Demographic Predictors of County-Level Election Results

Cindy Li & Maya Dayan

Final Report

PHP 2650

May 10, 2020

Introduction

This paper seeks to answer the following research questions: What demographic indicators are the best predictors of a county's voting patterns, both in general and primary elections? How strong are purely demographic prediction models (without polling data)?

The results of the 2016 Presidential election defied most prediction models based on public opinion data.[1] While each model was unique, two of the most prevalent issues were incorrect weighting and incorrect self-reporting. More specifically, a large number of polls did not weigh by education. Higher educated individuals are more likely to respond to pollsters and are also more likely to vote for Democrats. As these individuals were disproportionately represented in the samples, polls skew towards the Democratic candidate, Hillary Clinton.[2]

A more fundamental flaw with these models was the inaccuracy of the data itself. Later analysis revealed what experts called a "shy Trump effect." This describes a condition where Trump voters were more hesitant to identify themselves as Trump supporters, generally claiming they were undecided instead.[3] This is a serious concern for all prediction models relying on self-reported data, but only becomes a serious problem when the level of inaccuracy in the data correlates with the outcome.

This paper attempts to find an alternative or companion to public opinion polling using purely demographic indicators to build machine learning models. The electorate is often dissected into demographic-based voting blocks, with headlines like "Trump is Entering the

---

[1] Andrew Mercer et al. *Why 2016 Election Polls Missed Their Mark*. Pew Research Center, 9 Nov. 2016, www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/.
[2] Nate Cohn. *A 2016 Review: Why Key State Polls Were Wrong About Trump*. The New York Times, 31 May 2017, www.nytimes.com/2017/05/31/upshot/a-2016-review-why-key-state-polls-were-wrong-about-trump.html.
[3] Danielle Kurtzleben. *4 Possible Reasons The Polls Got It So Wrong This Year*. NPR, 14 Nov. 2016, www.npr.org/2016/11/14/502014643/4-possible-reasons-the-polls-got-it-so-wrong-this-year.

2020 General-Election Season with Key Demographics Moving Away from Him"[4] and "Older voters could offer Biden a new path to the White House"[5] already circulating around the 2020 election. After the 2016 election, various outlets ran reports about how various demographic groups voted, but they did not use this data as potential predictors.[6]

The paper uses the relationship between a counties demographics and how counties voted in the 2014 gubernatorial election to predict demographic's impact on 2016 presidential election outcomes in four key swing states. It also attempts a similar model for the 2020 Democratic Primary, predicting how counties will vote in later states based on how demographics impacted counties voting in earlier states. This was included to see if demographics are a strong indicator for more nuanced political distinctions or only larger party affiliation.

Methods

We chose to focus on the county-level analysis as we wanted to be able to compare our analysis with actual election results and the county was the smallest area in which we had both the demographic and election data. An individual-level analysis would only tell us how various demographic groups voted, information which already exists. Our model identifies the demographic characteristics most determinant of whether a county will vote for a Democrat or

[4] Philip Bump. "Analysis | Trump Is Entering the 2020 General-Election Season with Key Demographics Moving Away from Him." *The Washington Post*, WP Company, 9 Apr. 2020, www.washingtonpost.com/politics/2020/04/09/trump-enters-2020-general-with-key-demographics-moving-away-him/.

[5] Ronald Brownstein. "Older Voters Could Offer Biden a New Path to the White House." *CNN*, Cable News Network, 28 Apr. 2020, www.cnn.com/2020/04/28/politics/joe-biden-2020-older-voters/index.html.

[6] "An Examination of the 2016 Electorate, Based on Validated Voters." *Pew Research Center - U.S. Politics & Policy*, 7 Jan. 2020, www.people-press.org/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/.

Republican, and later the demographic characteristics making a county most likely to vote for a specific Democratic candidate.

The four states chosen for general election prediction are Iowa, Michigan, Pennsylvania and Wisconscin. All four of these states voted for Obama in 2012 and Trump in 2016, and are considered swing states. Michigan, Pennsylvania and Wisconscin are all considered critical swing states for 2020,[7] and Iowa was added because it is so critical in the primary process as it is the first state to vote. Furthermore, the other three states were generally incorrectly predicted to vote for Clinton while Iowa was correctly predicted to vote for Trump, so adding it to the analysis gives the model more external validity for a larger range of swing states.[8] Finally, all four states had gubernatorial elections in 2014, and the results from these races were used to create our training dataset.

