READ ME

Summary

This project uses the following scripts to generate multiple line plots in an effort to determine a pattern between population growth in certain congressional districts and the percent of votes for Republican candidates in U.S. House elections in those specific districts. This project will use racial data from the U.S. Census Bureau and election data from the MIT Election Data and Science Lab to analyze the changes over time in the percentages of white-only populations and the percentages of Republican-supporting votes in nine of Texas' congressional districts.

Input Data

1. Racial composition data will be obtained from the United States Census Bureau via its API. This project uses American Community Survey 5-Year data for the years 2012 through 2019 with the following calls:
    a. https://api.census.gov/data/2012/acs/acs5
    b. https://api.census.gov/data/2013/acs/acs5
    c. https://api.census.gov/data/2014/acs/acs5
    d. https://api.census.gov/data/2015/acs/acs5
    e. https://api.census.gov/data/2016/acs/acs5
    f. https://api.census.gov/data/2017/acs/acs5
    g. https://api.census.gov/data/2018/acs/acs5
    h. https://api.census.gov/data/2019/acs/acs5
2. Election data will be downloaded from the MIT Election Data and Science Lab website: https://electionlab.mit.edu/data. The resulting CSV file, 1976-2018-house3.csv, contains data for every election to the U.S. House of Representatives from 1976 to 2018.

Deliverables

There are a total of 3 scripts: cd_race.py, racedatajoin.py, and electiondata.py. The script cd_race.py generates a total of 8 CSV files, with each one containing the racial make-up of Texas' 36 congressional districts for a single year from 2012 to 2019. The script racedatajoin.py generates white.csv and CDpercentwhite.png. Finally, the electiondata.py script generates 2 CSV files (CDrepdemcontested.csv and join.csv) and 3 PNG files (CDpercentrepublican.png, CDpercentrepcontested.png, and whitevsrep.png). All in all, these 3 scripts generate 15 deliverables total, 11 of which are CSV files and 4 of which are PNG files.

Instructions

A) Sign up for a Census API Key
   1. To request an API key, go to https://api.census.gov/data/key_signup.html.

B) Script cd_race.py

   This script involves the querying of the U.S. Census Bureau with its API to retrieve racial data. This script creates eight CSV files: CDpercentwhite2012.csv, CDpercentwhite2013.csv, CDpercentwhite2014.csv, CDpercentwhite2015.csv, CDpercentwhite2016.csv, CDpercentwhite2017.csv, CDpercentwhite2018.csv, and CDpercentwhite2019.csv. Each file contains the percentages of the population in each of Texas' 36 congressional districts that identify as only white.

   1. Import requests and pandas as pd.
   2. Set the variable api to the American Community Survey 5-Year estimate endpoint for the year 2012, as shown: api = 'https://api.census.gov/data/2012/acs/acs5'
   3. Set the variable for_clause to 'congressional district:*' because that is the geographic unit we want data for.
   4. Set the variable in_clause to 'state:48' because Texas' FIPS code is 48.
   5. Set the variable key_value to the Census API you are given in quotes.
   6. Set the variable payload to a dictionary to the following keys and values because they set the parameters for the data request:
      a. 'get':"NAME,C02003_003E,C02003_001E"
         i. C02003_003E is the variable corresponding to the estimate of the total population of one race (white)
         ii. C02003_001E is the variable corresponding to the estimate of the total population
      b. 'for':for_clause
         i. This is the unit of observation, which we set to congressional districts above.
      c. 'in':in_clause
         i. This is the enclosed geographic area, which we set to state 48, or Texas, above.
      d. 'key':key_value
   7. Set the variable response to the call requests.get() with the arguments api and payload to collect the response of the census query.
   8. Set up if and else statements to see if the census request succeeded.
   9. Set the variable row_list to the call of .json() on the response variable to transform JSON returned by Census server into a list of rows. To organize the coming dataframe, set the variable colnames to the first row of row_list and set the variable datarows to the remaining rows of row_list.

