Chapter 1: Introduction

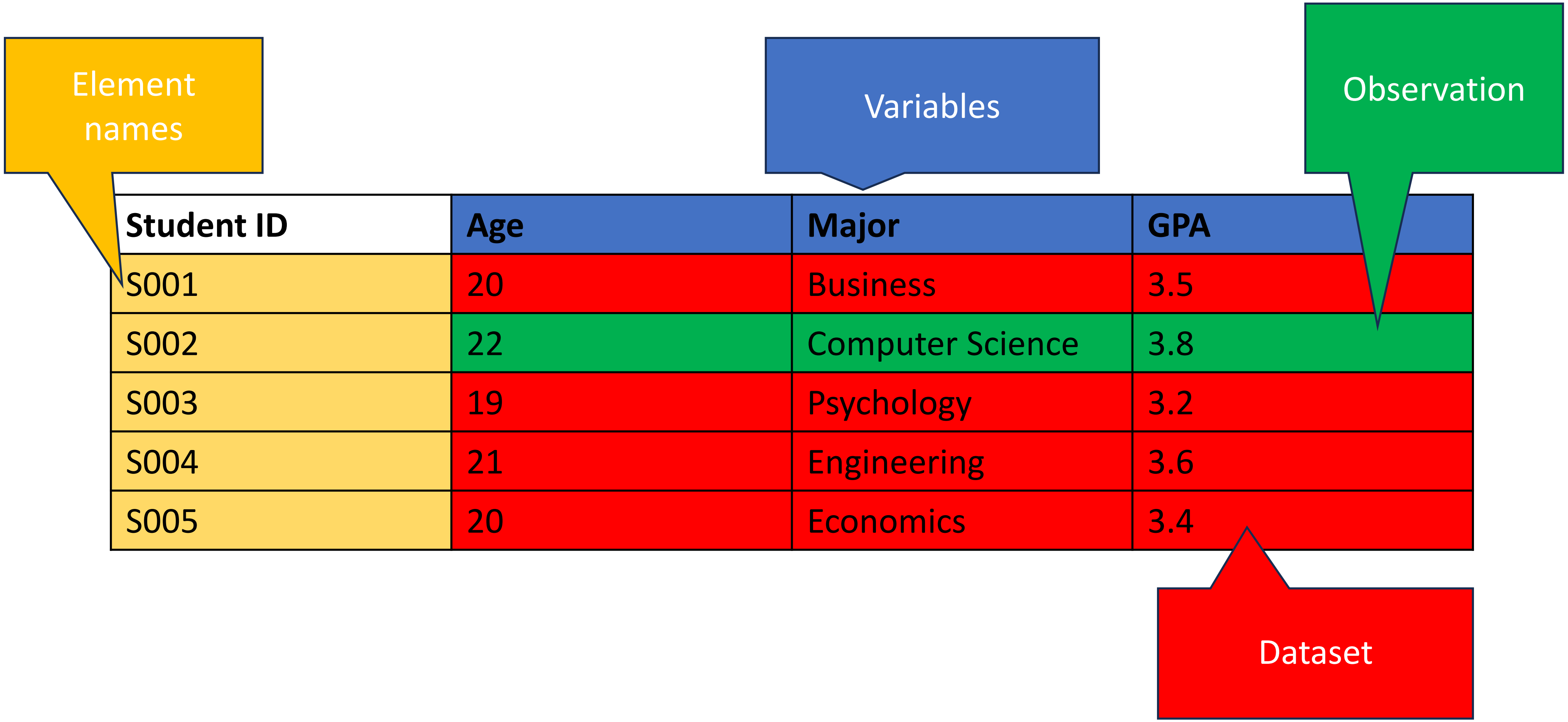
Agenda

1. Data Overview
   - What is data?
   - Types of Data: Qualitative and Quantitative
   - Measurement Scales
   - Data Sources: Primary and secondary
   - Post-Data Collection Procedures
2. Descriptive Statistics
   - Types of data: Cross-sectional, Time series and Panel
   - Tables for appropriate data representation by type
   - Descriptive statistics for each type of data
3. Applications
4. Review on statistics in business

What is Data?

- Data: set of values such as numbers, text, or symbols collected for storage, summary, analysis and decision-making.
- Dataset: collection of data collected in a particular study.
- Elements: entities on which data are collected.
- Variable: characteristic of interest for the elements.
- Observation: the set of measurements obtained for a particular element.
- A dataset with ***n*** elements contains ***n*** observations.
- The total number of data values in a complete dataset is the number of elements multiplied by the number of variables.

What is Data?



Element names

Variables

Observation

Student ID	Age	Major	GPA
S001	20	Business	3.5
S002	22	Computer Science	3.8
S003	19	Psychology	3.2
S004	21	Engineering	3.6
S005	20	Economics	3.4

Dataset

Types of Data

- Qualitative:
  - Labels or names used to identify an attribute of each element
  - Can also be called categorical data
  - Use either the nominal or ordinal scale of measurement
  - Can be either numerical or non-numerical
- Quantitative:
  - Indicate how many or how much
  - Is always numerical
  - Ordinary arithmetic operations (mean, median, etc.) are meaningful for this type of data

Quantitative Data Types

- Discrete: Countable values
  - Example: number of students in this classroom, the set of all positive integers
- Continuous: Infinite values within a range, cannot be counted
  - Example: temperature, time, weight, height

Example: Qualitative Data

Student ID	Favorite Color	Preferred Learning Style	Country of Origin
001	Blue	Visual	Canada
002	Green	Auditory	Brazil
003	Red	Kinesthetic	India
004	Yellow	Visual	Australia

Example: Quantitative Data

Student ID	Age	Test Score (%)	Hours Studied
001	20	88	10
002	22	75	7
003	19	92	12
004	21	81	9

Measurement Scales

- How variables are defined and categorized
- Describes the nature of data in the dataset
- Determines the type of analysis can be used for the data of interest
- There are 4 types:
  - Nominal scale
  - Ordinal scale
  - Interval scale
  - Ratio scale

Nominal Scale

- Classifies data into categories
- No intrinsic ranking or order within categories
- Can only be compared for equality (only A = B or A ≠ B)
- Example: Blood type (A, B, O, or AB), gender, nationality

