Tutorial 1 — Student Handout

Z3003 Data Mining: Scaling to Large Data Thursday 17 September 2026, 4:45–5:30 pm

1 Learning Outcomes

By the end of this tutorial, you will be able to:

- Load a real dataset into pandas and correctly report its shape and structure
- Classify every column as nominal, ordinal, binary, or numeric — and defend borderline cases
- Detect and quantify missing data, and reason about how that approach changes at 50 million rows
- Compute the full set of central-tendency/dispersion statistics from Friday's lecture on a real dataset
- Produce and interpret a histogram and a boxplot from real data

2 What You Need

- This handout (read alongside the notebook)
- The starter notebook: `W1_Tutorial1_STUDENT.ipynb`
- The dataset: `datasets/W1_zanzibar_crop_sample.csv`
- A working Python environment — choose **either** option below (both fully supported)

3 Environment Setup

3.1 Option A — Google Colab (no installation required)

1. Go to colab.research.google.com and sign in with any Google account.
2. Click **File → Upload notebook**, and upload `W1_Tutorial1_STUDENT.ipynb`.
3. In the Colab file browser (folder icon, left sidebar), click the upload icon and upload `W1_zanzibar_crop_sample.csv`. Create a folder named `datasets` first and upload it there, OR change the file path in the notebook to match wherever you uploaded it.
4. Run the first code cell (Section A) to confirm pandas/numpy/matplotlib are available — Colab includes these by default.

Note: files uploaded directly to Colab are deleted when your session ends. To persist your work across sessions, mount your Google Drive instead (folder icon → mount Drive) and save the dataset there.

3.2 Option B — Local Jupyter / VS Code

1. Confirm Python 3.10+ is installed: open a terminal and run `python --version`
2. Install required packages: `pip install pandas numpy matplotlib jupyter`
3. Place `W1_Tutorial1_STUDENT.ipynb` and the `datasets/` folder (containing the CSV) in the same parent folder.
4. Open the notebook either in VS Code (install the "Jupyter" extension, then open the `.ipynb` file directly) or by running `jupyter notebook` from that folder in a terminal.
5. Run the first code cell (Section A) to confirm your environment works.

4 Session Plan & Time Budget (45 minutes)

Time	Section	What you do
4:45–4:50	A. Environment check	Confirm your environment works (already done for you)
4:50–5:00	B. Load & classify	Load the CSV; classify every column by attribute type
5:00–5:08	C. Data quality	Quantify missing values; inspect affected rows
5:08–5:18	D. Statistics & viz	`describe()`, `groupby()`, one histogram, one boxplot
5:18–5:30	E. Graded exercise	5-part exercise — submit before you leave

5 Theory Recap Before You Start

Attribute types (Lecture 2): nominal ($=, \neq$ only) — binary (nominal, 2 states) — ordinal (adds $<, >$) — numeric (adds $+, -, \times, \div$).

Central tendency / dispersion (Lecture 3):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{IQR} = Q_3 - Q_1$$

A value is an outlier if it falls below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$.

Remember: pandas' `df.dtypes` reports *storage* type only (`int64`, `float64`, `object`) — this is *not* the same as the nominal/ordinal/numeric classification from Lecture 2. You must reason about the true type yourself.

6 The Dataset: Zanzibar Smallholder Farms

Column	Data-mining type	Notes

`farm_id`	Nominal (identifier)	Never average this — pandas won't stop you, but it's meaningless
`region`, `crop`	Nominal	5 categories each
`farm_size_ha`, `yield_tonnes`, `rainfall_mm`, `price_per_kg`, `revenue_usd`	Numeric (ratio)	`yield_tonnes` has ∼5% missing values (realistic field data)
`fertilizer_used`	Binary (symmetric)	Yes/No, both states equally important

7 Working Through the Notebook

Open `W1_Tutorial1_STUDENT.ipynb`. Section A is completed for you as a worked example — read it carefully to see the expected style and level of detail. From Section B onward, replace every `# TODO` with real code, and every `#>>>` line with your own written answer in the markdown cell provided.

- **Section B** — Load the dataset, inspect it, and classify all 9 columns by attribute type
- **Section C** — Quantify missing data and explain the float64/int64 dtype-promotion behavior
- **Section D** — Compute summary statistics and produce a histogram and a boxplot
- **Section E** — The 5-part graded exercise (see rubric below)
- **Optional (ungraded)** — a timing experiment comparing 500 rows vs. a 25-million-row replicated copy

8 Graded Exercise Rubric

#	Task	What "complete" looks like
1	Justify `fertilizer_used` as binary symmetric	2+ sentences referencing the formal definition from Lecture 2
2	1.5×IQR outlier count on `revenue_usd`	Shows Q_1, Q_3, IQR, upper bound, and the resulting filtered table
3	Mean `yield_tonnes` by region + crosstab	Both outputs shown, with a 1-sentence interpretation
4	Reasoning at 50 million rows	References the single-pass/streaming argument from Lectures 2–3, not just "it would be slow"

5	Original rainfall vs. yield chart	Any chart type; axes labeled; 1-sentence interpretation of the relationship

9 Common Errors

Symptom	Cause	Fix
`FileNotFoundError` on `read_csv`	Wrong working directory / dataset not uploaded (Colab)	Check the path; on Colab, confirm the CSV is uploaded to the exact folder referenced
Mean of `yield_tonnes` looks wrong / NaN	Column has missing values	Use `.mean()` (skips NaN by default), not a manual sum/len
`dtype` shows `float64` for an integer-like column	Any NaN forces promotion to `float64`	Expected behavior — explain it in Section C, don't drop rows to "fix" it
Plot doesn't appear	Non-interactive backend	Use `plt.savefig(...)` then view the file, or run `%matplotlib inline` first on Colab

10 Submission

- Save your notebook with **all outputs visible** (do not clear outputs before saving)
- File name: `YourName_W1_Tutorial1.ipynb`
- Submit via the course portal before 5:30pm today
- Not separately graded for marks, but the most direct preparation for Quiz 1 (Fri 6 Nov) and your mini-project's data-understanding section