INTRODUCTION TO DATA SCIENCE

Exercises & Solutions

Tina Moré

Chapter 2 Exercises & Challenges

Data — The Foundation of Data Science

Exercises and Challenges

Thinking Clearly About Data

These exercises are designed to strengthen your understanding of data as a concept, not just as an object. Focus on reasoning, judgment, and explanation rather than technical execution.

Exercise 1: Identifying Data Types in the Real World

For each of the following variables, identify:

1. Whether it is quantitative or qualitative
2. Whether it is discrete or continuous (if quantitative)
3. Its level of measurement (nominal, ordinal, interval, or ratio)

a) Number of children in a household
b) Customer satisfaction rating (Very Unsatisfied → Very Satisfied)
c) Temperature measured in Celsius
d) National ID number
e) Monthly household income

Exercise 2: Diagnosing Data Quality Problems

Below are short scenarios. Identify **which data quality dimension(s)** are violated.

a) A dataset records ages as -3, 25, and 400
b) A customer database lists the same person three times with slightly different names
c) Sales data uses USD in some rows and EUR in others, with no indicator
d) A survey has missing responses for education level in 55% of records
e) A hospital uses patient addresses last updated eight years ago

Exercise 3: The Data Lifecycle in Action

A mobile health app collects daily step counts from users.

Map the following actions to the correct **stage of the data lifecycle**:

a) Encrypting user step data on cloud servers
b) Removing outliers caused by faulty sensors
c) Asking users for permission during app installation
d) Creating a dashboard showing weekly trends
e) Deleting user data after account closure

Exercise 4: Ethics and Responsibility Check

Consider the following case:

A company uses historical hiring data to train a model that recommends job candidates. Past hiring practices favored one demographic group.

Questions:

1. What ethical risks are present?
2. Which principles from this chapter are being challenged?
3. What should a responsible data scientist do?

Exercise 5: Choosing the Right Data Format

For each scenario, recommend the **most appropriate data format** and explain why.

a) Sharing a small dataset with a non-technical manager
b) Receiving real-time data from a web API
c) Storing millions of transaction records with relationships
d) Logging application error messages
e) Exchanging data between two different software systems

Challenge 1: Spot the "Big Data" Myth

A student says:

"I need to learn Spark and Hadoop before I can become a data scientist."

Task:
Evaluate this statement using the concepts from this chapter.

Challenge 2: Reflective Thinking (Written Exercise)

Answer in 5–7 sentences:

How has this chapter changed the way you think about data?

(No single correct answer. Focus on insight, not jargon.)

Note: These exercises prioritize **judgment over memorization**. If a learner can explain *why* a choice is correct, they are developing real data intuition.

Solutions and Guided Answers

These are guided solutions, not templates. Thoughtful alternatives are encouraged if well-reasoned.

Solution: Exercise 1

a) Number of children

- Quantitative
- Discrete
- Ratio
  Explanation: It is numeric, countable, has a true zero, and ratios are meaningful.

b) Customer satisfaction rating

- Qualitative
- Not applicable (not numeric)
- Ordinal
  Explanation: Categories have a meaningful order, but distances between levels are not equal.

c) Temperature in Celsius

- Quantitative
- Continuous
- Interval
  Explanation: Differences are meaningful, but zero does not represent the absence of temperature.

d) National ID number

- Qualitative
- Not applicable

- Nominal
  Explanation: Although numeric in appearance, it is an identifier with no quantitative meaning.

e) Monthly household income

- Quantitative
- Continuous (conceptually)

Ratio

Explanation: Has a true zero and supports meaningful comparisons.

Solution: Exercise 2

a) Ages -3 and 400

- Validity
- Accuracy
  Explanation: Values violate logical and domain constraints.

b) Same person listed multiple times

- Uniqueness
  Explanation: Duplicate representations of a single real-world entity.

c) Mixed currencies without labels

- Consistency
  Explanation: Units and standards are not uniform across records.

d) Missing education level

- Completeness
  Explanation: Large portions of required data are absent.

e) Eight-year-old addresses

- Timeliness
  Explanation: Data is outdated relative to its intended use.

Solution: Exercise 3

a) Encrypting data

- Storage

b) Removing sensor outliers

- Processing / Preparation

c) Asking for permission

- Collection / Generation

d) Creating dashboards

- Analysis (and Sharing if distributed to users)

e) Deleting data

Deletion

Solution: Exercise 4

1. Ethical risks

- Reinforcement of historical bias
- Discrimination against underrepresented groups

2. Challenged principles

- Fairness
- Bias awareness
- Ethical responsibility

3. Responsible actions

- Audit the dataset for bias
- Assess model outcomes across demographic groups
- Consider re-weighting, re-sampling, or alternative features
- Communicate limitations transparently
- Possibly reject the use of the dataset if harm cannot be mitigated

Solution: Exercise 5

a) Non-technical manager

- Excel (.xlsx)
  Explanation: Familiar interface and formatting support.

b) Web API

- JSON
  Explanation: Native to APIs and supports nested structures.

c) Large transaction records

- SQL Database
  Explanation: Supports scale, relationships, integrity, and queries.

d) Application logs

- Text files (.log)

- Explanation: Sequential, unstructured, append-only data.

e) System-to-system exchange

- CSV

  Explanation: Lightweight, universal, and easy to parse.

Solution: Challenge 1

This statement reflects a **misunderstanding of Big Data**.