Ordinal Scale

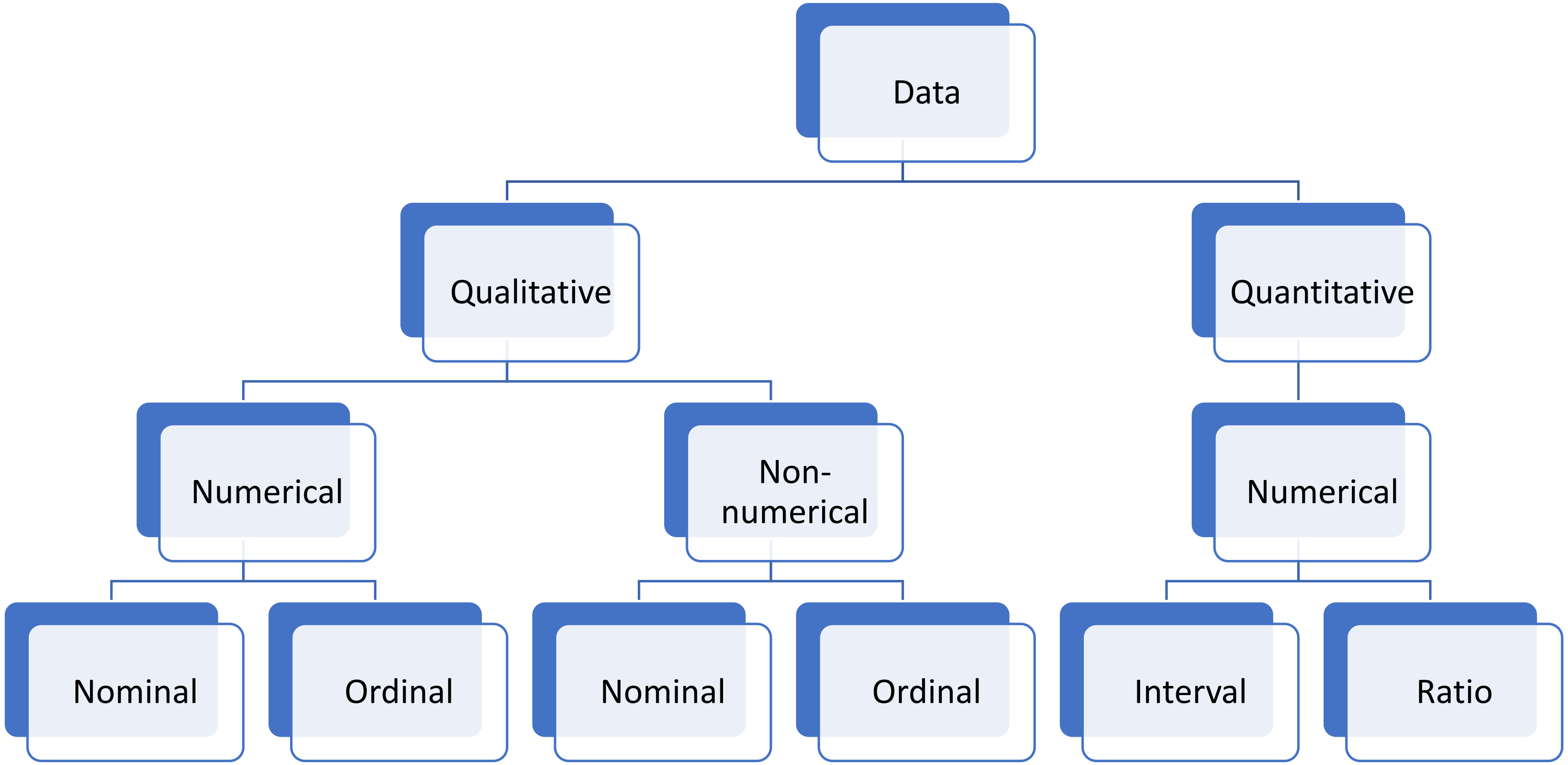
- Classifies data into categories
- There is ranking and order within categories
- The interval between ranking is not equal
- Example: Education level, rankings in a competition

Interval Scale

- Equal intervals between values
- No true zero point
- Addition and subtraction can be applied but not multiplication and division
- Example: Temperature in Celsius or Fahrenheit, IQ scores

Ratio Scale

- Equal intervals between values
- True zero point (point where 0 means “none”)
- Supports addition, subtraction, multiplication, and division
- Ratio between values has meaning
- Example: Height, age, distance



Data

Qualitative

Quantitative

Numerical

Non-numerical

Numerical

Nominal

Ordinal

Nominal

Ordinal

Interval

Ratio

Operation	Nominal Scale	Ordinal Scale	Interval Scale	Ratio Scale
Equality Comparison	✓	✓	✓	✓
Order Comparison		✓	✓	✓
Addition / Subtraction			✓	✓
Multiplication / Division				✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Other Statistical Operations				✓

Data Sources

- Data can be primary or secondary based on collection method.
- Primary data:
    - Data collected directly from original sources through methods such as surveys, interviews, experiments, or observations, for the specific purpose of a current research project.
    - Original and first-hand information
    - Tailored to the researcher's needs
    - Time-consuming and costly to gather
    - High accuracy and relevance

Data Sources

- Secondary data:
  - Data that has been previously gathered and made available by other sources, such as government agencies, organizations, or researchers, and is reused for a different analysis or purpose
  - Quick and cost-effective to obtain
  - May not perfectly match current research needs
  - Data quality varies depending on the original source
  - Example: Government data, academic journals, online databases

Data Sources

Aspect	**Primary Data**	**Secondary Data**
Definition	Data collected firsthand for a specific research purpose	Data previously collected by others for a different purpose
Source	Original sources such as surveys, interviews, observations	Existing sources such as books, reports, websites, and databases
Collection Method	Direct methods: surveys, experiments, observations	Indirect methods: accessing published data or reports
Time Required	Time-consuming	Time-saving
Cost	Expensive (requires resources and effort)	Usually low-cost or free
Relevance to Research	Specifically tailored to current research needs	May not fully match current research objectives
Data Accuracy	High (if properly collected)	Depends on the credibility of the original source
Control Over Data	Full control over how data is collected	No control over how data was originally collected
Examples	Field surveys, experiments, direct interviews	Government reports, academic articles, statistical databases

Post-Data Collection Procedures

- Data preparation
    - Data cleaning: identify and correct errors, missing values, and inconsistencies
    - Data coding: Convert qualitative data into numerical or symbolic codes
    - Data validation: Confirm the consistency, completeness, and accuracy of data
- Data analysis and interpretation
    - Data transformation: format data for analysis
    - Data analysis: apply statistical or thematics methods to the data
    - Data interpretation: extract findings and insights from the results of the analysis
- Documentation and storage
    - Report writing: present your findings in a structured report/presentation
    - Data storage and backup

Types of data

- Based on time dimension and number of units, data can be categorized into the following:
    - Cross-sectional data: Data collected at a single point in time from multiple units.
    - Time series data: Data collected over time from a single unit.
    - Panel data: Data collected over time from multiple units.

Cross-sectional data

- Data collected at a single point in time (or over a very short period) across multiple subjects.
- Captures a snapshot of a population or phenomenon.
- Allows for comparison between different subjects at one moment.
- Commonly used in surveys, market research, and census studies.

