Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

CS506 Final Project Report

U.S. election & Covid-19 - Tiger Yi

I. Project Proposal

https://docs.google.com/document/d/1Zi8tL5owRF5oEoj1VYDC-qYlUodmx52hu65b1xcZUy4/edit?usp=sharing

Background/Motivation

In my previous Project Proposal, I covered the main motivation of this individual final project is to analyze how Covid-19 is impacting the result of the 2020 U.S. general election and all the national polls that lead up to 2020-11-03. This project also delves deep into the U.S. demographic data on the county level and how exactly those metrics can help us to construct a mathematical model to predict the result of the 2020 election.

The goal of this project is to try to see how some of the factors, such as Covid 19, influence the past 2020 U.S. general election. The polls have largely overestimated Democratic performance in the general election, according to multiple national news such as Washington Post: https://www.washingtonpost.com/politics/2020/12/03/was-2020-really-disaster-polling/

But why is the result of the most polls biased toward Democrats? Did Covid-19 play a major role in biasing the poll and influencing the election? Or is it more of the demographic difference between various parts of the country?

And is it possible to predict and check some of the 2020 election results (%margin difference between REP and DEM) based on other 2020 results in the training set, the previous election result from 2016, the county level demographic data, and the Covid-19 positive rate on the 3,000+ counties level?

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

Questions

The overall question will be how did the Covid-19 impact the U.S. 2020 elections in general?

Did Covid-19 contribute to the polling error/bias of the election polls?

Many news sources suggest that Covid-19 skews the polls in democrat's favor because they are more likely to answer polls during the pandemic. And can we confirm this claim by some statistical analysis?
https://www.vox.com/policy-and-politics/2020/11/10/21551766/election-polls-results-wrong-david-shor

As the positive rate rises and the number of positive cases rises, how does this trend impact the margin of the polls between DEM and REP?

What is the correlation between the polling error vs. covid positive rate? And is the correlation statistically significant enough? (Captured as R^2)

Different counties in the U.S. have been impacted by Covid differently, did the counties that got hit by Covid the most see a big shift in the election margin?

Did other traditional factors, such as the difference between gender, race & ethnicity, employment & economy contribute to the result of the election more than Covid-19?

How exactly did the Covid impact the election? How the highly Covid affected areas voted differently compared to the relatively safer areas? Are those severely affected counties overall more Democratic or Republican? Are those counties having more minority presence?

Can we use the 2016 result, covid data, and demographic data on the county level to predict and recreate the result of the 2020 election? And is Covid as a factor statistically significant enough to be included in the model to predict the 2020 election results?

Try to split up the counties into a training set and test set to predict the final margin for the 2020 general election, based on existing 2020 results in the training set, the county demographic data and the covid data, as well as the previous results from the 2016 general election. To see if we can get a reasonable model to forecast the final percent of margin for the rest of the counties in the test set.

How to answer those questions?

The first stage is to combine several of those datasets in order to answer some of the questions. Make sure to get rid of the columns containing useless information.
The data cleaning step is mostly done by Pandas. Need to pivot longer/wider and join multiple datasets for both the state and county level.

I can compare the daily state polling data against daily state Covid data to graph for each state how the Covid positive rate affects the margin of the polls (DEM-REP).

Try to compare the Covid daily positive rate against the daily poll bias/error to see if there is any correlation when performing a linear regression.

The polling error is obtained with a few steps. There is existing daily polling data from each state to show what is the percentage of voters for REP or DEM, so we can easily obtain the polling margin by subtraction. Then we can compare this daily polling margin vs. the actual 2020 election result margin to know exactly how big the polling error is overtime.

In order to compare, we need to make scatter plots for the Covid positive rate and vote margin of the polls (and 2nd graph for the polling error), then we need to run linear regression and check R^2 if the model is statistically significant. Visualization with EDA will help, but also need the model and number. The comparison can be horizontal (across time) or vertical (across states). The data I have is cross sectional data, so I need to take advantage of that.

Try to split up the counties into a training set and test set to predict the final margin for the 2020 general election, based on existing 2020 results in the training set, the county demographic data and the covid data, as well as the previous results from the 2016 general election. To see if we can get a reasonable model to forecast the final percent of margin for the rest of the counties in the test set.

Basically to see if we have a training set and test set. With a few other factors, can we reasonably predict the test set? and how well is our prediction vs. the actual result based on MSE? Or RMSE?

Also in order to know what variable to be included, it will be wise to run multi linear regression to see which variables are statistically significant enough.

Sklearn, statsmodels.api must be used extensively for statistical analysis.

Use what I learned from class such as PCA, GMM, Regression for the quantitative models.

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

II. Data Collection

Already discussed in the Project Proposal. Copy and paste down here:

1. Datasets:
All my datasets will be downloaded into my google drive that i share with you. Link:
https://drive.google.com/drive/folders/19lNytOmjVmR7TMJJ2mieXGBSDe3QS3hI?usp=sharing

2. Source:

Election polls provided by FiveThirtyEight (2 datasets from here)
https://github.com/fivethirtyeight/data/tree/master/election-forecasts-2020
https://projects.fivethirtyeight.com/2020-general-data/presidential_polls_2020.csv
https://projects.fivethirtyeight.com/2020-general-data/presidential_poll_averages_2020.csv

General election results by Cook Politics as CSV on state level:
https://cookpolitical.com/2020-national-popular-vote-tracker

General election results from Kragle on the county level
https://www.kaggle.com/unanimad/us-election-2020?select=president_county_candidate.csv

Covid datasets from Kaggle:
https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_states_covid19_daily.csv

Covid data from Kaggle, JHU tracker
https://www.kaggle.com/headsortails/covid19-us-county-jhu-data-demographics?select=covid_us_county.csv

U.S. Demographic data by county:
https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv
https://www.kaggle.com/etsc9287/2020-general-election-polls?select=county_statistics.csv

Election, Covid, Demographic by County (this dataset has a lot of information columns)
https://www.kaggle.com/etsc9287/2020-general-election-polls?select=county_statistics.csv

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

III. Data Preparation

data_clean_up.ipynb

The Jupyter Notebook “data_clean_up.ipynb” is the main code for my data processing and data preparation.

The main technique of this section is predominantly using pandas package to try to parse out the useless information in many dataset and combine the dataset in a way that will be helpful in the data analysis section later.

Before going too deep into the data processing, allow me to present the result of this section.

Input datasets:

2017_demographic_county_data.csv	U.S. demographic data on the county level
covid_us_county.csv	U.S. Covid daily data on the county level
covid_us_states.csv	U.S. Covid daily data on the state level
presidential_poll_averages_2020.csv	U.S. election national polls daily average
popular_vote_by_states.csv	U.S. election result by state
president_by_county.csv	U.S. election result by state
county_statistics.csv	U.S. county economic and demographic data
usa_states_latitude_and_longitude.csv	U.S. state & county latitude and longitude data

Output datasets:

poll_tigeryi.csv	U.S. election polls on state level
election_tigeryi.csv	U.S. election results on the state level
covid_tigeryi.csv	U.S. covid data on the county level
poll_covid_tigeryi.csv	U.S. election poll and covid data on the state level
demographic_tigeryi.csv	U.S. demographic, election results on the county level
df_tigeryi.csv	More clean version for the demographic_tigeryi.csv

As discussed in the Project Proposal, this project is looking for data in the 4 areas:

U.S. election polls, U.S. election results, U.S. covid-19, and U.S. demographics

The entire process of the data cleaning isn't very organized in the code because I always come back to this section when I try to do visualization and analysis. You would never know what to clean and how to clean in the first try. Pandas module is huge in the section.

I dropped a lot of NA values, sometimes the entire row with it, as I weigh pros and cons. I also merged across a lot of datasets in order to have as many metrics as possible in the end, as I never know what metrics will become useful in later.

I also did lots of row operations and created a lot more new variables on demand, especially for the delta changes of the variable. For Covid in particular, data is cumulative, positive cases and deaths are rising daily. But what about day to day changes? What about the rolling average to smooth out the abnormal data point? What about calculating the proportion of the variable, as different states and counties have vastly different populations?

It actually took a long time for me to process the datasets to the point of my satisfaction. A good dataset I look forward to is having as many column entries as possible to capture information, but at the same time, as fewer empty cells/NANs as possible to avoid conflicts in the data analysis later.

I also incur some new problems when I clean the dataset and I have to undo what goes wrong. It is very annoying to take care of NANs and INF in the dataset. It is also not easy sometimes to convert the datetime properly. Be very careful when doing division as you might divide numbers by 0 to generate INF, which will completely mess up the graph visualization in the next section.

IV. Data Analysis & Visualization

exploratory_data_analysis.ipynb

The Jupyter Notebook “exploratory_data_analysis.ipynb” is the main code for my data analysis, model construction, and data visualization.

The exact code itself won’t be repeated as it is in the code itself. This section will primarily focus on how I come up with the way to approach the various questions. What data science knowledge and skills I’ve applied. And how do we interpret the graphs and equations in order to help us to answer the questions in the project proposal. Try to walk the readers through what I did, why I did, and how I did for this final project.

The first step is to import all the processed .csv files into exploratory_data_analysis.ipynb

Then I try to address the question of how does Covid-19 impact the U.S. election polls and the 2020 election result. But before that, we need to explore the relationship between the election polls and the election result itself in many states. Many sources argue that the election polls were heavily biased towards the Democrats compared to the final results. To test this claim, I will first use graphs for data visualization.

The methods “poll_state” basically will take input from the daily state level election polls and the final general election result data and create 2 graphs. The graph on the left shows the Democrat margin in the polls. The graph on the right shows the polling error based on the final election result.

The way to calculate the polling error is fairly trivial: first, to calculate the democrat’s margin in the polls (DEM % poll - REP % poll), then we calculate the democrat’s margin in the actual election results (DEM election result % - REP election result %). Finally, we subtract the election result margin from the election poll margin to get the poll error.

Poll error % = (DEM election result % - REP election result %) - (DEM % poll - REP % poll)

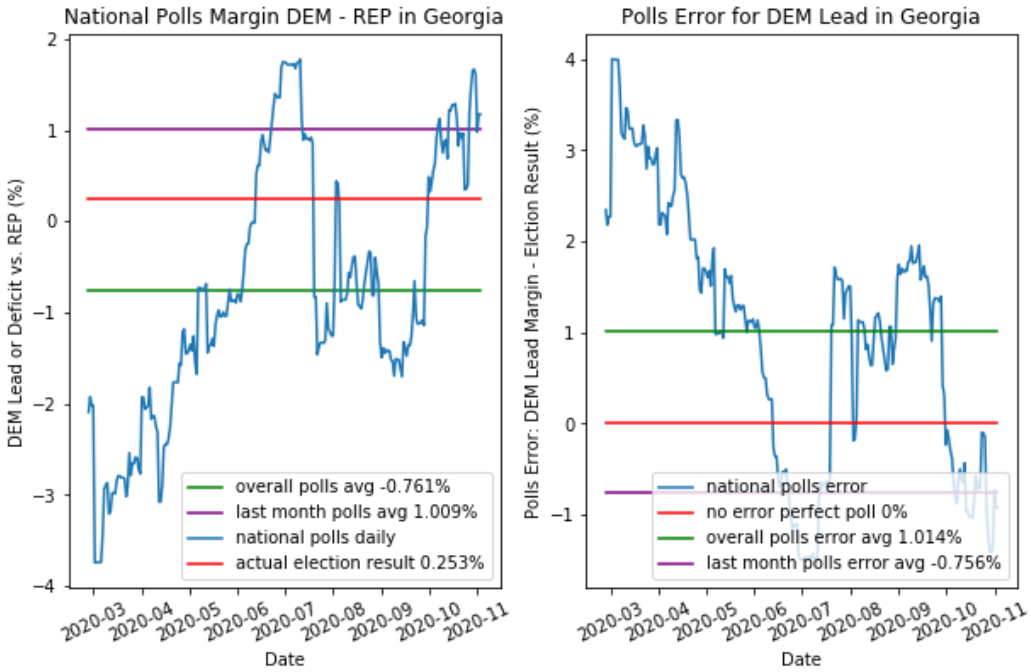
If Poll error % is positive, then it shows that the polls have underestimated Joe Biden’s support, or Joe Biden did better in the general election than the polls have projected.

