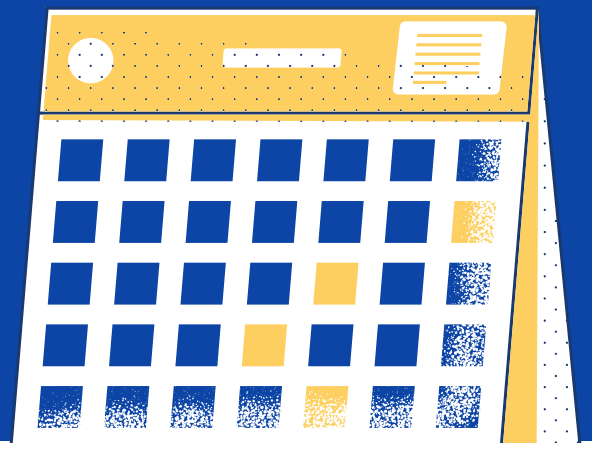
Data Fundamentals

Descriptive Statistics - Part 2

GOALS



Review descriptive Statistics - Part 1

Comprehension

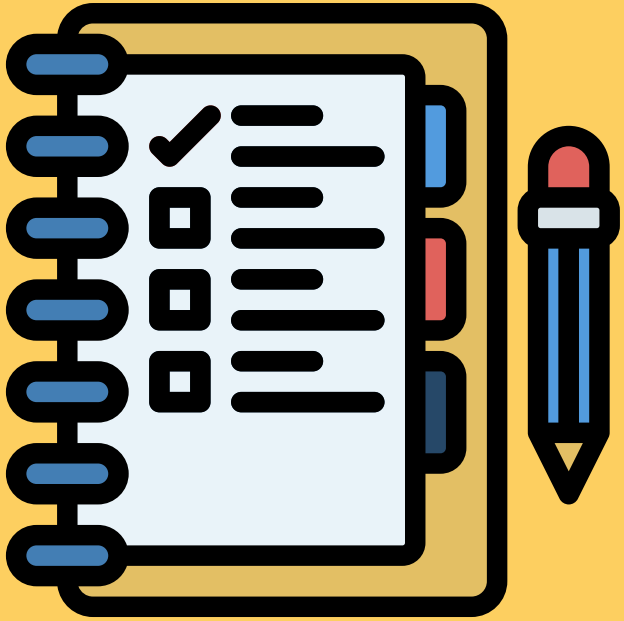
- Histograms
- Five Number Summary
- Box plots

