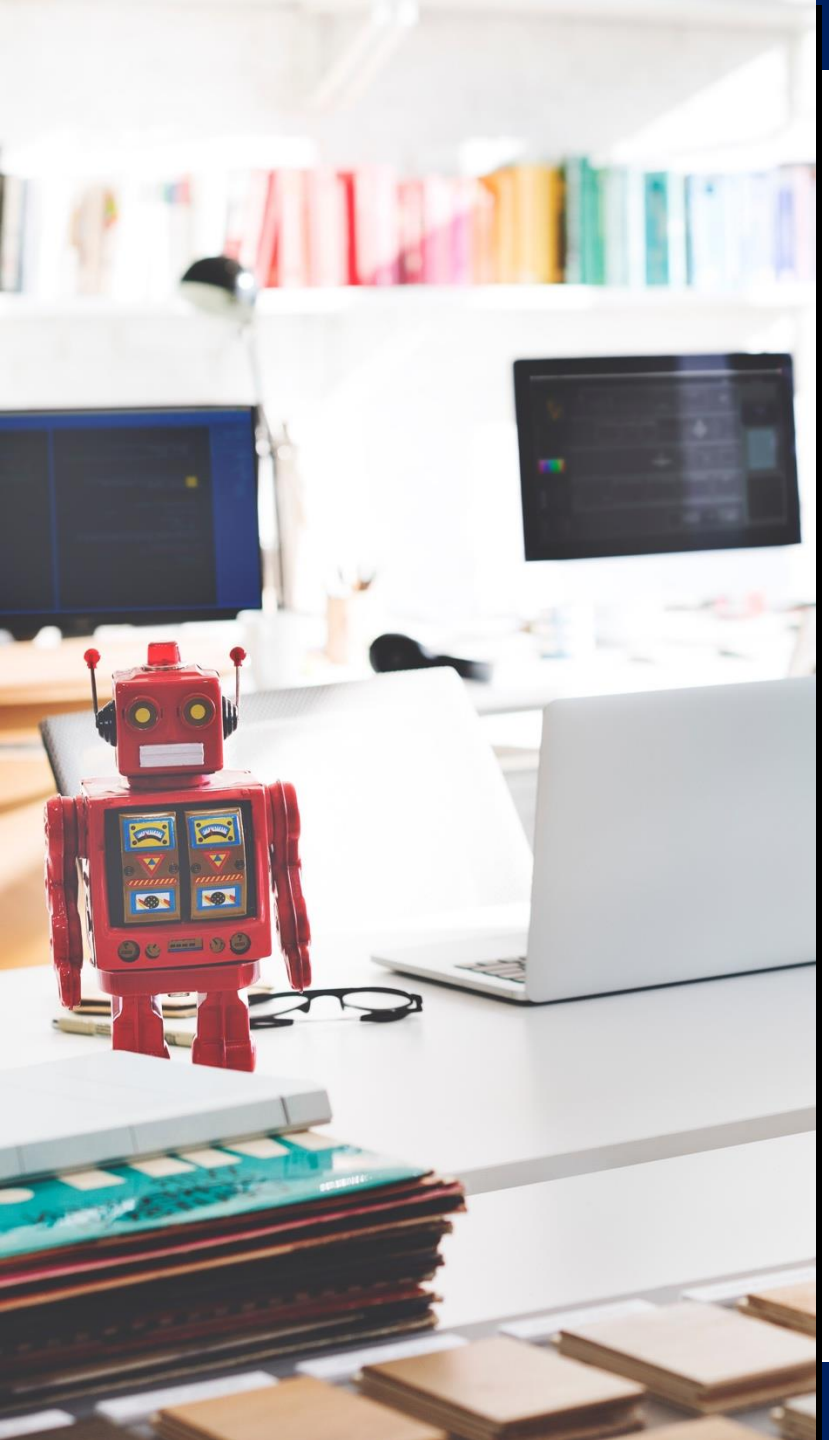
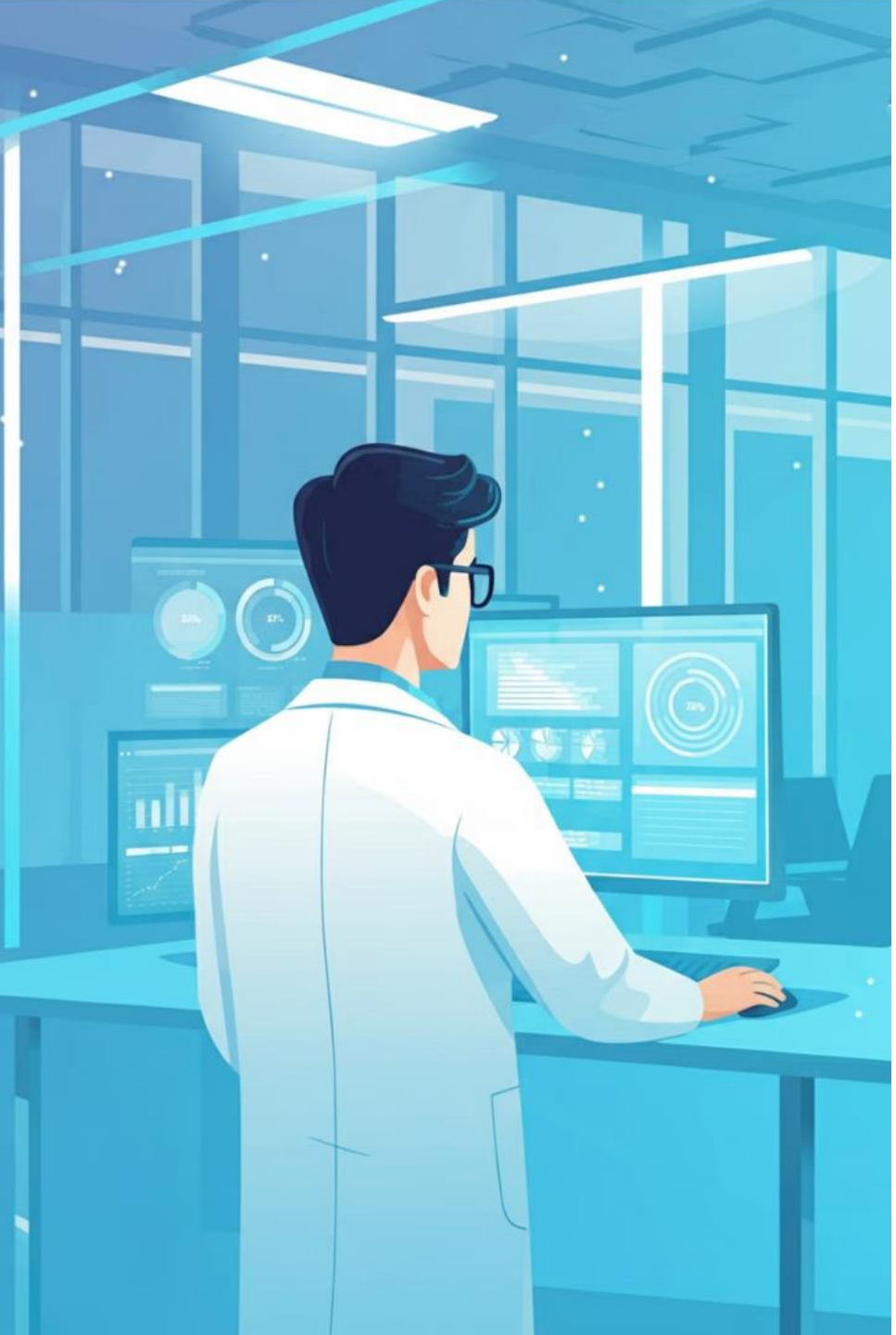


Statistics: The Foundation of Data Science

Introduction to basic concepts



Statistics

- Statistics is essential in data science for:
  - data analysis,
  - decision-making, and
  - building and validating predictive models

- It provides the mathematical foundation that enables data scientists to turn data into meaningful insights

Population and Samples

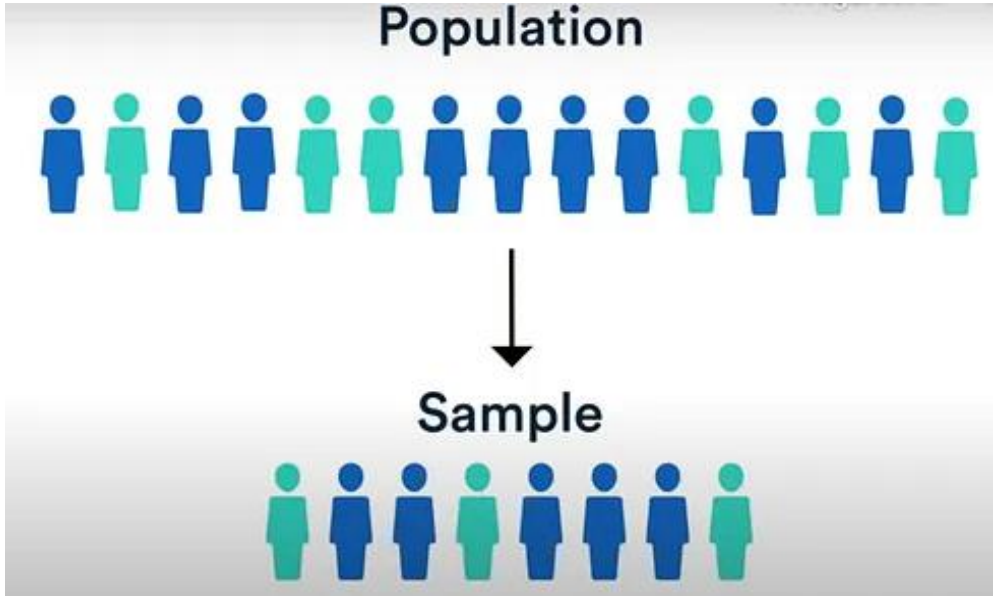
- A **population** refers to the entire group or set of individuals, items, or events of interest that you want to study or draw conclusions about
  - Think of this as a complete collection of data points
  - It includes all possible subjects that meet a specific criterion

- Examples:
  - All the citizens in a country
  - Every product produced by a factory in a year.
  - All students enrolled in a particular school

Population and Samples

- A **sample** is a subset of individuals, items, or events selected from a population
  - It is a smaller group that is representative of the population, used to draw conclusions about the entire group

- Examples:
  - A survey of 1,000 people to represent the opinions of a country's citizens
  - 100 randomly selected products tested for quality control from a day's production
  - A group of 50 students chosen from a school for a study on academic performance.



