**The GDP Change in Swing States After the 2016 Election**

**Group 3 - Priyanka Khanal, Madison Moye, Bhanuprasad Tudy, Yuxiang Wang, and Ethan Goldbeck**

*Background*

The 2016 U.S. presidential election was one of the most talked about election not only in the States but also across the world. Several factors helped distinguish the high magnitude and traction from this election between the two major political parties. One of the most crucial and exciting factors within this particular election was the uncertainty of “Swing States.” The term, *swing state*, describes a US state where the number of voters of the two different parties hold a similar number of votes. These states can be considered to be the most crucial when it comes to the overall result in a presidential election. In most states, the election results are usually predistinguised as almost every state supports one or the other political party. However, in this election, these were several swing states, which were the determining states for the winning party this election. Of the *swing states* that ended up voting for the Democratic Party, they consisted of Colorado, Virginia, Nevada, New Hampshire and North Carolina. The *swing states* that ended up voting for the Republican Party were Florida, Iowa, Michigan, Wisconsin, Ohio and Pennsylvania. These *swing states* brought up a lot of discussion on not only their status on which political direction they would go, but also the change in their economy. The 2016 presidential election generated a lot of buzz and ended up changing the United States’ economy.

*Introduction*

For this statistical analysis, we decided to take one of America’s monumental political events and explore the effect of its *swing states* in relation to Gross Domestic Product (GDP). GDP provides an “economic snapshot of a country, used to estimate the size of an economy and growth rate” (Chappelow). After the 2016 presidential election, we wanted to analyze the following dataset to see if the 2016 *swing states* had any or no effect on America’s GDP. **Does change in GDP have an effect on election outcomes? More specifically, does the effect of GDP affect the change in percentage of republican votes (2012-2016) for counties in the twelve *swing states***. After briefly assessing several different datasets, we found a compilation of data that would be beneficial in helping define and answer our research question.

*Datasets*

In order to answer our research question, we gathered our data from three different datasets. From Diego García, we grabbed the Case01 dataset, a compilation of data on the US presidential election (2000-2016) at the county level. Case01 contained the following variables: year, state, state po, county, FIPS (numeric code for county), office , candidate , party (Dem/Rep/Ind/etc), candidate votes, total votes, version (database version). This dataset provided us with the readily available information on the presidential election. Since we are specifically looking into the effect of GDP on the counties in these twelve *swing states*, it was essential to use this dataset to pull information from based on county. Our first step with this dataset was to index the candidate votes and the total votes. Then, individually, we could pull out this information based on our specific *swing state* and look at the percent change.

Our next dataset was taken from Kaggle, an online Google community where users can find and upload data sets. This user, MuonNeutrino, pulled this information from the DP03 and DP05 tables of the 2015 American Community Survey 5-year estimates. Within his kernel, there were two datasets worth noting: one 2015 and one 2017 on the U.S. census. Both datasets contained the following variables: censusID, state, county, total pop, gender, race, citizen, income, income err, income per cap, incomepercaperr, poverty, child poverty, job specification (professional, service, office, etc.), transportation to work, employed, private work, public work, self employed, and unemployed. In order to use these two datasets into our analysis, we merged the two datasets by using the cbind function. We pulled out import columns from the two datasets, so we used the census data with our own *swing state* analysis.