Measures of Spread

The Shape of the data

Outliers

Descriptive vs. Inferential Statistics

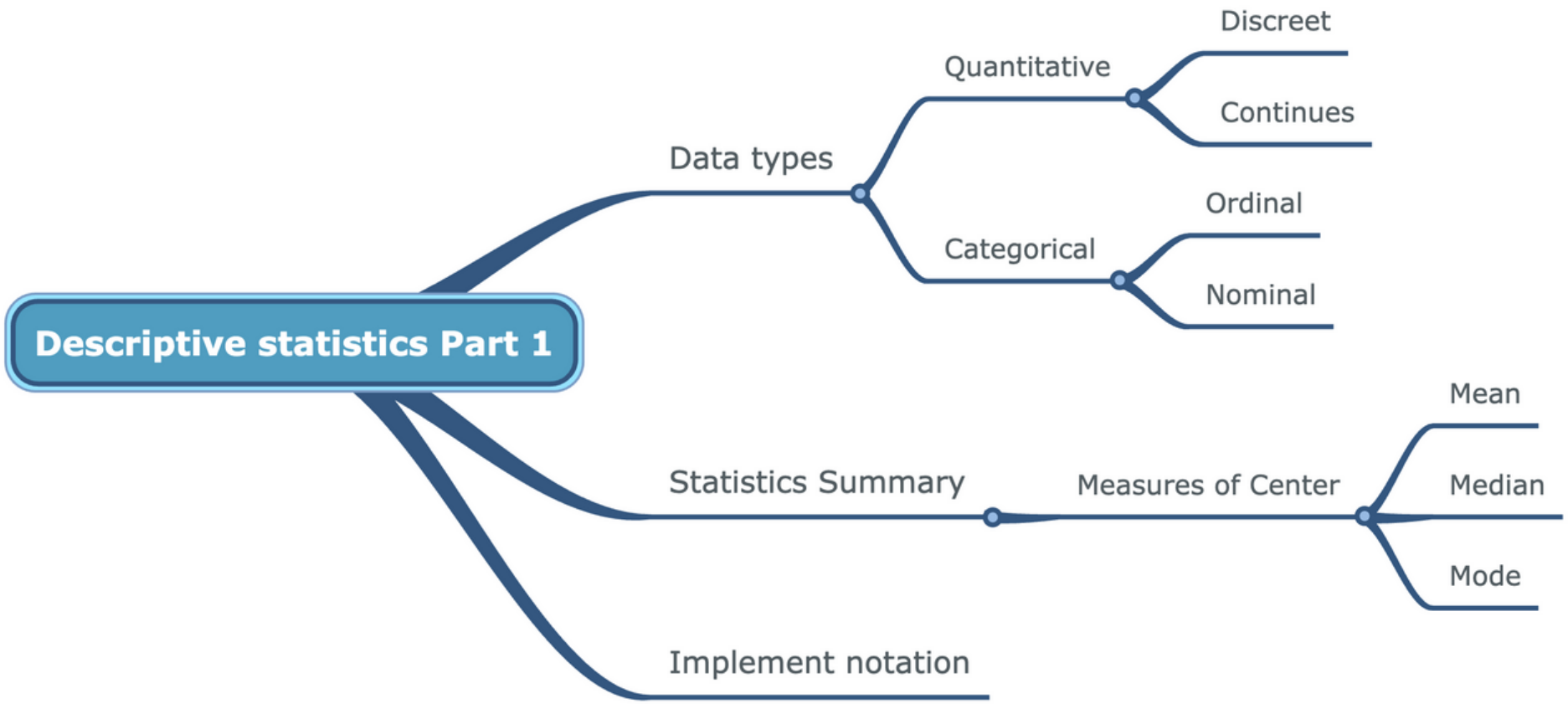


AGENDA

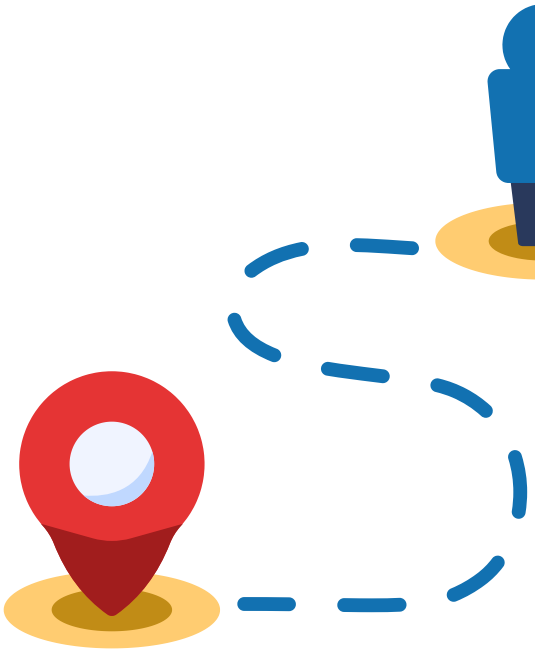
- Welcome
- Review & Roadmap
- Histograms
- Five Number Summary
- Box plots
- Measures of Spread
- The Shape of the data
- Outliers
- Descriptive vs. Inferential Statistics
- Q&A

Behind every data point, there's a story waiting to be told.

REVIEW


Descriptive statistics Part 1
Data types
Quantitative
Discreet
Continues
Categorical
Ordinal
Nominal
Statistics Summary
Measures of Center
Mean
Median
Mode
Implement notation

