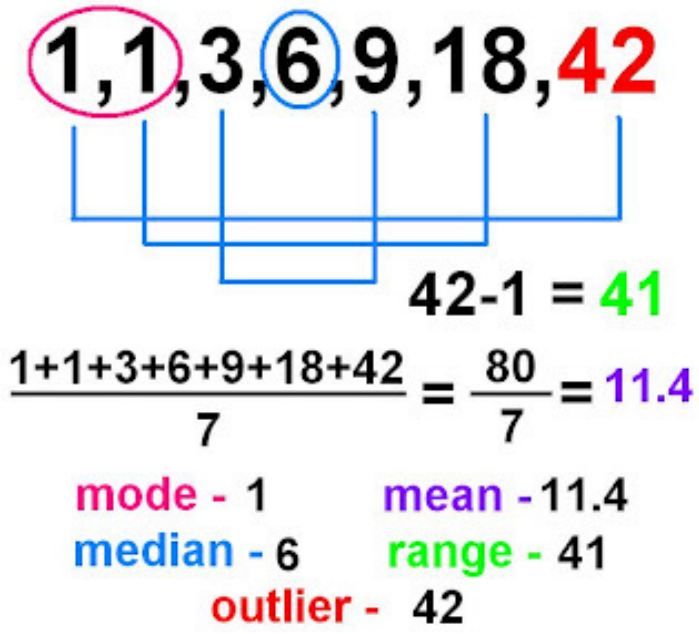


r on Unsplash

Mean: Graphical Visualization

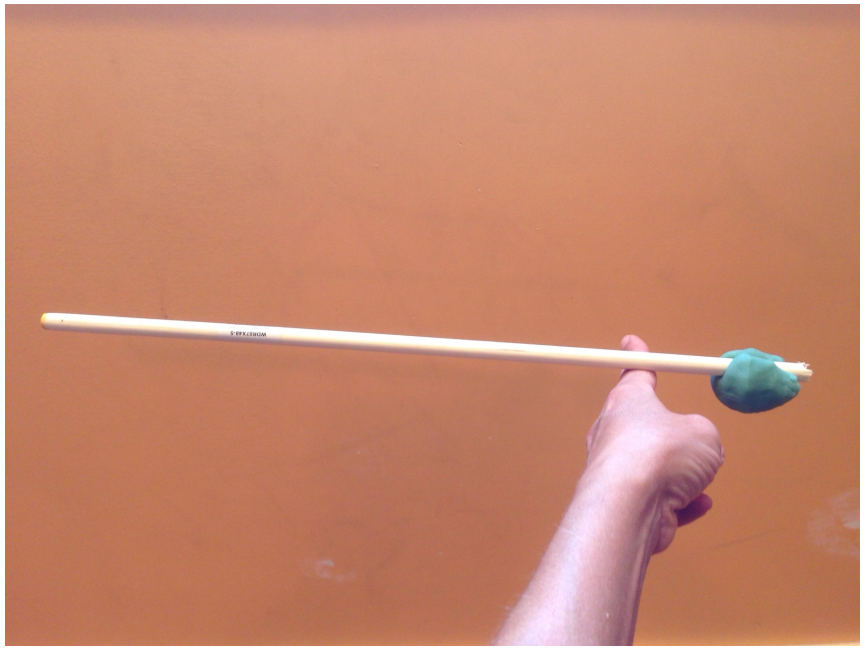
The Reference Line

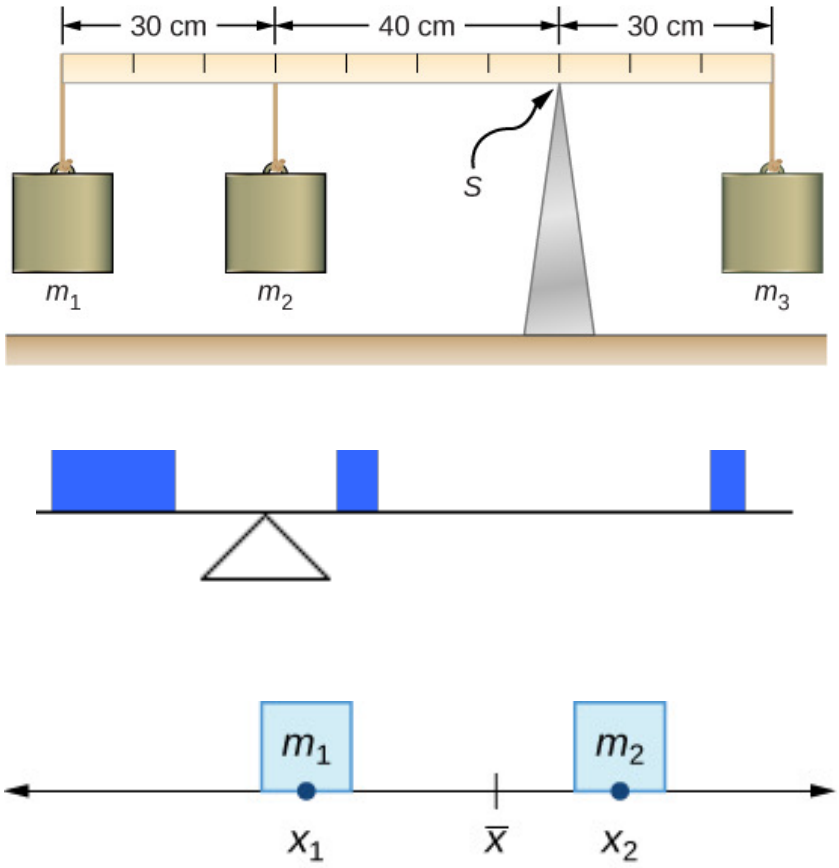




Mean: Graphical Visualization

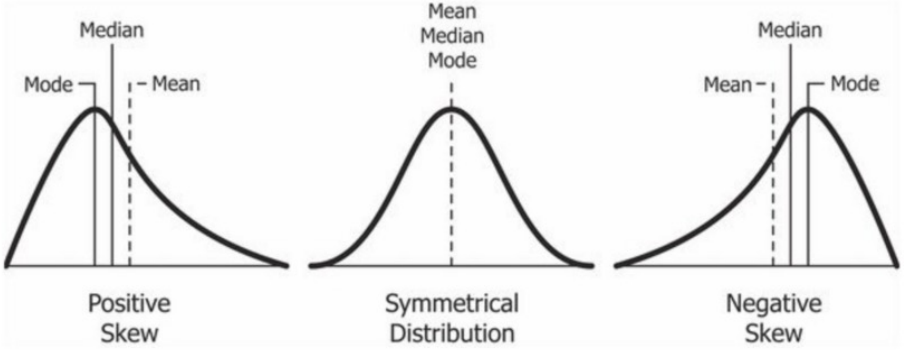
The Balance Point

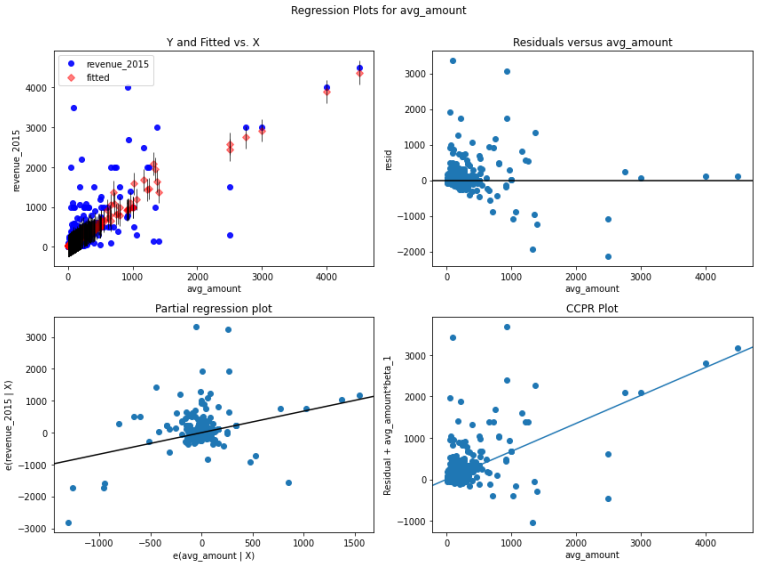


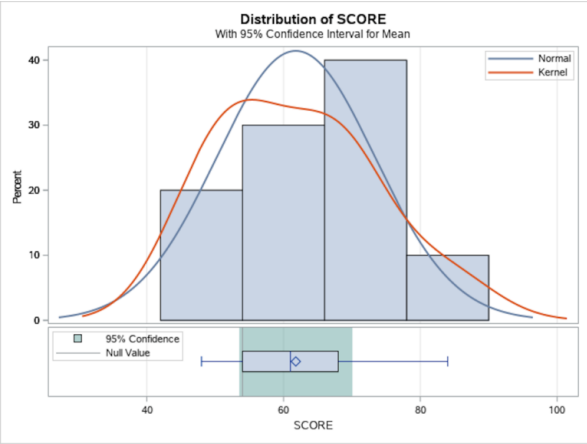


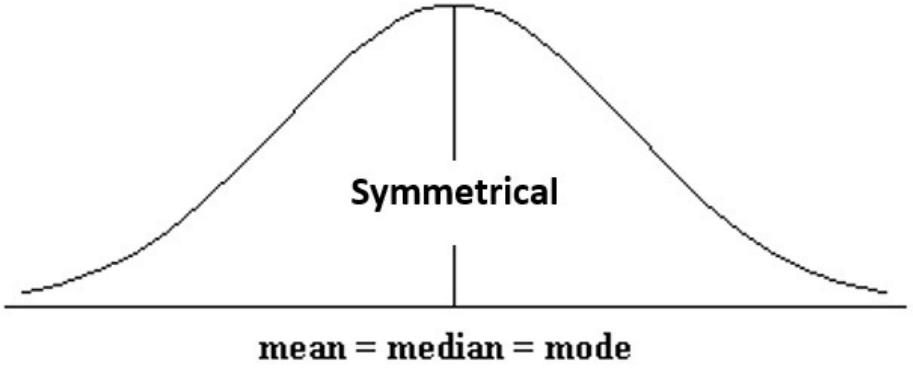
Mean: Graphical Visualization

Symmetry vs. Skew









Median: Concept & Formula

Middle Value in Ordered Data

Definition
The median is the middle value when data is ordered from smallest to largest.