Cross-sectional data

- Data on 4 households with the number of motorbikes of each households and the monthly electricity bill for the month of June 2025.

Household ID	Number of Motorbikes	Monthly Electricity Bill (VND)
H001	2	1,050,000
H002	1	870,000
H003	3	1,200,000
H004	2	950,000

Time Series Data

- Data collected on a single subject over multiple time periods.
- Focuses on how value changes over time.
- Each observation is tied to a timestamp.
- Can be used for trend, seasonality, and forecasting analysis.

Time Series Data

- Time series can be broken down into 4 components:
  - Trend: the general direction or pattern of change of the data, whether it be decreasing, increasing, or stable over a period of time. For example, inflation and price are generally trending up (increasing)
  - Seasonality: the repeating pattern that occur at predictable intervals, usually tied to seasons or calendar. Example: tourism increase in holiday season, electricity use increase in the summer.
  - Cyclic component: long-term, non-seasonal changes in a time series that occur in irregular but recurring cycles. Example: Economy cycle (Expansion → Peak → Recession → Trough), real estate pricing cycle (price booms followed by downturns).
  - Irregular remainder (white noise): the random, unpredictable fluctuations in a time series that cannot be explained by the trend, seasonality, or cyclic patterns. Example: natural disasters, unexpected economic shock created by new policies.

Time Series Data

Component	Description	Example
Trend (T)	Long-term movement or direction in the data over time	Gradual increase in average temperature over decades
Seasonality (S)	Repeating short-term pattern at regular intervals (e.g., monthly, yearly)	Higher tourism in holiday season
Cyclic Component (C)	Long-term up-and-down movements not of fixed period (often economic cycles)	Business cycles: boom → recession → recovery
Irregular/Random (I)	Unpredictable, short-term fluctuations due to random or unusual events	Sudden drop in stock price due to new policies

Time Series Data

- Electricity bill of a household in 2024

Month	Electricity Bill (VND)
Jan 2024	950,000
Feb 2024	870,000
Mar 2024	1,050,000
Apr 2024	1,200,000
May 2024	1,100,000

Panel Data

- Can also be called Longitudinal data.
- Data that observes multiple subjects over multiple time periods.
- Combines the elements of both cross-sectional and time series data.
- Enable analysis for changes within each subjects as well as differences between subjects.

Panel Data

- Data on 3 households, the number of motorbikes in each household and their monthly electricity bill in June and July 2025

Household ID	Month	Number of Motorbikes	Monthly Electricity Bill (VND)
H001	Jan 2025	2	950,000
H001	Feb 2025	2	870,000
H002	Jan 2025	0	1,000,000
H002	Feb 2025	1	880,000
H003	Jan 2025	3	1,200,000
H003	Feb 2025	2	1,250,000

Aspect	**Cross-Sectional Data**	**Time Series Data**	**Panel Data** (Longitudinal)
Definition	Data collected at a **single point in time** from **multiple units**	Data collected from a **single unit** over **multiple time periods**	Data collected from **multiple units** over **multiple time periods**
Time Component	❌ No	✅ Yes	✅ Yes
Observation Units	Many units (e.g., people, firms) at one time	One unit over time	Many units over time
Focus	Differences **between subjects**	Trends and patterns **over time**	Changes **within and between subjects over time**
Use Cases	Market research, opinion polls, census	Forecasting, trend analysis	Economic, behavioral, and policy studies
Data Structure	One-time snapshot	Time-based sequence	Multi-dimensional: units × time

Tables and charts for appropriate data representation by type

- Below is dataset of 10 students, with 5 variables: gender, satisfaction level, IQ score, weekly study hours, and grade.

ID	Name	Gender (Nominal)	Satisfaction Level (Ordinal: 1=Very Dissatisfied, 5=Very Satisfied)	IQ Score (Interval)	Weekly Study Hours (Ratio, Continuous)	Grade (GPA, Ratio, Continuous)
1	Alice	Female	4	112	12.5	3.4
2	Ben	Male	3	105	8.0	2.8
3	Gina	Female	5	120	18.0	3.9
4	Dan	Male	2	101	6.0	2.5
5	Cara	Female	5	118	15.0	3.8
6	Ivy	Female	2	102	7.0	2.7
7	Hugo	Male	3	107	9.5	3.0
8	Eva	Female	4	109	10.0	3.2
9	Jack	Male	4	115	14.0	3.6
10	Finn	Female	1	98	5.5	2.4

Tables and charts for appropriate data representation by type

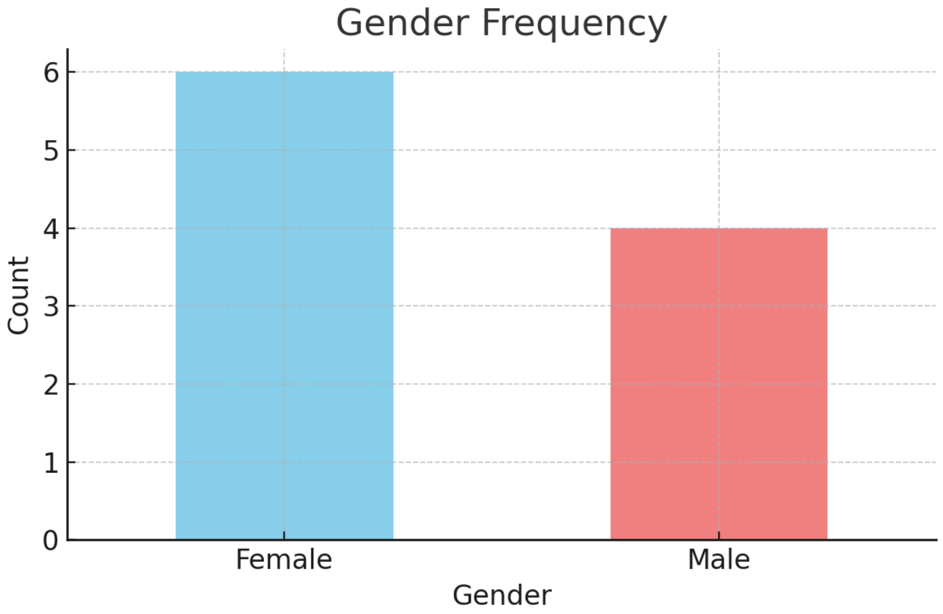
- For Qualitative (Categorical) data:
  - Nominal data:
    - Frequency table: presents the number of occurrences (counts) of each category within a dataset, helps summarize large amounts of categorical data into a readable format.