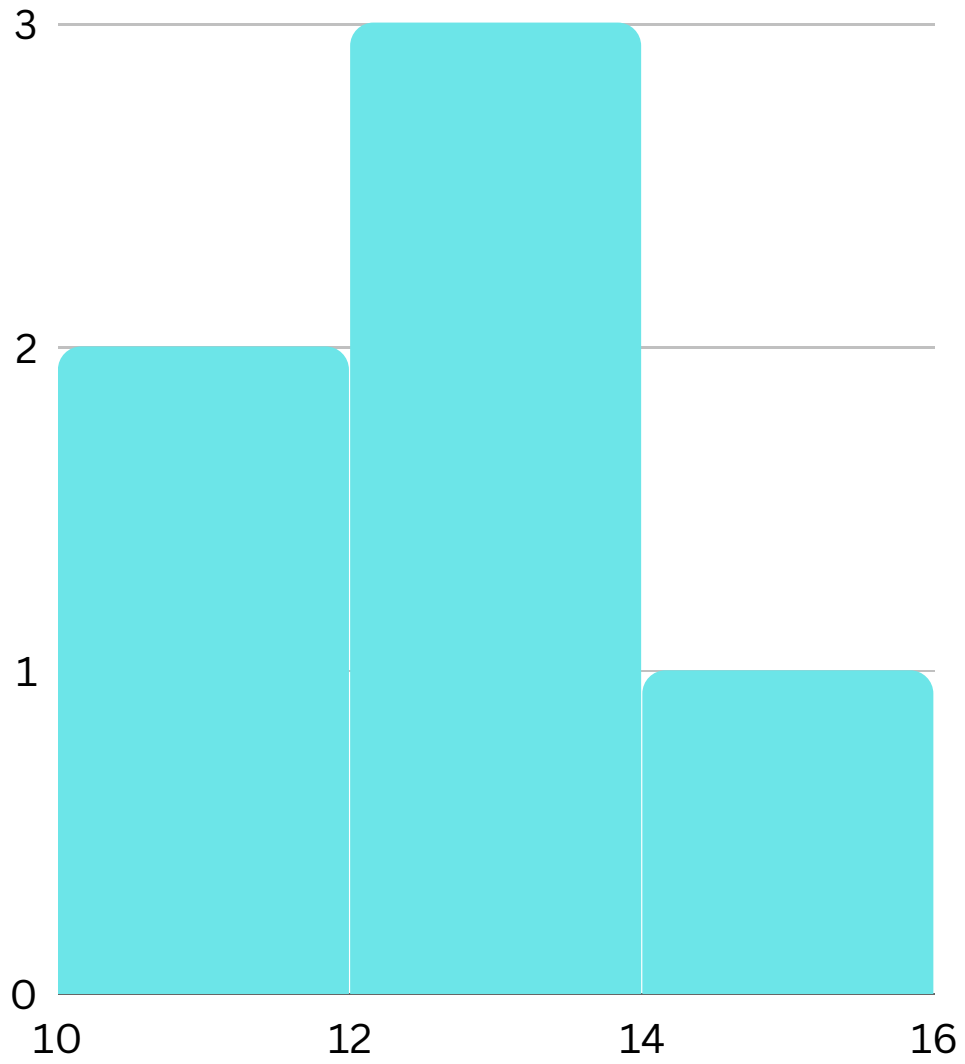
ROADMAP



- **Descriptive statistics**
  - Part1
    - Data types
      - Quantitative
        - Discreet
        - Continues
      - Categorical
        - Ordinal
        - Nominal
    - Statistics Summary
      - Measures of Center
        - Mean
        - Median
        - Mode
    - Implement notation
  - Part2
    - Histograms
    - Five Number Summary
    - Box plots
    - Statistics Summary
      - Measures of Spread
        - Range
        - Interquartile range (IQR)
        - Standard Deviation
        - Variance
    - The Shape of the data