For predicting primary election results, we chose Iowa, New Hampshire, Nevada, South Carolina to comprise our training set and California and Texas to comprise our testing set. Iowa and New Hampshire were selected for training since it is generally believed that the results of their caucus and primary, respectively, can have an influence in the results of other states' primaries, whereas Nevada and South Carolina were selected as they can be indicators of how certain candidates will fare among different racial demographic groups. California and Texas were chosen as the testing set as they have the largest number of delegates.

Our demographic data comes from the 2014, 2016, and 2018 American Community Survey, which provides extensive demographic data at the county level.[9] The 2018 data was used

---

[7] "The Swing Voters in 3 Key States Democrats Must Persuade in 2020." *Morning Consult*, 26 Nov. 2019, morningconsult.com/2019/11/21/the-swing-voters-in-3-key-states-democrats-must-persuade-in-2020/.
[8] "2016 Presidential Election Forecast Maps." *270 To Win*, www.270towin.com/2016-election-forecast-predictions/.
[9] US Census Bureau. "American Community Survey 5-Year Data (2009-2018)." *The United States Census Bureau*, 13 Nov. 2019, www.census.gov/data/developers/data-sets/acs-5year.html.

for the 2020 Democratic primary as it was the most recent complete dataset available. From this data, the percentage in a county of different races, ages, marital statuses, income levels, education levels, and citizenship statuses were used as predictors for political outcomes. The data of outcomes of the 2014 gubernatorial elections by county came from the New York Times.[10] The 2016 presidential election results were sourced from GitHub, which was compiled from TownHall.com.[11] Finally, the 2020 Democratic primary results data came from the New York Times as well.[12]

Analysis

For the general election prediction, we used a decision tree, for examining feature importance, and a random forest, to make the model more robust to overfitting. We also used an SVM with a radial kernel, logistic regression, and k-nearest neighbors, using cross-validation to select the hyperparameters for the SVM.

For the primary election prediction, we again used a decision tree for feature importance analysis and a random forest to help prevent overfitting, as well as k-nearest neighbors. We additionally used an xgboosted model to further try to prevent overfitting.

Results

[10] "2014 Governor Election Results." New York Times, Accessed May 10 2020, https://int.nyt.com/applications/elections/2014/data/2014-11-04/supermap/governor.json
[11] Tonmcg. "US County Level Election Results 08-16." *GitHub*, 7 Sept. 2018, github.com/tonmcg/US_County_Level_Election_Results_08-16.
[12] "2020 Iowa Democratic Presidential Primary Election Results." New York Times, Accessed May 10 2020, https://int.nyt.com/applications/elections/2020/data/api/2020-02-03/iowa/president/democrat.json

For the general election prediction problem, our best model was the random forest, which achieved an accuracy of 100% on the governor election data and an accuracy of ~90% on the presidential election data. The other models also performed well, with the decision tree having the lowest accuracy on the test set of ~80%, while the logistic regression and k-nearest neighbors models achieved accuracies around 87%.

For the primary election prediction problem, the highest accuracy achieved on the testing set was by the decision tree with an accuracy of only ~55%. The decision tree's training accuracy was around 85%. The other models' accuracies ranged from 37% to 48%.

Discussion

Our results also show that demographic data is a strong predictor for whether or not a county votes Democratic or Republican, but is less useful in predicting primary results. There are a few reasons that this would be the case, primarily that the differences between 2020 Democratic candidates are smaller so that very similar people could still vote different ways. Even a month before the Iowa and New Hampshire primaries, only around a third of voters considered themselves fully committed to a candidate.[13]

In addition, political residential segregation could make counties more homogenous in terms of party affiliation. When Americans move, they generally want to live in an area where people share the same ideology as them, meaning that Democrats and Republicans are less likely to live together today than a decade ago.[14] This means that the model likely benefited from the

[13] Ella Nilsen. "Less than One-Third of Iowa and New Hampshire Voters Have Settled on a Candidate." *Vox*, Vox, 11 Jan. 2020, www.vox.com/policy-and-politics/2020/1/11/21057416/iowa-and-new-hampshire-voters-undecided-2020-election.
[14] Alan Greenblatt. *How Republicans And Democrats Ended Up Living Apart*. NPR, 27 Nov. 2013, www.npr.org/sections/itsallpolitics/2013/11/26/247362143/how-republicans-and-democrats-ended-up-living-apart.

fact that not only do the Republican and Democratic parties disproportionately attract different demographic groups, but those groups are then more likely to live next to each other because of their political affiliation. The same cannot be said for primary voting.

