Session 3 Handout: Data Quality and Fairness

FAST-05 Applied Data Science and ML for Business Users · 7formations · Instructor: Wadood Chaudhary · All examples fictional

Why this session matters

A model can't be better than the data it learned from. Duplicates, blanks, inconsistent codes, and skewed samples quietly become wrong predictions, and sometimes unfair ones that land on the same group of people every time. You can spot most of these problems with a spreadsheet and a few good questions, before anyone trains anything.

Key concepts

1. Garbage in, garbage out

Problem	What it looks like	Why it matters
Duplicates	Same record twice	Overweights some cases; inflates counts
Missing values	Blanks, "N/A", 0 used as blank	If missing isn't random, filling or dropping creates bias
Inconsistent formats	03/04/2026 vs 2026-04-03; Y/N vs 1/0	Misread dates; mislabeled outcomes
Outliers and errors	Lead time of 900 days	Distorts averages and patterns

2. Sampling and bias

Who's in the data determines who the model works for. If 90% of training rows come from one office, one customer type, or one season, the model may perform poorly, or unfairly, for everyone else.

Proxies: even if a protected characteristic isn't a feature, others (ZIP code, insurance type, commute distance) can stand in for it. Ask: if this model is wrong, who bears the cost, and is it always the same group?

3. Drift

Models assume the future looks like the past. When something changes (a new reminder system, a big customer's new ordering pattern, a pandemic, a price change), predictions drift. Watch for errors that jump after a known change, and plan for retraining or human override.

4. Practical data checks (no coding)

- Count rows by group (PivotTable / COUNTIFS).
- Count blanks per column, and by group.
- List distinct values in key columns (codes, labels).
- Check date formats and ranges.
- Remove exact duplicates (on a copy).
- Ask the data owner about any process changes during the period.

AI tools can profile a table quickly. Confirm one number by hand.

Worked example: Charles River Family Dental no-show file (fictional, IDs only)

Finding	Size	Consequence
Skewed sample	Newton 90%, Waltham 10% (similar real volumes)	Waltham predictions rest on ~120 rows
Missing lead time	112 rows, 104 Waltham	Filling with 0 would make Waltham look like same-week bookings
Mixed formats	Waltham dates MM/DD; no_show 1/0	Misread dates; mixed labels
Duplicates	18 rows	Inflated counts
No Saturdays	Waltham opens Saturdays	Model never saw Saturday patterns
Drift	Text reminders since Jan 2025	Likely overpredicts no-shows now

Fairness note: Double-booking based on these scores would mostly affect Waltham patients with public insurance and long commutes, who could face longer waits. Lower-harm alternative: reminder calls for high scores, tested by office first.

Checklist: before trusting a dataset

- ☐ Rows counted by key groups; sample matches reality?
- ☐ Blanks counted, and by group.
- ☐ Codes and date formats consistent.
- ☐ Duplicates removed (on a copy).
- ☐ Process changes during or after the data period identified.
- ☐ Proxy features considered; affected groups named.

Do and don't

Do	Don't
Ask why values are missing	Fill blanks with zero by default
Compare the sample to reality	Assume "lots of data" means representative
Name who bears the cost of errors	Call a model "objective"
Prefer lower-harm actions on scores	Use scores to deny or downgrade service without review
Log process changes as drift risks	Assume the past predicts the future

Try this at work

1. Take one non-sensitive dataset your team relies on.
2. Run the six checks above.

3. Write one finding with size, location, and consequence.

Glossary

Term	Meaning
Skewed sample	Data over-representing some groups or conditions
Proxy	A feature that stands in for a protected or sensitive attribute
Drift	Change in the world that makes past patterns less valid
Missing not at random	Blanks concentrated in a group or condition
De-identification	Removing identifiers per specific rules (HIPAA has two methods)

Key takeaways

- Most model problems start as data problems you can see in a spreadsheet.
- Who's in the data decides who the model works for.
- Watch for drift after any change.