      - Right-skewed
      - Left-skewed
      - Symmetric
    - Outliers
    - Descriptive vs. Inferential Statistics

HISTOGRAM

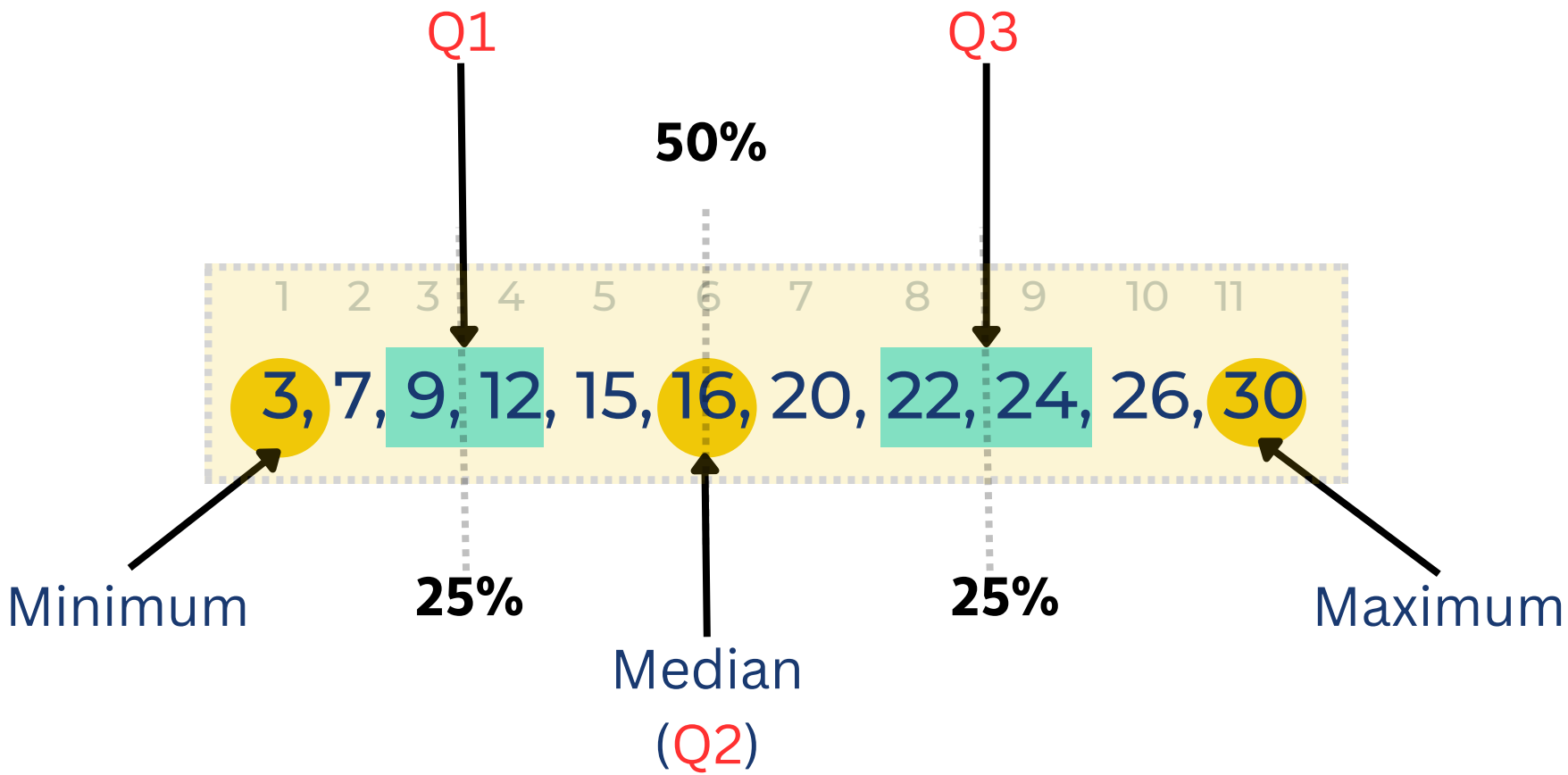
- **Data**: The data are sorted into "bins" or intervals. Each bin represents a specific range of values.
- **Bins**: These are represented by the x-axis and each bin corresponds to a bar's width.
- **Frequency**: The y-axis shows the frequency (number of data points) within each bin. The height of each bar represents this frequency.
- **Shape**: The overall shape of the histogram gives insights into the distribution of the data.



