Exploring the Relationship between Candidate Endorsements, Demographics, Financial Contributions and their Effect on Elections in the United States

Victoria Austin, Rebecca Li, David Xu, Robert Schwartz

Introduction

One area that statistics and data have always played a prominent role in the United States occurs every 4 years. Once every 4 years, many Americans, polling groups, and parties value insights about different political candidates and their odds of winning. Our team chose to focus on political data because of the importance of politics and we wanted to make the data more understandable for all. We used 2 different datasets for our analysis, one of which was collected by 538 and the other was collected by the Federal Election Commision.

Data Overview

The 538 dataset contains various demographic features of primary candidates in the Democratic races, as well as endorsement data. Some examples include whether or not a candidate was an Obama administration alumni, whether they were a gun sense candidate and whether or not they were endorsed by big name figures. The dataset by itself was kind of barebones especially considering not too many candidates had endorsements. Because of this we decided to introduce a second dataset in order to study and ask more advanced questions about the nature of US elections. This second dataset was from the FEC and tracks financial data of all primary candidates. While there were plenty of features this dataset introduced, the ones we were really focused on were about a candidate's total external and internal contributions. The reason being we don't have the domain expertise to do financial analysis on various tax and financial information, but total contribution was something we could understand and likely affected election outcomes. One unfortunate result of merging these two datasets was that the financial dataset only contained data on primary candidates running for House of Representatives. So our data became even more narrow in scope because we were no longer looking at Governor or Senate elections.

Data Scope and Concerns

The final merged data that we worked with represents a census of around 300 candidates running in the 2018 Democratic primaries. The information in the data was already publicly available, so all participants were aware of the data collection and the possibility of others using this data for inference or other purposes. Since our merged data was specifically focused on primary results rather than the Republican vs Democrat debate, we really weren't worried about the issues and bias that might come up from excluding Republican data. Our reasoning for choosing the Democrat data over the Republican data was centered around the fact that the Democrat data seemed more complete, and we were worried that the Republican data did not have sufficient demographic features needed to construct a useful model for prediction. The dataset had pretty high granularity, to match the granularity of our research questions. Each row of the data looked at each primary candidate and their campaign financials and various features like demographics and endorsement information. With the scope of the data in mind, we do need to be careful to not make over generalizations about entire groups based on our individualized granularity levels to avoid ecological fallacies. Our biggest concerns with the dataset is the fact that for a lot of our demographic and endorsement

data, there wasn't much data in the first place, meaning that our models may not be as strong and generalizable as we would like. This was solely due to the fact that endorsements are already quite rare, not every candidate running for House is going to get a Biden endorsement. We weren't too worried about the endorsement data being incorrect as most candidate endorsements are public and we doubt that a candidate would be hiding an endorsement. Even though our data allows us to do interesting analysis on Democat election data, we also wish that the Republican data had more demographic features as it would allow us to generalize our results for more candidates and gain more insight on the nature of elections in the United States.

Research Questions

In the beginning of our project, we began exploring some of the relationships between different variables, and decided to explore and focus on the following key questions:

GLM vs Black Box

1. How accurately can we predict total external contribution for a candidate running in the Democrat's primary given different demographic features such as veteran, STEM, endorsements, etc. using a parametric GLM compared to a black box model such as a random forest?

Answering this question provides a closer understanding as to why certain candidates receive more external contributions over others. Knowing whether a candidate of interest could have monetary disadvantages or advantages based on their demographics can help donors decide who they want to donate to and how much they want to donate. A GLM is appropriate for answering this question since it provides a more flexible generalization of linear regression. We are interested in comparing the accuracy between a parametric and non-parametric model since an accurate parametric model could provide more interpretability, but we don't want to be limited to a fixed number of parameters if it results in higher predictive accuracy.

Causal inference

2. Does a large total external contribution cause an increase in the probability a candidate will win in their primary?

If a large total external contribution is positively correlated to a candidate's election results, this would provide further motivation for donors to decide to donate to their desired candidate. Especially if a candidate has undesirable demographic features that might limit his or her total contribution and reduce the likelihood of winning. Causal inference is important for answering this question since we want to know if a large total external contribution *causes* an increase in win probability, independent of any confounding factors.