Most real-world data science problems:

- Fit comfortably on a single machine
- Use CSVs, databases, or spreadsheets
- Focus more on data quality, framing, and interpretation

Big Data tools are **situational**, not foundational. A data scientist should first master:

- Data understanding
- Cleaning and preparation
- Analysis and reasoning

Big Data technologies are learned **when scale demands them**, not before.

Model Answer (Example): Challenge 2

Before this chapter, I viewed data mainly as numbers to analyze. I now understand that data is a representation of reality shaped by human decisions. I see how quality, bias, ethics, and context influence every analysis. I also recognize that data is never perfect and that responsible work involves acknowledging limitations. This chapter has helped me see data science as a discipline of thinking, not just coding.

Chapter 3 Exercises & Challenges

Science — The Methodology of Data Science

Exercises and Challenges

Thinking Like a Scientist with Data

These exercises are designed to strengthen your scientific mindset. Focus less on finding "correct" answers and more on practicing careful reasoning, skepticism, and clarity of thought.

Exercise 1: Correlation or Causation?

For each scenario below, decide whether the relationship described is more likely **causal**, **correlational**, or **unclear**. Briefly justify your reasoning.

1. Cities with more firefighters tend to have more fire damage.
2. Students who attend exam-preparation classes score higher on tests.
3. Ice cream sales and heatstroke cases rise during the same months.
4. Employees who receive promotions report higher job satisfaction.
5. People who exercise regularly have lower rates of heart disease.

Exercise 2: Spot the Skeptical Questions

You are reviewing a report that claims:

"After implementing a new dashboard, company productivity increased by 20%."

List **at least five skeptical questions** you would ask before accepting this conclusion.

Exercise 3: Identify the Pitfall

Match each scenario to the most likely scientific pitfall involved:

Pitfalls (use each at most once):

- Selection bias
- Survivorship bias
- Overfitting

- Confirmation bias
- Confounding variables

Scenarios:

1. A startup accelerator studies only companies that completed its program and concludes the program guarantees success.
2. A model performs extremely well on training data but poorly on new data.
3. A researcher ignores data points that contradict their hypothesis.
4. A survey about internet usage is conducted only via email.
5. A study links coffee consumption to longevity without accounting for income or lifestyle.

Exercise 4: Domain Knowledge Check

For each situation, describe **one domain-specific insight** that would be essential for proper interpretation.

1. Analyzing hospital readmission rates
2. Studying employee turnover in a call center
3. Evaluating student performance across schools
4. Investigating sudden drops in mobile app usage

Exercise 5: Prediction vs. Causation

For each objective below, decide whether **prediction**, **causal understanding**, or **both** are required.

1. Forecasting tomorrow's electricity demand
2. Reducing customer churn
3. Detecting fraudulent credit card transactions
4. Designing a public health intervention
5. Recommending movies to users

Challenge 1: Design a Thought Experiment

You are asked to determine whether **remote work increases employee productivity**.

1. Describe an **ideal randomized experiment** to answer this question.
2. Explain why such an experiment might be difficult or impossible.
3. Propose **one observational approach** to approximate causality.

Challenge 2: The Skeptical Reviewer

Read the claim below and write a short critique (5–7 sentences):

"Our machine learning model increased profits by 30%, proving that AI-driven decisions outperform human judgment."

Your critique should address:

- Evidence quality
- Alternative explanations
- Missing information
- Causal vs. predictive claims

Solutions and Guided Answers

These are guided solutions, not templates. Thoughtful alternatives are encouraged if well-reasoned.

Solution 1: Correlation or Causation

1. **Correlation** – Larger cities have more fires and more firefighters; city size is a confounder.
2. **Unclear** – Motivated students may self-select into prep classes.
3. **Correlation** – Temperature is the confounding variable.
4. **Unclear / Possibly causal** – Promotion may increase satisfaction, but high performers may also be more satisfied beforehand.
5. **Likely causal (with caveats)** – Strong evidence exists, but lifestyle confounders must be considered.

Solution 2: Skeptical Questions

Examples include:

- How is "productivity" defined and measured?
- Were there other changes during the same period?
- Was there a control group without the dashboard?
- Could seasonal effects explain the increase?
- Did productivity improve temporarily or persist?

Solution 3: Scientific Pitfalls

1. Survivorship bias
2. Overfitting
3. Confirmation bias
4. Selection bias
5. Confounding variables

Solution 4: Domain Knowledge Insights

1. Healthcare – Readmissions may reflect case severity, not care quality.
2. Call centers – Turnover may be influenced by shift schedules or emotional labor.
3. Education – Differences in funding, curriculum, or student backgrounds matter.
4. Mobile apps – App updates, outages, or pricing changes may explain usage drops.

Solution 5: Prediction vs. Causation

1. Prediction
2. Both
3. Prediction (primarily)
4. Causation
5. Prediction

Challenge 1: Remote Work Thought Experiment

1. **Ideal experiment:** Randomly assign employees to remote or in-office work while keeping roles and workloads identical.
2. **Why difficult:** Ethical concerns, employee preferences, organizational constraints.
3. **Observational approach:** Use a difference-in-differences design comparing productivity before and after policy changes across teams.

Challenge 2: Skeptical Reviewer (Sample Answer)

The claim lacks sufficient evidence to establish causation. It is unclear how profits were measured and whether external factors influenced the increase. The model may be predictive without directly causing better decisions. No comparison group or time-based controls are mentioned. Additionally, “AI-driven decisions” is vague and may mask human oversight. Stronger causal evidence is needed before drawing such conclusions.