10. Convert the data into a dataframe. Set the variable percent_white to the pd.DataFrame() call with the arguments columns=colnames and data=datarows.
11. Create new column "num_white" in percent_white and set it equal to the "C02003_003E" column of percent_white, and then cast the pandas objects to a float data type with .astype() method.
12. Create new column "total" in percent_white and set it equal to the "C02003_001E" column of percent_white, and then cast the pandas objects to a float data type with .astype() method.
13. Create new column "percent_white" in percent_white and set it equal to dividing the "num_white" column of percent_white by the "total" column of percent_white, followed by multiplying the quotient by 100.
14. Write your percent_white pandas DataFrame to a CSV file using the .to_csv() method on percent_white to a file called 'CDpercentwhite2012.csv'.
15. Repeat steps 2-14 for the remaining years (2013-2019), remembering to replace the year in both the API survey year and the CSV file name.

C) Script racedatajoin.py

This script generates white.csv, which pools together the white-only population percentages for each of the specified congressional districts for the years 2012-2019. CDpercentwhite.png is a line plot of that data.

1. Import pandas as pd.
2. Set the variable dtype_CD to a dictionary using the column name 'congressional district' as a key with a value for the string data type (str). This will tell Pandas to load data in string form.
3. Read the Texas congressional districts' yearly white percentage data. Create dataframes with variable names identifying corresponding years and call pd.read_csv() on the different yearly CSV files with the argument dtype=dtype_DC.
4. Set the index of each dataframe to the column "congressional district" with the inplace=True parameter to make change and overwrite existing dataframe.
5. Create new dataframe with percentages of population that were white in the Texas congressional districts for each year from 2012 to 2019. Set white equal to the pd.DataFrame() function.
6. Create new columns in the white dataframe labeled by year that pull the 'percent_white' columns from each of the eight dataframes above (from CDpercentwhite2012 to CDpercentwhite2019).
7. Stack the white dataframe, reset its index, and set the columns of the dataframe to 'district', 'year', and 'w%' (a shorter name for the column containing the percentages of the districts that are white only).
8. Select specific congressional districts to analyze. Be careful with zeros before single-digit districts.

9. Write white dataframe to CSV file called 'white.csv'.
10. Import matplotlib.pyplot as plt and seaborn as sns.
11. Create a Seaborn line plot from the white dataframe that displays the changes in the percent of the white population in each of the selected Texas congressional districts from 2012 to 2019. The x-axis represents the year and the y-axis represents the percent of the white-only population. The legend indicates the districts that correspond to each of the nine plotted lines. Save the figure as 'CDpercentwhite.png'.

D) Script electiondata.py

This script uses the election data from MIT to plot the percentages of votes that were for Republican candidates in the specified Texas congressional districts (CDpercentrepublican.png). Uncontested election data is then eliminated (CDrepdemcontested.csv) and contested elections are then plotted in CDpercentrepcontested.png. Lastly, the racial composition data and the election data are merged in join.csv before the final line plot is generated in whitevsrep.png.

1. Import pandas as pd.
2. Set data types for specific columns by setting the variable fix_dtype to a dictionary with column names as keys and data types as values.
3. Read CSV file containing 1976 to 2018 U.S. House election data from MIT Election Data and Science Lab after you have downloaded it.
4. Filter the data down to only the 2012, 2014, 2016, and 2018 elections in Texas. Further filter that data down to only Republican and Democrat congressional candidates. Then filter that data down to the specific congressional districts you want to analyze.
    a. Be careful about not including zeros before single-digit districts when working with this dataset.
5. Create a new dataframe containing the year, the district, the number of votes per political party, the total number of votes per election, the percent of the votes for Republican candidates, the percent of the votes for Democratic candidates, and the party in control of the seat after the election. Set CDrep equal to the pd.DataFrame() function.
6. Import matplotlib.pyplot as plt and seaborn as sns.
7. Create a Seaborn line plot from the CDrep dataframe that displays the percentage of the votes for the Republican congressional candidates in each of the nine Texas congressional districts' U.S. House elections from 2012 to 2018. The x-axis represents the year and the y-axis represents the percent of the votes that went to the Republican candidate. The legend indicates the districts that correspond to each of the nine plotted lines. Save the figure as 'CDpercentrepublican.png'.