In addition, our demographic data describes the full county. While going from the full county to just voters is certainly a leap, the model seems to better account for it than the leap from the full county to only the Democratic voters who are relevant in the primary. The demographic composition of the full county is less representative of the Democratic primary voters than all voters.

The level of success achieved by our models shows that purely demographic predictors can go a very long way in predicting election results. This paper hasn't examined the mechanisms from expanding from county-level winners to the state and national level, but it has shown that demographic composition has strong predictive power over election results, and this model could serve as a good basis for election forecasts on its own and particularly in conjunction with public opinion polling. Our models, however, focused on four states, all of which voted for Trump in the 2016 election. In fact, in the case of the random forest model, the model often predicted Republican when the true label was Democrat.

Again, our model's success rate was very high, and the areas which incorrectly predicted the outcome were pretty evenly split between counties which voted for Trump and those which voted for Clinton. However, there were far fewer counties that did vote for Clinton, and those that did despite the model predicting they would vote for Trump were largely in Iowa and Wisconsin, while the counties the model predicted would vote for Clinton and really voted for

Trump were mostly in Michigan and Pennsylvania In the figure below, you can see the distribution of the inaccurate predictions.

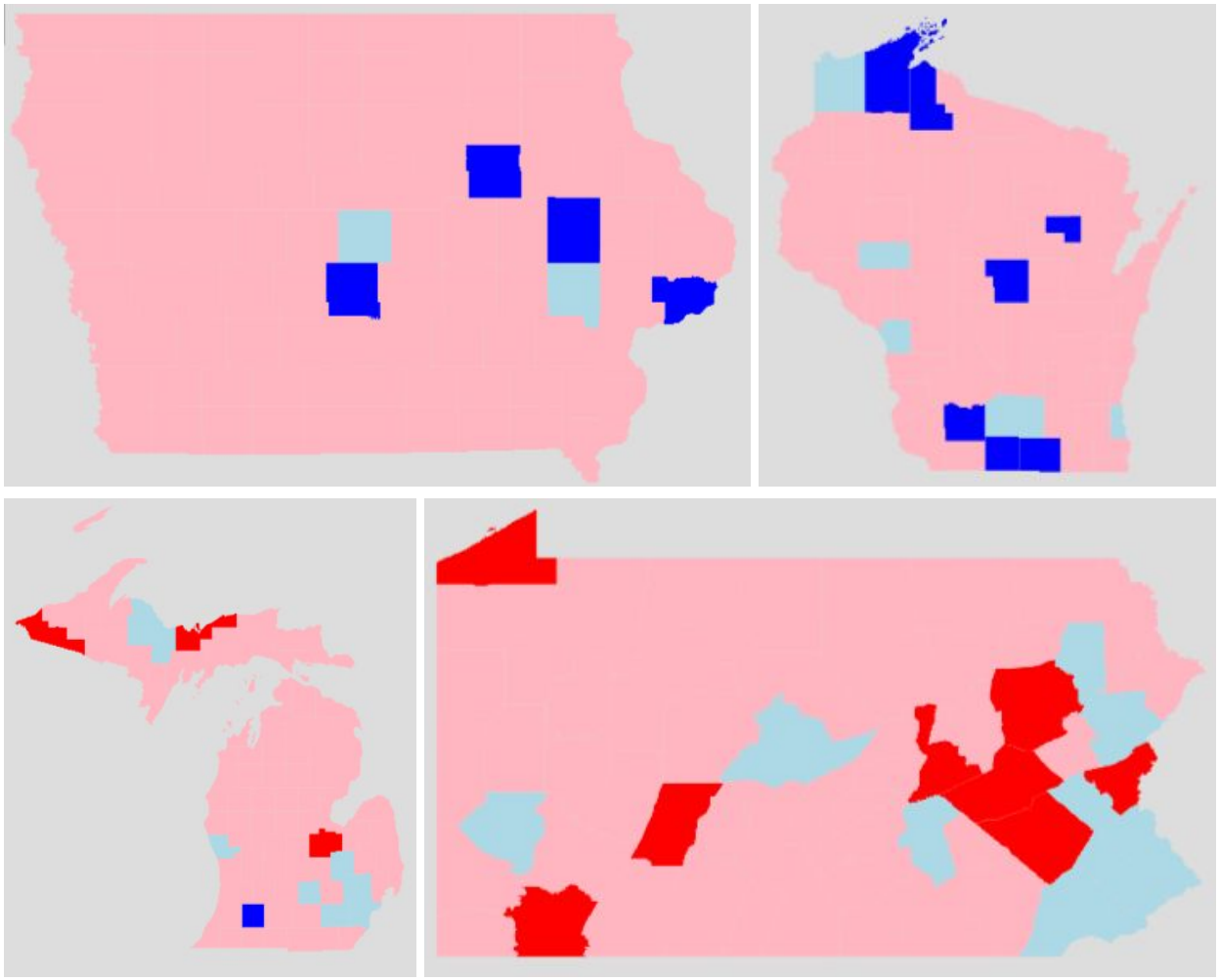

Figure 1. Graphs of each of the states and their county predictions. The bright red indicates counties that voted Republican but were incorrectly predicted to vote Democrat, while the dark blue indicates counties that voted Democrat but were incorrectly predicted to vote Republican. The lighter colors indicate a correct prediction.

We used the decision tree to determine the most important features for prediction. The random forest yielded similar results in terms of relative importance.