Final Reflection Prompt

Which scientific habit from this chapter—skepticism, domain awareness, or causal reasoning—do you currently underuse the most? How will you practice it intentionally going forward?

Chapter 4: Exercises and Challenges

The Language of Data — Statistics

Exercises & Challenges

Exercise 4.1: Speaking the Language of Statistics

For each scenario below, identify **whether descriptive statistics, inferential statistics, or both** would be most appropriate. Briefly explain your reasoning.

1. A school principal wants to summarize last year's exam results for all students.
2. A pharmaceutical company tests a new drug on 1,000 volunteers and wants to know if it will work for the general population.
3. A retail manager wants to know the average daily sales and how much they fluctuate.
4. A political analyst surveys 2,500 voters to predict national election results.

Exercise 4.2: Choosing the Right Measure of Central Tendency

For each dataset below, decide whether the **mean, median, or mode** is the most appropriate measure of central tendency. Explain why.

1. Household incomes in a city.
2. The most common shoe size sold in a store.
3. Ages of participants in a retirement community.
4. Customer ratings (1–5 stars) for a mobile app.

Exercise 4.3: Understanding Variability

Two delivery companies have the same average delivery time: **45 minutes**.

- Company A has delivery times ranging from **42 to 48 minutes**.
- Company B has delivery times ranging from **20 to 90 minutes**.

1. Which company has higher variability?
2. Why is variability important even when averages are the same?

3. Which company would you prefer as a customer? Why?

Exercise 4.4: Reading the Shape of Data

Match each situation to the most likely **distribution shape**: symmetrical, right-skewed, or left-skewed.

1. Salaries in a large corporation.
2. Test scores on a very easy exam.
3. Heights of adult men in a country.
4. Waiting times at a hospital emergency room.

Exercise 4.5: Mean vs. Median — A Critical Thinking Challenge

A company reports that the **average (mean) salary** of its employees is $80,000. However, you discover that most employees earn between $35,000 and $50,000, while a few executives earn millions.

1. Which measure of central tendency would better represent a "typical" employee's salary?
2. Why might the reported average be misleading?
3. What does this example teach you about statistics and communication?

Exercise 4.6: Spotting Outliers

Consider the following monthly sales figures (in units sold):

120, 130, 125, 128, 127, 129, 500

1. Identify the outlier.
2. Should the outlier automatically be removed? Why or why not?
3. Give one situation where such an outlier could be meaningful rather than an error.

Challenge 4.1: Statistics in Everyday Life

Choose **one real-world scenario** (e.g., salaries, exam scores, rainfall, social media likes, hospital wait times).

Answer the following:

1. What would you want to know about the **center**, **spread**, and **shape** of the data?
2. Which statistical measures would you use?

3. What wrong conclusion could someone draw if they ignored variability or skewness?

Challenge 4.2: From Description to Decision

You are advising a small business owner who only looks at **average monthly revenue** to make decisions.

1. Explain why averages alone are insufficient.
2. Which additional descriptive statistics would you recommend?
3. How could misunderstanding the data lead to poor business decisions?

Solutions & Explanations

Solution 4.1: Speaking the Language of Statistics

1. **Descriptive statistics** — summarizing known exam results.
2. **Inferential statistics** — generalizing from a sample to a population.
3. **Descriptive statistics** — averages and variability describe existing data.
4. **Inferential statistics** — using a sample to predict national outcomes.

Solution 4.2: Choosing the Right Measure

1. **Median** — income data is typically right-skewed with extreme high earners.
2. **Mode** — identifies the most common shoe size.
3. **Mean or Median** — depending on whether ages are evenly distributed.
4. **Mode or Median** — ratings are ordinal and often clustered.

Solution 4.3: Understanding Variability

1. **Company B** has higher variability.
2. Variability shows consistency and risk, not just typical performance.
3. Most customers would prefer **Company A** because delivery times are predictable.

Solution 4.4: Reading the Shape of Data

1. **Right-skewed** — a few very high salaries.
2. **Left-skewed** — most scores are high, few are low.
3. **Symmetrical** — biological traits often follow a normal distribution.

4. **Right-skewed** — long waits for a few patients.

Solution 4.5: Mean vs. Median

1. **Median** better represents the typical employee.
2. Extreme salaries pull the mean upward.
3. Statistics can mislead if context and distribution shape are ignored.

Solution 4.6: Spotting Outliers

1. The outlier is **500**.
2. No — it may reflect a special event or bulk purchase.
3. It could indicate a successful promotion or seasonal demand spike.

Solution to Challenge 4.1 (Sample Guidance)

- Center: mean or median income
- Spread: range or standard deviation
- Shape: skewness reveals inequality
- Ignoring skewness may exaggerate “typical” earnings

Solution to Challenge 4.2 (Sample Guidance)

- Averages hide volatility and risk.
- Recommend standard deviation, trends, and minimum/maximum values.
- Overconfidence during good months could lead to overspending or stock shortages.

Closing Note

These exercises are not about memorizing formulas—they are about **learning to listen to what data is really saying**. If you can interpret center, spread, and shape thoughtfully, you are already thinking like a data scientist.