Population (N) > Sample (n)

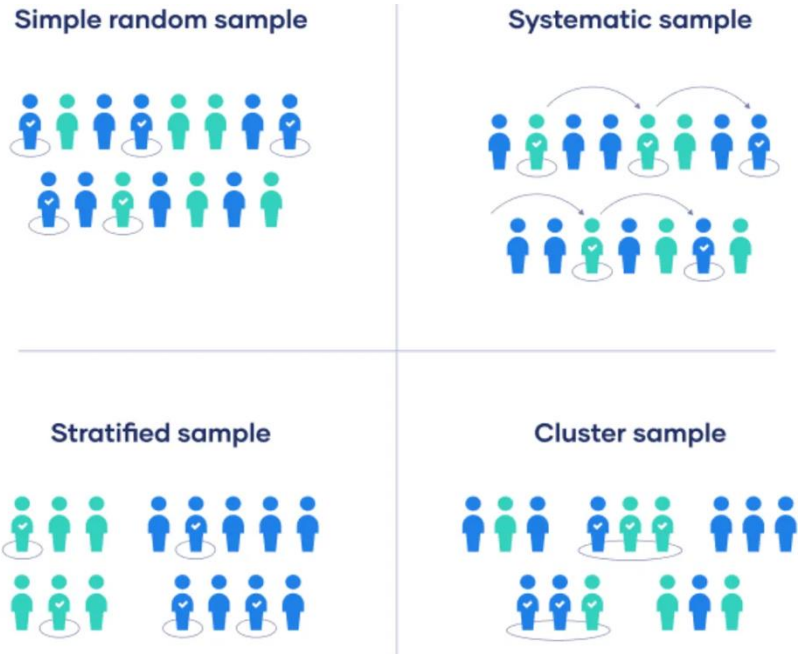
Sampling Techniques

Simple Random Sampling

It ensures that every member of the population has an equal chance of being selected, thereby guaranteeing fair representation

Systematic Sampling

The population is segmented into subgroups, and samples are drawn from each to maintain proportional representation



Stratified Sampling

Involves selecting samples at regular intervals, providing a streamlined approach to random sampling

Cluster Sampling

Dividing the population into clusters and randomly selecting entire clusters for a comprehensive study

Parametes and Statistics

1

Parameters (μ, σ^2)

These measures, like the mean (μ) or variance (σ^2), describe the entire population

They are often unknown because examining every data point in a population is impractical

2

Sample Statistics ($\bar{x}$, s^2)

These are estimates derived from the sample, such as the sample mean ($\bar{x}$) and variance (s^2), aimed at estimating their population counterparts

3

Inferential Statistics

Inferential statistics use sample statistics to make educated guesses about the population, relying on proper sampling methods to validate these guesses

4

Techniques

Techniques like hypothesis testing and confidence intervals are crucial for ensuring the accuracy of our conclusions.

Measures of Central Tendency

Mean

The mean (often called the average) is the sum of all the values in a dataset divided by the number of values

Median

The median is the middle value of a dataset when the values are arranged in ascending or descending order

If there is an even number of observations, the median is the average of the two middle values

Mode

The mode is the value that appears most frequently in a dataset

A dataset can have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode at all (if all values occur with the same frequency).

Variable Types

Categorical (Nominal and Ordinal)

Variables that represent data grouped into categories, often with no inherent order or ranking between the categories

Nominal

They have no natural order, such as colors

Ordinal

They have some order, such as education levels:

- high school < bachelor's < master's

Discrete

It can take on only specific, separate values

Discrete variables are often countable and typically represent whole numbers or categories

E.g.: number of students

Continuous

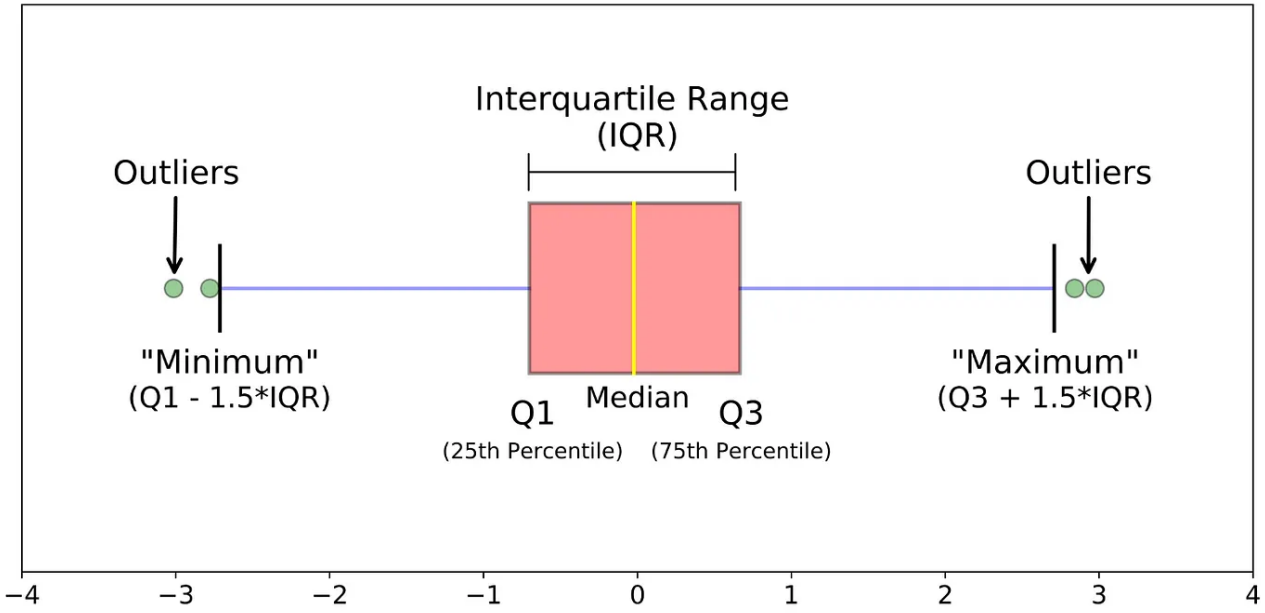
It can take on any value within a given range

They are measurable and can represent any value, including fractions and decimals, within a certain interval

E.g.: height

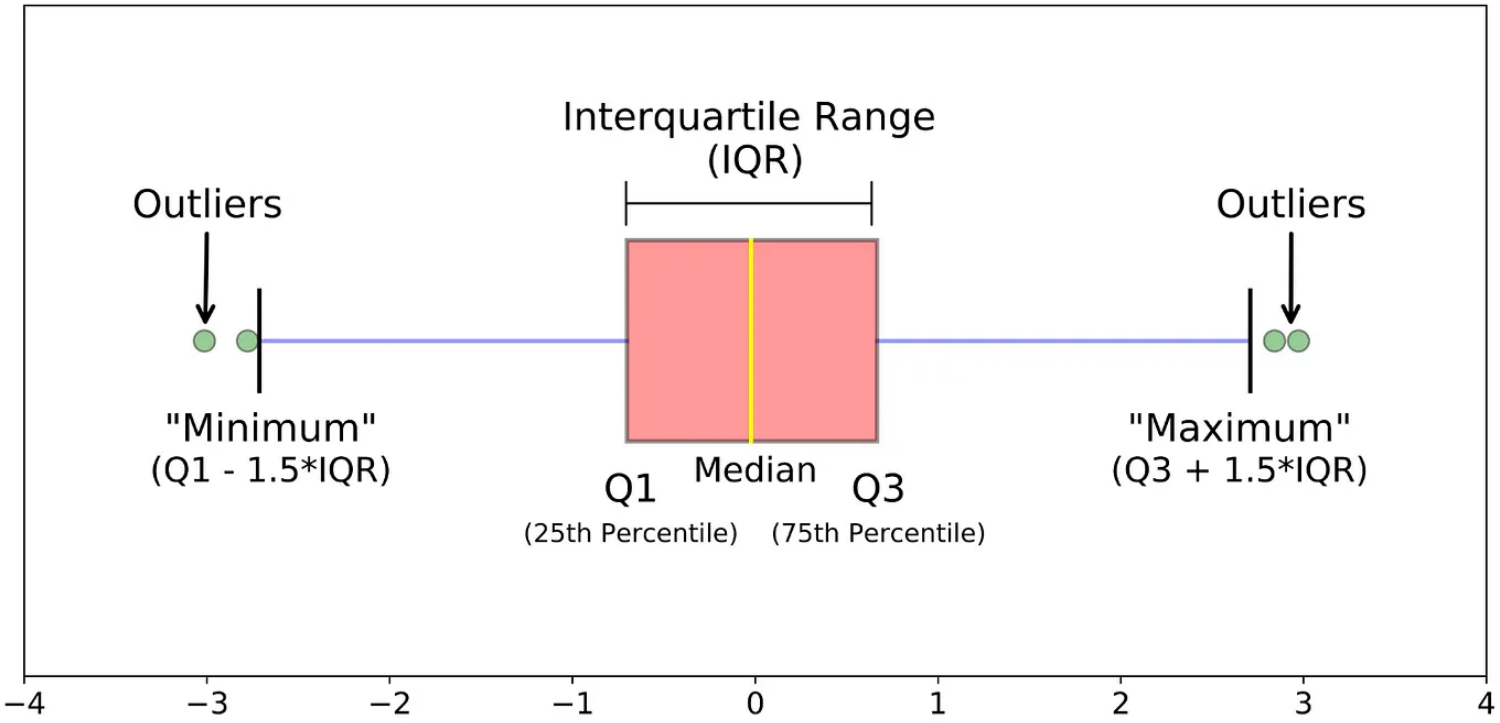
Quartiles

- Quartiles split a dataset into four intervals based on the rank order of values:
  - **First Quartile (Q1): (***aka* the lower quartile), Q1 is the value below which 25% of the data falls
    It is the median of the lower half of the dataset
  - **Second Quartile (Q2):** This is the median of the dataset, dividing the data into two equal halves
    It represents the value below which 50% of the data falls
  - **Third Quartile (Q3):** (*aka* the upper quartile) Q3 is the value below which 75% of the data falls
    It is the median of the upper half of the dataset



Interquartile Range (IRQ)

