Descriptive Statistics

From raw data to quick insights

“In today’s lesson, we will be discussing the fundamentals of Descriptive Statistics"

Lesson Objectives

By the end of this lesson, you will be able to:

1. Interpret different types of data visualizations
2. Describe the shape of a distribution
3. Calculate measures of center and spread
4. List the possible ways of identifying outliers

Agenda

Today we will:

- Motivation
- Data Visualizations
- Shapes & Distributions
- Measures of Center and Spread
- Outliers

Motivation

Probability vs. Statistics

Source : A Brief Intro to Experiments, Samples, and Statistical Inference

- **Probability:**
  You have a fair coin, and flip it 10 times.
  What's the probability of 3 consecutive heads?

- **Statistics**:
  You flip a coin 100 times and get heads 70 times.
  What conclusions could you make about the fairness of the coin?

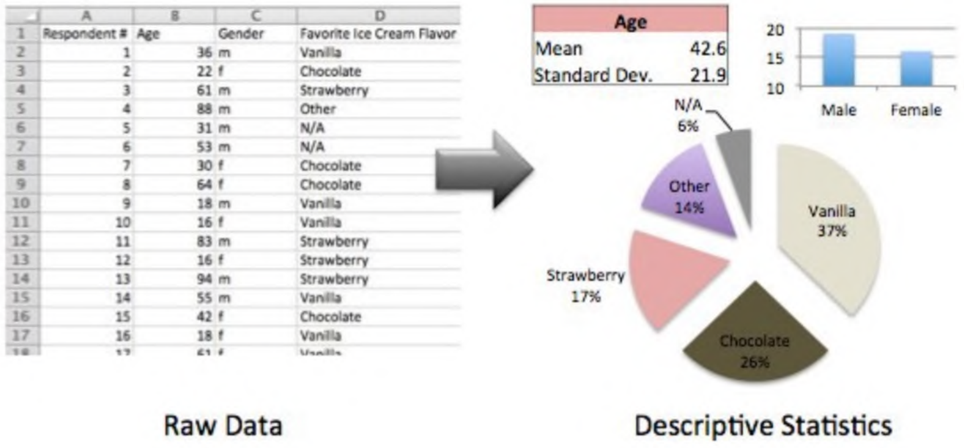
Descriptive vs. Inferential Statistics

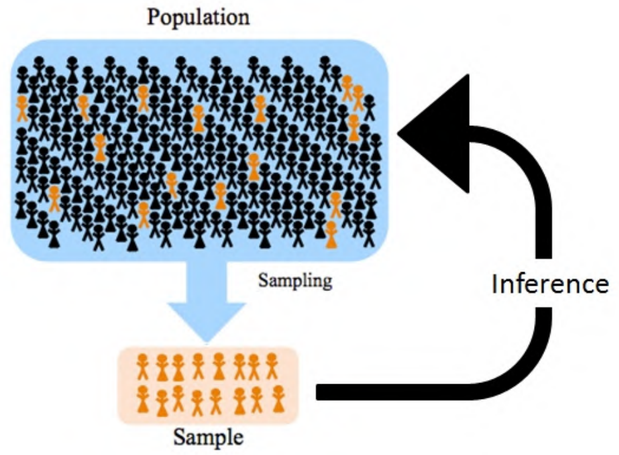
Descriptive Statistics: Summarize our collected data in an accurate way using charts, tables, and graphs

- Summary statistics: mean, median, mode, standard deviation, and variance

Inferential Statistics: Use our collected data (a sample) to make conclusions about a larger population

- Common methodologies: confidence intervals, hypothesis tests, analysis of variance, and regression

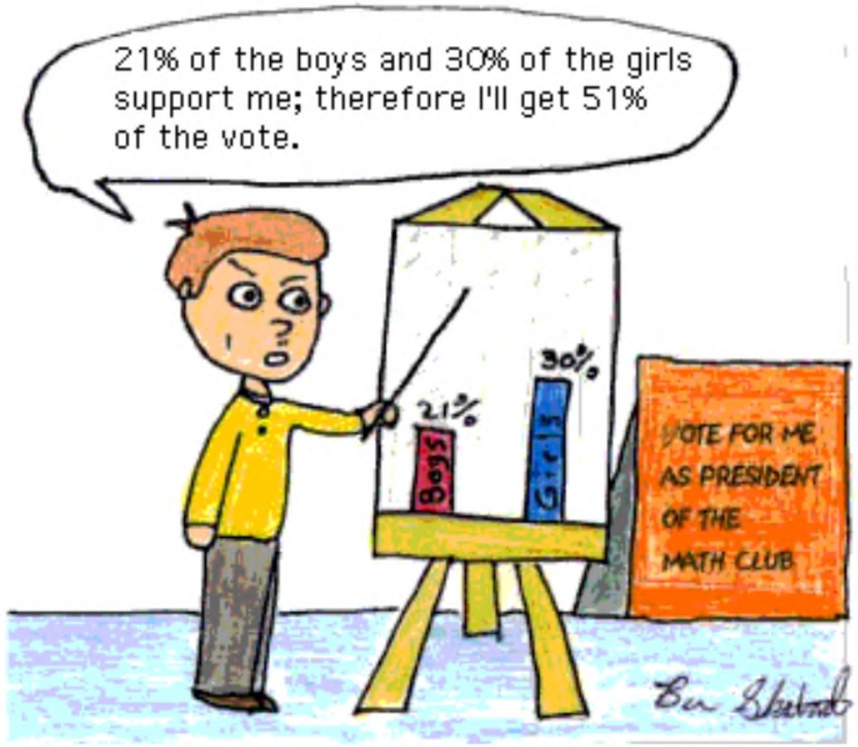




Motivation

Descriptive Statistics is a method to:

- Collect,
- Organize,
- Summarize,
- Display,
- and Analyze sample data taken from a population



Data Visualizations

Data Visualizations

- **What?** The practice of translating information into a visual context
- **Why?** To make it easier to identify patterns, trends and outliers in large data sets.

