APHRC

African Population and
Health Research Center
Transforming lives in Africa through research.

AFRICAN POPULATION AND HEALTH RESEARCH CENTER
APCC/SFA FOUNDATION

Senior Programme Officer - Data Science

Due date: 2026 Jan 28 (Wed) - 23:59 HRS

Instructions

1. Submit a **.pdf** file with all the derivations and results (**NO CODE CHUNKS INCLUDE**) and a separate **.Rmd** or **Python/Python notebook** file which contains codes used to generate the results (**the .pdf** file).
2. Try as much as possible to explain your outputs, especially graphs and any other outputs from the code.
3. AVOID using GPTs. In case you have to, kindly state how it was used. In case any other materials or online sources are used, kindly provide reference.

The Data

Identify a publicly available online dataset that satisfies all of the following criteria:

1. Public Health Focus:
   - The dataset must pertain to a topic in public health (e.g., epidemiology, health services, environmental health, disease surveillance, nutrition, etc.).
2. Longitudinal Design:
   - The data must include repeated measurements over time (e.g., cohort study, panel data, repeated surveys, follow-ups).
3. Variables:
   - At least one outcome (dependent) variable,
   - At least four predictor (independent) variables,
   - At least one predictor must be categorical.
4. Spatial Information:
   - The dataset must include spatial variables such as:
     - Geographic coordinates (latitude/longitude),
     - Administrative region codes (e.g., county, district, province),
     - Raster/shape identifiers that can be linked to spatial files.

Task

1. Describe where the dataset comes from, who collected or hosts it, and why it is suitable for this task, particularly in terms of its public health relevance, longitudinal nature, and availability of spatial information.
2. Describe the dataset, including the study population, time period covered, and the main variables. This should include the outcome variable, at least four predictor variables (with at least one categorical), and the spatial variables.
3. State a clear scientific question that reflects the outcome of interest, the key predictors, and the longitudinal and spatial nature of the data.
4. State the main objectives of your investigation and ensure they are aligned with your scientific question.
5. Perform and report all data preparation steps. If there are missing values, describe how they were handled. If there are no missing values, briefly describe general approaches for handling missing data, taking into account different data types.
6. Formulate an appropriate model that accounts for both the longitudinal and spatial structure of the data and write down its mathematical form. Estimate the model, present the results using graphs, and compare estimates across regions. Comment on whether you are confident in the estimates.
7. Describe how you would account for the longitudinal and spatial nature of the data in prediction. Suggest and implement at least two predictive models, select the best one, and explain how you chose it and all the steps taken to implement the models. Briefly discuss whether it can be compared with the model in **6**.
8. Using the estimates from Section 6, simulate data with the same structure as the original dataset. Run several replicates and use the simulated data to assess whether the selected model fits the data well.
9. Build a simple app (for example R shiny or streamlit app) to compare the models.

Submission

1. A report in .pdf format written in the style of a journal manuscript of your choice; please indicate the journal format used.
2. A link to the simple app developed in Section 9.
3. All code files with clear instructions on how to run them to reproduce your results. Ensuring reproducible results is very important.

——— GOOD LUCK ————