HCAI5DS02 –Data Analytics and Visualization.

Lecture – 01

Foundations of Statistical Thinking.

Understanding Data in Data Analytics.

Siman Giri

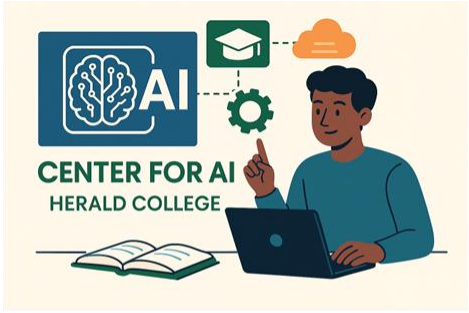


image generated via copilot.

1. Understanding Statistical Thinking.

"From Data to Decisions: A Critical Thinking Approach."

1.1 What is Statistical Thinking?

- "**Statistical thinking** is not about numbers; it is about **understanding the process that generated the data** and **using evidence** to **make decisions under uncertainty**."
  - George E.P. Box.
- To simplify Statistical Thinking best describes:
  1. **What process generated data?**
  2. **What summaries best describe it?**
  3. **What does the data suggest, and what does it leave uncertain?**

1

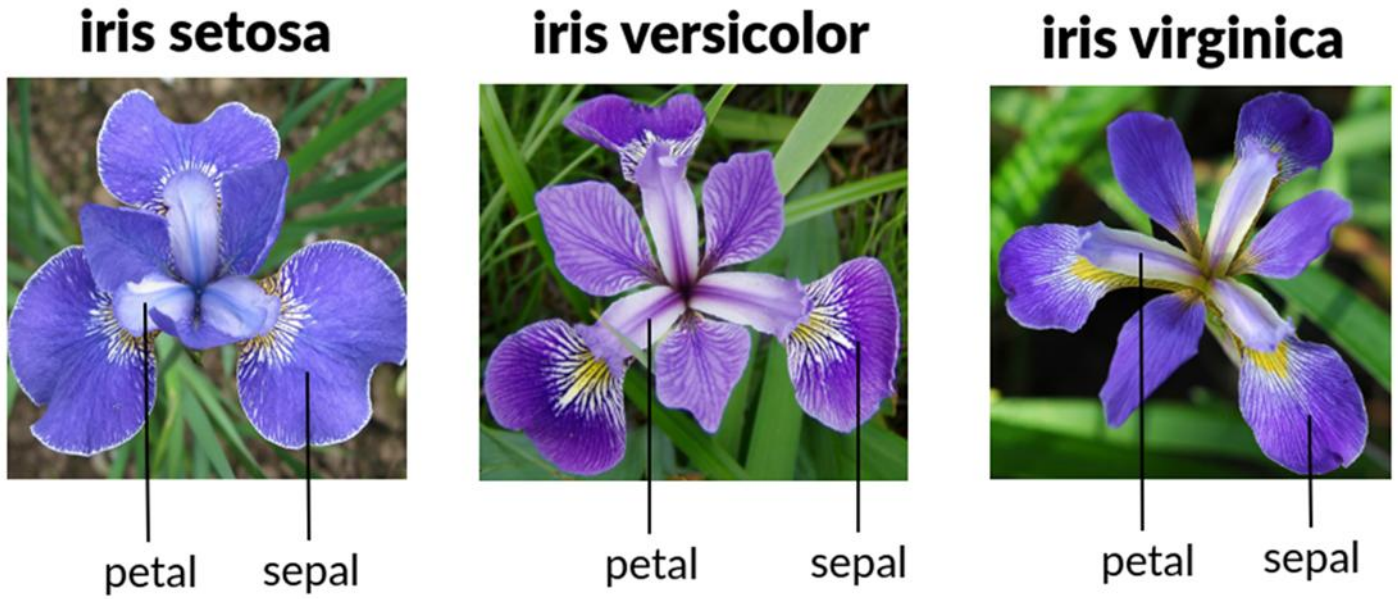
2

3

- Let's take an example.

1.2 Example: Exploring the Iris Dataset.

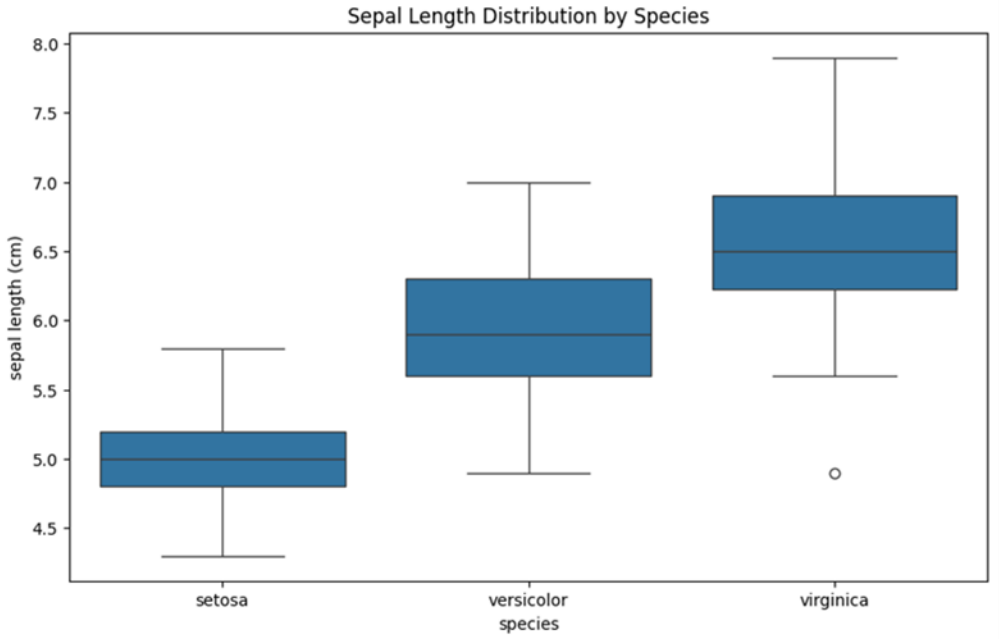
- Dataset Source:
  - The Iris dataset contains measurements of iris flowers from three species:
    - ***setosa, versicolor,*** and ***virginica.***
  - It has 150 samples with 4 numeric features: sepal length, sepal width, petal length, and petal width.

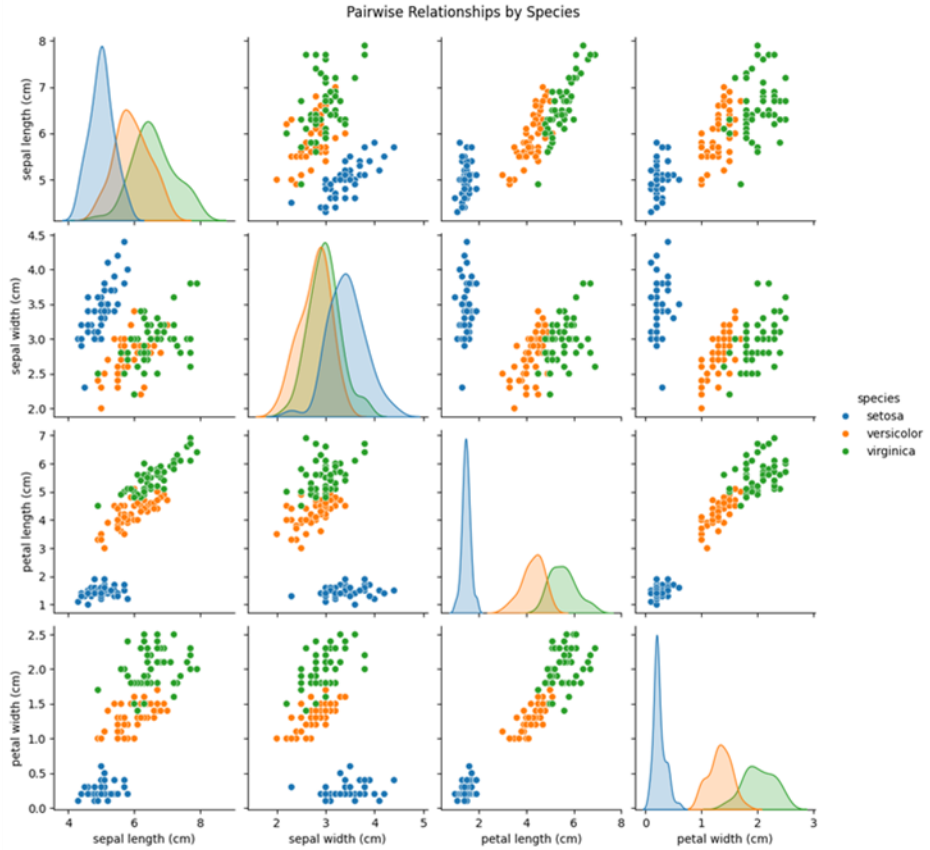


1.2.1 Example: What process generated this data?

- This data was collected by botanist Edgar Anderson and popularized by Ronald Fisher.
- It is observational data measured from flower specimens.
- Each sample corresponds to an individual flower; the species label was assigned by expert classification.
- No experimental manipulation, so results are descriptive and comparative rather than causal.