Even Data Sets
If the dataset size is even, the median is the average of the two middle values.

Resistant to Outliers
Less affected by extreme values compared to the mean.

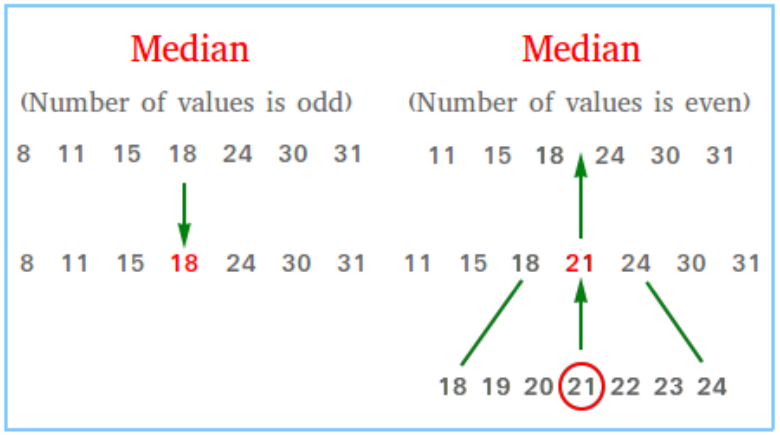
Median

n is odd,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

n is even,

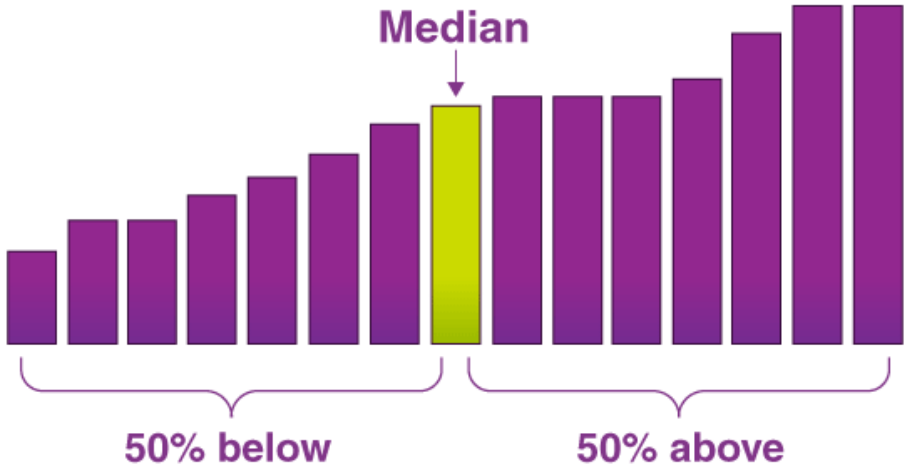
$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{ observation}}{2}$$



Median: Graphical Visualization

Box Plot Representation

- **Median Line:** Shown as the line inside the box of a box plot.
- **Data Halves:** Splits dataset into two equal parts.
- **Outlier Detection:** Box plots highlight extreme values beyond whiskers.



Mode: Concept & Formula

Most Frequent Value

- **Definition:** The mode is the value that appears most often in a dataset.
- **Multiplicity:** Can be unimodal, bimodal, or multimodal depending on data.
- **Best Use:** Ideal for categorical data and detecting common values.

1. **Mean** = Sum of scores divided by the number of scores (often referred to as the statistical average)

Pronounced "x-bar"

$$\bar{X} = \frac{\sum X}{N}$$

Capital Sigma for "Sum of"

"X" represents each score

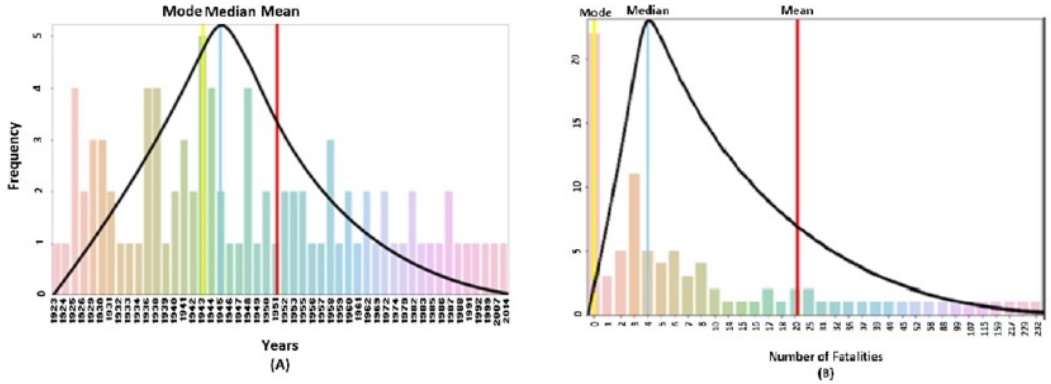
N represents the number of scores

2. **Median** = Middle Most Number

M_d

3. **Mode** = Most Frequently Occurring Number

M_o



Mode: Graphical Visualization

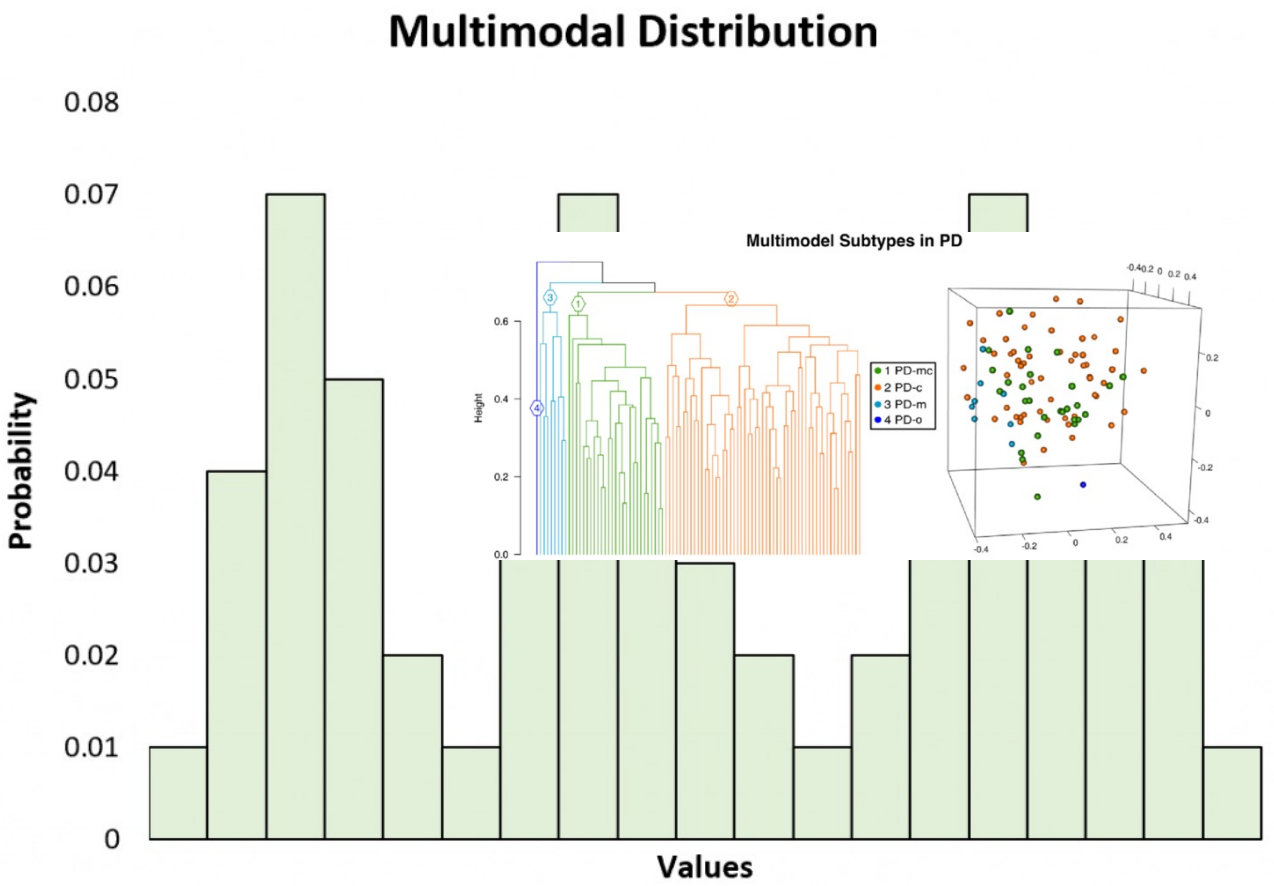
Frequency Peaks

Histogram Peaks
The highest bar in a histogram represents the mode.

