Data Analysis

Chapter 2

Data Collection & Preparation

Dr. Mahmoud Elsabagh

Contents

Chapter 1: Introduction to Data Analysis

Chapter 2: Data Collection & Preparation

Chapter 3: Exploratory Data Analysis (EDA)

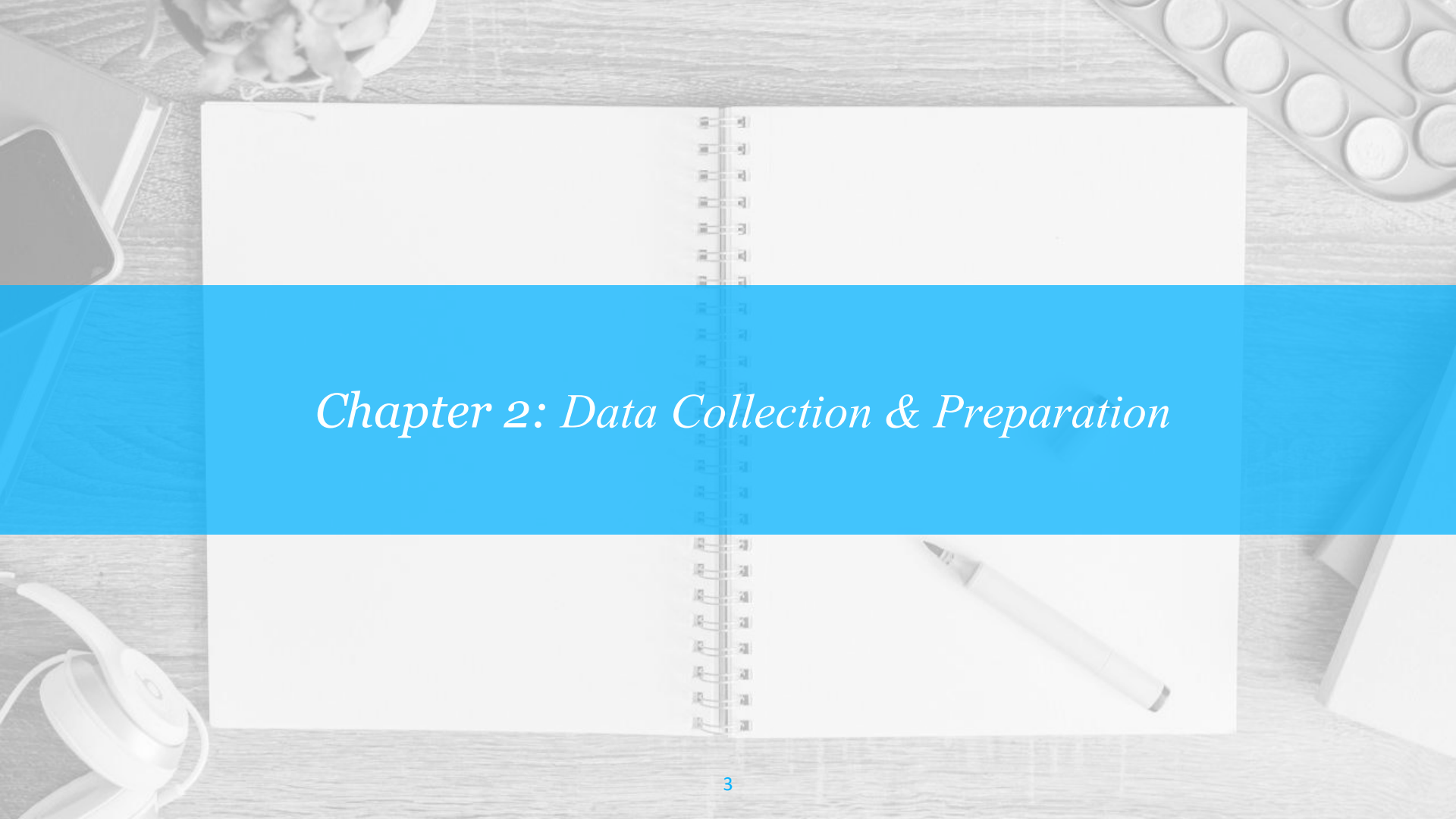
Chapter 4: Statistical Analysis

Chapter 5: Predictive Data Analysis

Chapter 6: Data Analysis Tools & Software

Chapter 7: Communicating Results

Chapter 8: Applications & Future Trends



Chapter 2: Data Collection & Preparation

1. Introduction / Hook

- Start with a question:
- 👉 *"Imagine you're cooking. If the ingredients are spoiled, will the dish taste good?"*
- Answer: *No, because the quality of the output depends on the quality of the input.*
- Message: Good analysis depends on good data. If data is wrong, results will also be wrong.

