**MINI-PROJECT | 20% OF THE FINAL GRADE**

# Mine a Real Dataset, End to End

*Announced: Thursday 1 October 2026 (Tutorial 3) | Dataset approval: Friday 23 October | Final submission: Friday 20 November 2026, 11:59 pm*

**Summary:** choose a real dataset, prepare it with the techniques of Weeks 1–3, apply ONE mining technique from Weeks 4–9, and explain what you found to someone who is not a data scientist.

## 1. Why this project

Quizzes test whether you can compute a z-score or an information gain. This project tests whether you can **use** them: decide what a messy real dataset needs, justify each choice, run a mining method correctly, and communicate the result honestly. It is the foundation for Part 2 of the course.

## 2. What you will hand in

| Deliverable | Format | File name |
|---|---|---|
| Analysis notebook: all code, outputs and written reasoning inline | .ipynb (outputs visible) | YourName_MiniProject.ipynb |
| One-page plain-language summary for a non-technical reader | PDF (or the notebook's final cell) | YourName_MiniProject_Summary.pdf |
| Dataset, or a link and description if larger than 50 MB | CSV / link | YourName_MiniProject_Data.csv |

## 3. Choosing your dataset

| Requirement | Minimum |
|---|---|
| Size | at least 300 rows and 5 attributes |
| Attribute types | numeric AND at least one of nominal / ordinal / binary (Lecture 2) |
| Source | real and public, or data you may legally use (e.g. a cooperative's records with written permission) |
| Zanzibar crop data | allowed only if you extend it meaningfully, e.g. merge real weather or price data, or ask a genuinely new question |

### Where to look

| Source | Good for |
|---|---|
| UCI Machine Learning Repository | classic, well-documented tables (classification and clustering) |
| Kaggle Datasets | wide variety; check the licence and the documentation |
| FAOSTAT (Food and Agriculture Organization) | crop production, yields, prices by country and year |
| World Bank Open Data | development indicators for Tanzania and the region |
| Humanitarian Data Exchange (HDX) | health, food security and population data for East Africa |
| National Bureau of Statistics, Tanzania | official survey and census tables |

**Checkpoint, Friday 23 October:** submit your dataset choice (name, source link, rows × columns, and the question you want to answer) for approval. Not separately graded, but starting without approval is at your own risk.

## 4. Required sections and what 'good' looks like

### 1. Data understanding (Tutorial 1)

- Shape of the data and what one row represents
- Every attribute classified as nominal / binary / ordinal / numeric, with a one-line justification
- Missing values counted per column, with a guess at WHY they are missing

**Example of a strong sentence:** "24 of 500 farms (4.8%) lack yield. All 24 are spread across crops, so the values appear to be missing at random."

### 2. Exploratory analysis (Tutorial 2)

- Summary statistics and at least two informative plots
- At least one correlation, with a named plausible confounder (Lecture 4)
- At least one distance or similarity suited to your attribute types (Lecture 5)

**Example of a strong sentence:** "Price and revenue correlate at r = 0.92, but crop type drives both: cloves are expensive AND high-revenue."

### 3. Preprocessing (Tutorial 3)

- Missing-data strategy, justified
- Normalization method chosen from the data's shape and fitted on training data only (no leakage)
- Discretization OR PCA, or a reasoned explanation of why neither is needed

**Example of a strong sentence:** "Revenue is right-skewed (mean 5,136 vs median 2,207), so we used the robust z-score. PCA: 4 components keep 99% of the variance."

### 4. Mining (choose ONE from Weeks 4–9)

- Association rules (Apriori / FP-Growth) with support and confidence thresholds justified, OR
- A classifier evaluated on a held-out test set (accuracy plus a confusion matrix), OR
- Clustering with a justified number of clusters and an evaluation measure (e.g. silhouette)

**Example of a strong sentence:** "k-means with k = 3 chosen by the elbow and silhouette (0.41); cluster 2 is small, high-value clove farms in Pemba."

### 5. Findings and discussion

- Results in plain language: what should a farmer, manager or ministry DO differently?
- One scaling consideration: what changes if your data were 1,000× larger?
- One ethical or social consideration specific to your data (Lecture 1)

**Example of a strong sentence:** "At 500,000 farms we would stratify-sample by region before clustering; exact PCA would be replaced by randomized PCA."

## 5. Grading rubric (100 marks)

| Section | Marks | Excellent (full marks) | Adequate (about half) | Weak (little credit) |
|---|---|---|---|---|
| Data understanding | 15 | every attribute typed and justified; missingness explained | types listed, little justification | types missing or wrong |
| Exploratory analysis | 20 | correct statistics; a genuine confounder; distance suited to the types | correct but generic | errors, or no interpretation |
| Preprocessing | 20 | every choice justified by the data; no leakage | sensible defaults, little justification | leakage or unjustified steps |
| Mining | 25 | correct method, justified parameters, proper evaluation | runs, but weak evaluation | incorrect or unevaluated |
| Findings & discussion | 20 | clear for non-experts; specific scaling and ethics points | clear but generic | jargon, or missing parts |

## 6. Timeline

| Date | Milestone |
|---|---|
| Thu 1 Oct | Brief released in Tutorial 3 |
| Fri 23 Oct | Dataset approval checkpoint |
| Week 5 and 6 tutorials | In-class project work sessions, bring questions |
| Fri 20 Nov, 11:59 pm | Final submission |

## 7. Policies

- No late submissions are accepted. Work submitted after the deadline receives zero
- Cite every external dataset, library and code snippet you use
- Work is individual unless your instructor approves a pair; you must be able to explain every line you submit
- AI assistants may be used for debugging and explanation; you must state how you used them in a final 'Tools used' cell

## 8. Frequently asked questions

### Can I use a dataset from my own research or job?

Yes, if you have permission to use it and can share at least an anonymised sample for grading.

### Do I have to use both discretization and PCA?

No. Use at least one, or a clear argument for why neither helps your data.

### My dataset has 300 000 rows and my laptop is slow.

Sample it (Lecture 8) and explain your sampling choice.

### Which mining technique should I choose?

The one that answers your question: 'what goes together?' → rules; 'predict a label' → classification; 'find natural groups' → clustering.

## 9. Submission checklist

| ☐ | Before you submit |
|---|---|
| ☐ | Notebook runs top to bottom without errors; all outputs visible |
| ☐ | Every preprocessing choice has a one-sentence justification |
| ☐ | Scalers and encoders fitted on training data only |
| ☐ | Mining result evaluated (test set / thresholds / silhouette) |

| | |
|---|---|
| ☐ | One-page plain-language summary included |
| ☐ | Scaling and ethics paragraphs are specific to YOUR data |
| ☐ | All sources and tools cited; files named correctly |