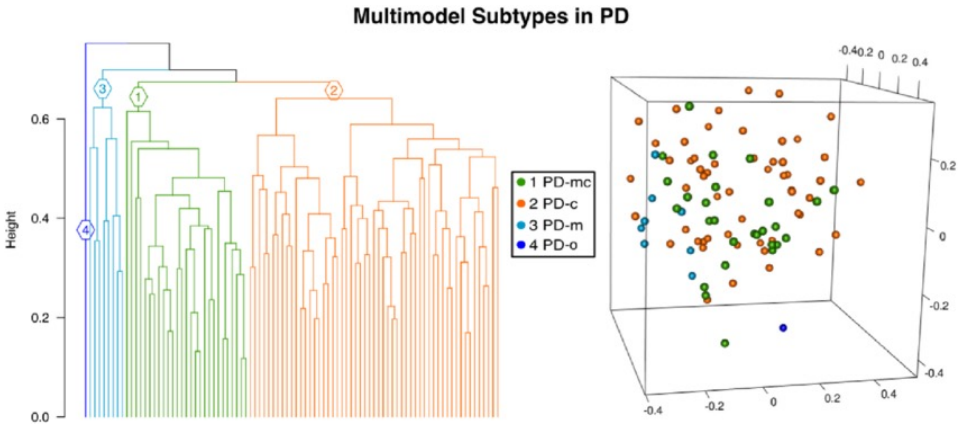
Multiple Modes
Multiple peaks indicate bimodal or multimodal distributions.

Pattern Detection
Modes reveal data clustering and subgroup patterns.

Mode: Graphical Visualization

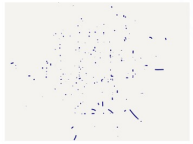


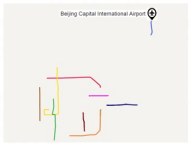
Mode: Graphical Visualization

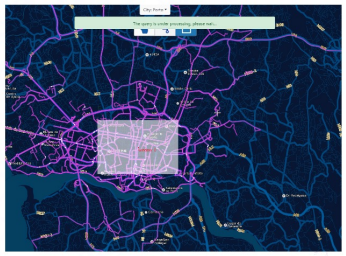


Clustering Millions of Trajectories

- Traffic Trend Analysis for Tier 1 Cities like Beijing



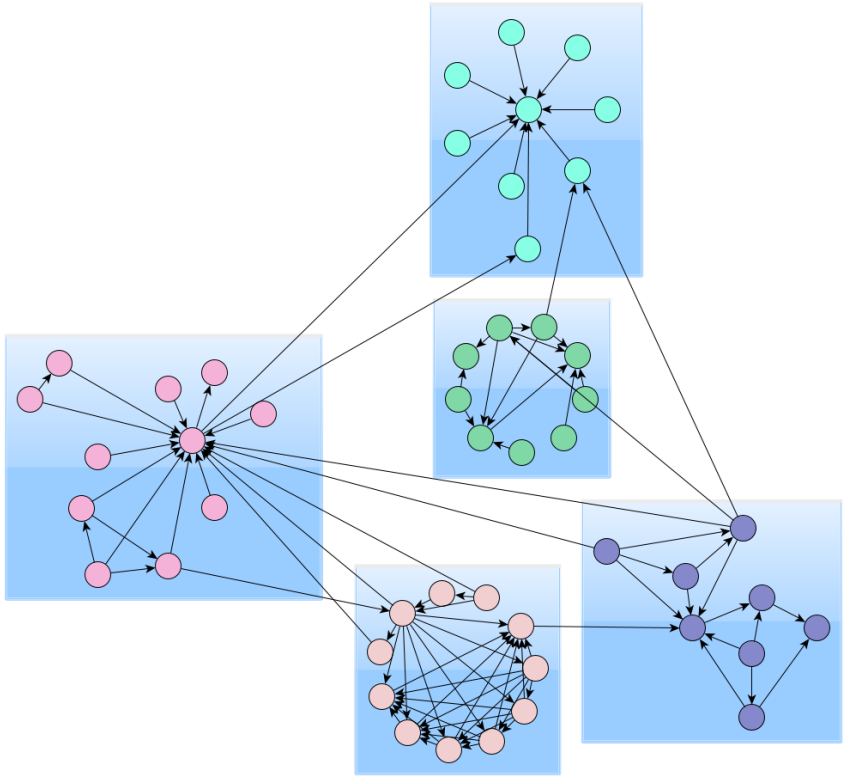
TRACLUS*

k-paths (k=10)

https://t4research.github.io/

(CCF A) PVLDB 2020. Sheng Wang, Zhifeng Bao, J. Shane Culpepper, Timos Sellis, Xiaolin Qin. Fast Large-Scale Trajectory Clustering.

*Jae-Gil Lee et al. Trajectory clustering: a partition-and-group framework. SIGMOD Conference 2007: 593-604



Introduction to Measures of Dispersion

Understanding Data Spread

- **Definition:** Measures of dispersion quantify how data points vary around the central value.
- **Common Measures:** Range, interquartile range, variance, and standard deviation.
- **Purpose:** Complements central tendency by showing variability and reliability.

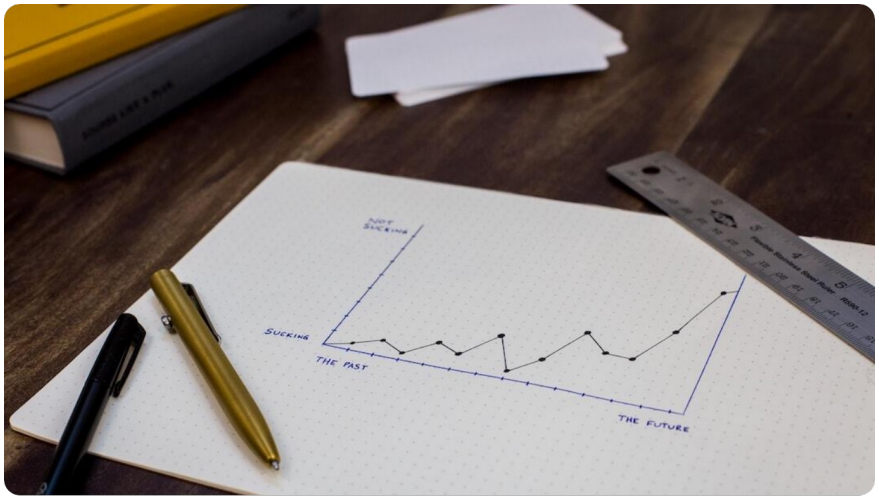
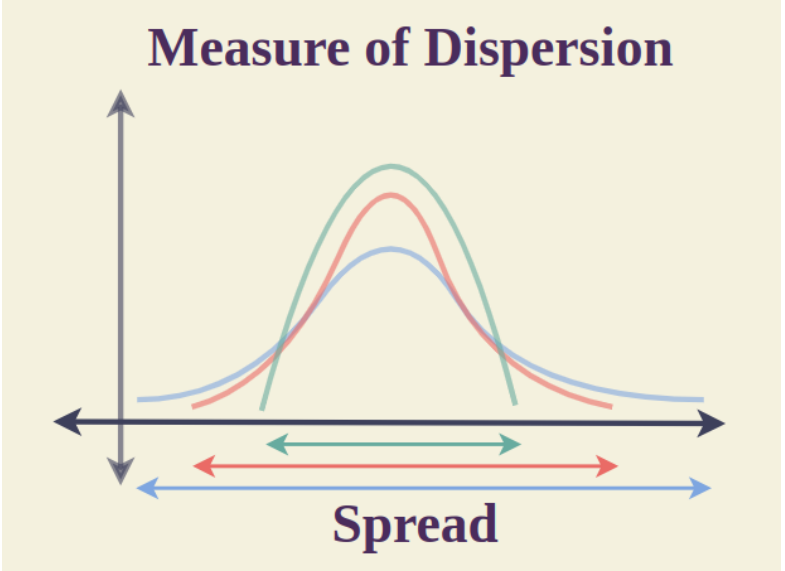
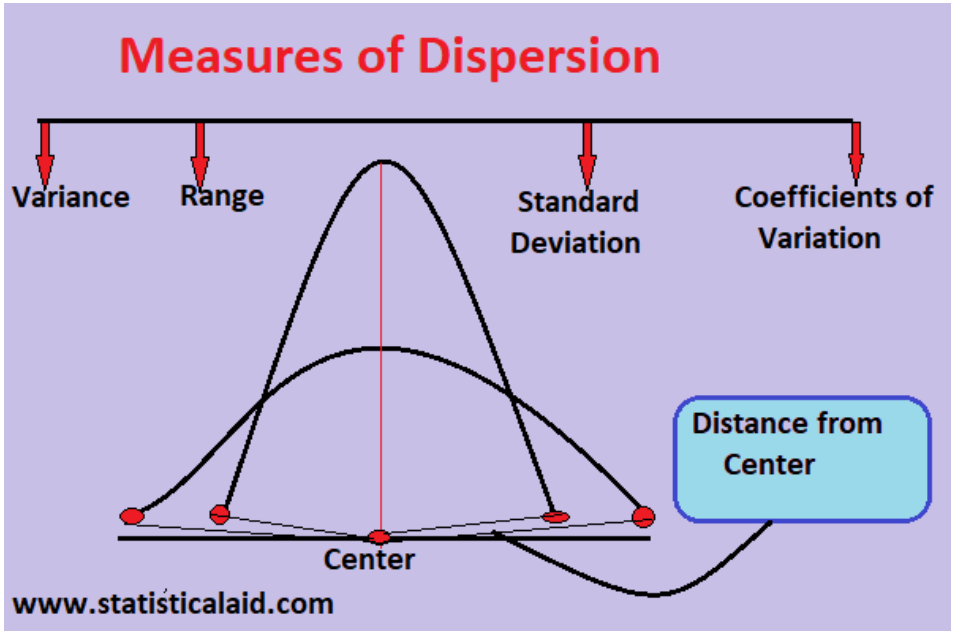