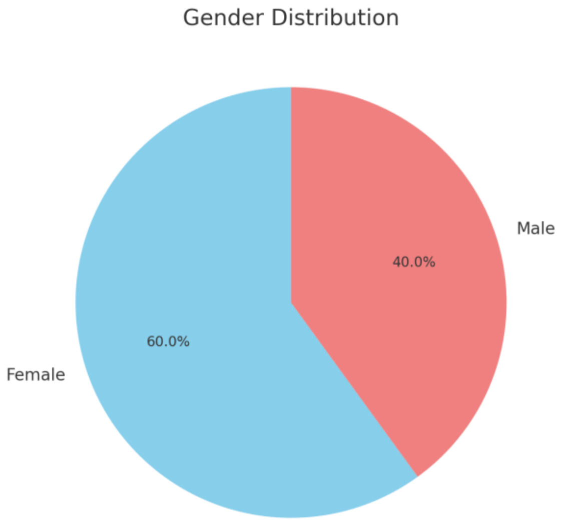
Gender	Frequency
Female	6
Male	4

Frequency Table

Tables and charts for appropriate data representation by type

- For Qualitative (Categorical) data:
  - Nominal data:
    - Bar chart: represents categorical data with rectangular bars. The length of each bar corresponds to the frequency or proportion of that category.
    - Pie chart: shows the proportion of each category as slices of a circle, useful for displaying part-to-whole relationships.





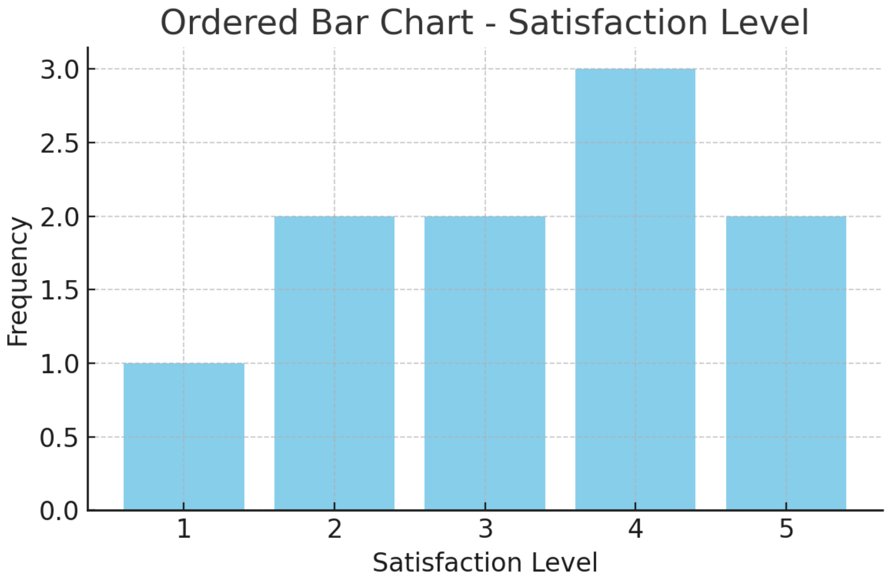
Tables and charts for appropriate data representation by type

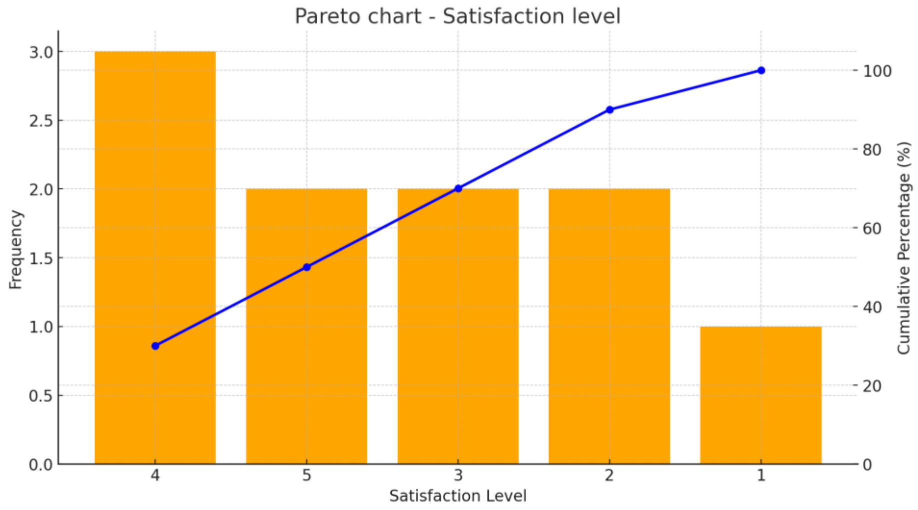
- For Qualitative (Categorical) data:
  - Ordinal data:
    - Ordered Frequency table: frequency table where categories are arranged based on their inherent order. It helps in understanding trends or patterns across ranked groups.
    - Cumulative Frequency Table: shows the running total of frequencies up to each category. It's useful for understanding how values accumulate across ordered categories.

Satisfaction Level	Frequency	Cumulative Frequency
1	1	1
2	2	3
3	2	5
4	3	8
5	2	10

Tables and charts for appropriate data representation by type

- For Qualitative (Categorical) data:
  - Ordinal data:
    - Ordered Bar Chart: arranges bars in a logical sequence based on category order. This enhances the readability of ordinal data trends.
    - Pareto Chart: bar chart sorted in descending order of frequency, often paired with a cumulative line. It helps identify the most significant categories contributing to a dataset.





Tables and charts for appropriate data representation by type

- For Quantitative data:
  - Discrete data:
    - Frequency Table: table for discrete data lists each unique value and its corresponding count. It’s suitable when the number of values is small and easily distinguishable.

Weekly Hours Worked	Frequency
20	1
28	1
32	1
36	1
40	1
42	1
45	1
50	1
Total	**8**

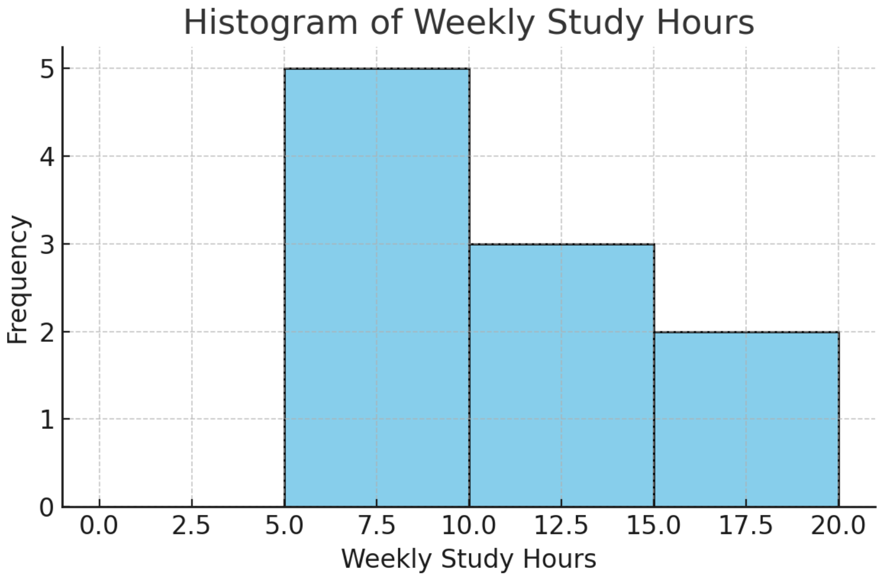
Tables and charts for appropriate data representation by type

- For Quantitative data:
  - Discrete data:
    - Grouped Frequency Table: when discrete data includes a wide range of values, similar values are grouped into intervals to simplify analysis. This is especially helpful for large datasets.