EDA

Data Cleaning

We cleaned our data based on the granularity, scope, and faithfulness of the data and also observed other issues with our data that we addressed during cleaning. Before working with our combined dataset, we used string manipulation on the candidate names in order to merge our financial data with our election endorsement data. There were a lot of missing values for demographic and endorsement variables and instead of removing these rows, we decided to replace the missing values with “No” since we inferred that the lack of a “Yes” under an endorsement for example, meant that the candidate did not get the endorsement. By doing this however, we must acknowledge that our data now groups candidates that were anti-endorsed with candidates that were simply not endorsed, which can introduce bias in our model and muddle the effects of endorsement on election outcomes. Next, we one-hot-encoded key categorical variables for candidate demographics and financial information.

Exploring Features

While our main questions center around how finances influence race results, we were just as curious to see if there were other factors that might be influencing races. Our data had many categorical features so we decided to calculate the different correlations between these features and primary vote percentage. Essentially, a higher correlation meant that there is an association between a categorical variable and an outcome variable, like primary vote percentage. As seen in Figure 1 below, it seemed like demographic features like whether or not a candidate was a veteran or LGBTQ didn’t influence results as much as party endorsements. Our correlations indicated that a candidate endorsed by a big party name, like Joe Biden, Emily Warren, or a party endorsement in general was more associated with winning a primary compared to non-endorsed candidates. This seems to be what we expected, and it does seem to mean that we could build some models to predict whether a candidate won or lost, or their primary vote percentages. The party endorsement seemed the most effective as it either meant the candidate was placed on the DCCC’s Red to Blue list before the primary, was endorsed by the DSCC before the primary, or if the DSCC/DCCC aired pre-primary ads in support of the candidate. Otherwise the candidate was specifically anti-endorsed.

One particular column that stood out was partisan lean. Partisan lean is described as being the average difference between how a state or district voted in the past two presidential elections and how the country voted overall, with 2016 results weighted 75 percent and 2012 results weighted 25 percent. So, it didn’t seem to make sense that a higher partisan lean was negatively correlated with winning primaries. Upon further analysis it seemed like there were a couple fallacies and confounding factors that needed to be accounted for. One theory we had for the negative correlation between partisan lean and primary vote percentage was that maybe there were many more candidates running in Democratic leaning states and so there were going to be more candidates competing and dragging the average primary percentage down. This ended

up providing some explanation as was the case in Alabama vs. California (Figures 2 and 3). The overall party lean in California is significantly lower in magnitude than in Alabama (Figure 4) but because there were more candidates running in California, the overall primary vote percentage would be lower. This is a huge confounding factor that influences both our models and our causal inference question. We'll dive a little deeper into it but we added a column that measured how many competitors a candidate faced.

Another huge issue we faced was whether or not our regression focused questions were viable. For example, while Biden endorsements had a high correlation with total external contribution, there were so few Biden endorsements it would be hard to fit any regression model to predict external contribution. One consideration we started thinking about during EDA was whether it would be smarter to change our question to a classification problem. If Biden endorsements are correlated with higher external contributions, can we classify which candidates are going to have a high contribution? (Rather than trying to predict the exact external contribution) Again, this is something we cover later but it was something that we had to account for as early on as in EDA.

Prediction with GLMs and Non-Parametric Methods

Methods

We are trying to predict a candidate's total external contribution using whether the candidate had party support, if they were endorsed by Emily, and if they were endorsed by Biden. We chose these features since they had the highest correlation to total external contribution. We found that if we added more features, there was too much noise in the data, which led to inaccurate predictions.

We decided to use a Gaussian GLM using Frequentist methods. We chose a Gaussian distribution since it was the best fit out of a series of distributions, with the lowest log-likelihood. Additionally, when looking at the histogram of total external contribution, the data appeared as a rightly-skewed distribution with most of its mass concentrated at lower values. Since we do not know the distribution of total external contribution, we used the central limit theorem to assume that with enough samples, the distribution is roughly normal. Under a GLM model, we also assumed independence between data points, homoscedasticity of variance, and that the error had a normal distribution independent of other points. We evaluated the performance of our model by analyzing the log-likelihood and Pearson chi-squared error to determine goodness of fit.

For the non-parametric method we chose a random forest since it is a versatile model that incorporates randomness to avoid overfitting. Random forests assume that splits are made based on the residuals for predicting that node by minimizing the mean squared error. We evaluate the performance of our model by calculating the mean accuracy of our test data.

