DATA ANALYTICS

IRON HACK

Exploratory Data Analysis

Descriptive Statistics & Outliers

Table of Contents

Part I: Introduction and Univariate Analysis

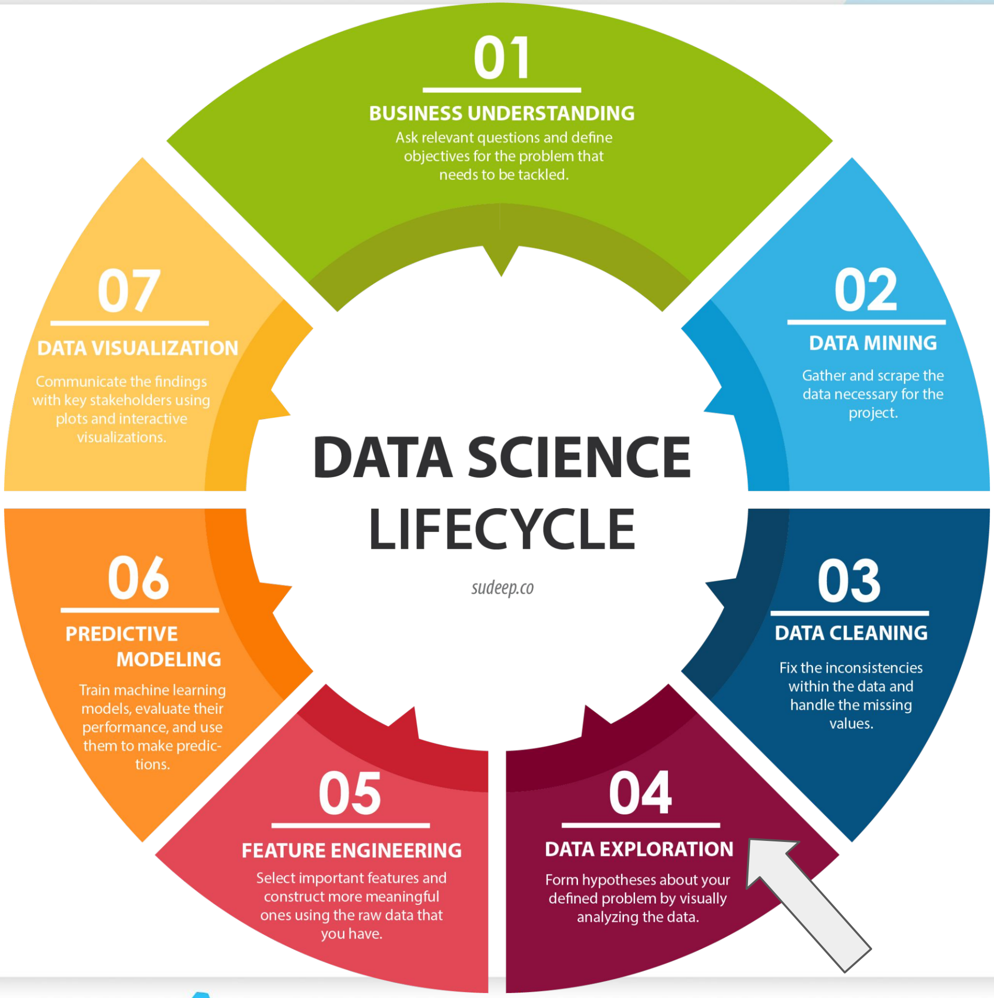
- What's EDA
- Key Concepts
- Data types
- EDA Framework
- Univariate Analysis techniques

Part II: Bivariate Analysis and Outliers

- Bivariate Analysis techniques
- Atypical values (Outliers)

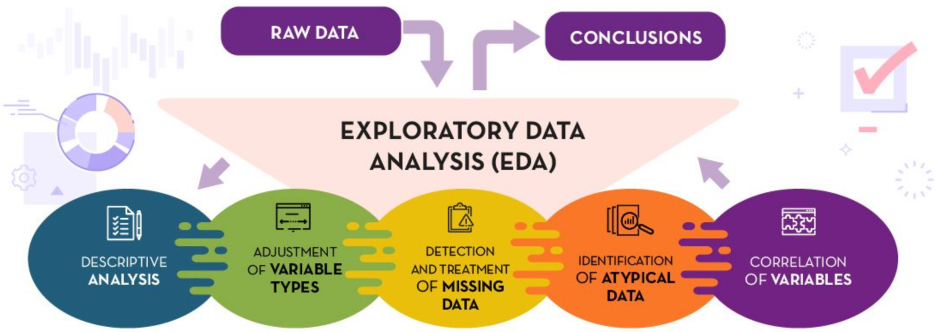
What's EDA (Exploratory Data Analysis)

Lifecycle


DATA SCIENCE LIFECYCLE
sudeep.co
01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predic-tions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

What is EDA?

