# Predicting the Results of the 2020 Election

Brady, E., Gomez, E., Hall, D., Sellar, S., & Yedidi, V.
Foundations of Data Science - Group 2
Rensselaer Polytechnic Institute
Professor Dorit Nevo
October 31, 2020

**Introduction** **3**
Introduction to Political Polls 3
2016 Election and Why Polls Failed 3
**Methods** **4**
**Results** **6**
**Conclusion** **7**
**References** **8**
**Appendix 1- Figures and Tables:** **11**
**Appendix 2 - R Code:** **16**
2.1- Objects: 16
2.2- Main Code: 20

# Introduction

## Introduction to Political Polls

According to a masterclass on political polling “a political poll is the use of survey instruments to elicit and record an individual’s opinions, attitudes, and personal information”[9]. Polls are a form of survey research using a random sample of a population to determine public opinion. In order to determine opinions and other qualitative information open-ended questions can be asked and in order to determine more quantitative information there are fixed options for the respondents' answers. There are different kinds of political polls and each one determines a different piece of information from the sample [9]. There are benchmark polls which are conducted at the beginning of a campaign to determine baseline opinions of a candidate. This kind of poll can be conducted before or immediately after the announcement of intent to run for an office. Another type of poll that is conducted is called a brushfire poll. These polls are to determine a candidate's popularity and support on important issues [9]. These polls can be helpful to determine opinions on public issues that a candidate is campaigning on. There also exists a tracking poll which are regular polls that are given out at specific intervals throughout a campaign [13]. Polls like this are used to track the ongoing opinion of candidates and are updated as the data is collected. Along with these predictive polls that are taken before an election takes place there are also exit polls which are given as voters exit a polling location [9]. These types of polls are used to predict results and to some extent they are used as a rough indicator of election fraud [13]. All of these polls are used at specific times to attempt to predict election results based on public opinion. Historically political polls have been good indicators of performance in an election however, there are times where the polls have been wrong in predicting the outcomes. One example of this is in the 1936 election between Franklin D.

Roosevelt and Alf Landon [3]. In the popular publication Literary Digest it was predicted that Landon would be victorious in the election however FDR emerged the president in a landslide win having the majority in all but two states. According to a history professor at the American University “It was not a scientific poll” and relied on using selective returns and collected the information from telephones and postcards [3]. Another notorious example of a poll being incorrect was the 1948 presidential election between Harry Truman and Thomas Dewey [4]. This election is notorious due to the fact that the poll predicted that Dewey would win with 50% of the popular vote. This poll was so influential that the Chicago Tribune ran an early edition stating that candidate Dewey had won the election when in fact Truman emerged victorious upsetting the entire nation. This error in the poll was due to the method of quota sampling, the samples polled should have all had the same groups represented however they did not and the poll was biased towards the republican candidate. There are a number of issues in polling that can influence their results and it is important for the polling office to consider all of these potential errors and biases that can affect their results.

## 2016 Election and Why Polls Failed

The 2016 election was considered a massive upset as most major prediction models predicted Hillary Clinton would win the election and become the President. After the election results were released pollsters began an in depth look into why the models used were incorrect. There are a number of different reasons why the polls were incorrect including nonresponse bias, limited data, and social desirability bias. The nonresponse bias is a very likely reason why the polls were inaccurate, this bias occurs when certain kinds of people systematically do not respond to surveys [1]. Some groups who were identified as key voters for Trump such as the less educated are less likely to respond to polls. Another reason the 2016 polls failed to predict

the outcome for the election can be due to the fact that the data used is limited in scope [12]. The data collected for these surveys was not a good predictor of future events rather they showed the climate at a specific moment in time. This kind of limited data can be used to make general projections however, there is more to the election results than what is just collected. Another reason for the incorrect polls could be due to the fact that people were not honest in their answers. There are a number of different theories that state that the idea of the secret trump vote was more realistic than people think. After many controversial things on the campaign trail people did not want to openly say that they were going to vote for Trump, this is known as social desirability bias which is the idea that voters give polling answers that they think will reflect well on themselves [5]. Another reason for the polls for the 2016 election predicting the wrong outcome can be attributed to voters who were listed as undecided in the surveys. Surveys taken after the election revealed that many more undecided and minor-party voters went with Trump than was expected, and while the prediction that Clinton would win was wrong, she did win the popular vote [10]. This problem could affect any polling and is hard to account for when it comes to predicting the final outcome, but is important to keep in mind as a deciding factor. Polls are a set of questions that pollsters ask to a sample of the population in an attempt to predict the result of an election. Polls have many different biases and margins of error that can affect these results. These are all potential reasons for why the polls failed to predict the outcome for the 2016 election and need to be considered when analyzing future political polls.

## Methods

To make an initial prediction, polling data was taken from FiveThirtyEight [7]. The file presidential_polls_2020.csv was used. It contains a conglomerate of polls taken in every single state from companies such as surveymonkey, YouGov, Marist College, ABC News, etc. The ‘pct’ column provides the chance each poll says each candidate has of winning.

Once the data was collected, it was determined that for each state, the data was slightly skewed. But when the log transform was taken, it was normalized. The data for the polling was split state by state, and within each state the data for each candidate was split up as well. Then for each state, a two sample, two tailed t test was done at a 10% confidence with a null hypothesis being that the democrats and republicans didn’t have significantly different results in the state, and the alternate hypothesis was that there was a significant difference between each candidate. If the null hypothesis held true, it showed that there isn’t enough conclusive evidence to predict a winner for that state based on the polling data. If the null hypothesis was rejected, it meant that a winner could be predicted as the data was significantly different, so the winner of the state was chosen as the party with the higher percentage of votes. The results for this model can be seen in Figure 2.

