SI 618 Project 2 Report

The 2020 Election, COVID-19, and Demographics

Motivation

The year 2020 has been tumultuous. Covid-19 has spread quickly across the US in multiple waves and the president's response to Covid-19 has become a topic of debate in recent months, especially in the context of the 2020 election. The 2020 election has just been called for Joe Biden. With the nation still divided, I am curious to know what we can learn about the attitudes of our nation. Specifically, I am interested in understanding the potential relationships between Covid-19, demographic data, and the 2020 presidential race. The three questions I will be examining in this report are:

1. Are Covid-19 Cases per County and the 2020 Presidential Election Results Related?
2. Did an Increased Covid-19 Presence Polarize Counties' Political Preferences?
3. What Can We Learn About the Election Through Demographic Information?

Initially, I had hoped to examine the relationship between Covid-19, the election results and various demographic features. However, the analysis of the first two questions showed no significant results and conducting an analysis became futile. I therefore shifted my focus on question three to examining demographic features in relation to the election. Additionally, I narrowed my scope to the presidential election, since congressional terms last 6 years.Is there a relationship between Covid-19 and the Presidential election results?

Data

Covid-19 Data
Location: here
CSV Name(s): CONVENIENT_us_confirmed_cases.csv, CONVENIENT_us_deaths.csv
Description: Contains covid data at the county level by day
Last updated: Nov 9, 2020

2020 Election Data
Location: here
CSV Name(s): president_county.csv
Description: This includes two CSV's reporting the breakdown of votes per political party by county in both the Presidential race and the Senate race.
Last updated: Nov 9th, 2020

2016 Election Data
Location: here
CSV Name(s): US_County_Level_Presidential_Results_12.csv
Description: This includes two CSV's reporting the breakdown of votes per political party by county in both the Presidential race and the Senate race.
Last updated: 2019

Census Data
Location: here

CSV Name(s): acs2015_county_data.csv
Description: This includes demographic information by county.
Last updated: 2017 (This is the most recent census data available)

Zip_county, county, and abbs were dataframes used to help convert county names and ids and can be found in the data folder.

Methods

Preparing the data for analysis was the most challenging part. Please see si618_project2_data_preprocessing_llluns.ipynb for data preparation and cleaning.

Covid-19 Data

The Covid-19 data used counties as columns and each row was a day in time. I needed the rows to be counties and columns to be the total number of Covid-19 cases or deaths up to two specific dates. I had to clean up the column names, create subsets of the tables that only included data up to the election date and up to a month before the election, transpose the data and then sum all the cases and deaths.

2020 Election Data

Initially, the data was structured with for rows per county, each row representing the statistics of a different party. I had to compute the % of votes each party got and change the structure of the data such that the rows are one county and the columns are the %'s of each party.

2016 Election Data

The first 28 rows of the election data were all the same data, so cut them out of the table. Instead of state and county names, there was the fips number. I had to find a csv file that allowed me to translate the fips number to a state and county. I then got the formatted state and county and created a dataframe with the relevant information.

Census Data

I selected out the relevant columns and created the id column so it could be joined.

Joining

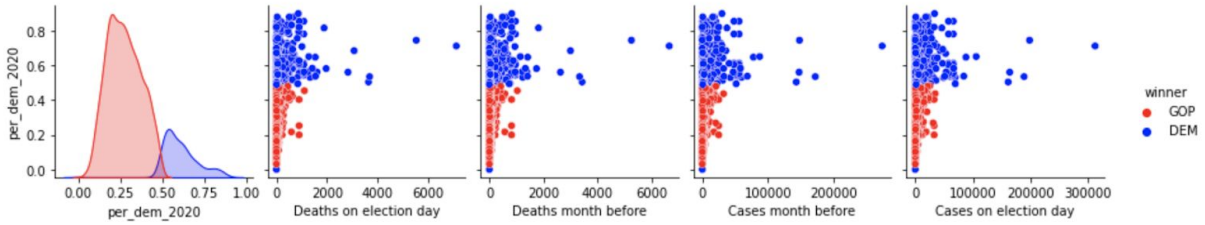
I first joined all data frames that had similar ids (the census and both election dataframes). This join lost 100 rows. I then created the id consisting of the state and first word of the county name and merged it with the Covid-19 data, which resulted in a loss of another 100 rows. There are 3001 counties in the US and I have only ~2800 rows. This means ~6% of my data is missing.