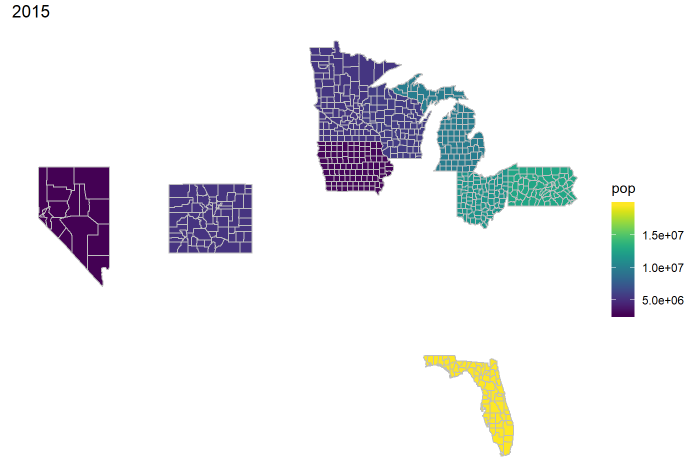
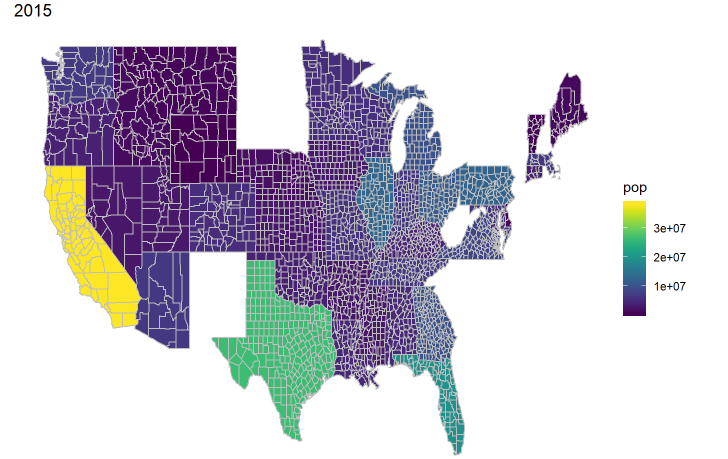
Our last dataset was our lagdp dataset, which shows the real gross domestic product by county (2012-2015). This dataset is broken into two categories: *real GDP (thousands of chained 2012 dollars)* and *percent change from preceding period* all of which are distinguished by state and county in rows. After uploading this GDP dataset, we were able to index only the rows we needed for specific *swing states.* We had to manipulate the data and use the as.numeric function to coerce the data into integers as well as using the gsub function to replace “ ” and spaces. From this dataset, we were able to analyze the percent change and GDP change by using the summary(lm).

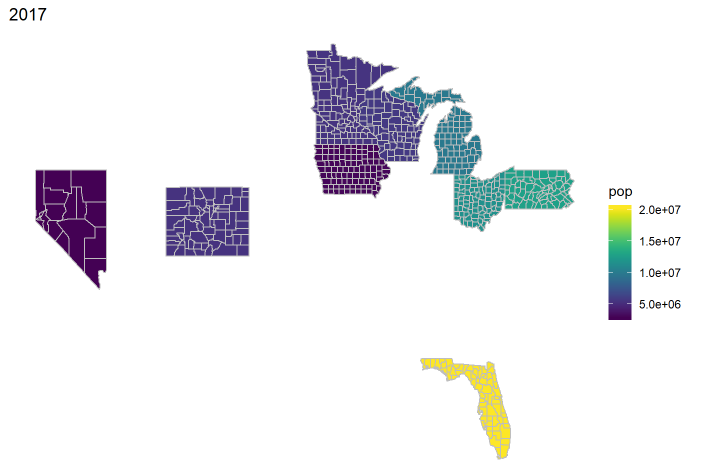
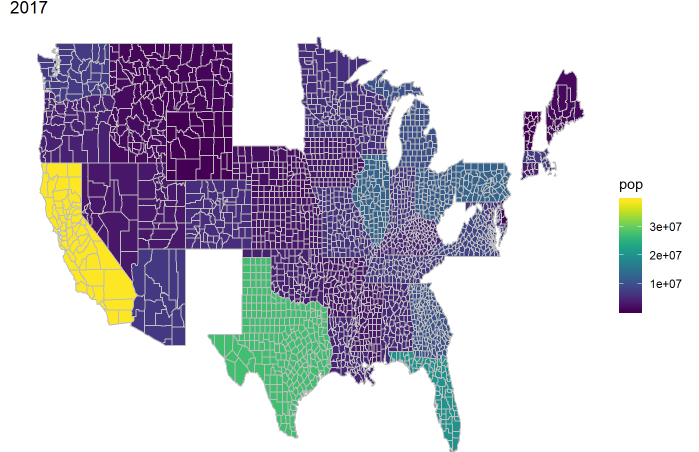
*Findings*

For our first analysis let’s look at the regression model for the change in GDP , vote percent change and Total Population change for the swing states. To calculate the simple regression model we used lm() function in R.