Results

Our Gaussian GLM had a log-likelihood of -5882.6 and chi-squared error of 9.27e+14. Unfortunately, these are not great results and a Gaussian GLM appears to not be the best model for predicting total external contribution from demographic data. The variance of our model is too high suggesting that we are overfitting our model to the random noise that is present in our training data. Therefore, any interpretation of our results is not necessarily representative of the relationship between endorsements and contribution, and rather representative of how we might've tweaked the model to find a better fit for our data. Based on our results, for every unit of increase in total external contribution, *Party Support?* increases by 3.143e+06%, *Emily Endorsed?* increases by 1.523e+06%, and *Biden Endorsed?* increases by 4.41e+06% on average. *Party Support?* values can range from 2.784e+06% to 3.502+e06%, *Emily Endorsed?* values can range from 1.219+e06% to 1.827+e06%, and *Biden Endorsed?* values can range from 3.699+e06% to 5.121+e06%. For our random forest, we achieved an accuracy of 50%.

Discussion

Our random forest performed better than our GLM at predicting Total External Contribution. We believe that this is due to not having a large enough sample as well as not enough candidates having any endorsements. This made it difficult to fit our model. Predicting a continuous variable with too few data points is difficult using a random forest and we recommend using classification to improve accuracy for applying our model to future datasets. For example, when we made Total External Contribution a categorical variable where a "high" contribution was above the median and a "low" contribution was below the median, we achieved an accuracy of 70%.

Causal Inference

Methods

Our treatment variable was "Contribution Level", which is defined as 1 if a candidate's total external contribution was above the median and 0 if it was below. Our treatment variable is observational, and was not assigned to a candidate randomly. The outcome variable is "Won Primary", which is either a 1 if a candidate won or 0 if the candidate lost their primary. Initially, our outcome variable was the proportion of votes a candidate received, but after further analysis, we decided not to use this variable since one candidate's percentage of votes could affect another candidate's percentage if they are in the same race, leading to biased results. Our confounding variables were 'Veteran?', 'LGBTQ?', 'Elected Official?', 'Self-Funder?', 'STEM?', 'Obama Alum?', 'Party Support?', 'Emily Endorsed?', 'Guns Sense Candidate?', 'Biden Endorsed?', 'Warren Endorsed? ', 'Sanders Endorsed?', 'Our Revolution Endorsed?, 'Justice Dems Endorsed?', 'PCCC Endorsed?', 'Indivisible Endorsed?', 'WFP Endorsed?', 'VoteVets Endorsed?', 'No Labels Support?'. The unconfoundedness assumption does not hold for our model since all the variables that could possibly affect our treatment and outcome variables can not be observed and controlled for. In order to control for confounding variables we used a

propensity score model. This allows us to estimate treatment effect accounting for propensity. To do this we fit a logistic regression model that predicts our treatment variable based on our confounders. We used the fitted model to predict the probability for each sample for each feature. We incorporated this value in our average treatment effect estimate.

Results

Initially, we calculated an average treatment effect, ATE, of 0.4182. This would suggest that total external contribution has a small positive causal effect on the probability that a candidate running for Representative will win in his or her primary. However, we were concerned that we were falling into Simpson's paradox. We considered that the ATE may be affected by the size of the race the candidate was running in, since candidates in larger races have more competition, and are more likely to have a lower primary percentage and more losses and a higher total external contribution. We broke the data into four groups where each group represented the 25th, 50th, 75th, and 100th percentile of the number of candidates running in a given race. We then performed the same methodology and estimated the ATE. Candidates in a race size in the bottom 25th percentile had an ATE of 0.4365. Candidates in a race size in between the 25th and 50th percentile had an ATE of 0.3550. Candidates in between the 50th and 75th percentile had an ATE of 0.4711. Candidates in between the 75th and 100th percentile had an ATE of 0.1373. By grouping the data, it appeared that total external contribution had a stronger positive effect on winning in races that are small to medium/large in size than for very large races.

In "Measuring Campaign Spending Effects in Post-Citizens United Congressional Elections", it was found that spending didn't effect wins on incumbents and that the effects were unclear for challengers (Barutt & Schofield, 2016). It was suggested that candidates that are more likely to win are the ones that attract money and not the other way around. Therefore, there is opposing evidence to cast doubt on whether having a high total external contribution causes a candidate to win. There are also other confounding variables that could affect both our treatment and outcome variables that were not observed in our data and are unaccounted for in our analysis. This gives further uncertainty on the accuracy of our causal effect.