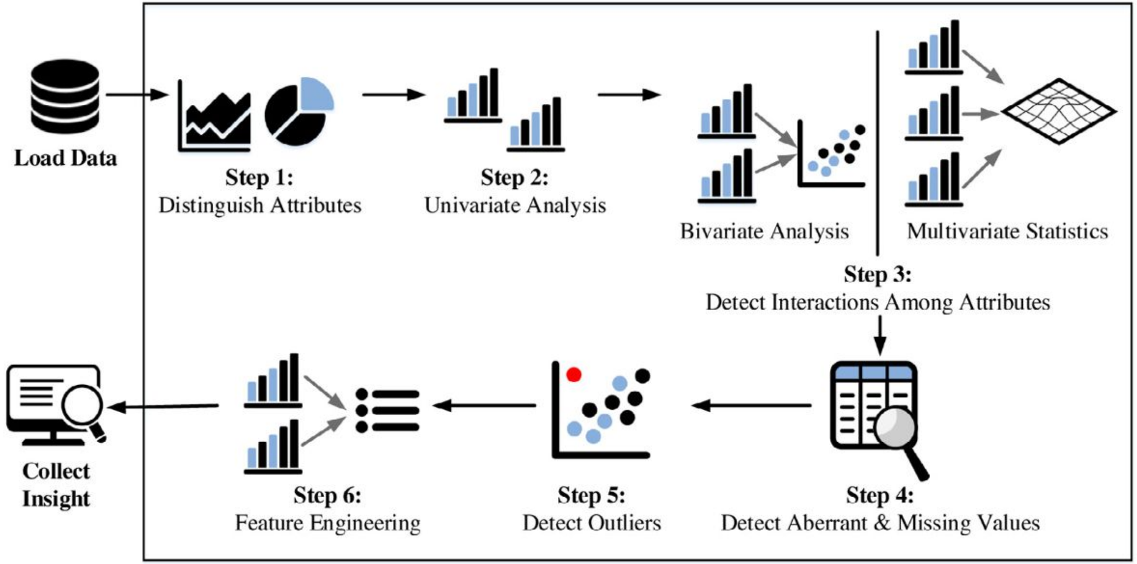
- EDA is an **iterative process** that requires a combination of **domain knowledge**, **intuition**, and **technical** skills. The objective is to **explore**, **summarize**, and **understand** the underlying **patterns**, **relationships**, **anomalies**, and **structures** in data to inform further analysis and hypothesis formulation.



What is EDA?

- EDA comprehends:
  - **Descriptive Statistics:** numerical and visual techniques that describe data
  - **Multivariate Analysis:** techniques such as **PCA**
  - **Pattern and Anomaly Detection**: this includes spotting **outliers** or unusual **clusters**.
  - **Data Cleaning**: understanding **missing** data, possible **errors**, or **inconsistencies** in the dataset. EDA often leads to data cleaning and preprocessing steps.
  - **Assumption Checking**: checks assumptions related to subsequent statistical tests or modeling. For instance, checking for **normality** or homoscedasticity.
  - **Complex Data Types Exploration**: such as **Time Series Analysis** or **Spatial Data Analysis.**
  - **Interactive Data Exploration**

What is EDA?



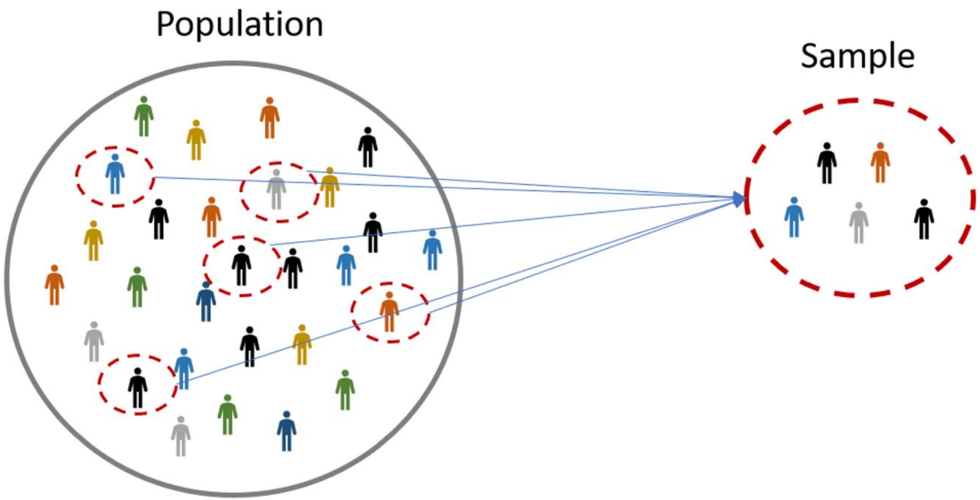
Load Data
Step 1:
Distinguish Attributes
Step 2:
Univariate Analysis
Bivariate Analysis
Multivariate Statistics
Step 3:
Detect Interactions Among Attributes
Step 4:
Detect Aberrant & Missing Values
Step 5:
Detect Outliers
Step 6:
Feature Engineering
Collect
Insight

Key Concepts

Population and Sample

- A **population** is the **entire group** or set of individuals, items, or data points of interest that one aims to study or describe.