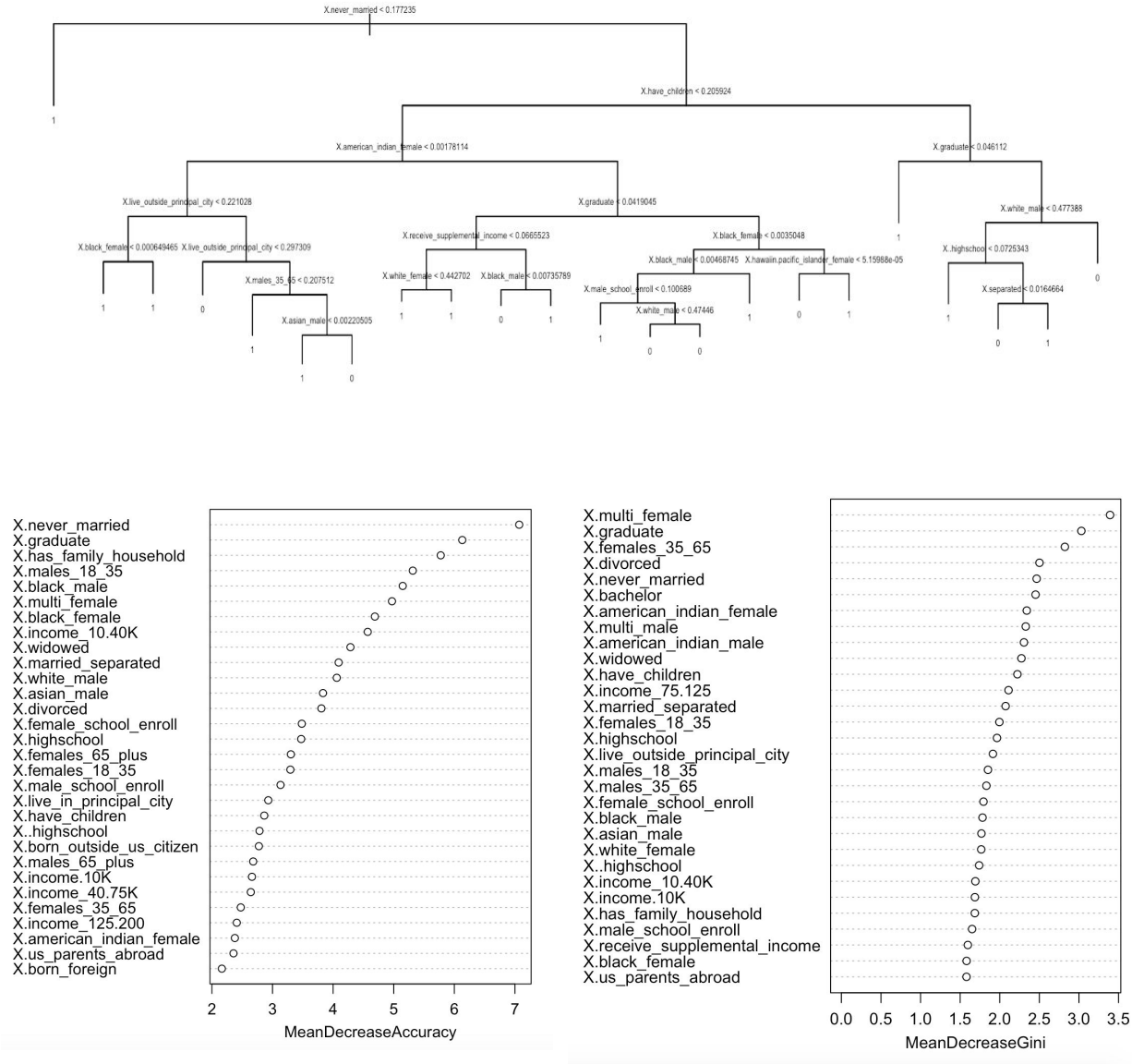

Figure 2. The top graph is the decision tree. The bottom two graphs are the feature importances, with the left being the mean decrease in accuracy for each feature and the bottom right being the mean decrease in gini index. We can see in all three graphs that the percentage of people never married seems to be an important feature.

We can see that the first split of the tree is based on the percentage of people who were never married, making it an important feature in determining whether a county will vote Democrat or Republican. It similarly had the highest mean decrease in accuracy and a relatively

high mean decrease in gini index. Another important feature appears to be the percent of people whose education level is graduate