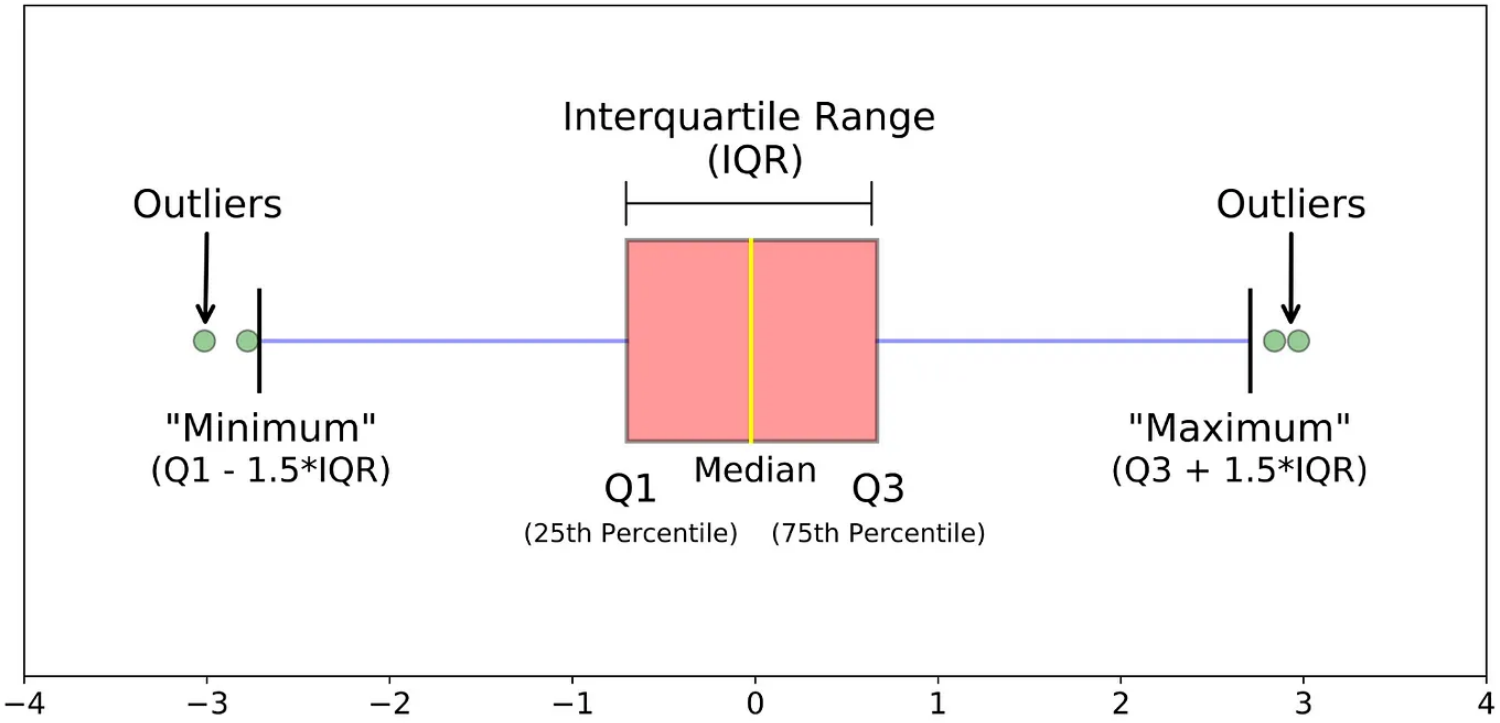
- Interquartile Range (IQR): The range between the first quartile (Q1) and the third quartile (Q3)
- It measures the spread of the middle 50% of the data
- A large IQR indicates that the data is more spread out, while a small IQR indicates that the data points are closer together

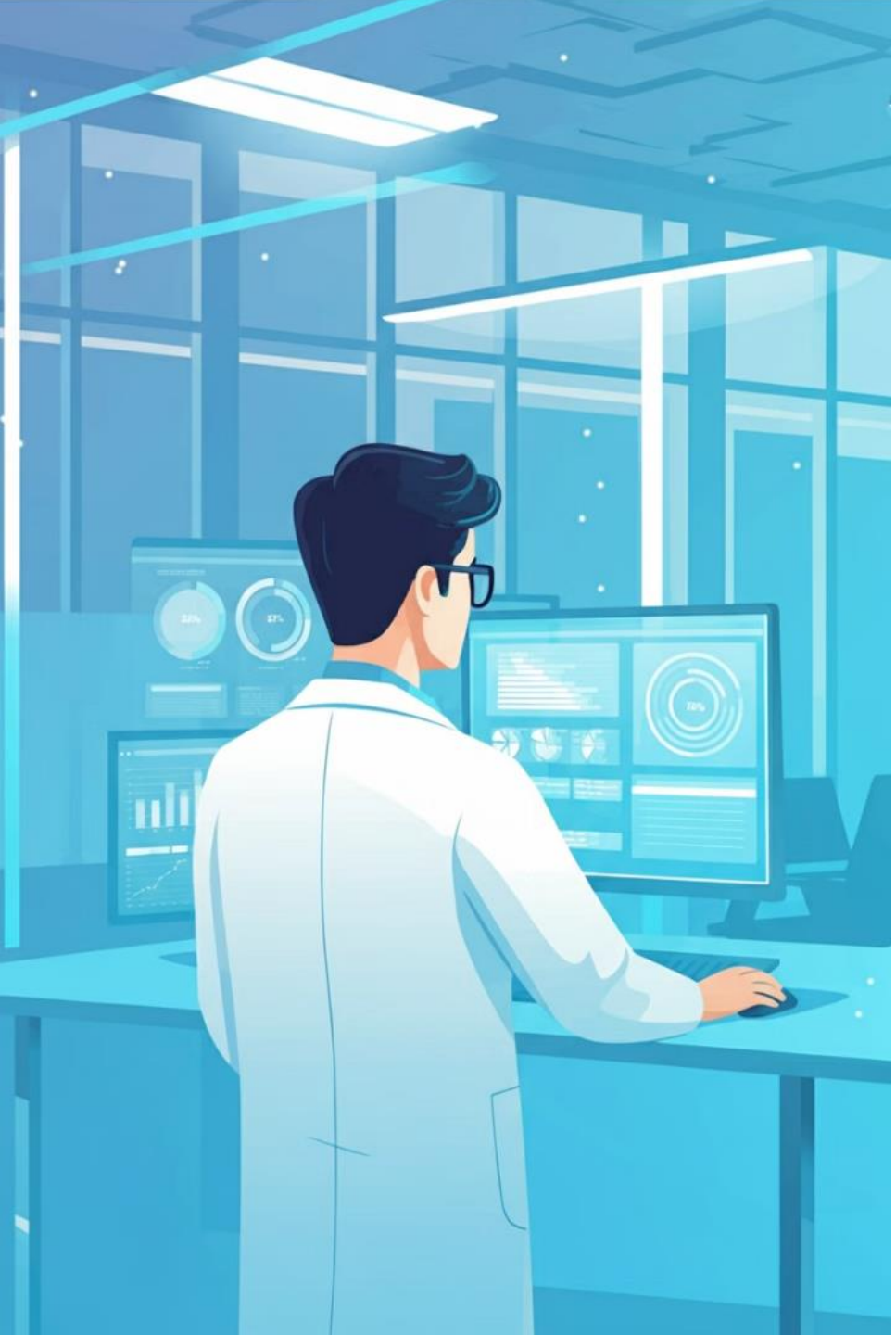


Outliers

- They are data points that differ significantly from the rest of the dataset
  - They are values that lie far outside the overall pattern of distribution, either much higher or much lower than most of the other data points

- Outliers can provide valuable insights, such as errors in data collection or unique phenomena
  - But they can also skew statistical analyses and lead to misleading conclusions.



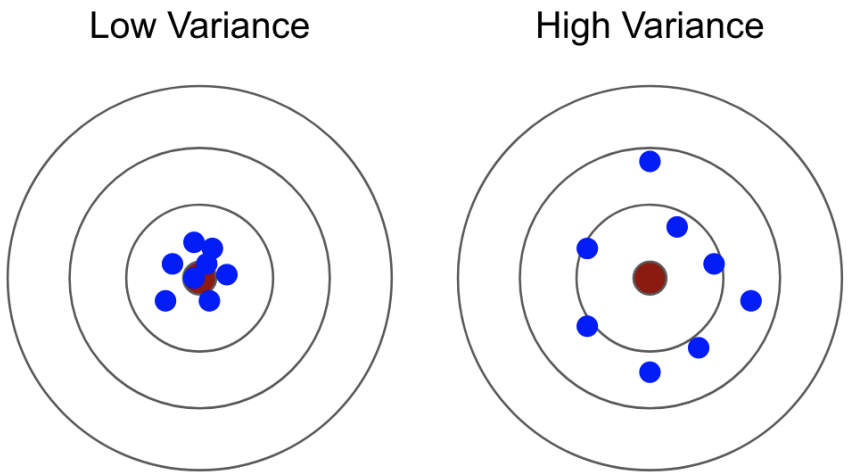


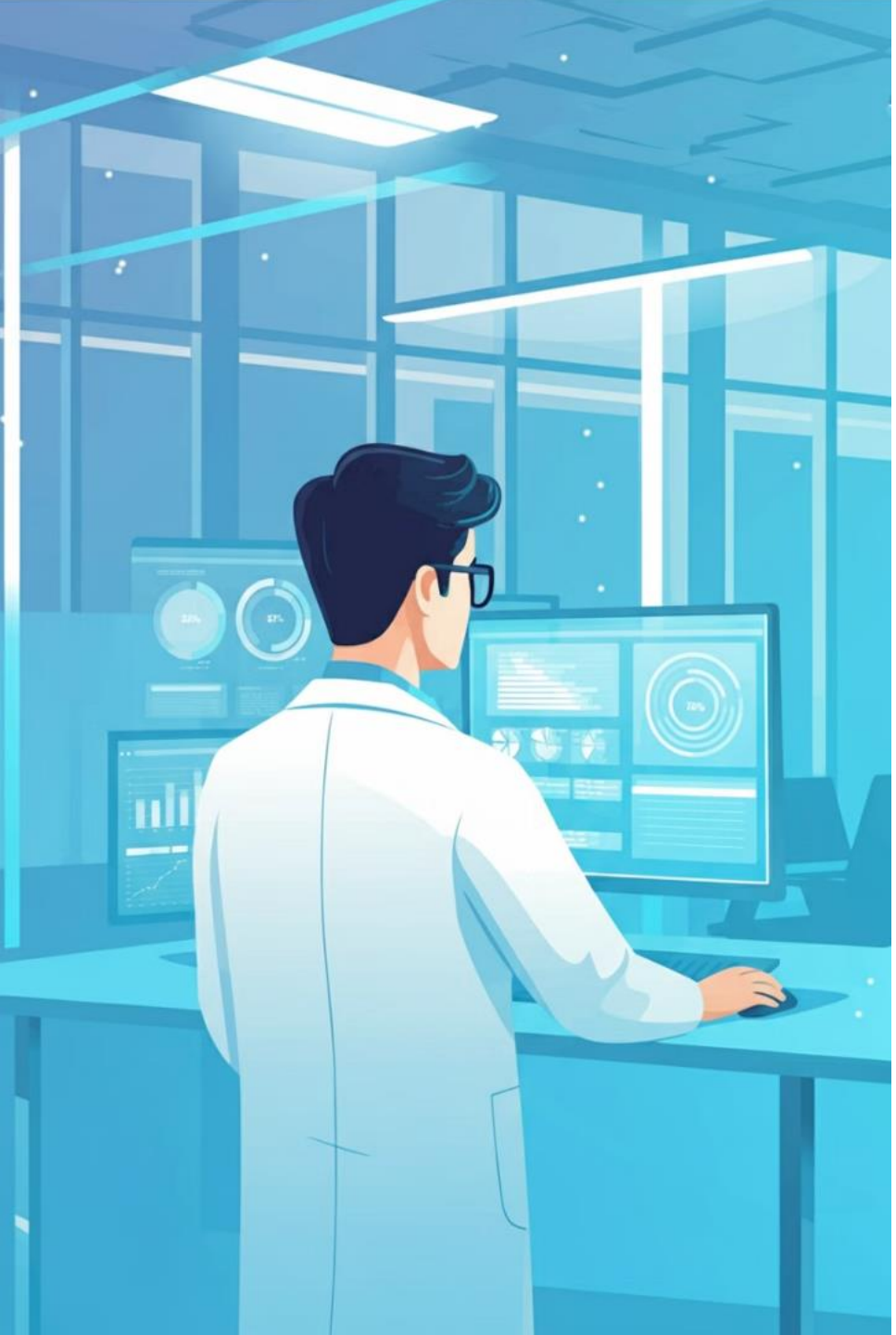
Variance

- **Variance** measures the average squared deviation of each data point from the mean