1.2.2 Example: What summaries best describe it?





- Boxplots reveal variation and outliers.
- Summary statistics (mean, median, std) show differences in flower measurements across species.
- Pair - plots help identify correlations and species separation patterns.

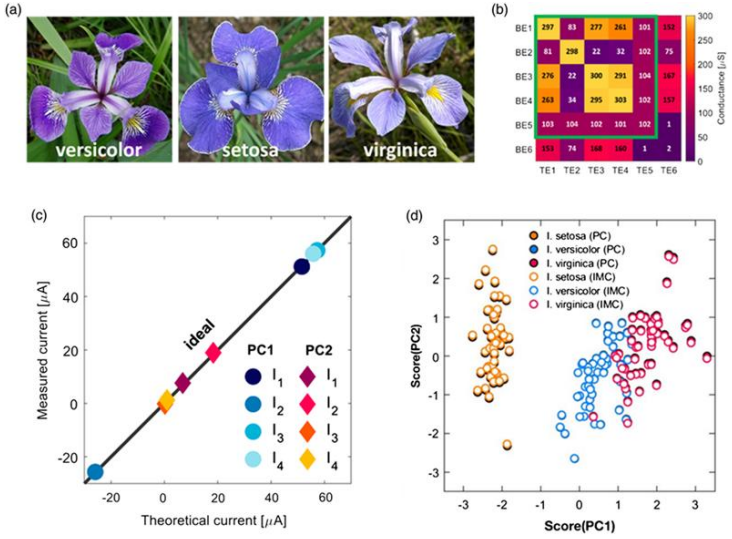
1.2.3 Example: What does the data suggest, and what does it leave uncertain?

Suggests:

- The three species show clear differences in petal and sepal dimensions.
- *Setosa* is clearly separable from the other two species based on petal length and width.
- There is correlation among some features, e.g., petal length and petal width.

Leaves Uncertain:

- Whether the differences are caused by genetics, environment, or other factors.
- The predictive power on new or mixed species outside this dataset.
- The boundary cases where species overlap are less clear.



1.3 Why statistical thinking is foundational to data science?

1. **Data Are Not Always What They Seem?** 🔍
   - Raw data can be misleading due to
     - **sampling errors**, **outliers**, or **biases**.
   - Statistical thinking helps you ask the right questions:
     - ***"Is this result due to chance?",***
     - ***"Is this a representative sample?"***

- "Without statistical reasoning, data science becomes data manipulation, not data understanding."

2. **Data Science = Inference + Prediction + Decision Making**
   - Data science is not **just about building models**
     - it's about ***drawing conclusions from data***.
   - Statistical thinking enables:
     - **Inference:**
       - What can we learn about the population from a sample?
     - **Prediction:**
       - How reliable are our predictions?
     - **Decision – making:**
       - Should we act on this result?

1.3 Why statistical thinking is foundational to data science?

3. **Distinguishing Signal from Noise** 📈:
   - Real world data is **messy and noisy**.
   - Statistical thinking helps identify:
     - True patterns vs. random fluctuations.
   - **Correlation** vs. **Causation**
- Without it there is a risk of overfitting, **p-hacking** or **false discoveries**.

4. **Thinking Critically About Models** 🧠:
   - A **data analyst** should *not just trust the output* of a **machine learning model**.
- Statistical thinking helps you ask:
  - ***"Is this model appropriate?"***
  - ***"Are the assumptions valid?"***
  - ***"How confident can we be in the result?"***

5. **Effective Communication with Uncertainty** 📊:
   - One of the most valuable skills in data science is **explaining results clearly**
     - including their limitations.
   - Statistical thinking helps frame **conclusions with nuance:**
     - ***"We are 95% confident that the effect is between X and Y."***
     - ***"This correlation does not imply causation."***

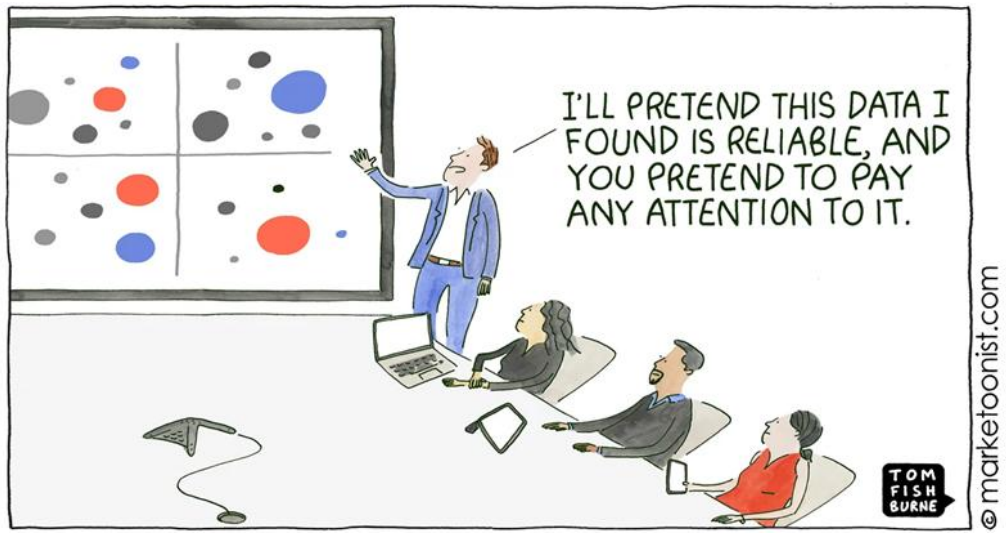
"You may well have data, Smithers, but I have strong opinions, and I pay your wages"

1.3 Why statistical thinking is foundational to data science?

6. **Understanding the Data Generating Process** 🧪:
   - Is the data from a controlled experiment or natural observation?
     - **Are there confounders?**
     - Missing data?

- Statistical thinking trains you to consider how the data was produced, which deeply influences analysis and conclusions.

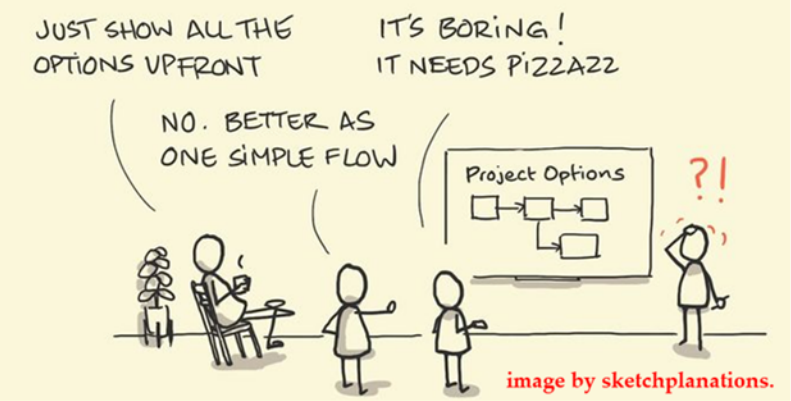
7. **Guarding Against Misinterpretation** 💭**:**
   - Misuse of statistics can lead to bad policy, misleading business insights, or public misinformation.
   - Statistical thinking builds the reflex to question conclusions, not just accept them.



2.Data: The Fuel of Insight.

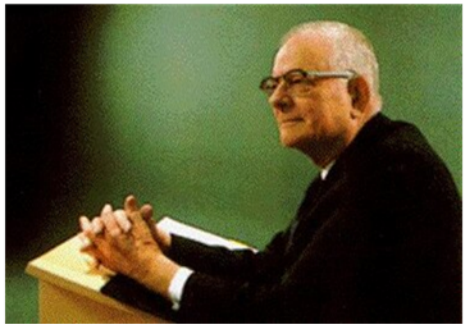
"Understanding What Data Is, Before we Analyze It."

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"



"**Without data, you're just another person with an opinion.**"

- William Edwards Deming.



🔍 What Exactly is Data, Really?