school. The random forest similarly had these two features as among the most important, with these two having the highest mean decrease in accuracy.

Though none of the 2020 Democratic primary prediction models performed particularly well, given the multitude of candidates, our model still performed much better than random guessing. Though our models did not have high accuracies, we can examine the tree trained on the four states to determine feature importance.

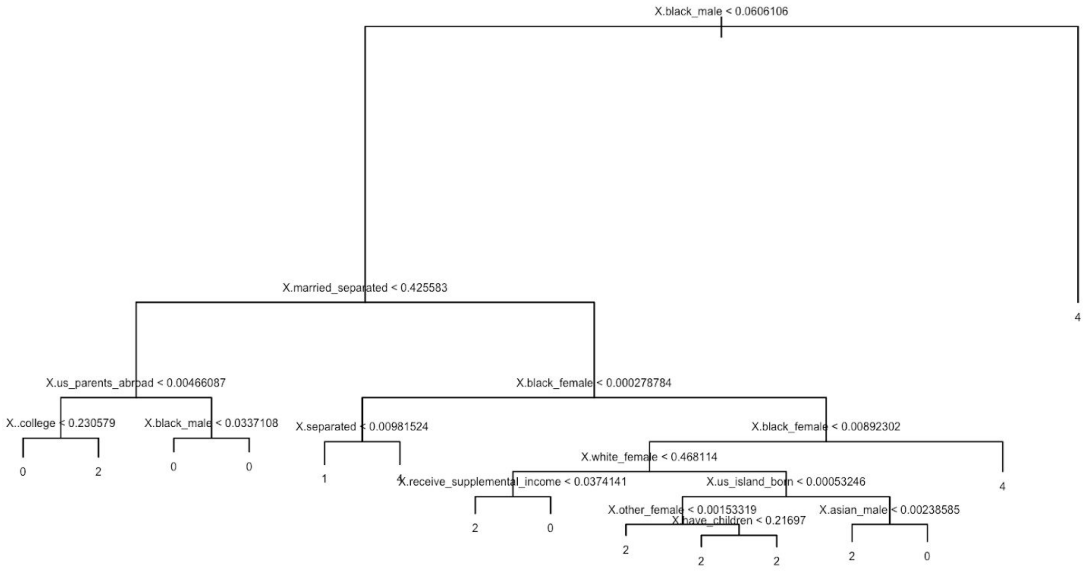

Figure 3. Decision tree for predicting primary election results. 0 represents Sanders, 1 indicates Klobuchar, 2 represents Buttigieg, and 4 represents Biden. Here we can see that the percent of the population that is black and male seems the be the most important feature.