Chapter 5 Exercises and Challenges

The Art of Asking The Right Question

Exercises and Challenges

Exercise 1: Decision First, Question Second

For each scenario below, identify:

1. The **decision** that needs to be made
2. One **poorly framed question**
3. One **well-framed data question** using the SAAV framework

Scenarios:

1. A mobile app company says: “Our users are not very active.”
2. A hospital administrator says: “Patient wait times feel too long.”
3. An online course creator says: “People aren’t finishing my courses.”

Exercise 2: Apply the SAAV Framework

Evaluate each question below. Identify which SAAV principles it violates and explain why.

1. “How can we make our customers happier?”
2. “Does marketing affect sales?”
3. “Which customer behaviors predict churn in the next 30 days?”
4. “What will the economy look like next year?”

Exercise 3: Stakeholder Translation

Translate each stakeholder statement into **two or more data-oriented questions**.

1. “We need to improve customer engagement.”
2. “Our marketing isn’t working.”
3. “Our product quality is declining.”

Exercise 4: Clarifying Questions in Practice

For each vague request below, write **three clarifying questions** that are:

1. Non-judgmental
2. Collaborative
3. Directional

Requests:

1. "Can you analyze our customer data?"
2. "We want to understand why revenue dropped."

Exercise 5: Spot the Hidden Assumption

Each question below contains at least one hidden assumption. Identify it and rewrite the question without the assumption.

1. "Why are customers unhappy with the new interface?"
2. "Which features are causing users to churn?"
3. "Why does our pricing model fail?"

Challenge: From Business Problem to Data Question

Choose **one real or hypothetical organization** (e.g., school, clinic, small business, NGO, startup).

1. Describe a **business problem** in plain language
2. Identify the **decision** that must be made
3. Write **three candidate data questions**
4. Select the best one and justify it using the SAAV framework

This challenge has no single correct answer—clarity and reasoning matter more than perfection.

Solutions and Sample Answers

Note: These are illustrative solutions. In practice, multiple answers may be valid if they are well-reasoned.

Solution 1: Decision First, Question Second

Scenario 1: Mobile App Usage

- Decision: Whether to redesign features, improve onboarding, or increase notifications
- Poor question: *"Why don't users like our app?"*
- Better question: *"Which user actions are most strongly associated with weekly active usage over the past three months?"*

Scenario 2: Hospital Wait Times

- Decision: Whether to change staffing levels or scheduling procedures
- Poor question: *"Are wait times bad?"*
- Better question: *"What is the average and 90th percentile patient wait time by department and time of day?"*

Scenario 3: Course Completion

- Decision: Whether to redesign content or adjust pacing
- Poor question: *"Why do learners quit?"*
- Better question: *"At which lesson do most learners drop off, and how does this differ by enrollment source?"*

Solution 2: SAAV Evaluation

1. "How can we make our customers happier?"
    - Violates: Specific, Actionable
2. "Does marketing affect sales?"
    - Violates: Specific, Answerable
3. "Which customer behaviors predict churn in the next 30 days?"
    - Meets all SAAV criteria
4. "What will the economy look like next year?"
    - Violates: Answerable, Actionable

Solution 3: Stakeholder Translation

Customer Engagement

- How often do users return within 7 and 30 days?
- Which features are used by our most active users?

Marketing Isn't Working

- Which channels generate leads that convert within 60 days?

- What is the cost per acquisition by channel?

Product Quality Decline

- Have return rates or support tickets increased over time?
- How have customer satisfaction scores changed by product version?

Solution 4: Clarifying Questions

Analyze Customer Data

- What decision are you hoping this analysis will inform?
- Which customer outcome matters most right now?
- What action would you take if the results surprised you?

Revenue Drop

- When did the drop begin, and was it gradual or sudden?
- Did this affect all products or specific segments?
- What explanations are you most concerned about?

Solution 5: Hidden Assumptions

1. Assumption: Customers are unhappy
    - Revised: *"How has user satisfaction changed since the interface update?"*
2. Assumption: Features cause churn
    - Revised: *"Which factors are associated with higher churn rates?"*
3. Assumption: Pricing is failing
    - Revised: *"How does conversion and retention vary across pricing tiers?"*

Challenge: Sample Response (Abbreviated)

Organization: Online Learning Platform

- Business Problem: Learners stop engaging after the first week
- Decision: Whether to redesign onboarding
- Candidate Questions:
    1. What percentage of learners return after 7 days?
    2. Which early actions predict course completion?
    3. How does engagement differ by signup source?

Best Question: *"Which learner behaviors in the first 72 hours are most predictive of course completion?"*

Why: Specific, actionable, answerable, and directly informs design decisions.

Chapter 6: Practical Exercises and Challenges

From Chaos to Clarity: Preparing Your Data

Practical Exercises and Challenges

Exercise 1: Identifying Data Quality Issues

You are given the following dataset excerpt:

Customer_ID	Name	Age	City	State	Join_Date	Spend
101	John Smith	29	Boston	ma	03/15/24	$120.50
102	mary jones		Seattle	WA	2024-03-16	89.99
103	Mike Brown	340	los angeles	CA	15/03/2024	45
104	John Smith	29	Boston	MA	03/15/24	$120.50

Task:

1. List at least five data quality problems you can identify.
2. Categorize each problem (missing data, inconsistency, error, duplicate, formatting issue).

Exercise 2: Handling Missing Values

In the dataset above, the **Age** column is missing for Customer 102.

Task:

1. List **three possible strategies** for handling this missing value.
2. For each strategy, explain **when it would be appropriate**.
3. Choose the strategy you would use if age is required for analysis and explain why.

