Capstone Project Battle of the Neighborhoods: A Study of New Jersey Congressional District Five

Karel Hoppe

September 1, 2019

Section 1

Introduction

There are twelve Congressional districts in New Jersey. Each one is unique. It would be great to figure out some ways to understand the different districts using data science. To start with focus can be brought to the fifth district as it is about 50% Republican and 50% Democrat.

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.

Section 2

The Data and Solving the Problem

As we are focused on the NJ5 we gathered data primarily from four 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

4. The Foursquare API. The location data on Foursquare is free. It is easy to get an API account to pull the data with and convert the JSON result into a data frame. Also the Foursquare data contains details about a community through telling us the numbers and types of restaurants, parks, transit centers, and other neighborhood details.

   https://foursquare.com/

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 set of CSV files 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.

The Fourquare data was very robust and did not clean up. It was imported as JSON and needed to be fed into Pandas.

Section 3

Methodology Using the Data to Solve the Problem

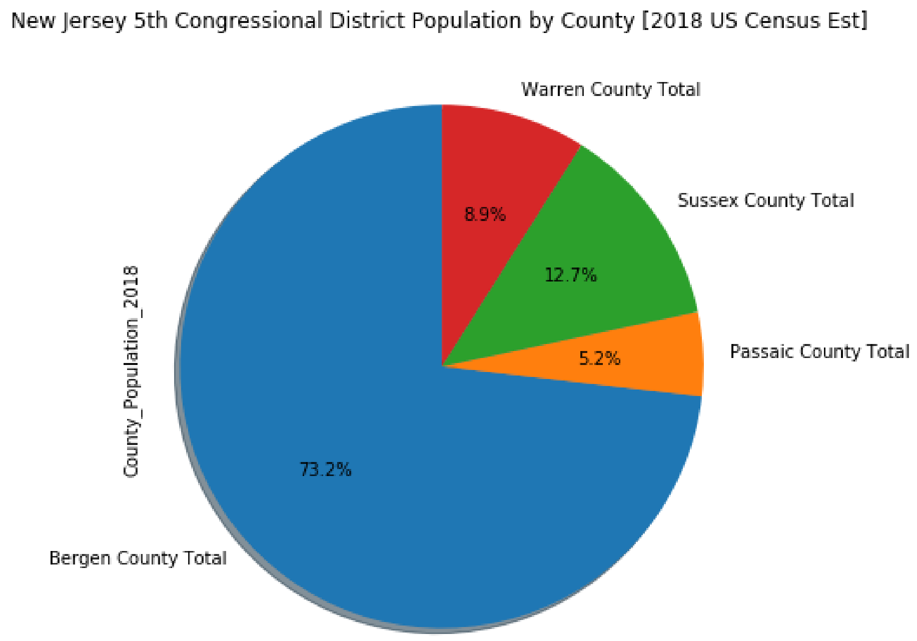
In order to solve the problem, we used the government provided data to understand NJ5 a bit better including clustering the communities using K Means. We compared the clustering to see if the clusters provide insight to the possibility of a community voting Democrat or Republican.

Upon building out the clusters from the government data, we ran the same process for the data from Foursquare. Again, we used clustering for the communities this time based on the Foursquare data and looked see if the clustering lead to insights if the community will vote Democrat or Republican.

Data Insights

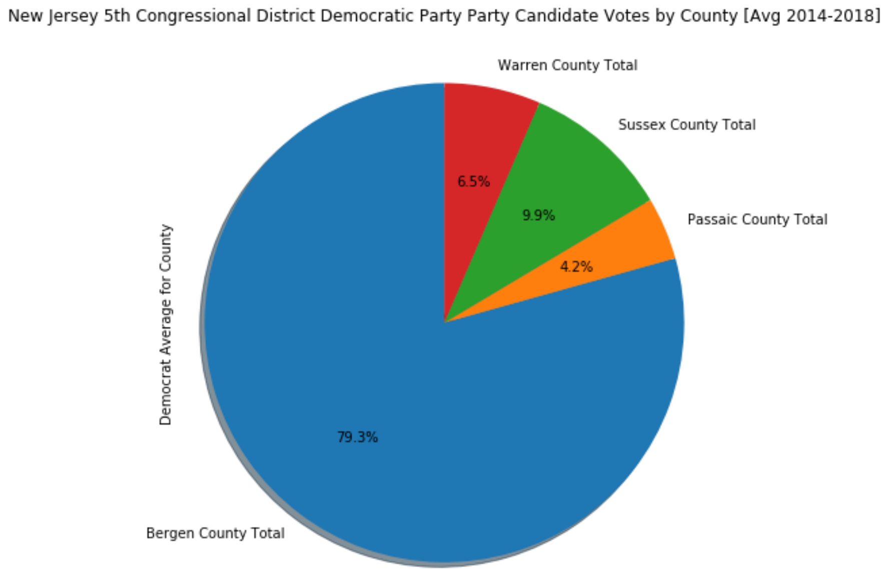
After importing all the data, we started seeing some interesting things. For example, we could see that Bergen County is the biggest county in NJ5 in terms of population.