8. Query the CDrep dataframe to filter out the uncontested elections (where the percentage of votes for Democratic or Republican candidates equaled zero).
9. Create a Seaborn line plot from the CDrepdemcontested dataframe that displays the percentage of the votes for the Republican congressional candidates in only the contested U.S. House elections (where both Democratic and Republican candidates ran for public office). The x-axis still represents the year and the y-axis represents the percent of the votes that went to the Republican candidate. The legend indicates the districts that correspond to each of the nine plotted lines. Save the figure as 'CDpercentrepcontested.png'.
10. Write the CDrepdemcontested dataframe to a CSV file and read it into a new variable. Read 'white.csv' into a new variable.
11. Merge the reppercent and whitepercent datasets into new join dataframe using a one-to-one inner join based on their shared "year" and "district" columns. Write join dataframe to CSV file.
12. Create a Seaborn line plot from the join dataframe that displays the percent of the white population in each of the selected Texas congressional districts opposite the percent of the votes that went to Republican candidates in contested U.S. House elections from 2012 to 2018. The x-axis represents the percent of the population that is white only and the y-axis represents the percent of votes that went to the Republican congressional candidate in a given election. The legend indicates the districts that correspond to each of the nine plotted lines. Save the figure as 'whitevsrep.png'.
    a. Ensure that the 'district' column is treated as a categorical variable by converting the 'district' column in the join dataframe to a string data type.

Results

The purpose of this project was to see if there was an identifiable relationship between the presence of fast-growing cities in a congressional district and that congressional district's voter turnout. I was motivated to see if the presence of these fast-growing cities in Texas' congressional districts was impacting election results, being that I am from a congressional district experiencing a large increase in population.

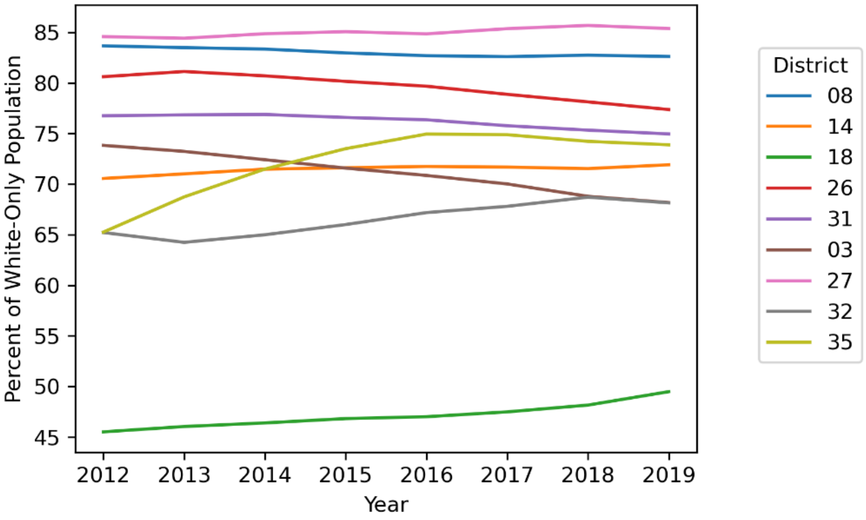
Six cities in Texas were identified by the U.S. Census Bureau as being within the top 15 fastest-growing large cities by percent change from 2010 to 2019, with large cities meaning a population of greater than 50,000 people. Those six cities were: (1) Frisco, TX, (2) New Braunfels, TX, (3) McKinney, TX, (4) Cedar Park, TX, (5) Conroe, TX, and (6) Round Rock, TX (https://www.census.gov/library/visualizations/2020/demo/fastest-growing-cities-2010-2019.html).

My initial hypothesis involved several assumptions: that population growth in a congressional district meant increased diversity, that increased diversity meant a higher percentage of Democratic voters, and that an increase in Democratic voters meant a greater likelihood of Democratic representation in those specific congressional districts. I expected to see congressional districts with fast-growing cities experience a decline in the percentage of their population that identifies as white only and a decline in the percentage of votes for Republican congressional candidates in U.S. House elections.

I first identified the specific Texas congressional districts I wanted to analyze. I looked at the congressional districts associated with the fast-growing cities, and I saw that two congressional districts had two of the fast-growing cities in them (although cities often do not fall perfectly within the boundaries of congressional districts). TX-3 had Frisco and McKinney and TX-31 had Cedar Park and Round Rock. I then identified three districts with only one fast-growing city that fell within its respective boundaries. Almost half of Frisco fell into TX-26, Conroe fell solely into TX-8, and New Braunfels fell across three (dare I say fairly-gerrymandered) districts, so I chose TX-35. I then proceeded to identify four congressional districts that not only did not have any of the fastest-growing cities in them, but that had also experienced decreases in population from 2010 to 2019. Given that the districts with the fastest-growing cities all tended to cover suburb areas around the Dallas, Austin/San Antonio, and Houston areas, I decided to focus on selecting my four congressional districts with population losses from those areas too. I picked TX-32 from the Dallas area, TX-27 from the Austin/San Antonio/Corpus Christi area (again, the congressional boundaries are drawn oddly), and TX-18 and TX-14 from the Houston area.