We did data cleaning before we built the model to analyze, there are some missing data like “NA” , so need to delete it or change them to 0s, we also changed all data type to numeric which makes data serve us better, we keep 0s because our analytic based on percentage, so 0’s are not affecting our final conclusion.

After data cleaning, we created a map for each state we selected about total population, its a visual way to know the overall population from 2015 and 2017. From the below maps, we found Folrida has the most amount of total population, and Pennsylvania at second place, we made a hypothesis , more total population changes resulted in more vote changes and gdb changes, we veerided our hypothesis in our later work. From the maps, we also found Total populations didn’t change much from 2015 to 2017.





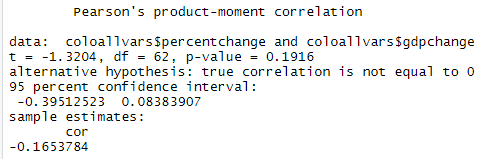
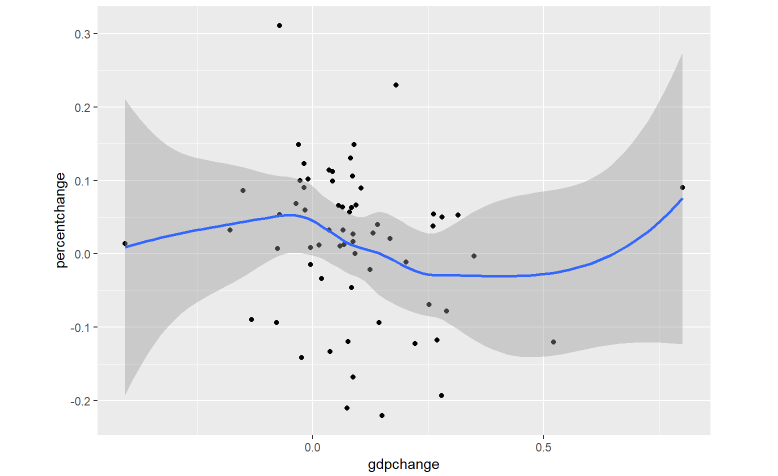
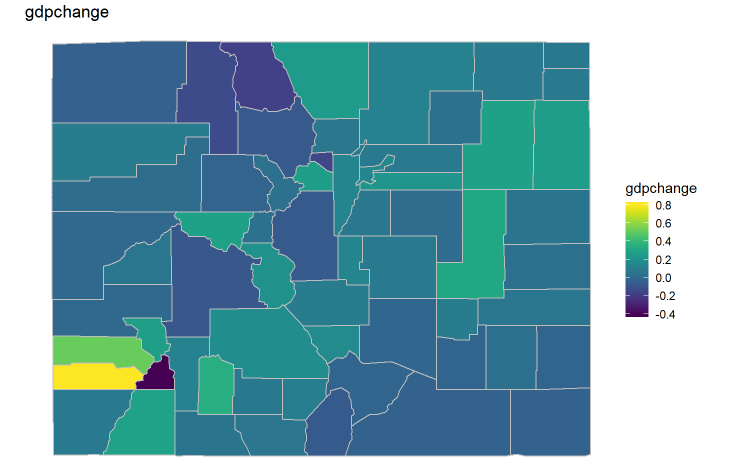
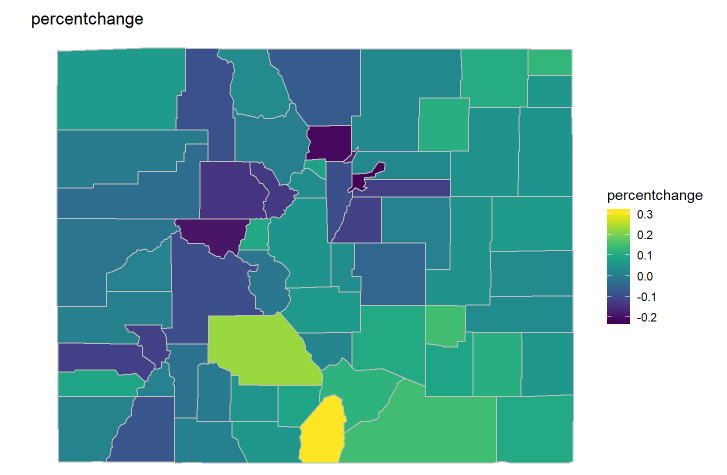
After analyzed the overall total population for each state, we started to build analyzed the correlations between each attributes for each state.First, we selected 2012 and 2016 of each state which we selected from Case01 dataset, and use these to calculate the percentage of vote changing for each state, after that, we pop out the data from gdp dataset which matches up the state we selected, then calculated the percentage of gdp changing, then build a regression model to analyze the correlation between vote changing and gdp changing. Finally, we select the data from census data which we combined 2015 and 2017 as one dataset to calculate the percentage of population changing, then add population changing to our regression model, but we didn’t find strong relation to analyze. Therefore, we add all attribution form census data to see which factor has the most effects for vote changing.