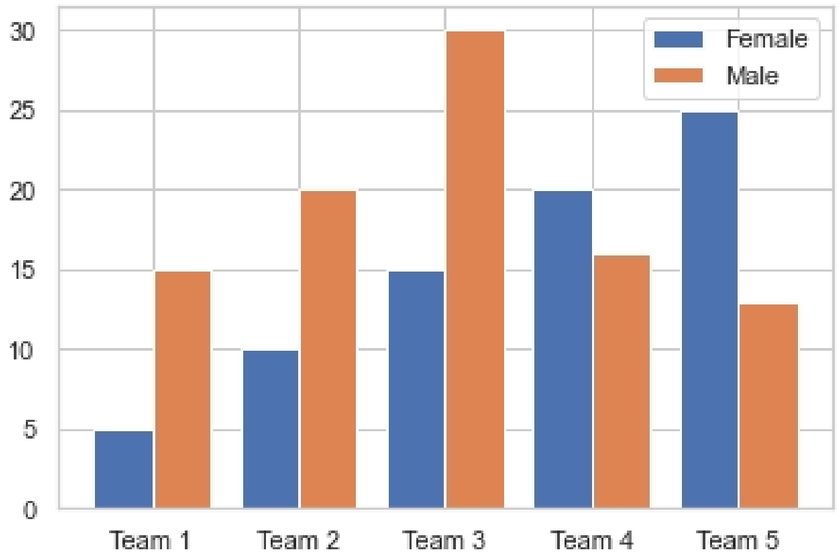
Data Visualizations

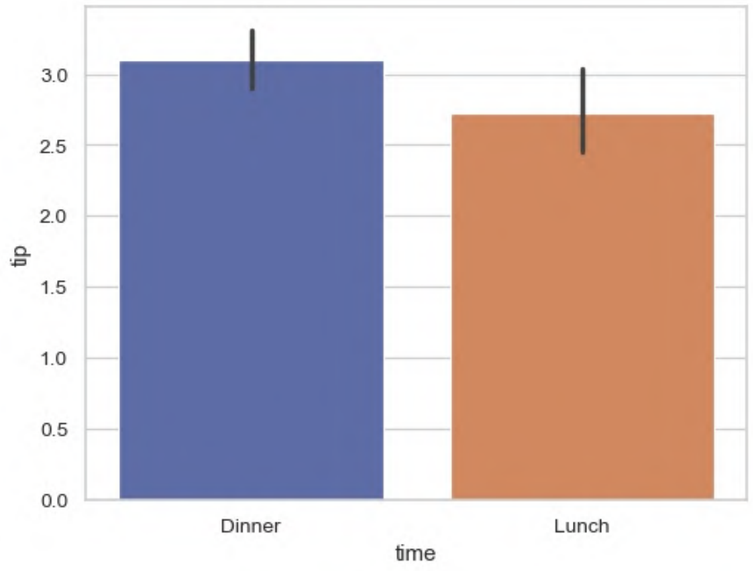
Can also be referred to as "charts", "plots", or "graphs"

- Bar Plot
- Line Plot
- Scatterplot
- Histogram
- Boxplot
- Many more...

Bar Plots

The classic bar plot uses either horizontal or vertical bars (column chart) to show discrete, numerical comparisons across categories (categorical data).

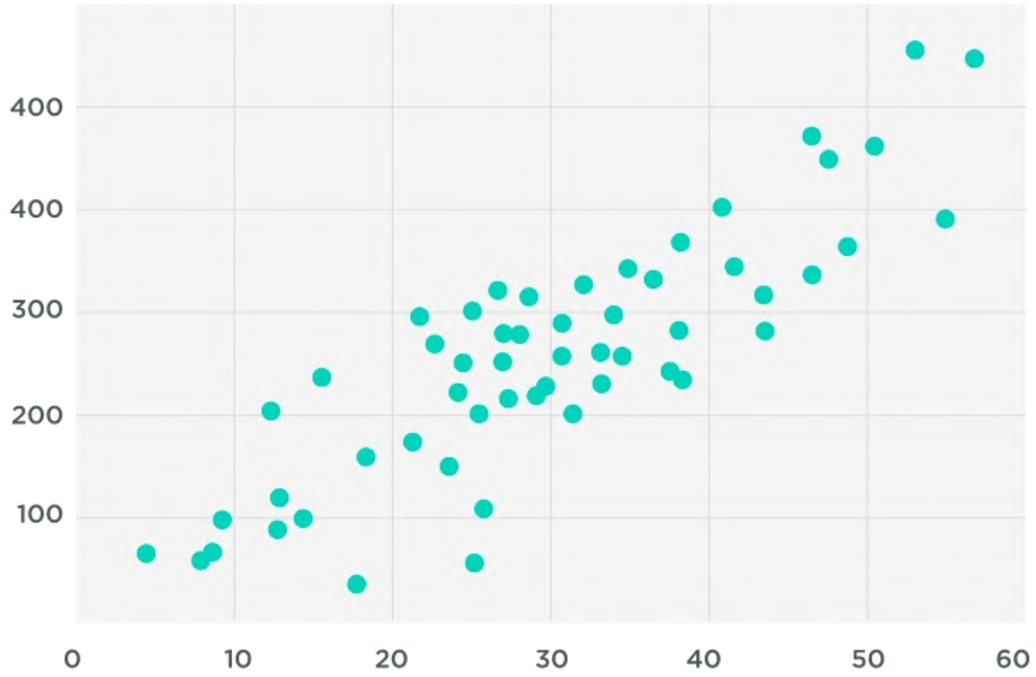




Scatterplots

Scatterplots display two quantitive variables against each other.