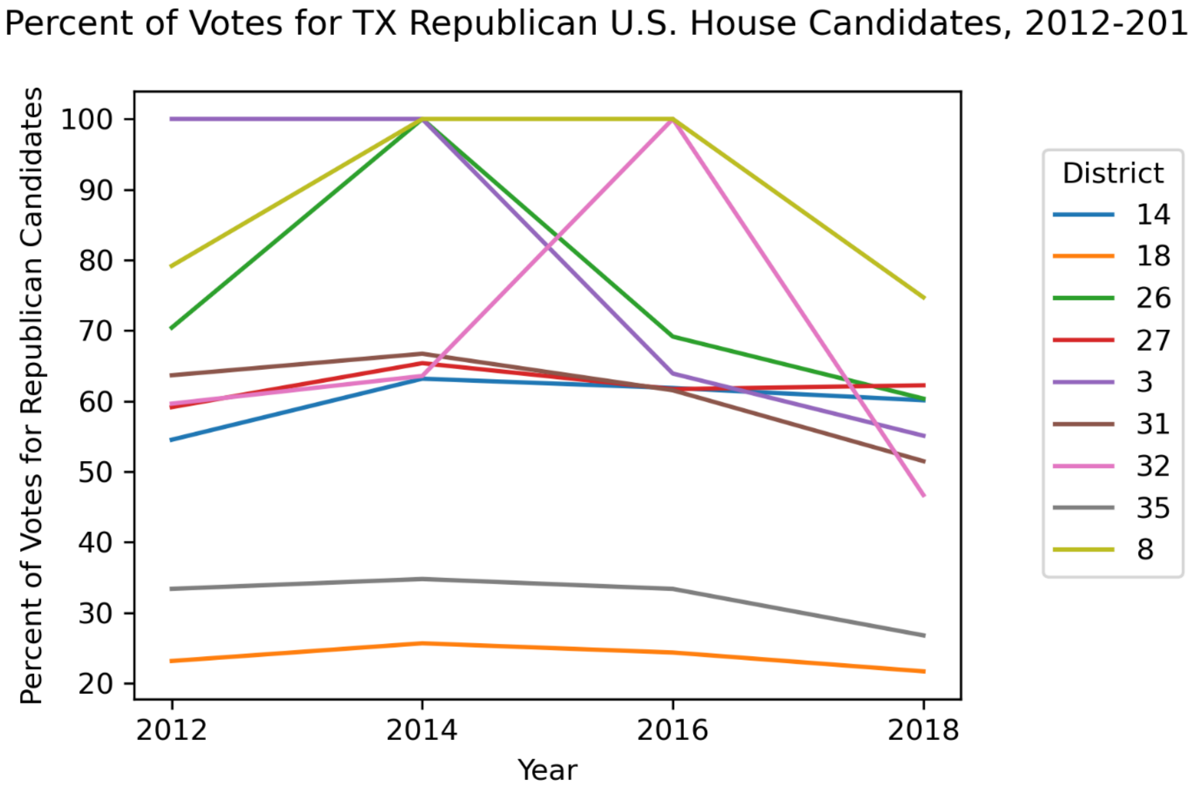
Once I had my nine Texas congressional districts selected, I gathered the racial data for each of them from 2012 to 2019. I started with 2012 because that was the first year that the redrawn districts from the 2010 census were used in U.S. House elections.