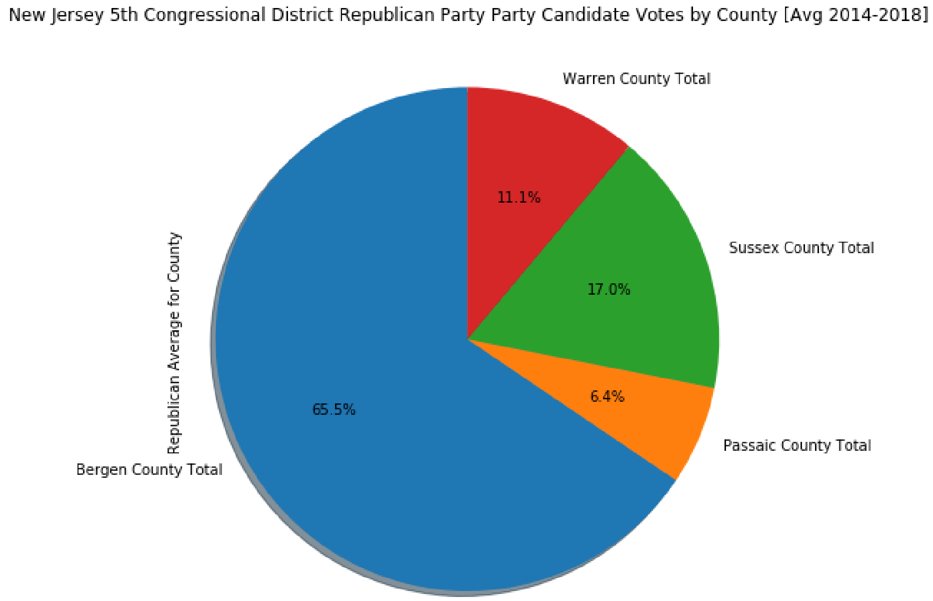
Figure 1



What makes this even more interesting is we can see the Democratic party candidates for US Congress have done better in Bergen County than the other counties over the last three elections.

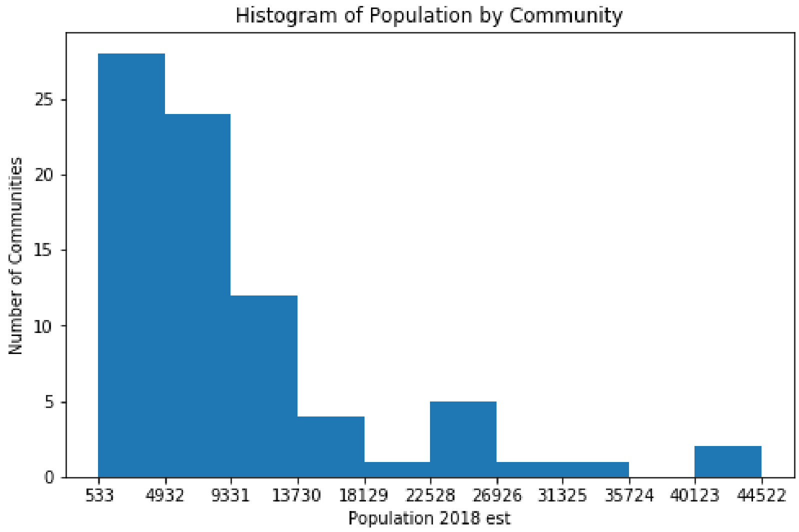
Figure 2 and 3


New Jersey 5th Congressional District Democratic Party Party Candidate Votes by County [Avg 2014-2018]