Unfortunately, as with any data collection, there is going to be a bias. FiveThirtyEight found that certain polling companies that do the polling have a tendency to collect data from people that lean in certain directions. As seen in Figure 1, it’s clear that research done by companies such as the Pew Research center or Public Policy Polling lean heavily left, while Gallup and Quinnipiac are much more right leaning. [11]

In order to adjust for this polling bias FiveThirtyEight adjusted the data, weighted together with a 'house adjustment' to compensate for the polling company's bias[11]. The same method as described earlier was used to create a prediction for each state based on the house adjusted polling data. The results can be seen in Figure 3.

Another effect that can skew the polling data is known as the trend effect [11]. Data for the polling is collected at different times throughout the election cycle; some polls were done very early on, and some were completed near the end of the election cycle. So FiveThirtyEight also includes trend adjusted data, which accounts for how early in the election cycle the poll was taken. The earlier the poll was done, the less weight it has on the final prediction. A model was created using this adjustment on top of the house adjustment, using the same method to predict the final outcomes as with the first two models. The final results for this can be seen in Figure 4.

Unfortunately, as the 2016 election showed us, polling results can wildly misrepresent the final election results. So another model was created based on historical popular vote data by state for elections from 1976 to 2016. Although there were third party candidates over the years who made a significant amount of noise, James Campbell and Thomas Mann found while making their own election forecasts back in 1996, that these third party candidates never really made enough of an impact on the final results to be considered significant [15]. This holds increasingly more true since then, as the country has become more polarized, leaving less room for third party

candidates. So for this model, the third party candidates were filtered out, and only the popular vote data for democrats and republicans remained.

As with the previous 3 models, the data was split into subsets by state and by candidate, and the mean popular vote over all the election cycles was taken for each data subset. This data was found initially to not be normal, but when the log transform was taken, the data became normalized. A two sample, two tailed t test was performed on the means for each state, where the null hypothesis was that there is no difference between the republican and the democrat popular vote. The alternate hypothesis was that there is a difference between the popular vote for the two parties. If the null hypothesis held true, it proved that the votes for each state did not necessarily lean towards one party, meaning a prediction cannot be made on which way the state is going to lean on 2020. The alternate hypothesis proved that the state did historically lean in one direction, so the mean of the popular votes were compared, and the higher mean was chosen as the direction the state might lean in 2020. Figure 5 shows the final predictions for this model.

To create the most robust model possible, the predictions from the historical popular votes (Figure 5) was combined with the predictions from the house and trend adjusted data (Figure 4). If both models predicted a state would lean in a certain direction, then this model showed that state as leaning in that direction. If there was any discrepancy between the two models for a state, then the state was listed as undecided in this model. One drawback of this method is that it forecasts more states as unpredictable than the historical predictions and the trend and house adjusted predictions do individually. But an advantage of combining those two predictions is that for the states where a final prediction can be made, the chances of this

prediction being correct is much higher than any of the other models created. This model can be found in Figure 6.

None of these models were able to give a prediction for all the different states, so it did not lead to a final prediction for which party is going to win in 2020. A paper written by Matthew B. Incantalupo [19] showed that economic variables such as unemployment and gdp growth rates can impact the final results. So a multiple linear regression model was created using gdp, unemployment, and inflation data found at thebalance.com [20]. The results of this regression with the democratic popular vote percentage being the dependent variable can be seen in Figure 7. This model was able to give an overall prediction of the popular vote, without having to go state by state. This has the advantage of being able to give a direct prediction of who wins the popular vote, but the limitations are that the popular vote is not a direct predictor of the electoral college winner.

## Results

By using the initial polling data taken from fivethirtyeight an initial unadjusted polling prediction was made. As seen in Figure 2, 27 states are designated as democrat 20 are republican and 3 are undetermined. After adjusting the data with the polling bias shown in Figure 3 that 26 states are democratic, 22 are republican and 2 are undetermined. This result is different because when different organizations collect polling data they are biased and this needs to be considered. In looking at the polling predictions for 2020 the data was adjusted by house and trend (Figure 4) to determine which candidate would win each state. 28 states were won by Biden (Democratic/Blue), 20 states were won by Trump (Republican/Red), and 3 states were undecided (Green).Taking the results from the polling predictions for 2020 adjusted by house and trend

(Figure 4), the number of electoral college votes were calculated for each candidate by state (Table 2A). Subsequently, the total number of electoral college votes were summed for each candidate to determine who would win the election (Table 2B). Three states (OH, IA, and TX) were undecided on the analysis of who would win, but in looking at the total votes, these three states would not affect the overall outcome of the election. Using historical data collected from 1976 to the present day it can be seen in Figures 5 and 6 that many states are considered swing states and they have no historical basis in predicting which candidate they will side with based on political party. These differences in party majorities by state add a large amount of uncertainty in predicting results and have a large effect on how predictions are made.

A linear regression using the GDP growth, unemployment rate, and inflation it was determined that for the democratic popular vote there was a significant correlation a R^2 value of 0.5807 which is below the R^2 generally required for a valid linear regression, but since social data is very difficult to model by nature due to a high variance, an R^2 of above 0.35 can be considered statistically significant for this case[14]. This regression model shows that there is correlation between the factors listed above and which candidate holds the majority for each state. This kind of model gives an idea of which factors influence the swing states for each election. Using an unemployment rate of 0.079 [18], a gdp of -0.049 [17], and an inflation rate of 0.062 [16], the linear regression model predicted a democratic popular vote of 0.5182458 for 2020.

