KOMAZA

Work Assignment: Data Quality Assurance Officer

Confidentiality Notice: This assignment and associated links are Confidential. Do not share.

Overview

At Komaza we deeply value clear communication, collaboration, structured thinking, and systems building. While flexibility and adaptability are a way of life for us, our top performers are able to approach challenges through the development of rational assumptions, quick-fire research, and simple, scalable systems that stand up to the trials of our operating conditions.

These tasks are designed to give us an idea about your analytical and communication skills. We know you may not have time to complete everything, but we hope this will give you some exposure to the sort of work you might be involved in at Komaza, and it'll give us a sense of your experience. Please complete this task on your own.

Thank you for your time and we look forward to receiving your work assignment.

SECTION 1: CODING

Background

As the Data Quality Assurance Officer, you will be required to quality check high volumes of data. Currently a lot of approval processes are very manual and we would like to automate these processes. Coding is the quickest way to review high volumes of data with a high degree of accuracy and efficiency. Tasks that can be carried out by 3-6 people in many days can be simplified by a simple code run in seconds.

Context:

Field Operations is one of the major departments supported by the Data Operations Team. One of the major Field Operations Activities is planting and shamba management. These operational processes are described as follows:

- **Seedlings Delivery:** Komaza delivers seedlings to the farmer. The seedlings delivered are based on the FSE(Full Shamba Equivalent or shamba size)
- **Planting:** Planting the tree in the ground.
- **Shamba Management:** This is broken down to 2 cycles
  - **Cycle 1:** FA records seedlings alive, dead and missing in the first month: April
  - **Cycle 2:** FA records seedlings alive, dead and missing in the second month: May

All of the steps above are conducted by our field assistants, front line staff working directly with farmers and collecting the data. The field facilitators are managed by Field Officers, who are in turn managed by Field Managers.

A farmer is expected to plant all seedlings however during shamba management some deedlings do not survive hence why data is collected to assess how many seedlings survive post planting. All seedlings will either survive till cycle 2 or some will not survive. We however do not expect the seedlings to increase from the originally planted numbers. This simply means: total alive, dead and missing seedlings will either be equal to seedlings planted or they will be less but never more.

Scenero:
The Field Operations Director suspects the FA's have been submitting wrong data. He instructs you to carry out a data quality check on the planting and shamba management data and provide a report on your assessment of the data quality.

Deliverable 1:
For purposes of this exercise, we have prepared planting and shamba management dataset entitled: Data Quality Assurance (Work Assignment Data) . Use this dataset to:

1. Write a script in R or Python executing a detailed review of the data quality with a focus on the following data quality issues:
   a. Missing values
   b. Duplicates
   c. Cycle 1 reported seedlings(alive+dead+missing) greater than or less than + or - 10 respectively compared to seedlings planted
   d. Cycle 2 reported seedlings greater than or less than + or - 10 respectively compared to seedlings planted
   e. Cycle 2 reported seedlings greater than or less than + or - 10 respectively compared to Cycle 1 reported
   f. Cycle 2 GPS greater than 150 m difference
2. A final dataset containing all records with any/all of the above errors

You will be expected to share both the code and the dataset for this section of the assignment.

SECTION 2: DATA QA(Quality Assurance) DASHBOARD

Data quality dashboards are quick snapshots indicating the health of a dataset. The Field Operations Director asks you to create a dashboard indicating the errors **a** to **f** in **Deliverable 1** to serve as the report on data quality. He is interested in a breakdown of the errors grouped by Field Manager.

Deliverable 2:
Using the same Data Quality Assurance (Work Assignment Data) create a data quality dashboard highlighting what you think would be most relevant to the Field Operations Director. Feel free to use a readable tool of your choice e.g; G-sheets or Excel or Google Data Studio.

Explain the rationale for choosing specific statistics for each of the dashboards.

Bonus points:

- **Retrospective Insight:** Share a one pager with highlights on your findings to the data team indicating; what did you learn that would be interesting to discuss internally in terms of quality of data, method used, potential biases, etc