Center for Epidemiological and Modelling Analysis (CEMA)

Practical Session - Data Science Research Assistant Positions

Time: 24th October 2024

Congratulations on your selection to this stage of the interview process for the position of **Data Science Research Assistant** at CEMA. We received several hundred applications for this position. You and only a few others were impressive on your experience and CV and we are delighted to engage further on your suitability for the position.

We aim to select three final candidates to be offered the current available positions. This decision will be based on:

1. Demonstrated capability to wrangle, visualize and analyze data using a programming language e.g., R
2. The assessment during the verbal interview session

To provide you an opportunity to showcase your skills in a) above, we have prepared the practical exercise below. Before starting, please note the following:

- You have three hours for the exercise (*this is an open book challenge - you may consult internet sources as you perform the tasks*)
- You are required to submit all your input codes and outputs for each of part of the exercise.
    - We recommend you submit
        - **An RMD document with the codes (named using your first and last name)**
        - **A html document with the output**
    - This should be submitted to ***info@cema.africa*** by 11.00am

<u>The Challenge</u>

1) Data Wrangling Challenge

- You are provided with a subset of a dataset obtained from a randomized control trial with three study arms. The study aims to determine the impact of two interventions delivered at household level on nutritional outcomes of children.
    - Study arm 1: Households received livestock feed
    - Study arm 2: Households received enhanced nutritional counseling in addition to the livestock feed
    - Study arm 3: Control arm households
- You can access the data using this link: https://github.com/cema-uonbi/L4H_sample_data.git

(Hint: In case you are not familiar with Github, click code, local then download zip.

- Import the data
- Clean the column names of each dataset using function *clean_names* from the package *janitor*
- Filter data in baseline household data where the hh_eligible is 1 (use the data_dictionary to check what that stands for)
- Merge the datasets
    - To do this, use number (in baseline individual data) and number__0 (in baseline mother data) to merge these data then,
    - Use the household_id to merge the data above with baseline household data
- Use the data dictionary (location - https://github.com/cema-uonbi/L4H_sample_data.git) to recode the following variables:
    - Reason_for_eligibility
    - Rspntgndr
    - H_hfrml_eductn
    - Rspndtmarital
    - Rspndt_eductn
    - Maincme
- Separate the following variables:
    - lvstckown using space as a separator
    - herdynamics using space as a separator
- Create a new column called study_arm and assign the values for each row of data to either *Study arm 1*, *study arm 2* or *Study arm 3* based on column named **villages** as below:
    - **Study arm 1:** Lependera, Gobb Arbelle, Nahgan-ngusa, Sulate, Saale-Sambakah, Namarei, Manyatta Lengima, Lokoshula, TubchaDakhane, Rengumo-Gargule
    - **Study arm 2:** Galthelian-Torrder, Uyam village, Galthelan Elemo, Nebey, Rongumo_kurkum, Urawen_Kurkum, Eisimatacho, Manyatta K.A.G, Ltepes Ooodo, Lorokushu, Marti, Manyatta Juu West/East, Lbaarok1
    - **Study arm 3:** All the other villages
- Create an object named *herd_dynamics* that has the following column names **interview_date, household_id, study_arm, cwsbrth, shpbrth, goatsbrth, cmlsbrth, calves_death, bulls_death, cows_death, sheep_death, msheep_death, fsheep_death, goats_death, mgoats_death, fgoats_death, camels_death, mcamels_death, fcamels_death, cowsgft, sheepgfts, goatsgft and cmlsgft variables.**

- Create a new column named *monthyear* from the variable *interview_date* that has only month and year (e.g., 2019-12-28 to 2019-12)
- For each *study_arm*, and for each *monthyear*, calculate the number of animals born, died, gifted and given out for each species (cows, sheep, goats and camels) regardless of age or sex of animal
- Create a subset of the dataset with the following variables: study arm, monthyear, columns of cows, sheep, goats and camels births, deaths, gifts and given away, which have been calculated above. Ensure you remove duplicates in the data.

2) Data Visualization Challenge

- Using the new wrangled dataset *herd_dynamics* from 1) above, develop a single graphic to show the frequencies and changes over time in animal births, deaths, gifts in and given out for the different species
    - Hint 1 - you may need the data in long format (e.g., using function pivot_longer)
    - Hint 2 - geom_col() function in ggplot2 package and facet_grid() may be useful . Use *monthyear* as the x-axis and *fill color* to show the different animal species
- The Kenya Demographic Health Survey is conducted every five years to understand the health and social trends of the population. The last survey was conducted in 2022. You have been provided with data on teenage pregnancies at a county level which can be accessed from this link: https://raw.githubusercontent.com/cema-uonbi/L4H_sample_data/main/table6_teenpregnancybycounty.csv
    - Draw a map of Kenya showing the percentage of teenagers who have ever been pregnant by county (use *Ever_pregnant* column)
    - You may get the shapefiles from this link: https://github.com/cema-uonbi/L4H_sample_data/tree/main/shapefiles (Hint: you need to download all the files).
    - Hint: Use the st_read() function in sf package to read the shapefiles into R.

3) Data Analytics Challenge

You have been provided with a subset of data from the IDEAL study which contains data collected over time for 30 calves. each calf has a unique ID (CalfID) and multiple visits (VisitWeek). The data may be accessed using this link: https://raw.githubusercontent.com/cema-uonbi/L4H_sample_data/main/ideal3a.csv

- The variables in the dataset are:
    - CalfID: Unique ID of the calf
    - CalfSex: calf sex: 1=Male, 2=Female
    - CADOB- Calf Date of Birth
    - Education- Highest level of education attained by the household head
    - Distance_water: Distance of the household to the water point
    - Recruitweight: Weight of the calf during the recruitment time
    - Weight: Weight of the calf during the different visits

    - ManualPCV to Q.Strongyle.eggs: Qualitative and quantitative health indicators of the calves
    - Age: Age of the calves

- Conduct a logistic regression to determine the factors associated with likelihood of death (ReasonsLoss1) and interpret the results
- Determine the factors associated with rate of growth of the calf (ADWG) and interpret the results. (Hint: use a linear regression model). Include model diagnostics.