



Data Analysis

Chapter 3

Exploratory Data Analysis (EDA)

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 3: Exploratory Data Analysis (EDA)

1. Introduction / Hook

→ Start with a question:

👉 “*If you receive a dataset with 1,000 student grades, what’s the first thing you do?*”

→ Answer: You don’t start modeling immediately; you explore the data.

→ Analogy: *EDA is like meeting a new friend – you ask questions, get to know them, and understand their habits before making assumptions.*

→ Message: EDA is the first step to make sense of raw data.

2. What is EDA?

→ Definition:

“Exploratory Data Analysis is the process of examining datasets to summarize their main characteristics, often using visual methods.”

→ Purpose:

Understand data distribution.

Detect errors or anomalies.

Find relationships between variables.

Guide next steps in analysis or modeling.

→ Quote (John Tukey, father of EDA):

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

3. Steps in EDA

- Understand your variables – What do columns represent? Which are categorical/numerical?
- Descriptive statistics – Compute measures of central tendency and spread.
- Visualizations – Plot graphs to spot patterns and outliers.
- Correlation & relationships – Check how variables are connected.
- Anomaly detection – Identify unusual data points.
- Hypothesis generation – Ask new questions from insights.
-  EDA = *Summarize, Visualize, Interpret*

4. Descriptive Statistics

- Measures of Central Tendency

- Mean (Average): Sum of values ÷ Number of values.
- Median: Middle value when data is sorted.
- Mode: Most frequent value.

 Example: Student test scores = [60, 65, 65, 70, 80]

- Mean = 68, Median = 65, Mode = 65.

- Measures of Spread

- Range: Max – Min.
- Variance & Standard Deviation (SD): Measure how spread out values are.
- Quartiles & IQR: Useful for detecting outliers.

 Example: If SD is small → data is consistent. If large → data is spread out.

5. Data Distribution

- Normal Distribution: Bell-shaped curve (common in exam scores, heights).
- Skewed Distribution:
 - Right-skewed (e.g., income distribution – most people earn less, few earn very high).
 - Left-skewed (e.g., retirement age – most people retire around 60–65).
- (Visual: Show bell curve + skewed curves.)

6. Data Visualization Techniques

- Univariate Analysis (1 variable at a time)

→ Histogram – shows frequency distribution.

→ Bar chart – for categorical data.

→ Pie chart – for proportions.

- Bivariate Analysis (2 variables)

→ Scatter plot – relationships between two numerical variables.

→ Boxplot – shows spread, median, and outliers.

- Multivariate Analysis (more than 2 variables)

→ Heatmaps – correlation between multiple variables.

→ Pair plots – relationships between all variable pairs.

→  Example: Student study hours vs. exam score → Scatter plot shows positive correlation.

7. Detecting Outliers

→ Outlier: A value far away from the rest.

→ Methods:

Boxplot (points outside whiskers).

Z-score (values > 3 SD away from mean).

→ Example:

Class exam scores = mostly 40–90, but one student = 5 → Outlier.

⚠ Note: Outliers aren't always “bad” – they can reveal important info (e.g., fraud in bank transactions).

8. Correlation Analysis

→ Correlation coefficient (r):

$+1 \rightarrow$ strong positive relationship.

$-1 \rightarrow$ strong negative relationship.

$0 \rightarrow$ no relationship.

→ Example:

Study hours & grades → positive correlation.

Exercise & weight → negative correlation.

Shoe size & intelligence → no correlation.

→💡 Visual: Correlation heatmap of multiple student performance metrics.

9. Case Study: Student Performance Dataset

- Dataset:

→ Variables: Hours studied, Attendance %, Final exam score.

- Analysis:

→ Mean exam score = 72, SD = 10.

→ Histogram: Most students scored between 65–80.

→ Scatter plot: Positive correlation between hours studied & exam score.

→ Outlier: One student studied 20 hours but scored 40 (possible health issue/test anxiety).

- Outcome: EDA helps teacher design better interventions.

10. Tools for EDA

- Excel/Google Sheets: Pivot tables, charts.
- Python (Pandas, Matplotlib, Seaborn): Professional-level EDA.
- R: Strong in statistics & visualization.
- Tableau/Power BI: Business dashboards.

 In-class Demo:

- Load small dataset in Excel → Create histogram & scatter plot.
- Show how visualization makes patterns visible instantly.

11. Challenges in EDA

- Large datasets → harder to visualize.
- Mixed data types (categorical + numerical).
- Missing or inconsistent data.
- Human bias: Analysts may “see” what they want to see.

Summary

- EDA = exploring data before formal analysis.
- Steps: Summarize → Visualize → Interpret.
- Use descriptive stats + visualizations to spot patterns.
- Outliers and correlations provide deeper insights.
- EDA guides future modeling and decision-making.

Thanks!

Any questions?