2.1 What is Data, Really?

- "*Data is not just numbers — it's numbers with a context*."
  - *David S. Moore*
- **Data** is a structured **collection** of ***facts***, ***observations***, or ***measurements*** about **objects**, **events**, or **phenomena**
  - used to **derive** **insight**, **make predictions**, or **guide decision-making**.
- In this course, we define:
  - **Data** = **Varaibles (what we observe)** + **Measurements (how we quatify it)**.
    - **Variables** refer to **attributes**, **characteristics** or **properties** of **objects or phenomena**
    - **Measurements** refer to the **observed values** of **those variables**.
- Examples:

Phenomenon	Variable	Measurement
Color of the Sky	Color	"Blue"
Height of Mount Everest	Elevation	8848 meters
Yesterday's Temperatures	Temperature (Min – Max)	28 – 31 C.



2.2 Type of Data Measurements.

Quantitative

- It can be expressed as **a number**, so it can be **quantified**.
- Easier to **manipulate and represent** using various **statistical** tools.
- There are two major types:
  - **Discrete Data**
    - A numerical variable results in discrete data if the possible values of the variable correspond to isolated points on the number line.
    - Countable
  - **Continuous Data**
    - A numerical variable results in continuous data id the set of possible values forms an entire interval on the number line.

Qualitative

- Sometimes also called **Categorical**
  - This type of data can't be **counted or measured** easily using numbers and therefore **divided into categories**.
  - The gender of a person (male, female, or others) is a good example of this data type.
- There are two major types:
  - **Nominal Data**
    - These are the set of values that **don't possess a natural ordering**.
    - The **color** of a smartphone can be considered as a nominal data type as we can't compare one color with others.
  - **Ordinal Data**
    - These types of values have a natural ordering while maintaining their class of values.
    - If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of **small < medium < large**.

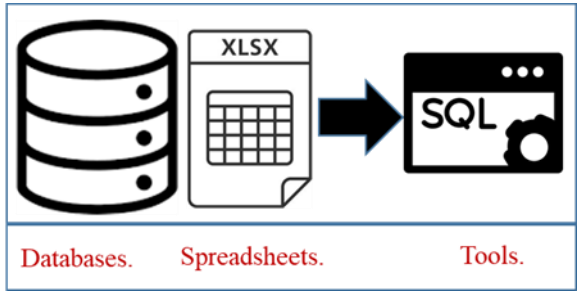
2.3 How Data are Organized?

- **Data Needs to be Organized:**
  - Data presented in various forms and formats are then organized before being **processed for analysis**:
    - Data in general are organized two way:
      - **Structured**
      - **Unstructured**

2.3.1 Structured Vs. Unstructured.

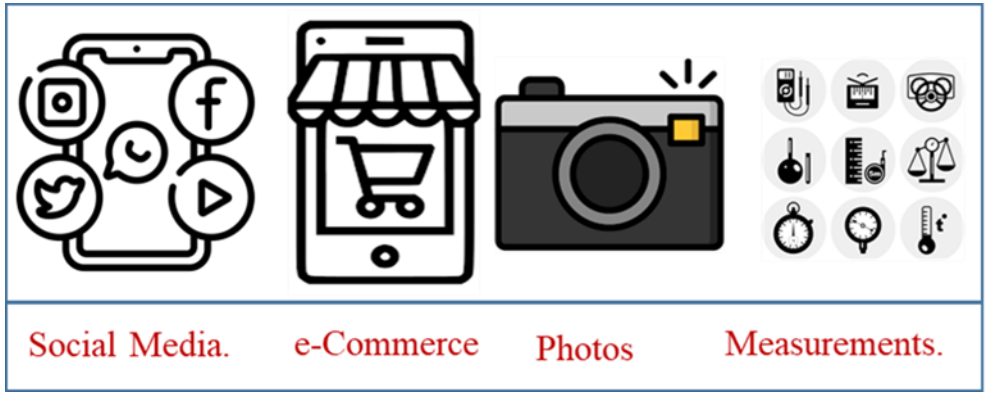
Structured Data

- Structured data refers to data that is organized in a well-defined format, with a clear and consistent structure that can be easily processed and analyzed by computers or computer-based tools.
- This type of data is typically stored in databases or spreadsheets, and can be easily queried using **SQL** (Structured Query Language).
- Examples of structured data include **financial records**, **customer data**, and **inventory data**.



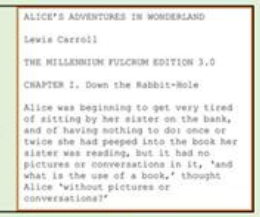
Unstructured Data

- Unstructured data refers to data that does not have a clear and consistent structure, and is not organized in a predefined format.
- Unstructured data is often more difficult to analyze and process than structured data, as it requires more complex algorithms to make sense of the information.



2.4 Data Storage.

Plain Text:
Ends in .txt (generally)
No formatting, font type, font size, color, etc.
Text position is povided by whitespace characters (space, tab, return)



JSON

XML

.csv format
Tab-separated (.tsv/.csv)

Delimiter: character that seperates.

```
Bill #, Jack Reed (RI), Elizabeth Warren (MA)
Bill 27, Yay, Yay
Bill 28, Yay, Nay
Bill 30,, Nay
Bill 47, Nay, Nay
Bill 91, Nay, Nay
Bill 105, Yay,Yay
```

Fig: Storing Data: Different formats.

For this course:
We are interested in data in tabular form stored as .csv file

3. The Data Analytics Process.

{"From Questions to Insight: A Six Step Journey."}

3.1 What is "The Data Analytics Process"?

- The **data analytics process** is a **structured framework** that guides us **from identifying a problem** to making informed, data-driven **decisions**. It involves two critical tasks:

Task ⚙️	Description 💡
Data Collection 📥	Gathering relevant and reliable data using proper methods
Data Analysis 📊	Using tools and techniques to extract patterns and meaning

- **Cautions:**
  - Raw data without analysis is noise.
  - {Even advance sophisticated} Analysis without proper data is misleading.
  - Thus:
    - **Good analysis needs good data and Good data needs proper collection methods.**
- Both data collection and data analysis are essential. When done well, they transform scattered facts into knowledge – and knowledge into action.

Facts 📋 → 💡 Insights → 📚 Knowledge → Decisions ⚖️

3.2 What is the purpose of Data Analytics?

- **Data analytics** allows us to **transform questions** about the world into **evidence – based answers**.
- **For Example:**
  - 💉 Is a new flu vaccine effective in preventing illness?
    - **Data collected from clinical trials, analyzed to measure reduction in illness.**
  - 🏍️ Are motorcycle injuries less severe for helmeted riders?
    - **Accident data helps compare injury severity between helmeted and non – helmeted riders.**
  - 🛍️ Why are weekend sales lower than weekday sales?
    - **Sales and Customer footfall data can reveal behavioral trends.**
- To answer them reliably and responsibly, we need:
  - **Good Data → Collected through careful planning.**
  - **Statistical Thinking → To understand variability and bias.**
  - **Data Analytics Process → Brings it all together.**
    - **Ensuring** that both data collection and analysis are conducted rigorously, so **decisions** are based on **evidence** rather than **the guesswork**.

3.3 The Data Analytics Process:

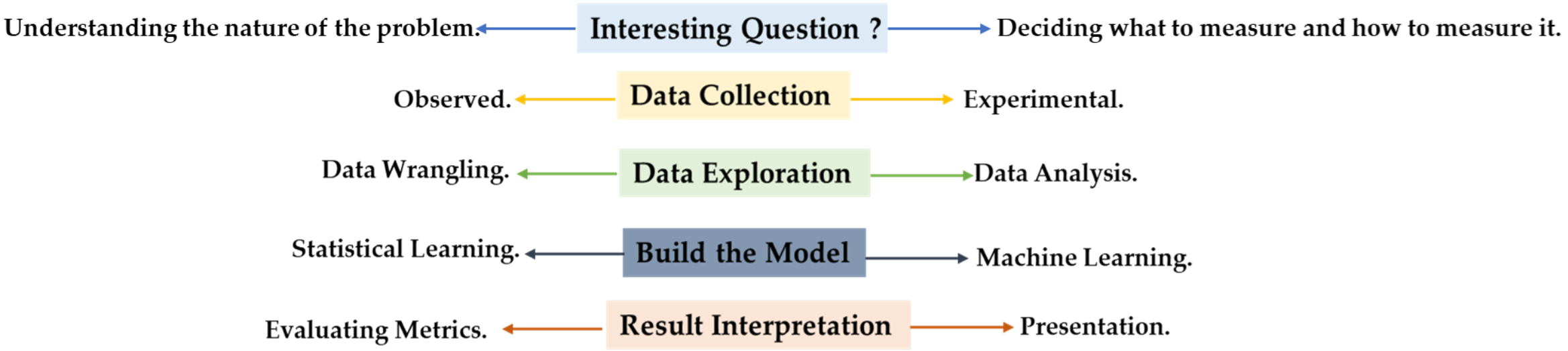


Fig: Elements of a Data Analytics Process.