Similarly, if Poll error % is negative, then it shows that the polls have underestimated Donald Trump ’s support, and Trump did better in the general election than the polls have projected.

The election poll data is cross sectional. It has the horizontal dimension of datetime, it also has vertical dimension across the 50 states. To understand the election polls dynamic, it is useful to analyze from both directions.

National polls data is the daily average multiple polls weighted by their quality score.

Here are the graphs for a couple battleground states in this 2020 election cycle.

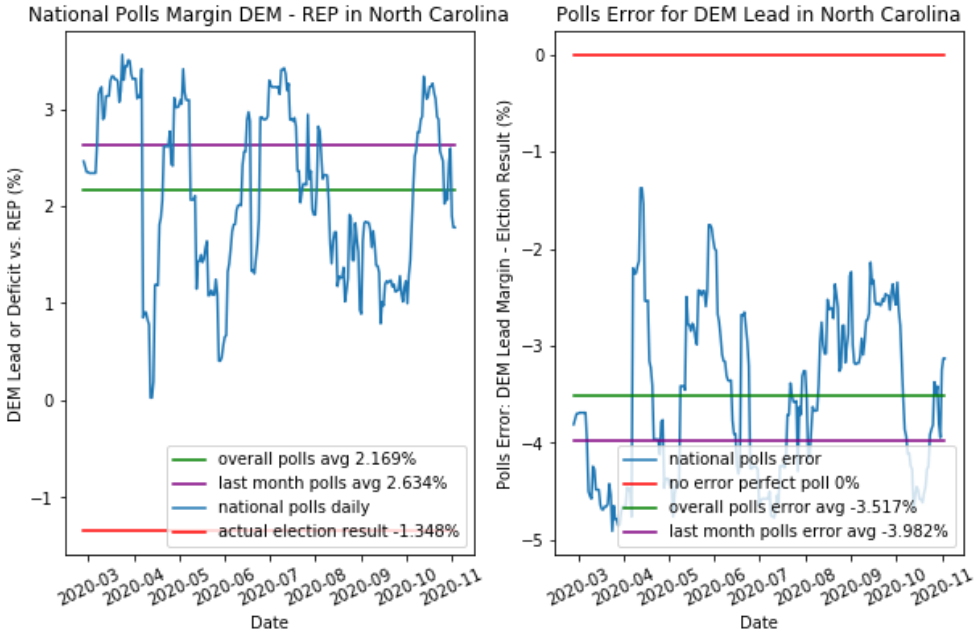


The red line on the left shows the actual election result, in this case for Georgia. Joe Biden narrowly beats Trump in Georgia by 0.253%. The national polls (the blue curve) here is fairly accurate. Overall, the polls forecast Biden would've lost Georgia by 0.761%, and for the last month prior to the election date 2020-11-03, the polls forecast Biden would've won by 1.0%.

For the polling error graph on the right, ideally, you would want the blue curve to stay as close to the red line at 0 as possible in order for the polls to be accurate. For Georgia, in the early days, national polls have underestimated Biden's support by 3-4%. And as the time gets closer to the election day, national polls start to slightly favor Biden in a tight race. And in the very last days, polls only overestimated Biden's lead by 0.7% roughly (-0.756% on the graph), which is very accurate compared to the final result.

But unfortunately, for the rest of the battleground states, the national polls did horribly at forecasting the election result. More specifically, national polls have all overestimated Joe Biden's support by a huge margin. Here are some examples:

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

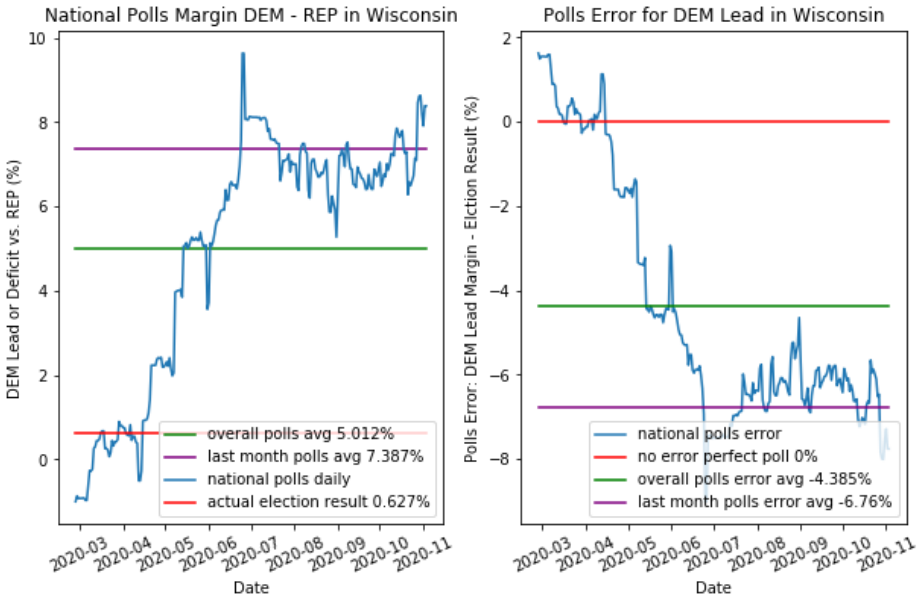


In North Carolina, all national polls project that Joe Biden would win the state by over 2% the entire time. The final result is, in fact, that Donald Trump won the state by 1.35%.

Therefore, the national polls for North Carolina have overestimated Joe Biden's election margin by 3.5%-4%.

It is quite interesting that unlike Georgia, the polls in North Carolina didn't adjust better as time approaches election day. The national polls in N.C. were off the whole time!

There must be some factors underneath the surface, which constantly biased the national polls in Joe Biden's favor. Whatever the factor(s) might be, it must've behaved quite differently between Georgia and North Carolina. Because the poll in Georgia has corrected itself overtime, but the polling error in North Carolina remains almost at the same level from March to November 2020.



But the polls in Wisconsin are even more inaccurate!

Joe Biden narrowly won Wisconsin by just 0.627% in the election. But the polls overall projects he would've won by at least 5%. And especially when it's close to the election day, polls suggest that Biden's lead would've been even wider, at 7.38% over Trump.

Unfortunately, for the polls, as time gets closer to the election, the polls become worse and worse. The polls in the last month overestimated Biden's lead by over 6.76%. What thought to be a blowout win for Biden, ended up a razor thin tossup in Wisconsin.

There are more graphic examples in the code itself but I will leave those out of the discussion.

It's a very consistent trend that the national polls are not only off by a lot, but also always overestimate Joe Biden's margin compared to the actual election result, which is statistically impossible to be just a coincidence.

There must be some underlying factors which biased the national polls in Biden's favor. And by exploring those factors, would be hugely beneficial in the future predictions.

Before moving on to the next section, I need to bring your attention to the different trends of polling error across those states. As the time gets closer to 11-03, polls in Georgia are becoming more and more accurate as the race there tightens up. However, the polls in North Carolina, even though wrong, remain largely at the same level. The polls for Wisconsin become more and more unreliable and inaccurate as the time approaches the election.

The dichotomy between those 3 trends must come from the same underlying conditions which all bias the polls in the same direction, but at the same time, different from states to states.

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

The main goal of my final project is to try to find why the national polls are off by quite a significant margin, why the polls are more likely to bias toward the Democrats, and why the polls in some states become more accurate, whereas the polls in other states worsen.

The key to the puzzle is Covid-19 pandemic.

It is very reasonable to attribute Covid-19 to the election poll errors and the shift in election result fundamentally. Covid could explain why many national polls across almost every state have overestimated Biden's lead over Trump: at the height of the Covid pandemic, Democrats are more energized politically compared to Republicans, so it is more likely for Democrats to answer the phone calls for polls than Republicans. Covid biases how the polls sample the likely votes for the election.

On the issue of Covid, Republicans tend to have less trust in the poll questions than Democrats, which further skew the polls in the Democrat's favor in 2020. Republicans might not answer the polls to the extent what Democrats do; however, those Republicans are still going to vote on election day, which is why Joe Biden's victory in the actual election result is much tighter than the polls suggested.

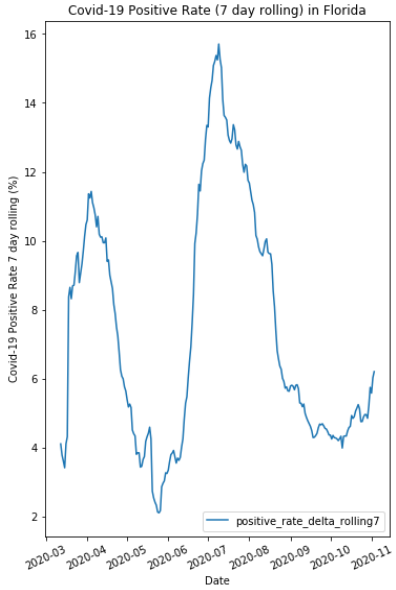
Furthermore, as to explain how the polling errors vary across different states, many states suffer from various degrees of Covid-19 pandemic at different months. The higher the Covid-19 positive rate is, the more the national polls overestimates Joe Biden's lead over President Trump.

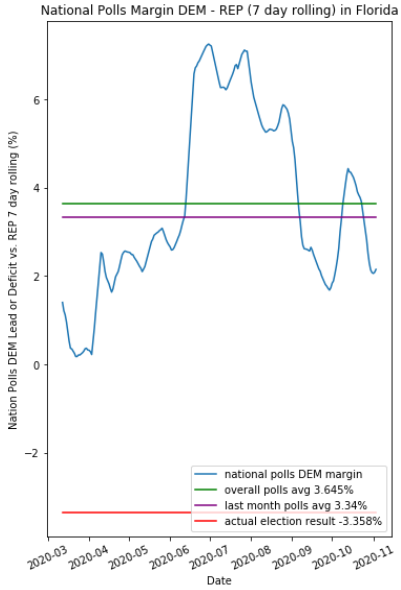
In order to verify my theory of Covid-19 impact on the election polls and result, I expand my graph visualization in the code, and add in the weighted least square linear regression for quantitative data analysis.