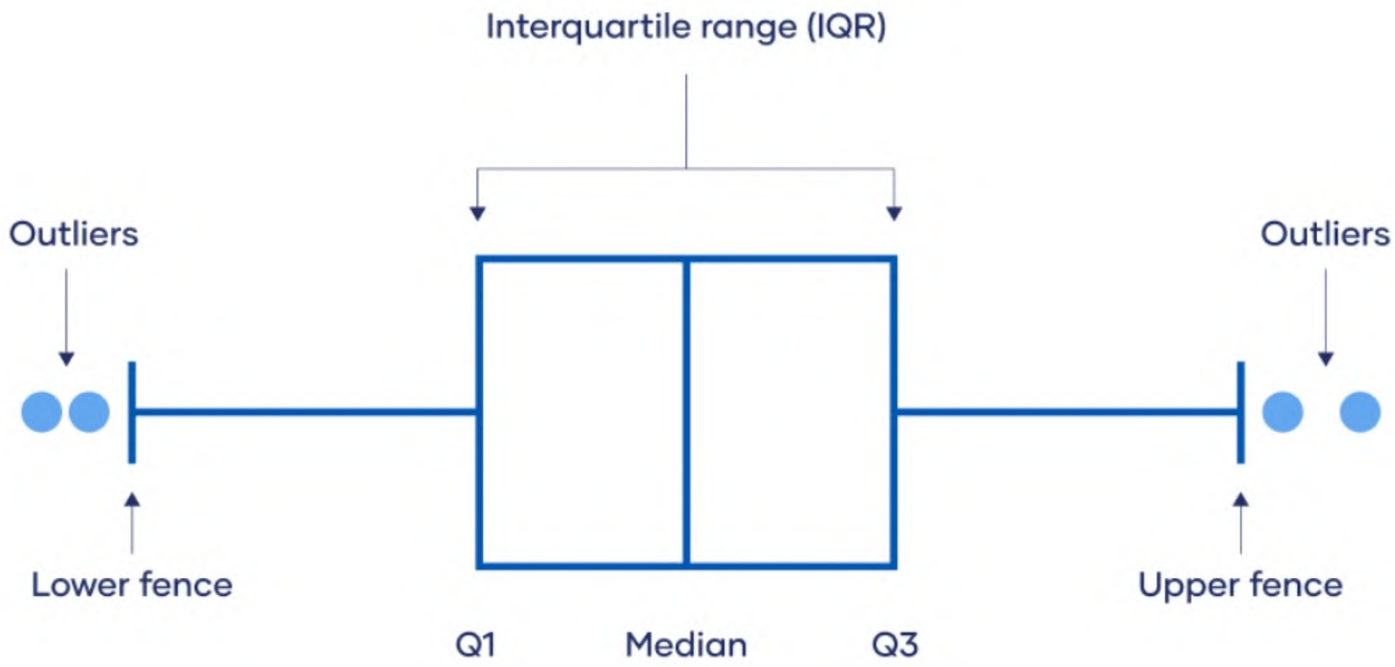
A scatterplot is most frequently used to show relationships between variables.



BoxPlots

A boxplot visualizes the distribution of numeric data using the 5-number summary.

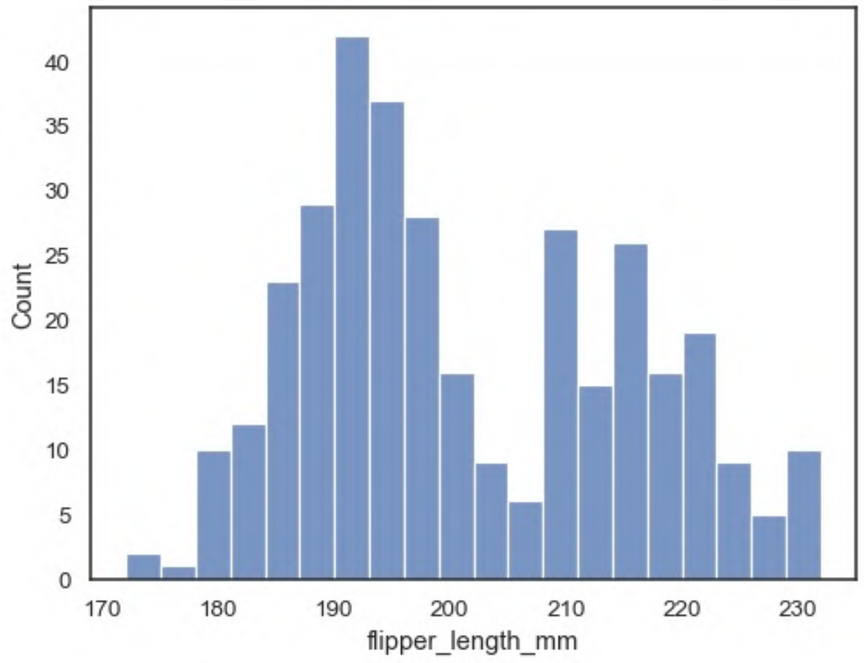
They are useful for giving a rough view of the probability distribution in a compact form.