We can see that the proportion of black males in a county is the most important feature for predicting the primary winner, and that counties with higher proportions of black males are more inclined to vote for Biden. This makes sense given that Biden dominated in South Carolina

and South Carolina has a larger black population. Again, the percentage of people never married seems to be a relatively important feature, given it is the second split.

Conclusion

Our models can predict which party counties will vote for in a presidential election based on their demographics and which party they voted for in the prior gubernatorial elections with high accuracy. Furthermore, it seems that the proportion of the population who have never been married before, as well as the percent of the population who attended graduate school are important features in predicting the party a district votes for. However, the states we focused on voted mostly Republican, and thus may have caused the models to become biased towards predicting Republican. Introducing more Democratic states would be a good next step to see if the model can still predict accurately. Another good next step might be to see whether our models can continue to accurately predict election results in other years, such as using the 2010 gubernatorial results to predict 2012 presidential election results, or whether it was only in this one year that the midterm elections were indicative of presidential election results.

Our models did not fare well with predicting which Democratic candidate a county would vote for in the primaries as it did with predicting which party a county would vote in the presidential election results. Again, this may be because the data we used was county-wide and not specific to the Democratic portion of the population, so future work might focus on using demographic data within parties to see if then the Democratic primary results can be predicted.

Works Cited

"An Examination of the 2016 Electorate, Based on Validated Voters." *Pew Research Center - U.S. Politics & Policy*, 7 Jan. 2020, www.people-press.org/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/.

Brownstein, Ronald. "Older Voters Could Offer Biden a New Path to the White House." *CNN*, Cable News Network, 28 Apr. 2020, www.cnn.com/2020/04/28/politics/joe-biden-2020-older-voters/index.html.

Bump, Philip. "Analysis | Trump Is Entering the 2020 General-Election Season with Key Demographics Moving Away from Him." *The Washington Post*, WP Company, 9 Apr. 2020, www.washingtonpost.com/politics/2020/04/09/trump-enters-2020-general-with-key-demographics-moving-away-him/.

Cohn, Nate. *A 2016 Review: Why Key State Polls Were Wrong About Trump*. The New York Times, 31 May 2017, www.nytimes.com/2017/05/31/upshot/a-2016-review-why-key-state-polls-were-wrong-about-trump.html.

Greenblatt, Alan. *How Republicans And Democrats Ended Up Living Apart*. NPR, 27 Nov. 2013, www.npr.org/sections/itsallpolitics/2013/11/26/247362143/how-republicans-and-democrats-ended-up-living-apart.

Kurtzleben, Danielle. *4 Possible Reasons The Polls Got It So Wrong This Year*. NPR, 14 Nov. 2016, www.npr.org/2016/11/14/502014643/4-possible-reasons-the-polls-got-it-so-wrong-this-year.

Nilsen, Ella. "Less than One-Third of Iowa and New Hampshire Voters Have Settled on a Candidate." *Vox*, Vox, 11 Jan. 2020, www.vox.com/policy-and-politics/2020/1/11/21057416/iowa-and-new-hampshire-voters-undecided-2020-election.

Mercer, Andrew, et al. *Why 2016 Election Polls Missed Their Mark*. Pew Research Center, 9 Nov. 2016, www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/.

"The Swing Voters in 3 Key States Democrats Must Persuade in 2020." *Morning Consult*, 26 Nov. 2019, morningconsult.com/2019/11/21/the-swing-voters-in-3-key-states-democrats-must-persuade-in-2020/.

Tonmcg. “US County Level Election Results 08-16.” *GitHub*, 7 Sept. 2018, github.com/tonmcg/US_County_Level_Election_Results_08-16.

US Census Bureau. “American Community Survey 5-Year Data (2009-2018).” *The United States Census Bureau*, 13 Nov. 2019, www.census.gov/data/developers/data-sets/acs-5year.html.

“2014 Governor Election Results.” New York Times, Accessed May 10 2020, https://int.nyt.com/applications/elections/2014/data/2014-11-04/supermap/governor.json

“2016 Presidential Election Forecast Maps.” *270 To Win*, www.270towin.com/2016-election-forecast-predictions/.

“2020 Iowa Democratic Presidential Primary Election Results.” New York Times, Accessed May 10 2020, https://int.nyt.com/applications/elections/2020/data/api/2020-02-03/iowa/president/democrat.json