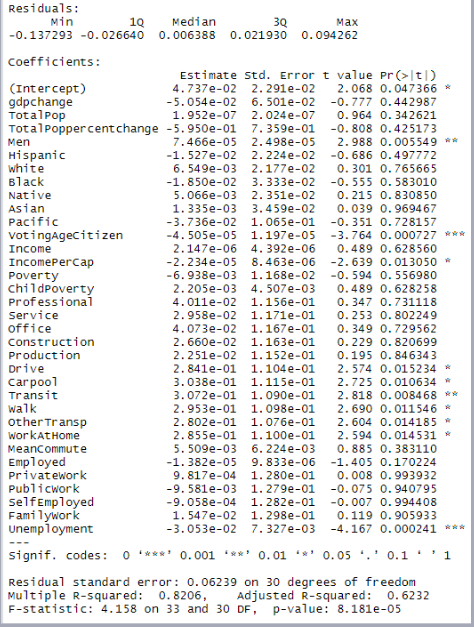
Colorado:

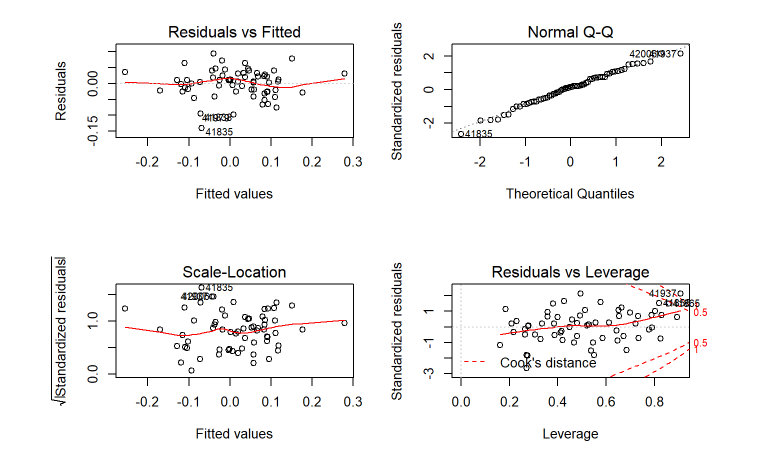
As for Colorado, we calculated the vote percentage change, total population percentage change and gdp percentage change for each country in colorado. We also created a regression model to analyze, then display the result by graphics.

From the below graphics, we can found the correlation almost 0 between the vote changes and gdp changes, from the map, the color from light to dark means the percentage from high to low.We found Costilla has the highest percentage of vote changes but the percentage of gdp changes is really small. Dolores has the highest percentage of gdp changes meanwhile the percentage of vote changes is also high. As for Routt, both of votes changes and gdp changes are pretty low, so its pretty hard to find the linear correlation. Because of this, we built a correlation scatter plot to analyze the overall correlation, from the plot, we found the range of points is quite diverse, and correlation is pretty small. We the correlation is -0.16 which is pretty small and p-value 0.1916 > 5% which is quite big, and there is no linear relation between vote changes and gdp changes.