FIVE NUMBER SUMMARY

1. **Minimum:** The smallest number in the dataset.
2. **First Quartile (Q1):** This is the middle number between the minimum and the median. 25% of the data fall below this point.
3. **Median (Q2):** This is the middle number of the dataset. If the dataset has an even number of observations, the median is the average of the two middle numbers.
4. **Third Quartile (Q3):** This is the middle value between the median and the maximum. 75% of the data fall below this point.
5. **Maximum:** The largest number in the dataset.

Five Number Summary



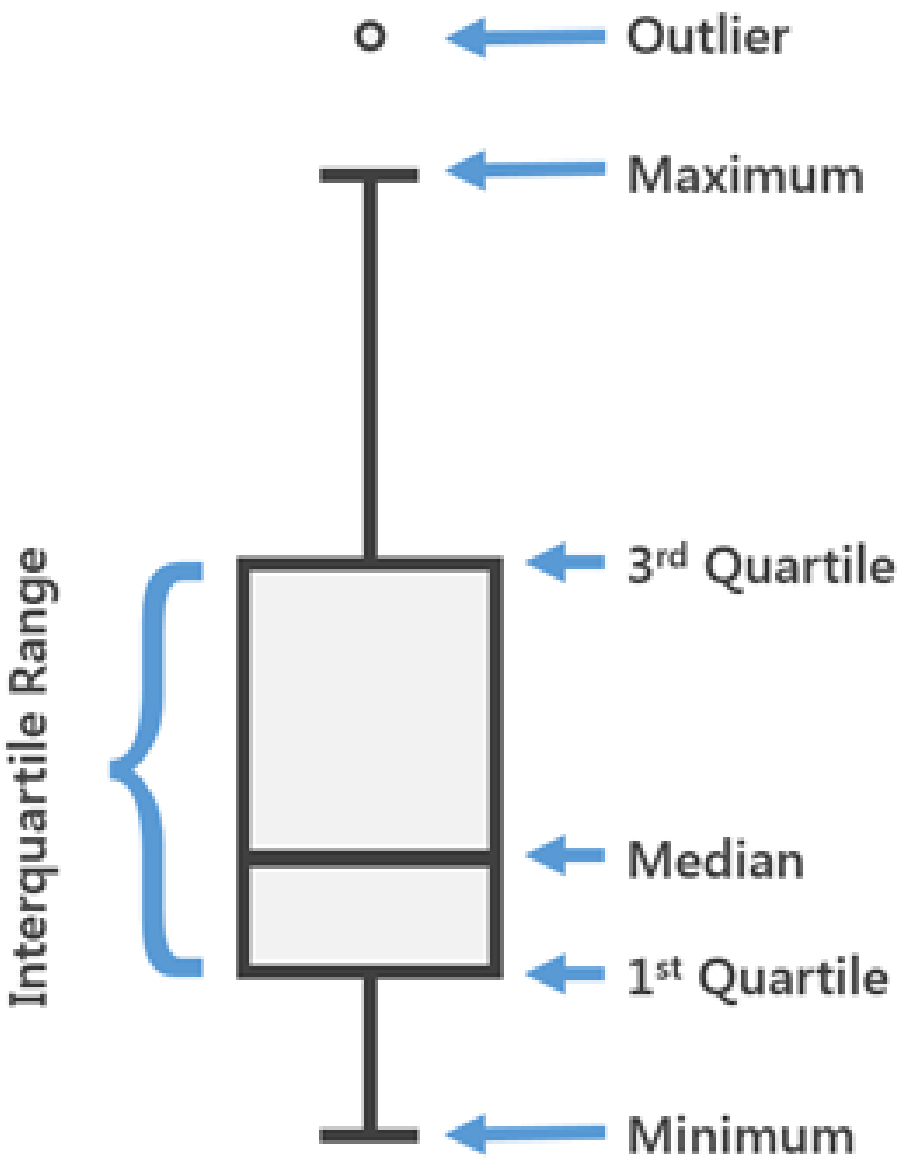
Five Number Summary

SOLUTION

- **Minimum**: The smallest number in the dataset is 3.
- **First Quartile (Q1)**: The 25th percentile of the data. Since the count of numbers (11) is odd and not divisible by 4, we can average the 3rd (9) and 4th (12) smallest numbers to get Q1, which is 10.5.
- **Median (Q2)**: The middle number in a sorted, ordered list of the data. For this dataset, since there are 11 numbers, the median is the 6th number, which is 16.
- **Third Quartile (Q3)**: The 75th percentile of the data. Similarly, we can average the 8th (22) and 9th (24) smallest numbers to get Q3, which is 23.
- **Maximum:** The largest number in the dataset is 30.

BOX PLOTS

1. **Box**: The main body, or box, represents the interquartile range (IQR), which is the range between the first quartile (Q1, 25th percentile) and the third quartile (Q3, 75th percentile). The box's length visually demonstrates the range of the middle half of the data.
2. **Line in the Box**: A line inside the box marks the median (Q2, 50th percentile) — the middle value of the data set.
3. **Whiskers**: Lines extending from the box (known as whiskers) indicate variability outside the upper and lower quartiles, hence they are also part of the data distribution. They typically extend to the minimum and maximum data values, or can represent a certain range around the box (like 1.5 times the IQR) to help identify outliers.
4. **Outliers**: Outliers (if any) are usually plotted as individual points beyond the whiskers.

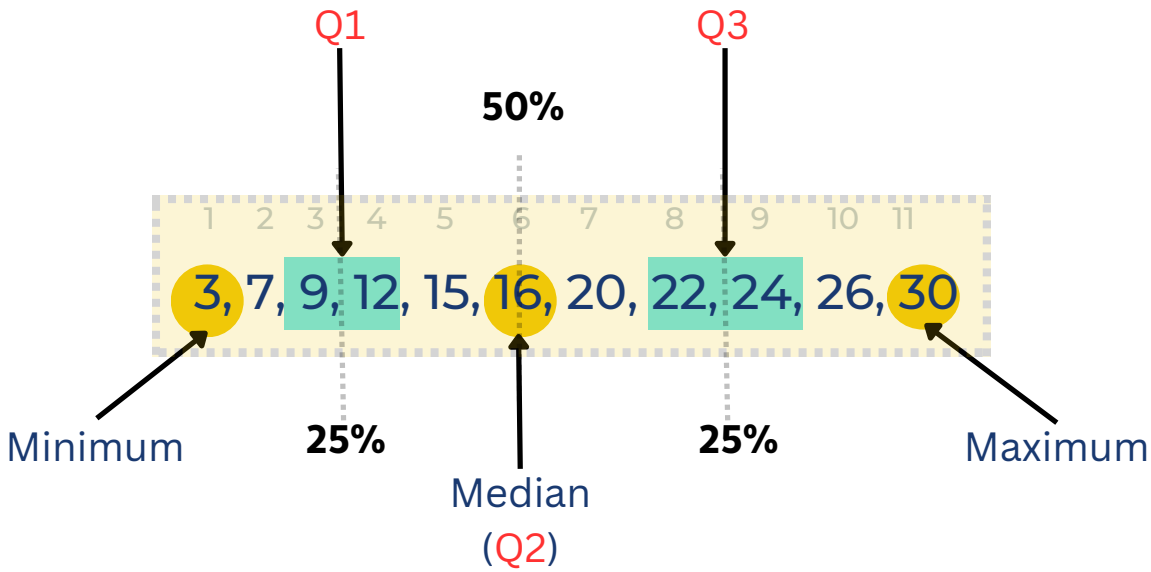


MEASURES OF SPREAD

1.