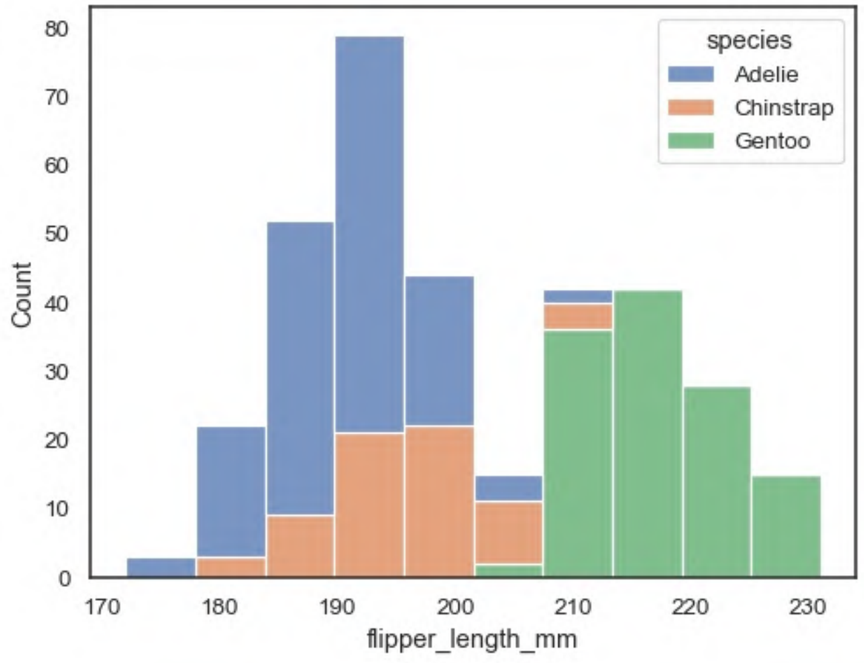


Histograms

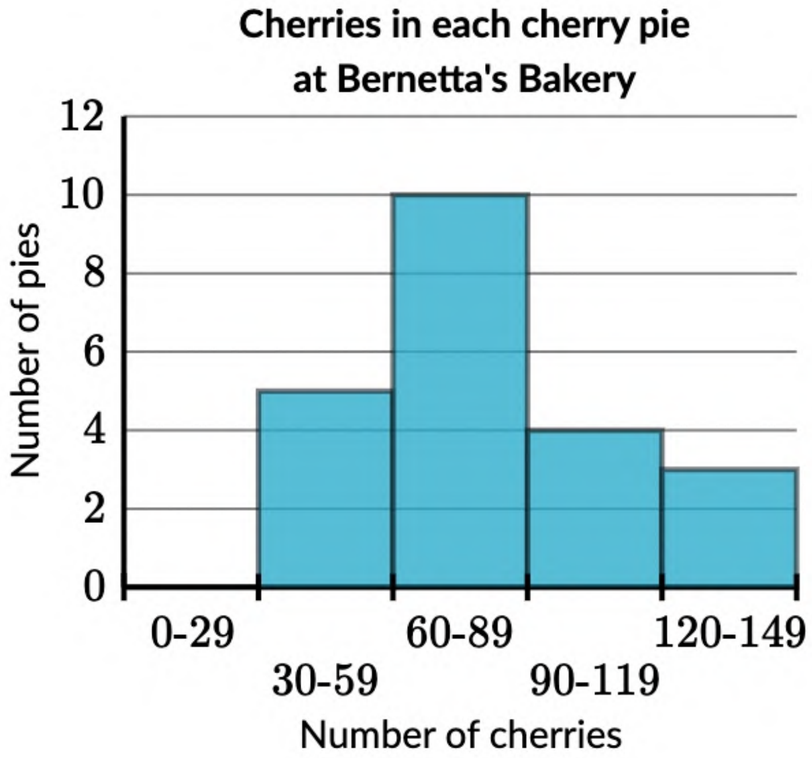
A Histogram visualises the distribution of numeric data by binning the values on the x-axis and showing the frequency on the vertical axis.

They are useful for giving a view of the probability distribution.





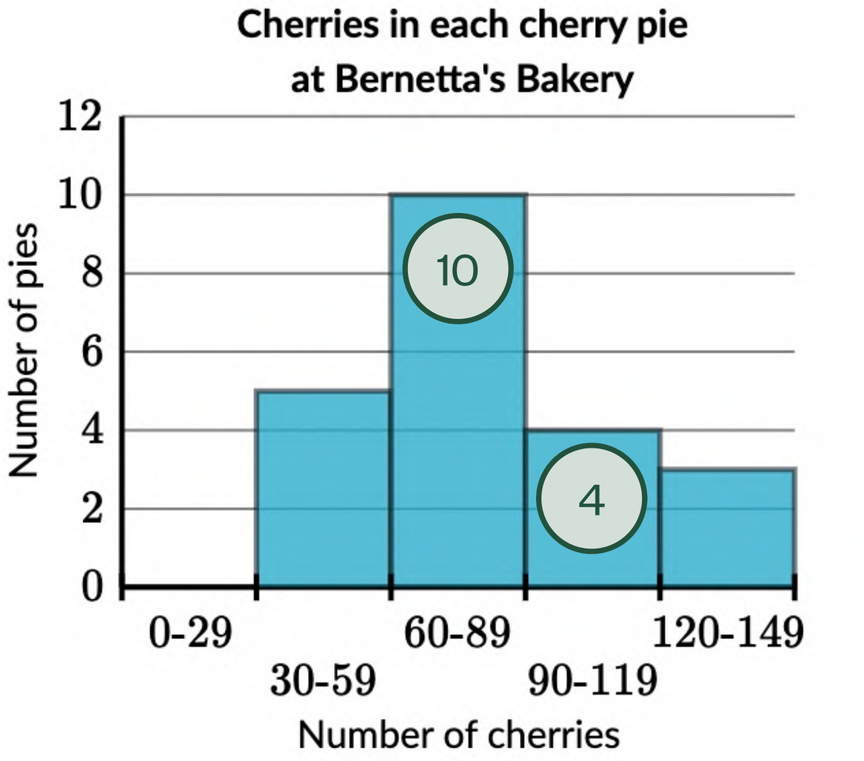
Learning Check



How many more cherry pies have 60 to 89 cherries than 90 to 119 cherries?

slido

Learning Check



How many more cherry pies have 60 to 89 cherries than 90 to 119 cherries?