Weekly Study Hours Range	Frequency
0–5	0
6–10	6
11–15	3
16–20	1

Tables and charts for appropriate data representation by type

- For Quantitative data:
  - Discrete data:
    - Bar Chart: work well for discrete data with a limited number of values. Each bar represents a specific value and its frequency.
    - Histogram (with gaps): a histogram with visible gaps between bars is sometimes used for discrete data to emphasize the discrete nature of the values.



Tables and charts for appropriate data representation by type

- For Quantitative data:
  - Continuous data:
    - Grouped Frequency Table with Class Intervals: data is grouped into intervals (called "classes") to summarize distributions. Each interval represents a range of values, and frequencies show how many data points fall into each range.

Class Interval	Frequency
5-9	1
10-14	2
15-19	3
20-24	2
25-29	1
30-34	1

Tables and charts for appropriate data representation by type

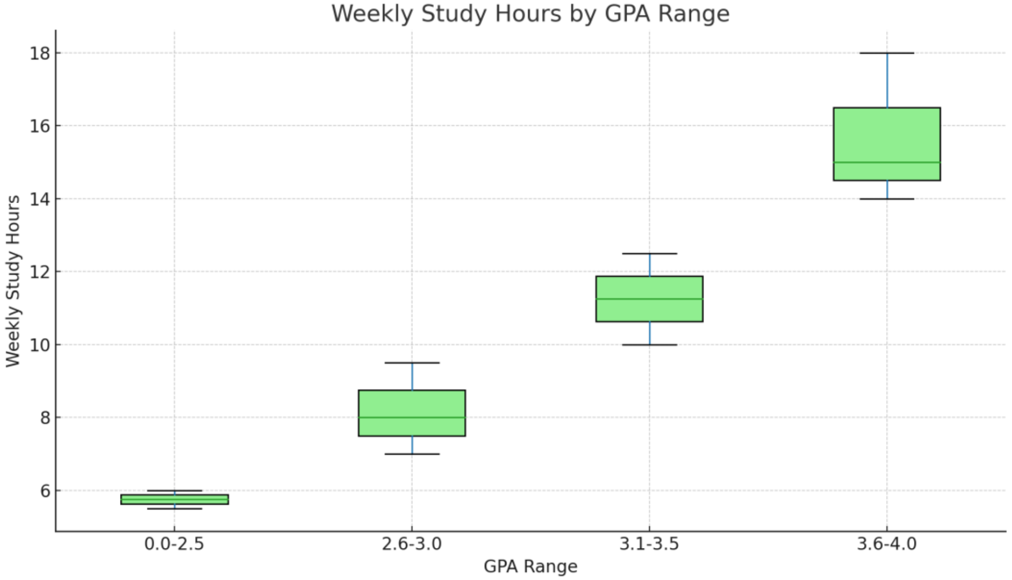
- For Quantitative data:
  - Continuous data:
    - Relative Frequency Table: shows the proportion or percentage of data points in each category or class interval, rather than raw counts, useful for comparing distributions across different sample sizes.

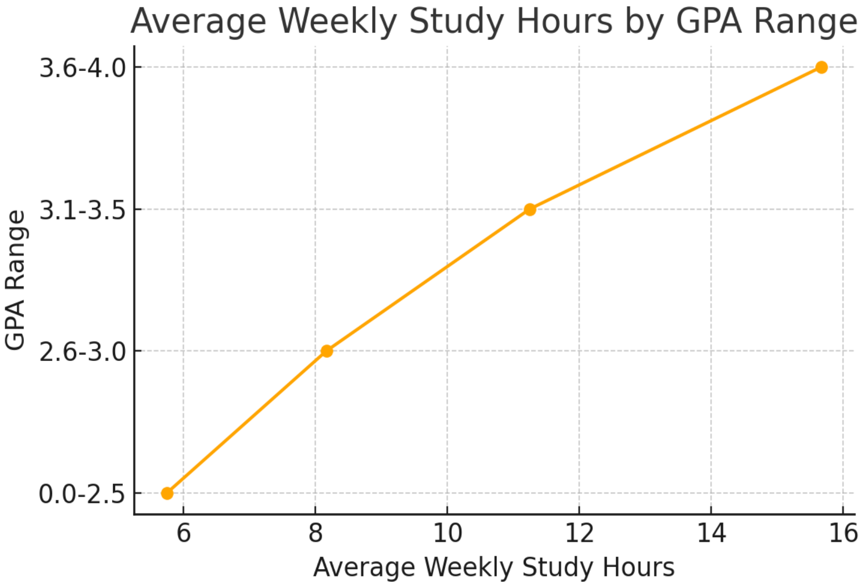
GPA Range	Relative Frequency
0.0–2.5	0.2
2.6–3.0	0.3
3.1–3.5	0.2
3.6–4.0	0.3

Tables and charts for appropriate data representation by type

- For Quantitative data:
  - Continuous data:
    - Histogram: displays the frequency distribution of continuous data using adjacent bars. Unlike bar charts, the bars touch to represent the continuous nature of the data.
    - Box Plot (Box-and-Whisker Plot): visually summarizes the distribution of continuous data using five key values: minimum, lower quartile, median, upper quartile, and maximum. It also highlights potential outliers.
    - Line Chart: connects data points with lines, making it ideal for tracking trends over time in continuous variables.

Tables and charts for appropriate data representation by type





Tables and charts for appropriate data representation by type

- In some cases, we may be interested in more than 1 variable, the data that look at more than 1 variable is called Bivariate data (for 2 variables) or Multivariate data (for multiple variables). This type data is useful for investigation of the relationships, correlations, or comparisons among multiple variables.

Tables and charts for appropriate data representation by type

- Contingency Table: summarizes the relationship between two categorical variables. It allows analysis of possible associations or dependencies.