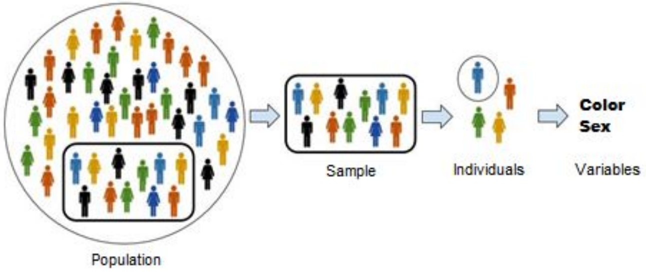
- A **sample** is a **subset** of the population selected for investigation, used to infer or make generalizations about the entire population.



Variables and Individuals

- An **individual** is a single entity or **member of a population or sample** for which data is observed or collected.

- A **variable** is a characteristic or **attribute** that can assume different values or categories across observations.



Variable

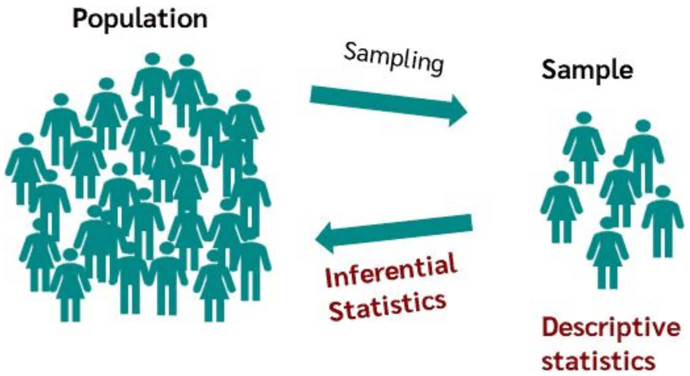
	No.	Tax Refund	Civil Status	Income (€)	Fraudster
Individual	1	Yes	Single	125K	No
	2	No	Married	100K	No
	3	No	Single	70K	No
	4	Yes	Married	120K	No
	5	No	Divorced	95K	Yes
	6	No	Married	60K	No
	7	Yes	Divorced	220K	No
	8	No	Single	85K	Yes
	9	No	Married	75K	No
	10	No	Single	90K	Yes

What is Statistics?

Statistics can be broken down into two areas:

- **Descriptive statistics**:
  - **Describes and summarizes** data.
  - Example: the average SAT score for incoming freshmen; racial makeup of the student body

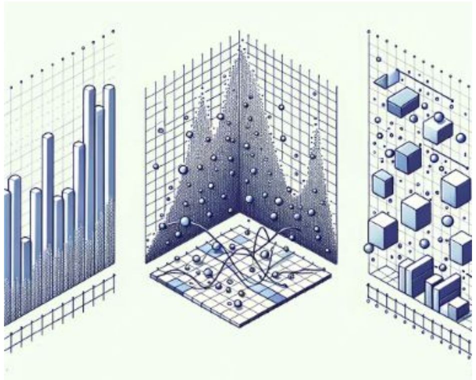
- **Inferential statistics**:
  - Makes inferences about **populations** (e.g. all universities in the country) using data drawn from **sample** data (e.g. from one university) of that population.
  - Includes hypothesis testing, confidence intervals, and regression analysis.



Descriptive Statistics

According to the **number of variables** being **analyzed simultaneously**:

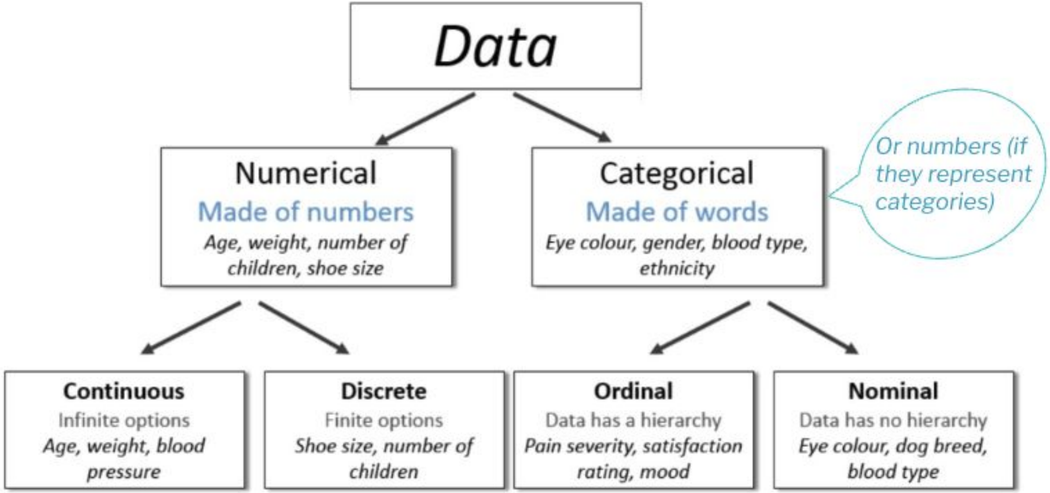
- **Univariate Analysis:**
  - Focuses on a **single variable**. The aim is to understand the **distribution, central tendency, and variability** of the data.
- **Bivariate Analysis**:
  - Examines the **relationship** or **association** between **two variables**.
- **Multivariate Analysis**:
  - Involves analyzing **three or more variables** simultaneously. The aim is often to understand the **relationships** among multiple variables or to **reduce** the number of variables.