After that , we built a regression model to predict the dist value about the vote changes with gdp changes,and all attributes from census dataset, ( TotalPop , Men , Hispanic , White , Black , Native , Asian , Pacific , VotingAgeCitizen , Income , IncomePerCap , Poverty , ChildPoverty , Professional , Service , Office , Construction , Production , Drive , Carpool , Transit , Walk , OtherTransp , WorkAtHome , MeanCommute , Employed , PrivateWork , PublicWork , SelfEmployed , FamilyWork , Unemployment),



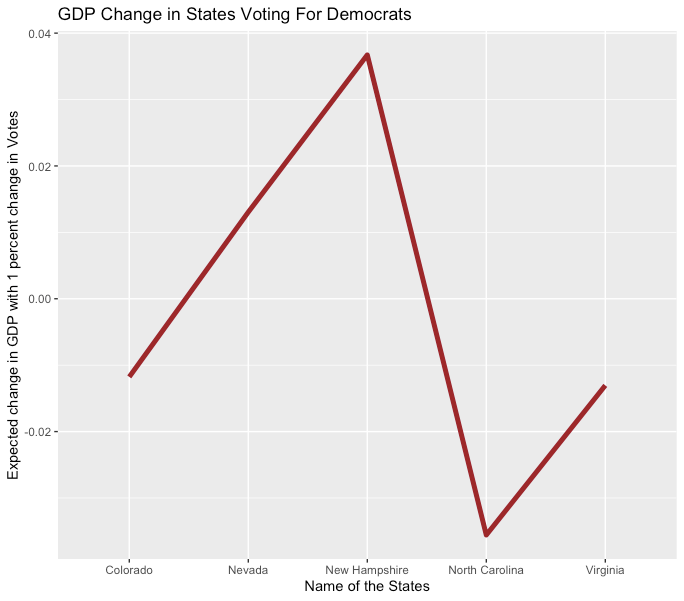
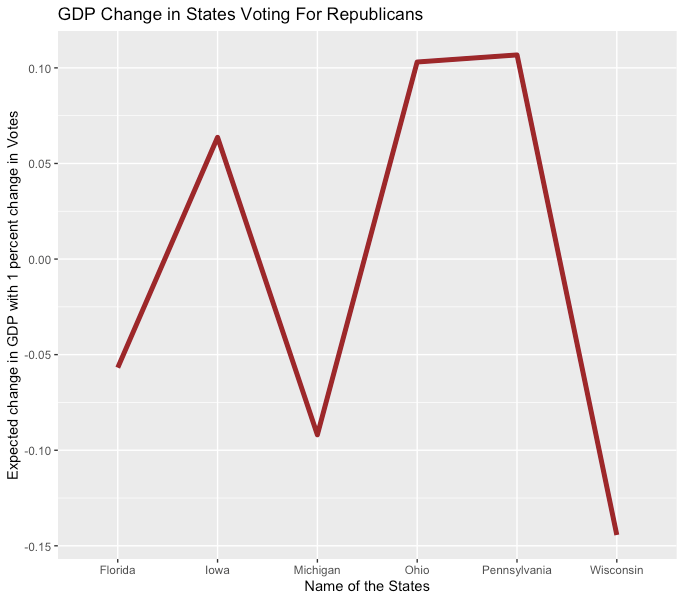




From our regression model,we can see that the distribution of the residuals do kinds of appear to be strongly symmetrical which means That means that the model predicts certain points that fall NOT far away from the actual observed points.From the out-put for our regression , we found bot of gdp changes and total population changes have negative relation with vote changes, but the p-value of couple attributes we are interested in are pretty big, which is > 5% which means we can’t reject the null hypothesis , and some attributes are not statistically significant . So it's pretty hard to predict by using these attributes. We also found some good attributes which can be used for predicting, like unemployment which p-value is 0.000241 and absolute value of t -test is greater than 2. The R^2 for our model is not bad which is 62% which means our data is pretty fitting for model.

