

5

Exploratory Data Analysis and Visualization

The Purpose of EDA

Exploratory data analysis (EDA), a term coined by John Tukey, is the disciplined practice of summarising and visualising data before committing to a formal model. In health data science, EDA has a specific and non-optional role: it is often the stage at which data quality problems left over from Chapter 4's cleaning process reveal themselves, and it is the stage at which an analyst forms hypotheses about relationships (does severity of adolescent depressive symptoms differ by household food security status? does test positivity vary by rainfall season?) that the formal modelling chapters will later test rigorously.

Univariate Summaries

For a single variable, the choice of summary depends on its type (Section 4.1). Continuous variables such as haemoglobin concentration are summarised by measures of centre (mean, median), spread (standard deviation, interquartile range), and shape (skewness); the median and IQR are generally preferred over the mean and standard deviation for health measurements known to be skewed, such as clinic wait times or CD4 counts. Categorical variables such as diagnosis category are summarised by frequency tables and proportions, always reported alongside the denominator

(a common error in reporting Nigerian facility-level data is presenting a percentage without stating how many tests or visits it was computed from).

Bivariate and Multivariate Exploration

Relationships between two variables are explored through cross-tabulations (for two categorical variables), grouped boxplots or violin plots (for a continuous variable across categories, such as birthweight by maternal ANC attendance category), and scatterplots with a smoothed trend line (for two continuous variables, such as rainfall and monthly malaria case counts). Correlation coefficients (Pearson's r for linear association, Spearman's ρ for monotonic association robust to outliers) provide a numeric summary but must never be interpreted causally — a point re-emphasised throughout the regression chapters, since a strong correlation between distance to the nearest health facility and undernutrition prevalence does not by itself establish that distance causes undernutrition; both may be driven by underlying poverty.

Visualization Principles for Health Communication

A visualisation intended for a technical audience (a co-author, a methods reviewer) can carry more statistical detail than one intended for a policy audience (an LGA health officer, a State Ministry of Health briefing). Good practice that applies to both: always label axes and units, always show the denominator behind a rate or proportion, avoid 3-D effects and unnecessary decoration ('chartjunk') that obscure the data, choose a colour palette that remains interpretable to colour-blind readers, and be deliberate about the baseline of a y-axis, since a truncated axis can visually exaggerate a modest change in, for example, immunisation coverage.

PYTHON example

```
# Python: visualising monthly malaria TPR trend for a state, with a rainfall overlay
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("state_malaria_monthly.csv", parse_dates=["month"])

fig, ax1 = plt.subplots(figsize=(9, 5))
ax1.plot(df["month"], df["tpr"], color="firebrick", marker="o", label="Test positivity rate")
ax1.set_ylabel("Test Positivity Rate (%)")
ax1.set_xlabel("Month")

ax2 = ax1.twinx()
ax2.bar(df["month"], df["rainfall_mm"], alpha=0.25, color="steelblue", label="Rainfall (mm)")
ax2.set_ylabel("Rainfall (mm)")

fig.suptitle("Malaria Test Positivity vs. Rainfall, [State], 2023-2024")
fig.tight_layout()
plt.savefig("tpr_rainfall_trend.png", dpi=200)
```

EDA as Hypothesis Generation, Not Hypothesis Confirmation

A discipline this chapter insists on: patterns found through open-ended exploration of a dataset should not be reported with the same confidence as a pre-specified hypothesis test, because searching many possible relationships in a dataset inflates the chance of finding an apparently striking but spurious pattern (a manifestation of the multiple comparisons problem formalised in Chapter 7). Good practice is to treat EDA findings as candidate hypotheses to be confirmed — ideally on an independent dataset or a held-out subsample — rather than as conclusions in their own right.

Worked Example**A Complete Univariate-to-Multivariate EDA Walkthrough**

Consider a cleaned dataset of 500 adolescents from a two-wave cohort, with variables for depressive symptom score (continuous, 0-27), household food security (categorical: secure/mildly insecure/severely insecure), sex, and school attendance (yes/no).