Exercise 3: Standardizing Text Data

You are cleaning a **City** column with the following values:

Boston

boston

BOSTON

Los Angeles

los angeles

Los Angeles

Task:

1. Define a standard format for the City column.
2. Describe the steps you would take to transform all values into that standard.
3. Explain why text standardization is critical before analysis.

Exercise 4: Fixing Dates

You encounter these date values in an **Order_Date** column:

03/15/2024

15/03/2024

2024-03-15

03-15-24

40/15/2024

Task:

1. Identify which dates are valid and which are invalid.
2. Choose a standard date format and justify your choice.
3. Explain how you would handle the invalid date.

Exercise 5: Duplicate Detection

Consider the following records:

Order_ID	Customer	Email	Order_Date
201	Sarah J.	sarah@email.com	2024-03-10
201	Sarah J.	sarah@email.com	2024-03-10
202	Sarah J.	sarah@email.com	2024-03-15

Task:

1. Identify which records are true duplicates.
2. Explain why the remaining records should or should not be removed.
3. Define what makes a record “unique” in this dataset.

Exercise 6: Error or Outlier?

You find the following values in an **Age** column:

-3, 0, 27, 105, 400

Task:

1. Identify which values are clearly errors.
2. Identify which values might be legitimate but unusual.
3. Explain how you would handle each value and why.

Exercise 7: Feature Engineering

You have the following columns:

- Date_of_Birth
- Order_Date
- Order_Amount
- Number_of_Items

Task:

1. Propose **three new features** you could engineer from this data.
2. Explain how each engineered feature could improve analysis or modeling.

Exercise 8: Reshaping Data

You receive sales data in the following format:

Product	Jan	Feb	Mar
A	120	135	150
B	90	110	130

Task:

1. Explain why this is considered *wide format*.
2. Describe how you would reshape it into *long format*.
3. Explain which format you would prefer for trend analysis and why.

Challenge: The Data Cleaning Plan

You are given a **new, unfamiliar dataset** from a marketing team.

Task:
Create a **step-by-step data cleaning plan** that includes:

- Initial assessment checks

- Strategies for missing data
- Rules for inconsistencies
- Duplicate handling
- Validation steps
- Documentation practices

Focus on **decision-making**, not tools or code.

Solutions and Explanations

Solution 1: Identifying Data Quality Issues

Problems identified:

1. Missing age (Customer 102) → Missing data
2. Age = 340 → Obvious error
3. State codes inconsistent (ma vs MA) → Formatting inconsistency
4. City capitalization inconsistent → Text inconsistency
5. Mixed date formats → Date inconsistency
6. Spend stored as text with $ → Data type issue
7. Duplicate record (John Smith appears twice with identical data) → Duplicate

Solution 2: Handling Missing Values

Possible strategies:

1. **Delete the record** – appropriate if missing values are rare and non-critical
2. **Impute with median age** – appropriate when age is required and similar records exist
3. **Flag and analyze separately** – appropriate when missingness may be meaningful

Chosen strategy: Impute with median age and create a flag column, because age is required and only one value is missing.

Solution 3: Standardizing Text Data

Standard format: Proper case (e.g., "Los Angeles")

Steps:

1. Trim leading/trailing spaces

2. Collapse multiple spaces
3. Convert all text to proper case

Why it matters: Without standardization, identical cities are treated as different categories, leading to incorrect counts and misleading results.

Solution 4: Fixing Dates

Valid dates:

- 03/15/2024
- 15/03/2024
- 2024-03-15
- 03-15-24

Invalid date:

- 40/15/2024 (impossible month)

Standard chosen: ISO format (YYYY-MM-DD)

Handling invalid date: Flag for review rather than guessing, and document the issue.

Solution 5: Duplicate Detection

True duplicate:

- Order 201 appears twice identically → remove one

Remaining record:

- Order 202 is a separate transaction → keep

Uniqueness definition: Order_ID uniquely identifies each order.

Solution 6: Error or Outlier?

Value	Decision
-3	Error → mark missing or remove
0	Context-dependent → investigate
27	Valid
105	Rare but possible → flag
400	Error → investigate or mark missing

Key principle: **Do not remove unusual values unless you are confident they are wrong.**

Solution 7: Feature Engineering

Possible engineered features:

1. **Customer Age** = Order_Date − Date_of_Birth
2. **Average Price per Item** = Order_Amount ÷ Number_of_Items
3. **Customer Tenure** = Order_Date − First_Purchase_Date

These features provide richer behavioral and temporal insights than raw variables.

Solution 8: Reshaping Data

Why wide format: Each month is stored in a separate column.

Long format structure:

Product	Month	Sales

Preferred format for trends: Long format—because it simplifies filtering, aggregation, and visualization over time.

Challenge Solution: Data Cleaning Plan

A strong plan includes:

1. Initial profiling (row counts, missing values, ranges, …)
2. Duplicate detection and rules for uniqueness
3. Standardization of text, dates, and categories
4. Thoughtful handling of missing and erroneous data
5. Creation of flags for uncertainty
6. Validation checks
7. Documentation of every decision

The goal is **trustworthy, analyzable data—not perfection**.