One particular challenge was joining the Covid-19 data with the rest of the data. Using the state_county as a key worked to join the census and election data. However, the Covid-19 data formatting the counties differently. The census and election data labeled counties as "Autauga County" whereas Covid-19 data labeled it as "Autauga". A seemingly simple fix for this would be to append " County" to the counties, except not every county is a county, some have different labels. When I first tried this method, a lot of rows were dropped in the join. To get around this, I created a new ID that joined the state and the first word of the county. Using .unique() I ensured there were no duplications with this method. I then merged this on the label.

Analysis and Results

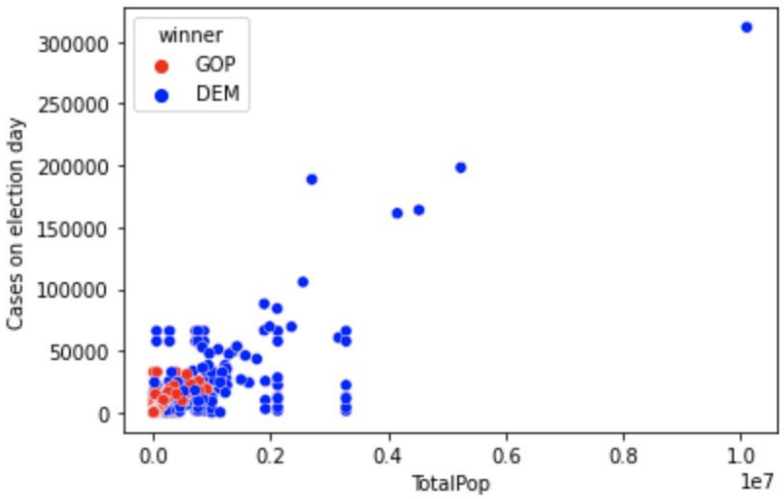
1. Are Covid-19 Cases per County and the 2020 Presidential Election Results Related?

First, I added a few extra columns to help in my analysis. These columns told me the winners of the 2020 and 2016 election(useful for plotting hues), *winner & winner16* respectively. Additionally I calculated the change in percentage points for the democratic candidate between 2016 and 2020 and added it to the column *delta*. The first thing I wanted to examine was if there was a visible relationship between Covid-19 and the presidential election results. Since I created 4 subsets of Covid-19 data (deaths and cases looked at up to election day and up to a month before election day), I wanted to examine all four at once against the percent of residents who voted for the democratic candidate (general trends for the republican percent can be inferred as the inverse of the democratic data). I also labeled the hue of each data point based on the *winner* column, to visualize which candidate won in each county.