2. Data Collection – What is it?

Definition:

- ➜ *"Data collection is the systematic process of gathering information from various sources to address a research question, test a hypothesis, or support decision-making."*
- ➜ Analogy: *Like collecting puzzle pieces before solving the puzzle.*
- ➜ Example:

  A school wants to study student performance.

  Data collection: exam results, attendance, participation, and demographics.

3. Methods of Data Collection

a) Primary Data Collection (data collected first-hand)

- ➜ Surveys & Questionnaires – collecting feedback from students/customers.
- ➜ Experiments – conducting controlled studies.
- ➜ Interviews – one-on-one discussions.
- ➜ Observations – watching behavior (e.g., customer in a shop).

3. Methods of Data Collection

b) Secondary Data Collection (data from existing sources)

➜ Databases – government census, company records.

➜ Research articles – published academic papers.

➜ Public datasets – Kaggle, World Bank, WHO, Open Data portals.

➜ Web data – scraping e-commerce sites, social media analytics.

➜ 💡 Example for students:

➜ Primary = *You ask your classmates directly about their favorite study method.*

➜ Secondary = *You download a dataset from Kaggle about student study habits.*

4. Data Sources in the Real World

- Business: Sales transactions, customer feedback.
- Healthcare: Patient records, lab test results, wearable devices.
- Education: Student grades, attendance, online learning logs.
- Government: Census data, tax records.
- Internet: Tweets, blog posts, YouTube comments, website logs.

5. Data Formats

➜ Structured Data (organized in rows and columns)

CSV, Excel, SQL databases

➜ Semi-structured Data (not fully tabular, but has some structure)

JSON, XML, logs

➜ Unstructured Data (no predefined structure)

Text documents, images, videos, audio

➜ Time-series Data (data collected over time)

Stock prices, temperature readings, website traffic

➜ 💡 Example:

➜ CSV → “student_scores.csv” (structured).

➜ JSON → “weather_data.json” (semi-structured).

➜ Tweets or Instagram photos (unstructured).

6. Data Cleaning (Preprocessing)

After collecting, data is usually messy. Cleaning makes it ready for analysis.

- Common problems and solutions:

➜ Missing Values

Cause: Human error, sensor failure.

Solutions: Fill with mean/median, remove rows, or use prediction models.

➜ Duplicates

Cause: Multiple entries for the same record.

Solution: Remove duplicate rows.

➜ Inconsistent Formats

Cause: Different date formats (“12/05/25” vs “2025-05-12”).

Solution: Standardize to one format.

6. Data Cleaning (Preprocessing)

➜ Outliers (unusual values)

Cause: Data entry error, fraud, or genuine rare event.

Solution: Investigate and decide whether to remove or keep.

➜ Noise

Random, irrelevant data.

Solution: Use smoothing techniques, filters, or domain knowledge.

7. Data Transformation

- Normalization: Rescaling values to a small range (e.g., 0–1).

  Example: Salary values from $1,000–$100,000 → scaled down.

- Standardization: Centering data around mean = 0, SD = 1.

  Example: Z-scores in statistics.

- Encoding Categorical Data: Converting text into numbers.

  Example: Gender → Male = 0, Female = 1.

- Feature Engineering: Creating new variables from existing ones.

  Example: From “Date of Birth” → calculate “Age.”

8. Tools for Data Collection & Preparation

- ➜ Excel/Google Sheets – small datasets, basic cleaning.
- ➜ SQL – querying databases.
- ➜ Python (Pandas, NumPy) – professional data wrangling.
- ➜ R – statistical analysis.
- ➜ ETL Tools (Extract, Transform, Load) – Talend, Apache Nifi.
- ➜ Data Visualization Tools – Tableau, Power BI (for quick validation).
- ➜ 💡 Live Demo idea: open a CSV in Excel and remove duplicates.

9. Challenges in Data Collection & Preparation

- ➜ Data Quality Issues: Inaccurate, incomplete, or outdated information.
- ➜ Data Integration: Combining data from multiple sources (Excel + SQL + APIs).
- ➜ Data Volume: Handling big data from sensors, social media.
- ➜ Bias: If the sample isn't representative, results will be misleading.
- ➜ Ethics & Privacy: Handling personal information responsibly (GDPR, HIPAA).

10. Case Study: Online Retail Store

- Data collected: Customer details, transactions, website clicks.
- Problems:
  - Missing email addresses.
  - Duplicate orders.
  - Product categories written differently (“T-shirt” vs “Tee Shirt”).
- Cleaning:
  - Fill missing values, remove duplicates, standardize categories.
- Outcome: Clean dataset → Better customer segmentation → Targeted marketing.

Summary

- Data collection = gathering raw information (primary & secondary).
- Data comes in different formats: structured, semi-structured, unstructured.
- Data cleaning ensures quality and reliability.
- Transformation prepares data for analysis.
- Without good data preparation, even the best analysis is meaningless.

Thanks!

Any questions?