3.4 Asking an Interesting Question:

- Every data analytics journey begins with a **meaningful question**. A **strong question** helps us:
  - **1. Understand the Nature of the Problem:**
  - **2: Decide What to measure and How:**

1. **Understand the Nature of the Problem:**
   - Effective **data analysis** begins with **understanding what problem** we are trying to **solve**.
   - We need clarity on:
     - The goal of the analysis.
     - The questions we hope to answer.
   - A **well – defined question** keeps the project focused and avoids wasted effort.
   - Without a clear question, even a well-run analysis may miss the mark. For example:
     - **Good Question**: **" Does daily screen time affect student academic performance?"**
       - **Nature of the Problem**: Investigate whether there is a relationship between screen use and grades.
     - **Bad Question**: **" Does eating an apple a day keeps doctor away?"**
       - **Nature of the Problem**: Too vague and based on a proverb. It lacks clearly defined variables or measurable outcomes. It also oversimplifies complex health dynamics.

3.4 Asking an Interesting Question:

2. **Decide What to Measure and How:**
   - Ask: What information is required to answer the question?
   - Identify:
     - The variables involved
     - The method and tools of measurement
     - The population or sample scope
   - These decisions ensure that **our data** will be **useful and valid**.
   - Example:
     - **Question:** " Does daily screen time affect student academic performance?"
     - **What to measure:**
       - **Variable 1: Average daily screen time in hours.**
       - **Variable 2: Academic performance GPA or test scores.**
     - **How to measure:**
       - Use surveys or digital tracking for screen time.
       - Collect grades from school records or self – reports.

3.5 But … The Question Delima …

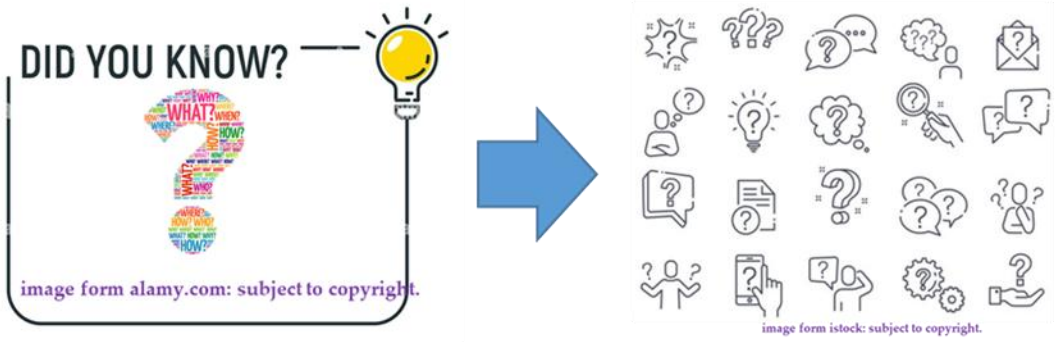


Fig: Question → Search For Data.

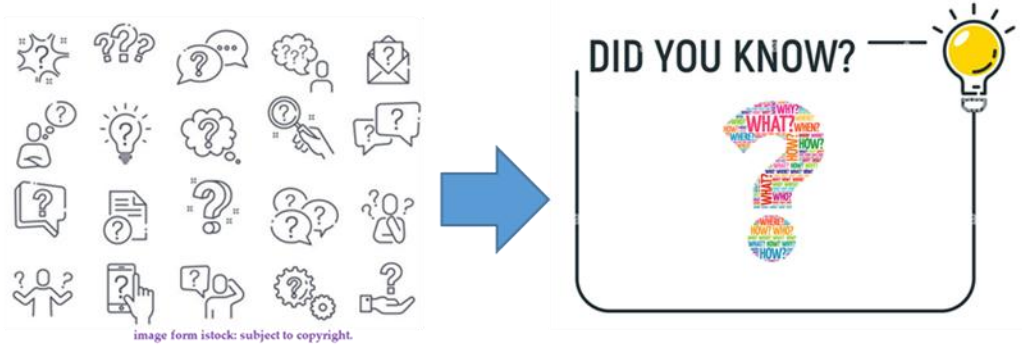


Fig: Data → What question you can answer?

Fig: Question to Data or Data to Question

"We tried crowdsourcing the data, but the crowd turned into an angry mob and threw bogus data at us instead."

4. Collection of Data.

{Where Data Comes From & Why it Matters?}

4.1 Data Collection Methods …

Method 🧪	Definition 📖	Example 💡
Observational Study	Data Collection through passive observation without intervention	Analyzing test scores across different schools.
Experimental Study	Data Collection involving deliberate manipulation of variable to assess effects.	Conducting A/B testing on a website layout.
Survey	Data Collection via structured questionnaires administered to participants.	Employee Satisfaction survey.
Census	Comprehensive data collection from the entire population under study.	National population count.
Sampling	Data Collection from a representative subset of the population.	Polling 1,000 voters before an election.

📌 **Note:**
While both experimental and observational studies are essential to data science, **this week we focus only on *observational studies*. Experimental methods will be covered in upcoming weeks.**

4.2 Observational Study: What Is it?

- An ***observational study*** involves **watching and recording data without influencing** the **environment** or **subjects** being studied.
  - You collect data as it naturally occurs, without applying any treatments or interventions.

Table: Types of Observational Studies with examples.

Type	Description 📝	Example 💡
Surveys & Questionnaires	Self – reported data on preferences or behavior	Student screen – time survey
Census or Population Data	Comprehensive data about a population	Government demographic data
Observational Logs	Passive data recording	Tracking clicks on a website
Case – control Studies	Comparing groups based on outcome	Comparing smokers and non – smokers with lung cancer
Cohort Studies	Tracking a group over time	Studying diet patterns and heart disease in adults

⚠️ Please Note with Cautions:

- **Survey**, **Census**, and **Sampling** are ***methods*** of **data collection**.
- They usually involve observing or recording responses or characteristics from people or populations **without** intervention.
  - Therefore, **they are often considered types or methods within observational studies**, because you are typically just observing or recording data without experimental manipulations.
- However, **if a survey or sampling involves manipulating variables or treatment** (e.g. randomly assigning different conditions), then it might be part of an **Experimental Study**.

4.3 Getting Observational Data.

Fig: Classical Approaches in Data Collection.

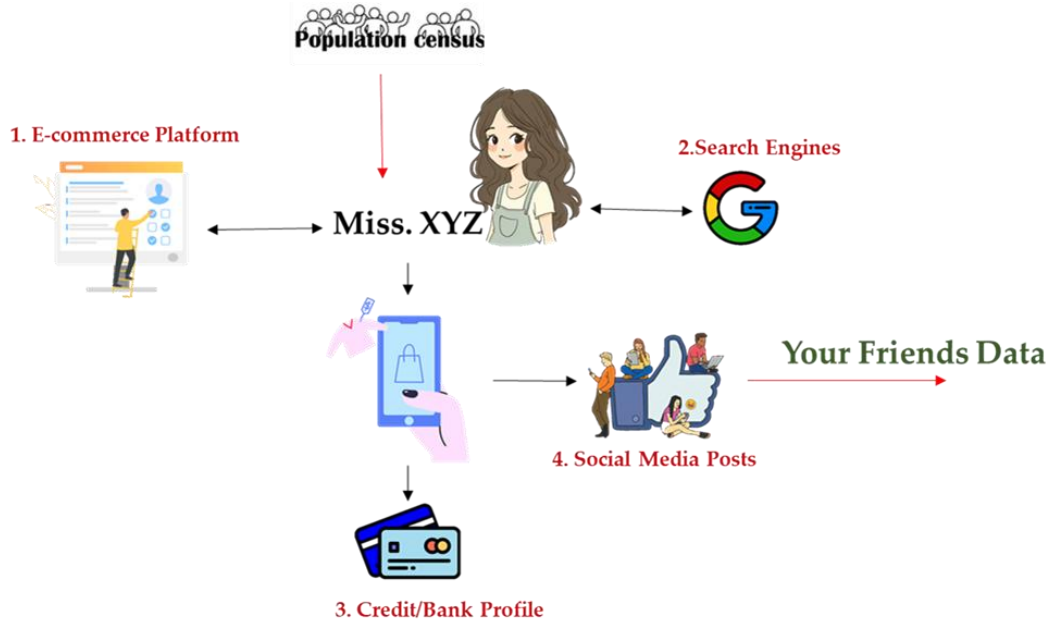


Fig: Points where data can be collected for single online order.

5.1 Primary Source of Collecting Data: A Survey.

Designing a Good Survey: A practical Guide.