## Conclusion

Using the data collected and the linear regression model that was created, it was determined that candidate Former Vice President Joe Biden is predicted to win the popular vote

and the electoral college vote of the 2020 election. Even though Biden is predicted to win the election we cannot conclusively say that he will as historically even if a candidate wins the popular vote the electoral college vote is the deciding factor in who will be the next president. The electoral college vote depends on the states a candidate wins. The data used in this model has many sources of error that can skew the results and it is important to understand that polling predictions are just collections of how a sample of the population is at a specific point in time during the election cycle. The results of the election may be different due to uncontrollable influences such as voter fraud, external manipulation or potential mishandling of mail in ballots. This is especially true for the 2020 election as the US continues to battle the COVID-19 pandemic. In an effort to protect the health of  at risk individuals, some may choose to avoid the polls on election day which could produce results that contrast with this model. Election predictions are done at many different points during an election year and can give an idea of which candidate the population is leaning towards and which states are poised to have a certain majority, by understanding the limitations of these predictions they may aid in determining who will win a presidential election.

## References


[1] A. Mercer, C. Deane, and K. McGeeney, “Why 2016 election polls missed their mark,” *Pew Research Center*, 14-Aug-2020. [Online]. Available: https://www.pewresearch.org/fact-tank/2016/11/09/why-2016-election-polls-missed-their-mark/. [Accessed: 28-Oct-2020].

[2] A. O'Neill, “Share of electoral and popular votes by US president 1789-2016,” *Statista*, 30-Jul-2020. [Online]. Available: https://www.statista.com/statistics/1034688/share-electoral-popular-votes-each-president-since-1789/. [Accessed: 28-Oct-2020].

[3] B. E. C. K. Y. Little, “Four of History's Worst Political Predictions,” *National Geographic*, 07-Nov-2016. [Online]. Available: https://www.nationalgeographic.com/news/2016/11/presidential-election-predictions-history/. [Accessed: 28-Oct-2020].

[4] “Case Study 2: The 1948 Presidential Election,” *math.Upenn.edu*. [Online]. Available: https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case2.html. [Accessed: 28-Oct-2020].

[5] D. Kurtzleben, “4 Possible Reasons The Polls Got It So Wrong This Year,” *NPR*, 14-Nov-2016. [Online]. Available: https://www.npr.org/2016/11/14/502014643/4-possible-reasons-the-polls-got-it-so-wrong-this-year. [Accessed: 28-Oct-2020].

[6] D. Walther, “Picking the winner(s): Forecasting elections in multiparty systems,” *Electoral Studies*, vol. 40, pp. 1–13, 2015.

[7] Fivethirtyeight, “fivethirtyeight/data,” *GitHub*. [Online]. Available: https://github.com/fivethirtyeight/data/tree/master/election-forecasts-2020. [Accessed: 28-Oct-2020].

[8]J. Gramlich, “What the 2020 electorate looks like by party, race and ethnicity, age, education and religion,” *Pew Research Center*, 26-Oct-2020. [Online]. Available: https://www.pewresearch.org/fact-tank/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion/. [Accessed: 28-Oct-2020].

[9] Master Class, “What Is an Election Poll? Understanding the Various Methods Politicians Use to Poll and Survey Voters - 2020,” *MasterClass*, 02-Oct-2020. [Online]. Available: https://www.masterclass.com/articles/what-is-an-election-poll. [Accessed: 28-Oct-2020].

[10] N. Cohn, “A 2016 Review: Why Key State Polls Were Wrong About Trump,” *The New York Times*, 31-May-2017. [Online]. Available: https://www.nytimes.com/2017/05/31/upshot/a-2016-review-why-key-state-polls-were-wrong-about-trump.html. [Accessed: 28-Oct-2020].

[11] natesilver538, “Calculating 'House Effects' of Polling Firms,” *FiveThirtyEight*, 22-Jun-2012. [Online]. Available: https://fivethirtyeight.com/features/calculating-house-effects-of-polling-firms/. [Accessed: 28-Oct-2020].

[12] O. Christie, “Why Polling Dramatically Failed to Predict the 2016 US Election.,” *Medium*, 12-Nov-2016. [Online]. Available: https://towardsdatascience.com/why-polling-dramatically-failed-to-predict-the-2016-us-election-85d344a38357. [Accessed: 28-Oct-2020].

[13] *Types of Polls*. [Online]. Available: http://oer2go.org/mods/en-boundless/www.boundless.com/political-science/textbooks/boundless-political-science-textbook/public-opinion-6/measuring-public-opinion-46/types-of-polls-269-1480/index.html. [Accessed: 28-Oct-2020].

[14] “Linear Correlation” [Online]. Available: https://condor.depaul.edu/sjost/it223/documents/correlation.html. [Accessed: 28-Oct-2020].

[15] J. E. Campbell and T. E. Mann, “Forecasting the Presidential Election: What can we learn from the models?,” 28-Jul-2016. [Online]. Available: https://www.brookings.edu/articles/forecasting-the-presidential-election-what-can-we-learn-from-the-models/. [Accessed: 29-Oct-2020].

[16] P. by E. Duffin and M. 7, “U.S. - projected inflation rate 2008-2024,” *Statista*, 07-May-2020. [Online]. Available: https://www.statista.com/statistics/244983/projected-inflation-rate-in-the-united-states/. [Accessed: 29-Oct-2020].

[17] “Bureau of Labor Statistics,” 02-Oct-2002.

[18] D. Payne, “Recovery Has Begun, but Progress May Slow,” *Kiplinger*, 27-Aug-2020. [Online]. Available: https://www.kiplinger.com/economic-forecasts/gdp. [Accessed: 29-Oct-2020].

