Tutorial 2 Student Handout

Thursday 24 September 2026, 4:45–5:30pm

1 Learning Outcomes

- Compute a correlation matrix and interpret its strongest and weakest pairs
- Build and interpret a Q-Q plot to assess normality
- Compute Euclidean, Manhattan, and Jaccard distance between two real records
- Run and interpret a χ^2 test of independence between two nominal attributes

2 What You Need

- This handout
- The starter notebook: `W2_Tutorial2_STUDENT.ipynb`
- The dataset (same as Tutorial 1): `datasets/W1_zanzibar_crop_sample.csv`
- Additional package: `scipy` — install with `pip install scipy` if not already present (Colab has it by default)

3 Session Plan & Time Budget (45 minutes)

Time	Section	What you do
4:45–4:50	A. Correlation matrix	Worked example, done for you
4:50–5:00	B. Scatter & Q-Q plots	Visualize the strongest correlation; check normality
5:00–5:10	C. Distance & similarity	Euclidean, Manhattan, Jaccard between two farms
5:10–5:20	D. Chi-square test	Test independence of fertilizer_used and crop
5:20–5:30	E. Graded exercise	5-part exercise — submit before you leave

4 Theory Recap Before You Start

Correlation (Lecture 4):

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}, \qquad \mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$r \in [-1, 1]$; magnitude close to ± 1 means a strong linear relationship. **Remember Lecture 4's warning: correlation is not causation — always consider confounders.**

Distance measures (Lecture 5):

$$\text{Euclidean: } \sqrt{\sum_k (x_k - y_k)^2} \qquad \text{Manhattan: } \sum_k |x_k - y_k| \qquad \text{Jaccard distance: } \frac{r+s}{q+r+s}$$

where q = both-1, r = only-first-1, s = only-second-1 in a binary contingency table.

Chi-square test (previewed Lecture 4, developed Lecture 6):

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

A large χ^2 (small p-value) means the two nominal attributes are significantly associated, not independent.

5 Working Through the Notebook

Open `W2_Tutorial2_STUDENT.ipynb`. Section A (the correlation matrix and heatmap) is completed for you. From Section B onward, replace every `# TODO` with real code, and every `#>>>` line with your own written answer.

- **Section B** — Scatter plot of the strongest correlated pair; Q-Q plot of `revenue_usd`
- **Section C** — Distance/similarity between two farms (remember: drop missing values first!)
- **Section D** — χ^2 independence test between `fertilizer_used` and `crop`
- **Section E** — The 5-part graded exercise

6 Graded Exercise Rubric

#	Task	What "complete" looks like
1	Correlation: farm_size_ha vs. revenue_usd	Value computed, sign and magnitude interpreted in 1–2 sentences
2	Euclidean distance, two specific farms	Correct normalization shown, distance value reported
3	Q-Q plot for `yield_tonnes`	Plot produced, shape compared to `revenue_usd`'s in 1–2 sentences
4	χ^2 test: region vs. crop	Statistic, dof, and p-value reported with a conclusion
5	Reasoning at 5 million rows	References the $O(d^2 \times n)$ argument from Lecture 4, not just "it would be slow"

7 Common Errors

Symptom	Cause	Fix

`ValueError: array must not contain infs or NaNs` on a distance function	A chosen farm has a missing `yield_tonnes` value	Drop incomplete rows first: `df.dropna(subset=[...])`
`ModuleNotFoundError: scipy`	scipy not installed locally	`pip install scipy`; already available on Colab
Q-Q plot looks empty or wrong shape	Missing values not dropped before `stats.probplot`	Use `.dropna()` on the column first

8 Submission

- Save your notebook with all outputs visible
- File name: `YourName_W2_Tutorial2.ipynb`
- Direct preparation for Quiz 1 and your mini-project's exploratory data analysis section