90-119 cherries: 4 pies
60-89 cherries: 10 pies
10 - 4 = 6

Data Distributions

Distribution of Numeric Data

***Source**: Cuemath*

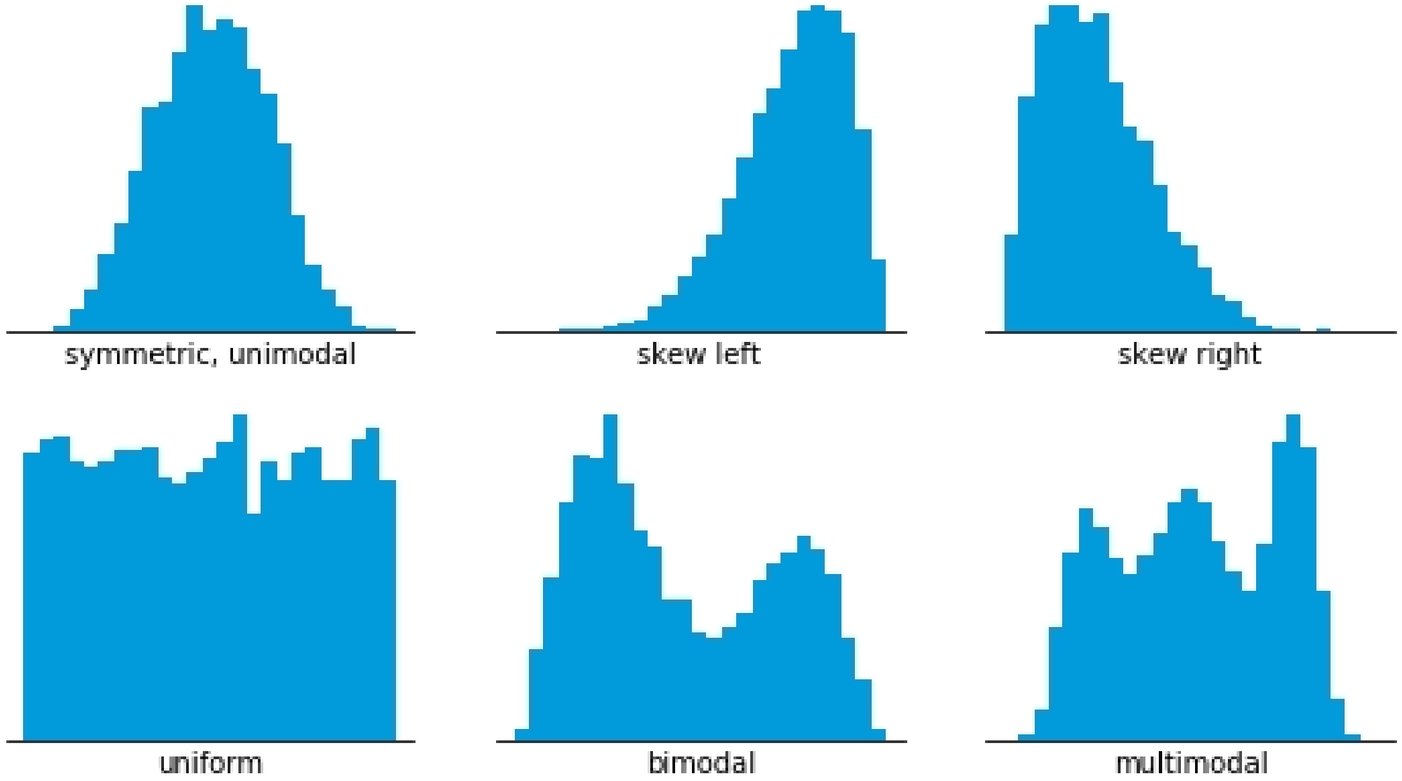
- The **distribution** of numeric data is represented graphically as:
  - Data values on the x-axis
  - Frequency on the y-axis

- **Example:** How many books did you read last summer?
- **Dot plot:** all individuals in the sample are a single dot
- We can observe the **shape** of the distribution

Distribution Shapes

- **Histogram:** data values are grouped into bins with frequency

- Characteristics of distribution shape:
  - **Symmetry:** whether you can divide the distribution into two mirrored halfs (symmetrical vs. skewed)
  - **Modality:** the number of peaks it contains (unimodal, bimodal, multimodal)



Measures of Center and Spread

Summary Statistics

Used to summarize a set of observations in order to communicate the largest amount of information as simply as possible.

1. **Measures of Central Tendency**:
   mean, median, mode

2. **Measures of Spread**:
   standard deviation, range, IQR

Measure of Central Tendency

- Different ways to summarize the middle or typical value of a dataset.

- **Mean/Average:** Computed by taking the sum of all the values in the dataset divided by the total number of values.

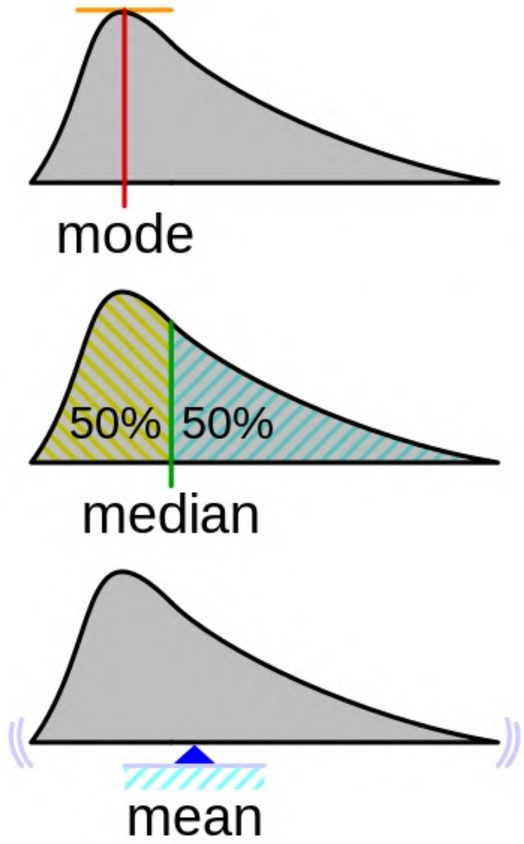
  (1 + 2 + 2 + 4 + 7 + 10) / 6 = 4.33

- **Median:** The value that lies in the middle of the dataset when arranged in ascending or descending order. If there are an even number of values, average the middle two.

  (2 + 4) / 2 = 3

- **Mode:** The most occurring value in the dataset.

  2



Source: *Wikipedia*

Measure of Central Tendency

- Different ways to summarize the middle or typical value of a dataset.

- **Mean/Average:** Computed by taking the sum of all the values in the dataset divided by the total number of values.

  (1 + 2 + 2 + 4 + 7 + 10) / 6 = 4.33

- **Median:** The value that lies in the middle of the dataset when arranged in ascending or descending order. If there are an even number of values, average the middle two.