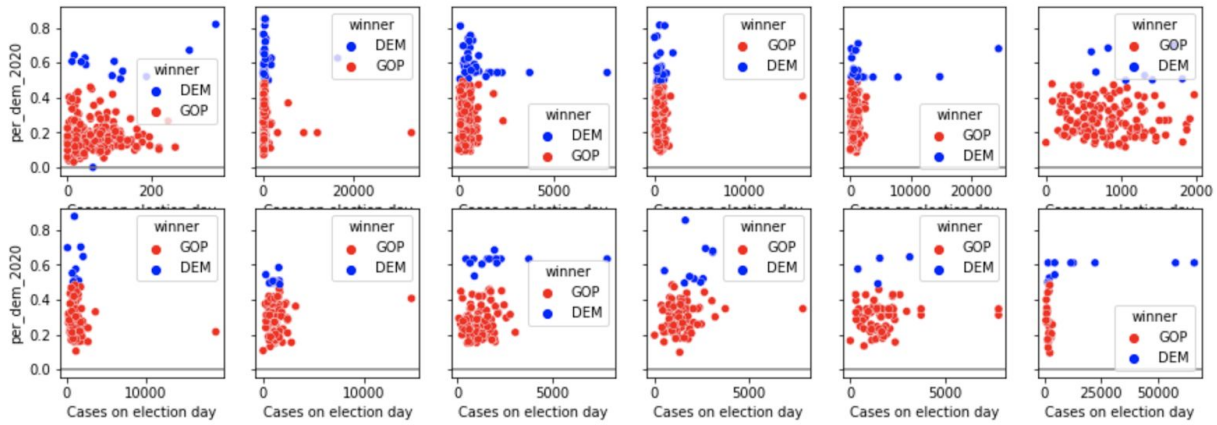


What I found was that there was very little difference in patterns between each of the four subsets of Covid-19 data on the relationship with the percent of voters who voted for the democratic candidate. Furthermore, I noticed more variation of democratic counties, indicating a possible relationship between Covid-19 and the election. I wanted to explore this more.

But first, I wanted to understand if there was a relationship between Covid-19 cases and population size. Intuitively, areas with a greater population density will get more Covid-19 . Since liberal voters frequently prefer more densely populated areas, it was important to understand this relationship to identify a possible confounding variable.



As suspected, there's a pretty strong relationship between Covid-19 and population total per county. Because of this, I wanted to examine the relationship between Covid-19 and the election while holding the population steady. To do this, I employed stratification, a method by which you examine data where the confounding variable has the same value to remove the effect (Pourhoseingholi, 2012). Instead of looking at just two similar data points, I wanted to increase the accuracy and bin the data into similar groups. I used a histogram of the population data to identify a reasonable bin size of increments of 5,000 people. Binning increases the amount of data which increases accuracy, but also increases the noise since all the population values aren't the same, just close. I took slices of the data spanning 5,000 people in population at a time. For instance, the first graph looks at counties with 0-5,000 people and compares Covid-19 to the % of democratic voters in the county. I did this 12 times hoping to find a consistent pattern.

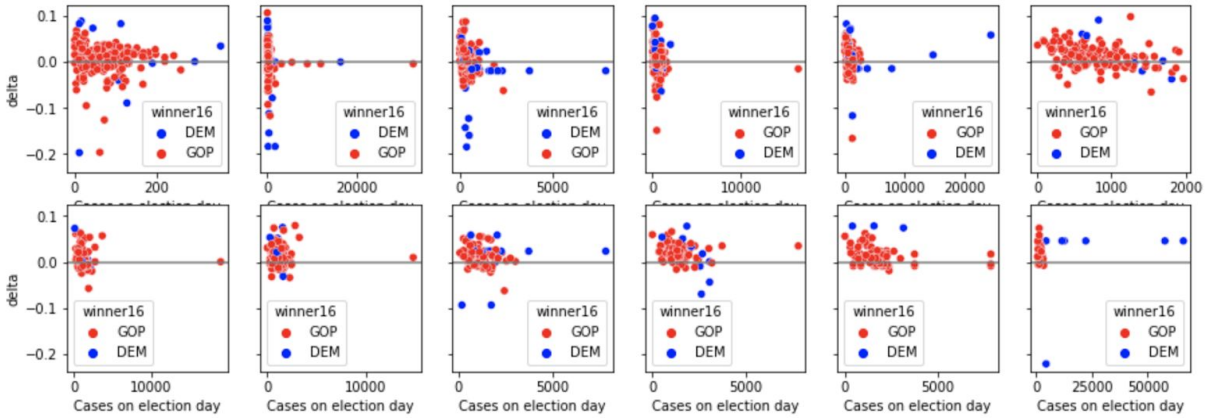


Unfortunately, no clear pattern arose. I also calculated the correlation coefficients for each slice and found little to no relationship across all slices.