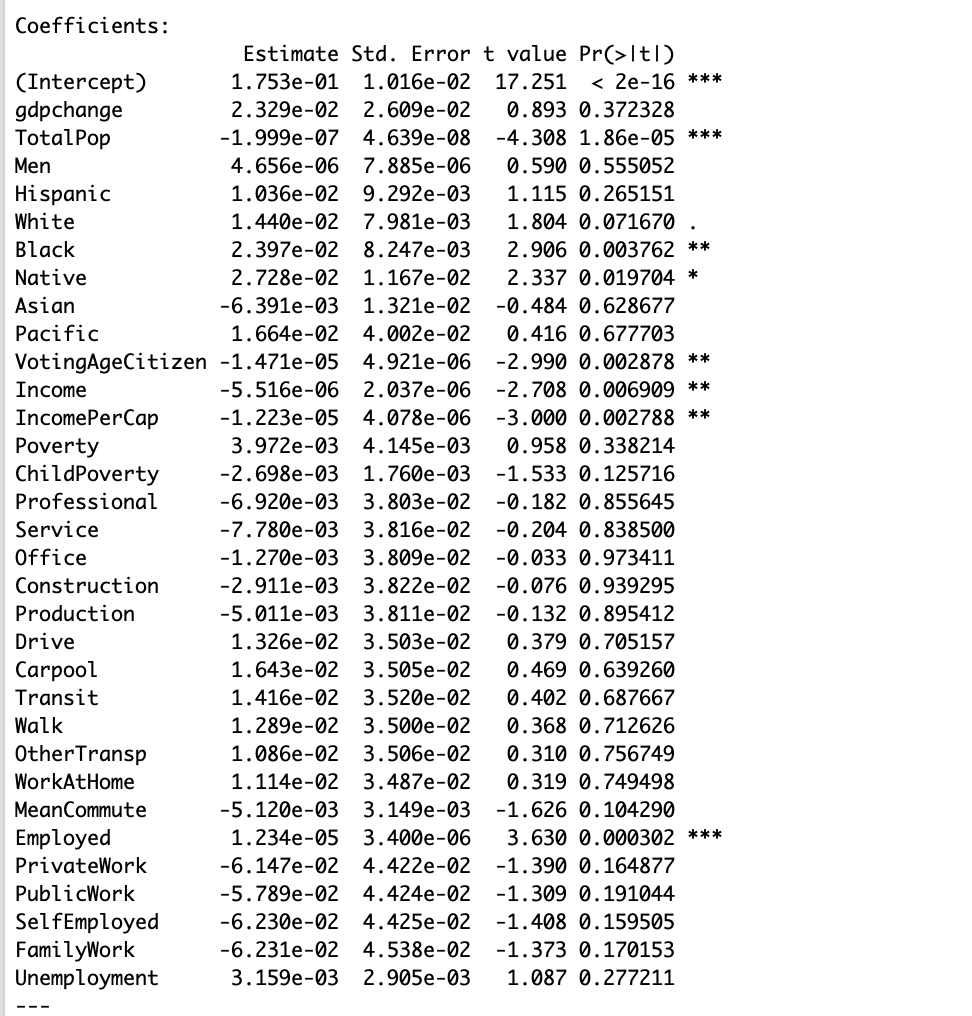
In order to evaluate our output of regression model, we created a regression plot. From the Residuals vs Fitted, we found the regression assumptions is pretty good for our model, the shape is liner then we can make sure our model is unbiased. The Q-Q plot shows the linear shape which is great, which means our data is ideally and residuals are normally distributed. The scale- location ,This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points.The residuals vs leverage plot helps us to find influential cases if any. Not all outliers are influential in linear regression analysis . Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn’t be much different if we either include or exclude them from analysis. From our data for Colorado, we find there is no points outside the cook’s distance lines, but some points are pretty close to this line at the top right of plot, there might be some outliers would affect our final result.

The graph below shows the change in GDP in these swing states that voted for Republican party.