A complete EDA walkthrough proceeds: (1) univariate — histogram of the symptom score reveals a right-skewed distribution with a long tail of severely affected adolescents, so the median (score 6) and IQR (3-11) are reported rather than the mean; (2) bivariate — a grouped boxplot of symptom score by food security category shows a visibly higher median score in the severely insecure group, and a cross-tabulation of school attendance by food security category shows lower attendance in the severely insecure group; (3) multivariate — a scatterplot of baseline against follow-up symptom score, coloured by food security category, suggests that the food-security gap present at baseline widens by follow-up rather than narrowing. None of these three observations is yet a tested finding — they are candidate hypotheses that Chapter 8’s regression modelling (adjusting for sex, school attendance, and baseline score simultaneously) will test formally.

Figure 5.1 — Symptom Score by Household Food Security (Worked Example, Sec.

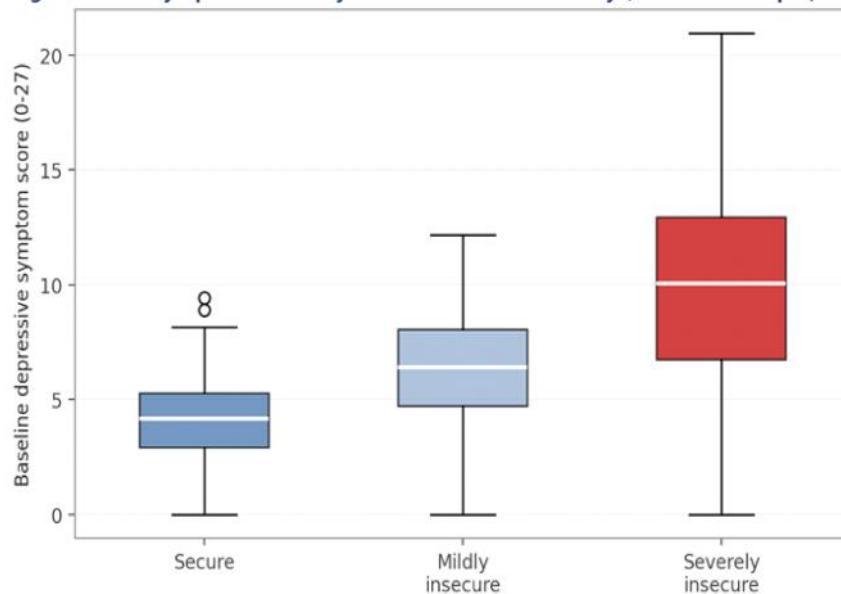


Figure 5.1 — Depressive symptom score by household food security (Section 5.6 worked example).

R example

```
# R: a compact EDA walkthrough matching the worked example above
library(dplyr); library(ggplot2)

# Univariate
summary(cohort$symptom_score_w1)
ggplot(cohort, aes(x = symptom_score_w1)) + geom_histogram(bins = 15, fill = "steelblue")

# Bivariate: boxplot by food security
ggplot(cohort, aes(x = food_security, y = symptom_score_w1, fill = food_security)) +
  geom_boxplot() + theme_minimal() + labs(y = "Baseline symptom score", x = "")

# Cross-tabulation
table(cohort$food_security, cohort$school_attendance)

# Multivariate: baseline vs follow-up, coloured by food security
ggplot(cohort, aes(x = symptom_score_w1, y = symptom_score_w2, colour = food_security)) +
  geom_point(alpha = 0.6) + geom_smooth(method = "lm", se = FALSE) + theme_minimal()
```

Common EDA Pitfalls in Practice

- Reporting a mean and standard deviation for a visibly skewed variable (e.g., clinic wait time) rather than the more appropriate median and IQR, giving a misleading picture of a ‘typical’ value.
- Interpreting a strong Pearson correlation as evidence of a linear relationship without first plotting the scatterplot, which can reveal a non-linear pattern (e.g., a threshold effect) or an outlier driving the entire correlation.
- Presenting a percentage change (e.g., ‘30% increase in cases’) without the underlying counts, which can dramatically overstate or understate practical importance when the base count is very small (a rise from 3 to 4 cases is technically a 33% increase).
- Treating a pattern found by examining dozens of cross-tabulations as a confirmed finding, rather than as one candidate hypothesis among many that must survive formal testing (Chapter 7) and, ideally, replication.