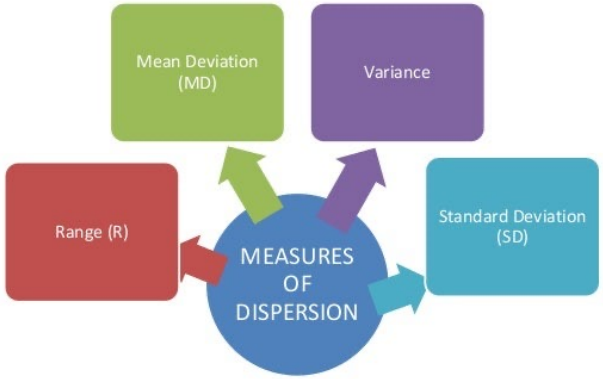
Photo by Isaac Smith on Unsplash

Introduction to Measures of Dispersion

Understanding Data Spread



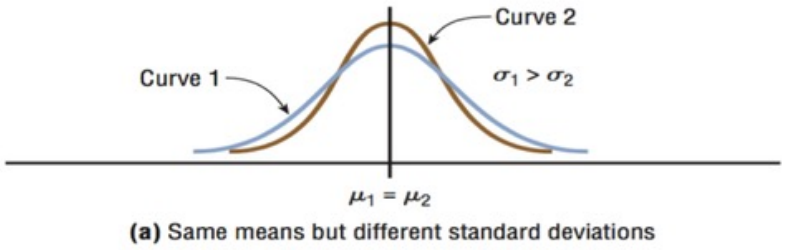




Significance of Understanding Data Spread

Variability and Reliability

- **Reliability:** Data spread affects the consistency of conclusions.
- **Outlier Detection:** Helps identify extreme values that may distort analysis.
- **Same Mean, Different Spread:** Two datasets with identical means can differ significantly in variability.



(a) Same means but different standard deviations

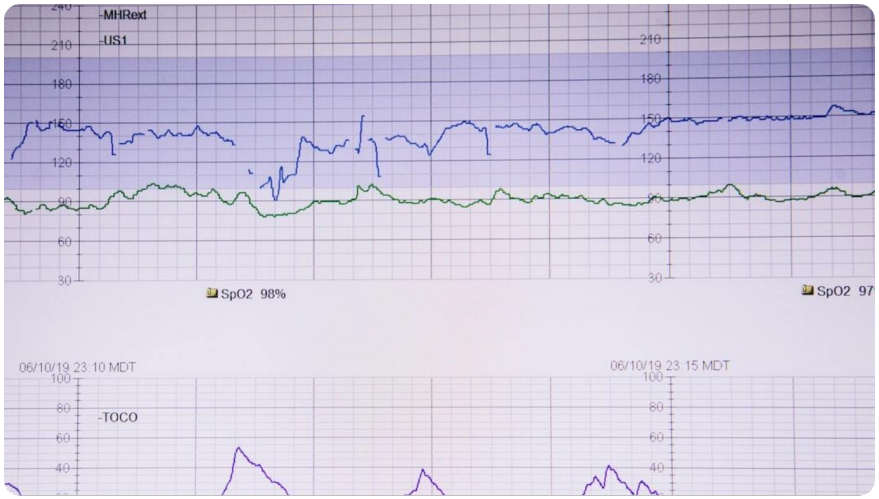


Photo by Alexander Grey on Unsplash

Range

Simplest Measure of Spread

- **Formula:** Range = Maximum Value − Minimum Value.
- **Quick Measure:** Easy to compute but sensitive to outliers.
- **Example:** [5, 7, 8, 12] ➜ Range = 12 − 5 = 7.

Photo by Sven Mieke on Unsplash

Interquartile Range (IQR)

Middle 50% Spread

- **Formula:** IQR = Q3 − Q1.
- **Robustness:** Less affected by extreme values than range.
- **Example:** Q1 = 5, Q3 = 15 ➔ IQR = 10.

Photo by Ricardo Arce on Unsplash

Visualizing Range, Quartile Range, and IQR

Box Plot Example with Dataset: 4, 7, 9, 15, 20

Range
The full spread from minimum to maximum value (20 - 4 = 16), shown as whiskers in the plot.

Quartile Range
Spread between any two quartiles; in this example, Q3 - Q2 or Q2 - Q1 could be shown in shaded segments.

Interquartile Range (IQR)
The middle 50% of the data, Q3 - Q1 (15 - 7 = 8), highlighted as the central box in the plot.

Understanding Range in Statistics

Definition, Calculation, and Uses

Definition
Range is the difference between the largest and smallest values in a dataset.

Example
For the dataset 4, 7, 9, 15, 20 → Range = 20 - 4 = 16.

Uses and Limitation
Quick measure of spread; highly sensitive to extreme values or outliers.

Quartile Range & Interquartile Range (IQR)

Breaking Down Quartiles and Middle Spread

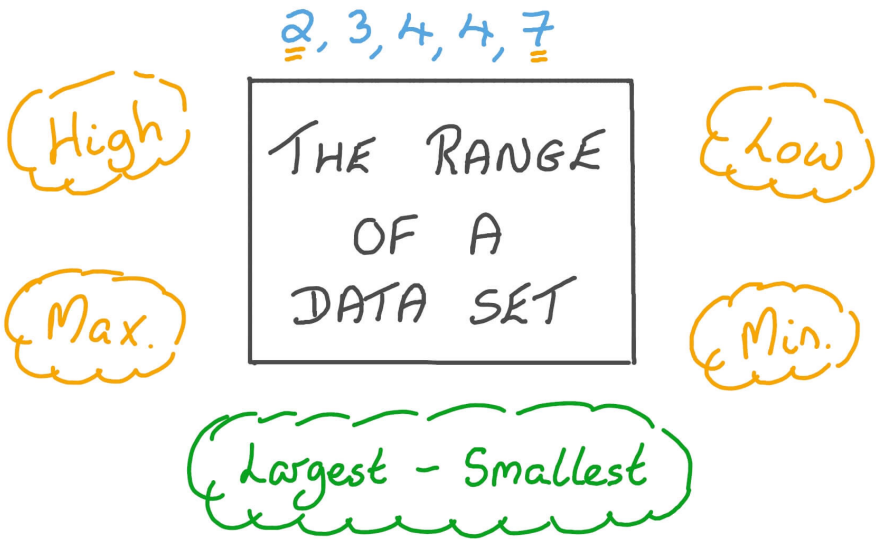
Quartiles
Q1 (25%), Q2 (median, 50%), and Q3 (75%) divide ordered data into four equal parts.

Quartile Range
Spread between any two quartiles (e.g., Q3 – Q2 or Q2 – Q1).

IQR
Middle 50% of data, calculated as Q3 – Q1. Robust to outliers.