Gender	0.0–2.5	2.6–3.0	3.1–3.5	3.6–4.0
Female	2	2	2	0
Male	0	1	2	1

Contingency Table (Gender × GPA range)

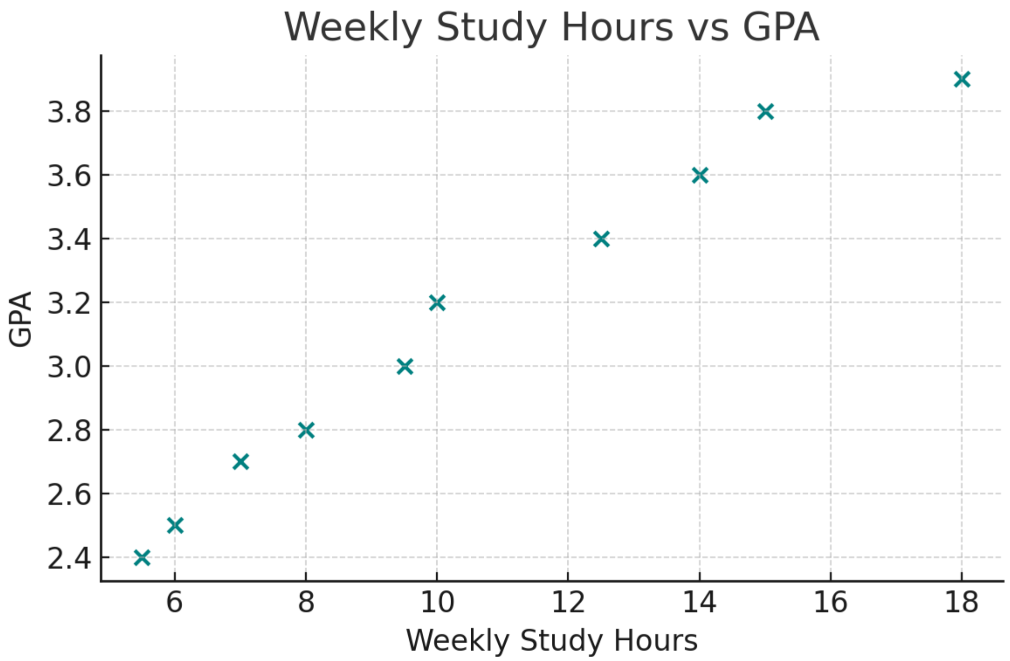
Tables and charts for appropriate data representation by type

- Summary Statistics Table: includes descriptive statistics such as mean, median, mode, and standard deviation for numerical variables, often broken down by categories (e.g., average income by education level).

Statistic	IQ Score	Weekly Study Hours	GPA
Count	10.0	10.0	10.0
Mean	106.4	10.3	3.02
Std	6.84	3.91	0.46
Min	96.0	5.0	2.3
25%	101.25	7.0	2.75
50%	105.0	10.5	3.05
75%	111.25	13.0	3.35
Max	118.0	17.0	3.8

Tables and charts for appropriate data representation by type

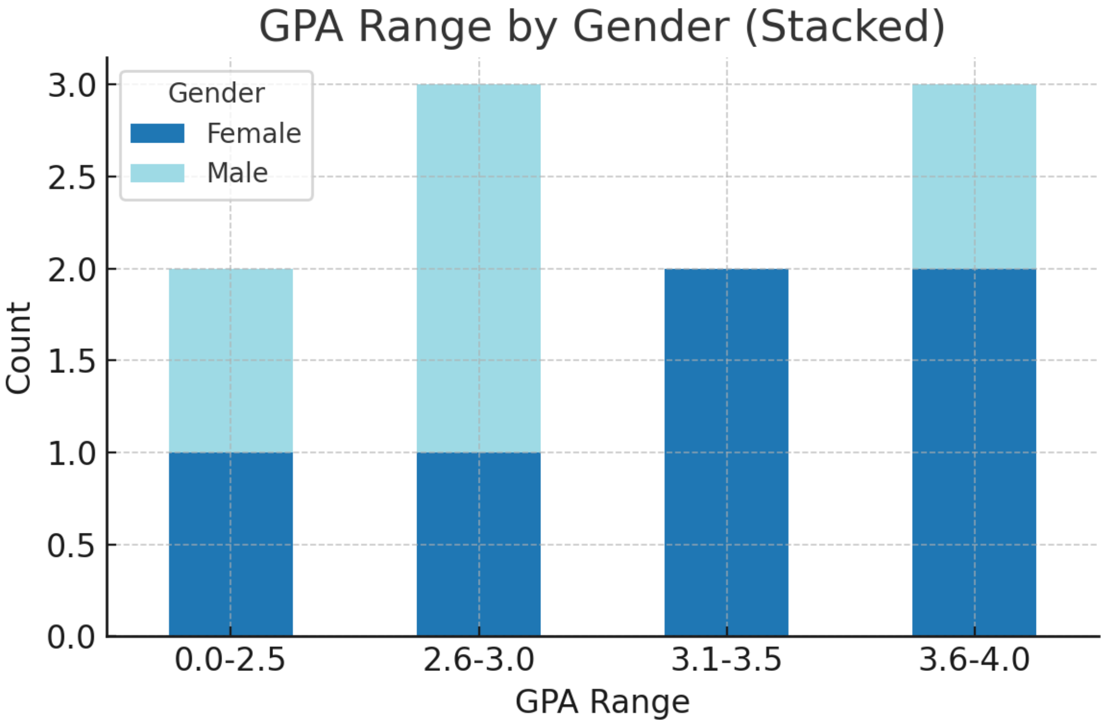
- Scatter Plot: displays the relationship between two numerical variables. Each point represents one observation, making it useful for spotting trends, patterns, or correlations.