Chapter 7: Practical Exercises and Challenges

Exploring Data: Uncovering Hidden Stories (Conceptual EDA)

Practical Exercises and Challenges

Exercise 7.1: Understanding Your Dataset at First Sight

You are given a dataset containing the following variables from a retail coffee shop:

- Date
- Daily_Sales
- Customer_Count
- Avg_Transaction_Value
- Temperature
- Wait_Time

Tasks:

1. List the first five EDA questions you would ask before looking at any visualizations.
2. Identify which variables are numerical and which are categorical or time-based.
3. Explain why answering these questions is critical before deeper analysis.

Exercise 7.2: One Variable, Many Clues

Focus on the variable Daily_Sales.

Tasks:

1. Describe which visualization you would use to explore Daily_Sales and why.
2. List at least three distribution characteristics you would examine.
3. Identify two possible real-world reasons for extreme values (outliers).

Exercise 7.3: Exploring Relationships Between Variables

You suspect that both Customer_Count and Temperature influence Daily_Sales.

Tasks:

1. Identify the most appropriate plot for analyzing the relationship between:
    - Daily_Sales and Customer_Count
    - Daily_Sales and Temperature
2. Describe what a strong positive relationship would look like.
3. Explain how you would tell whether a relationship is weak or misleading.

Exercise 7.4: Adding Context with a Third Variable

Assume the dataset also includes a categorical variable called Day_Type with values:

- Weekday
- Weekend

Tasks:

1. Explain how you would incorporate Day_Type into a scatter plot.
2. Describe two insights that might emerge from this multivariate view.
3. Explain why this insight might be invisible in a bivariate analysis.

Exercise 7.5: Interpreting a Correlation Matrix

You calculate the following correlations:

Variable Pair	Correlation (r)
Daily_Sales & Customer_Count	0.94
Daily_Sales & Temperature	-0.81
Customer_Count & Wait_Time	-0.65
Avg_Transaction_Value & Wait_Time	-0.10

Tasks:

1. Identify the strongest and weakest relationships.
2. Explain what these correlations imply operationally.
3. Identify one pair that could signal a bottleneck or risk.

Exercise 7.6: Segment Analysis Challenge

The business segments customers into:

- Morning Regulars
- Lunch Crowd
- Weekend Socials
- Grab-and-Go

Tasks:

1. Explain why segmenting data is important during EDA.
2. Identify one metric you would compare across all segments.
3. Suggest one actionable decision for each segment based on EDA findings.

Exercise 7.7: Spotting Simpson's Paradox

Overall analysis shows that customer satisfaction has increased over time. However, when broken down by store location, satisfaction has decreased at most individual stores.

Tasks:

1. Explain how Simpson's Paradox could be occurring here.
2. Identify what additional data you would examine.
3. Describe one incorrect conclusion and one correct conclusion.

Exercise 7.8: EDA Time-Boxing Strategy

You have only **6 hours** to perform EDA on a new dataset.

Tasks:

1. Propose how you would allocate time across EDA phases.
2. Explain what you would *not* do in this time-boxed scenario.
3. Define what "enough understanding" means in practical terms.

Solutions and Explanations

Solution 7.1: Understanding Your Dataset

1. Key questions:
    - What does each variable represent?
    - What is the time range?
    - Are values complete and realistic?
    - What is the unit of measurement?
    - What decision will this data support?
2. Numerical: Daily_Sales, Customer_Count, Avg_Transaction_Value, Temperature, Wait_Time
   Time-based: Date

3. These questions prevent misinterpretation, incorrect assumptions, and wasted analysis effort.

Solution 7.2: One Variable Exploration

1. A **histogram** best reveals distribution shape and spread.
2. Key characteristics:
    - Central tendency
    - Skewness
    - Outliers
3. Possible outliers:
    - Holidays or promotions
    - Store closures or data entry errors

Solution 7.3: Relationship Exploration

1. Scatter plots for both relationships.
2. Strong positive relationship:
    - Points form an upward-sloping cloud
3. Weak or misleading:
    - Wide scatter
    - No consistent pattern
    - Influence driven by a few extreme points

Solution 7.4: Multivariate Insight

1. Use color to distinguish Day_Type.
2. Possible insights:
    - Weekend sales may be higher at the same customer count
    - Temperature effects may differ by day type
3. Without segmentation, these behavioral differences are averaged away.

Solution 7.5: Correlation Interpretation

1. Strongest: Daily_Sales & Customer_Count (0.94) Weakest: Avg_Transaction_Value & Wait_Time (-0.10)
2. Implications:
    - Sales growth depends heavily on traffic

    - Long waits reduce customer flow
3. Risk signal:
    - Customer_Count & Wait_Time → service capacity issue

Solution 7.6: Segment Analysis

1. Segments reveal hidden behavioral differences.
2. Compare:
    - Average transaction value
3. Actions:
    - Morning Regulars: Loyalty rewards
    - Lunch Crowd: Faster service
    - Weekend Socials: Premium offers
    - Grab-and-Go: Improve satisfaction speed

Solution 7.7: Simpson's Paradox

1. The mix of customers shifted toward high-satisfaction stores.
2. Examine:
    - Store-level trends
    - Customer volume per store
3. Incorrect conclusion: "All stores are improving"
   Correct conclusion: "Overall satisfaction rose due to customer redistribution"

Solution 7.8: Time-Boxed EDA

1. Suggested allocation:
    - Basics: 1 hour
    - Univariate visuals: 1.5 hours
    - Relationships: 2 hours
    - Segmentation: 1 hour
    - Documentation: 0.5 hour
2. Skip:
    - Deep modeling
    - Edge-case optimization
3. Enough understanding:
    - Key drivers identified

- Major risks spotted
- Clear next questions defined

Chapter 8: Practical Exercises and Challenges

From Patterns to Predictions: Introduction to Modeling

Practical Exercises and Challenges

Exercise 1: Is This a Model?