  (2 + 4) / 2 = 3

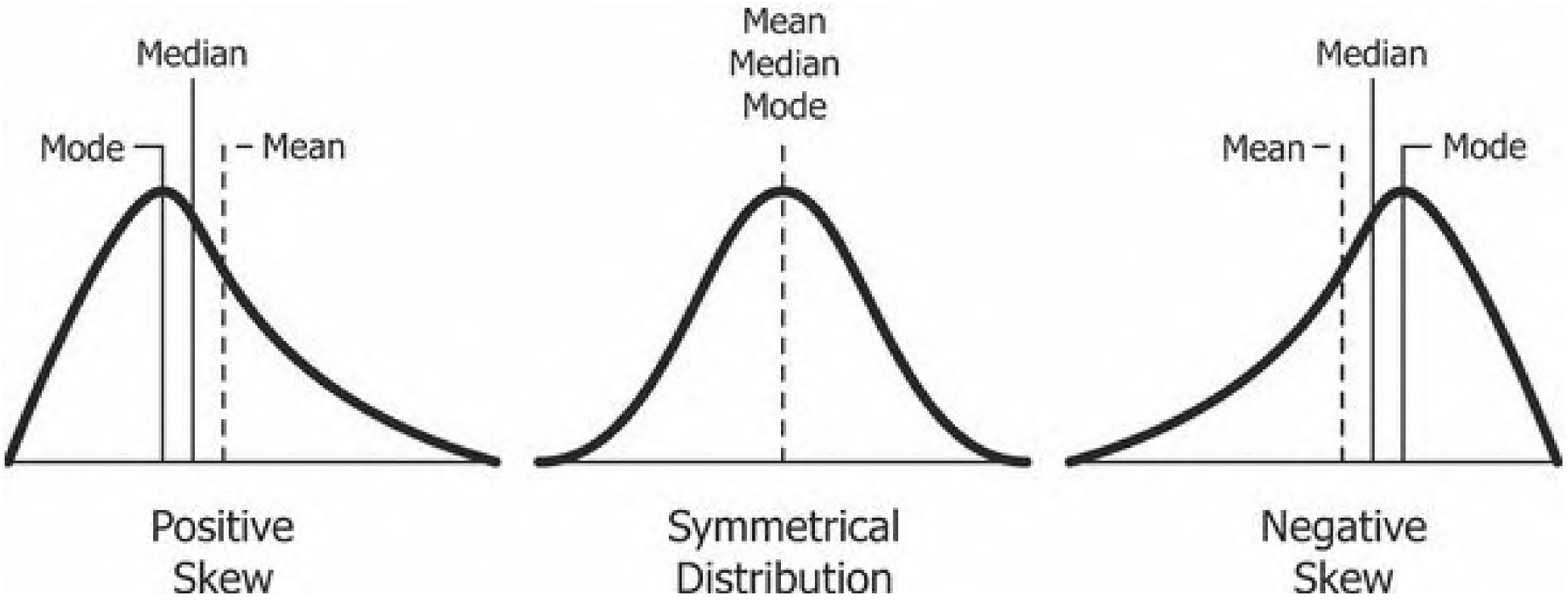
- **Mode:** The most occurring value in the dataset.

  2

A dataset of 6 values

1, 2, 2, 4, 7, 10

Measure of Central Tendency



Source: Skew and Kurtosis: 2 Important Statistics terms you need to know in Data Science

slido

Learning Check

Calculating the mean

You have found the following ages (in years) of all 4 lions at your local zoo:

[5, 4, 6, 39]

What is the average age?

1. 6 years old
2. 11.5 years old
3. 13.5 years old
4. 10 years old

slido

Learning Check

Calculating the mean

You have found the following ages (in years) of all 4 lions at your local zoo:

[5, 4, 6, 39]

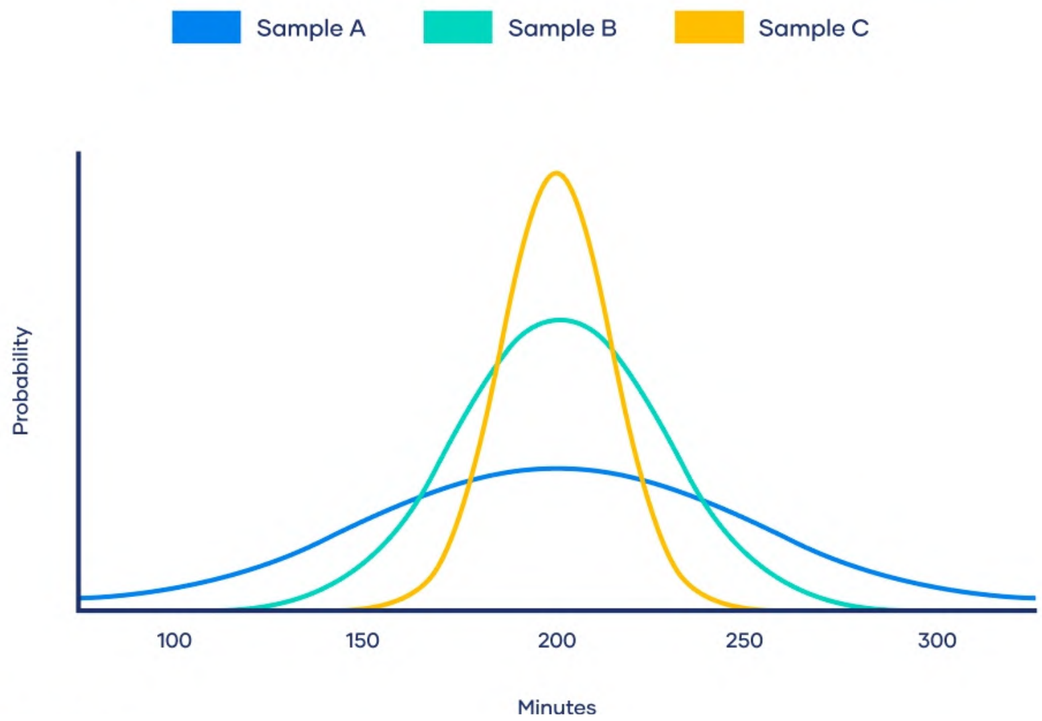
What is the average age?