White Percent of Population in TX Congressional Districts, 2012-2019

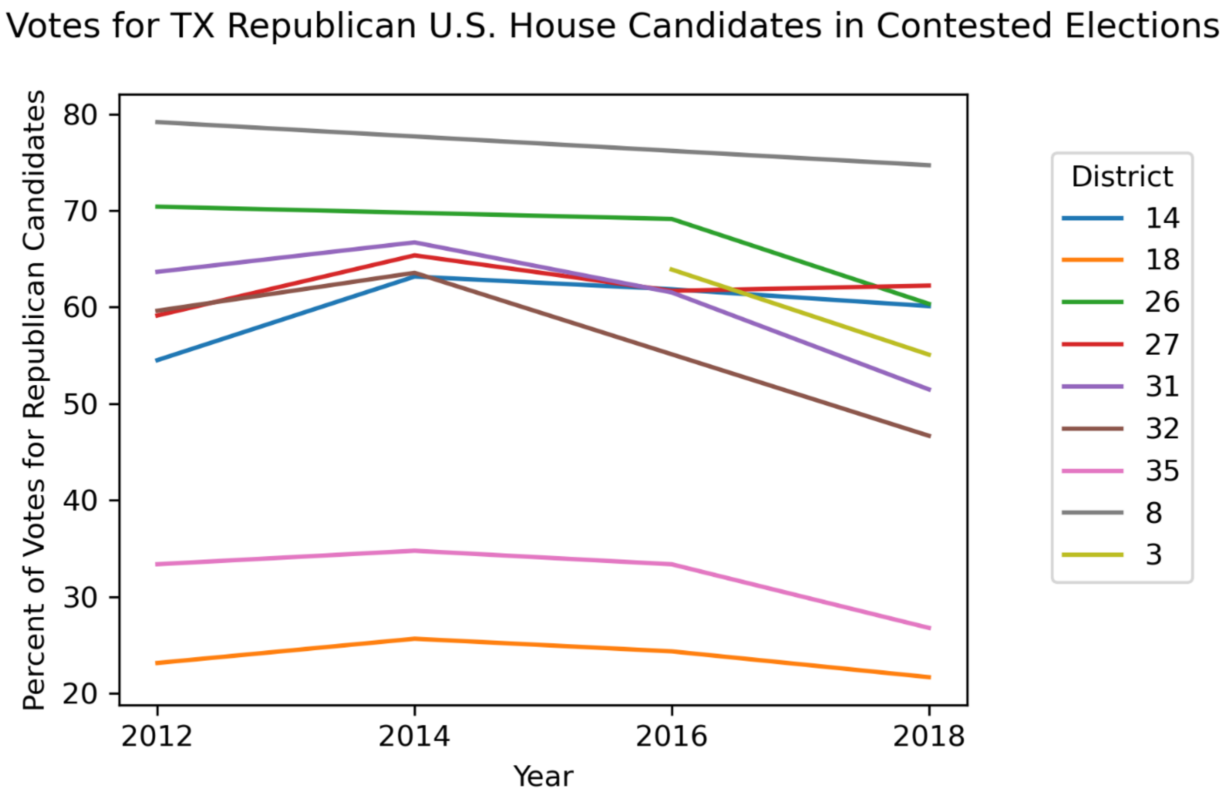


I expected to see a decline in the percentages of the white-only populations in the congressional districts with fast-growing cities (TX-3, 31, 26, 8, and 35) and an increase in the percentage of the white-only population in the congressional districts that had population declines (TX-32, 27, 18, and 14). This was mostly the case. The two districts with two fast-growing cities each (TX-3 and 31) both saw decreases in their white-only populations, while each congressional district with population loss saw increases in their white-only population percentages. The one outlier in this group was TX-35, which not only had a population increase, but saw a significant increase in its white-only population (so a lot of white-identifying individuals moved to this district).