- It gives an overall idea of how spread out the data points are around the mean

- Interpretation:
  - A lower variance indicates that the data points are closer to the mean
  - A higher variance indicates that the data points are more spread out around the mean



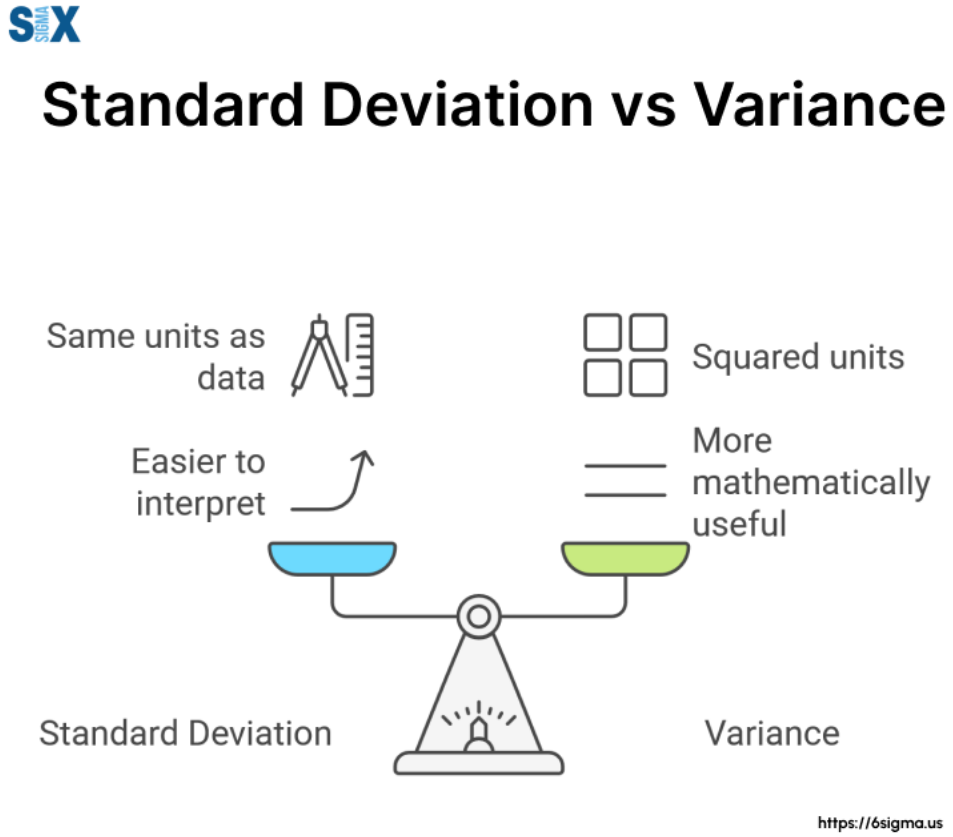


Standard Deviation

- **Standard Deviation** is the square root of the variance
- It is expressed in the same units as the original data, making it more interpretable than variance
- It provides a measure of how spread out the values are around the mean
- Interpretation:
  - A smaller standard deviation indicates that the data points are closer to the mean
  - A larger standard deviation indicates that the data points are more spread out

Standard Deviation vs Variance

	Standard Deviation	Variance
What Is It?	The square root of the variance	The average of the squared differences from the mean
What Does It Indicate?	The spread between numbers in a data set	The average degree to which each point differs from the mean
How Is It Expressed?	The same as the units in the data set	In squared units or as a percentage
What Does It Mean?	A low standard deviation (spread) means low volatility, while a high standard deviation (spread) means higher volatility	The degree to which returns vary or change over time



Source: https://www.investopedia.com/ask/answers/021215/what-difference-between-standard-deviation-and-variance.asp

Correlation

- It describes the strength and direction of a relationship between two variables
  - It tells us how two things are related to each other and whether changes in one variable are associated with changes in another

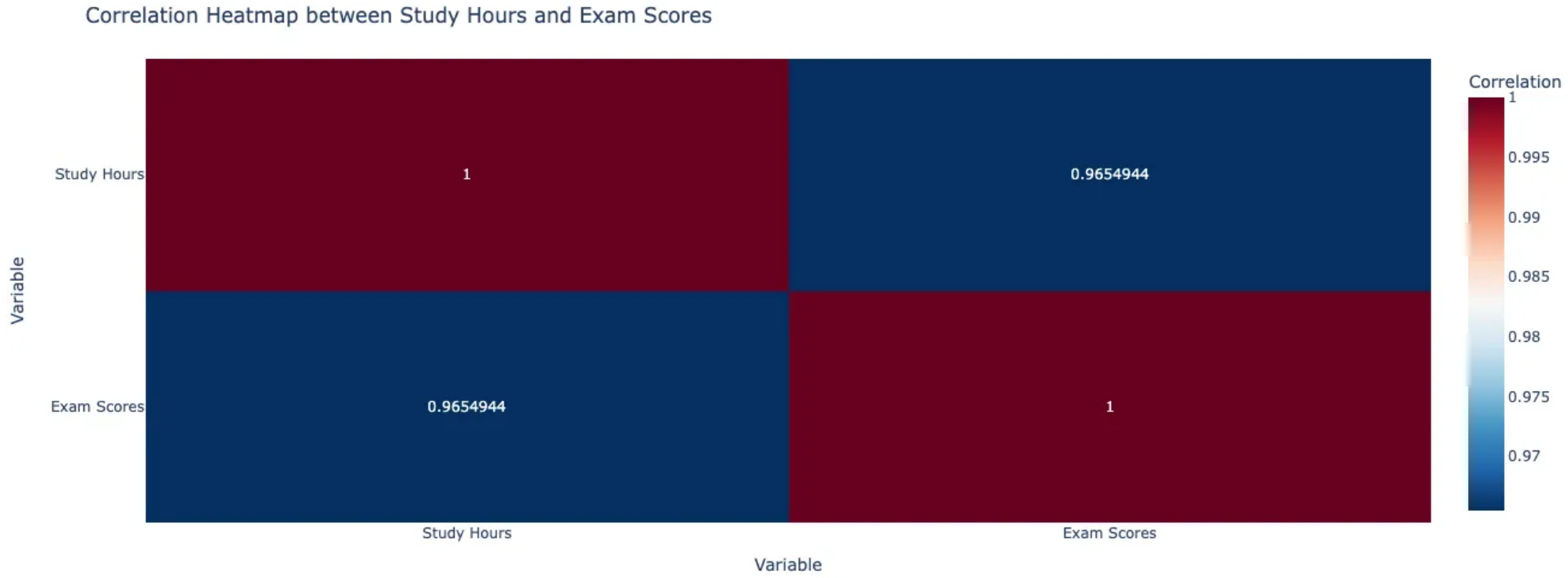
- **Positive Correlation:** When both variables increase or decrease together
  - Example: The more you exercise, the more calories you burn
    - Both exercise time and calories burned go up together

- **Negative Correlation:** When one variable increases while the other decreases.
  - Example: The more time you spend studying, the fewer mistakes you make on a test
    - As study time goes up, mistakes go down

- **Zero Correlation:** When there is no predictable relationship between the two variables.
  - Example: The amount of coffee you drink and the score you get on a quiz are unrelated

Correlation Matrix

- A table that shows the correlation coefficients between multiple variables
- It helps to understand the relationships between all pairs of variables in a dataset immediately
- Each cell in the matrix represents the correlation coefficient (r) between two variables



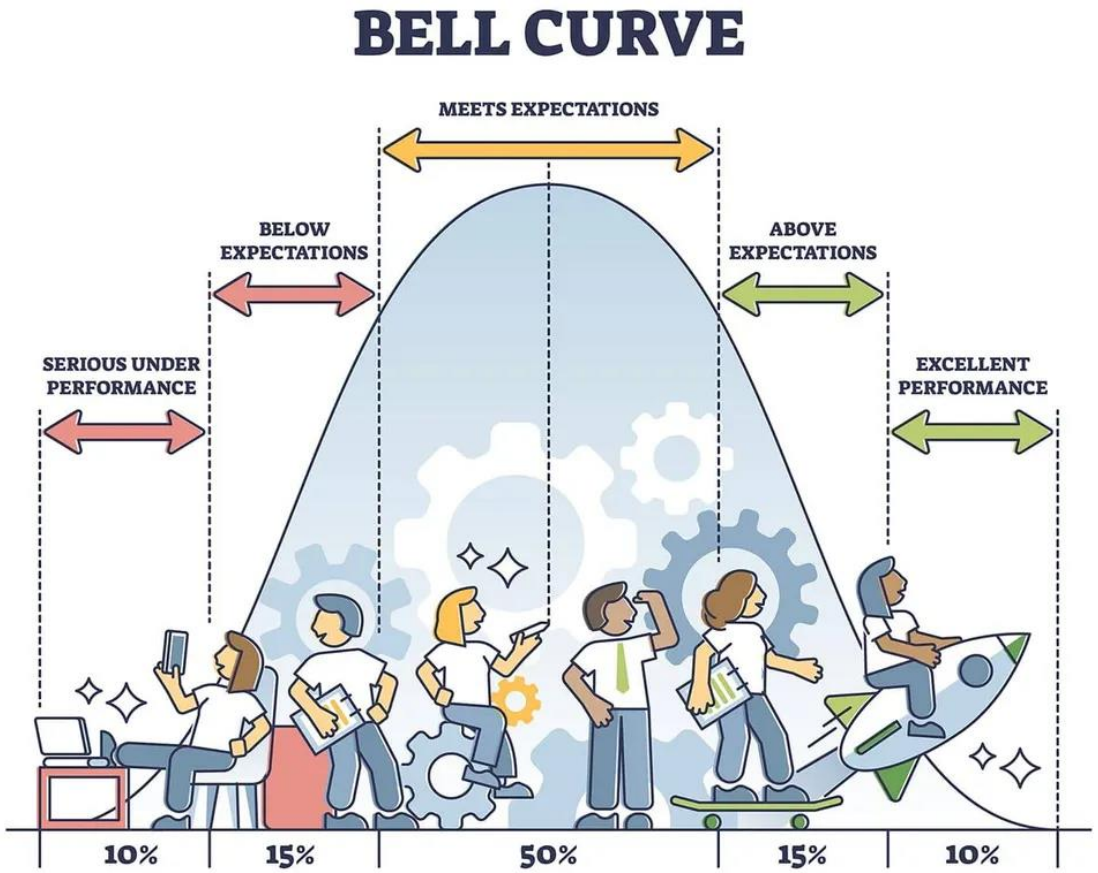
Normal Distribution

Normal Distribution

- A normal curve, or bell curve, shows how common different values are

- This pattern, where most things are average and only a few are extreme, looks like a bell
  - That’s why we call it a bell curve