1. 6 years old
2. 11.5 years old
3. **13.5 years old** ✓
4. 10 years old

avg = (5+4+6+39)/4 = 13.5

Measure of Spread

Also referred to as "variability" or "dispersion"



***Source**: Scribbr*

- How much the data varies, or is "spread out"
- **Standard Deviation:** the average distance from the mean
- **Range:** the difference between the lowest and highest values
- **Quartile:** The values that mark the 25th, 50th, 75th, and 100th percentiles of the data, which are Q1, Q2, Q3, and Q4, respectively
- **Interquartile Range (IQR):** describes the difference between the first and the third quartile, Q3 -Q1

Standard Deviation

Two ways

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Population standard deviation

- For a whole population
- Divide by N

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

Sample standard deviation

- For a sample of a population
- Divide by N-1

Why? It's complicated...

Learning Check

Calculating the standard deviation

slido

You have found the following ages (in years) of all 4 lions at your local zoo:

[5, 4, 6, 39]

What is the standard deviation?

1. 4
2. 11.5
3. 14.7
4. 11

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Learning Check

Calculating the standard deviation

You have found the following ages (in years) of all 4 lions at your local zoo:

[5, 4, 6, 39]

What is the standard deviation?

$$\mu = 13.5$$

$$\sigma = \sqrt{\frac{(5-13.5)^2 + (4-13.5)^2 + (6-13.5)^2 + (39-13.5)^2}{4}}$$

$$= \sqrt{\frac{869}{4}} = 14.7$$ ✓

slido

Learning Check

Calculating IQR

Find the interquartile range (IQR) of the data in the dataset below.

40	39	44	34	32	35	47	35

1. IQR: 7.5
2. IQR: 37.5
3. IQR: 5
4. IQR: 35

Learning Check

Calculating IQR

Find the interquartile range (IQR) of the data in the dataset below.

40	39	44	34	32	35	47	35

1) Sort the data

32 34 35 35 39 40 44 47

2) Split data in half

3) Split halves to get quartiles

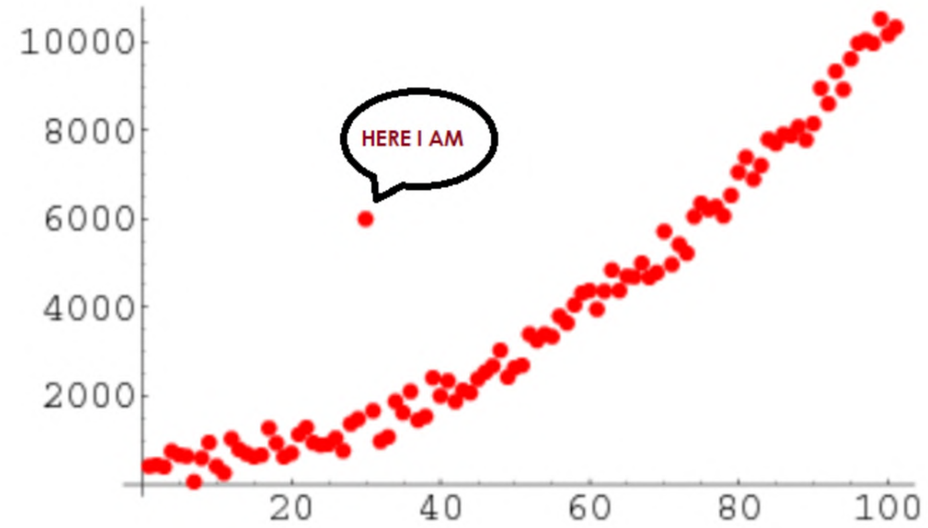
4) Calculate:

Q1 = (35+34)/2 = 34.5
Q3 = (44+40)/2 = 42
IQR = Q3 - Q1 = 42 - 34.5 = **7.5** ✓

Outliers

Outliers

- Outliers are extreme values that differ from most other data points in a dataset.
- They can have a big impact on your descriptive statistics analysis.



How to find outliers?

Sorting Method

Your dataset for a pilot experiment consists of 8 values.