For each of the scenarios below, decide whether it represents a **model** as defined in this chapter. If yes, briefly explain *what is being modelled* and *what is being simplified*.

1. A supermarket manager estimates tomorrow's sales by averaging the last four Saturdays.
2. A GPS app predicts your arrival time using distance, traffic, and speed limits.
3. A checklist that says: "If it's raining, carry an umbrella."
4. A spreadsheet that predicts monthly electricity costs using past usage and temperature.

Exercise 2: Regression or Classification?

For each question below, identify whether the task is best solved using **regression** or **classification**.

1. How many units will we sell next month?
2. Will this customer churn? (Yes / No)
3. What will the delivery time be (in minutes)?
4. Is this transaction fraudulent, suspicious, or legitimate?
5. How many students will pass the exam?

Exercise 3: Supervised or Unsupervised?

Determine whether each scenario describes **supervised learning** or **unsupervised learning**, and explain why.

1. A bank trains a model using past loan applications labeled "approved" or "rejected."

2. A retailer groups customers based on shopping behavior without predefined labels.
3. A hospital predicts patient diagnoses using historical patient records with confirmed outcomes.
4. A cybersecurity team detects unusual network activity without knowing in advance what attacks look like.

Exercise 4: Training vs. Testing — Spot the Mistake

A data analyst reports:

"My model predicts customer churn with 96% accuracy."

You later discover that the model was evaluated on the **same dataset it was trained on**.

1. Why is this a problem?
2. What risk does this pose in real-world use?

Exercise 5: Generalization Challenge

Two students build models to predict exam scores.

- **Student A** builds a very complex model that predicts training data perfectly.
- **Student B** builds a simpler model that makes small errors on training data.

Which model is more likely to perform better on new data—and why?

Exercise 6: Overfitting or Underfitting?

Identify whether each situation is an example of **overfitting**, **underfitting**, or **good fit**.

1. A model predicts nearly the same value for all inputs and performs poorly everywhere.
2. A model performs extremely well on training data but fails badly on new data.
3. A model performs reasonably well on both training and testing data.

Exercise 7: Choosing the Right Model Type

You are given the following business questions. For each, recommend a **model type** and briefly justify your choice.

1. Forecast daily coffee sales.
2. Identify groups of customers with similar purchasing habits.

3. Decide whether to approve, reject, or review a loan application.
4. Estimate the lifetime value of a customer.

Exercise 8: Model Evaluation Thinking

A medical diagnosis model has:

- High accuracy
- Low recall for detecting a serious disease

1. Why could this be dangerous?
2. What metric should the team prioritize improving—and why?

Exercise 9: The Human-in-the-Loop Scenario

A model predicts very high sales for next Tuesday.

Later, a manager overrides the prediction because:

"Next Tuesday is a national holiday, and the store will be closed."

1. Did the model fail?
2. What does this example teach about the role of humans in modeling?

Exercise 10: Reflection Challenge (Short Answer)

In your own words, complete the sentence:

"Models should inform decisions, not replace them, because..."

Solutions and Explanations

Solution 1: Is This a Model?

1. **Yes** — It models future sales using historical averages, simplifying demand patterns.
2. **Yes** — It models travel time, simplifying complex traffic dynamics.
3. **Yes (very simple model)** — It models weather-based decision-making with a single rule.
4. **Yes** — It models electricity cost using past usage and temperature, ignoring many real-world variables.

Key idea: A model does not have to be mathematical or complex—it just has to simplify reality usefully.

Solution 2: Regression or Classification?

1. Regression
2. Classification
3. Regression
4. Classification
5. Regression

Rule of thumb:

- Predicting a **number** → Regression
- Predicting a **category** → Classification

Solution 3: Supervised or Unsupervised?

1. Supervised — outcomes are labeled.
2. Unsupervised — no predefined labels.
3. Supervised — known diagnoses guide learning.
4. Unsupervised — the model discovers anomalies without labels.

Solution 4: Training vs. Testing

1. The model was tested on data it already saw, so performance is misleading.
2. The model may fail badly on real customers because it memorized patterns instead of learning general ones.

This is a classic case of overconfidence caused by data leakage.

Solution 5: Generalization Challenge

Student B's model is more likely to perform better on new data.

Why:

- Small training errors suggest learning real patterns.
- Perfect training performance often signals memorization (overfitting).

Solution 6: Overfitting or Underfitting?

1. **Underfitting** — model is too simple.

2. **Overfitting** — model memorizes noise.
3. **Good fit** — balanced learning and generalization.

Solution 7: Choosing the Right Model Type

1. **Regression** — predicting numerical sales values.
2. **Clustering** — discovering natural customer segments.
3. **Classification** — discrete decision categories.
4. **Regression** — predicting a numerical lifetime value.

Solution 8: Model Evaluation Thinking

1. High accuracy but low recall means many sick patients are missed — a serious risk.
2. **Recall** should be prioritized to ensure most actual cases are detected.

In healthcare, missing positives is often worse than false alarms.

Solution 9: Human-in-the-Loop Scenario

1. No — the model worked correctly within the limits of its data.
2. It shows that:
    - Models lack context,
    - Humans provide situational awareness,
    - Decisions require judgment beyond predictions.