Small Multiples: Comparing Many Groups at Once

When exploring a variable across many groups simultaneously — for example, monthly TPR trends across all 36 states plus the FCT — a single overlaid plot with 37 lines quickly becomes unreadable. The ‘small multiples’ technique instead produces a grid of many small, identically-scaled panels, one per group, allowing a reader to scan for both the overall pattern and any state that stands out as an outlier, without the visual clutter of a single overcrowded chart. This technique, alongside the reusable-function approach of Chapter 6, is one of the most practically useful visualisation habits for a Nigerian health data scientist working with state- or LGA-level panel data, since it scales gracefully to dozens of groups in a way that a single combined plot does not.

R example

```
# R: small multiples of TPR trend by state using ggplot2 facets
library(ggplot2)

ggplot(malaria_clean, aes(x = report_month, y = tpr)) +
  geom_line(colour = "firebrick") +
  facet_wrap(~ state, ncol = 6) +
  theme_minimal(base_size = 8) +
  labs(title = "Monthly TPR Trend by State", x = NULL, y = "TPR")
```

Extended Case Study

When EDA Reveals a Data Collection Problem, Not a Health Pattern

Perhaps the single most valuable real-world function of exploratory data analysis, more valuable in practice than any individual chart type, is its capacity to reveal that an apparent ‘finding’ is actually an artefact of how the data was collected — and this extended case study works through a realistic example of exactly that discovery process. An analyst exploring a multi-site adolescent health cohort dataset notices, via a simple bar chart of mean depressive symptom score by data collection site, that one site’s mean score (4.2) is dramatically lower than all seven other

sites (ranging from 9.8 to 13.1) — an apparent finding that, taken at face value, might suggest that site’s adolescent population is genuinely, and remarkably, less symptomatic than every other site in the study.

Following the EDA discipline of this chapter (treating a striking pattern as a hypothesis to be interrogated, not a finding to be reported), the analyst next examines the distribution of individual item-level responses at the low-scoring site specifically, rather than just its summary mean, and discovers that the response distribution is suspiciously concentrated on the lowest possible response option for nearly every item on the symptom checklist — a pattern much more consistent with a systematic data collection or translation problem (for example, an interviewer at that site consistently anchoring respondents toward the lowest response option, or a translation of the instrument into a local language that inadvertently softened the wording of several items) than with a genuine underlying difference in adolescent mental health at that site. A follow-up conversation with the field team confirms that this particular site had, in fact, used a locally back-translated version of the instrument that had not gone through the same formal validation process as the other sites’ materials. Reporting the original site-level difference as a substantive finding — without this EDA-driven investigation — would have produced a published claim about geographic variation in adolescent depressive symptoms that was, in reality, entirely an artefact of instrument translation quality, a mistake that a purely confirmatory analysis (jumping straight to a regression model without this exploratory step) would have had no natural opportunity to catch.

A Short Checklist for Reviewing Someone Else’s Exploratory Plots

Reviewing a colleague’s or student’s exploratory analysis is a routine task for anyone in a supervisory or peer-review role, and

a short, consistent checklist makes this review faster and more thorough than an unstructured read-through. Useful questions to ask of any exploratory plot presented to you: Does the axis start at zero, or is the baseline chosen in a way that could exaggerate an apparent pattern (Section 5.4)? Is the denominator behind any displayed rate or percentage stated somewhere in the figure or its caption? If a mean is shown, does the underlying distribution's shape (visible in an accompanying histogram, or inferable from context) justify using the mean rather than the median? And, most importantly, is this plot being presented as an established finding, or is it appropriately framed as one exploratory observation among several that still requires formal testing or independent confirmation (Section 5.5)?

Interactive Exploration Versus Static Reporting

The static charts emphasised throughout this chapter (histograms, boxplots, scatterplots) remain the right choice for a fixed report or publication, but an increasingly valuable complementary tool for the EXPLORATION phase itself is an interactive dashboard (built, for example, with R Shiny or Python's Streamlit/Dash) that allows an analyst — or, more importantly, a non-technical domain expert collaborator — to filter, drill down, and re-plot a dataset in real time without needing to write new code for each new question that arises during a collaborative review meeting. This is particularly valuable in the Nigerian context when reviewing a multi-state or multi-facility dataset together with programme staff who have direct, on-the-ground knowledge of specific facilities or states, since they can immediately flag an implausible or interesting pattern in a specific facility of interest during the meeting itself, rather than needing to submit a follow-up request and wait for a new static chart to be produced and circulated afterward.