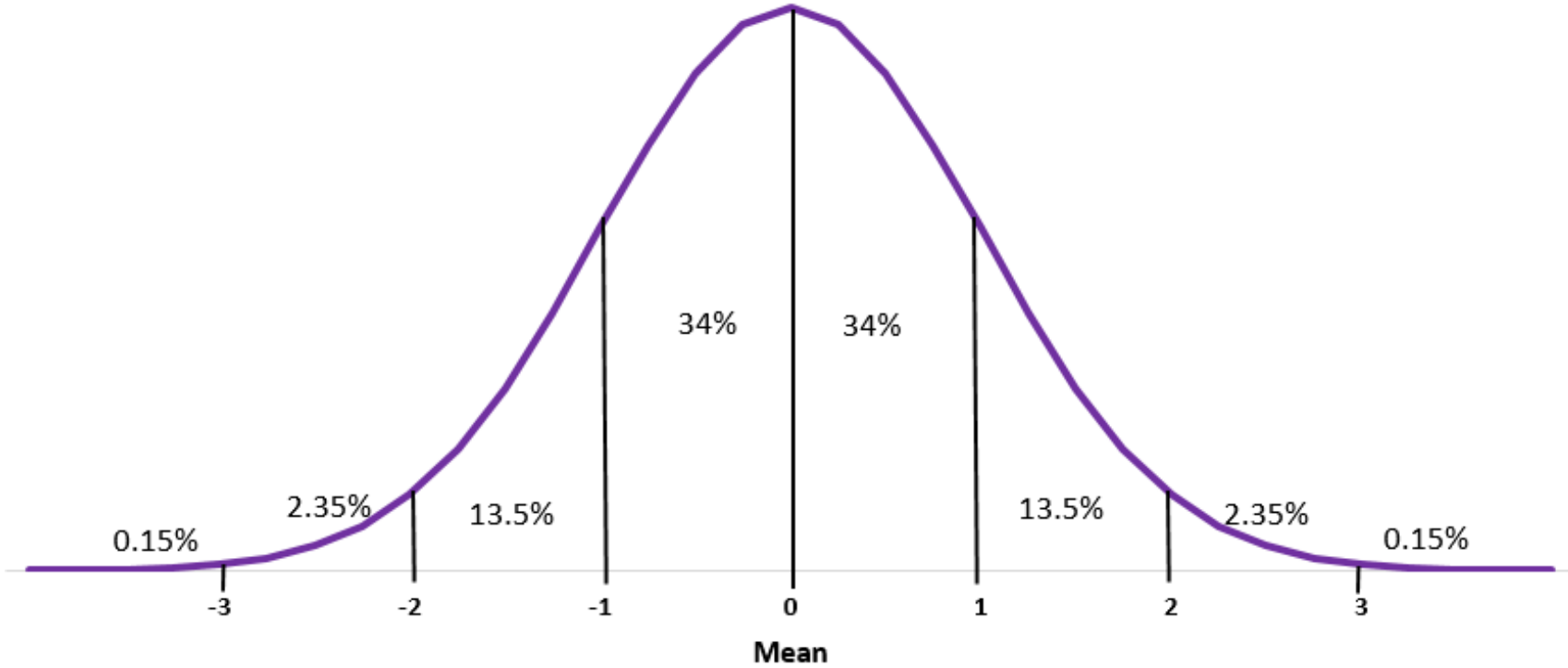
- The bell curve, or normal distribution, is important because it helps us understand how data are spread around the average.



Normal Distribution

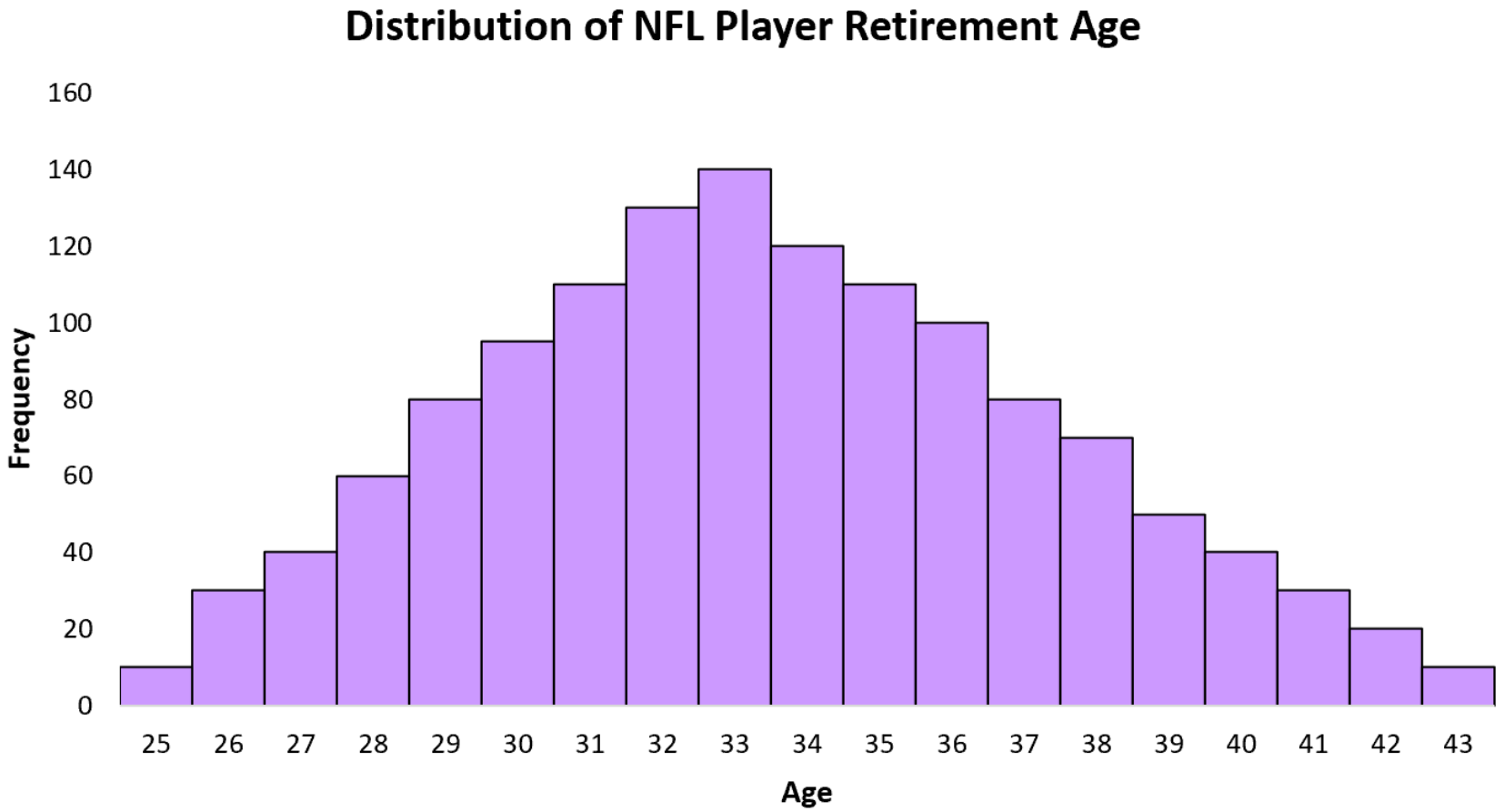
- Bell shape
- Symmetrical
- Mean and median are equal; both are located at the center of the distribution
- About 68% of data falls within one standard deviation of the mean
- About 95% of data falls within two standard deviations of the mean
- About 99.7% of data falls within three standard deviations of the mean

Empirical Rule
(68-95-99.7 rule)



Average NFL Player Retirement Age

- The distribution of retirement age for NFL players is normally distributed with a mean of 33 years old and a standard deviation of about 2 years



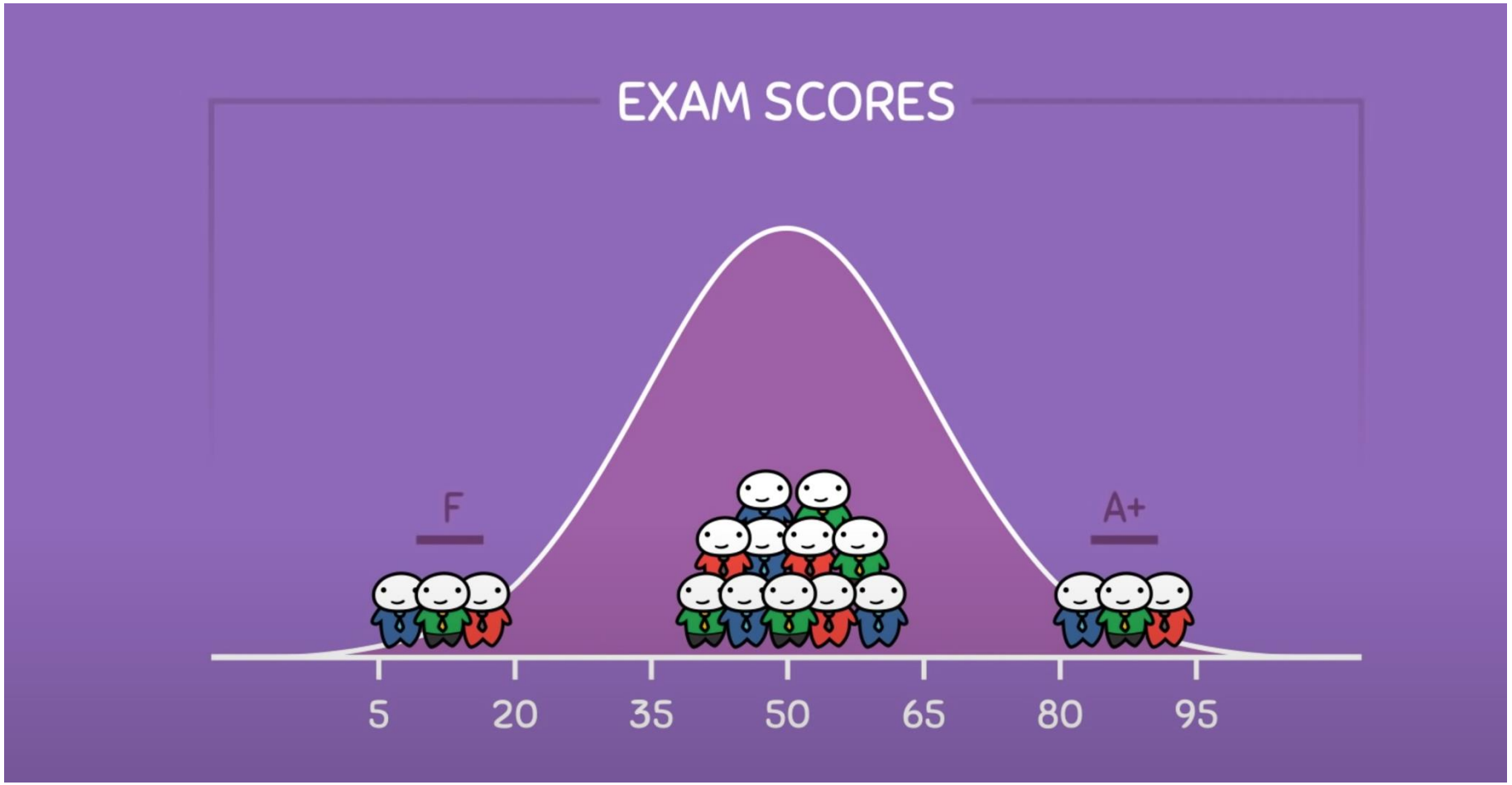
A histogram of this distribution exhibits a classical bell shape

Source: https://www.statology.org/example-of-normal-distribution/

Why is normal distribution important?