Discussion

We experienced Simpson's paradox during our analysis. When all race sizes were combined we experienced an ATE that was higher than it was for races between the 25th to 50th percentile and 75th to 100th percentile. With the most notable disparity for very large races. Since all of our data was observed, it is impossible to control for all potential confounders that could affect election outcomes. Additional data that would be useful would be other races beyond representative positions, financial and election data for Republicans, and elections not just from 2018. Due to not many candidates having a high total external contribution, having candidates from larger and smaller races would be helpful to further analyze the causal relationship between different race sizes and funding categories. Our results indicate that there is a small positive causal relationship

between total external contribution and election wins. However, given the limitations of our observational data, we are not confident that we controlled for all potential confounding variables and suggest further investigation to determine causality.

Conclusion

Our first research question was to discover how accurately we could predict total external contribution given endorsement features such as Party Support? and Biden Endorsed? using a GLM and non-parametric methods. We found that a GLM was not a good fit for this research question given our data. There weren't enough candidates with endorsements, leaving very little data to predict total external contribution from. The variance in our model was too high, suggesting that we were overfitting to the random noise that was present in our data. Our random forest using the same features performed at around 50% accuracy. In order to improve the accuracy of our model for future data, we suggest making total external contribution a categorical variable and use classification rather than regression for prediction. With more data and using a random forest classifier, accurate predictions could help to inform donors which candidates will receive high external contribution given their endorsements in order to make data driven decisions about who they want to donate to and how much.

Our second research question was to find if there was a causal relationship between total external contribution and primary wins. We found that total external contribution has a small positive average treatment effect, suggesting there is a small positive relationship between our treatment and outcome variables. However, while we were able to control for confounding factors that were present in our data, there are many more confounders that were unobserved that could lead our finding to be inaccurate. Past research suggests an unclear relationship between candidate spending and election results, suggesting that there is more to the story than our data was able to provide. For future studies, we encourage incorporating more confounding variables into the model. We also suggest further exploration of the effect of election race size on our treatment and outcome variable, as it appears there may be disparities in the magnitude that total external contribution affects the increase in probability of a primary win.

Merging candidate election outcomes and demographic data with financial data is a powerful combination that can help us learn more about how money affects election outcomes. This relationship is important for U.S. politics since as a country, it values being able to achieve your dreams given hard work, regardless of how much money you have or where you come from. The U.S. has a history of placing people in power who come from wealthy and privileged backgrounds. Being able to quantitatively measure the magnitude that this affects election outcomes could influence how future U.S. elections are organized and designed in order to create a more even ground for all candidates to be able to win. If future studies are able to show that we can predict a candidate’s total external contribution given their demographic features and that there is a causal relationship between total external contribution and election outcome, this analysis would be the grounds for reimagining how democratic elections in the U.S. should be run.

