Nicolson C. Panos

Professor Nicole Dalzell

STA 381

9 December 2022

Write Up on Political Views Data Set

**Section 1: Goal**

The goal for this portion of the project was to find/create a data set that would work for finding the political views of each of the counties in North Carolina. The reason we had this goal in mind was that we and our clients viewed that the political views of the people of the counties could play a role in both of our Y1 and Y2 response variables, thus it was important to us to find a measure and data set that could be built into our large data set, so as to make modeling much easier. To approach this goal, we tried to figure out numerous ways for measuring such views as there are many different ways to define such a goal. After corresponding with our clients, we figured the best way to proceed for finding a data set to estimate the political views of the people in the counties in North Carolina was by election data, specifically Presidential election data. Since a majority of the counties in North Carolina tend to vote for the Republican party candidate in these elections, we agreed that the important measure would be to find the percentage of votes that went for the Republican candidate in each of the 2008, 2012, 2016, and 2020 elections. Then, once we were able to get all this data information, the data would be applied to the closest year before the next election so our data that corresponded with the years 2010 and 2011 would have the Presidential election data from 2008 attached to it, and this guide would remain true throughout. The county data would still be matched with the corresponding percentage and would then be added fully through to the data set. With the plan to tackle our goal firmly established, the next step in the process was finding where best to get this election data information.

**Section 2: Data**

After an exhaustive search through many different resources, the best data place I found to find the percentage of voters who voted Republican came from the North Carolina State Board of Elections Database. Within the official website of this important committee, they had a helpful section of their website dedicated to the election results and data for people to work with. The data is all stored on a “dashboard” where people looking for specific elections of interest can match the data, going back to 2002, of the election and can select the specific county, office, and contest of interest, which proved very helpful for the collection of the data. However, unfortunately, this data was not downloadable in the sense of easily getting the number of votes into an excel spreadsheet, which did mean inputting all the 400 rows of data, each row representing an election year and NC county, with their correct percentage republican. This meant the data cleaning, processing, and creation would take a fair amount of time, which made working with this explanatory variable the longest process compared to the other explanatory variables we were giving special attention. Certainly, I had to make sure to not get too beat down with all this data, but I was glad to have found the proper data instead of having to deal with other resources I found, which frequently gave inconclusive results such as missing counties or missing election years.

**Section 3: Data Cleaning, Processing, and Creation**

The creation of the data set, which would come to be titled North Carolina Election Data.xlsx, involved spending a fair amount of hours having the excel spreadsheet on one side of my screen, while I looked over at the NC State Board of Election Dashboard on the other side of my screen. For each of the counties in the elections of 2008, 2012, 2016, and 2020, respectively, I inputted the Republican candidate's name (just to make sure that data is in there in case of interest in a future study having the candidate be a variable of explanation), the number of votes that candidate received, the total number of votes, and then did a simple formula in excel of Republican candidate votes divided by the total number of votes for each of the rows in the data set. I continued this process until the completion of all 400 rows of data. I then created four other files: PoliticalViewsNCData2008.xlsx, PoliticalViewsNCData2012.xlsx, PoliticalViewsNCData2016.xlsx, and PoliticalViewsNCData2020.xlsx, so that myself, my group, my clients, and future people who would want to look into this explanatory variable could have access to each of the election years in question, seeing if they wanted to mess with possibly have more weighted interest in the years 2016 and 2020 versus 2008 and 2012 to deal with the interest in changing views from Senators John McCain and Mitt Romney to Former President Donald Trump.

The next part of this process was to combine the previous full explanatory values data set, called Combine1, from R and add in my data on the Percentage Republican. Instead of trying to do this process on R, which I would have found difficult as, for example, trying to match the data from 2010 on Forsyth County with the percentage of Republican votes in Forsyth County from the 2008 election year would be very difficult through R, or I would fear mistakes R could make in that process. To combat this problem, I, instead, downloaded the Combine1 data set into an excel spreadsheet and then manually inputted the data values, mainly copying and pasting over from one spreadsheet to the next. After this was complete, I then uploaded the new Combine data set after adding my explanatory variable of interest, and then made that data set the new Combine1 dataset in R, but still titled the same as the previous one. With my work finally done on the data set, we were finally able to have all of our explanatory variables ready to go for the next stage of the project, modeling.