Visualizing Range, Quartile Range, and IQR

Range



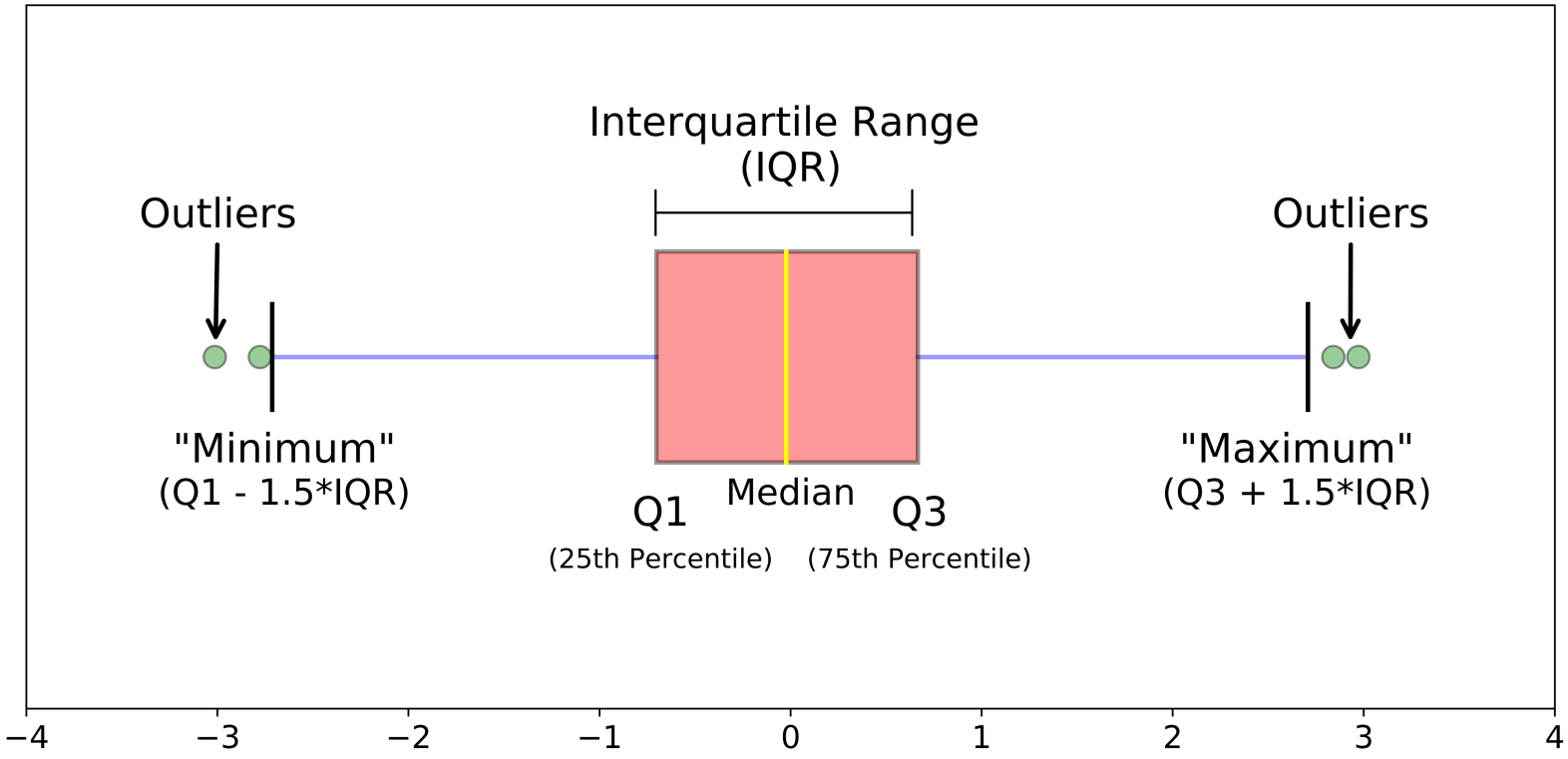
A: 1, 1, 8, 8, 8, 10

B: 1, 3, 6, 8, 12, 12

mean	median	mode	range
6	**8**	8	9
7	**7**	12	11

Visualizing Range, Quartile Range, and IQR

Quartile and Interquartile Range



Visualizing Range, Quartile Range, and IQR

Quartile and Interquartile Range

Interquartile Range

The **interquartile range** is the difference between the upper quartile and the lower quartile in a data set.

$$IQR = Q3 - Q1$$

3 | 4 | 4 | 5 | 6 | 8 | 10

Q1 = 4, M = 5, Q3 = 8

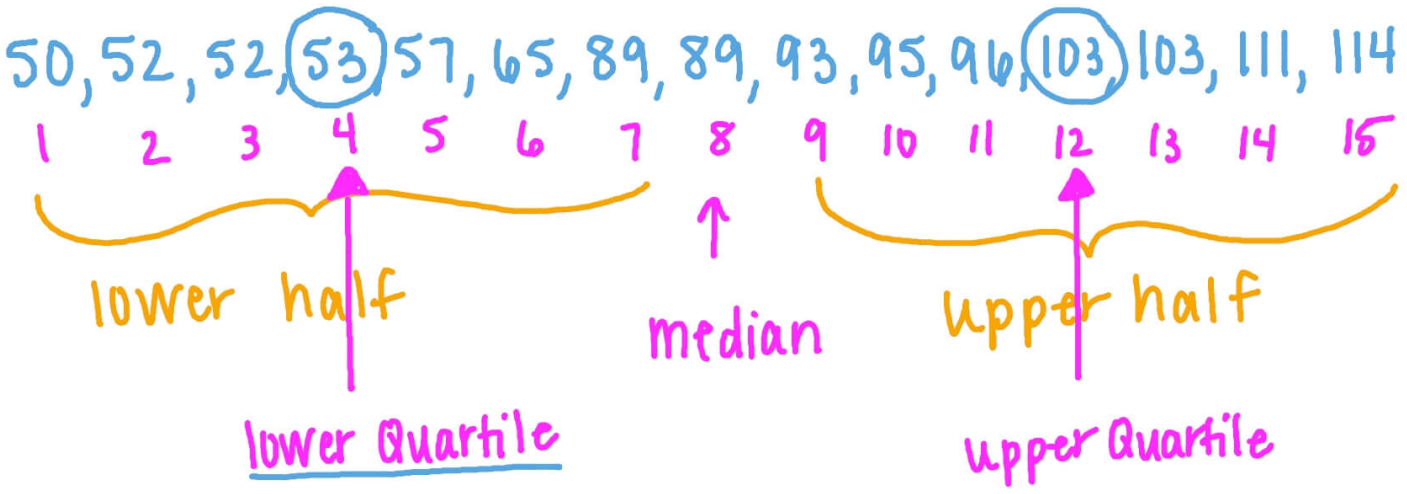
For this data set,
$IQR = Q3 - Q1 = 8 - 4 = 4$

The five-number summary for this data is:

Statistic	Value
Lowest Value	3
Lower Quartile (Q1)	4
Median (M or Q2)	5
Upper Quartile (Q3)	8
Highest Value	10

Visualizing Range, Quartile Range, and IQR

Determine the upper and lower quartiles of the following set of data:
114, 103, 50, 52, 96, 103, 93, 53, 65, 57, 52, 89, 111, 89 and 96.



Lower Quartile is 53 and Upper Quartile is 103

Visualizing Range, Quartile Range, and IQR

Quartile and Interquartile Range

Example 1:
Find the interquartile range of the data below:

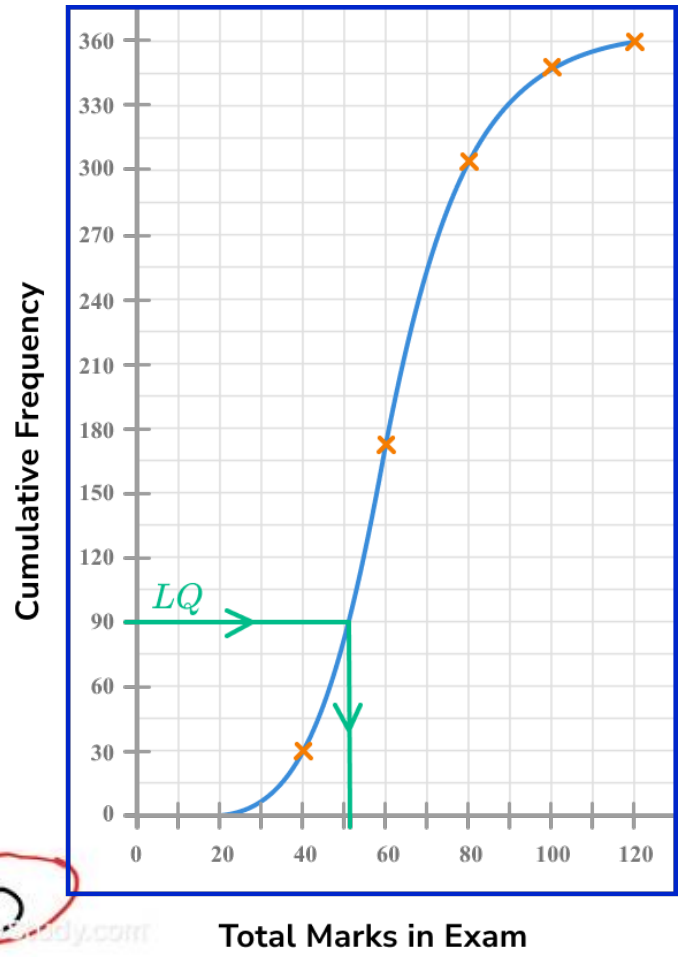
15, 27, 16, 18, 20, 25, 13

13 15 16 18 20 25 27

Q1 Q3

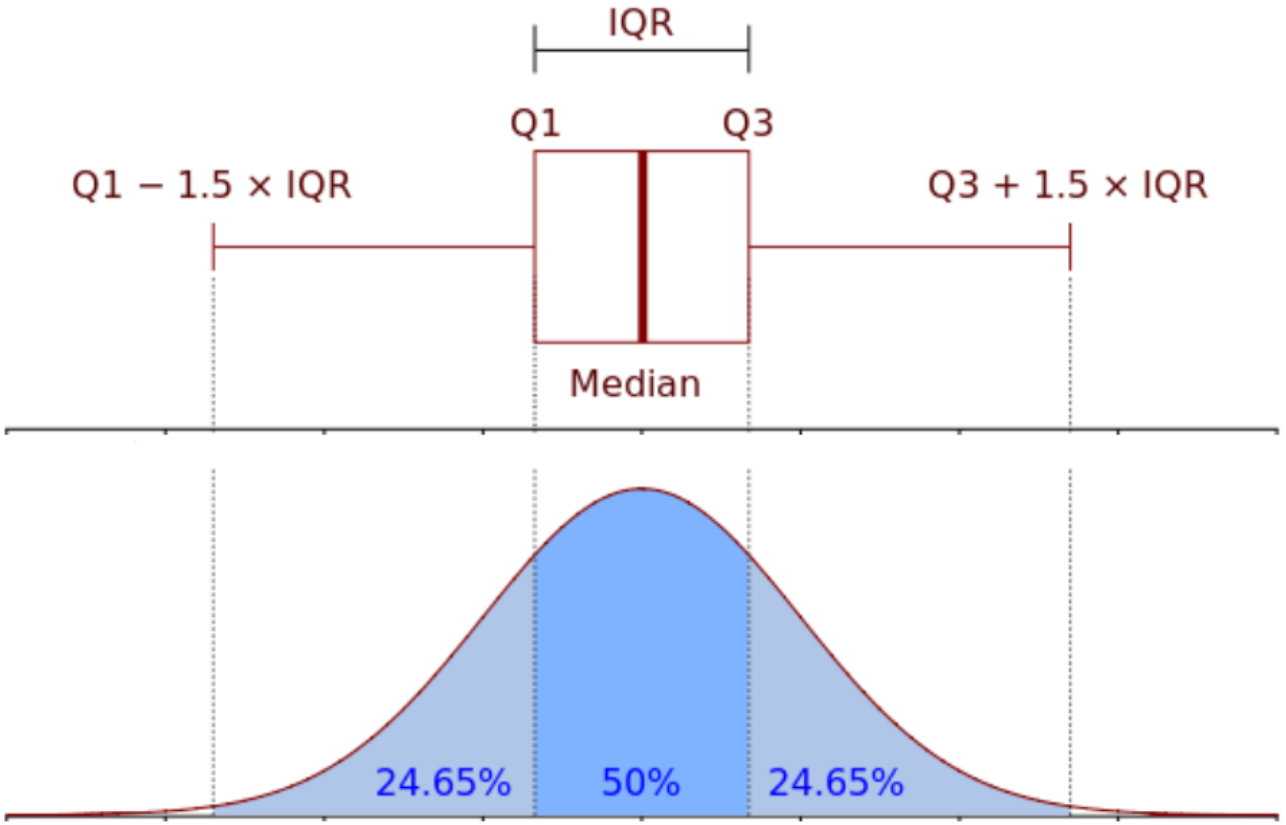
median

IQR = Q3 − Q1 = 25 − 15 = 10



Visualizing Range, Quartile Range, and IQR

Quartile and Interquartile Range



Variance

Average Squared Deviation

Definition
Variance is the mean of squared differences from the mean.

Formula
$\sigma^2 = \Sigma(x_i - \bar{x})^2 / n$

Use Cases
Applied in statistical modeling, ANOVA, and risk measurement.

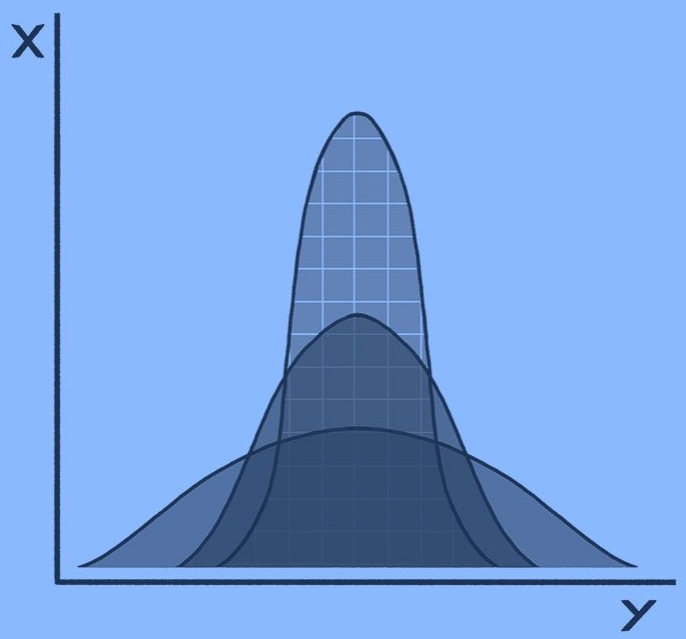
Variance

Formulas for Variance

cuemath
THE MATH EXPERT

	Population	Sample
Ungrouped	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}$
Grouped	$\sigma^2 = \sum_{i=1}^{N} \frac{f(M_i - \bar{X})^2}{N}$	$\sigma^2 = \sum_{i=1}^{N} \frac{f(M_i - \bar{X})^2}{N - 1}$

Variance

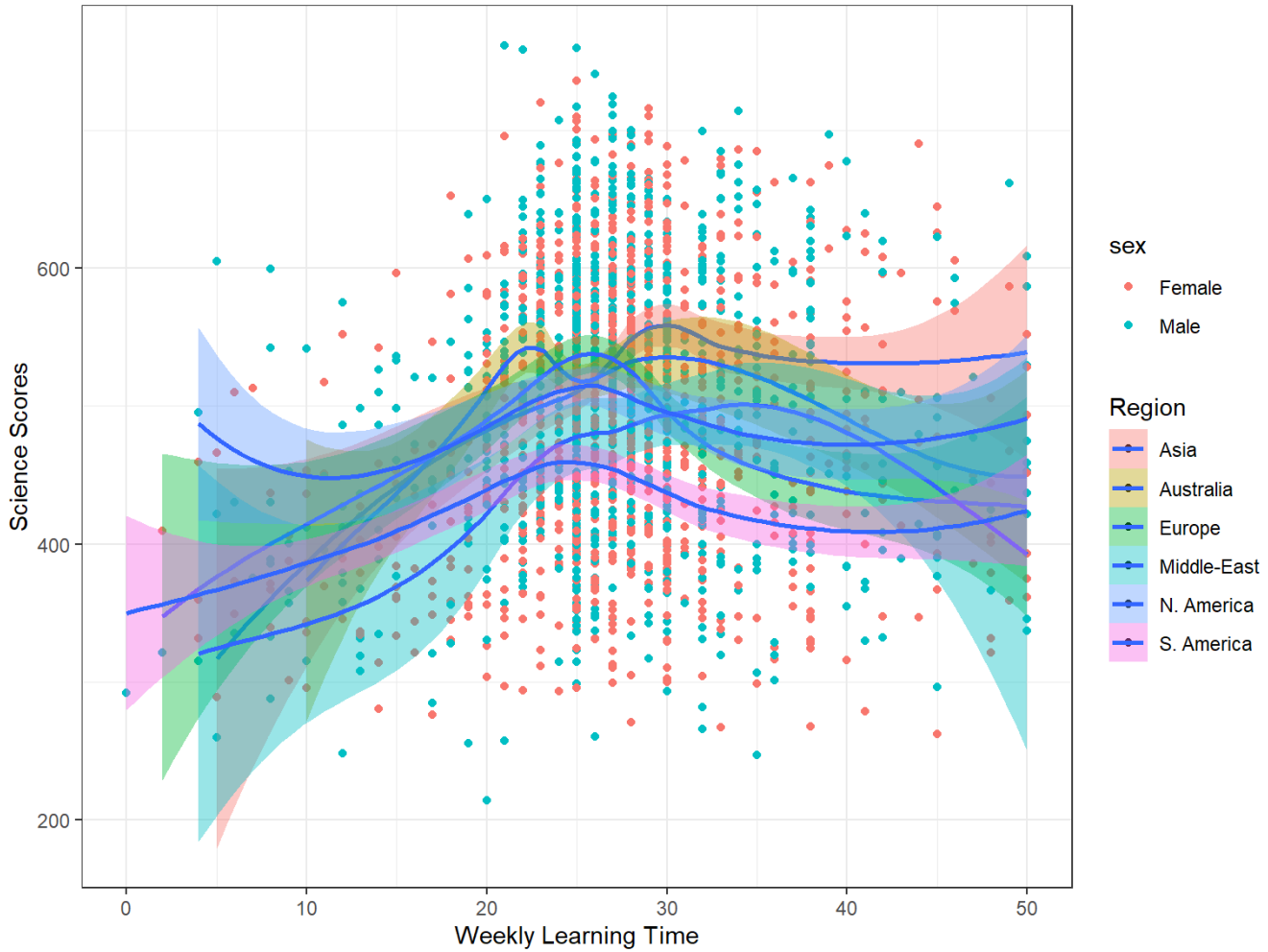


Variance

['ver-ē-ən(t)s]

A measurement of how far each number in a data set is from the mean (average), and thus from every other number in the set.