2. Did an Increased Covid-19 Presence Polarize Counties' Political preferences?

I was curious to see if perhaps the relationship between Covid-19 and the changes of voting habits had any stronger relationship. To explore this, I used the aforementioned *delta* column to see the changes in voting preferences across counties. If the delta was positive by 0.02, that meant the county's democratic voters increased by two points from 2016 to 2020. I applied the same stratification techniques to the delta data, except looked hue based off of who won in 2016, so I could quickly visually identify if there were any groups of counties that moved

significantly.

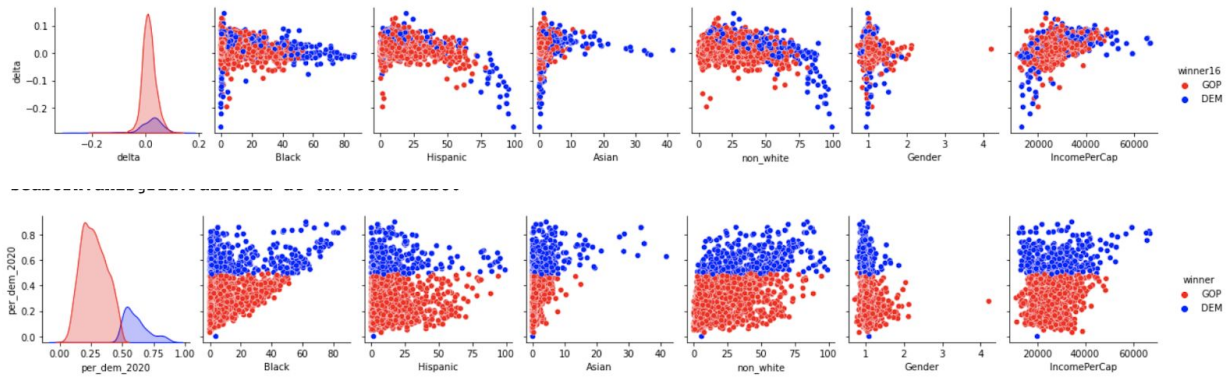


We know from question 1 If there was polarization due to Covid-19 , if democrats were more democratic and vice versa, we would see blue dots be more positive as cases increase and red dots be more negative. However, we don't visually see this, meaning it is unlikely that the presence of Covid-19 polarized political preference.

A positive correlation between the delta value and Covid-19 cases would indicate that an increased presence of the virus led counties to be more democratic. A negative correlation would indicate counties were responding to the virus by voting more republican. However, when a correlation coefficient was calculated for each of the slices, no correlation was found as the values averaged -0.08.

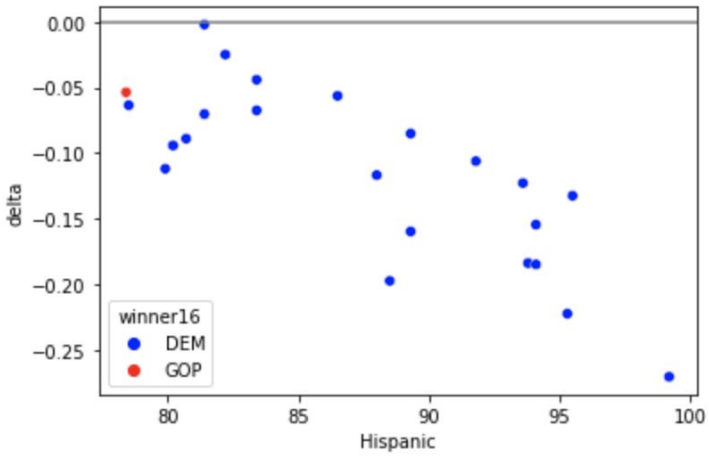
3. What Can We Learn About the Election Through Demographic Information?