5.1 Designing Effective Surveys.

- Surveys are a cornerstone of data collection, but poor design can introduce **bias, inaccuracies, and low response rates**, leading to unreliable insights.
- As a **data analyst**, your role in survey design is critical to ensuring **high-quality, actionable data**.
- Below are **key strategies** to optimize your surveys:
  1. **Define Clear Objectives:**
     - Before drafting questions, ask:
       - What is the research goal? {Exploratory, descriptive, casual?}
       - What specific insights are needed? { Customer satisfaction, user behavior, market trends?}
       - How will the data be analyzed? { Statistical tests, regression, NLP?}
     - Good surveys begin with a focused goal.
     - Align each question with a **measurable objective** to avoid "**nice-to-know**" but irrelevant data.
     - Examples:
       - **"How many hours do university students spend coding per week?"**
       - **"What Challenges do first – year CS students face when learning Python?"**
       - **Avoid Vague goals like "What do students think?"**

5.1 Designing Effective Surveys.

2. Optimize Question Design:

- **Avoid common Question Pitfalls.**

Problem	Bad Example ❌	Good Solution 👍
Leading Questions	"Don't you love our new product?"	"How would you rate our new product?"
Double – barreled Questions	"Do you find our app fast and user – friendly?"	Split into two Questions: 1) Is our app fast? 2) Is our app user friendly?
Ambiguous Wording	"Do you exercise often?"	"How many days per week do you exercise?"
Overlapping scales	"Age Groups: 10 – 20, 20 – 30, 30 – 40"	"Age Groups: 10 – 19, 20 – 29, 30 – 39"

- **Choose the Right Question Type**

Type	When to Use 👇	Example ✨
Multiple – choice (single)	When only one answer is possible.	"Which OS do you use most?"
Multiple – choice (multi)	When multiple answers apply.	"Which social media platforms do you use?" (select all)
Likert Scale	Measuring attitudes/agreement	"Rate your satisfaction:" (1= very dissatisfied, 5= very satisfied)"
Open – ended	Exploratory insights	"What improvements would you suggest?"

5.1 Designing Effective Surveys.

3. Think About Measurement Scales:

- **Why this Matters?**
  - Different types of data require different types of visualizations, statistical tests, and interpretation strategies. Getting this wrong can lead to invalid conclusions.

Measurement Scale	Definition	Example Question	Option
Nominal Scale 🔹	Categorizes data without any order or ranking.	What is your primary programming language?	Python, Java, C++, …
Ordinal Scale 📶	Data with a natural order, but intervals between values are not meaningful or uniform	How confident do you feel debugging code?	1) Not Confident, 2) Neutral, 3) Confident.
Interval Scale 📏	Numeric data with equal spacing between values, but no true zero point.	What is the average room temperature where you usually code?	**−0°C** →It does not mean "no temperature"
Ratio Scale 📐	Numeric data with equal intervals and a meaningful true zero.	How many hours did you spend coding yesterday?	**0 hours** → It mean no time spent

5.1 Designing Effective Surveys.

4. **Minimize Bias & Maximize Response Quality:**
   - **Sampling Bias Prevention:**
     - Random Sampling: if possible, to ensure representatives.
     - Stratified Sampling: if subgroups e.g. age, region need proportional representation.
     - Avoid self – selection bias: only surveying volunteers.
   - **Response Bias Mitigation:**
     - Anonymous surveys for sensitive topics like student's feedback.
     - Neutral wording avoid emotionally charged language.
     - Balanced scales "very poor" to "Excellent" not just "Good" to "Great".
   - *"We'll explore **sampling techniques** and **bias mitigation** in detail in the upcoming sections."*
   - **Increased Response Rates:**
     - Keep it short { 5 – 10 minutes max.}
     - Offer incentives { gift cards, discounts etc.}
     - Send reminders { Follow – ups improve completion rates.}

5.1 Designing Effective Surveys.

5. Pilot Test Before Launch 🛠️:

- Test with 5 – 10 people from your target audience.
- Check for:
  - Clarity of questions.
  - Technical issues (e.g. broken skip logic, missing options)
  - Average Completion time.
- Refine based on feedback.

6. Ethical Considerations 🔐:

- **Informed Consent:**
  - Explain how data will be used.
- **Privacy Compliance:**
  - Avoid collecting unnecessary personal identifiable information.
- **Transparency:**
  - Share aggregated results if promised.

5.2 Secondary Source of Collecting Data: Sourced.

Using Pre – Existing Data to Drive New Insights.

5.2 What is Secondary Data?

- **Secondary data** refers to data that was **collected by someone** else for a **different purpose** but can be used for your **own research or analysis**.
- It can come from **government agencies, research institutes**, **company reports**, **public datasets**, and more.
- **Examples of Secondary Data Sources** 📚:
  - **Government Databases**: Census data, labor statistics, healthy surveys.
  - **Research Repositories**: Kaggle datasets, UCI Machine Learning Repository.
  - **Company and Product Data**: App reviews, GitHub activity logs.
  - **Web and Social Media Data**: Twitter trends, Reddit posts via APIs.

Advantages ✅	Challenges ⚠️
Saves time and cost of primary data collection.	Data may be outdated or irrelevant for your specific question.
Offers access to large, well curated datasets.	Lack of control over how it was collected.
Enable historical analysis and benchmarking.	Potential quality issues i.e. missing values, unclear definitions.
	Licensing or ethical constraints - especially with web-scraped data.

6. Working with Real – World Data: Challenges.

"Understanding Data Quality, Context, and Collection Bias."

6.1 Key Terminologies When Working with Data.

Population

- Comprehensive collection of all **possible measurements or outcomes** for any given **context/objectives** is called the **population**.
- **Example:**
  - A survey was carried to estimate the average BMI of all undergrad students majoring in computer science.
  - Population: All of the student studying computer science.

Sample

- A **sample** is a **subset** of the **population**,
  - selected for analysis when studying the full population is impractical or impossible.
- In **most real-world data analytics tasks**, we work with **sample data** — not the **full population**.
- **Cautions ⚠️:**
  - **Random Sampling**:
    - Helps ensure that the sample is representative of the entire population.
  - **Bias**:
    - A poorly chosen sample can misrepresent the population and lead to inaccurate conclusions.
- Sample Example:
  - selected 300 students from herald college.

6.2 Random Sampling: Key Characteristics.

- Almost **every dataset** used in **Data Analytics** or **Data Science** or **Machine Learning** originates from some form of **sampling**.
  - Therefore, it's essential to understand the **key characteristics** of good sampling:
  - **Sampling Process** 🔁:
    - The method or mechanism through which data is selected from the population.
  - **Randomness** 🎲:
    - Every individual in the population must have an **equal chance** of being selected.
    - Example: Randomly selecting 100 students from a full college roster using a random number generator.
  - **Biasness** ⚠️:
    - Bias occurs when certain outcomes or individuals are favored or excluded.
    - May leads to invalid or misleading conclusions.
  - **Representativeness** 👥:
    - A representative sample is one in which the sampling process is free of bias, so that each subgroup in the population has a fair choice of being included.
    - Example
      - **Goal: To conduct a survey on average height of Herald's Student.**
        - You Choose your subgroup to be Herald's Basketball Team.
          - **Did your selected sub-group be called representative?**

6.3 Method of Sampling: Simple Random Sampling.

- **What is it?**
  - A sampling technique where every member of the population has an equal and independent chance of being selected.
    - It is the **method of selection**, not the **composition of the final sample**, that defines **SRS**.
- **Formal Definition:**
  - **"A simple random sample of size n is a sample selected from a population such that every possible sample of that size n has an equal probability of being chosen."**
- **Why it Matters?**
  - Ensures unbiased representation.
  - Helps generalize results to the population.
  - Foundation for other statistical methods.

6.3.1 Simple Random Sampling: An Example.

- A university wants to **conduct a survey** to understand **the opinions** of its students about a new campus recreation center. The university has **a population of 10,000 students**.
- To conduct the survey, the researcher decides to **use simple random sampling**.
  - The researcher obtains a **list of all 10,000 students { aka sampling frame}**.
    - They use a **random number generator** to select a **random sample of 500 students** from the list.
      - [The random number generator assigns a unique number to each student in the list and then selects 500 numbers at random.]
- Every student has an equal chance of ending up in the sample.
  - **How this Ensures key Characteristics of Simple Random Sampling:**

Characteristic:	**How the Example Ensures It:**
Sampling Process	• The researcher clearly defines the process: • get a list of 10000 students – the sampling frame, • then use a random number generator to select 500.
Randomness	• A random number generator ensures that every student has an equal chance of being selected. • {Every student has a 1 in 20 (or probability of 0.05) chance of being selected.}
Representativeness	• Since the selection is purely random and drawn from the full population, the sample is likely to reflect the population's diversity.
Bias Avoidance	• Because the selection is not based on judgement, convenience, or subgroups, the process minimizes systematic biases.