[19] M. B. Incantalupo, Princeton University, rep., 2010.

[20] K. Amadeo, “Compare Today's Unemployment with the Past,” *The Balance*, 03-Apr-2020. [Online]. Available: https://www.thebalance.com/unemployment-rate-by-year-3305506. [Accessed: 30-Oct-2020].

# Appendix 1: Figures and Tables

| Pollster | House Effect |
|---|---|
| Pew Research | D +3.2 |
| Public Policy Polling (PPP) | D +3.1 |
| Ipsos | D +2.9 |
| SurveyUSA | D +2.4 |
| Marist (NBC/Marist) | D +1.9 |
| YouGov | D +0.8 |
| CNN (Opinion Research) | D +0.4 |
| Rasmussen Reports | R +1.3 |
| Washington Post / ABC News | R +1.4 |
| Fox News (Robbins & Shaw) | R +1.5 |
| Quinnipiac | R +1.7 |
| Gallup | R +2.5 |

**Figure 1:** Polling Bias

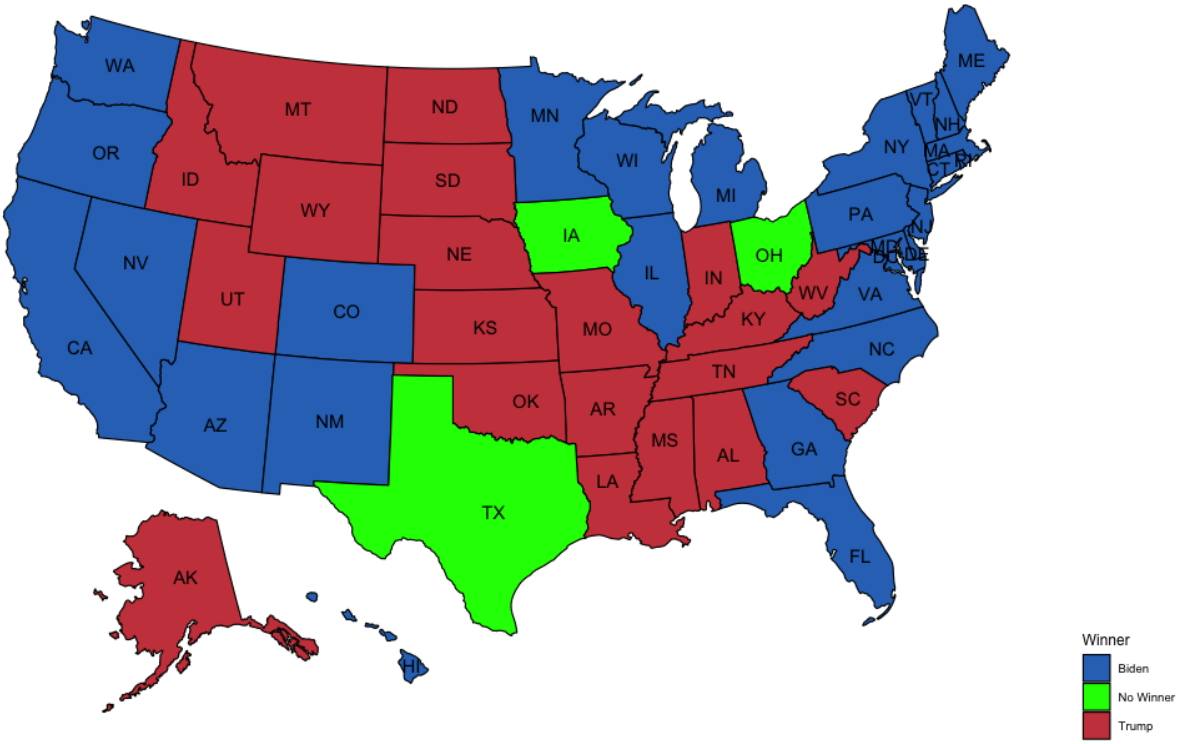


**Figure 2:** Unadjusted polling predictions by state for 2020.

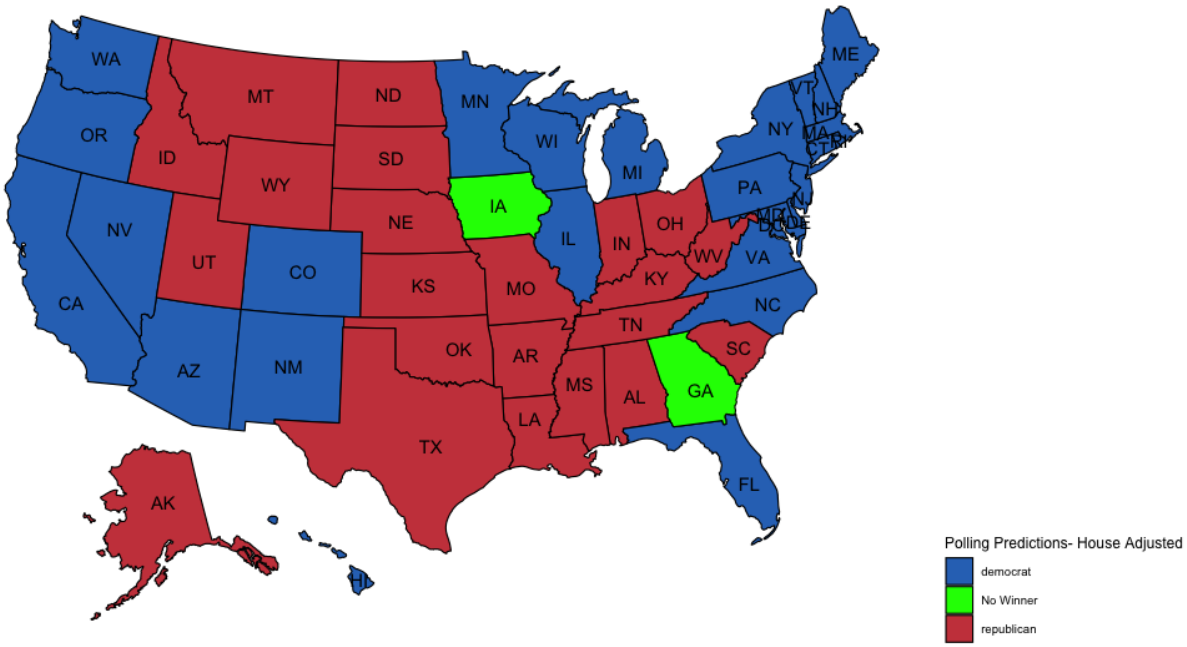


**Figure 3:** Polling predictions for 2020 by state adjusted for housing bias.

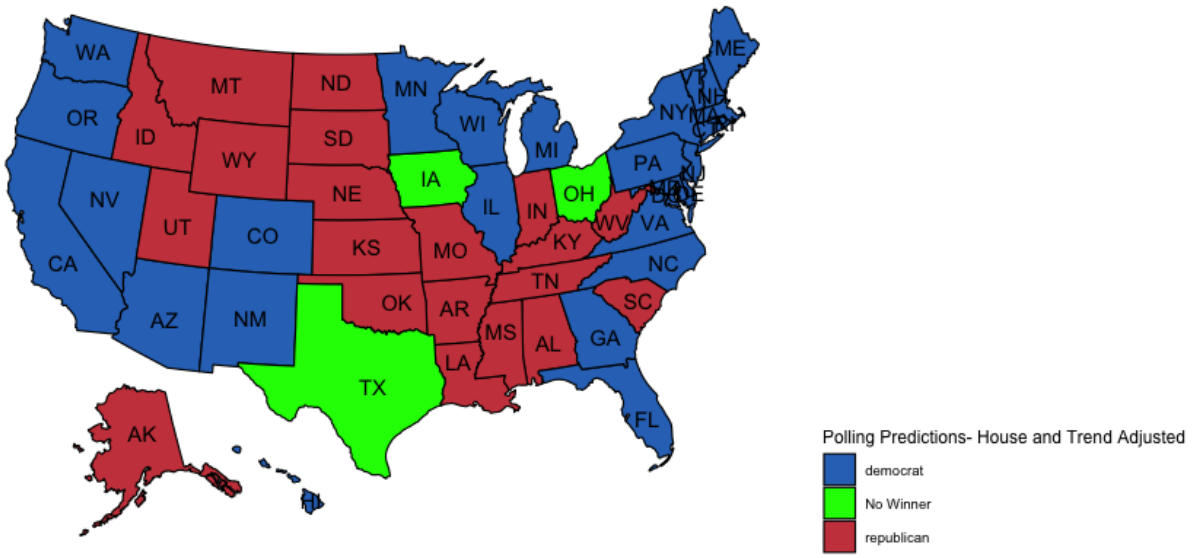


**Figure 4:** Polling predictions for 2020 adjusted by house and by trend.

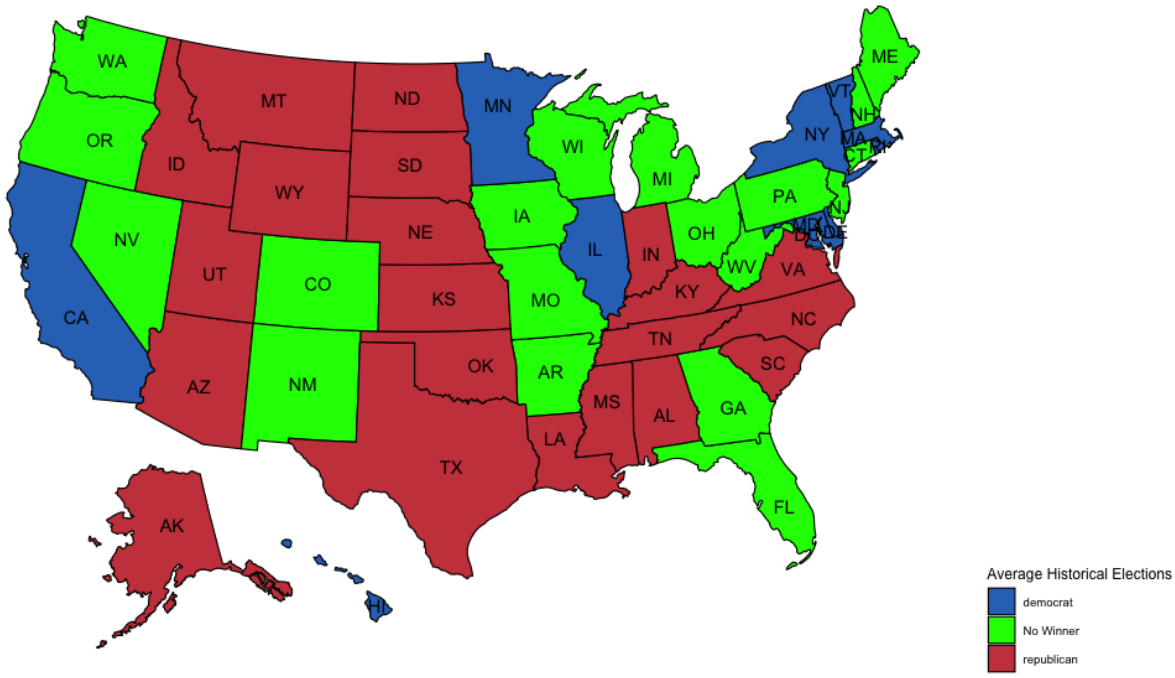


**Figure 5:** Average Historical predictions by state from 1976 to present day.

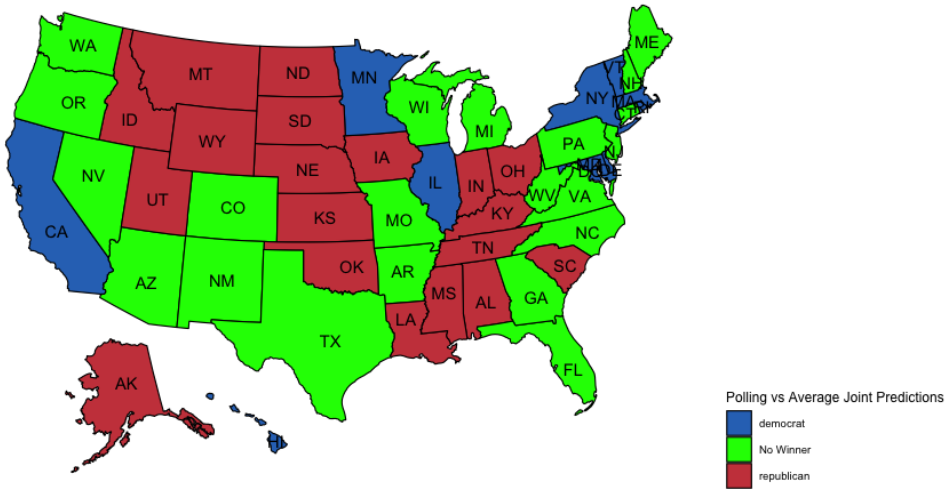


**Figure 6:** Prediction by state based on 2020 polling and Historical Popular vote data.

```
Call:
lm(formula = dem_popular ~ unemployment_rate + gdp_growth + inflation,