RANGE

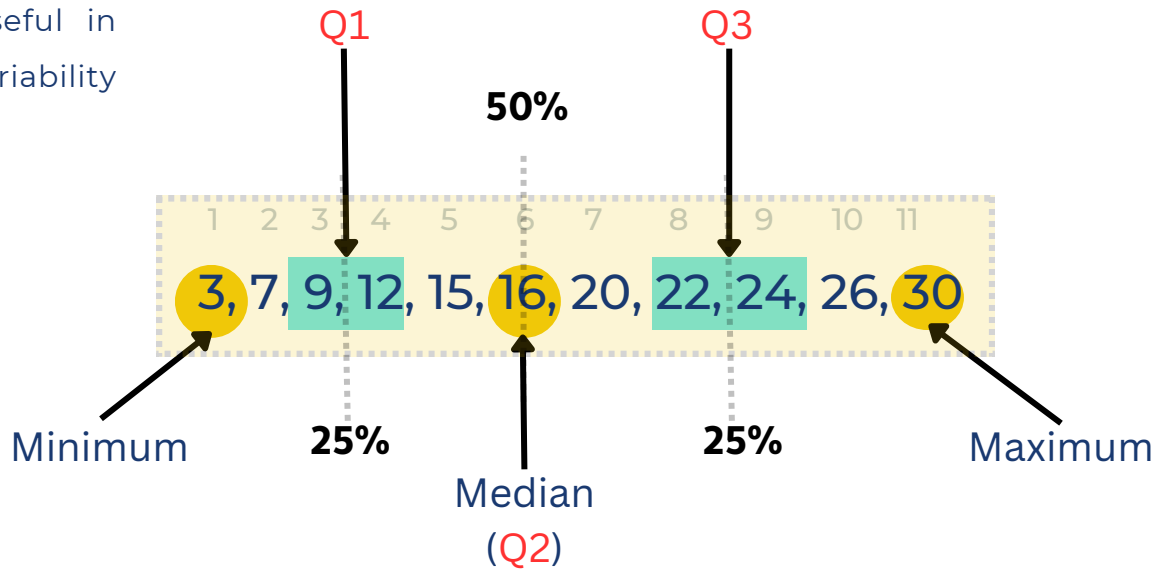
The simplest measure of spread, it's the difference between the largest and smallest values in a dataset.



Range = Max - Min

IQR

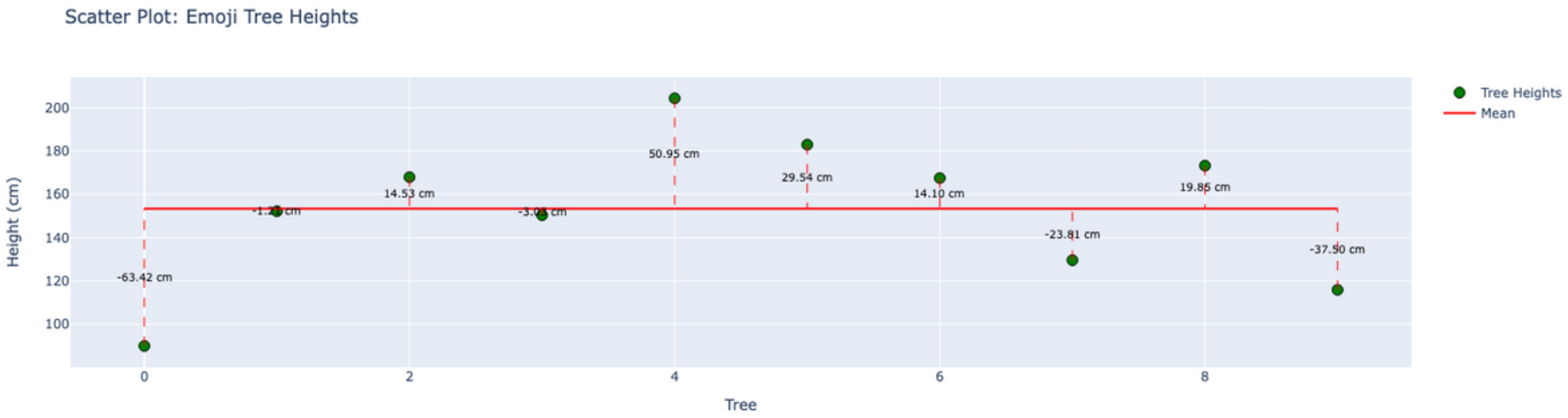
The interquartile range (IQR) is the range where the middle 50% of your data lies. It's used to measure how spread out the data points in a set are from the mean of the data set. The IQR is particularly useful in identifying outliers and understanding the variability of your data.



$$IQR = Q3 - Q1$$

VARIANCE

Variance measures how far each number in the set is from the mean (average) and thus from every other number in the set. It's the average of the squared differences from the mean.