Variance


Science Scores
600
400
200
0
10
20
30
40
50
Weekly Learning Time
sex
Female
Male
Region
Asia
Australia
Europe
Middle-East
N. America
S. America

Standard Deviation

Square Root of Variance

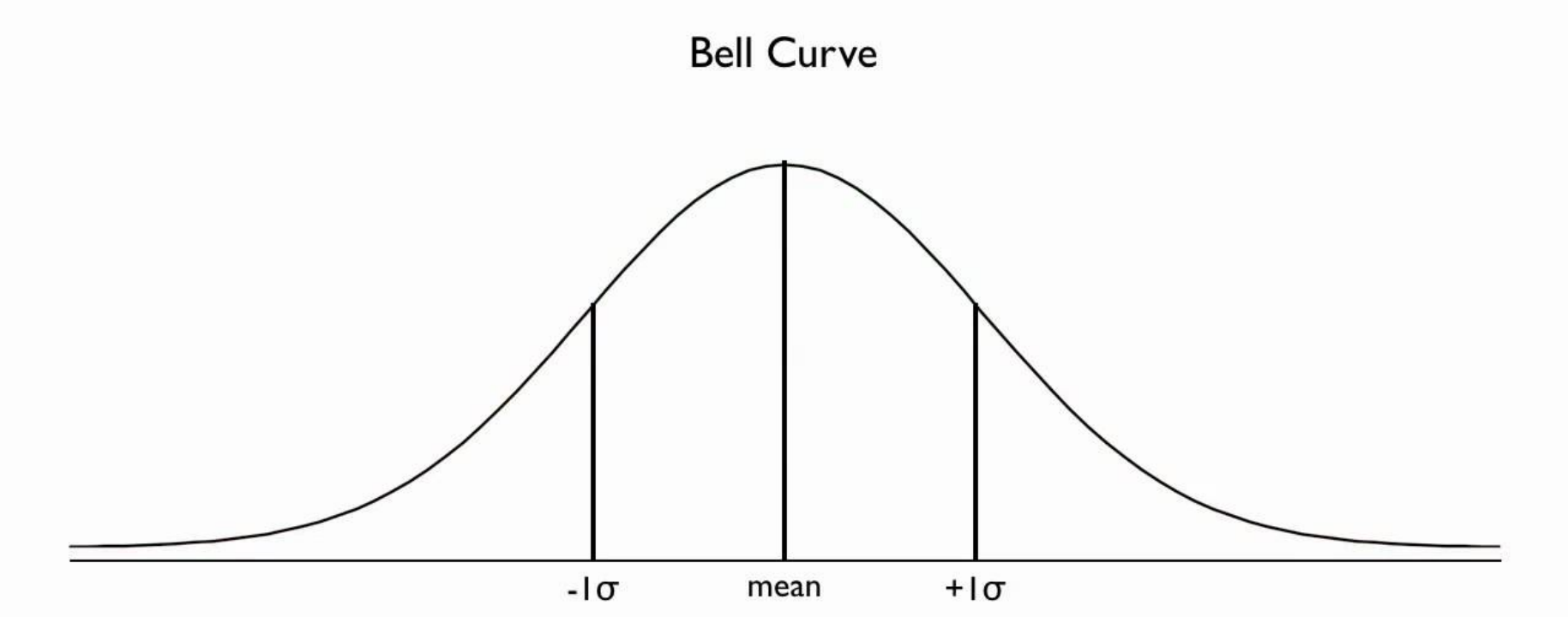
Definition
Square root of variance, in the same units as data.

Interpretation
Represents average distance from the mean.

Applications
Quality control, finance risk assessment, and statistical inference.