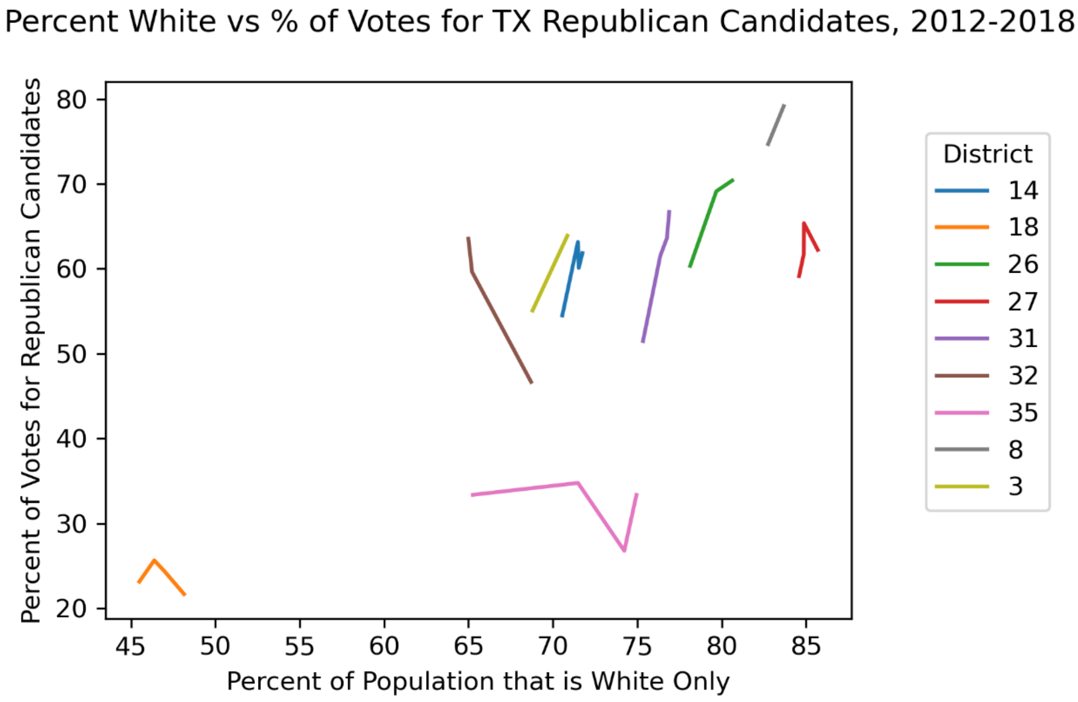
I moved on to calculating the percentage of the votes in each of the nine Texas congressional districts' U.S. House elections from 2012 to 2018 that were for Republican congressional candidates. Again, I expected to see a decrease in the percentage of votes for Republicans for fast-growing districts and an increase in slow-growing districts. But I ran into an issue because I had made another assumption. I had assumed that every election was contested, in that every election had both Republican and Democratic congressional candidates to vote for. This was an incorrect assumption. Six elections ended up having only a Republican candidate and no Democratic candidate, resulting in 100% of the votes being counted as going toward the Republican candidate.


Percent of Votes for TX Republican U.S. House Candidates, 2012-2018

Since this was not useful for my analysis, I decided to eliminate the uncontested elections from the dataset and plotted only the contested elections. Contested elections would show me whether the percentage of votes for Republican candidates was increasing or decreasing as the percentage of white-only individuals in each district changed.



The elimination of left me with just a few points in some of the districts, but we carried on. TX-3 and 31 both saw decreases in the percent of votes for Republican candidates over time (although two points for TX-3 is not a lot of information to go off of). TX-26 and TX-8 both decreased over time as well, but TX-35 continued to go against my assumption. TX-35 saw an increase in its white-only population but it saw a decrease in percent of Republican votes, so it can be assumed that many liberal-leaning white-identifying individuals moved to the district (which makes sense given the district's positioning outside the generally liberal-leaning capital of Austin). The four districts with population decreases did not have consistent trends. TX-32 switched from electing a Republican in 2014 to electing a Democrat in 2018 (something I did not expect for a population-declining district). TX-18 remained strongly Democrat-leaning, and both TX-27 and 14 increased and decreased in Republican-supporting vote percentages over the years.



When I plotted white-only population percentages against vote percentages for Republican candidates in contested elections, I was not surprised to get semi-inconclusive results, given the variance in patterns observed earlier. While four of the districts (TX-3, 31, 26, and 8), all of which were fast-growing, tended to show the relationship I assumed in higher white-only populations being correlated with higher percentages of votes for Republican candidates, the other districts tended not to follow any pattern. And this makes sense.

I think part of the reason this graph is a bit inconclusive is that there are a whole host of other factors that can impact voter turnout and subsequently election results. It cannot be assumed that everyone that moves to a new district votes. Other factors like which party the current President is part of, how long an incumbent has been in office, and how accessible voting laws are (Texas certainly has what some would consider voter suppression laws) can also impact election results. Future analyses would probably need to control for other variables to find a meaningful relationship.

Regarding future uses, one could look at how the changing populations of these fast-growing cities impact the election results of mayoral elections rather than congressional districts. However, if sticking with congressional districts, I think it would be interesting to dive deeper into the block group level and analyze the racial composition of different neighborhoods in fast-growing areas. It should be noted, though, that congressional district lines are being redrawn at the moment. Many states do not have independent commissions in charge of redistricting, so a study like this could actually be used to influence the way future districts are drawn (based on trends in population growth and racial composition). This could lead to gerrymandering, which is something I would want future researchers to be wary of.