Since there was no interesting take away from the Covid-19 analysis, I couldn't go as in depth as I had originally planned, and therefore shifted my focus to looking at the election through demographic features. To do this, I first calculated a ratio of men/women to standardize gender across population size. I then took the gender ratio, %of Black residents, %of Asian residents, %of Hispanic residents, %of Non-white residents, and the Income per capita and plotted them against the % dem and the delta change from 2016 to 2020. The hue of the data is the winning party in 2016 when comparing the delta, and the winning party in 2020 when comparing the %dem.

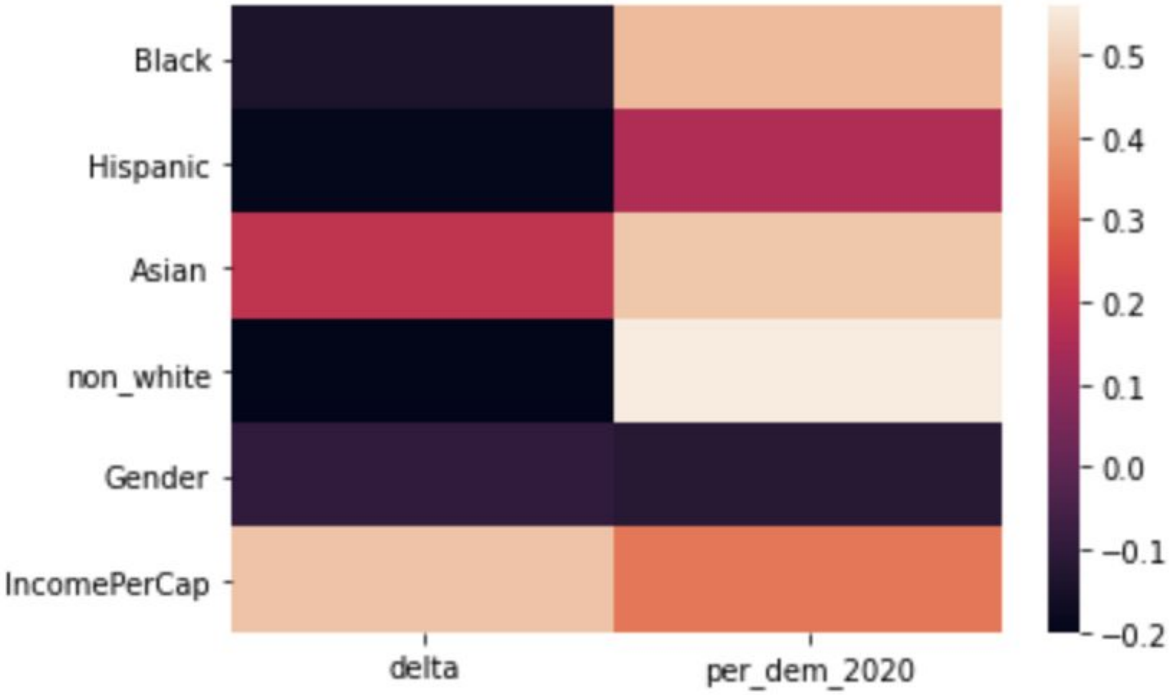


What I found interesting was the comparison of Hispanic counties in 2016 that were initially blue, trended more republican in 2020. This finding aligns with the push in campaign

spending the Trump administration made to try to recruit Hispanic and Latinx voters (Valdes, 2020). To zoom in on this finding, I created a subset of data where Hispanic population is above 75% and plotted it against the change in % dem.



Returning to the larger view, I also mapped the correlations between the demographic factors and the two election metrics. Another interesting observation at this stage is the moderate correlation between income per capita and becoming more democratic. This correlation suggests that wealthier counties shifted more blue in 2020.



Sources

Pourhoseingholi, Mohamad Amin, et al. “How to Control Confounding Effects by Statistical Analysis.” Gastroenterology and Hepatology from Bed to Bench, Research Institute for Gastroenterology and Liver Diseases, 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC4017459/.

Valdes, Marcela. “The Fight to Win Latino Voters for the G.O.P.” *The New York Times*, The New York Times, 23 Nov. 2020, www.nytimes.com/2020/11/23/magazine/latino-voters.html.