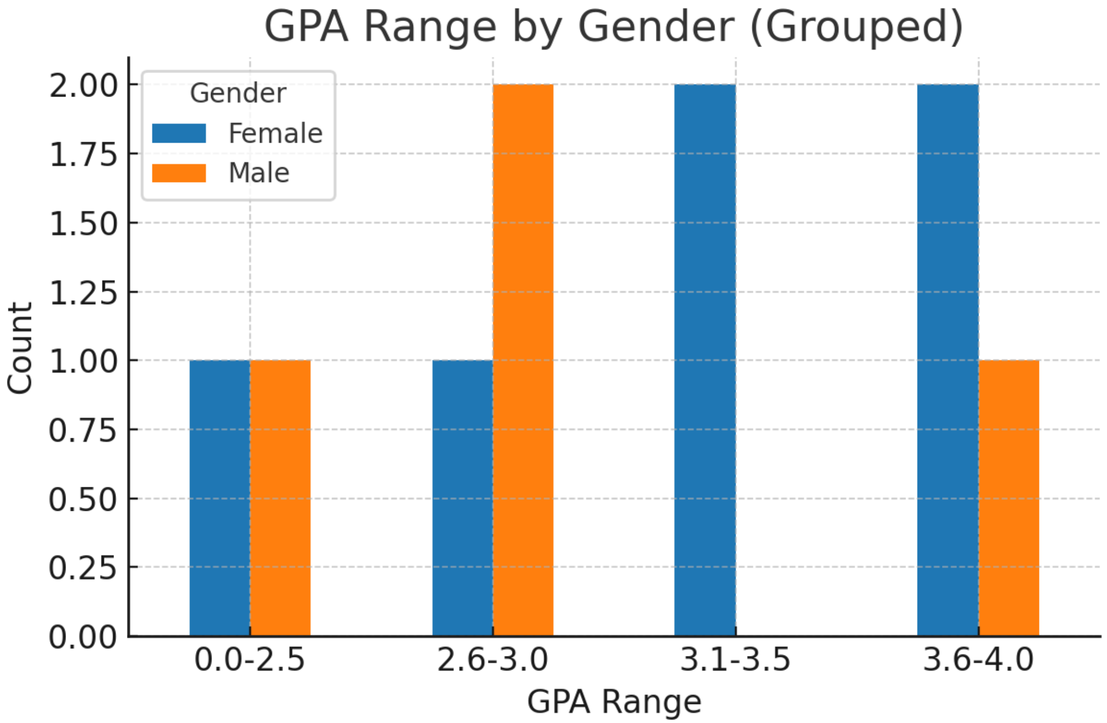


Tables and charts for appropriate data representation by type

- Stacked Bar Chart: shows how individual parts contribute to the whole across categories. It’s useful for comparing distributions within a group while showing total values.
- Grouped Bar Chart: compares categories across multiple groups (e.g., test scores by gender and grade level). Each group of bars represents a different subgroup.

Tables and charts for appropriate data representation by type


GPA Range by Gender (Stacked)


GPA Range by Gender (Grouped)

Descriptive statistics

- Descriptive statistics are used to summarize and present data so that patterns and key features become easier to understand.
- Methods involve numerical measures such as mean, median, mode, range, and standard deviation.
- Descriptive statistics describe the dataset at hand without making predictions or generalizations about a larger population.
- Descriptive statistics can be divided into 4 subgroups:
  - Measures of Central Tendency
  - Measures of Dispersion
  - Measures of Position
  - Shape Descriptors

Descriptive statistics

- Measures of Central Tendency: describe the “center” or the typical value of the dataset
  - Mean: sum of all values divided by the number of observations.
    - Represents the balance point of the distribution of the data
    - Sensitive to outliers (extreme values)
    - Formula: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
    - Where x_i = each data point, and n is the number of observations
  - Median: the middle value when data is arranged in ascending order. If there is an even number of observations, it is the average of the two middle values

Descriptive statistics

- Measures of Central Tendency: describe the “center” or the typical value of the dataset
  - Median: the middle value when data is arranged in ascending order. If there is an even number of observations, it is the average of the two middle values.
    - Less affected by outliers.
    - Formula:
      - If n is odd: $Median = x_{\left(\frac{n+1}{2}\right)}$
      - If n is even:$Median = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+1}{2}\right)}}{2}$
  - Mode: the value of category that appears most frequently in the dataset, a dataset can have no mode, one mode (unimodal), or multiple modes (bimodal/multimodal).

Descriptive statistics

- Measures of Dispersion (Variability): describe how spread out the data is.
  - Range: the difference between the largest and the smallest value in a dataset. Highly sensitive to outliers.
    - Formula: $Range = x_{max} - x_{min}$
  - Variance: measures the average squared deviation of each value from the mean. Higher variance indicates that data point are more spread out from the mean.
    - Formula: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Descriptive statistics

- Measures of Dispersion (Variability): describe how spread out the data is.
  - Standard deviation: the square root of variance, measure the spread of the data from the mean in the same units as the data. Clearer sense of deviation from the mean.
    - Formula: $\mathrm{s} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
  - Interquartile range (IQR): the difference between the third quartile (Q3) and the first quartile (Q1). Measures the spread of the middle 50% of the data, resistant to outliers.
    - Formula: $\mathrm{IQR} = Q_3 - Q_1$

Descriptive statistics

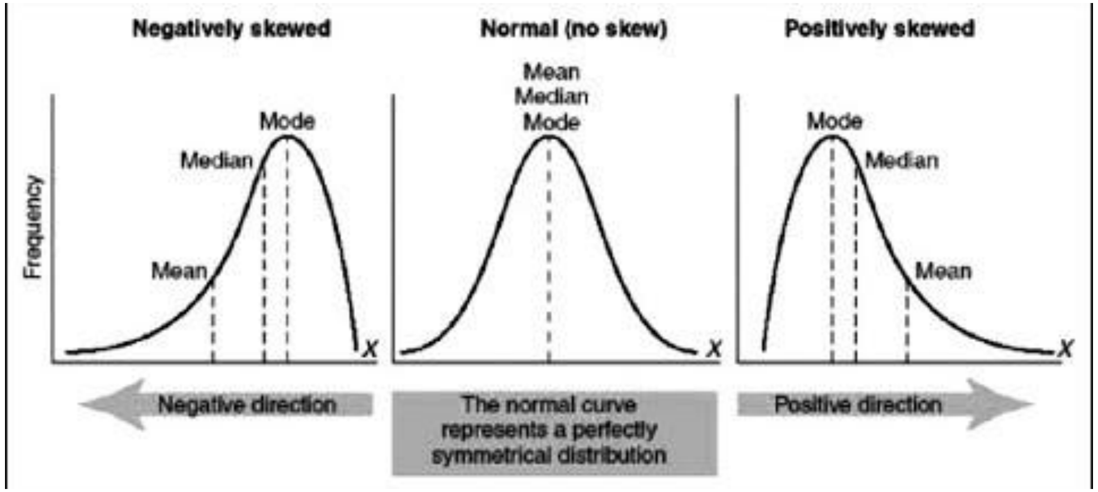
- Measures of Position:
  - Percentiles: indicates the value below which a given percentage of observations fall.
    - Formula: $P_k = \frac{k(n+1)}{100}$
    - Where P is the percentile rank position and k is the desired percentile
    - For example: a ranked dataset of 100 value from 1 to 100 {1,2,3,…,100} will have the 90th percentile at position 90 which also has the value of 90.
  - Quartiles: similar to percentile but instead divide the data set into 4 parts with Q1 being the 25th percentile, Q2 is the 50th percentile (median), and Q3 is the 75th percentile.