6.4 Simple Random Sampling: Techniques.

- The fact that every **individual has an equal chance of selection**, by itself is not enough to guarantee that **the sample is a simple random sample**.
- **Selecting a simple Random sample:**
  - **Sampling without Replacement:**
    - Once an individual from the population is selected for inclusion in the sample, it may not be selected again in sampling process.
    - A sample selected without replacement includes 'n' distinct individuals from the population.
  - **Sampling with Replacement:**
    - After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process.
    - A sample selected with replacement might include any particular individual from the population more than once.

Fig: Sampling with Replacement

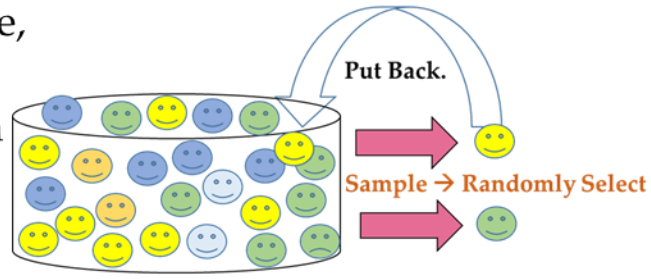


Fig: Sampling without Replacement

6.5 Popular Sampling Type

1. **Stratified Sampling:**
   - The population is divided into **strata** (subgroups) based on a shared characteristic (e.g., gender, major, year level).
   - Then, **random samples** are taken **from each stratum**.
   - Why use it?
     - Ensures representation from all subgroups.
   - **Example:**
     - From a university population: 40% female, 60% male
     - You divide into gender groups,
       - then randomly sample 100 Students ensuring 40 are female and 60 are male.

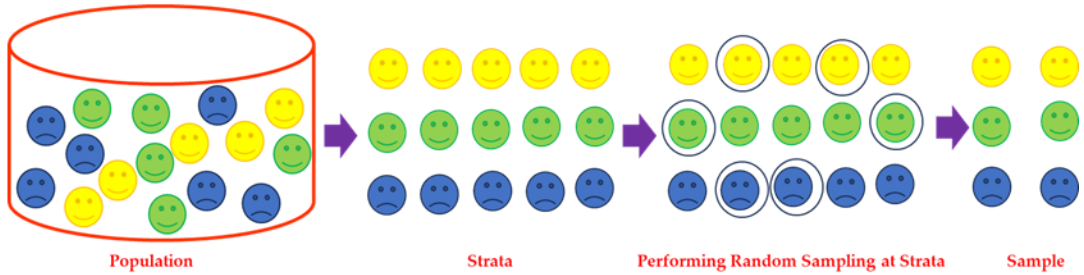


Fig: Stratified Random Sampling.

   - Types of Stratified Sampling:
     - **Proportional Stratified Sampling**:
       - The sample size from each stratum is proportional to the stratum's size in the overall population.
     - **Disproportional Stratified Sampling**:
       - The sample size from each stratum is not proportional to its size in the population, often to ensure adequate representation of small groups.

6.5 Popular Sampling Type

2. **Cluster Sampling:**
   - The population is divided into clusters
     - often but not compulsory clusters are often **selected for convenience** e.g. geographic regions or section and they may be internally heterogeneous or homogeneous.
       - Cautions: This can introduce bias if the selected clusters do not capture the diversity of the population.
     - So, while it is cost effective, the representativeness depends heavily on how clusters are formed and selected.
   - **Why use it?**
     - Useful when the population is large and spread out geographically.
     - Example: A university has 200 CS Classes. Randomly select 10 classes (clusters) and survey all students in those classes.
       - Note: Unlike stratified sampling, cluster members **are ideally heterogeneous**.

6.5 Popular Sampling Type

3. Systematic Sampling:

- Select every **k – th** unit from a list after a **random start.**
- Example:
  - From a list of 1,000 students, choose every 10^{th} student after randomly starting at the 4^{th} resulting in a sample of 100.
  - Condition: **The list must be randomly ordered**.

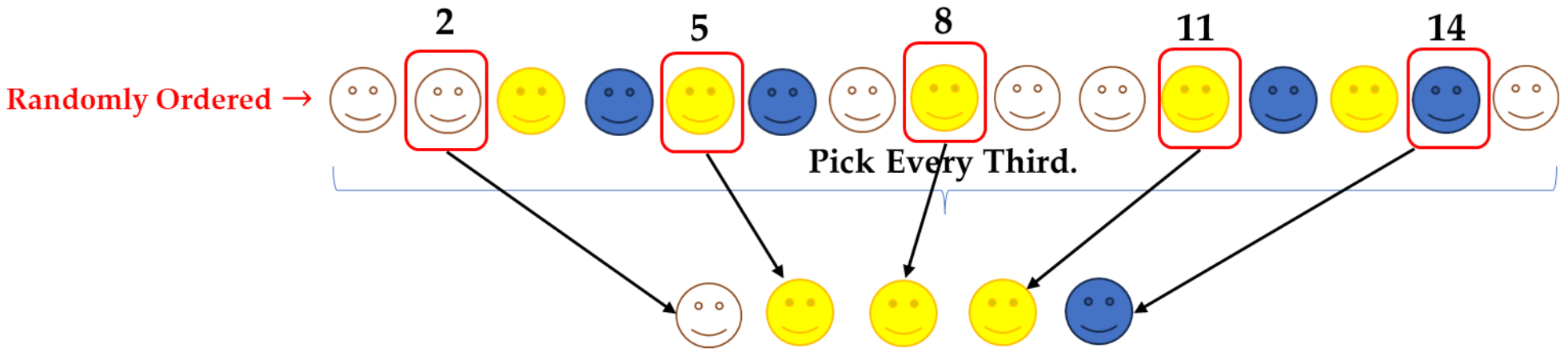


Fig: Systematic Sampling (every 3^{rd}).

Sampling Techniques: Cheat Sheet.

Sampling Type	Divides Population	Selection Method	Ensures Representations?	Use Case
Simple Random	No ❌	Randomly picks individuals	Not guaranteed	Small homogeneous group
Stratified	Yes {Strata} ✅	Random from each Strata	Yes	Subgroup analysis (e.g. gender, major)
Cluster	Yes {Clusters} ✅	Randomly pick whole clusters	Not necessarily	Geographically spread populations
Systematic	No ❌	Every k-th item after random start	Not always	Ordered lists, assembly lines

6.6 Biasness induced by Sampling.

- Biases may not necessarily be **intentional.**
- Even if you don't *think* your **over-/ under-representation of a subpopulation** will impact your results**, it's still a bias**.
- Always strive to **minimize any biases** in your data collection procedures.
- **Common biases in data (sampled dataset):**
  - **Selection Biases:**
    - Occurs when certain members of the population are **more likely to be included** in the sample than others.
    - Example: Surveying only students in the computer lab may miss those who study at home.
  - **Survivor Bias:**
    - Happens when **only the subjects that "survived" a process are considered**, ignoring those that didn't make it.
    - Example: Studying successful startups and ignoring the failed ones leads to overestimating success rates.
  - **Simpsons Paradox:**
    - An observed trend appears in different groups of data but **disappears or reverses when these groups are combined**.
    - Example: A treatment appears effective for both men and women individually, but not when the data is combined due to uneven group sizes.

6.6.1 Think & Discuss: Is this Sample Bias - free.

- **Scenario:**
  - A policy group lobbying for the alcohol industry conducted an **exit poll** outside of a **popular restaurant late at night**. Patrons were asked:
    - **"Do you think driving after drinking is a serious problem?"**
      - Based on responses, the group **concluded that driving after drinking is not a serious problem**.
- **What's Wrong Here?**
- **What kind of Bias it may cause?**

6.6.1 Think & Discuss: Is this Sample Bias - free.

- **Scenario:**
  - A policy group lobbying for the alcohol industry conducted an **exit poll** outside of a **popular restaurant late at night**. Patrons were asked:
    - **"Do you think driving after drinking is a serious problem?"**
      - Based on responses, the group **concluded that driving after drinking is not a serious problem**.
- **What's Wrong Here?**
  - 🧩 Who was included in this sample?
    - People leaving a **popular restaurant late at night**—likely to be those who **drink socially**.
  - ❌ Who might have been excluded from the sample?
    - People who do **not go out late at night**, those who **don't drink**, or individuals who **actively avoid alcohol-related environments**.
  - 📊 Is this sample likely to reflect the opinions of the entire population?
    - **No**. It is skewed toward people who are more likely to **underestimate** the dangers of drinking and driving because of their **environment and habits**.
  - 🚨 How might timing, location or affiliation of the poll influence the results?
    - **Timing** (late at night) and **location** (outside a drinking venue) increase the chance of surveying people with **favorable views toward alcohol use**.
    - The **affiliation** with the alcohol industry introduces **bias** in framing and intent, possibly influencing the question wording or interpretation.
- **What kind of Bias it may cause?**
  - This is **selection bias** because the **method of sample collection systematically excludes** diverse opinions and overrepresents a subgroup (social drinkers), leading to **misleading conclusions** about public opinion.