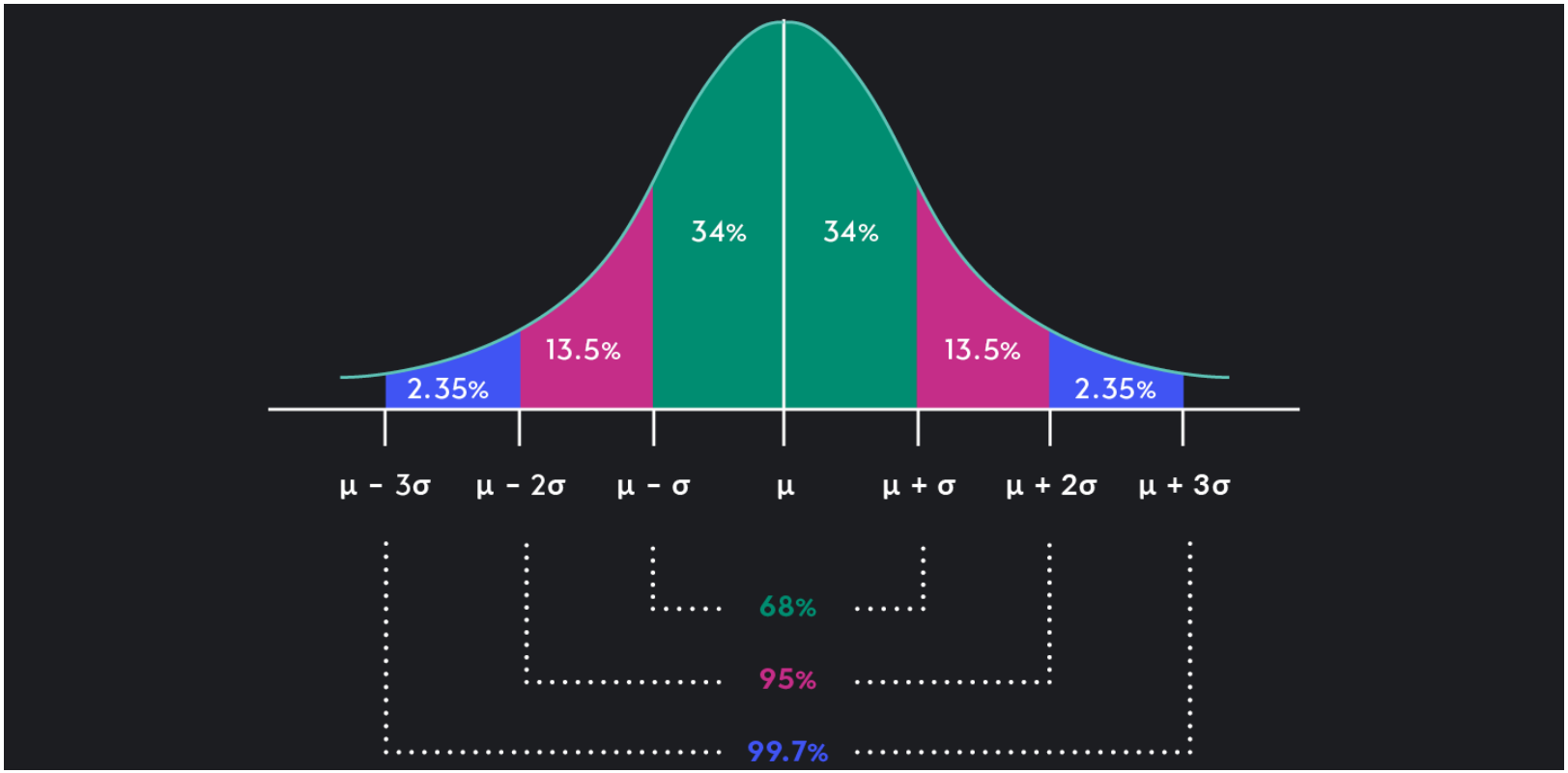
Standard Deviation

Square Root of Variance

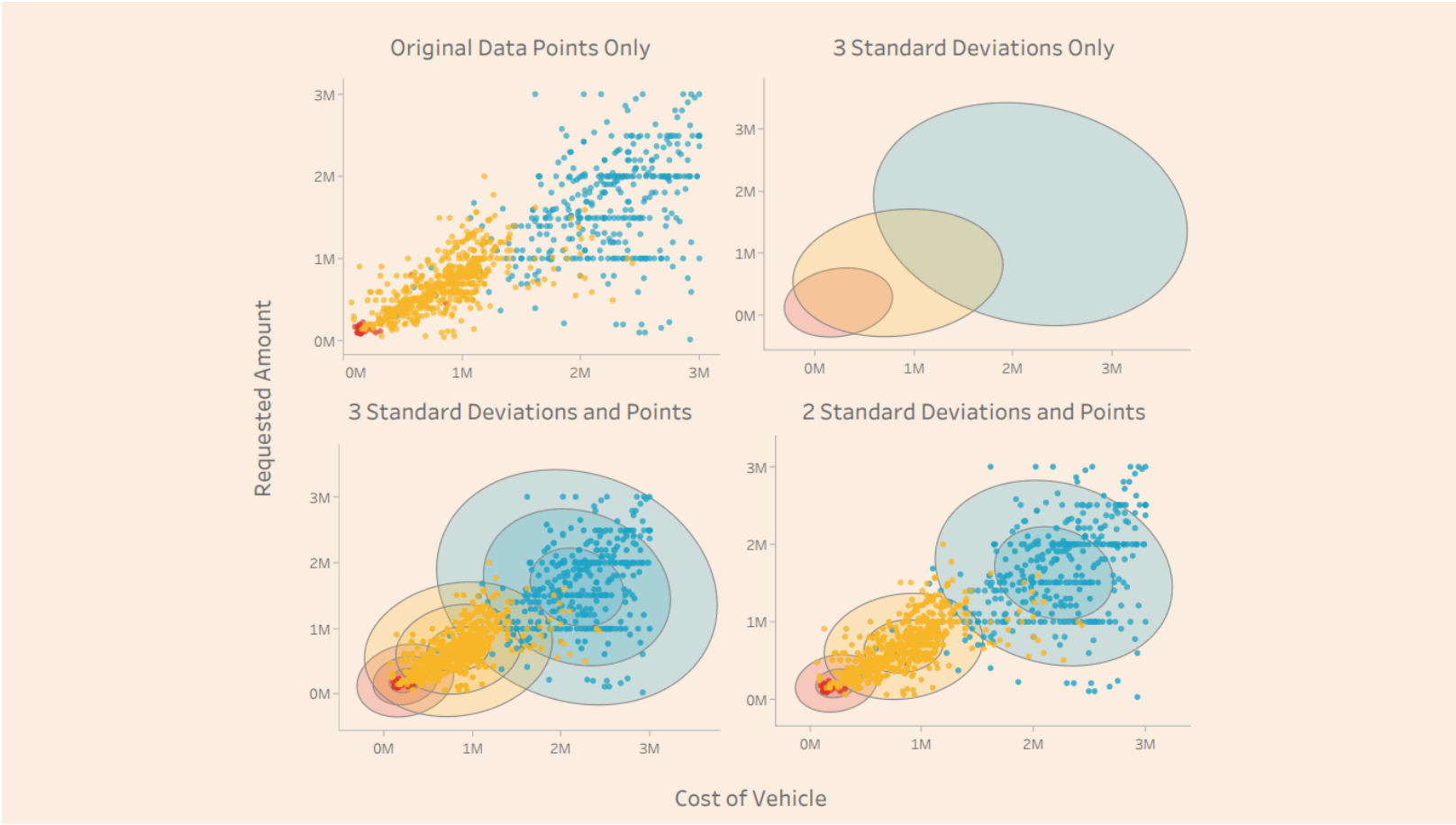


Standard Deviation

Square Root of Variance



Standard Deviation



Practical Examples & Visualization Ideas

Dispersion in Action

- **Box Plots:** Show spread, quartiles, and outliers.
- **Histograms:** Display distribution shape and frequency.
- **Scatter Plots:** Reveal variability patterns between variables.

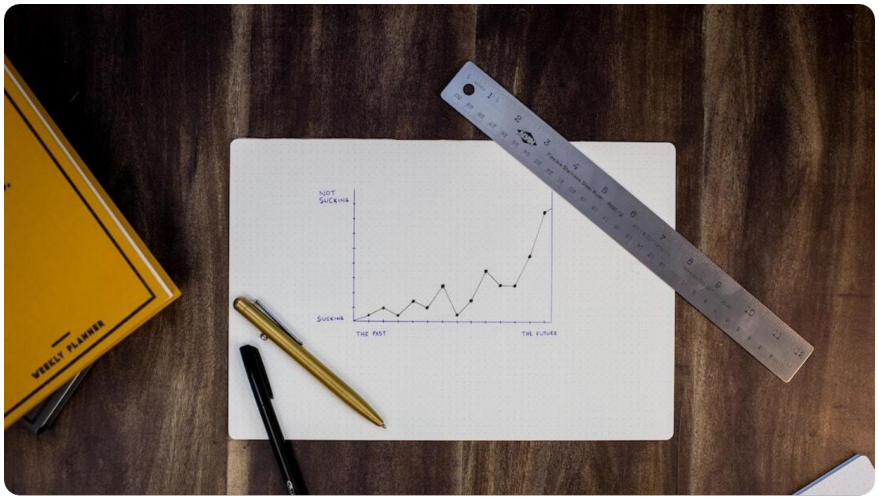


Photo by Isaac Smith on Unsplash

Summary & Real-World Relevance

Complete Statistical Picture

Central Tendency + Dispersion
Together they provide a holistic understanding of data.

Decision-Making
Essential for accurate forecasting, quality control, and risk assessment.

Practical Applications
Used in business, healthcare, manufacturing, and research.

Summary & Real-World Relevance

Central Tendency

- This is a method of calculating the **average** of a set of data
- The average represents the **centre** of the distribution
- These types of statistics are **descriptive** which means they seek to summarise the data

- The three common methods of finding the average are:

1. The **MEAN**
2. The **MEDIAN**
3. The **MODE**

Dispersion

- The MEAN summarises the 'centre' of a distribution, but on its own it may not be informative enough.
- **It is often useful to show how far figures differ from the average. This measure is known as DISPERSION.**

- Methods of showing dispersion:

1. Range
2. The inter-quartile range
3. Standard deviation

In your exam you will only have to calculate the range.

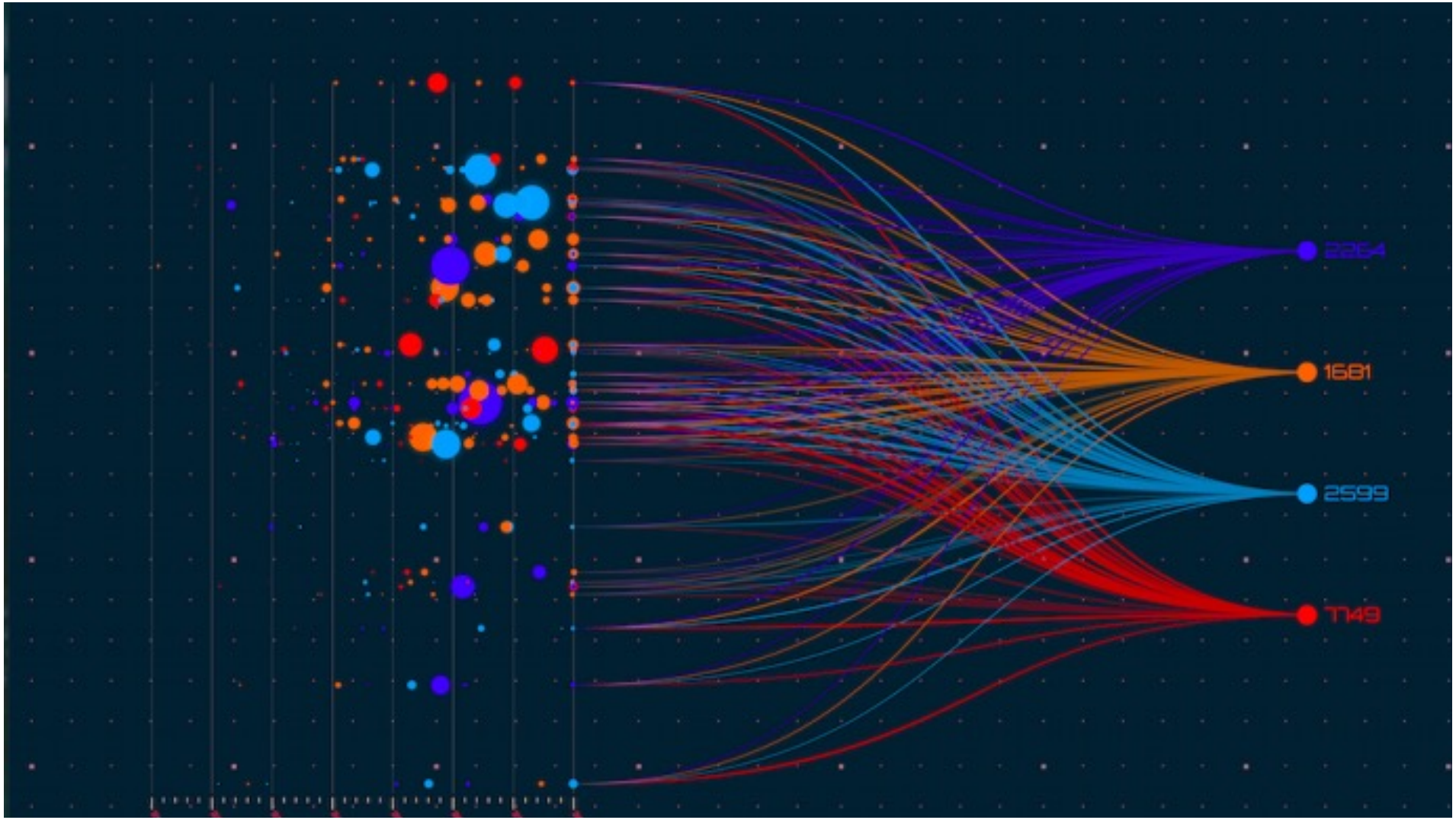
- **Range = the difference between the lowest and highest values in the data set.**

- ***Calculate the range for channel depth in cm: 45, 36, 36, 28, 24, 19, 16, 16, 12, 7, 3, 3, 3, 1.***

- Highest – lowest = range
- 45 - 1 = 44cm

- **Advantages:**
- Easy to calculate
- Shows the spread of data
- When used with the mean it shows the distribution of values around the mean – statistically more useful

- **Limitations:**
- Depends on only two values and ignores the rest. A particular problem if extreme values are atypical.
- The range tends to increase as the sample size increases.


2264
1681
2599
7749