Appendix

Figure 1

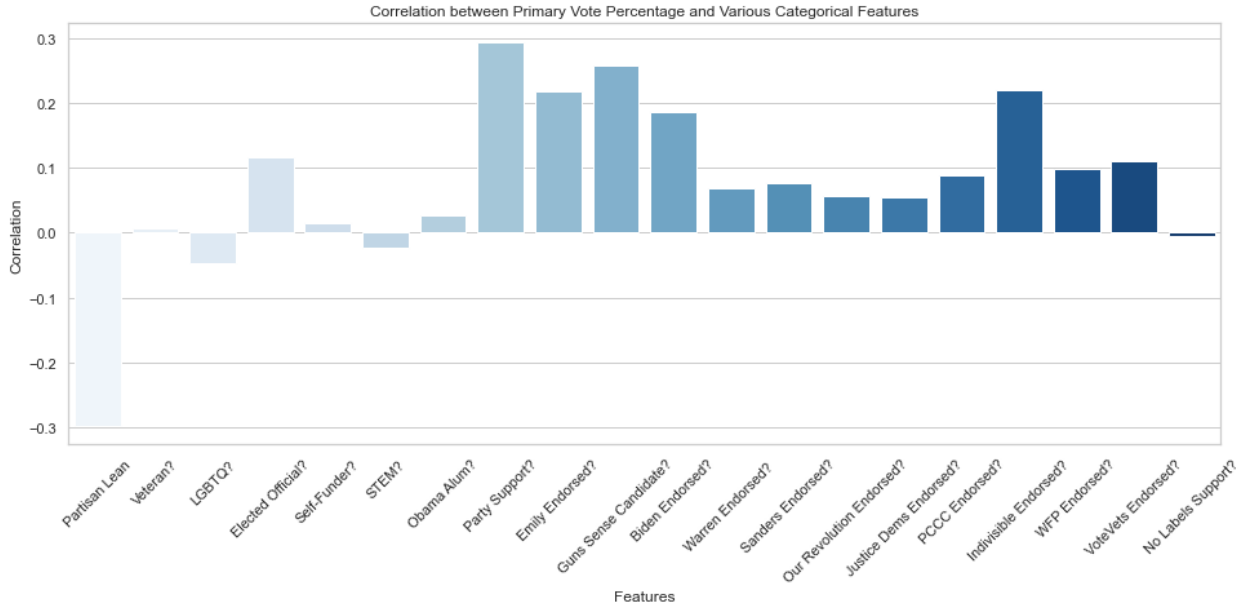


Figure 2

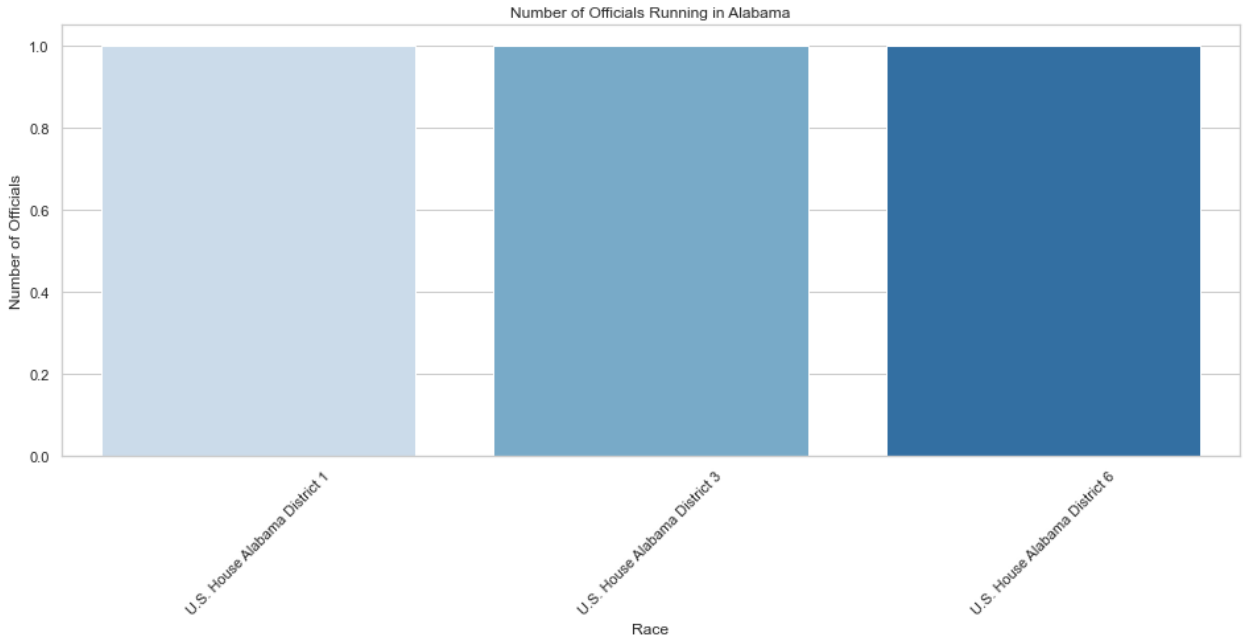


Figure 3

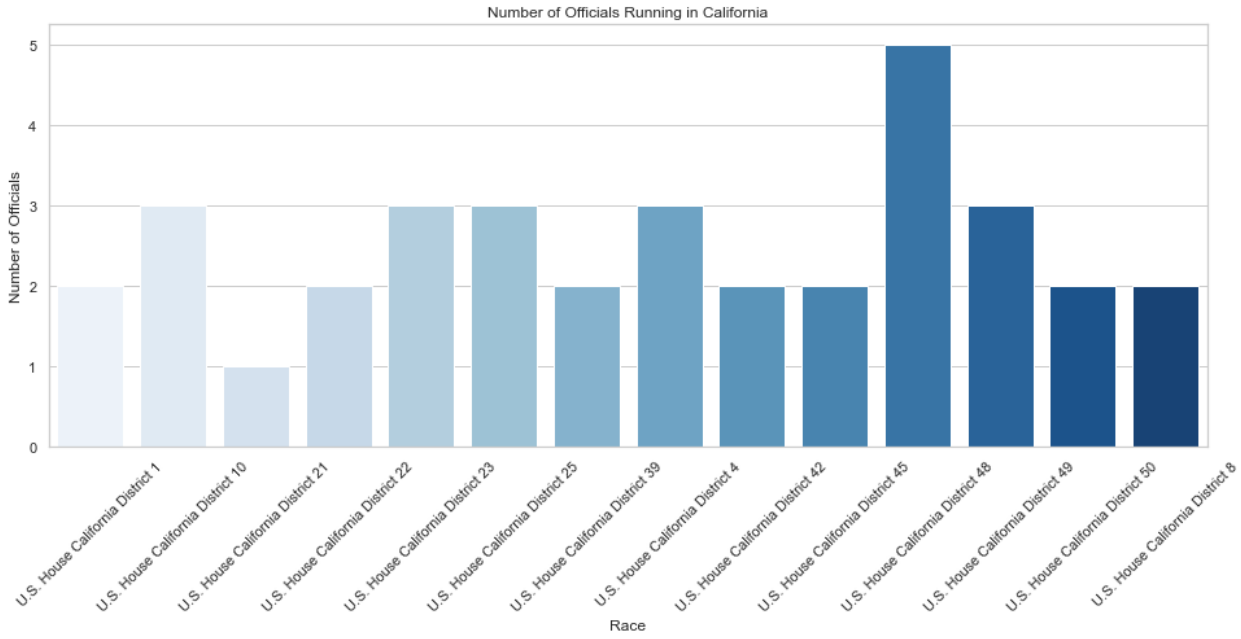
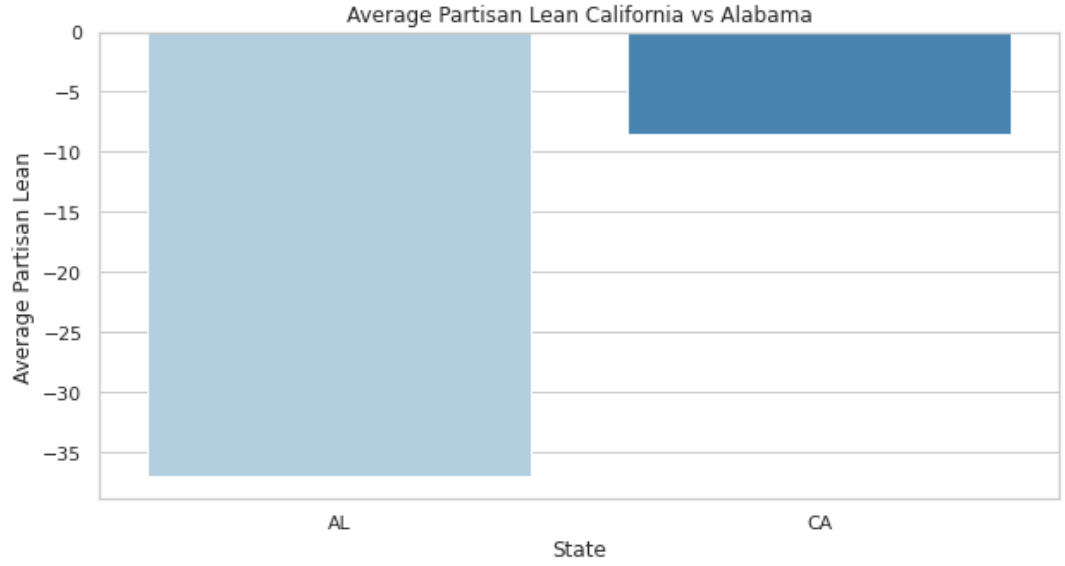


Figure 4



Bibliography:

Barutt B., Schofield N. (2016) Measuring Campaign Spending Effects in Post-Citizens United Congressional Elections. In: Gallego M., Schofield N. (eds) The Political Economy of Social Choices. Studies in Political Economy. Springer, Cham. https://doi.org/10.1007/978-3-319-40118-8_9