BUSARA
CENTER FOR
BEHAVIORAL
ECONOMICS

Data Analyst Assessment Task

November 2019

Instructions

You have **48 hours** to attempt the two questions in this task. All necessary files for completing this test are included in the attached folder. If at any point you are stuck, explain (preferably commented in your output file) what you would have done had you had more time or knew the correct commands for doing it. Try to get through as much of the test as you can in the time allotted; even if your answer depends on previous steps that you were unable to do, you will still get points for demonstrating that you would have gotten the correct answer if you had successfully completed all previous steps.

You should submit a compressed folder consisting of

A. Your analysis script, with explanatory comments and instructions for replication.
B. Answers to written questions (e.g. Microsoft Word, RMarkdown knitted file, or Jupyter Notebook).
C. Output files (graphs and clean data)

Name the folder *{Surname_Firstname_Busara_Q4_2019}* and attach it to your email reply.

The Busara data team uses R for most analysis of small, static datasets and prefer for you to use RMarkdown for this task. If so, please include a knitted HTML of PDF file. You may also use Stata or Python for analysis; doing so will not be held against you. However, please note we will not accept SAS, SPSS, Excel/VBA or any other type of statistical software except R, Stata, and Python.

Please list any resources that you consulted in the analysis report, such as publications, tutorials or package documentation.

Question 1

The “HealthInsurance.xlsx” is a cross-sectional dataset originating from a medical expenditures survey conducted in 2010 in some African countries. The main objective of the study was to estimate health insurance uptake and also find out which factors influence this uptake.

1. Import the dataset and clean the data of any anomalies.
2. What is the distribution of the sample in terms of (i) gender and (ii) age?
3. Ideally, we expect respondents with poor health status to have health insurance. Is this the case according to this data? Using an appropriate statistical test, investigate the relationship between the two variables. What is the null hypothesis of this test?
4. Which ethnic group is most likely to be insured? Can we reject the hypothesis that the insured proportion for this group is equal to any other group?
5. How many self-employed people have insurance?
6. An insurance company would like to use this data to understand what factors drive insurance uptake. Select an appropriate statistical model for this task and give reasons for the choice.
7. Fit the model on the data and construct 95-percent confidence intervals for the coefficients of the explanatory variables. What conclusions can you draw from these results?
8. What advice would you give the insurance company?
9. Pretend you have to present these results to a health ministry official who is not familiar with statistical inference. Demonstrate how you would present the regression output to them, adding any notes you think are necessary to explain the model specifications, output and implications.
10. Use a likelihood-ratio test to test the null hypothesis that none of the explanatory variables influences insurance uptake.

Question 2

“Livestock keepers dataset.xlsx” is a dataset on pastoralists. The client would like to understand the groups they fall into so as to design more targeted interventions for them.

1. Describe what data quality issues you’ve identified on the provided dataset and how you’ve addressed them.
2. Segment this population of pastoralists into their natural groups. What is the basis of your segmentation and what does it show us?
3. Describe your segmentation process and why you’ve chosen this particular route.
4. How many segments have you identified and what makes you think that this is an appropriate number of segments? Provide both a business and statistical rationale for your answer and use graphs to the best of your ability.
5. Visualise these segments in a graph.
6. Describe the nature of each segment you’ve identified in terms of composition/profile in a paragraph.
7. What correlations do we have between the data we have and the groups you’ve identified? Identify the significant correlations and analyse what the correlations imply.
8. What are they key take-aways from this exercise?
9. What other analyses do you think can be performed to obtain further insights?