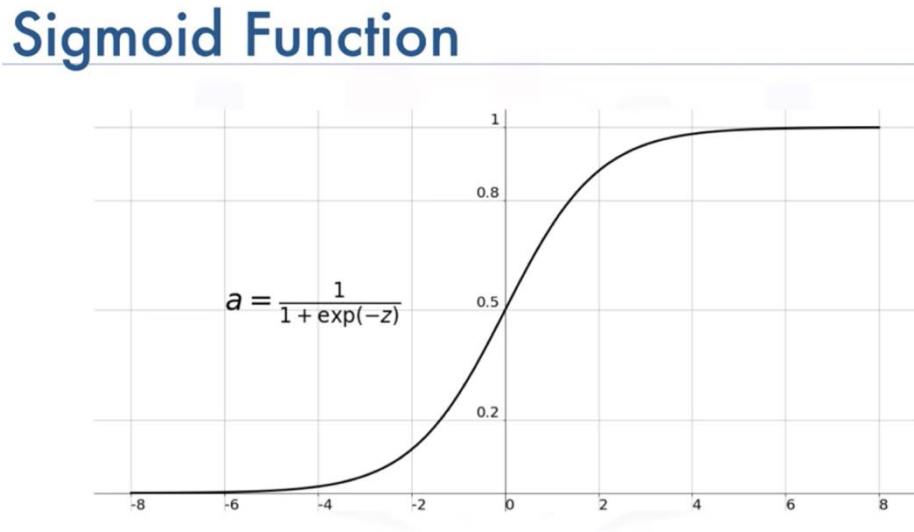
Importance of Normal Distribution

- Many random variables in the real world follow a normal distribution



Importance of Normal Distribution (cont'd)

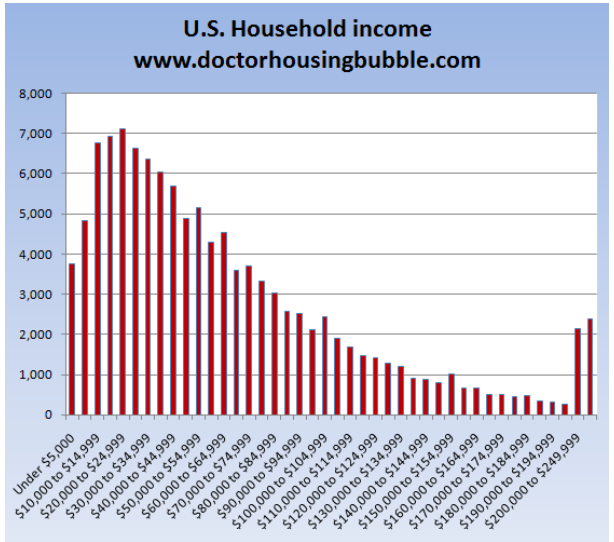
- We can make inference about a variable value if it is approximately normally distributed
- Some **statistical hypothesis test** assumes that the data follows a normal distribution
- It is beneficial for model building since it makes math easier
  - Models like LDA, Gaussian Naive Bayes, Logistic Regression, Linear Regression, etc., are explicitly calculated from the assumption that the distribution is normal
  - *Sigmoid functions* work most naturally with normally distributed data



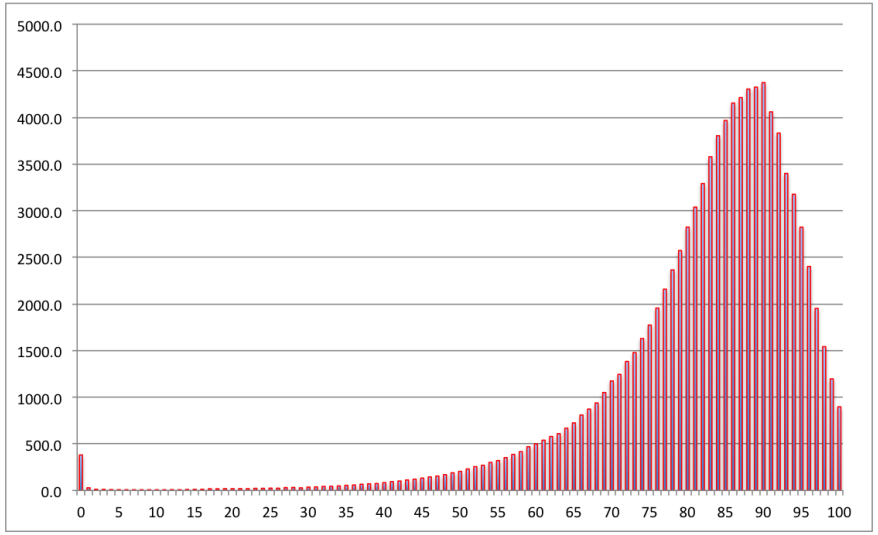
Skewed Data

- Skewness is a measurement of the distortion of symmetrical distribution or asymmetry in a data set

- Skewness is demonstrated on a bell curve when data points are not distributed symmetrically to the left and right sides of the median on a bell curve
  - Skewed data occurs when curve appears distorted or skewed either to the right or to the left (statistical distribution)