Descriptive statistics

- Measures of Position:
  - Z-score (standard score): expresses how many standard deviations a value is from the mean. Positive scores are above the mean, and negative scores are below.
    - Formula: $z = \frac{x - \bar{x}}{s}$

Descriptive statistics

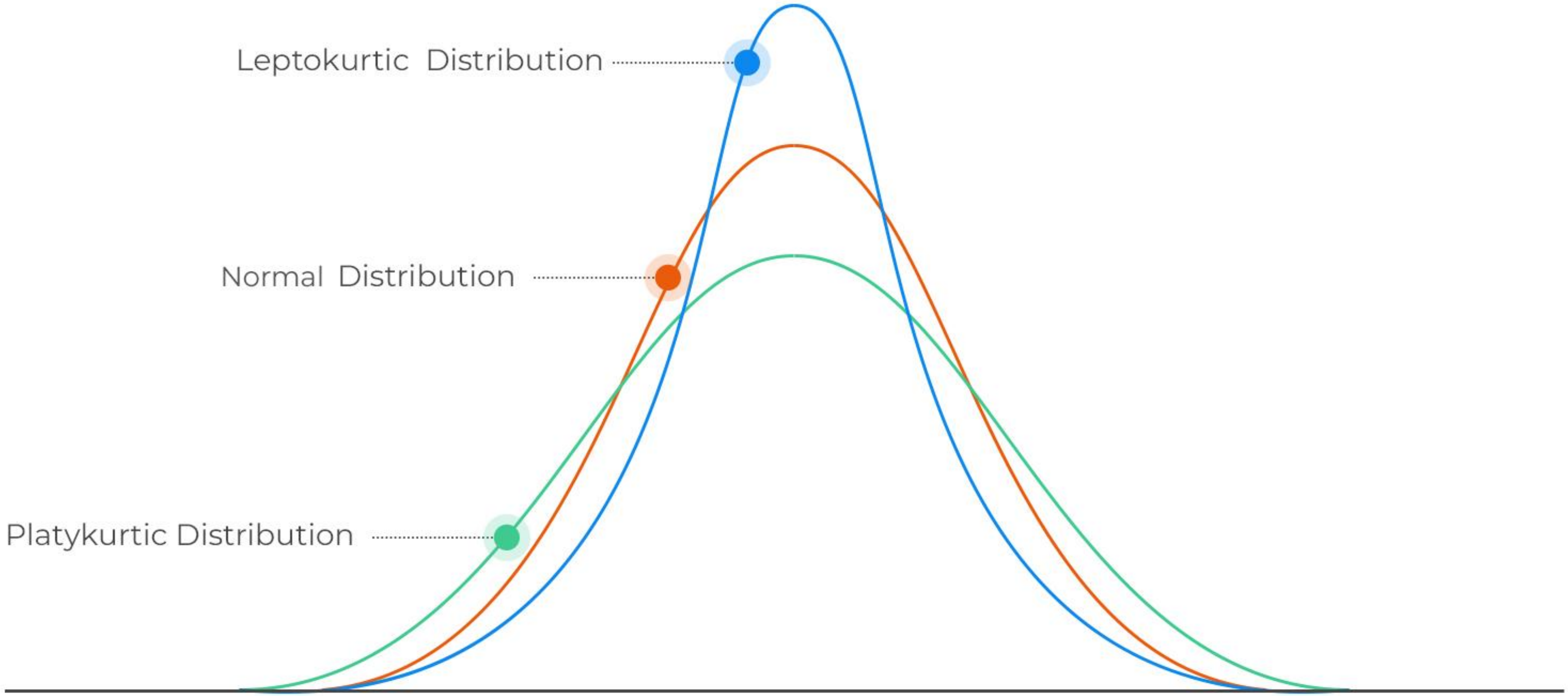
- Shape descriptors (distribution shape):
  - Skewness: measures the degree of asymmetry in the distribution of the data. Positive skew means the tail is on the right while negative skew means the tail is on the left.
    - Formula: $\text{Skewness} = \frac{\sum_{\mathrm{i}=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$



Descriptive statistics

- Shape descriptors (distribution shape):
  - Kurtosis: measures the "tailedness" of the data distribution, indicating whether data are heavy-tailed of light-tailed compare to a normal distribution. Higher kurtosis means that the data has more extreme values.
    - Formula: $\text{Kurtosis} = \frac{\sum_{\mathrm{i}=1}^{n}(x_i - \bar{x})^4}{(n-1)s^4}$
    - If the kurtosis value is > 3, the data is Leptokurtic, meaning data is more clustered around the mean. If the kurtosis value is < 3, the data is Platykurtic, meaning data is less clustered around the mean. If the kurtosis value is = 3, the data is Mesokurtic, meaning the distribution follows the normal distribution curve.

Kurtosis


Leptokurtic Distribution
Normal Distribution
Platykurtic Distribution

Applications of Data Analysis

- Qualitative analysis:
  - Examines non-numerical data to understand concepts, opinions, behaviors, or experiences.
  - Focuses on studying themes and patterns of the subject.
  - Example: analyzing customers’ feedback to identify where improvements are needed.
- Quantitative analysis:
  - Examines numerical data to measure quantities, compare groups, and identify relationships using statistical methods.
  - Focuses on the numerical and statistical interpretation.
  - Example: Calculating students’ average GPA and testing if there is a correlation between GPA and weekly study hours.

Real-world application

- Customer segmentation: identify group of customers with similar buying habits.
- Sales forecast: predict demand to plan production accordingly.
- Evaluation of marketing campaign: measure the return on investment (ROI) of the campaign.
- Risk assessment: evaluate the risk that a loan applicant carry (credit score).
- Clinical trials: evaluate the effectiveness of new drugs/prevention capabilities of new vaccines.

Real-world application

- Curriculum effectiveness: measure and compare the impact different teaching methods has on students’ performance.
- Census analysis: investigate population trend.
- Example:
    - During COVID-19, Vietnam’s Ministry of Health used case data and mobility tracking to forecast outbreak hotspots and allocate resources effectively.
    - Loyalty card programs used by supermarkets to create personalized promotion to customers based on their purchasing habits.
    - Tracking on monthly spending to create informed budgeting choices.

Ethical concerns

- Privacy and confidentiality: data collected must have the owner’s consent and must not be identifiable.
- Accuracy and integrity: data must be up to date and complete. Absolutely do not “cherry-pick” – use data that supports a specific agenda.
- Bias and fairness: data must try to reflect the population instead of trying to reinforce bias/stereotypes.
- Example:
    - U.S National Security Agency mass surveillance program where they collected data from citizens without consent.
    - Data trading of your phone number, name, personal information to scammers/marketers which results in scam/spam calls.

=> All entities whether private or governments can misuse data.