Capstone Project Battle of the Neighborhoods

The Problem

There is a lot of data available to study the Congressional Districts in the United States. The data can come from many sources both free and subscription based. The data also comes in multiple formats. It can be time consuming, error prone, and expensive to gather this data. This project is to focus on one district – the 5th Congressional District of New Jersey (NJ5).

The intent of this project is to see if the Foursquare location data API can be used to help easily cluster the different towns in NJ5 and compare these results to clustering done on both United Status Census data and New Jersey State election results data. We should be able to confirm if the Foursquare location data can help in not only describing NJ5 but also to see if it is possible to use the Foursquare data to describe all the 435 Congressional Districts in the United States. Gathering the Foursquare data in a friendly API is much easier than all the different file formats from the various US states and US Census data providers.

Ideally this analysis could be used to help people running political campaigns to understand the neighborhoods where voters live in each Congressional District and possibly predict voter tendency to vote for the Democratic or Republican party.

The Background

The average US Congressional District has over 700,000 residents. Running a campaign in a district involves analyzing large amounts of data to understand who the people are and what concerns them the most. For example voters in a rural farming community could be expected to be more concerned with farming regulations than voters in a highly urbanized area with poor mass transit options. Many other factors influence voting such as age, religion, and countless other things. Therefore understanding the neighborhoods is a major component of any US Congressional campaign.

The Data and Solving the Problem

As we are focused on the NJ5 we gathered data primarily from three resources.

1. Wikipedia has a good article describing NJ5 and lists out all the communities in NJ5. This can be used as a driver to ensure we have an independent source of information on the communities. The link for this site is:

   https://en.wikipedia.org/wiki/New_Jersey%27s_5th_congressional_district

2. The US Census has a good website. There is a quick facts section where CSV file format data can be downloaded. All community names listed in the Wikipedia site were searched for demographic data. Unfortunately, the US Census has a policy of only proving data on communities with a population of 5,000 or more. Therefore, some communities, particularly in

rural areas did not have demographics collected specifically for them. The data was sourced here and pulled into CSV by community name, county, and state.

https://www.census.gov/quickfacts/fact/table/washingtontownshipwarrencountynewjersey/PST045218

3. The State of New Jersey Division of Elections has NJ5 election results published in PDF format on its website. This was pulled from the site and converted from PDF to CSV. A link to the site is below.

   https://njelections.org/election-information-2012.shtml#general

Data Clean Up

The government provided data from the Census and Division of Elections needed to be cleaned up. The names of the towns were not consistent and the towns with under 5,000 in population only had total population and voting results data available. All the data was merged into a single CSV so it could be imported into the Python Pandas module in a Jupiter notebook.

All data was checked and rechecked using the above three sources to ensure there were no inaccuracies with the source information.

Using the Data to Solve the Problem

In order to solve the problem, we plan to use the government provided data to understand NJ5 a bit better including clustering the communities using K Means. We will compare the clustering to see if the clusters can provide insight to the possibility of a community voting Democrat or Republican.

After we have built out the clusters from the government data, we plan to run the same process for the data from Foursquare. We will cluster the communities based on the Foursquare data and see if the clustering can lead to insights if the community will vote Democrat or Republican in a given elections. If this works as well or better than the government provided data, we will have found a quick and easy work around to collection and cleaning of large amounts of data to get a sense of voter behavior.