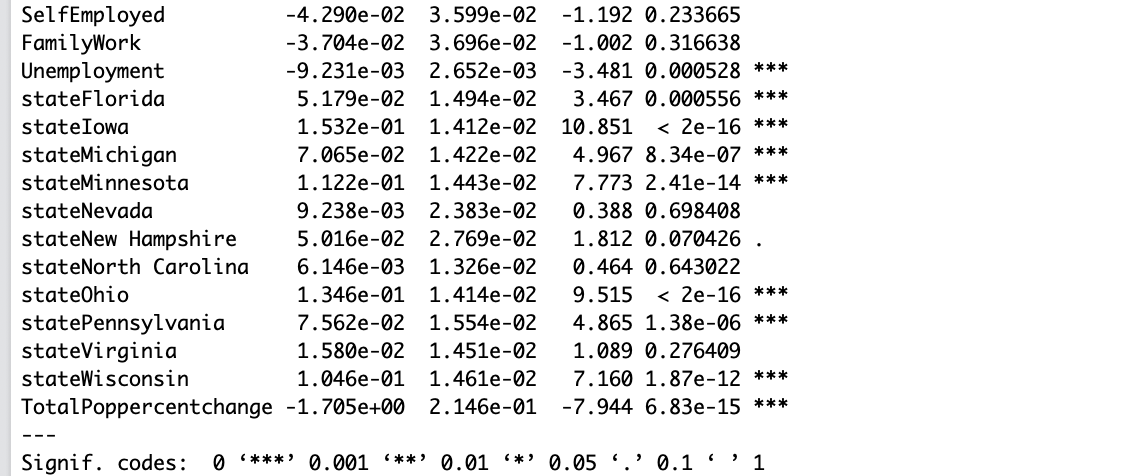


As we can see the GPD change in Florida, Michigan and Wisconsin is in negative whereas the GDP change in Iowa, Ohio and Pennsylvania is positive.Now, let us look at the GPD change for the states that voted for Democratic party. As we can see in the graphs above, the GDP change in Colorado, North Carolina and Virginia is in negatives and the GDP change in Nevada and New Hampshire is in positive.

The first two graphs only focused on the change in GDP with the percent change in votes. Next, we would like to analyze the change in demographics in these swing states. For this, we divided the states in three different categories. The states were classified as small if the total population is less than 20400. The states were classifies medium if the population was greater than 20400 or less than 61750 and the state was classified as large if the population was greater than 61750. Along with that, we added some predictive variables like Total population, gender, race, Voting age, Income, poverty, mode of transportation in the regression model to see if they have any significant effect on the percent change in votes in these swing states.