Data Types

Data Types

- **Data** can be broadly **classified** into **different types**, and the **techniques** we apply for analysis **depend** on the **data type** at hand.

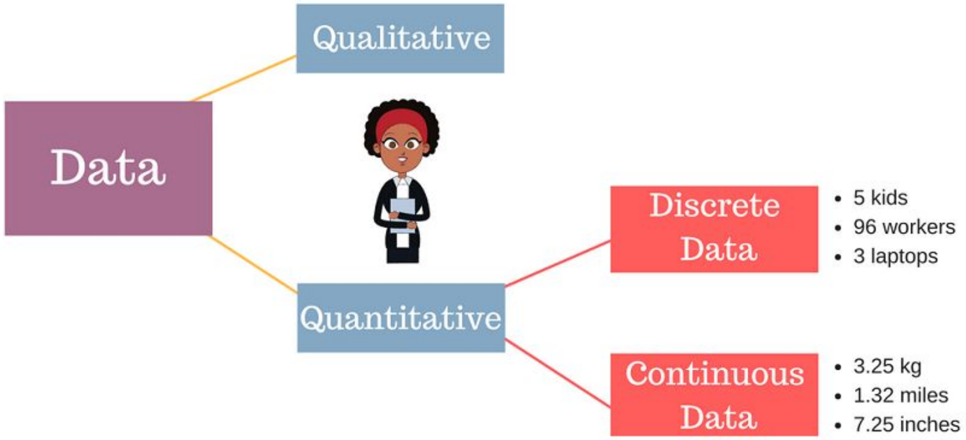


Numerical or Quantitative Data

Consists of values that can be **measured or counted**.

- **Discrete**: can only take on **specific values** within a defined range or set. These values are often **whole numbers** and **cannot** be further **subdivided**.

- **Continuous**: can take on **any value** within a **specified range**. It is not limited to whole numbers and can include decimal values.

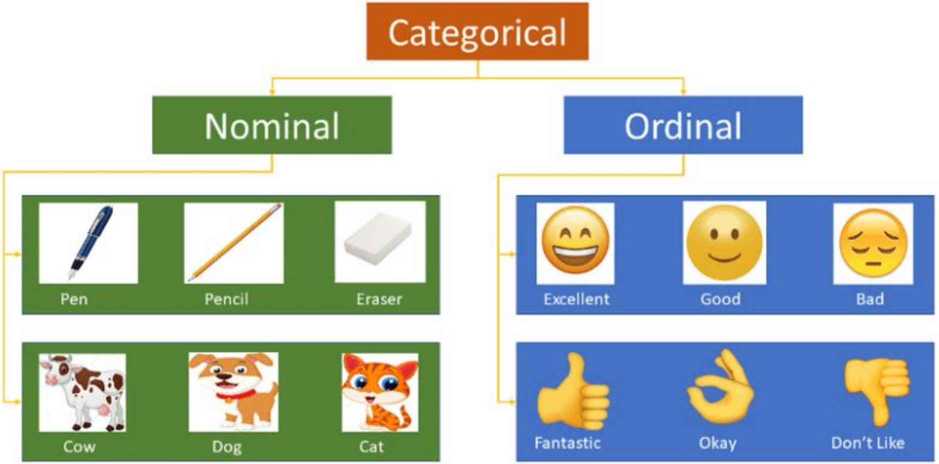


Categorical Data

Variables that are **divided into distinct categories or groups**.

- **Nominal:** have **no inherent order** or ranking. Each category is distinct and independent, without any numerical or hierarchical relationship between them.

- **Ordinal**: have a **natural order or ranking**. The categories possess a qualitative relationship of "more" or "less" compared to others but do not have a consistent or measurable difference between them.



Important Note

- In some cases, a **numerical or quantitative variable may represent a categorical or qualitative variable**.
- For example: if a dataset includes a column with numerical values representing different categories or labels, such as "0" for "male" and "1" for "female," it should be treated as a categorical variable rather than a true numerical variable.

Gender
1
0
0
1

Gender
Female
Male
Male
Female

- Always consider the **context** and **meaning** of the data when determining the appropriate data type.

Important Note: Tricks

Some tips and tricks to help you discern the nature of a numerical variable:

- **Unique Values:** If it has a **very small number** of unique values, relative to the sample size, it might be **categorical**.
- **Operational Sense**: If **performing arithmetic** operations (like addition or multiplication) on the variable **doesn't make practical sense**, it might be **categorical**. For instance, adding two ZIP codes doesn't have real-world meaning.
- **Context and Meaning:** For example, the "**number of doors in a car**" is a **discrete** numerical variable that often gets treated as **categorical** because there are a **limited number of common options** (typically 2-door, 4-door, etc.), and these **categories** have **specific implications** in terms of car type (e.g., coupe vs. sedan). The choice of how to treat it depends on the analysis goals and the nature of the data at hand:
    - **As a Numerical Variable**: If you're studying the average number of doors in cars across different years or regions, then you might treat it as numerical.
    - **As a Categorical Variable**: If you're comparing preferences, sales, or other attributes between 2-door and 4-door cars, it might make more sense to treat it as a categorical variable, because you're essentially comparing two distinct groups.