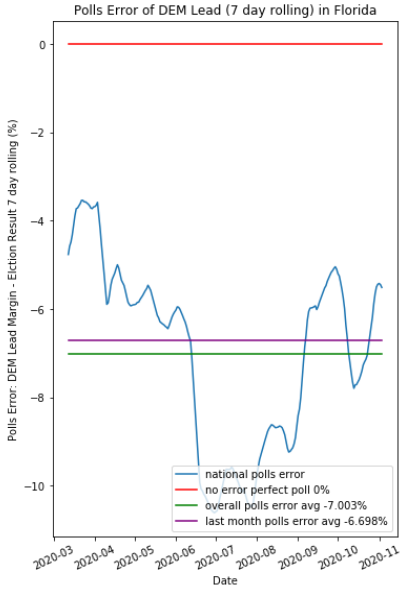
Our first example in this section will be the state of Florida, which demonstrates a strong correlation between Covid-19 positive rate and the national poll misses.

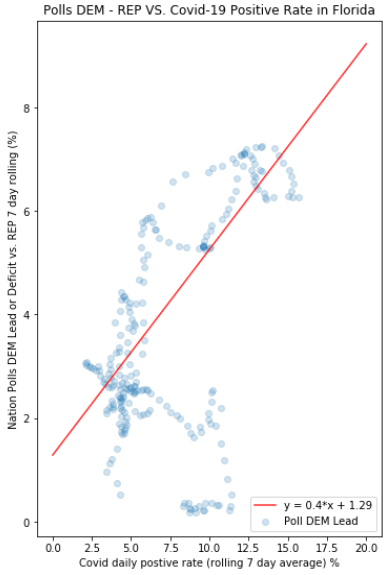
The 1st in the next page shows how the Covid-19 is developed in Florida, as you can see there are two major spikes of Covid cases: one in late March to April, another bigger spike in July 2020. What is very interesting to note is that, in the 2nd and 3rd graph from previously, there are also the upticks of democrat's lead margin in the polls and polling error at the similar time during April and July. And once we perform linear regression on the polling error and DEM lead on the Covid-19 positive rate, it shows adj $R^2 = 0.576$ and $P > |t| = 0.00$ for both the intercept and the dependent variables, which means the relationship between Covid and election polls are statistically significant and positively correlated at 95% confidence level.

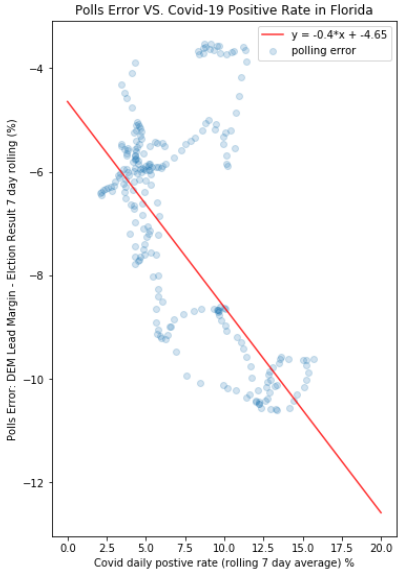
Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti











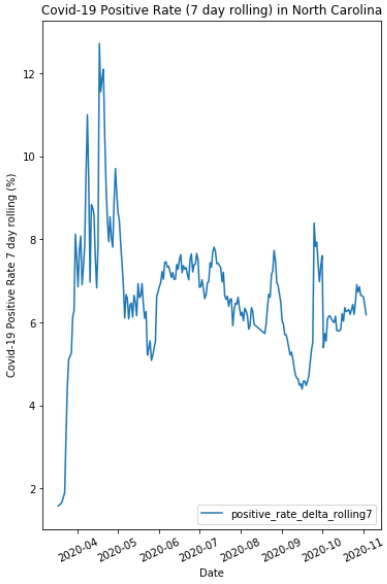
WLS Regression Results

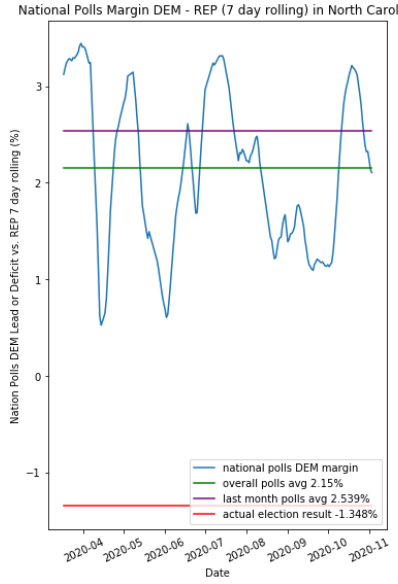
Dep. Variable:	dem_lead_rolling_7_diff	R-squared:	0.577
Model:	WLS	Adj. R-squared:	0.576
Method:	Least Squares	F-statistic:	321.1
Date:	Sat, 12 Dec 2020	Prob (F-statistic):	7.57e-46
Time:	15:30:31	Log-Likelihood:	-inf
No. Observations:	237	AIC:	inf
Df Residuals:	235	BIC:	inf
Df Model:	1		
Covariance Type:	nonrobust		

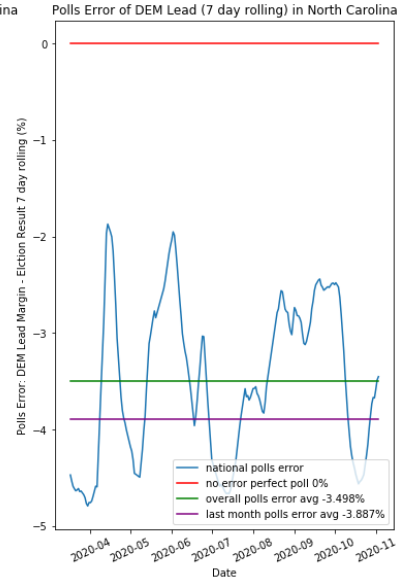
	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.6460	0.174	-26.747	0.000	-4.988	-4.304
positive_rate_delta_rolling7	-0.3970	0.022	-17.919	0.000	-0.441	-0.353

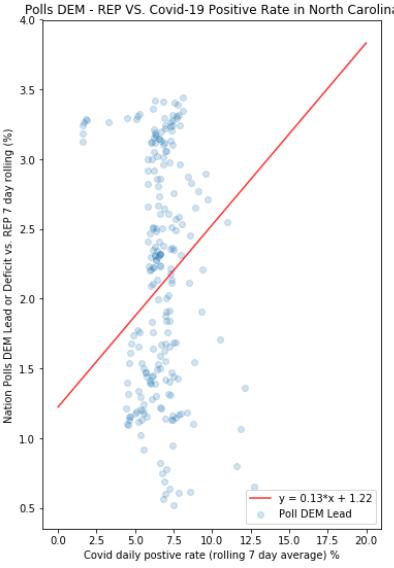
Omnibus:	6.409	Durbin-Watson:	0.025
Prob(Omnibus):	0.041	Jarque-Bera (JB):	4.602
Skew:	-0.210	Prob(JB):	0.100
Kurtosis:	2.461	Cond. No.	18.0

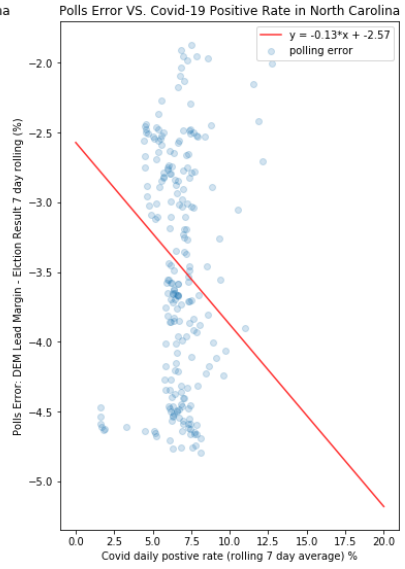
Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti











WLS Regression Results

Dep. Variable:	dem_lead_rolling_7_diff	R-squared:	0.032
Model:	WLS	Adj. R-squared:	0.028
Method:	Least Squares	F-statistic:	7.444
Date:	Sat, 12 Dec 2020	Prob (F-statistic):	0.00687
Time:	15:30:45	Log-Likelihood:	-inf
No. Observations:	225	AIC:	inf
Df Residuals:	223	BIC:	inf
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.5721	0.313	-8.227	0.000	-3.188	-1.956
positive_rate_delta_rolling7	-0.1304	0.048	-2.728	0.007	-0.225	-0.036

Omnibus:	15.067	Durbin-Watson:	0.034
Prob(Omnibus):	0.001	Jarque-Bera (JB):	6.072
Skew:	-0.090	Prob(JB):	0.0480
Kurtosis:	2.215	Cond. No.	42.2

Going back to North Carolina, as discussed before, the national polls were “constantly wrong” at the same level for the entire year. Even as the time approaches election day, national polls overestimate Democrat’s margin by 3-4%, the same level of the mistake as back in April.

The bottom 2 graphs and the regression line can explain this almost perfectly. For North Carolina, except for the short period of Covid positive rate spike in April, its Covid positive rate has been very steady and consistent at 5-8%. The slope for the best fit line of the polling error is at -0.13, which is a lot flatter and less steep than the -0.4% slope in Florida.

And considering the P>t value is still extremely low and 0 not in the 95% confidence interval, from the math standpoint, it shows that the relationship between Covid and election polls are still statistically significant.

For North Carolina, the slope of -0.13 means that even if the Covid 19 positive rate rises fast, the overestimation of DEM lead in the polls and the polling error are still very minimal. This explains why the DEM polling lead and error remains largely constant throughout the year: it is because the Covid-19 daily positive rate has controlled and flat all the way until the election.

The interpretation of this slope is this: In North Carolina, if the Covid-19 positive rate will rise by 1%, then the expected polling error and overestimation of DEM lead margin will increase by 0.13%, as compared to the 0.4% in Florida.

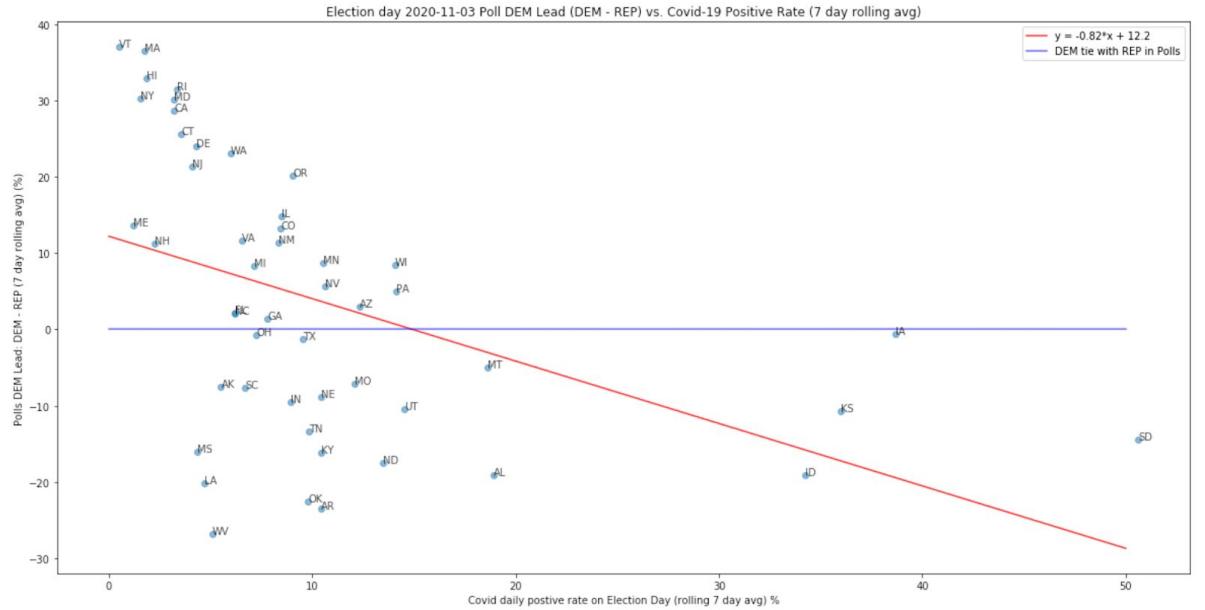
The result of this discovery is truly profound. There are more examples and graphs in the ipython notebook as well. This demonstrates, not only graphically and qualitatively, but also quantitatively how exactly the Covid-19 correlates with the democrat’s margin and its error in the national election polls:

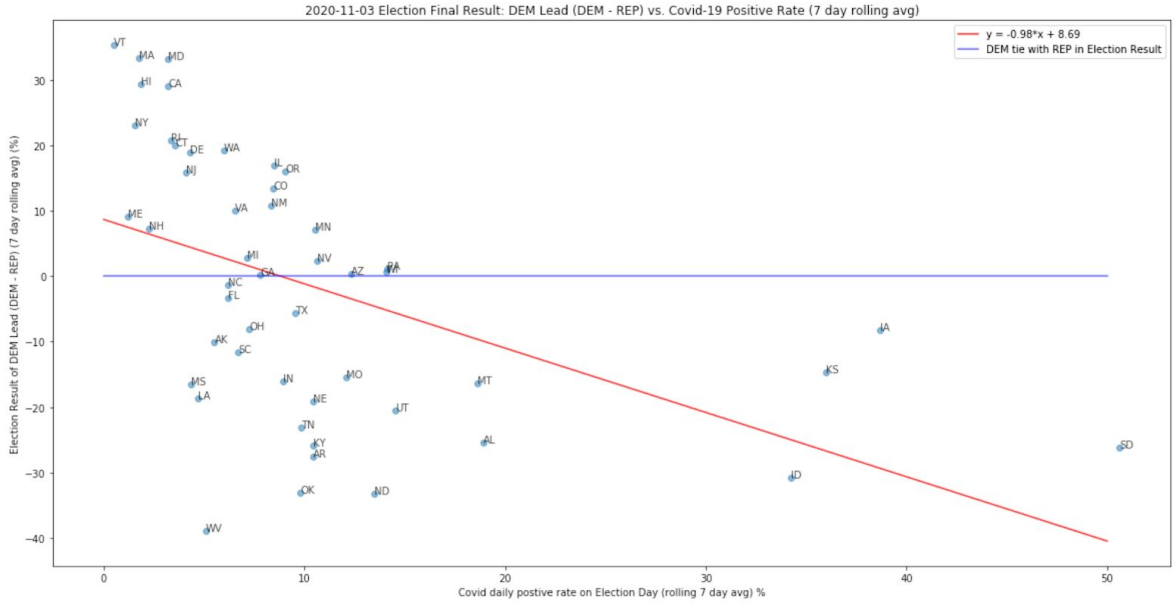
In most cases, Covid-19 itself biases the polls in the DEM favor, as we suspect because Democrats are more willing to answer the poll during Covid-19 pandemic compared to Republicans. In addition, the more severe Covid pandemic gets, the more margin Democrats have in the polls, and more likely the polls will be wrong and overestimate the Democrat’s performance in the general election.

As we are looking more on the horizontal level of the data: on the daily basis. It is also useful to analyze the data vertically: across 50 states. For the graphs on the next few pages, we will look at how states differ from each other during the Covid-19 pandemic and how the polls perform in those states.

On the next page, Covid positive rate is the 7-day rolling average on the election day for 50 states.The first diagram is election poll vs. covid, and the second for election result vs. covid.

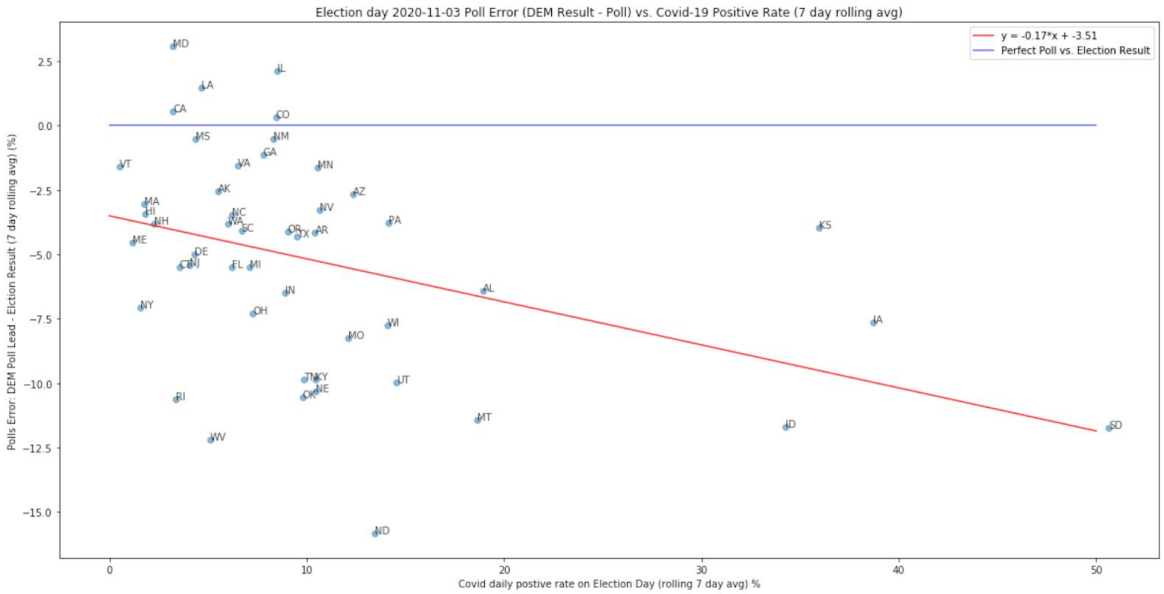
Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti





One of the key takeaways from these 2 graphs is that, for the states where Democrats (dots above the blue lines) has the lead in the polls (1st graph) or DEM wins the election (2nd graph), those states tend to have much lower Covid-19 positive rate compared to the states won by the Republicans. Especially for those states where REP won in the midwest, such as South Dakota, Kansas, Iowa, the Covid pandemic is rampant as the covid positive rate rises above 30% during the final 7 days leading up to the general election. Both linear regressions are in the python code to show the statistical significance.

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti



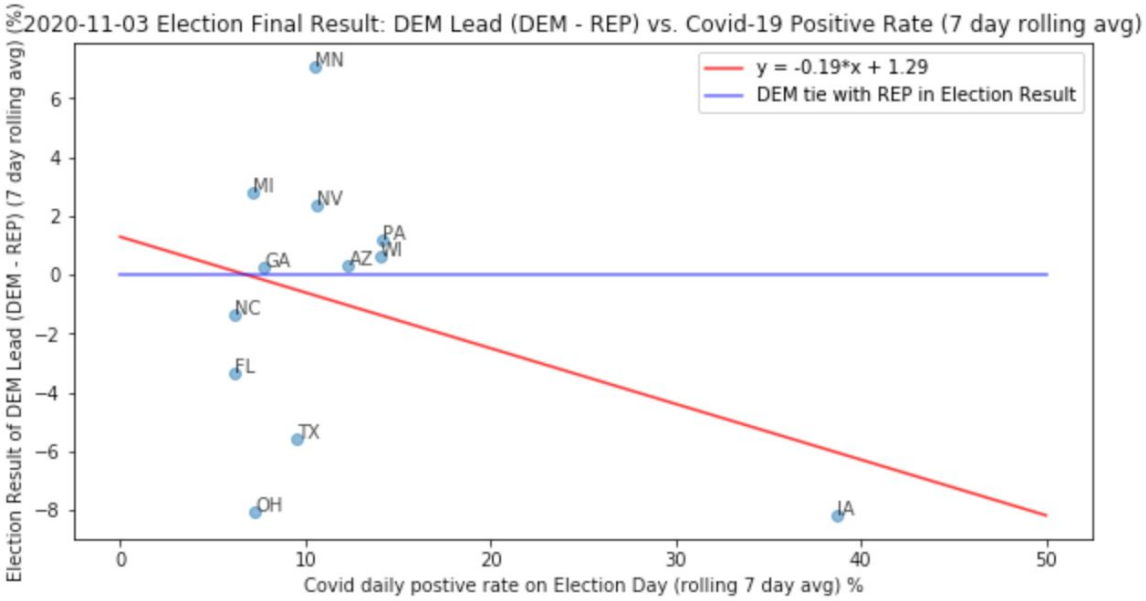
This 3rd graph shows that the national polling error has overestimated the Democrat's performance in the general election in the vast majority of states.

There are only 5 states above the blue line where the national polls have underestimated Joe Biden's performance. For the rest 40+ states, the polls all have overestimated Joe Biden's performance in the general election.

And it is true that, for a given state having a high Covid-19 daily positive rate (x-axis to the right), national polls are more likely to make bigger errors and overestimate Democrat's margin in the general election (y-axis to the bottom). And this information is captured in the download sloping red line of the OLS linear regression model.

However, before moving on to the next big topic and leave the polling behind, I also find an interesting take away that I want to highlight in this project. The above 3 graphs are on the national level, for all the 50+ states. But, if we would zoom in and only focus on the battleground states, the same graphs would tell us a very different story which will help me to reach a new educated speculation.

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti



I have discussed how the covid-19 is related to the election result on the national level 2 pages ago. Now let's focus on only the battleground states.

When looking on the national level, we have found that the states Democrats won tend to have lower levels of Covid positive rate than those Republican states.

But, when we zoom in on only the battleground states, it shows that Trump won states with relatively lower daily positive rates of Covid-19, such as Ohio, Florida, North Carolina etc.

Joe Biden, on the other hand, won the battleground states that have a slightly higher level of Covid positive rate, such as Arizona, Wisconsin, and Pennsylvania.

This is a total reversal of the trend on the national level. So what is going on here? It should be the case that those states won by republican have fewer travel restrictions, and therefore, higher covid positive rate than the Democratic blue states. But clearly, this does not apply for most of the battleground states, with the exception of Iowa on the bottom right. If we would remove Iowa as a battleground state, the red line will flip to have a positive slope.

To explain this conflict, I want to argue that those battleground states flip from Trump 2016 to Biden 2020, partially because of the high positive rate Covid-19. The mishandling of Covid by President Trump might very well damage his chance of reelection.

To test my claim, I want to use county level data on U.S. demographic, election, and Covid-19 to see which factors impact and how they contribute to the final result of the 2020 election.

So far, we have analyzed very thoroughly on why the national polls have overestimated the Democratic performance in the 2020 general election and how exactly the Covid-19 pandemic contributed to the error of the national polls on the state level, with lots of graph visualizations and statistical regression for quantitative analysis.

Now in the next big chapter, I would like to explore the election result beyond just Covid-19. I have cleaned a few dataset “demographic_tigeryi.csv” and “df_tigeryi.csv” for the data analysis of this next section. I will start to explore the more detailed data on the microscopic county level, as opposed to the state level earlier. Also, this new section will include a lot more new kinds of data beyond just Covid-19 as the factor, such as: race, ethnicity, income, job, gender, 2016 election, and 2020 election on the county level. And without a further due, let’s get started.