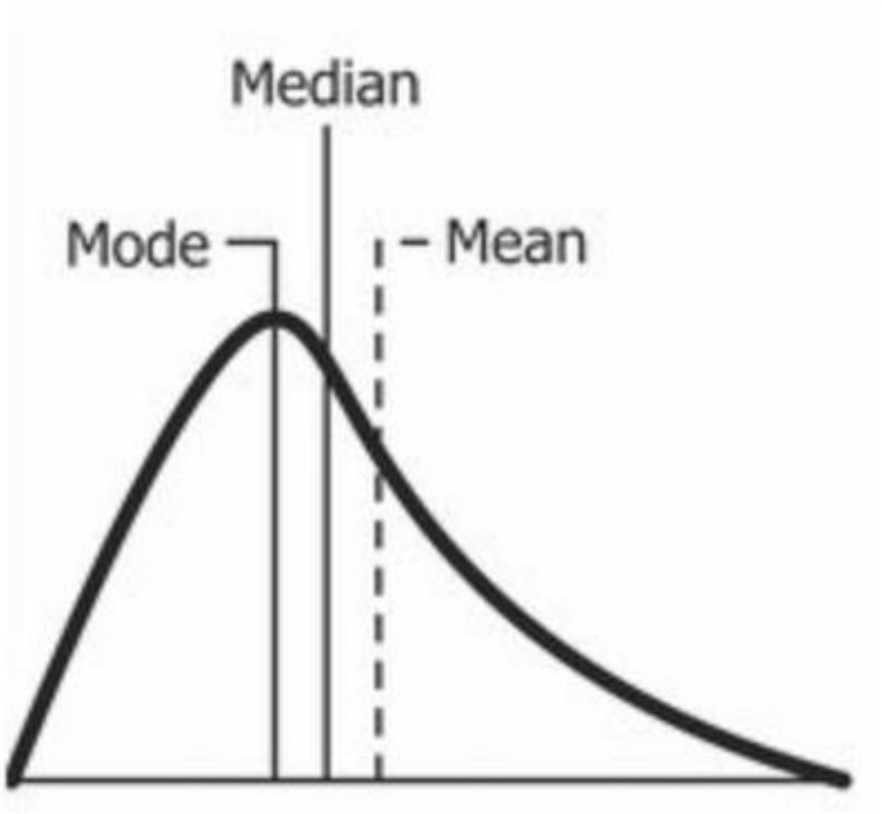
U.S. Household income

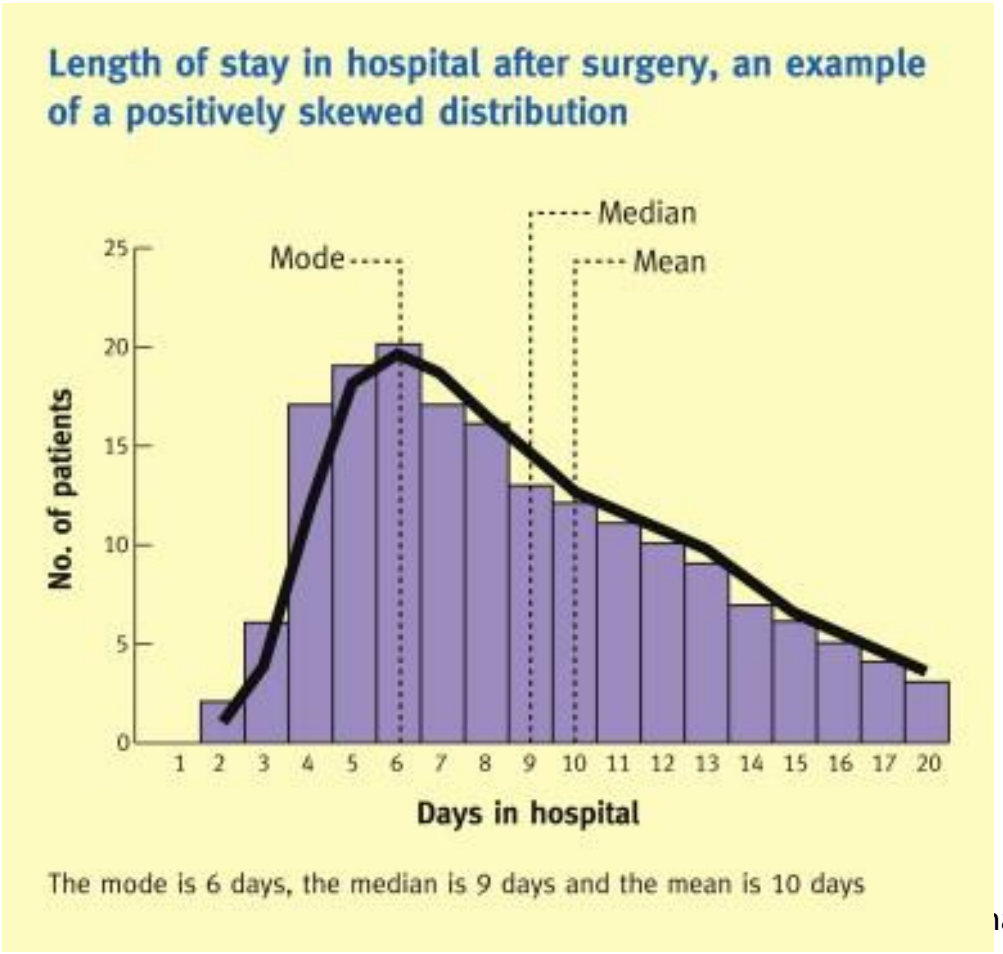


Mortality Distribution Graph

Positively Skewed Distribution

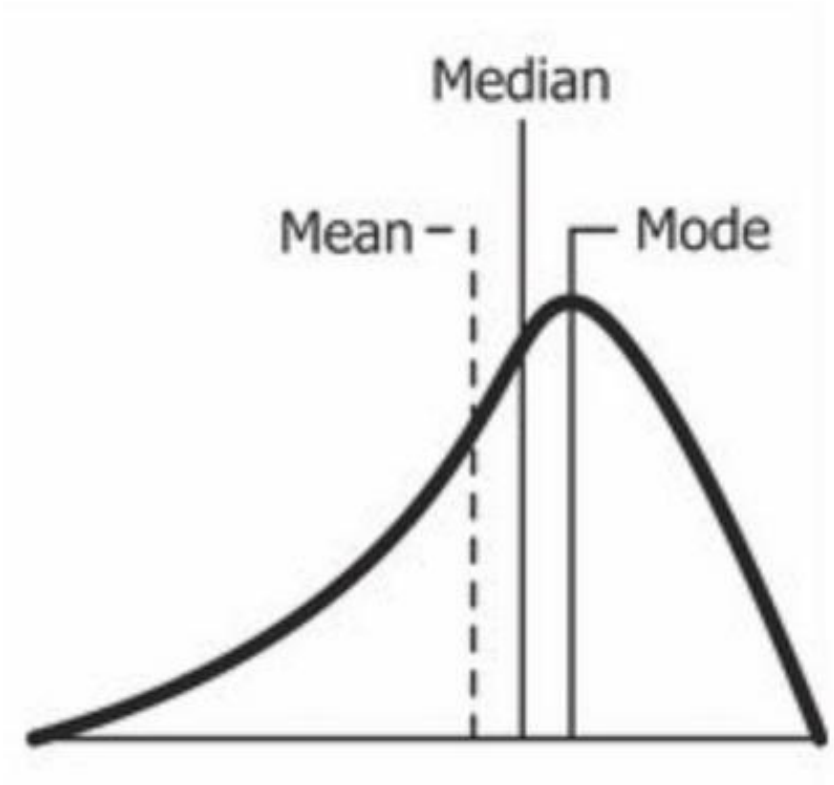
- Most values are clustered around the **left tail** of the distribution
  - Mean > Median > Mode