Check For Understanding: is the course number in academic settings, numerical or categorical?

Transforming Variables

Categorical to Numerical and Vice Versa

- In some cases, we need to convert between variable types. Here is why.
- **Categorical to Numerical:**
  - **Machine Learning Models:** Many algorithms require numerical input.
  - **Mathematical Operations:** To perform calculations, aggregations, or statistical tests that require numerical values.
  - **Feature Engineering:** Creating new features or leveraging patterns that emerge only when categories are numerically encoded.
- **Numerical to Categorical:**
  - **Data Binning:** Grouping a continuous variable into intervals can make analysis more intuitive and can highlight patterns better in some cases.
  - **Handle Outliers**: Transforming numerical data into categories can diminish the impact of outliers.

Transforming Variables

Numerical to Categorical

Fixed-Width Binning: Divide the range of the data into intervals of the same width. E.g., Age grouped into 0-18, 19-35, 36-60, and 60+.

Quantile Binning: Create bins such that each bin has (approximately) the same number of data points. E.g., quartiles.

Custom Binning: Define custom intervals based on domain knowledge or specific requirements. E.g., young, adult, elderly.

The resulting variable should strike a **balance** between detailed **granularity** and a **concise** overview, ensuring that the distribution's information isn't lost.

Sex	Age
male	22
female	38
female	26
female	35
male	35
male	80
male	54
male	2
female	27
female	14
female	4
female	58

Sex	Age
male	Adult
female	Adult
female	Adult
female	Adult
male	Adult
male	Elderly
male	Adult
male	Toddler/baby
female	Adult
female	Child
female	Toddler/baby
female	Adult

Transforming Variables

Categorical to Numerical

Label Encoding: Assign each category a unique number. This works well for **ordinal** data where there's a clear order.

	id	Gender	Age	Department	Rating
0	101	M	21	QA	A
1	102	M	25	QA	B
2	103	M	24	Dev	B
3	104	F	28	Dev	C
4	105	F	25	UI	B

	id	Gender	Age	Department	Rating
0	101	M	21	QA	3
1	102	M	25	QA	2
2	103	M	24	Dev	2
3	104	F	28	Dev	1
4	105	F	25	UI	2

One-Hot Encoding (Dummy Variables): For each category, create a new binary column (0 or 1). This avoids imposing an artificial order but increases data **dimensionality**.

gender	gender_m	gender_f
male	1	0
female	0	1
male	1	0
male	1	0
female	0	1
male	1	0
female	0	1
male	1	0
female	0	1

Gender
Female
Male
Male
Female

Gender
1
0
0
1

Univariate Analysis

Exploring Single Variables in Depth

Univariate Analysis

- When working with **categorical** data, we focus on understanding the **frequency counts** and **proportions** within each category.

- When working with **numerical** data, we focus on understanding the **central tendency** (measures of centrality), **the variability** (measures of dispersion) within the dataset and the **distributions** of variables.

Univariate Analysis

Categorical & Discrete Variables

Categorical & Discrete Variables

Frequency Tables

- Frequency Counts: Number of occurrences for each category.
- Frequency types:
  - **Absolute Frequency**: The number of times a particular value appears in a dataset.
  - **Relative Frequency**: The proportion or fraction of times a value occurs.
  - **Cumulative Frequency**: The sum of the absolute frequencies of all values less than or equal to the current one.
  - **Cumulative Relative Frequency:** The sum of the relative frequencies of all values less than or equal to the current one.

OS	Absolute frequency	Relative frequency
Android	230	0.46
iOS	250	0.50
Windows	15	0.03
Other	5	0.01
Total	**500**	**1**

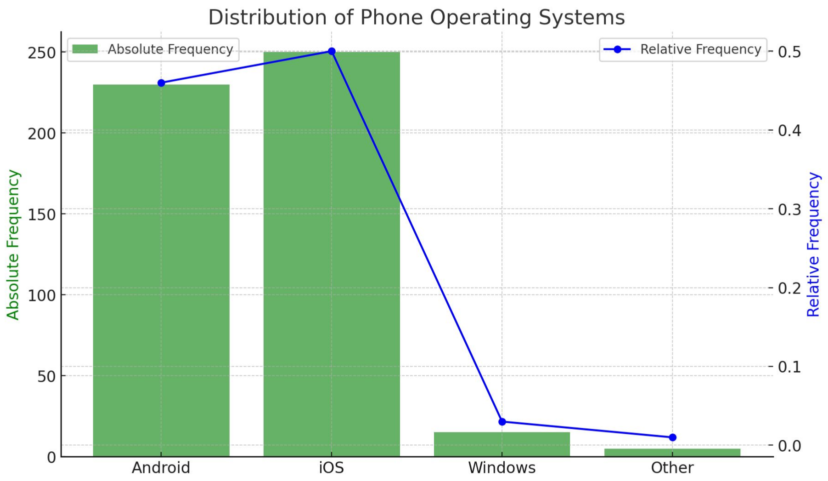
Survey of 500 users about the operating system on their primary mobile device.