6.6.2 Think & Discuss: Is this Sample Bias - free.

- **Scenario:**
  - Dr. Joseph Rhine conducted an experiment with 500 participants, asking them to guess the order of cards in a shuffled deck.
    - After each round, only those who guessed correctly moved on to the next round.
      - Eventually, the final participant who guessed correctly in all rounds was labeled as having telepathic abilities.
- **What's Wrong Here?**
- **What kind of Bias it may cause?**

6.6.2 Think & Discuss: Is this Sample Bias - free.

- **Scenario:**
  - Dr. Joseph Rhine conducted an experiment with 500 participants, asking them to guess the order of cards in a shuffled deck.
    - After each round, only those who guessed correctly moved on to the next round.
      - Eventually, the final participant who guessed correctly in all rounds was labeled as having telepathic abilities.
- **What's Wrong Here?**
  - This study is a classic case of **survivor bias** because it focuses **only on the participant who succeeded** through all the rounds, ignoring the hundreds who failed earlier.
    - ✅ The "survivor" (final participant) is **assumed to be special**, without accounting for how **random chance** could lead someone to guess correctly multiple times.
    - ❌ The experiment **disregards the failed participants**, who are essential for understanding the **full probability distribution**.
    - 🎲 The outcome could be explained by **pure luck**, not telepathic ability.
- **To Conclude:**
  - Survivor bias **skews conclusions** by ignoring those who don't make it through a selection process. In this case, **focusing only on the winner** leads to **misinterpretation of statistical chance as extraordinary ability**.

6.6.3 Think & Discuss: Is this Sample Bias - free.

- **Scenario:**
  - A research group tracked a cohort of individuals over 40 years and concluded that those who drank red wine were wealthier, more educated, had lower cholesterol ratios, and a reduced risk of heart attacks compared to non-drinkers.
- **What's Wrong Here?**
- **What kind of Bias it may cause?**

6.6.3 Think & Discuss: Is this Sample Bias - free.

- **Scenario:**
  - A research group tracked a cohort of individuals over 40 years and concluded that those who drank red wine were wealthier, more educated, had lower cholesterol ratios, and a reduced risk of heart attacks compared to non-drinkers.
- **What's Wrong Here?**
  - This is an example of **Simpson's Paradox**, where an observed trend in combined data **reverses or disappears when broken down into subgroups**.
    - ❗ While red wine consumption appears to be linked with better health, this could be a **spurious correlation**.
    - 🎓 The actual **confounding factors**—like higher **education** and **income** levels—might be driving both red wine consumption and improved health.
    - 🔄 If data were analyzed separately by **income or education groups**, the **positive effect** of red wine could **disappear or reverse**.
- **To Conclude:**
  - Simpson's Paradox reminds us to **look deeper into subgroup data**.
    - A correlation in aggregate data might **mask the true relationships** and lead to misleading conclusions if important **confounders are ignored**.

6.7 Collecting Data – Bias Free: Word of Cautions!!!

- **Considerations** when choosing a dataset:
  - What data is necessary to answer our question?
  - **How difficult is it to analyze a dataset?**
  - Is the source authoritative? (.com, .NET, .org, .gov, .name)
  - **Comprehensive data vs. sampled data?**
  - **Biases**
  - What is the allowed usage of data under its license (Copyright issues)?
  - **Who collected the data?**
  - **When was the data collected?**
  - **How was the data collected?**
  - How is the data formatted?
  - Does your data collection procedures need to be approved by an IRB(Review Board)?
  - Confidentiality/Privacy Concerns

7. After Collecting Data: Making it Useful.

{"From Raw to Ready: Cleaning, Organizing, and Exploring Data for Insightful Analysis."}

7.1 Data Exploration

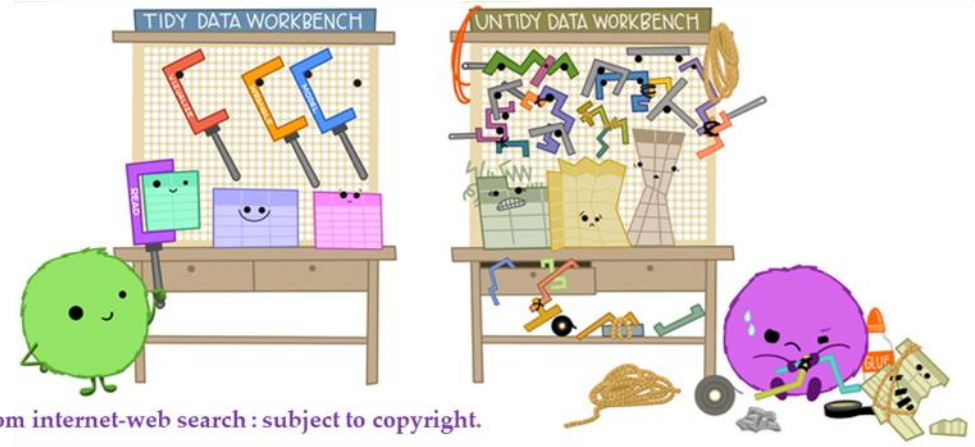


image from internet-web search : subject to copyright.

Fig: Tidying the Data.

Data Wrangling

image from internet – freshspectrum.com : subject to copyright.

Fig: Making Sense of Data.

Data Analysis

⚠ **Caution:**

- While this module briefly references **Data Wrangling** and **Data Analysis**, these topics will **not be explored in depth**.
- Students are **expected to have prior knowledge** of these foundational concepts, which were covered in **Week 2 of Module 5CS037**.
- To ensure a smooth progression through the current material, you are **strongly advised to review** the associated **lecture slides**, **tutorial exercises**, and **workshop activities** from that week.
- A solid understanding of how to clean, transform, and explore datasets is essential for effectively applying statistical and analytical methods in real-world contexts.

7.2 Data Wrangling: A Revision.

- The Process of transforming {raw} data into data that can be consolidated for it's **intended analytical** use case.
- In general, this constitutes of all the pre-processing steps we applied to a {raw} data but can be grouped in following:

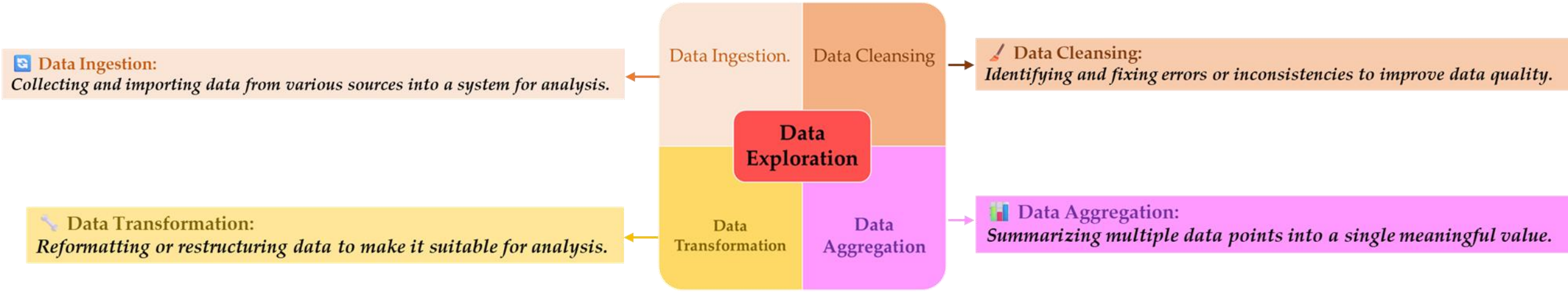


Fig: Major Steps in Data Wrangling Process

7.3 Data Analysis: Understanding Data with Statistics.

- **Statistics:**
  - Statistics is a science whose focus is on **collecting**, **analyzing** and **drawing conclusion** from **data**.
- **Branch of Statistics:**
  - **Descriptive Statistics {Please look into week 2 5CS037} :**
    - describe the **features or characteristics** of data.
    - Summarize and delivers quantitative insights about the data through **numerical** or **graphical representations**.
  - **Inferential Statistics {We will discuss in upcoming weeks}:**
    - used to make **conclusions or inferences** of **entire populations** from the available **sample data**.