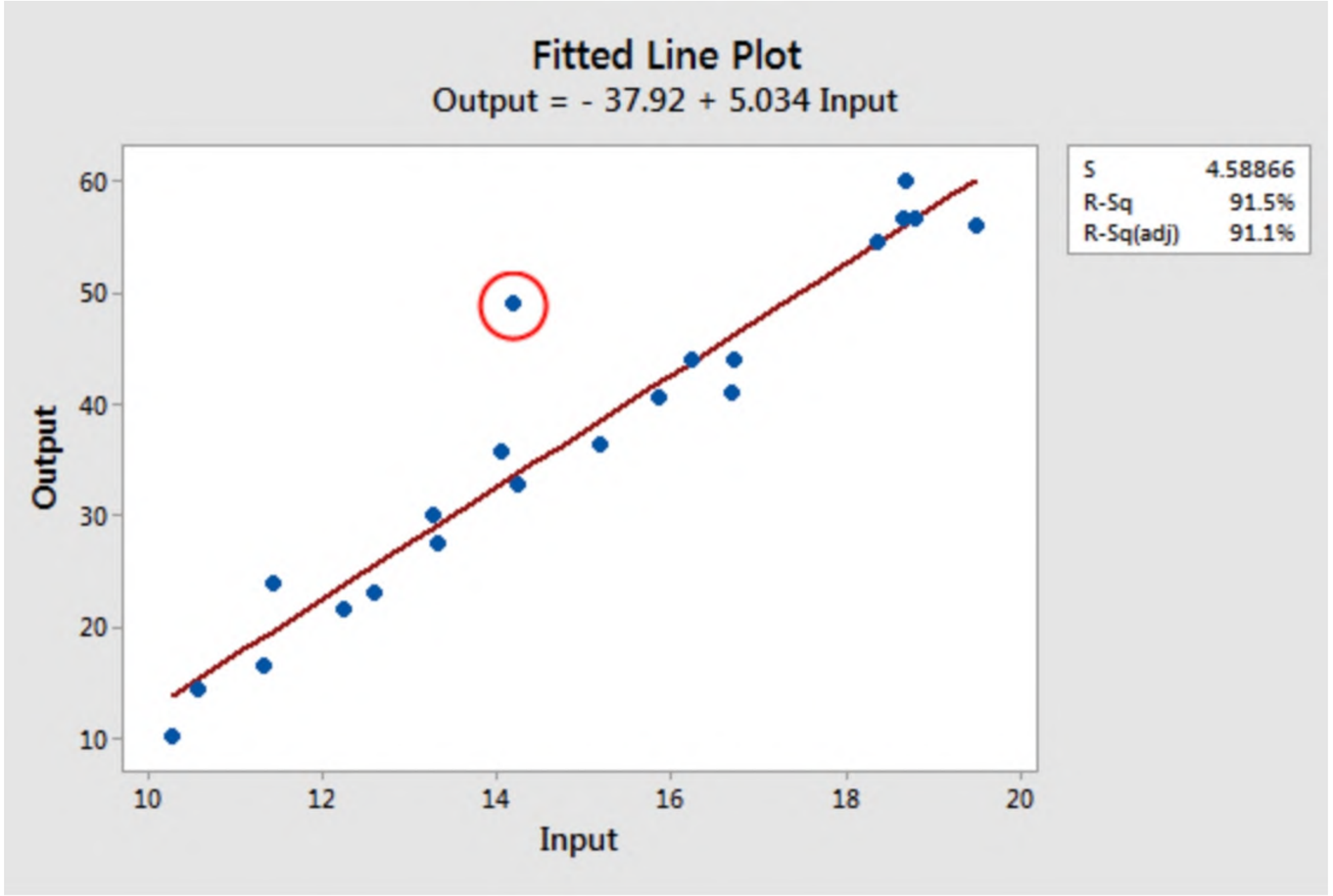
180	156	9	176	163	1827	166	171

You sort the values from low to high and scan for extreme values.

9	156	163	166	171	176	180	**1872**

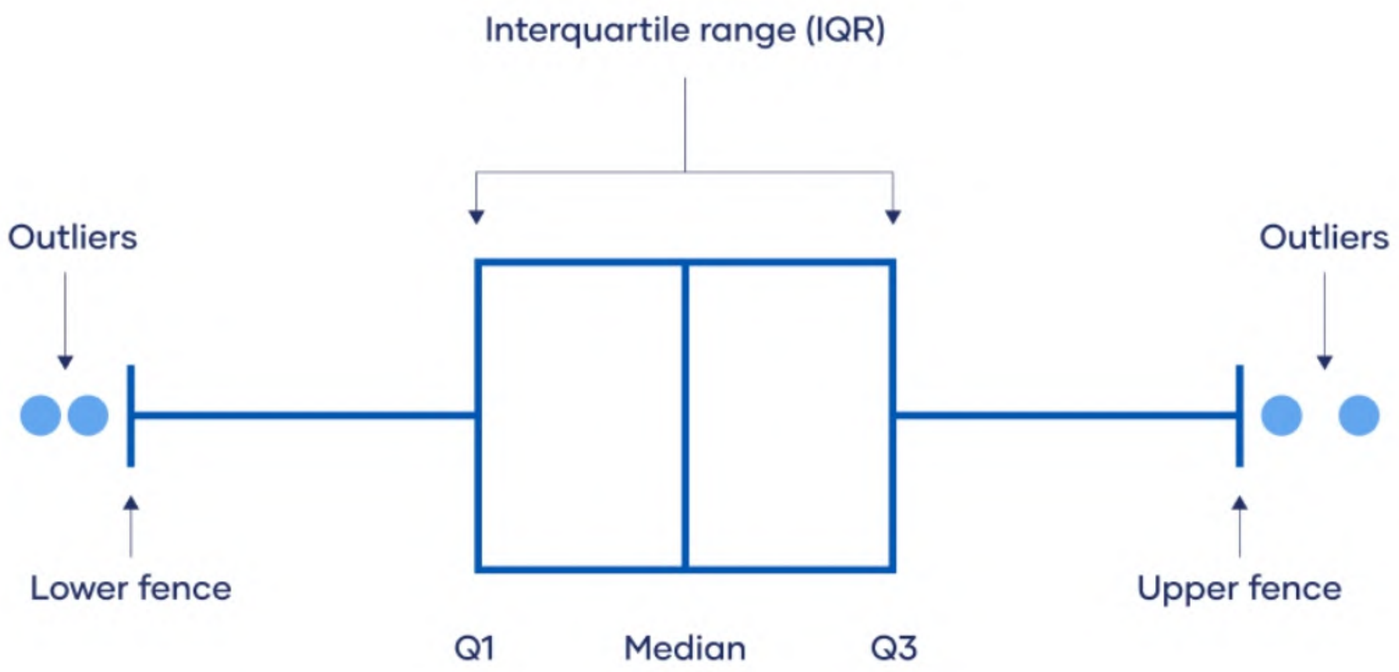
How to find outliers?

Using Visualizations



How to find outliers?

IQR Method



Transforming Data and Outliers

If we transform our data or remove outliers, different measures will be impacted differently.

- **Shifting:** Adding or subtracting a value from every point in the data set. Shifting changes the mean, median, and mode, but not the standard deviation, range, and IQR.

- **Scaling:** Multiplying or dividing a value from every point in the data set. Scaling changes the mean, median, mode, standard deviation, range, and IQR.

- **Outlier:** Removing an outlier will typically change the mean, standard deviation, and range, but not the median, mode, and IQR.

Lesson Summary

Closure

1. **Data Visualizations:** bar plot, line plot, scatterplot, boxplot, histogram
2. **Distribution Shapes:** symmetric/skewed, unimodal/multimodal
3. **Measures of Central Tendency:** mean, median, mode
4. **Measures of Spread:** standard deviation, range, IQR
5. **Outliers:** extreme values that differ from most other data points in a dataset

Closure

Today we:

- Motivation
- Data Visualizations
- Shapes & Distributions
- Measures of Center and Spread
- Outliers

What questions do you have?

Feel free to raise your emoji hand, unmute and speak, or type into the chat.

Thank You

And Keep Practicing!