Categorical & Discrete Variables

Visualization –Bar Charts

Display frequency or proportion of each category.

- The length of each bar corresponds to the quantity or magnitude of the category it represents.

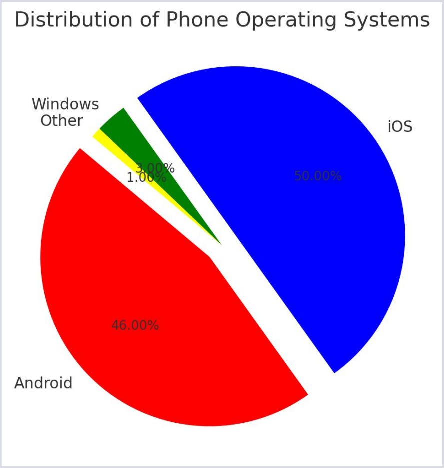


- The green bars represent the Absolute Frequency of each operating system.
- The blue line with markers represents the Relative Frequency of each operating system.

Categorical & Discrete Variables

Visualization –Pie Charts

Show proportion of each category relative to the whole. Use sparingly and **only** when categories are few.



- The chart provides a visual representation of the proportion each operating system holds relative to the whole.
- The percentages on each slice represent the relative frequencies (or proportions) of each operating system.

Univariate Analysis

Numerical Continuous Variables

Continuous Variables

Central Tendency

Measures of centrality provide insights into the **central or typical value** of a dataset.

Mean: Average of all values. *Sensitive to outliers, which can skew the mean.*

Median: Middle value when data is sorted (for even number of values, it's the average of the two middle numbers). *More resistant to outliers.*

Mode: Most frequently occurring value. Applicable to both numerical and categorical data.

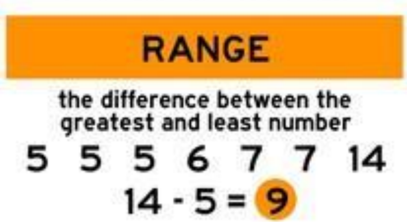
- A dataset can be unimodal (one mode), bimodal (two modes), or multimodal (more than two modes).

Continuous Variables

Measures of Spread

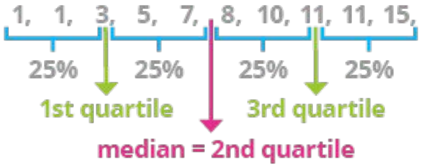
Measures of spread describe how **spread out or dispersed** a set of data is.

- **Minimum, Maximum and Range**:
  - **Range**: Difference between maximum and minimum values.
  - Very sensitive to outliers.



- **Percentiles & Quartiles**: Divide data into parts to understand distribution.
  - **Quartiles:** Quartiles split your data into four equal parts.
    - First Quartile (**Q1**): 25% of values are below this.
    - Second Quartile (**Q2**): This is the middle value, with half Below and half above (also known as the median).
    - Third Quartile (**Q3**): 75% of values are below this.

  - **Percentiles:** Same as quartiles but with any value between 1 and 99. Example: If your exam score is at the 80th percentile, it means you did better than 80% of people.