    data = winner_economy)

Residuals:
      Min        1Q    Median        3Q       Max
-0.042744 -0.011162 -0.003052  0.011944  0.051566

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.52372    0.04958  10.563 1.49e-05 ***
unemployment_rate 0.09922    0.71811   0.138   0.8940
gdp_growth       -0.82624    0.47863  -1.726   0.1279
inflation        -0.86775    0.30665  -2.830   0.0254 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03117 on 7 degrees of freedom
Multiple R-squared:  0.5807,    Adjusted R-squared:  0.401
F-statistic: 3.232 on 3 and 7 DF,  p-value: 0.0911

> |
```

**Figure 7:** The results of the linear regression. The dependent variable is the democratic popular vote. The independent variables are unemployment rate, gdp growth, and inflation.

**Table 1:** Number of states each candidate will win per model

| **Model** | **Democrat** | **Republican** | **Undecided** |
|---|---|---|---|
| Model 1: Percent | 27 | 22 | 2 |
| Model 2: House Adjusted Percent | 27 | 22 | 2 |
| Model 3: House and Trend Adjusted Percent | 28 | 20 | 3 |
| Model 4: Historical Percent | 11 | 21 | 19 |
| Model 5: Historical and House/Trend Adjusted Percent Joined | 11 | 19 | 21 |

**Table 2A:** From polling predictions for 2020 adjusted by house and by trend (Figure 4), electoral college votes were determined by state.

| State | Winner | Electoral College Votes |
