Module 10: Capstone project

Data analysis with Python for health specialists

Yaé Ulrich Gaba

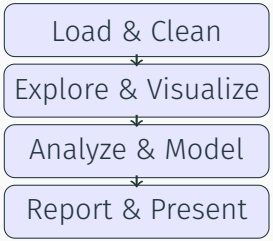
2026

AIRINA Labs

The capstone project

Apply every skill from this course to a real health question:

- Real public data (no simulated datasets)
- Reproducible Jupyter notebook
- At least 3 visualizations
- At least 1 statistical test or model
- Written report (5–8 pages)
- 5-minute oral presentation



5 suggested projects

#	Topic	Key datasets
1	Diabetes risk factors	Pima Indians (UCI), CDC BRFSS
2	COVID-19 across Africa	JHU CSSE, OWID, World Bank
3	Maternal mortality	WHO GHO (MMR, SBA), World Bank, DHS
4	Heart disease prediction	UCI Heart Disease (Cleveland)
5	Malaria burden & interventions	WHO GHO, Malaria Atlas, DHS

You may also **propose your own project** (subject to instructor approval).

Project 1: Diabetes risk factors

Objective: Identify strongest risk factors and build a predictive model.

Required analyses:

1. Clean Pima dataset (handle zeros as missing)
2. EDA: distributions, correlations, box plots by diabetes status
3. Logistic regression: odds ratios with 95% CIs
4. Compare logistic regression vs. random forest (accuracy, AUC-ROC)

Deliverables: correlation heatmap, ROC curves, OR table, clinical interpretation.

Project 2: COVID-19 impact in Africa

Objective: Analyze differential COVID-19 impact and correlations with health system capacity.

Required analyses:

1. Select 10–15 African countries across subregions
2. Epidemic curves (7-day rolling average), identify wave patterns
3. Key metrics: cases/million, deaths/million, CFR, vaccination rate
4. Merge with World Bank indicators; correlations with health expenditure
5. Choropleth map of case fatality rates

Deliverables: epi curves, scatter plots, choropleth, summary table.

Project 3: Maternal mortality determinants

Objective: Which health system and socioeconomic factors predict MMR?

Required analyses:

1. Load WHO GHO MMR data; identify 20 worst countries
2. Merge with skilled birth attendance, antenatal care, GDP, female literacy
3. Scatter plots + Pearson/Spearman correlations
4. Multiple linear regression predicting MMR; check assumptions
5. Choropleth map of MMR across Africa

Deliverables: scatter plots with regression lines, regression table, choropleth, policy discussion.

Projects 4 & 5

Project 4 — Heart disease prediction:

- Compare 4 models (logistic, decision tree, random forest, KNN)
- ROC curves on one figure, confusion matrix, feature importance
- Clinical interpretation: do top predictors match cardiology guidelines?

Project 5 — Malaria burden & interventions:

- Incidence trends (2000–2022) for 10 high-burden countries
- Correlate changes in ITN use / IRS coverage with incidence decline
- Choropleth of current malaria incidence
- Regression: does intervention coverage predict incidence change?

Report template

1. **Introduction** (0.5–1 page): health context, research question, dataset overview
2. **Data description** (1 page): source, variables, summary stats, missing data handling
3. **Methods** (1 page): cleaning steps, tests/models chosen and why, validation strategy
4. **Results** (1.5–2 pages): figures, tables, test statistics, model metrics
5. **Discussion** (1–1.5 pages): clinical meaning, comparison with literature, implications
6. **Limitations** (0.5 page): data quality, generalizability, future work

Figures rule
Every figure must have a title, labeled axes, and a caption. Figures without labels = lost marks.

Grading rubric

Component	Points
Research question	10
Data handling (loading, cleaning, documentation)	15
Exploratory analysis (stats + 3 visualizations)	15
Statistical analysis / modeling	20
Report quality (writing, figures, limitations)	20
Code quality (reproducible, commented)	10
Oral presentation (5 min, answers questions)	10
Total	**100**

Presentation guidelines

5 minutes + 2–3 min questions. Structure:

1. **Slide 1:** Title, your name, date
2. **Slide 2:** Motivation — why does this health question matter?
3. **Slide 3:** Data — source, *n*, key variables
4. **Slides 4–5:** Key results — your 2–3 best figures
5. **Slide 6:** Conclusions — 1–2 take-home messages

Tips: One message per slide. No code on slides. Speak to the audience, not the screen. Practice with a timer.

Today: get started

Complete this checklist before you leave:

1. Choose your project (1–5 or custom)
2. Download dataset, run `.head()` and `.shape`
3. Write your research question in one sentence
4. List 3 planned visualizations and 1 statistical test
5. Identify potential data problems
6. Create your project folder:

```
capstone_project/
    data/raw/
    data/cleaned/
    notebooks/
    figures/
    report.pdf
```

Timeline

Week	Milestone
Week 1	Choose project, download data, initial exploration
Week 2	Complete analysis, generate figures/tables, draft report
Week 3	Finalize report, prepare and deliver 5-minute presentation

🚀 You have all the skills. Now apply them.

Exercises (Hour 3)

1. **Pair review:** Present your research question and initial data exploration to a partner; give and receive feedback
2. **Data check:** Verify your dataset loads correctly, identify missing values, and plan your cleaning steps
3. **Analysis plan:** Write a 1-paragraph plan describing your 3 visualizations and 1 model/test
4. **Submit:** Project choice + research question to the instructor by end of session