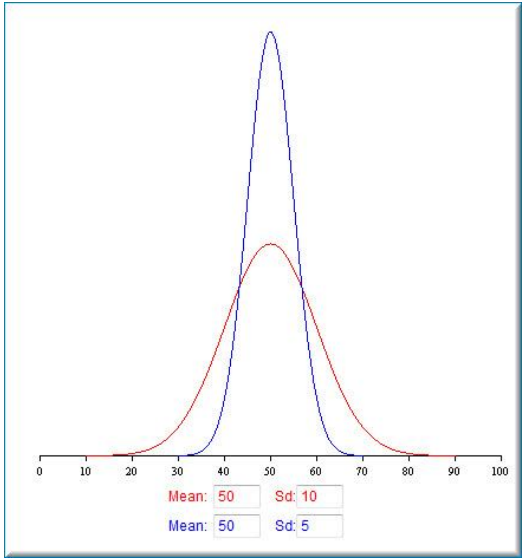


Continuous Variables

Measures of Spread

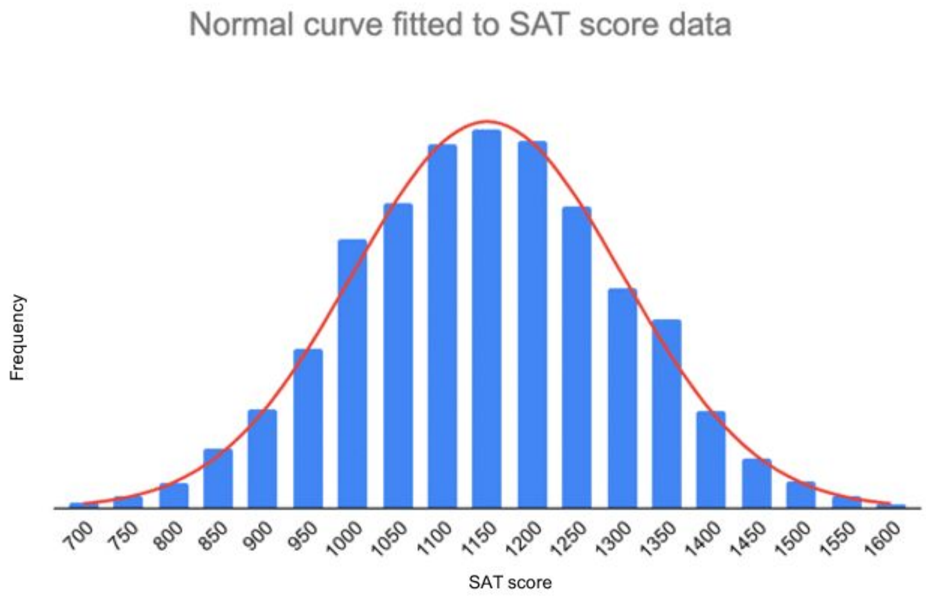
- **Variance**:
  - Average of the squared differences from the mean.
- **Standard Deviation**:
  - Square root of the variance. Indicates the spread of data around the mean.

Here we see two distributions. They both have the same mean.
The distribution that is more spread out and lower in the middle has the larger standard deviation.



Continuous Variables

Note: example of a Distribution

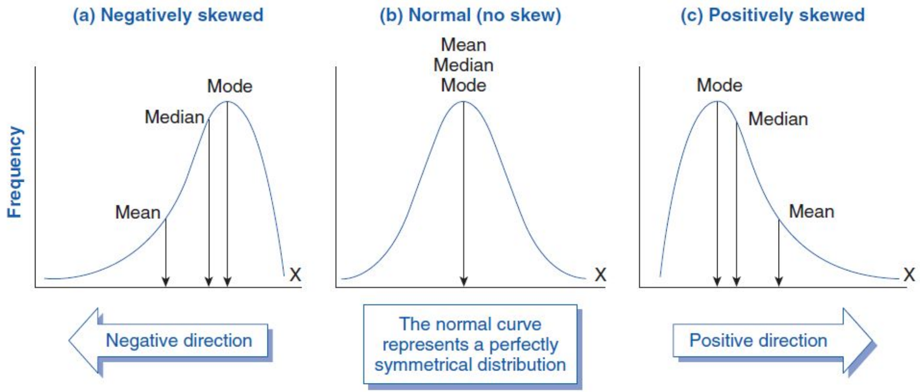


Continuous Variables

Shape of the Distribution

Measures that indicate the shape of the distribution without needing a visual display.

- **Skewness**: Measure of the **asymmetry** of the distribution.
    - **Positive skew**: Tail on the right.
      Mean > Median > Mode
    - **Negative skew**: Tail on the left.
      Mode > Median > Mean

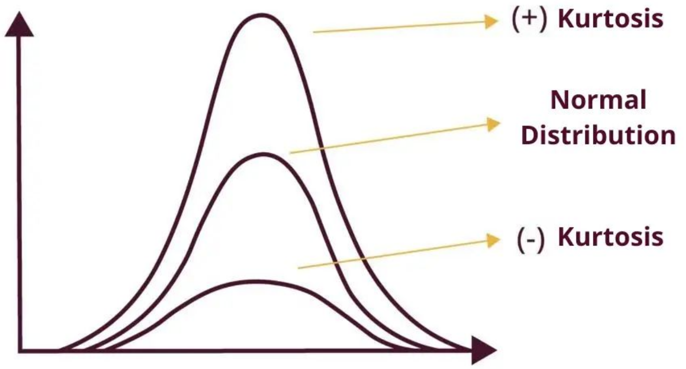


Continuous Variables

Shape of the Distribution

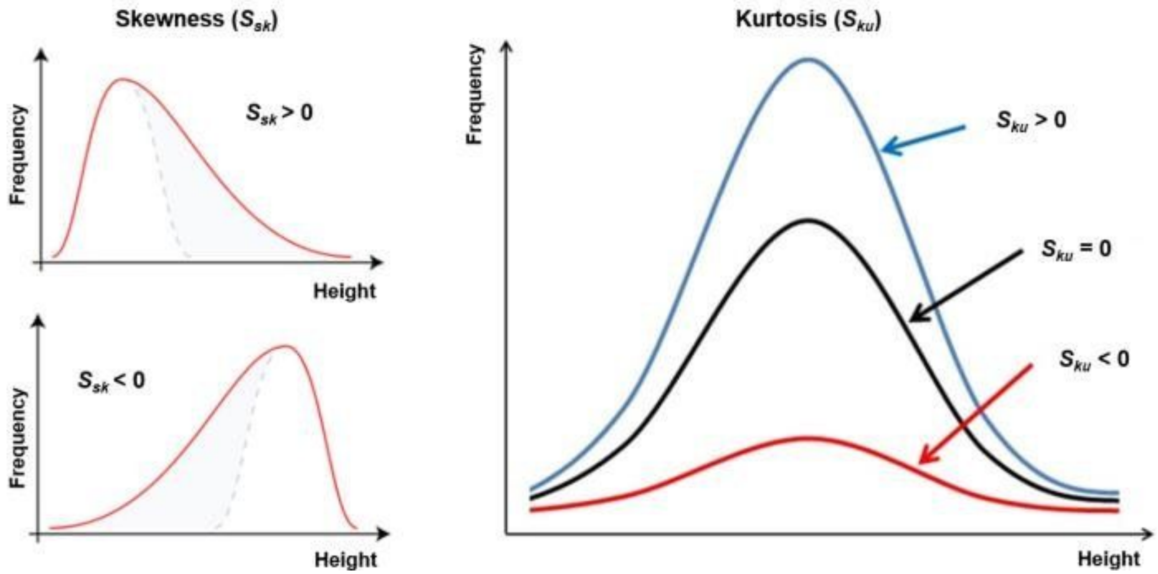
Measures that indicate the shape of the distribution without needing a visual display.