New Jersey 5th Congressional District Republican Party Party Candidate Votes by County [Avg 2014-2018]

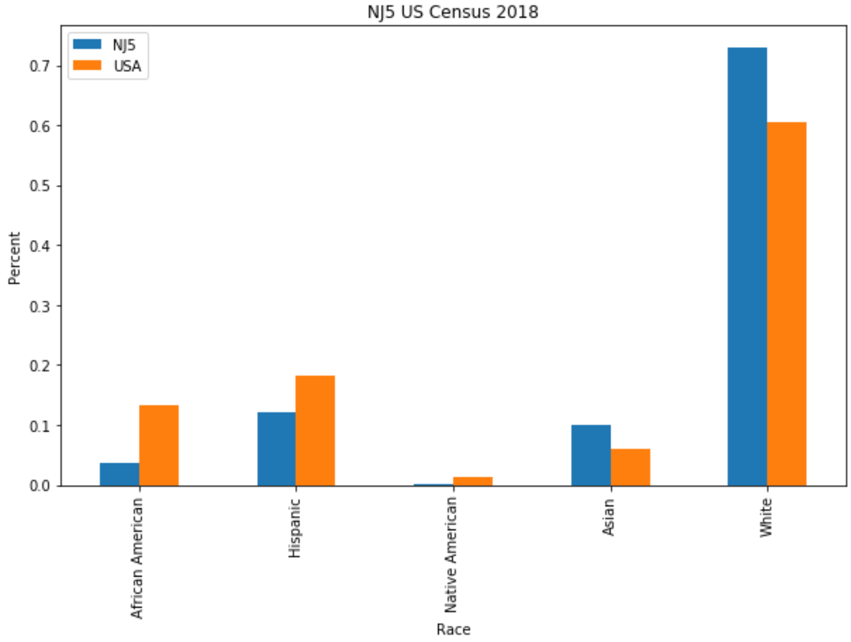
One of the things we noticed is there is not a big city in particular which makes Bergen County the biggest by population. For example, we can see the distribution of communities by size which shows no communities greater than 45,000 people and most communities are under 10,000.

Figure 4



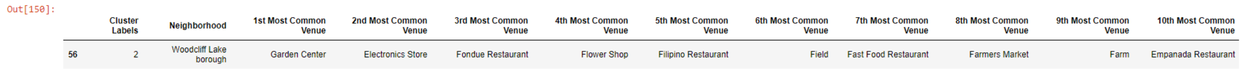
Further we were able to review the data in the US Census for insights about NJ5. For example we can see as compared to the rest of the United States NJ5 has a higher proportion of Asian and White ethnic groups.

Figure 5



The data from FourSquare also held some significant insights. For example, by examining the top venues in a community we were able to get a feel for the culture.

Figure 6



Out[150]:

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
56	2	Woodcliff Lake borough	Garden Center	Electronics Store	Fondue Restaurant	Flower Shop	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Farm	Empanada Restaurant

Section 4

Results with K Means

Using K Means we were able to cluster the data from the US Census and Foursquare. The clusters found were interesting. It was clear K Means is able to sort through the data and generally cluster communities together based on any numeric statistics we feed in. The concern found was just that. Using 23 data points from the US Census was perhaps too many and we did not see any clear pattern showing why the communities might tend to vote Republican or Democrat.   For the Foursquare data again, we didn't see any pattern which mimics the voting of Republican or Democrat for NJ5.

Discussion and Future Directions

The model using K Means appears to need some work if we are to use it to help determine a vote for Democrat or Republican. For example, it might be we should run K Mean for each US Census data point (23) and see if some work better for clustering independently. Then maybe create a model using the best performing data points.

Section 5

Conclusion

In the end we saw a few things. First, we can use Data Science effectively in the field of politics. We can get a good overview of any area and create interesting, relevant and graphical information. Second, we can use machine learning tools such as K means to help us build models. This second point should not be taken lightly as building an accurate model involves much work and can be time consuming in order to figure out which works the best.

In the end this exercise was very helpful in order to learn a lot more about NJ5 and the people who live there. Just showing the data and clustering by different communities can help anyone interested in politics understand the area better. Further this work should lead to further ideas around model building.