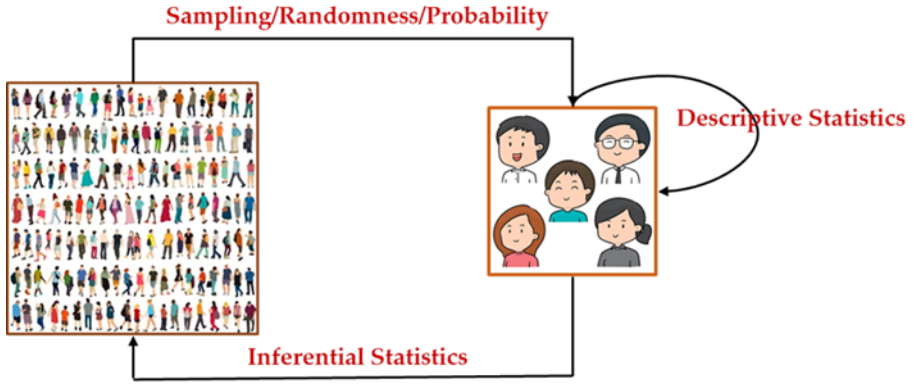


Fig: "Central Dogma" of Statistics.

7.3.1 Descriptive Statistics: A Revision.

- **Describes or summarizes** the **data** i.e. provides **quantitative insights** about **the data** through **numerical** or **graphical** representation.
  - It only **describes or summarize a data** {variables} upon which it is **applied** to.
- There are two ways descriptive statistics of any **data**{variable} can be analyzed:
  - **Visualization aka Graphical Methods {⚜ Discussed in second part}**
  - **Numerical Analysis:**
    - **Measures of Central Tendency {Univariate analysis}: Mean, Median, and Mode**
    - **Measures of Dispersion {Univariate analysis}: Range, Variance, Standard Deviation**
    - **Measures of Relation{Bivariate analysis}: Co-variance and Co-relation.{Will cover upcoming week.}**

7.4 Measure of Central Tendency: A Revision.

Measure	Definition	Population Formula	Sample Formula
Mean	Arithmetic average; Sensitive to Outliers	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	Middle Value when sorted	Middle value of all N values	• Middle value of all n values: • If n is odd: **Median is** $x_{\frac{n+1}{2}}$ • If n is even: **Median** $= \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$
Mode	Most frequently occurring	same	same

Notations	Definition
x_i:	Each individual data point.
n:	Sample Size.
N:	Population Size.
$\bar{x}$:	Sample Mean.
μ:	Population Mean.

7.5 Measure of Dispersion: A Revision.

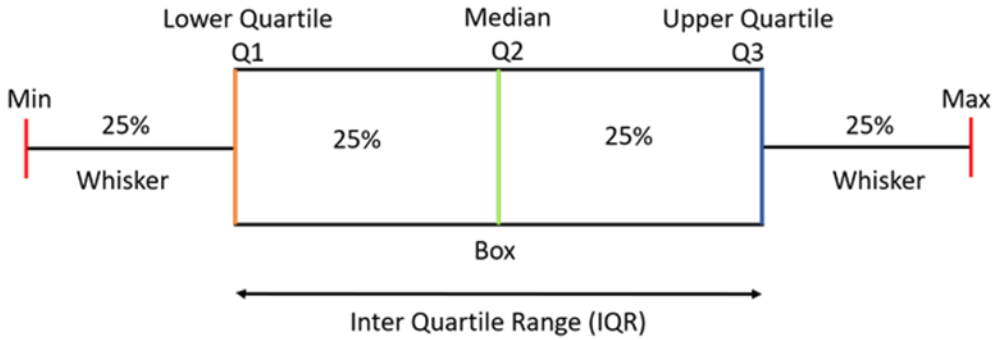
Measure	Definition	Population Formula	Sample Formula
Range	Max - Min	$\textbf{Range} = x_{\max} - x_{\min}$	same
Variance	Avg. squared deviation from center(mean)	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard Deviation	Spread around the mean	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
Interquartile Range	Spread of middle 50%	$IQR = Q_3 - Q_1$	same

Notations	Definition
x_i:	Each individual data point.
n:	Sample Size.
N:	Population Size.
$\bar{x}$:	Sample Mean.
μ:	Population Mean.
s^2, s	Sample Variance and std. dev.
σ^2, σ	Population Variance and std. dev.

7.6 Five Number Summaries.

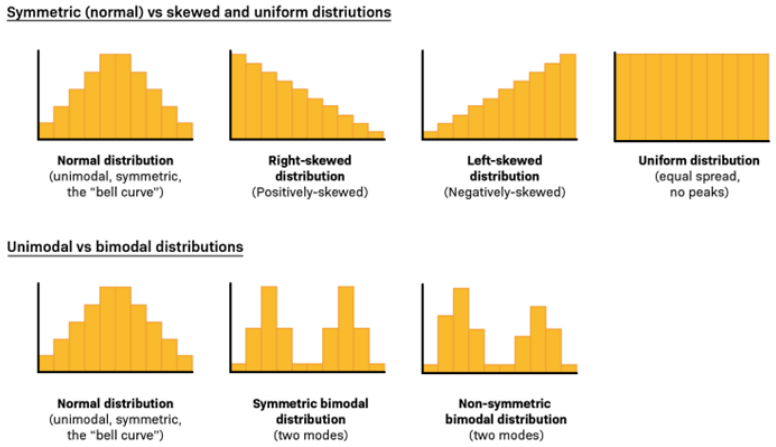
Component	Meaning
Minimum	Smallest value
Q1 – First Quartile	25th percentile
Q2 – Median – Second Quartile	50th percentile
Q3 – Third Quartile	75th percentile
Maximum	Largest value

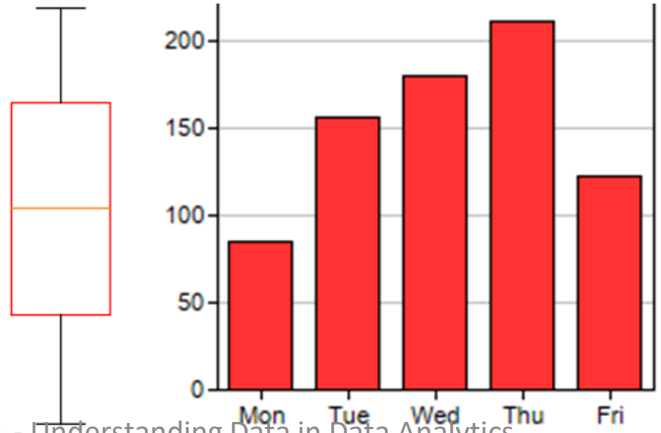
➡️ **Used to create boxplots and assess data distribution.**

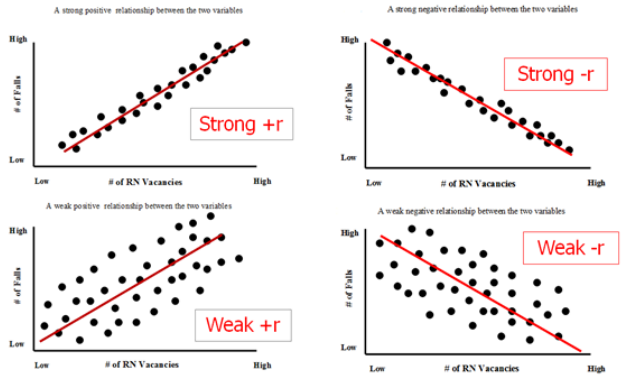


7.7 Visual Summaries of Data.

Plot Type	Best For	Definition
Histogram	Quantitative Data (univariate)	A graphical representations showing the frequency distribution of numerical data by grouping data into bins. Useful to visualize the shape (e.g. skewness, modality) of the distribution.
Boxplot	Spread, Outliers & Comparison	Displays the five – number summary (Min, Q1, Median, Q3, Max) and highlights potential outliers. Helps compare distributions across groups.
Bar chart	Categorical Data	Represents categories using rectangular bars with lengths proportional to the values they represent. Ideal for comparing frequency or proportion across categories.
Scatterplot	Bivariate Relationships	Plots pairs of numerical data to examine potential relationships or patterns (e.g. correlation, clusters, outliers) between two variables.







To Summarize …

Navigating Real-World Data for Meaningful Insights.

- 📥 **Understanding the Source:**
  - Real-world datasets are rarely perfect.
  - Data may be collected via primary or secondary sources.
  - Always assess how the data was generated or sampled.
- 🎯 **Sampling Matters:**
  - Population vs. Sample: Most analysis is based on samples.
  - Use **Random**, **Stratified**, or **Cluster Sampling** as appropriate.
- 🧹 **Data Quality & Wrangling:**
  - **Ingestion → Cleaning → Transformation → Aggregation.**
  - Good data practices improve reliability and reproducibility.
- 📊 **Descriptive Statistics:**
  - Use **Central Tendency** (Mean, Median, Mode) and **Spread** (Range, IQR, Variance, SD).
  - Visual summaries help spot patterns and anomalies:
    - Histogram, Boxplot, Bar Chart, Scatterplot.

💡 **Final Thought:**

"A successful data analytics endeavor begins not with sophisticated algorithms, but with a rigorous understanding of the data's origin, quality, and structure."

The – End.