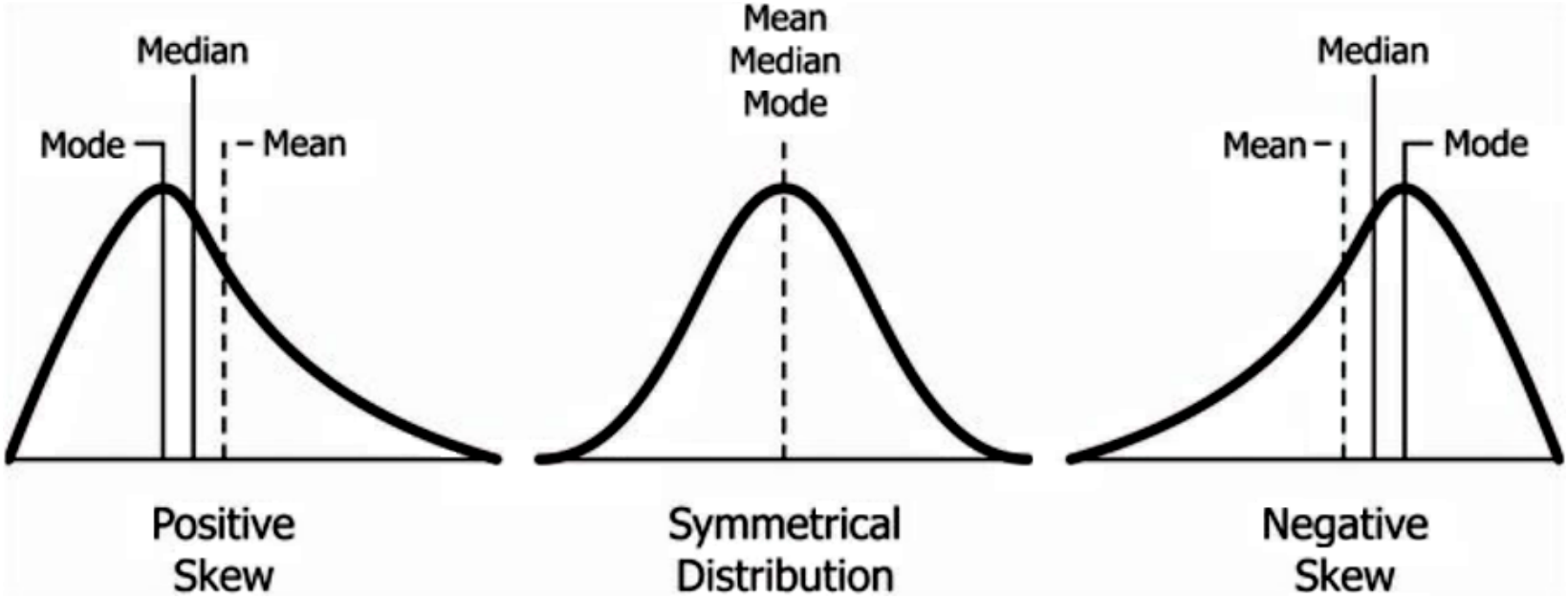


STANDARD DEVIATION

Standard deviation is a measure of how spread out numbers are in a dataset. It shows how much variation or dispersion exists from the average (mean), or expected value. A low standard deviation means that most numbers are close to the average. A high standard deviation means that numbers are more spread out.