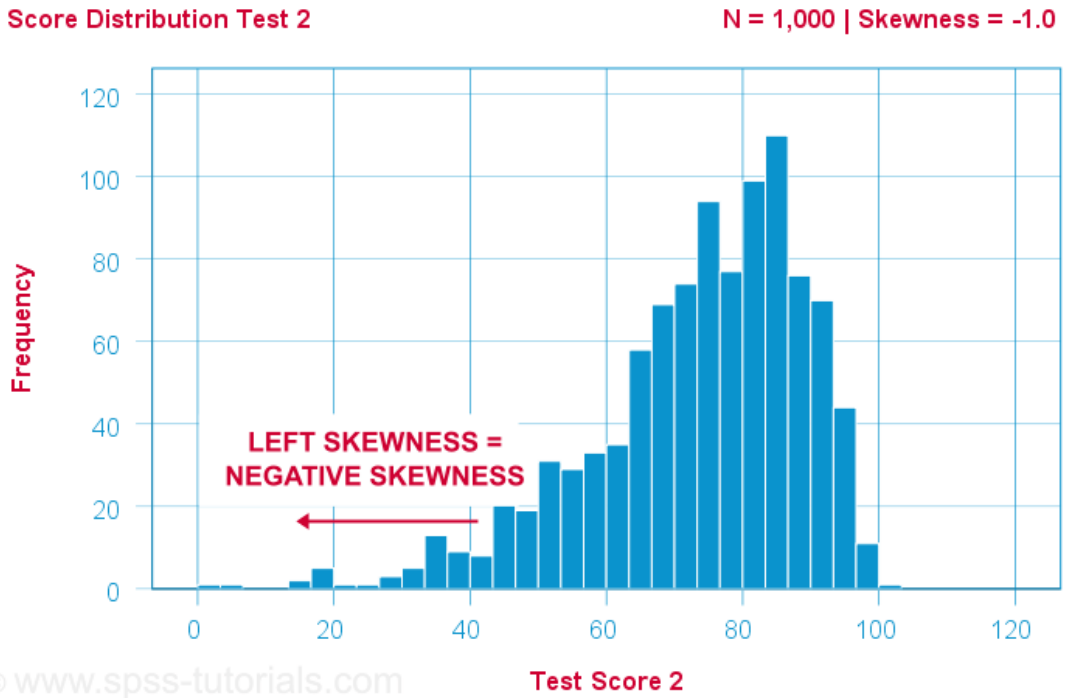


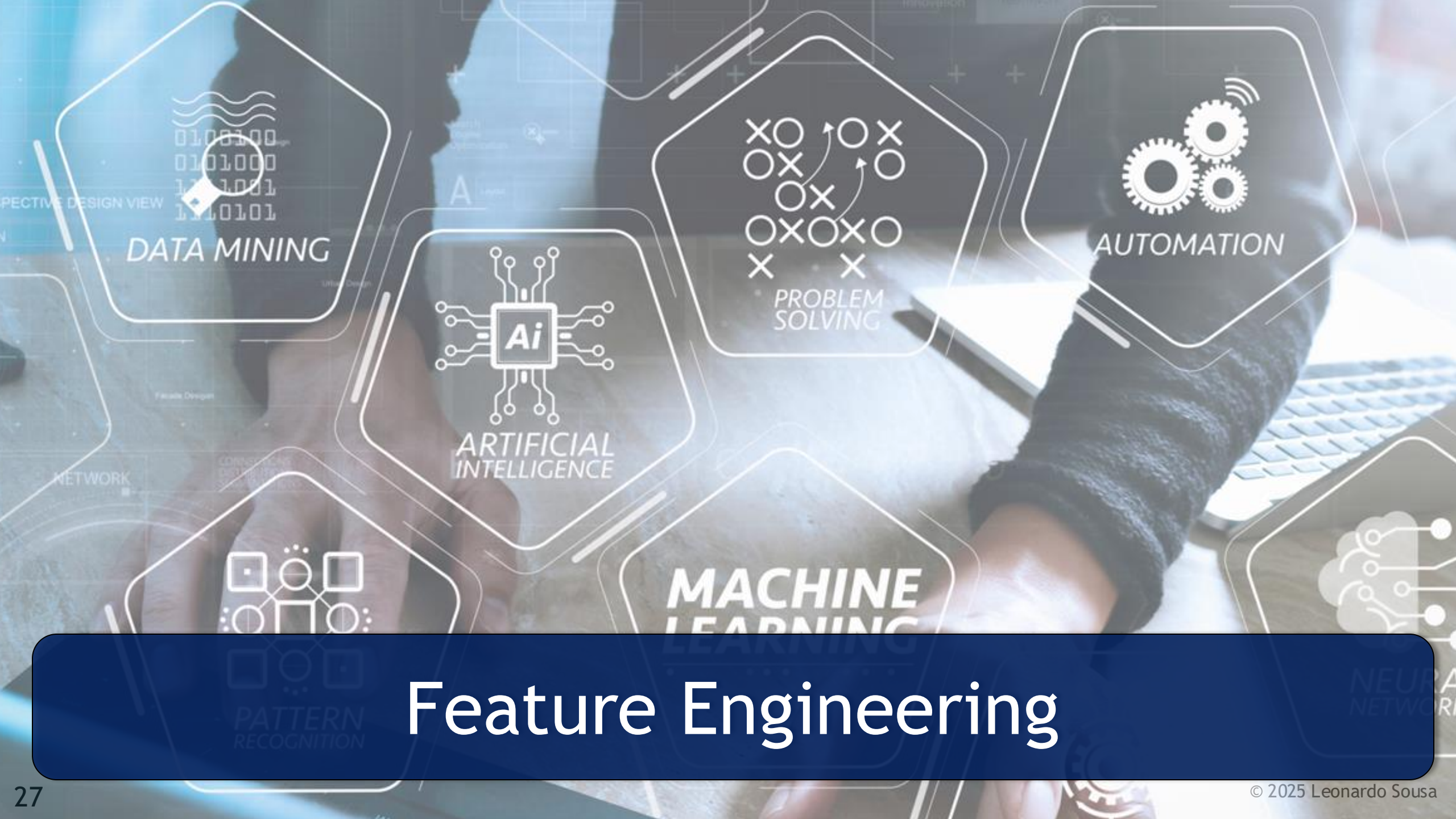


Negatively Skewed Distribution

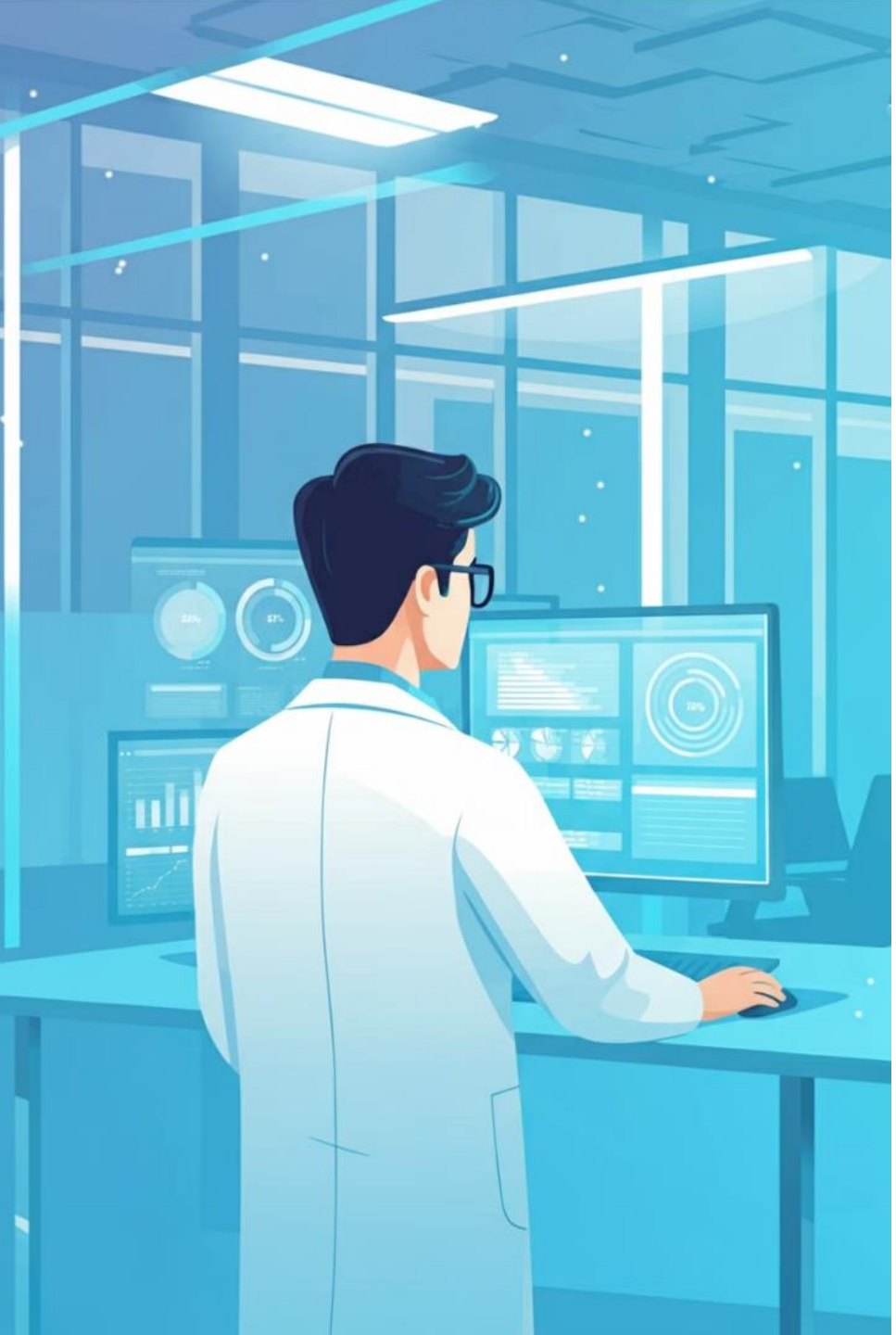
- Most values are clustered around the **right tail** of the distribution
  - Mode > Median > Mean







Feature Engineering



Feature Engineering

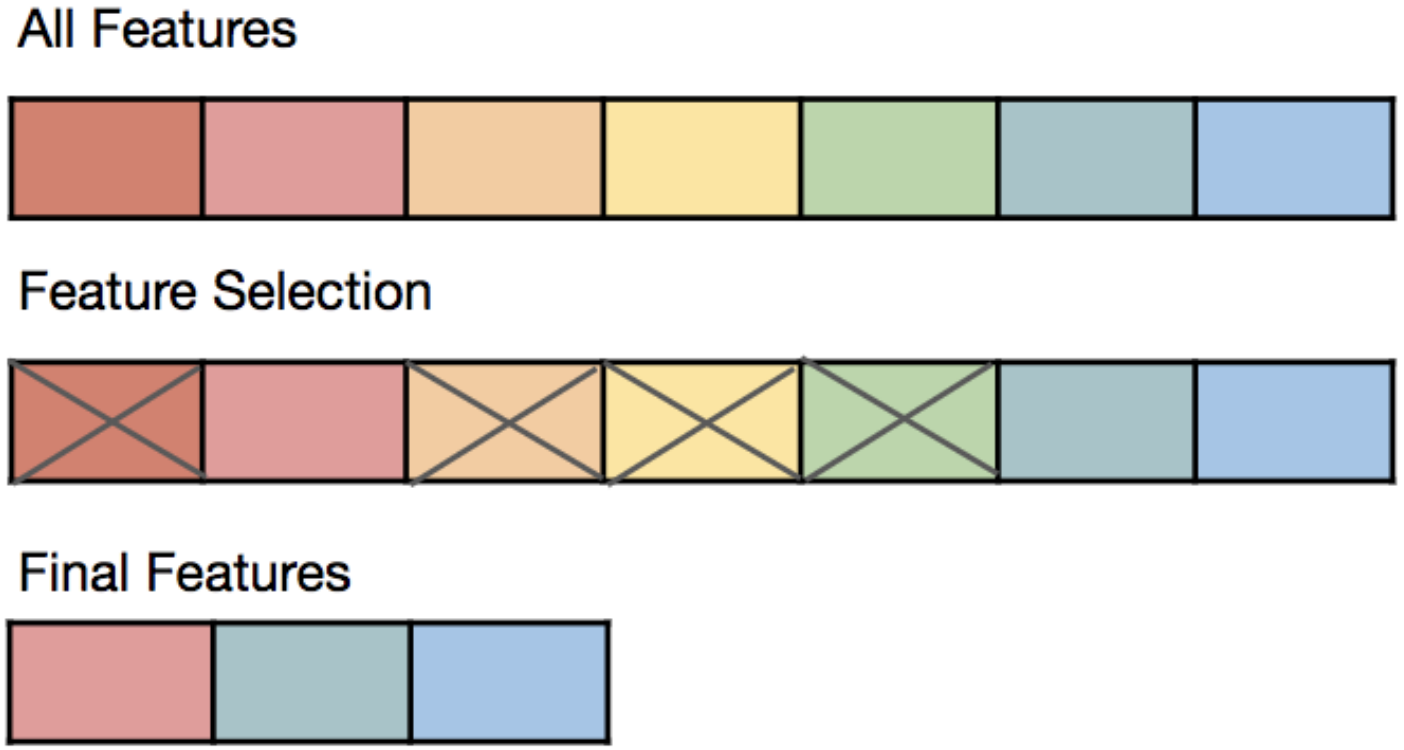
- The process of creating, selecting, and transforming variables (features) from raw data to improve the performance of machine learning models

- It involves using:
  - domain knowledge,
  - data analysis, and
  - creativity to create new input features that
- help the model learn patterns more effectively, ultimately improving the model's predictive power

- Benefits:
  - ↑ Improves Model Performance
  - ↑ Reduces Overfitting
  - ↑ Handles Data Limitations
  - ↑ Makes Data Suitable for Machine Learning

Feature Selection (Introduction)

- Process of selecting essential features that are more uniform, non-redundant, and relevant for your ML model



Why do we need feature selection?

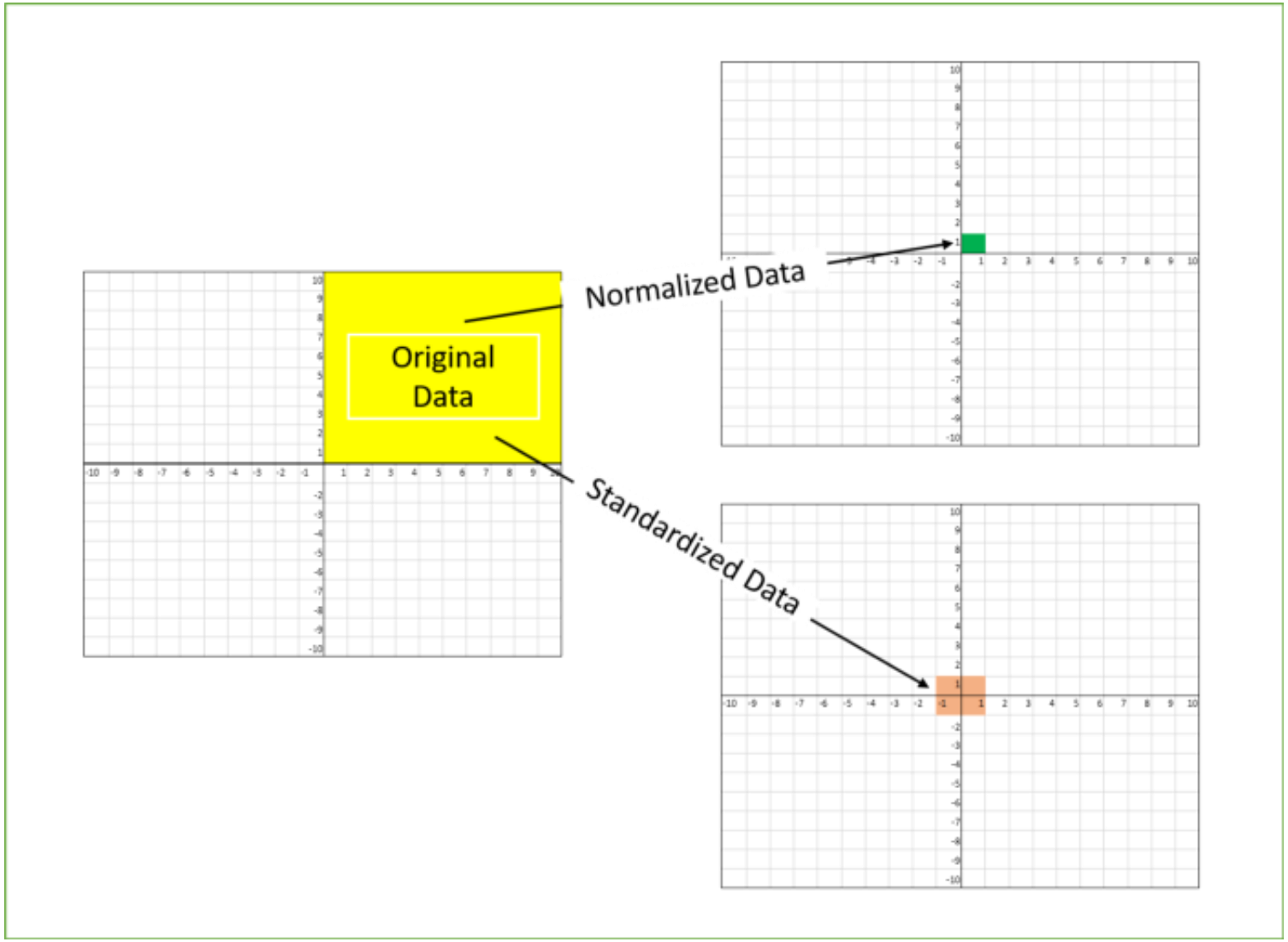
Feature Scaling

- Features with different ranges
  - Metrics that varies from 0 to 1
  - Metrics that varies from 0 to 188

- Feature scaling[1]
  - method used to normalize the range of independent variables or features of data
  - *Normalization*: values are shifted and rescaled so that they end up ranging between 0 and 1
  - *Standadization*: values are transformed to have zero mean and a variance of 1

1. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Feature Scaling



Source: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Feature Scaling

- **Normalization** when you know that the distribution of your data does not follow a normal distribution
  - Useful in algorithms that do not assume any distribution of the data
  - E.g., K-Nearest Neighbors and Neural Networks

- **Standardization** when the data follows a normal distribution
  - However, this does not have to be necessarily true
  - E.g., Linear Regression and SVM

- Unlike normalization, standardization does not have a bounding range
  - Even if you have outliers in your data, they will not be affected by standardization

1. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35