There are many fields in these datasets and I have processed the datasets so that the data frames contain only the integer or float values, so it will be a lot easier later to construct machine learning models to predict the 2020 general election result. I have to make sure that non-float columns such as “state” and “columns” are converted into data frame indices. I also drop out all the INF and NAN values, in case the outliers will blow up my regression models.

This is definitely the more challenging and exciting 2nd half of the data analysis and visualization. I will walk you through what I did and how I made the code work.

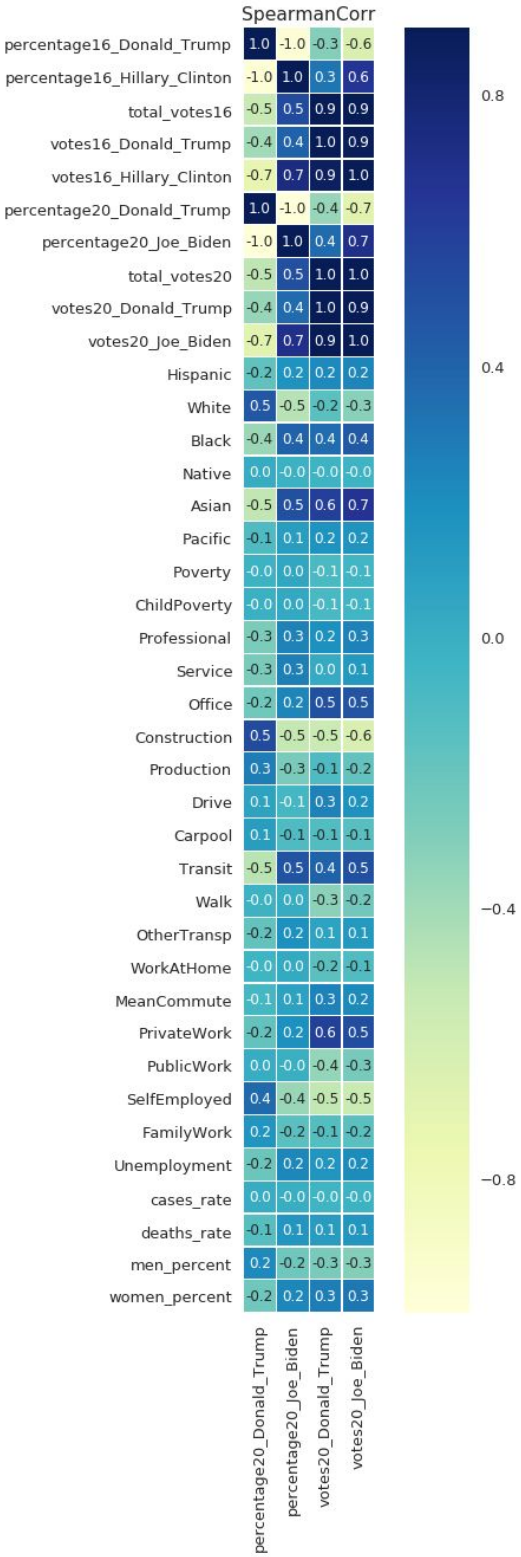
First, I identify the columns that will be my “dependent variable y” when split the dataset into training set and a test set. I make a decision, at first, to choose 4 columns as the attributes I want to create the model on.

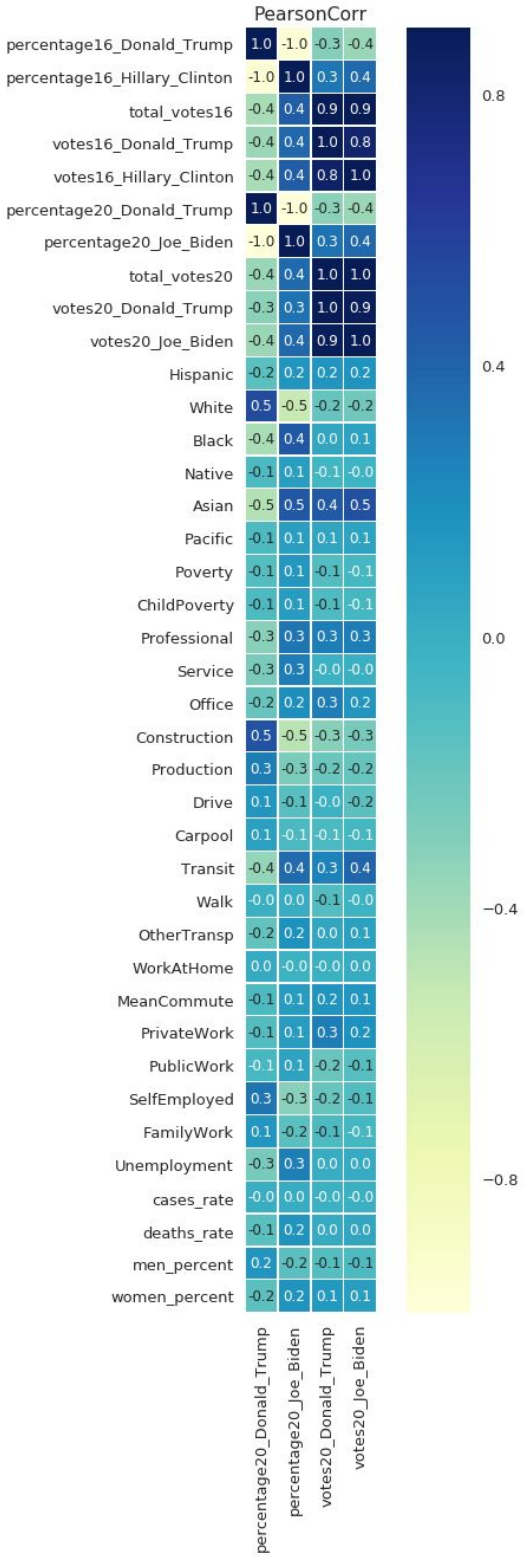
I select “percentage20_Donald_Trump”, “percentage20_Joe_Biden”, “votes20_Donald_Trump”, and “votes20_Joe_Biden” as the dependent variables.

Basically, the goal of this section is to train a regression model to make a prediction on those columns, given other independent variables such as demographic, Covid-19, 2016 election.

The first thing I’ve done is to construct a correlation matrix between all the columns of variables I decide to be included in my regression model. One for “pearson” correlation, another one for “spearman” correlation. Those are 2 different ways to calculate the correlation of the data frame. Most people taking statistics are more familiar with “pearson” coefficients which is calculated as the standard deviation of x times standard deviation of y, and divided by the covariance of both x and y. But “spearman” is a more complicated method to calculate the correlation and it performs slightly better when the data is large. I will present the result as a seaborn heatmap, and I will elaborate on how to interpret the correlation later.

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti





The seaborn heatmap in the previous page is a really clean way to demonstrate how each dependent variable (y) is dependent on the many independent variables (x) and how each independent variable (x) influences the dependent variable (y).

1.0 means a completely positive correlation between the variables, -1.0 means a completely negative correlation between the variables; and all other values between -1 and 1 will show the dependency for various degrees.

If we just look at the “percentage20_Donald_Trump”, which denotes the percent% of votes Trump won in that particular county, we will see this variable has the 100% positive correlation with 2016 Trump election percentage and 100% negative correlation with 2016 Clinton percentage when he faced Hillary Clinton.

It’s going to become obvious that for our data science model for prediction, the prior 2016 election percentage result between Trump and Clinton will play the dominant role in the 2020 election between Trump and Biden. This is expected as the previous election results are the benchmark for the 2020 election. It is very rare that the county on average would see a huge margin shift of party affiliation, from DEM to REP or REP to DEM. Therefore, I will keep the 2016 election results for my models and the previous results will play the biggest part in predicting 2020 election results.

What is also important are those demographic, covid, and economic variables. “Case_rate” denotes the cumulative Covid-19 positive rate on the county level. This is different from the positive rate I have been using before in previous sections. And since here I am not interested in date time horizontal analysis, I only include the covid positive rate on 2020-11-03 for all the counties in the U.S. I also include “death_rate” which captures how many people died from Covid-19 pandemic per the entire population of that county. “Women_percent” and “men_percent” capture the percentage of women and men respectively for genders.

Spearman tends to be the more accurate table so I will only discuss the one on the left.

As you can see, the correlation between Trump 2020 vote percentage and White people is 0.5%, which is a very significant number. This means that for every additional 1% of white people in a given county, Trump is likely to get 0.5% more votes in the election than Joe Biden. Biden, on the other hand, has more popularity with the minority: 0.4 with Black, 0.5 with Asian, and 0.2 with Latinos. The great thing about this heat map is to show the readers how exactly each race contributes to the 2020 election with numbers.

Everyone knows Republicans are more popular among White and Democrats among minorities. But without the correlation matrix, you can’t speak anything analytically.

In addition, for jobs and economics, Trump tends to be very popular among construction, production, self-employed workers, at 0.5, 0.3, and 0.4 respectfully. Joe Biden, on the other hand, has more support in transit, professional, and service jobs, at 0.5, 0.3, 0.2 respectively.

This is something new to me at least. I did not know how career occupation has an impact on political affiliation, and it is very cool to learn this result quantitatively, with the various magnitudes from each factor.

This table also shows that Trump has more support among men. The correlation with the men percentage is 0.2 for Trump. So if the county has more men than women, on average, Trump is more likely to beat Biden there. The opposite is true for Biden among women, coefficient is 0.2 for Biden among women. So if the county has 1% more women than men, controlling all other factors, Biden is expected to beat Trump by 0.2% there.

Last but not the least is Covid-19 pandemic. Here, we reaffirm my assumption that Covid-19 pandemic does hurt Trump in the election based on stats and numbers. The coefficient of death rate is -0.1, which means if the county has 1% higher death rate, Trump is to expect to lose to Biden on this front by 0.1%.

Unemployment is another factor which hurts Trump in the election, coefficient is at -0.2. Therefore, if a county is experiencing a high rate of unemployment, the chance is Trump is going to lose support to Joe Biden.

Now, I understand how each component and factor is working for or against in Trump's favor.

It is time to move on to the model construction.

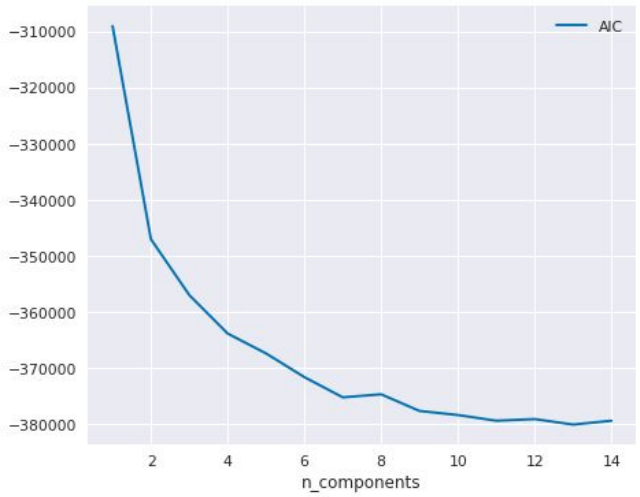
But before anything, it is very important to note that for all those fields / independent variables, the scale of them are quite different.