|---|---|---|
| AK | Trump | 3 |
| AL | Trump | 9 |
| AR | Trump | 6 |
| AZ | Biden | 11 |
| CA | Biden | 55 |
| CO | Biden | 9 |
| CT | Biden | 7 |
| DC | Biden | 3 |
| DE | Biden | 3 |
| FL | Biden | 29 |
| GA | Biden | 16 |
| HI | Biden | 4 |
| IA | Undecided | 6 |
| ID | Trump | 4 |
| IL | Biden | 20 |
| IN | Trump | 11 |
| KS | Trump | 6 |
| KY | Trump | 8 |
| LA | Trump | 8 |
| MA | Biden | 11 |
| MD | Biden | 10 |
| ME | Biden | 4 |
| MI | Biden | 16 |
| MN | Biden | 10 |
| MO | Trump | 10 |
| MS | Trump | 6 |
| MT | Trump | 3 |
| NC | Biden | 15 |

| | | |
|---|---|---|
| ND | Trump | 3 |
| NE | Trump | 5 |
| NH | Biden | 4 |
| NJ | Biden | 14 |
| NM | Biden | 5 |
| NV | Biden | 6 |
| NY | Biden | 29 |
| OH | Undecided | 18 |
| OK | Trump | 7 |
| OR | Biden | 7 |
| PA | Biden | 20 |
| RI | Biden | 4 |
| SC | Trump | 9 |
| SD | Trump | 3 |
| TN | Trump | 11 |
| TX | Undecided | 38 |
| UT | Trump | 6 |
| VA | Biden | 13 |
| VT | Biden | 3 |
| WA | Biden | 12 |
| WI | Biden | 10 |
| WV | Trump | 5 |
| WY | Trump | 3 |

**Table 2B:** Total electoral college votes for each candidate and undecided states.

| **Candidate** | **Electoral College Votes** |
|---|---|
| Trump | 126 |
| Biden | 350 |
| Undecided | 62 |

# Appendix 2: R Code

Github Link: https://github.com/yedidv/Election-Predictions for raw code and Excel Files used for data analysis and model generation.