- **Kurtosis**: Measure of the "**tailedness**" of the distribution.
    - **High kurtosis**: Heavy tails (more disperse), more outliers.
    - **Low kurtosis**: Light tails, fewer outliers.
    - **Zero or Moderate Kurtosis**: The distribution has a shape relatively equivalent to a normal distribution.



Continuous Variables

Shape of the Distribution - Comparison

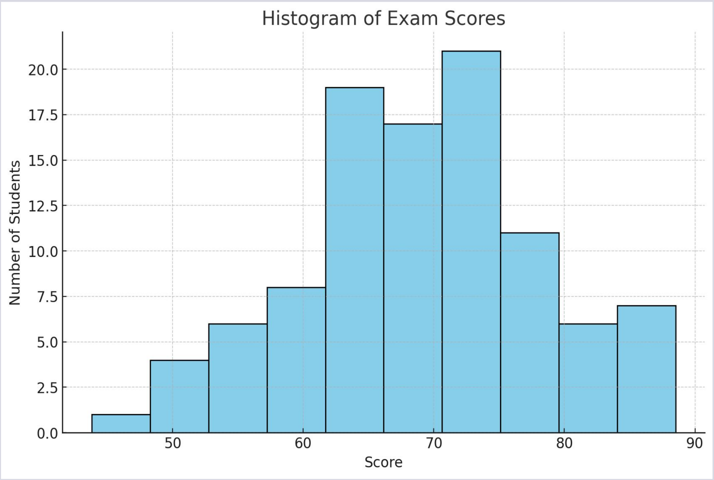


Continuous Variables

Visualization - Histograms

An histogram is a graphical representation of the distribution of a dataset.

- The data range is divided into intervals **(bins)**.
- The **width** of the **bin** affects the histogram:
    - **Narrow** bins might show too much **noise**
    - **Wide** bins might **hide** important details.
- **Difference from Bar Chart**: In histograms, bars are adjacent with no gap between them, while bar charts (for categorical and discrete) have distinct bars separated by spaces.



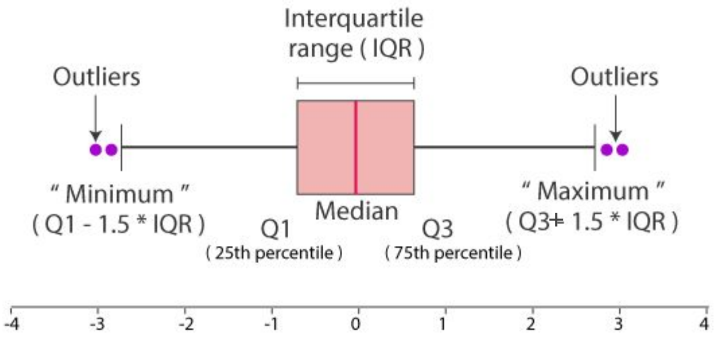
Histogram illustrating the distribution of exam scores for 100 students

Continuous Variables

Visualization - Box Plots

Box plots provide a visual summary of the data's **distribution**, including its **central value**, **variability**, and any potential **outliers**.
They are especially useful for **comparing distributions across different groups** (displayed in parallel).

- **Box**: spans from Q1 to Q3. This represents the Interquartile Range (IQR = Q3 - Q1), where the middle 50% of the data lies.
- **Whiskers**: lines that extend from both ends of the box (Q3 1.5 *±IQR).
- **Outliers**: Data points that fall outside the whiskers' range.



Different parts of boxplot

Summary

Summary

- **Univariate Analysis:** Focuses on a **single variable**
    - **Categorical** variables:
        - **Frequency** tables. Counts and proportions.
        - Visualizations: **Bar charts, pie charts**
    - **Numerical** variables:
        - Measures of centrality:
            - **Mean, median, mode**
        - Measures of dispersion:
            - **Variance, standard deviation, minimum, maximum, range, quantiles**
        - Shape of the Distribution:
            - **Symmetry and kurtosis**
        - Visualizations: **Histograms, box plots**