In order to make a more precious model, I must Min Max Scale the entire dataset from its original scale to standardized scale from -1 to 1 for all variables.

This return of benefit will come as a cost as well. For the scaled dataset, it is very difficult to interpret the result later, as everything is in the scale of -1 to 1, and it is extremely difficult to rescale 30-40 variables back.

Therefore, what I decided is to use both a scaled and an unscaled model. For the analysis of model coefficients during regression, I will use the scaled version because the model is more precious. For the interpretation of the number itself, like the vote percentage for Trump, I will use the unscaled model.

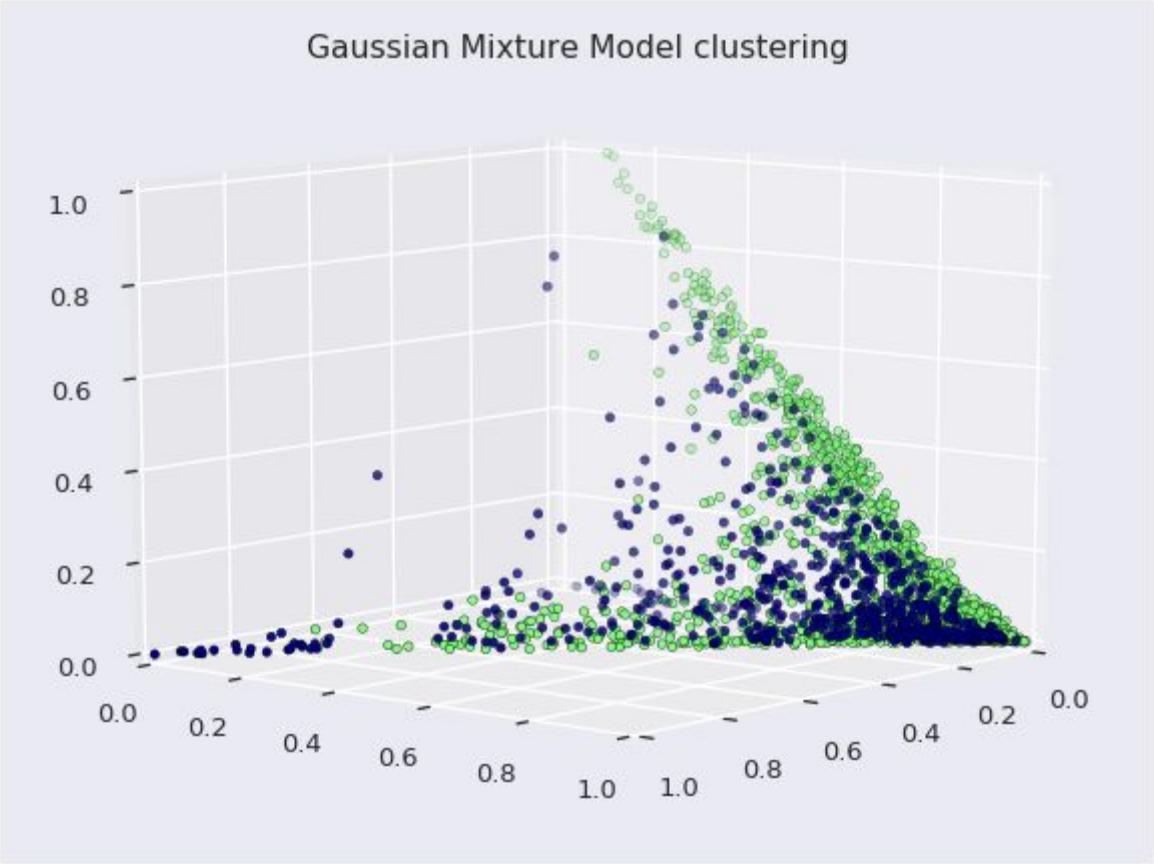
I will drop the variables like votes total, DEM votes in 2016 etc. and reduce the dataset for Gaussian mixture clustering model (GMM) on the min max scaled data to see how exactly the clusters are distributed.



I used GaussianMixture from sklearn package from CS506 to determine that for this model, it is best to have 2 clusters for GMM. And it makes sense because after all, we should only need to use 2 clusters to capture the difference between DEM and REP.

To visualize the cluster for the scaled data, I plot a 3D graph to understand better visually.

As you might see from the graph below, there are clearly 2 clusters of data from DEM and REP each. Then, I will attach the cluster IDs (in this case, only 0 and 1) to the scaled data in order for better model performance later in the next few steps.

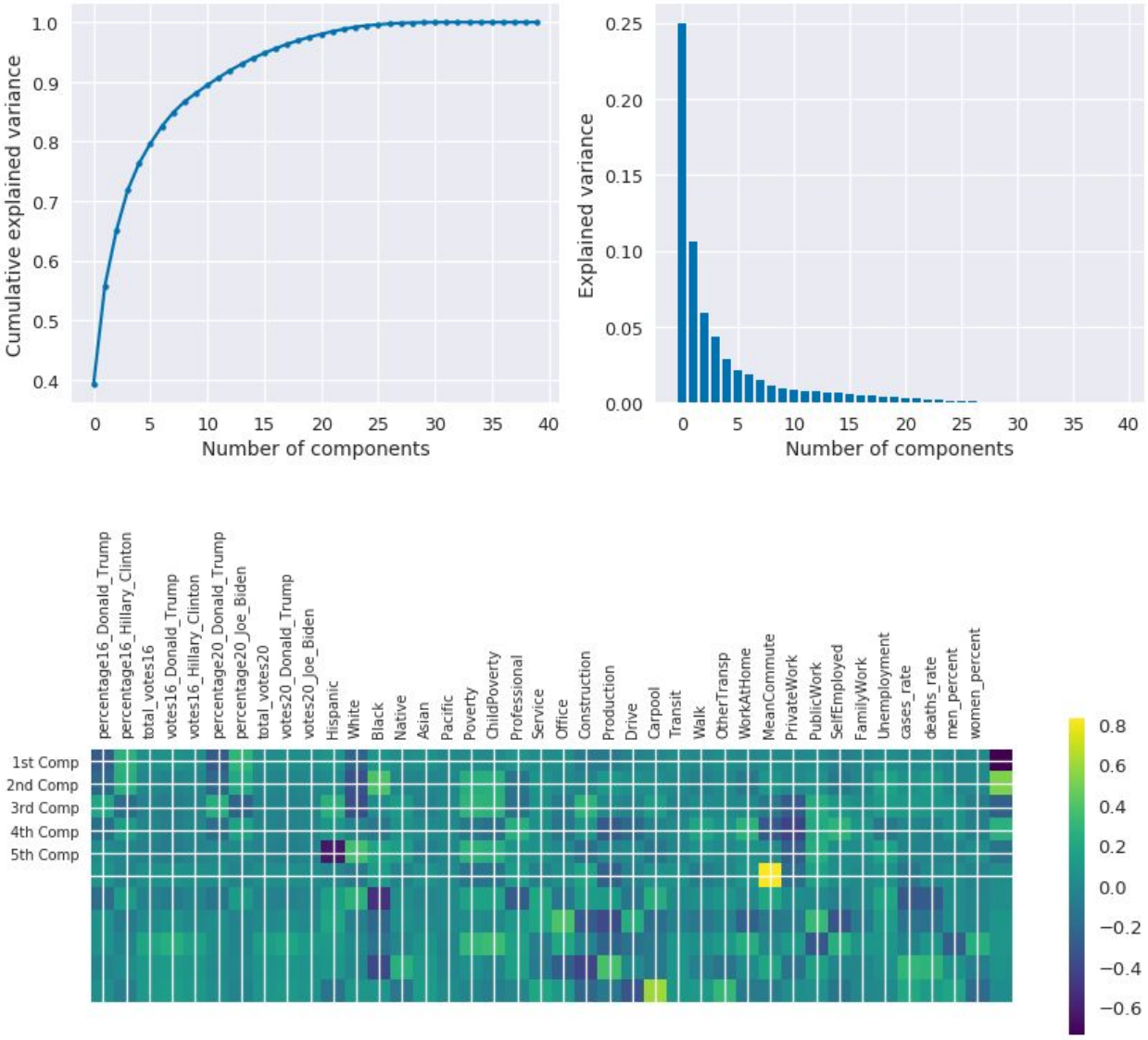


The next step for Data Science analysis is the Principle Component Analysis (PCA). The sklearn package can help me to identify the various degrees of the principle components which can help to explain the variance, and isolate out the correlation and interdependence for each variable. The linear transformation and single out individual component will be helpful to decompose the multi-dimensional space visually.

Like the GMM clustering before, we need to determine the number of principal components that can explain as much variance as possible with the minimum number of components for accuracy of the data.

As the graph shown below, the number of components really have a diminishing return on how much the variance that can be explained in the data. 90% of the variance will be a reasonable cut off, and this implies to use about 11 principle components in the PCA.

Now we will just fit PCA(11) on our data and again generate a seaborn heatmap to view.



Again, as said before, due to how the past 2016 election results shape up the current 2020 election, it is obvious that in the 1st component, 2016 and 2020 Trump vote percent has a distinctive negative contribution to Joe Biden, as well as “White” people in the first several components. Those negative contributions are very dark blue.

In contrast, 2016 Hillary Clinton percentage, “Black” people, issue of poverty and unemployment, all seem to work in Joe Biden’s favor. Those are much lighter colors compared to the shaded ones for Trump. It is also curious to note for commuters in the 6th component plays a huge role in Biden's victory.

Many factors, men vs. women, didn't differentiate until the very last components, which means that their contribution swayed very little in the election margin. It could be the men to woman ratio in most counties are fairly balanced, and the difference isn't as big as race.

For PCA, I also try to print out the factors that contribute to Biden the most:
Again, all the contributions have been explained in details and are very much expected.

	Name	Max absolute contribution
31	PrivateWork	0.828690
26	Transit	0.605653
22	Construction	0.393574
24	Drive	0.373126
13	Black	0.369373
19	Professional	0.341196
33	SelfEmployed	0.317437
29	WorkAtHome	0.304277
37	deaths_rate	0.302476
18	ChildPoverty	0.302208
8	total_votes20	0.294081
38	men_percent	0.279977
3	total_votes16	0.278326
30	MeanCommute	0.274190
15	Asian	0.257044
5	votes16_Hillary_Clinton	0.248183
21	Office	0.203204
39	women_percent	-0.252851
20	Service	-0.275662
2	percentage16_Hillary_Clinton	-0.289022
7	percentage20_Joe_Biden	-0.294472
34	FamilyWork	-0.323693
25	Carpool	-0.323823
23	Production	-0.406732
32	PublicWork	-0.413831
14	Native	-0.514235
12	White	-0.636059

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

Now it is at the very last steps of my final project. The goal is to create some models to predict the 2020 election margin for REP and DEM. I believe that to predict 2020 REP vote% for Trump is already good enough. To get 2020 DEM vote%, you can basically subtract 100% from the REP% to get the DEM% roughly.

The scaled and unscaled data has its own merits and flaws. Therefore, for this final section of my final project, I will do models on both the scaled and the unscaled datasets.

“percentage20_Donald_Trump” is the column in which we will construct models and predict on, using other data from the 2016 election, Covid positive and death rates, as well as the many demographic and economic data in our models.