## 2.1- Objects:

```
SelectCandidateState <- function(df, party_name, state_name){
  ## Input the dataframe, the candidate we want to filter by, and the state we want to filter by.
  ## Returns a subset of the polling data that meets the specified parameters

  return(subset.data.frame(df, (party == party_name) & (state == state_name)))
}


PollingResults <- function(polling_data, adjusted){
  ### Iterate through the data, for each state collect the trends for trump, biden, and see if they
are
  ### statistically different from each other. If they are, then pick the one with the highest mean
as the winner.


  polling_winner_columns <- c("state", 'p_Value', 'winner')
  polling_winner <- data.frame(matrix(ncol = length(polling_winner_columns), nrow =
length(unique(polling$state))))
  colnames(polling_winner) <- polling_winner_columns


  i <- 1
  for (state in unique(polling_data$state)){
    trump_data <- SelectCandidateState(polling_data, unique(polling$party)[2], state)
    biden_data <- SelectCandidateState(polling_data, unique(polling$party)[1], state)

    ## Select which polling data we are using
    if(adjusted == 'trend_and_house'){
      trump_polling <- trump_data$logtrendhouse
      biden_polling <- biden_data$logtrendhouse
    }
    else if(adjusted == 'house'){
```

```
    trump_polling <- trump_data$loghouse
    biden_polling <- biden_data$loghouse
   }
   else if(adjusted == 0) {
    trump_polling <- trump_data$logpct
    biden_polling <- biden_data$logpct
   }
   else{
    return('Invalid polling data')
   }

   ## Perform t test to see if the data is significant
   p_test <- t.test(trump_polling, biden_polling)[3]
   winning_candidate <- ifelse(p_test <= 0.05, ifelse(mean(biden_polling) >
mean(trump_polling), 'democrat', 'republican'), 'No Winner')


   polling_winner$state[i] <- state
   polling_winner$p_Value[i] <- p_test
   polling_winner$winner[i] <- winning_candidate
   i <- i + 1

  }

  return(polling_winner)

}



HistoricalPrediction <- function(historical_polls){
  states <- unique(historical_polls$state)
  historical_winner_columns <- c("state", 'p_value', 'winner')
  historical_winner<- data.frame(matrix(ncol = length(historical_winner_columns), nrow =
length(states)))
  colnames(historical_winner) <- historical_winner_columns
  head(historical_winner)


  ## Create a for loop to group together the data for each state and make a prediction
  ## based on the p - value of the t - test.
  ## Box plots were made for each state's data to see if it was normal. It was shown not to be,
  ## so we took the log transform of the data to ensure the normality assumption holds when
performing
```

```
  ## the t-test.
  j <- 1
  for (state in states){
    dem <- SelectCandidateState(historical_polls, 'democrat', state)
    rep <- SelectCandidateState(historical_polls, 'republican', state)
    p_value <- t.test(dem$log, rep$log)[3]
    winner <- ifelse(p_value < 0.05, ifelse(mean(dem$percentvotes) > mean(rep$percentvotes),
'democrat', 'republican'), 'No Winner')

    historical_winner$state[j] <- state
    historical_winner$p_value[j] <- p_value
    historical_winner$winner[j] <- winner

    j <- j + 1
  }
  return(historical_winner)

}


WinnerPopular <- function(historical_polls){
  winner_columns <- c('party', 'year', 'dem_popular', 'rep_popular')
  winner_economy <- data.frame(matrix(ncol = length(winner_columns),
                                      nrow = length(unique(historical_polls$year))))
  colnames(winner_economy) <- winner_columns


  j <- 1
  for (cycle in unique(historical_polls$year)){

    winner <- subset.data.frame(historical_polls, year == cycle)
    dem <- subset.data.frame(winner, party == 'democrat')
    rep <- subset.data.frame(winner, party == 'republican')
    dem_popular <- sum(dem$candidatevotes) / sum(dem$totalvotes)
    rep_popular <- sum(rep$candidatevotes) / sum(rep$totalvotes)
    winner <- ifelse(dem_popular > rep_popular, 1, 0)

    winner_economy$year[j] <- cycle
    winner_economy$dem_popular[j] <- dem_popular
    winner_economy$rep_popular[j] <- rep_popular
    winner_economy$party[j] <- ifelse(mean(dem_popular) > mean(rep_popular), 1, 0)
    j <- j + 1

  }
```

```
  return(winner_economy)
}



UsMapPlot <- function(winning_data, data_type){
  ### US Map plot for the predictions
  library(ggplot2)
  library(usmap)
  us_map <- plot_usmap(data = winning_data, values = 'winner', regions = 'states', labels =
TRUE)
  us_map <- us_map + theme(legend.position = 'right')
  us_map <- us_map + scale_fill_manual(name = data_type,
                                       values = c('democrat' = '#2E74C0', 'republican' = '#CB454A', 'No
Winner' = 'Green'))
  return(us_map)
}



ReadCsv <- function(path){
  return (as.data.frame(read.csv(path)))
}


CountWinner <- function(data, party){
  ## Count how many states each candidate wins in the predictions.
  return(length(subset.data.frame(data, (winner == party) & (state != 'National') )$state))
}
```