Solution 10: Reflection Challenge (Sample Answer)

“Models should inform decisions, not replace them, because they lack context, ethical reasoning, and awareness of real-world exceptions that humans understand.”

(Other well-reasoned answers are equally valid.)

Chapter 9: Practical Exercises and Challenges

From Insight to Impact: Communicating with Data

Practical Exercises and Challenges

Exercise 1: From Findings to Story

You conducted an analysis and discovered the following:

- Customer churn increased from 8% to 12% over six months
- The increase is concentrated among customers using the mobile app
- App users report slower performance after a recent update

Task:
Rewrite these findings as:

1. A facts-only summary
2. A short data story that includes context, insight, and action

Exercise 2: Choosing the Right Visualization

For each scenario below, choose the **most appropriate visualization** and briefly explain why:

a) Comparing revenue across 10 product categories
b) Showing how daily website traffic changes over a year
c) Understanding the distribution of customer wait times
d) Examining the relationship between marketing spend and sales
e) Showing how revenue contributions from products change over time

Exercise 3: Spot the Misleading Chart

You are shown a bar chart comparing profits for two departments:

- Department A: $4.8M
- Department B: $5.1M

The y-axis starts at $4.5M instead of zero, making Department B's bar appear nearly twice as tall.

Task:

1. Explain why this visualization is misleading
2. Describe how you would fix it

Exercise 4: Executive Summary Writing

You analyzed customer support data and found:

- Password reset tickets increased 60%
- Most resets come from users logged out overnight
- A new security policy shortened session duration

Task:
Write a **one-page executive summary** (200–300 words max) that includes:

- The business question
- The key insight
- A clear recommendation
- Expected impact

Exercise 5: Dashboard Design Challenge

You are asked to design a dashboard for a retail operations manager who checks performance **daily**.

Task:
Define:

- 5 key metrics to include
- One visualization for each metric
- One example of context (comparison, trend, or target) for each metric

Challenge 6 (Advanced): Ethics in Communication

You discover a statistically significant result, but:

- The effect size is small
- The sample excludes an important subgroup

Task:
Explain how you would:

1. Communicate this result honestly
2. Visualize it responsibly
3. Prevent misinterpretation by non-technical stakeholders

Solutions and Explanations

Solution 1: From Findings to Story

Facts-only summary: Customer churn increased from 8% to 12% over six months. The increase is concentrated among mobile app users. App performance complaints increased after a recent update.

Story-based version: Over the past six months, customer churn has risen sharply, from 8% to 12—a trend that initially puzzled the team. When we segmented churn by platform, a clear pattern emerged: nearly all of the increase came from mobile app users. Desktop users showed no meaningful change. Digging deeper into customer feedback revealed a consistent complaint—slower app performance following the latest update. This suggests the churn increase is not a loss of interest, but a response to degraded user experience. By optimizing app performance and addressing load-time issues, we have a clear opportunity to reduce churn and retain high-value mobile customers.

Solution 2: Choosing the Right Visualization

a) Bar chart – Best for comparing quantities across categories
b) Line chart – Best for showing change over continuous time
c) Histogram – Reveals distribution, skewness, and outliers
d) Scatter plot – Shows relationship between two continuous variables
e) Stacked area chart – Shows how parts contribute to a whole over time

Solution 3: Spot the Misleading Chart

1. **Why it's misleading:** Starting the y-axis at $4.5M exaggerates the difference between departments. The actual difference is modest (~6%), but the chart visually implies a dramatic gap.
2. **How to fix it:**
    - Start the y-axis at zero, or

- Clearly mark the axis break and annotate the true percentage difference

Solution 4: Executive Summary

Executive Summary

We investigated the cause of a 60% increase in customer support tickets related to password resets over the past quarter. Analysis of 38,000 support interactions revealed that most resets occur when users are unexpectedly logged out overnight. This behavior coincides with the implementation of a new security policy that reduced session duration from 30 days to 8 hours.

Our findings indicate that the increase in support volume is not driven by user confusion or security breaches, but by friction introduced by the new policy—particularly affecting users who return daily but not continuously.

We recommend extending session duration to 72 hours for trusted devices while maintaining stricter limits for high-risk scenarios. This change balances security with usability and aligns with industry best practices.

If implemented, we estimate a 40–50% reduction in password reset tickets, saving approximately $150,000 annually in support costs while improving user satisfaction and retention.

Solution 5: Dashboard Design Challenge

Key Metrics and Visualizations

1. Daily sales revenue → KPI card with trend vs. yesterday
2. Order volume → Line chart with 7-day rolling average
3. Fulfillment time → Bar chart vs. SLA target (Service level agreements targets)
4. Inventory stockouts → Table or alert card highlighting exceptions
5. Return rate → Line chart compared to last month

Each metric includes:

- A comparison to target or prior period
- Clear visual hierarchy
- Minimal text for quick scanning

Solution 6: Ethics in Communication

1. **Honest communication:** Clearly state that while the result is statistically significant, the effect size is small and may have limited practical impact.
2. **Responsible visualization:** Use appropriate scales, show confidence intervals, and avoid exaggerated visual emphasis.
3. **Preventing misinterpretation:** Explicitly explain limitations, note the excluded subgroup, and recommend further analysis before decision-making.

This approach maintains credibility, protects stakeholders from overconfidence, and aligns with ethical data practice.