I will use sklearn train_test_split() method to split the dataset into 4 sets of data:

x_train, x_test, y_train, y_test (Min Max Scaled)

x_train_origin, x_test_origin, y_train_origin, y_test_origin (Unscaled)

For example, here is what my y_train looks like for the scaled data:

		percentage20_Donald_Trump
state	county	
Tennessee	Trousdale	0.739445
Missouri	Webster	0.805858
Arkansas	Lawrence	0.792022
Texas	Victoria	0.680401
Missouri	Camden	0.768877

Here is the x_train data I have generated for the scaled data:

		percentage16_Donald_Trump	**percentage16_Hillary_Clinton**	**...**
state	county			
Tennessee	Trousdale	0.672517	0.311750	...
Missouri	Webster	0.788263	0.184102	
Arkansas	Lawrence	0.726374	0.221289	
Texas	Victoria	0.691875	0.294156	
Missouri	Camden	0.771175	0.209675	

After we successfully split up the dataset into 2 training sets and 2 test sets, I select models.

I have decided to use the Generalized Linear Model and the Robust Linear Model.

The reason why I chose these models is because the training set has many independent variables. It is good to use a model that can ultimately report back how statistically significant are from each individual variable.

Both models are from statsmodel. One is sm.GLM() and the other is sm.RLM()

For the GLM, I will use the Gaussian family as input for multivariate linear models.

For RLM, I will just use the default as it will calculate the robustness of each dependency.

We will be looking at the statistical summary of each model.

In addition, we will use the trained models to predict the y_test based on the x_test.

After we gather all the y_test from different models between the unscaled and scaled dataset, we will compute the Root Mean Square Error (RMSE) for each column of y_test. And we will make a final decision of which model is the best to use.

First, it is the GLM model on the scaled data with its statistical reports:

Now let's analyze the report from GLM scaled. It is very obvious that the 2016 Trump and 2016 Clinton vote% are the most dominant factors in this model, as we can tell from the P>|z| = 0.000 to know those 2 variables are extremely statistically significant.

And maybe to your disappointment or surprise, most of the demographics and economic data plays the minimal role in the model, as the z score is way too low or P>|z| value way too high to be statistically significant in the linear model.

A few of them do stand out, in particular, is the Covid-19 death rate for each county, which is extremely statistically significant at 1%: P>|z| = 0.010. Looking at its coefficient, we can confidently conclude that Covid-19 not only plays an important role in the 2020 election, but also it is a much more important factor than any race and demographic variable, which has a massive implication. The Covid-19 case rate is also statistically significant at 5% level. And those metrics do damage Trump's vote% significantly.

The last significant variables are a few minorities, such as Hispanic, Pacific, though not as statistically significant as what mentioned before on this page.

Generalized Linear Model Regression Results

Dep. Variable:	percentage20_Donald_Trump	No. Observations:	2364
Model:	GLM	Df Residuals:	2332
Model Family:	Gaussian	Df Model:	31
Link Function:	identity	Scale:	0.00097887
Method:	IRLS	Log-Likelihood:	4851.9
Date:	Sun, 13 Dec 2020	Deviance:	2.2827
Time:	13:04:33	Pearson chi2:	2.28
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
percentage16_Donald_Trump	0.4892	0.026	19.017	0.000	0.439	0.540
percentage16_Hillary_Clinton	-0.4747	0.027	-17.465	0.000	-0.528	-0.421
Hispanic	0.0943	0.047	2.012	0.044	0.002	0.186
White	0.0556	0.047	1.175	0.240	-0.037	0.148
Black	0.0270	0.042	0.645	0.519	-0.055	0.109
Native	-0.0064	0.042	-0.154	0.878	-0.088	0.076
Asian	-0.0326	0.027	-1.205	0.228	-0.086	0.020
Pacific	0.0633	0.025	2.510	0.012	0.014	0.113
Poverty	0.0408	0.015	2.691	0.007	0.011	0.071
ChildPoverty	0.0019	0.014	0.136	0.892	-0.026	0.030
Professional	-0.2235	0.567	-0.394	0.693	-1.334	0.888
Service	-0.1431	0.457	-0.313	0.754	-1.038	0.752
Office	-0.1084	0.319	-0.340	0.734	-0.733	0.517
Construction	-0.0678	0.337	-0.201	0.840	-0.728	0.592
Production	-0.0808	0.479	-0.169	0.866	-1.020	0.859
Drive	-0.3347	0.858	-0.390	0.696	-2.016	1.346
Carpool	-0.1261	0.276	-0.457	0.647	-0.667	0.414
Transit	-0.1564	0.581	-0.269	0.788	-1.296	0.983
Walk	-0.1470	0.399	-0.369	0.712	-0.928	0.634
OtherTransp	-0.0788	0.130	-0.607	0.544	-0.333	0.176
WorkAtHome	-0.1782	0.310	-0.574	0.566	-0.787	0.430
MeanCommute	0.0145	0.006	2.538	0.011	0.003	0.026
PrivateWork	0.4182	0.637	0.656	0.512	-0.831	1.667
PublicWork	0.4304	0.606	0.710	0.478	-0.757	1.618
SelfEmployed	0.3602	0.427	0.844	0.399	-0.477	1.197
FamilyWork	0.0704	0.090	0.779	0.436	-0.107	0.248
Unemployment	-0.0108	0.010	-1.133	0.257	-0.029	0.008
cases_rate	0.0192	0.010	1.965	0.049	4.84e-05	0.038
deaths_rate	0.0234	0.009	2.589	0.010	0.006	0.041
men_percent	0.5352	1.395	0.384	0.701	-2.199	3.270
women_percent	0.5200	1.395	0.373	0.709	-2.214	3.254
Party_Cluster	0.0013	0.002	0.603	0.546	-0.003	0.006

Then I will use this model to predict on x_test and will compare with y_test

state county	percentage20_Donald_Trump	Model_Average_Trump	Model_GLM_Trump
New Mexico DeBaca	0.732496	0.649853	0.737774
Illinois Saline	0.735554	0.649853	0.757657
Arkansas Lafayette	0.649822	0.649853	0.627183
Colorado Lincoln	0.820898	0.649853	0.808537
MinnesotaMartin	0.674376	0.649853	0.688849

I will also build RLM Robust Linear Model, in addition to GLM Generalized linear model

Robust linear Model Regression Results

Dep. Variable:	percentage20_Donald_Trump	No. Observations:	2364
Model:	RLM	Df Residuals:	2332
Method:	IRLS	Df Model:	31
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Sun, 13 Dec 2020		
Time:	13:18:53		
No. Iterations:	50		

	coef	std err	z	P>|z|	[0.025	0.975]
percentage16_Donald_Trump	0.5202	0.017	29.981	0.000	0.486	0.554
percentage16_Hillary_Clinton	-0.4550	0.018	-24.815	0.000	-0.491	-0.419
Hispanic	0.0097	0.032	0.307	0.758	-0.052	0.072
White	-0.0070	0.032	-0.218	0.827	-0.069	0.056
Black	-0.0170	0.028	-0.603	0.546	-0.072	0.038
Native	-0.0449	0.028	-1.592	0.111	-0.100	0.010
Asian	-0.0481	0.018	-2.635	0.008	-0.084	-0.012
Pacific	0.0416	0.017	2.448	0.014	0.008	0.075
Poverty	0.0226	0.010	2.207	0.027	0.003	0.043
ChildPoverty	0.0064	0.010	0.661	0.509	-0.012	0.025
Professional	-0.0436	0.382	-0.114	0.909	-0.793	0.706
Service	0.0135	0.308	0.044	0.965	-0.590	0.617
Office	-0.0115	0.215	-0.054	0.957	-0.433	0.410
Construction	0.0407	0.227	0.179	0.858	-0.404	0.486
Production	0.0685	0.323	0.212	0.832	-0.565	0.702
Drive	-0.3161	0.578	-0.546	0.585	-1.450	0.818
Carpool	-0.1245	0.186	-0.669	0.503	-0.489	0.240
Transit	-0.1279	0.392	-0.326	0.744	-0.897	0.641
Walk	-0.1475	0.269	-0.548	0.583	-0.675	0.380
OtherTransp	-0.0688	0.088	-0.786	0.432	-0.240	0.103
WorkAtHome	-0.1633	0.209	-0.780	0.435	-0.574	0.247
MeanCommute	0.0127	0.004	3.308	0.001	0.005	0.020
PrivateWork	-0.0477	0.430	-0.111	0.912	-0.890	0.795
PublicWork	-0.0155	0.409	-0.038	0.970	-0.817	0.786
SelfEmployed	0.0378	0.288	0.131	0.895	-0.527	0.602
FamilyWork	0.0110	0.061	0.181	0.856	-0.108	0.131
Unemployment	-0.0105	0.006	-1.639	0.101	-0.023	0.002
cases_rate	0.0152	0.007	2.309	0.021	0.002	0.028
deaths_rate	0.0164	0.006	2.703	0.007	0.005	0.028
men_percent	0.8277	0.941	0.880	0.379	-1.017	2.672
women_percent	0.8072	0.941	0.858	0.391	-1.037	2.651
Party_Cluster	0.0020	0.001	1.345	0.179	-0.001	0.005

The best of RLM over GLM is more computationally intensive as it will use the much more complicated covariance equation to factor in the robustness of each variable. In general, RLM is the superior model over GLM in most cases, especially here with many interrelated variables in this dataset. This model is the best model I've obtained for this project.

Here are the min max scaled models I've attempted. Now let's use GLM and RLM on the unscaled dataset.

Generalized Linear Model Regression Results

Dep. Variable:	percentage20_Donald_Trump	No. Observations:	2364
Model:	GLM	Df Residuals:	2333
Model Family:	Gaussian	Df Model:	30
Link Function:	identity	Scale:	7.4715
Method:	IRLS	Log-Likelihood:	-5715.9
Date:	Sun, 13 Dec 2020	Deviance:	17431.
Time:	13:27:32	Pearson chi2:	1.74e+04
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
percentage16_Donald_Trump	0.4948	0.025	19.524	0.000	0.445	0.544
percentage16_Hillary_Clinton	-0.4789	0.027	-17.559	0.000	-0.532	-0.425
Hispanic	0.0875	0.041	2.153	0.031	0.008	0.167
White	0.0539	0.041	1.325	0.185	-0.026	0.134
Black	0.0323	0.041	0.785	0.433	-0.048	0.113
Native	-0.0036	0.044	-0.082	0.935	-0.091	0.083
Asian	-0.0656	0.056	-1.163	0.245	-0.176	0.045
Pacific	0.4954	0.192	2.575	0.010	0.118	0.872
Poverty	0.0721	0.027	2.697	0.007	0.020	0.125
ChildPoverty	0.0020	0.016	0.123	0.902	-0.030	0.034
Professional	-0.3270	0.860	-0.380	0.704	-2.012	1.358
Service	-0.2563	0.859	-0.298	0.766	-1.941	1.428
Office	-0.2805	0.860	-0.326	0.744	-1.965	1.404
Construction	-0.1602	0.860	-0.186	0.852	-1.845	1.525
Production	-0.1313	0.860	-0.153	0.879	-1.816	1.553
Drive	-0.3105	0.821	-0.378	0.705	-1.920	1.299
Carpool	-0.3661	0.822	-0.445	0.656	-1.977	1.245
Transit	-0.2120	0.822	-0.258	0.796	-1.823	1.399
Walk	-0.2929	0.821	-0.357	0.721	-1.903	1.317
OtherTransp	-0.4887	0.821	-0.595	0.552	-2.099	1.121
WorkAtHome	-0.4619	0.822	-0.562	0.574	-2.072	1.148
MeanCommute	0.0333	0.013	2.508	0.012	0.007	0.059
PrivateWork	0.6257	0.982	0.637	0.524	-1.298	2.550
PublicWork	0.6799	0.982	0.693	0.489	-1.244	2.604
SelfEmployed	0.8103	0.981	0.826	0.409	-1.113	2.733
FamilyWork	0.7534	0.987	0.763	0.445	-1.181	2.688
Unemployment	-0.0331	0.029	-1.144	0.253	-0.090	0.024
cases_rate	0.0802	0.041	1.942	0.052	-0.001	0.161
deaths_rate	3.1912	1.233	2.588	0.010	0.774	5.608
men_percent	0.3593	1.546	0.232	0.816	-2.671	3.390
women_percent	0.3238	1.546	0.209	0.834	-2.706	3.353

Much like the GLM for scaled dataset before, here we train the GLM on an unscaled dataset.

