CEGIS Recruitment: Assignment

Overview

Thank you for applying to the Program Manager position with the Outcome Measurement team. The goal of this assignment is to give candidates an opportunity to showcase their approach to a few typical situations and display the data analysis skills that will be necessary to be successful in this role.

You are required to submit your response in an email with relevant attachments **within a week of receiving** this assignment.

Please note that this exercise is confidential. Feel free to ask us questions by writing to hiring@cegis.org if something is not clear.

Thank you. We appreciate your time and look forward to reading your responses!

Introduction

To complete your application to this CEGIS position, you are required to submit an assignment described below.

There are two parts to this assignment: a) data exercise and b) presentation to a policymaker. **Responses to both should be submitted in one single PDF document.**

- For (a), you must attach a well-documented code file, any intermediate datasets generated, and any external data sources directly used in the analysis. For components that need narrative answers, tables, charts, etc., organize your responses in a Word file with clearly labelled questions and responses matching the numbering in this document, and use font Arial, font size 11, line spacing 1.15.
    - You may choose between STATA, R, and Python for this part, and preferably complete all tasks on the same platform. Solutions using only Microsoft Excel / Google Sheets will not be accepted.
    - There are often many ways to perform a given task so try to use relatively time-efficient and neat ways - ideally exploiting the strengths of the platform you have chosen.
    - Irrespective of the technical sophistication of your method, ensure you explain your methodology, results, and interpretation in simple language.
    - If you get stuck in one question and you can move on to the next one, go ahead. We encourage you to solve as many questions as you can and for incomplete and partial answers please try to provide a narrative answer about your planned approach, constraints faced, data/information that could enable you to implement that.

    - **You will be evaluated on the quality of your code and approach and not purely on the accuracy of results.**
- For (b), you are encouraged to choose a presentation style appropriate for the audience specified.
- Throughout, all sources used must be cited clearly, following any popular citation style. Plagiarism in text and data analysis - including from your current or past professional work - will lead to immediate disqualification. Follow the instructions for each part carefully.

A. Data Exercise

You have been provided two files with nationally representative data on India's labour market as of January 2020. ('Labour1.csv' and 'Labour2.xlsx' [Link]) Both these datasets contain individual level information. In particular:

- Labour1 contains basic background information about each individual such as state, sector, household size, religion, and social group.
- Labour2 contains information about each individual's education, training, and employment status.

The variable 'Person ID' uniquely identifies individuals in both datasets, and can be used to merge the datasets, if required. You will find the descriptions for the variables in the datasets in the appropriate codebooks, also shared.

Using this data, please answer the questions below. Present your responses effectively using text, tables, and charts. While no word limit has been specified, you will be evaluated on your ability to highlight key insights concisely while responding to this part of the assignment. Feel free to support your responses with any relevant literature, data sources, and analysis of any other datasets. Remember to attach all working files relevant to your data analysis (well-documented code file, any intermediate datasets generated, any external data sources directly used in the analysis, etc.) to your response.

1. Load the data and merge the two parts to create a single dataset. To ensure all observations and variables are valid, identify and rectify any inconsistencies in the data, describing any assumptions made.
2. Prepare the dataset for analysis.
    a. Generate new unique household and individual IDs based on geographical information and briefly explain the structure.
    b. Appropriately label all variables you expect to use in your analysis below with a meaningful label and encode any categorical variables (i.e. provide a label for each level) you expect to use in your analysis below referring to codebooks (e.g. relationship_head, sex, state, district, etc.) and at least one

categorical variable using relevant external documentation (e.g. type of occupation).

   c. Generate new variables classifying a) Individual's nature of employment as per principal activity (Self employed, Salaried, Casual labour, Unemployed, or Not part of the labour force); b) Individual's usual location of work (rural or urban)

3. Briefly describe the profile of the individuals and households provided here - including but not limited to demographic, geographical, financial, and employment aspects - highlighting any interesting trends, outliers, logical inconsistencies.
4. Comment on the different ways in which unemployment can be measured from this dataset and on whether they yield similar estimates. (Hint: Employment may be seasonal, so think about the time period that different variables cover).
5. What do you think is the most appropriate way to measure unemployment using this dataset? Using the same, calculate the national and state-wise unemployment rates.
6. Validate your estimate against other relevant unemployment estimates. Describe a few factors from measurement (i.e. survey instrument and data collection process) and data perspective that could explain any differences in the estimates.
7. Choose any two comparable states, justify your choice, and comment on how their performance compares with Telangana's performance on these variables:
   a. Unemployment rate
   b. Female labour force participation
   c. Wages
   d. Benefits and job security
8. Identify the determinants of wages within the available dataset. You may refer to relevant theory and literature. Explain your methodology, suitably present your results, and explain your inference from those results. (Hint: Be cognizant of how certain variables may cause wages to change and certain variables are related to changes in wages)
9. Is there any difference in the wages earned by men and women? Does level of education have any influence on this difference? Specify an appropriate regression equation(s) to test this, implement to generate results, and explain your inference from those results and limitations, if any.
10. Clearly state a hypothesis to test that could be relevant to informing policymakers about the employment situation in Telangana. Can you test this within the available dataset? Describe any additional data points that you might need beyond the available dataset to accomplish this objective.

B. Presentation to a policymaker

Imagine you are designing a large-scale household survey in Telangana to generate estimates of certain outcome indicators, disaggregated at mandal/block level. The primary aim of this survey is to provide various levels of the State machinery visibility on outcomes and use that for goal setting and performance monitoring.

You have been asked to make a **presentation to the Commissioner of a certain Department in the Government of Telangana. Prepare a 10-slide deck for ONE among two themes:** a) Child Malnutrition; b) Unemployment. Ensure that you touch upon the below mentioned aspects such that it is salient for the specified audience, without diluting the technical rigour:

1. Value of this data to the Government (i.e. technically rigorous, high-quality, geographically disaggregated data, etc.)
2. Key outcome(s) to measure
3. Methodology for measuring the key outcome(s) at household/individual level
4. Survey design, including aspects like geographical coverage, sampling design, etc.
5. Usage of the collected data at different hierarchical levels. Hints for this:
    a. Think creatively about leveraging technology, tools, and communication channels. For example, consider summary reports, dashboards, in-person meetings, social media, etc.
    b. Keep in mind various consumers of this data by seniority, education, geography. For example, how will your strategy differ for a Secretary vis-à-vis an Anganwadi worker.

NOTE: You need NOT cover aspects of the survey itself (e.g. instrument, piloting for local context, etc.) or survey operations (e.g. team composition, financials, timelines, etc.).