Annotated Exploratory Plots

Turning EDA into a Communication Tool

While Section 5.1 framed EDA primarily as a private, hypothesis-generating activity for the analyst, a well-annotated exploratory plot can also serve as an effective communication tool for a technical audience, provided it retains explicit statistical caveats rather than being polished into a false appearance of confirmed finding. A practical technique is to add a brief textual annotation directly onto an exploratory plot — for example, a note reading ‘pattern observed in n=42 facility-months; not yet tested for statistical or practical significance’ directly on a suggestive scatterplot — which preserves the plot’s exploratory honesty even as it circulates beyond the analyst’s own working notes to a supervisor or a co-author who might otherwise mistake a preliminary pattern for a confirmed result.

Dashboards vs. Reports

Choosing the Right Format for the Audience

An interactive dashboard (allowing a user to filter by state, LGA, or time period on demand) and a static written report (presenting a fixed, curated set of findings with accompanying narrative interpretation) serve genuinely different purposes, and this chapter’s exploratory analysis skills feed into both, but the choice between them should be deliberate rather than default. A dashboard suits an audience that will return to the data repeatedly with different, evolving questions (a State M&E team monitoring an ongoing programme); a written report suits an audience that needs a specific, interpreted, decision-ready answer to a specific question at a specific point in time (a technical working group deciding whether to adjust a treatment protocol). A common and avoidable mistake is building an elaborate interactive dashboard for an audience that actually needed a two-page interpreted memo, or writing a static report for an audience that will need to re-slice the same data dozens of times over the following months.

A Reusable EDA Summary Function

Rather than re-typing the univariate and bivariate exploration commands from Section 5.6 for every new dataset, wrapping them into a single reusable function that returns a structured summary — one that can itself be inspected, tested, or rendered into a report template — pays off quickly once EDA becomes a routine, repeated step across many datasets and projects, exactly the same efficiency argument made for reusable reporting functions in Chapter 6.

PYTHON example

```
# Python: a compact, reusable EDA summary function
def eda_summary(df, outcome_col):
    summary = {
        "n": len(df),
        "n_missing_outcome": int(df[outcome_col].isna().sum()),
        "median": df[outcome_col].median(),
        "iqr": (df[outcome_col].quantile(0.25), df[outcome_col].quantile(0.75)),
        "skew": df[outcome_col].skew(),
    }
    summary["likely_skewed"] = abs(summary["skew"]) > 1
    return summary

result = eda_summary(cohort_df, "symptom_score_w1")
print(result) # e.g. flags likely_skewed=True, prompting median/IQR over mean/SD (Section 5.2)
```

Chapter Summary

- EDA summarises and visualises data to detect quality issues and generate hypotheses before formal modelling.
- Choice of summary statistic (mean/SD vs. median/IQR) should match the variable's distribution and type.
- Correlation does not establish causation; bivariate exploration should be interpreted cautiously.
- Visualisations for policy audiences must be honest about axes, denominators, and scale.
- Patterns discovered through open-ended EDA require independent confirmation before being treated as findings.

- A complete EDA walkthrough moves systematically from univariate, to bivariate, to multivariate exploration, generating — not confirming — hypotheses at each stage.

Mini-Project

Mini-Project 5: A Full EDA Report for a Health Dataset

Produce a structured, three-part EDA report (univariate, bivariate, multivariate) for a dataset of your choice, following the pattern of Section 5.6.

Step 1. Select a dataset with at least one continuous outcome and at least two categorical or continuous predictors of interest.

Step 2. Produce and interpret at least two univariate summaries, choosing appropriate statistics based on distribution shape.

Step 3. Produce and interpret at least two bivariate visualisations (one categorical-by-continuous, one continuous-by-continuous).

Step 4. Produce one multivariate visualisation (e.g., a scatterplot coloured or faceted by a third variable).

Step 5. List three candidate hypotheses your EDA suggests, explicitly labelling them as hypotheses to be tested (not findings), and state which chapter's method you would use to test each.

Deliverable: A short EDA report (plots plus 300-500 words of interpretation) ending with three clearly labelled candidate hypotheses.