## 2.2- Main Code:

```
source('finalfunctions.r')


##### INITIAL PREDICTIONS: USING POLLING TO PREDICT THE FINAL ELECTIONS
#####

### Read Polling CSV File
```

```
polling <- ReadCsv('presidential_polls_2020.csv')

## Subset the Dataframe to remove NE-1, NE-2, ME-1, ME-2
polling <- subset.data.frame(polling,  !(polling$state %in% c('NE-1', 'ME-2', 'NE-2', 'ME-1')))
tail(polling)


polling$party <- ifelse(polling$candidate_name  == unique(polling$candidate_name)[1],
'democrat', 'republican')
polling$logpct <- log(polling$pct)
polling$loghouse <- log(polling$house_adjusted_pct)
polling$logtrendhouse <- log(polling$trend_and_house_adjusted_pct)



### Iterate through the data, for each state collect the trends for trump, biden, and see if they
are
### statistically different from each other. If they are, then pick the one with the highest mean as
the winner.
### Start with percentages
polling_pct_winner <- PollingResults(polling, 0)
UsMapPlot(polling_pct_winner, 'Polling Predictions')

## Count how many states each candidate won.
dem_winner_count <- CountWinner(polling_pct_winner)
rep_winner_count <- CountWinner(polling_pct_winner)


## House Adjusted Polling Averages
polling_house_pct_winner <- PollingResults(polling, 'house')
UsMapPlot(polling_pct_winner, 'Polling Predictions- House Adjusted')

## Count how many states each candidate won.
dem_winner_house_count <- CountWinner(polling_pct_winner)
rep_winner_house_count <- CountWinner(polling_pct_winner)


## House and Trend Adjusted Polling Averages
polling_house_trend_pct_winner <- PollingResults(polling, 'trend_and_house')
UsMapPlot(polling_house_trend_pct_winner, 'Polling Predictions- House and Trend Adjusted')

## Count how many states each candidate won.
dem_winner_house_trend_count <- CountWinner(polling_pct_winner)
rep_winner_house_trend_count <- CountWinner(polling_pct_winner)
```

```
##### Prediction 2: Using Historical Data to Try and Predict Elections #######
historical_polls <- ReadCsv('1976_2016_president.csv')
head(historical_polls)
## Convert to percentages of votes per state to make data comparison easier.
historical_polls$percentvotes <- historical_polls$candidatevotes / historical_polls$totalvotes

## We determine that any candidate with less than 5% of the votes is insignificant towards the
final results.
historical_polls <- subset.data.frame(historical_polls, percentvotes > 0.05)
historical_polls$log <- log(historical_polls$percentvotes)
parties <- unique(historical_polls$party)
head(parties)
head(historical_polls)

### We can see we have election results for all years since 1976.
## We can put this data through the HistoricalPrediction function created
## which would return the predicted winner for each state based on historical data
historical_winner <- HistoricalPrediction(historical_polls)
UsMapPlot(historical_winner, 'Average Historical Elections')

## Count how many states each candidate won.
dem_winner_house_historical_count_ <- CountWinner(historical_winner)
rep_winner_house_historical_count <- CountWinner(historical_winner)

## In order to find states we are sure will go to Trump or to Biden, we can
## merge the two tables of the winner of the polls and of this historical polls
## in a full outer join.
average_polling_merge <- merge(x = polling_house_trend_pct_winner,
                        y =  historical_winner, by = 'state',
                        all = TRUE)
head(average_polling_merge)

## This determines the states where both the polling and the historical data predict
## the same winner, meaning we can say with high confidence that these states are
## accurately predicted.
average_polling_merge$winner <- ifelse((average_polling_merge$winner.x ==
average_polling_merge$winner.y),
                              ifelse((average_polling_merge$winner.x == 'democrat'),
                                    'democrat',
                                    'republican'),
                              'No Winner')
UsMapPlot(average_polling_merge, 'Polling vs Average Joint Predictions')
```

######## Prediction 3: Using health and economic data to predict the outcomes for the remaining states. #######

```
## Read unemployment file - contains unemployment, gdp, and inflation rates for every year
unemployment <- ReadCsv('unemployment.csv')
head(unemployment)

## Unemployment data during election year. Year must be a multiple of 4 and after the year 1975
## (to fit with the historical data set)
unemployment_electionyear <- subset.data.frame(unemployment, (year %% 4 == 0) & (year > 1975) )
head(unemployment_electionyear)


## Put through a function that outputs a dataframe with the popular vote winner for each election,
## what the popular vote was, and the election cycle.
winner_economy <- WinnerPopular(historical_polls)

## Merge data for the winner with the data for the
winner_economy <- merge(x = unemployment_electionyear,
                        y = winner_economy,
                        by = 'year',
                        all = TRUE)


head(winner_economy)

## We find there is a moderate correlation between the democratic popular vote and the unemployment factors
summary(lm(dem_popular ~ unemployment_rate + gdp_growth + inflation, data = winner_economy))
summary(lm(rep_popular ~ unemployment_rate + gdp_growth + inflation, data = winner_economy))
summary(lm(party ~ unemployment_rate + gdp_growth + inflation, data = winner_economy))

dem_lm <- lm(dem_popular ~ unemployment_rate + gdp_growth + inflation, data = winner_economy)
```

```
dem_winner_pct <- CountWinner(polling_pct_winner, 'democrat')
rep_winner_pct <- CountWinner(polling_pct_winner, 'republican')
dem_winner_house  <- CountWinner(polling_house_pct_winner, 'democrat')
rep_winner_house  <- CountWinner(polling_house_pct_winner, 'republican')
dem_winner_trend_house  <- CountWinner(polling_house_trend_pct_winner, 'democrat')
rep_winner_trend_house  <- CountWinner(polling_house_trend_pct_winner, 'republican')
dem_winner_historical  <- CountWinner(historical_winner, 'democrat')
rep_winner_historical  <- CountWinner(historical_winner, 'republican')
dem_winner_trend_house_historical  <- CountWinner(average_polling_merge, 'democrat')
rep_winner_trend_house_historical  <- CountWinner(average_polling_merge, 'republican')
```