Programming Task

This document outlines a set of tasks to evaluate your skillset as a potential research assistant for data science projects. Intermediate knowledge of python, R or Stata is presumed. **You may use the internet to look up documentation and ggplot examples.**

You may use a combination of python, R, Stata to do these tasks; most tasks are possible in any combination of these languages. It should be fairly easy to switch between languages once you have the basic tasks nailed down in one programming language. **However, please note that we prefer people who can demonstrate an advanced ability in python and R.**

Attempt each section to the best of your ability. Install the relevant libraries using conda, pip or ssc install.

**Data Management, Visualization, and Spatial Operations**

A big share of empirical political science work involves cleaning and merging data, conducting initial descriptive analysis, and

1. Task 1: Candidacy in elections
   1. Load in the delhi 2015 and 2020 election data (delhi-2015-2020-results)
   2. For each election year, calculate the number of candidates for each assembly constituency (AC) in Delhi (there are 70 ACs so you should create a data frame that has the numbers of candidates for each of the 70 ACs in 2015 and 2020)
   3. Identify the top 5 ACs with the most candidates in 2015 and 2020. Produce a table with their AC number, name, and the number of candidates.
   4. Identify how many of the candidates in the top 5 constituencies for 2020 ran without a party label (these are candidates with the Party “IND” meaning independent) – if you look through the election-codebook you can get more information on what each variable means.
2. Task 2: India Names
   1. Load in the India names sample dataset. This includes a random sample of voters names across the 70 ACs in Delhi. We are going to pretend for this exercise that these are all the voters in Delhi.
   2. The India names sample dataset is incomplete. There are 70 ACs in Delhi but there are only 67 ACs in this dataset. Figure out which 3 ACs are missing
   3. We want to know if places that have more Muslim voters are more likely to have more independent candidates. Produce a scatter plot for 2015 and for 2020 where each dot is an AC and the x-axis is the percent of the population in that AC that is Muslim and the y-axis is the number of independent candidates in that AC. This task will involve merging data from the election dataset with the india names dataset.
   4. We want to look at how names have changed over time. Create a dataframe that identifies the average “voter\_value” (this is the religiousity score for the voter in the dataset) for each year for both Muslims and Hindus (you will need to subset out any voters from other religious groups).
   5. Use that dataframe to create a figure where you plot a geom\_smooth line (this is a function in the R package ggplot) with birth\_year is the x axis and the average voter\_value is the y-axis and the color of the line is the religious group. You do not have to use ggplot but if you do geom\_smooth is the function you will want to use.
3. Task 3: Spatial
   1. For the state of Delhi, merge the number of candidates for each constituency in the most recent election (2020) with the constituency shapefile and make a choropleth map.
   2. For one constituency of your choice, identify the centroid of the constituency and draw a straight line passing by the centroid and perpendicular to the geographic north. This line will cut the constituency in a northern and a southern half.
   3. Calculate the area of the constituency in the southern half (bottom).

**Submit your code and responses in PDF format.**