End-of-Chapter Exercises

1. Given a facility-level dataset with monthly TPR, propose one univariate, one bivariate, and one multivariate exploratory plot, and state what each is intended to reveal.
2. A news report shows a bar chart of vaccination coverage with the y-axis starting at 70% rather than 0%. Explain how this could mislead a reader, using a concrete numeric example.
3. Why should a relationship discovered by scanning 30 variables for correlations with an outcome be treated differently from a single pre-specified hypothesis test?
4. In the worked example of Section 5.6, explain why the median and IQR were chosen over the mean and standard deviation for the depressive symptom score.

5. A report states ‘malaria cases in LGA X rose 33% this month.’ You learn the actual counts were 3 cases last month and 4 this month. Write two sentences explaining why the percentage figure alone is misleading here.
6. Using the R code in Section 5.6, describe what change in the boxplot you would expect if severely food-insecure adolescents had, in reality, both a higher median symptom score and much greater variability than the other groups.
7. List two of the EDA pitfalls in Section 5.7 that you have personally observed (or could imagine occurring) in a report you have read or written, and explain briefly how you would correct each.
8. Explain why a small-multiples plot (Section 5.8) would be more useful than a single overlaid line chart for comparing TPR trends across all 36 states plus the FCT.
9. Propose one dataset from your own work where a small-multiples approach would be more informative than a single combined chart, and explain why.
10. What is one disadvantage of small multiples compared to a single combined chart, particularly when comparing exact values between two specific groups?
11. In the Section 5.9 case study, explain why looking at the distribution of individual item-level responses (not just the mean) at the low-scoring site was the key diagnostic step.
12. Propose one additional check, beyond the item-level distribution, that could have helped confirm or rule out the translation-quality explanation earlier.
13. Explain why jumping directly to a regression model comparing sites, without first doing the exploratory work

in Section 5.9, would likely have missed this problem entirely.

14. If the low-scoring site's item-level distribution had instead looked similar in shape to the other sites' distributions (just shifted lower), would this change your interpretation? Explain why or why not.
15. Describe how you would report this discovered instrument-translation issue in a study's limitations section, in two to three sentences.
16. Using the checklist in Section 5.10, review a chart you have seen recently (in a report, news article, or presentation) and identify one question from the checklist it fails.
17. Why is 'is the denominator stated' one of the most important questions on this checklist, in your view?
18. Explain why a reviewer should ask whether a plot is presented as a finding or a hypothesis, even if the plot itself is technically accurate.
19. Propose one additional question you would add to the checklist in Section 5.10 for reviewing a health-related exploratory plot.
20. How would you handle giving this kind of critical feedback to a junior colleague or student in a constructive, non-discouraging way?
21. Using Section 5.11, explain one advantage an interactive dashboard offers during a collaborative review meeting that a set of static charts cannot.
22. Why might a non-technical programme staff member using an interactive dashboard catch an implausible pattern that a data scientist working alone might miss?

23. Explain why static charts remain the right choice for a final published report, even if interactive tools were used during exploration.
24. Propose one risk of relying only on an interactive dashboard for exploration, without ever producing a static, fixed record of what was found.
25. Would you recommend an interactive dashboard for the mini-project in Section 5.9's kind of instrument-translation investigation? Explain your reasoning.
26. Explain why this chapter treats EDA as a distinct, essential stage rather than something to skip in favour of going straight to formal modelling.
27. Summarise the common thread linking the instrument-translation case study (Section 5.9) and the reviewing-someone-else's-plots checklist (Section 5.10).
28. Why does this chapter warn against treating a striking EDA pattern as a finding, even when it looks statistically obvious?
29. Propose one exploratory check you will now perform routinely, having read this chapter, that you did not previously perform.
30. If you had only one exploratory plot to make before trusting a new dataset, which type would you choose, and why?
31. Using Section 5.13, describe a situation from your own work where a dashboard would be more appropriate than a static report, and vice versa.
32. What is the risk of building an elaborate interactive dashboard for an audience that only needed a two-page memo?

33. Explain why an audience that will ‘re-slice’ the same data repeatedly is better served by a dashboard than a fixed report.
34. Propose one hybrid approach that could combine some benefits of both a dashboard and a written report for a specific audience.
35. How should the choice between a dashboard and a report affect the EDA techniques (Sections 5.2-5.11) an analyst emphasises when preparing it?