To compare this unscaled model to the previous scaled one, I can already tell this model is less accurate than the min max scaled version, because the P>|z| values are greater here on most of the variables, which means they are less likely to be statistically significant in the model.

But nevertheless, it is still very accurate, and it predicts the vote% for Trump in 2020 on the correct 100% scale. Problem with min max scale is the inability to rescale back to the original scale.

Robust linear Model Regression Results

Dep. Variable:	percentage20_Donald_Trump	No. Observations:	2364
Model:	RLM	Df Residuals:	2333
Method:	IRLS	Df Model:	30
Norm:	HuberT		
Scale Est.:	mad		
Cov Type:	H1		
Date:	Sun, 13 Dec 2020		
Time:	13:27:59		
No. Iterations:	33		

	coef	std err	z	P>|z|	[0.025	0.975]
percentage16_Donald_Trump	0.5246	0.017	30.680	0.000	0.491	0.558
percentage16_Hillary_Clinton	-0.4604	0.018	-25.013	0.000	-0.496	-0.424
Hispanic	0.0143	0.027	0.521	0.602	-0.039	0.068
White	0.0006	0.027	0.020	0.984	-0.053	0.054
Black	-0.0103	0.028	-0.371	0.710	-0.065	0.044
Native	-0.0445	0.030	-1.485	0.137	-0.103	0.014
Asian	-0.1020	0.038	-2.680	0.007	-0.177	-0.027
Pacific	0.3339	0.130	2.572	0.010	0.079	0.588
Poverty	0.0404	0.018	2.237	0.025	0.005	0.076
ChildPoverty	0.0067	0.011	0.613	0.540	-0.015	0.028
Professional	-0.0693	0.580	-0.120	0.905	-1.206	1.067
Service	0.0236	0.580	0.041	0.968	-1.113	1.160
Office	-0.0348	0.580	-0.060	0.952	-1.172	1.102
Construction	0.1019	0.580	0.176	0.861	-1.035	1.239
Production	0.1220	0.580	0.210	0.833	-1.015	1.259
Drive	-0.2788	0.554	-0.503	0.615	-1.365	0.808
Carpool	-0.3476	0.555	-0.627	0.531	-1.435	0.740
Transit	-0.1580	0.554	-0.285	0.776	-1.245	0.929
Walk	-0.2806	0.554	-0.506	0.613	-1.367	0.806
OtherTransp	-0.4124	0.554	-0.744	0.457	-1.499	0.674
WorkAtHome	-0.4085	0.554	-0.737	0.461	-1.495	0.678
MeanCommute	0.0294	0.009	3.278	0.001	0.012	0.047
PrivateWork	-0.0849	0.662	-0.128	0.898	-1.383	1.213
PublicWork	-0.0350	0.662	-0.053	0.958	-1.333	1.263
SelfEmployed	0.0767	0.662	0.116	0.908	-1.221	1.374
FamilyWork	0.1137	0.666	0.171	0.864	-1.192	1.419
Unemployment	-0.0322	0.020	-1.648	0.099	-0.070	0.006
cases_rate	0.0617	0.028	2.216	0.027	0.007	0.116
deaths_rate	2.2911	0.832	2.754	0.006	0.660	3.922
men_percent	0.8214	1.043	0.787	0.431	-1.223	2.866
women_percent	0.7733	1.043	0.741	0.458	-1.271	2.817

The robust linear model for the unscaled version. Nothing new is worthy to discuss as I will only repeat what I've said in the scaled version.

The last part of this project is to compare and contrast how well each model performs.

And here is the table which summarizes the Root Mean Square Error (RMSE) of the 4 models.

Root Mean Square Error	models	not_scaled	scaled
Average of y_test	Average	15.394538	0.176184
Generalized Linear Model	GLM	2.492202	0.028572
Robust Linear Model	RLM	2.483165	0.028506

The table really helps to analyze the merits and flaws of each model.

The data after adjusting the scale on Min Max of each column really improves the RMSE quite significantly, for both the GLM and RLM models.

The robust linear model on the adjusted scale only has RMSE = 0.03, which is an incredibly accurate model at predicting the 2020 election Trump vote%.

		percentage20_Donald_Trump	Model_Average_Trump	Model_GLM_Trump	Model_RLM_Trump
state	**county**				
New Mexico	**De Baca**	0.732496	0.649853	0.737774	0.728339
Illinois	**Saline**	0.735554	0.649853	0.757657	0.758150
Arkansas	**Lafayette**	0.649822	0.649853	0.627183	0.627234
Colorado	**Lincoln**	0.820898	0.649853	0.808537	0.808237
Minnesota	**Martin**	0.674376	0.649853	0.688849	0.689165

But unfortunately for the highly accurate model after adjusting the scale, it is not intuitive to look at all the numbers between 0 and 1. So here is the unscaled result below:

		percentage20_Donald_Trump	Model_Average_Trump	Model_GLM_Trump	Model_RLM_Trump
state	**county**				
New Mexico	**De Baca**	72.807991	65.586883	73.251640	72.440118
Illinois	**Saline**	73.075187	65.586883	74.996622	75.035972
Arkansas	**Lafayette**	65.584173	65.586883	63.610829	63.616602
Colorado	**Lincoln**	80.532319	65.586883	79.452636	79.432277
Minnesota	**Martin**	67.729661	65.586883	68.972087	68.993832

Here we can clearly see the actual result of De Beca county, NM for Trump is 72.8%, our model predicts at 72.44%, which is again very accurate, even if not as precious as the scaled version. The unscaled models are easier to interpret the context behind the number.
And this is the end of my data analysis, visualization, and machine learning modeling.

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

V. Recap Results & Key Takeaway

The results and key takeaways are covered in the data analysis & visualization section.
But here is the recap for all the highlights and conclusions of this final project:

1. Almost all the national polls were off by a lot from the actual election results.

2. Majority of the national polls have the bias to favor Democrats and Joe Biden in the forecast. Most likely due to Covid-19 that REP isn't as exciting as DEM to answer phone calls.

3. The reason behind why the national polls overestimate the Democrat's support in the general election is most likely due to the Covid-19 pandemic, which led to REP to distrust the polls and skew the poll sample of the voting population.

4. It is statistically significant that as the Covid-19 daily positive rate rises, the polling margin for Democrats becomes wider, and it is more likely for the national polls to overestimate Democrat's actual performance in the general election on Nov. 3rd, 2020.

5. On the national level, Democrats won the states where they tend to have lower daily Covid 19 positive rates compared to those Republican red states. This is because Republican states tend to have fewer travel and mask restrictions compared to Democratic states.

However, the trend is completely flipped if looking at only the battleground states. One possible explanation for this conflict is that Covid-19 might help Joe Biden to flip those battleground states where the positive rate remains high. In another word, Covid-19 might damage Trump's prospect for reelection, as he was blamed for mishandling the pandemic.

6. From the correlation matrix and heatmap, we can know quantitatively how each individual factor contributes to the 2020 election.

Trump is popular among Whites and Biden is popular among minorities. Trump is popular with construction and production workers, whereas Biden has more support in the Retail, Service, Transit industry. Trump is more popular among men; Biden is more popular among women. Biden is also more popular among unemployed people.

7. Min Max Scaling of the dataset will dramatically improve the accuracy of the regression model, and to decrease the Root Mean Square Error, but in return, make it very difficult to interpret the variables afterward. Because for a scaled dataset, every value is scaled from -1 to 1. In order to know what percentage of votes for Trump or Biden won in each county, the unscaled dataset is still useful, despite being inferior and less accurate.

8. The use of Gaussian Mixture Model (GMM) and the Principal Component Analysis (PCA) from the class can really help to visualize how the 2 clusters of DEM and REP are distributed and which independent variables contribute to the election result the most.

9. The models in this project: GLM (Generalized Linear Model) and RLM (Robust Linear Model) works a lot better in the min max scaled dataset, compared to the unscaled version. And the RLM is the superior model over GLM due to its statistical robustness.

10. The 2016 vote% for Trump and Clinton plays the predominant roles of all the models, with a very high level of statistical significance.

11. Surprisingly, most of the demographic variables aren't statistically significant enough to be included in the linear regression model. However, both the Covid-19 death rates and positive rates are statistically significant, which indeed damage Trump's reelection odds. A few minority groups are statistically significant once we control other variables, which again work against Trump's favor.

Sicheng Yi (Tiger Yi)
U431888754 tigeryi@bu.edu
CS506 Prof. Lance Galletti

VI. Limitations & Potential Future Work

The time of me working on this individual project is fairly limited. I would like to apologize to any error in this project report and the code itself, as I am not a perfect person.

But jokes aside, this project is quite interesting. It is very fresh, and since I started the final project quite late, I am lucky that I am able to carry out a data science project on the subject of the 2020 U.S. general election.

One of the biggest limitations is the data I gathered. Many datasets I gathered are extremely fresh, like just a few weeks old. The election result data is NOT the finalized and certified data so some of the data points could be incomplete or inaccurate.

I wished I would have more time to spare on the modeling. I've tried GLM and RLM models which are already incredibly accurate at predicting the 2020 Trump vote% on the 3,000+ U.S. state counties. However, those models include TOO MANY variables that are statistically insignificant. A more thorough modeling will be dropping variables and comparing and contrasting the performance on RMSE.

Another limitation is mostly individual habit. My code in general tends to be messy and out of chronological order, as I have to go back to fix the code over and over again.

I hope this 35 page final report will help to clear out any confusion for my final project.

There are tons of potential in this project. It's been a hot topic that the elections polls are inaccurate. Misinformation and misunderstandings on how the polls are conducting is the result of the degradation of social trust and confidence in our society. To make the elections polls more accurate next time will play a huge role in social concord and peace.

I hope this final project is not only interesting but also eye opening. I have based the inaccuracy of the elections mostly on the Covid-19 pandemic in this project. But I am certain there are a lot more variables that are yet to be explored. This project is just a starting point to make the election polls more accurate and trustworthy on the quantitative level. I look forward to people to pick up my project and expand upon it in the future.

We all love to know what is going to happen in the 2022 midterm election or 2024 general election. But first, we must improve on our ability of prediction. Readers, it is now your turn to carry on where I left off and take the future in our hands.