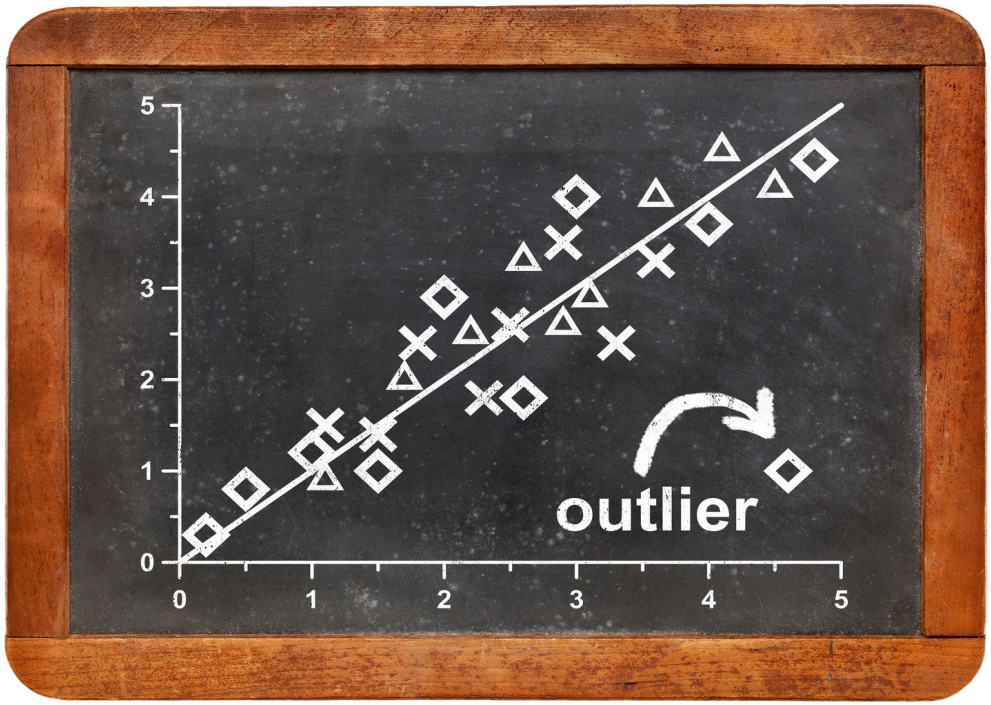
THE SHAPE OF THE DATA

1. **Symmetric:** A symmetric distribution has a shape that looks the same on both sides of the center (like a bell curve). The mean, median, and mode are usually the same in a symmetric distribution.
2. **Skewed Right/Positive Skew:** A distribution is skewed right when the tail on the right side of the distribution is longer or fatter than the left side. In a positively skewed distribution, the mean is usually greater than the median.
3. **Skewed Left/Negative Skew:** A distribution is skewed left when the tail on the left side of the distribution is longer or fatter than the right side. In a negatively skewed distribution, the mean is usually less than the median.



OUTLIERS

Outliers are data points that are significantly different or distant from other observations in a dataset. They can occur naturally due to variability in the data, or they can be caused by errors in data collection, recording, or processing.



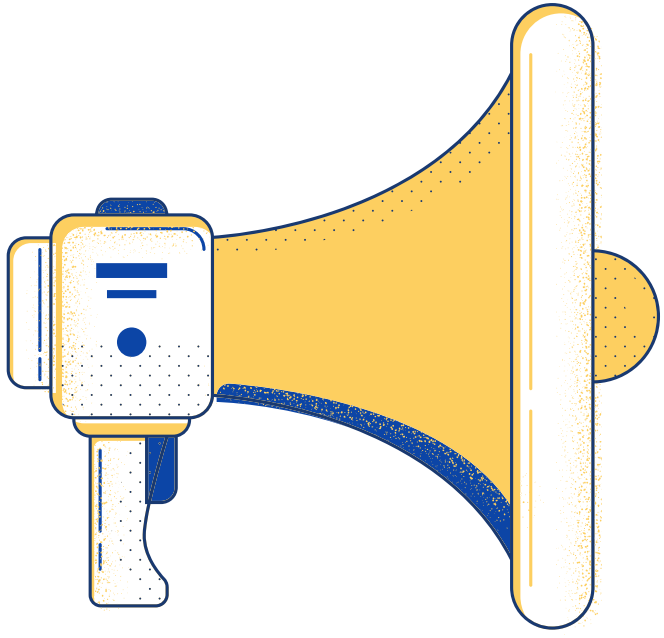
DESCRIPTIVE VS. INFERENTIAL STATISTICS

Descriptive statistics: summarize and organize data to understand its main features, using measures like mean, median, and standard deviation.

Inferential statistics: make predictions or inferences about a larger population from a sample, using methods like hypothesis testing and regression analysis.

NOTATION

Mean ($\bar{x}$)	$\bar{x} = \frac{\sum x}{n}$
Median (M)	If n is odd, then $M = \left(\frac{n+1}{2}\right)^{th}$ term If n is even, then $M = \frac{\left(\frac{n}{2}\right)^{th} \text{ term} + \left(\frac{n}{2}+1\right)^{th} \text{ term}}{2}$
Mode	The value which occurs most frequently
Variance (σ^2)	$\sigma^2 = \frac{\sum(x-\bar{x})^2}{n}$
Standarad Deviation (S)	$S = \sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$



Q&A Session: Let's explore and understand together

RESOURCES

- Measures of spread
- Descriptive statistic roadmap
- Mean, Median, and Mode of Apples
- StatKey
- Maths is fun
- Petrol prices

Your presence today has added value to our shared learning journey. Thank you for joining us!