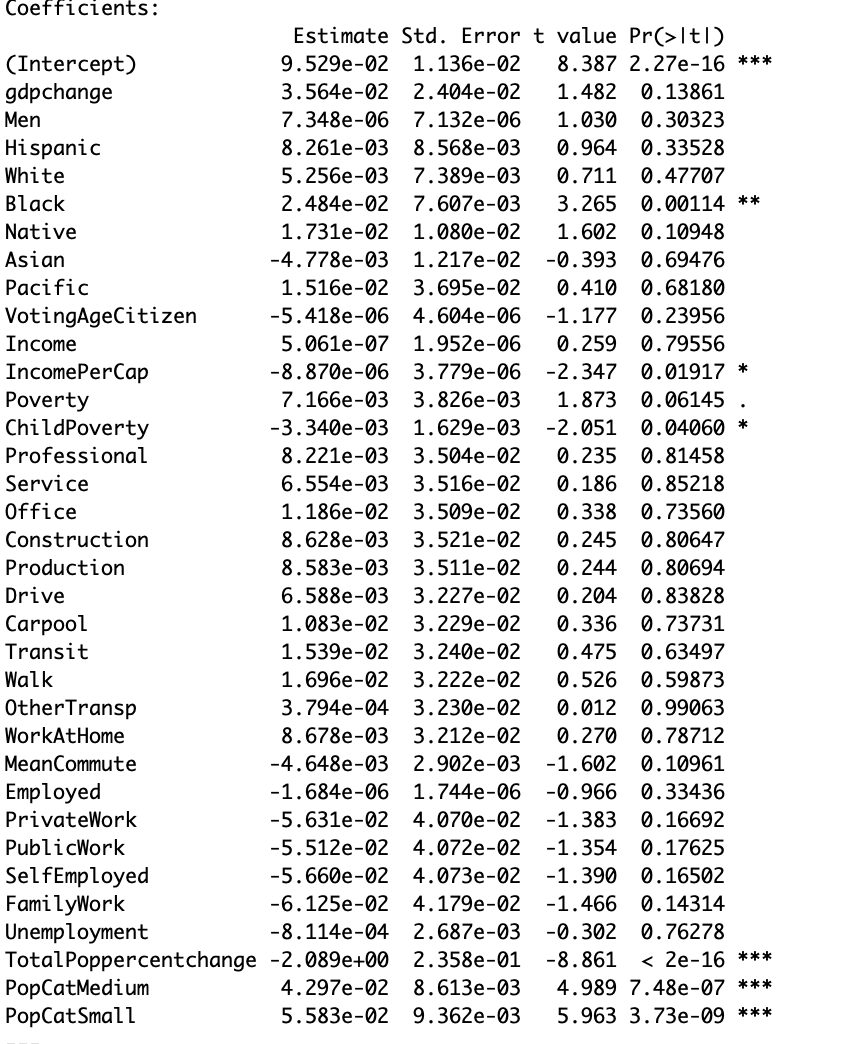


According to our regression model, Total population and employment rate play a significant role in change in votes.Total population has negative correlation and employment has positive. Now, we added the states as a categorical variable to analyze if the percent change in votes was more significant in some states than the other with same other predictive variables.

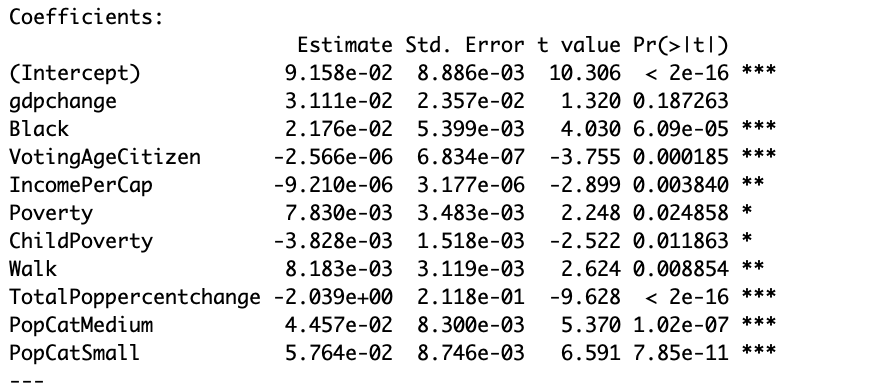


As we can see in the snippet of our regression model, the states that had more significance in change in votes were Florida, Iowa, Michigan, Minnesota, Ohio, Pennsylvania and Wisconsin. This is very interesting because most of these states voted for a republican party in 2016 election.

Next,let’s look at the regression model where the total population of the state is categorical.



As we can see in the regression model is, the total population plays a significant role in percent change no matter of the size of the population. The regression model is too big and a lot of predictors are not significant. So, we build another regression model with the demographic predictors that are significant.



In this regression model, we can see the population of the state, Voting age citizens, the population of black people plays a significant role in change in voting percentage.

*Conclusion*

The main goal of this research was to see has the change in voting percentage has been affected by the GDP change in these swing states. As we can see, the GDP in most of the swing states is negatively affected by the change in voting percent. After that, we added some demographic predictors to see if that changes the voting preferences of these swing states. As per the regression models we created, we can conclude that the size of the population, the population of black people and population of Voting age citizen plays a significant role in the change in voting percentage in these swing states.

**References**

Chappelow, Jim. “The ABC on GDP: All You Need to Know About Gross Domestic Product.” Investopedia, Investopedia, 18 Nov. 2019, <https://www.investopedia.com/terms/g/gdp.asp>.

García, Diego. Leeds School of Business. 18 Nov. 2019.