DATA ANALYSIS TEST

Instructions

You should spend **at most two hours** on this test[1]. All necessary files for completing this test are included in this email. Your final script files should be stored on a GitHub public repository[2]. The final dataset and any deliverables (graph, regression model) must be sent in a zip file to **a20141676@pucp.edu.pe**.

Break-up of marks

Section	Topic	Marks
1	**Data Cleaning**	**20**
2	**Summary statistics**	**20**
3	**Regressions**	**20**
4	**Version control/Coding Management (Optional)**	**10**
***40 points** will be allotted based on the overall proficiency, efficiency, clarity and replicability of your work.* ***A bonus of 10 points** can be obtained by displaying a good understanding on version control using GitHub*		
Total Marks - 100		

You may consult any resources you like, except for other people. Please list any resources that you consult. If at any point you are stuck, explain (preferably commented in your script) what you would have done had you had more time or knew the correct commands for doing it. **Except for the regression model and the graph, note down all your answers to the relevant questions in the do file itself.** 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.

The Dataset

In 2010, researchers conducted a randomized control trial (RCT) to increase voter turnout, with an emphasis on female turnout, during an election in India. The RCT was conducted in 27 towns, with approximately half of the polling booths in each town randomly selected for treatment. The treatment was rolled out in two phases. The outcomes of interest were total turnout (the number of votes cast at each polling booth) and female turnout (the number of votes cast by women at each polling booth). Data was also collected on the number of registered voters at each polling booth, disaggregated by gender, but for some polling booths

1 You can spend at most 5 hours, but there will be a penalty for sending it after the first 2 hours and data tasks sent after the time limit are not going to be evaluated.

2 This is not mandatory, but strongly suggested.

this data could not be obtained, and so data entry operators entered "-999" whenever they were missing data.

Problems

- **Section 1 - DATA CLEANING (20 Marks)**

  Make the dataset ready for use in analysis. This involves the following tasks:

  1. Import the data
  2. Merge with supplementary file with town names and districts
     - Make sure that all towns are named and drop any irrelevant towns.
  3. The district variable is currently 'string'. Create a district variable such that it is numerical
  4. Create a unique ID for each observation. Create the ID such that the first three digits are the town id
  5. Are there any variables in your dataset with missing data? Identify the variables that have missing values and deal with the missing values in your dataset so that it does not affect your analysis.
  6. Create a dummy variable for each value of Town ID.
  7. Label all variables as either "ID variable", "Electoral data" or "Intervention".
  8. Label values for the treatment variable appropriately.

- **Section 2 – DESCRIPTIVE STATISTICS (20 marks)**

  9. What is the average total turnout rate? Also note down the highest and lowest turnout rates recorded. How many polling booths recorded the highest turnout rate?
  10. By treatment, tabulate the number of booths in phases 1 and 2 of the study
  11. Tabulate the average turnout rate for females for each district which has a total turnout rate of 75% or above.
  12. Is the average turnout rate for females notably higher in treatment polling booths than control? Can you say the difference is significant? How would you test for it?
  13. Create one simple, clearly-labeled bar graph that shows the difference in turnout between treatment and control polling booths by gender as well as total turnout. Please output your results in the clearest form possible.

- **Section 3 – REGRESSION (10 marks)**

Your PI is going to present the results from the experiment at an academic conference, and needs you to create one table to show the effects of **Treatment on total turnout.**

Take into account (control for) the following variables – **all town_id dummies and registered turn_out** to improve the model (reduce noise)

14. Please output your results in Excel/Word in the clearest form possible. It is not necessary to show the coefficients on the control variables. However, do show the coefficient on registered voters.
15. What is the mean turnout for the control group?
16. Note down the dependent variable.
17. What is the change in the dependent variable after the intervention?
18. Is the difference in turnout between the treatment and control booths statistically significant? Explain in no more than 50 words how you would assess that.

❖ **Section 4 – INSTRUMENTAL VARIABLES (10 marks)**

Now assume that take-up of the intervention was not complete, meaning that in some of the polling booths that were assigned to get a voter turnout campaign, the turnout campaign did not actually happen.

The researchers would like to assess the effect of actually receiving the treatment instead of an intent-to-treat effect. The variable showing where take-up occurred is called take_up, equal to 0 if the campaign did not happen and 1 if it did.

19. Is there a variable in this dataset that is plausibly an instrumental variable for the presence of the voter turnout campaign?
20. Please state the relevance condition of instrumental variables and discuss/show why it would hold or doesn't hold in this case.
21. Please state the exogeneity condition for instrumental variables and provide evidence on whether it holds. Hint: the best variable in the data set to use for testing the exogeneity condition is registered_total, so you can just use that one.
22. Please run the instrumental variables regression showing the effect of take_up on turnout using an instrumental variables approach and discuss the magnitude of this effect relative to the effect you found previously in question 18.