DATE 102 FINAL PROJECT:
2018 Primary Election Analysis

Angelo Punzalan, Leslie Romo, Carrie Hu, and Reynolds Zhang

In an era of highly polarized elections, the 2018 United States Primary Elections were the arena for the battle of the political direction of the United States. We believe that the primary season held valuable information for the political mood of the nation and looked at different political endorsements as a measure of the types of candidates Americans wanted to advance to higher levels of government. In this paper, we investigate the impact that public endorsements and identity signifiers on the electability of primary candidates.

Data Overview

How was the data collected?

The key dataset that was being utilized for this project was cleaned and assembled by FiveThirtyEight. The dataset was primarily sourced from Ballotpedia, candidate websites, FEC fundraising data, and other reliable sources. Subsequently, endorsements and support details were gathered from the respective candidate or organization who endorsed the candidates.

This data represents a census. There are a finite number of people who run for elected office during each election. The generalizability of subsequent conclusions from this dataset will be limited to how Democrats and Republicans performed in the 2018 primary elections. Any claims made beyond 2018 will be extrapolated trends.

Row Representation

Each row of the dataset represents an individual candidate who participated in the specified primary elections for the U.S. Senate, U.S. House, and governor races. Each row includes various attributes (columns) providing details about the candidate, such as their name, state, district, office type, race type, election dates, primary status, endorsements, and other relevant information.

Granularity

The granularity allows us to analyze the 2018 primaries at the level of individual candidates which will help us study the characteristics, performance, and outcome of each candidate. It will also allow us to see patterns and trends that occur across groups of candidates such as in multiple hypothesis testing when looking into those who identified as part of the LGBTQ community.

Selection Bias, Measurement error, Convenience sampling?

For the data that is available, there is limited reason to be concerned about the impact of selection bias, measurement error, or convenience sampling. First, selection bias is not an issue because we are working with a census of politicians running for office in 2018. Issues of convenience sampling are also not relevant because the dataset is a census. Finally, issues of measurement error could be an issue but should be tempered. This dataset holds a set of features that establish whether groups endorsed particular candidates, but the information was sourced based on the information those groups published.

Excluded Groups?

There are no groups that were systematically excluded from the data because the dataset holds information about every candidate in both the Republican and Democratic primaries. It is a survey of people who ran for office, the quantity is finite and known.

Data Accessibility Awareness

Candidates will be aware that data relevant to their campaign will be easily accessible to the public. This can primarily be seen through federal campaign law, as the Federal Election Commission tracks campaign financing for all federal campaigns and makes that information readily available to the public to increase accountability and transparency. Additionally, the data compiled by FiveThirtyEight is related to the public endorsements of candidates and political organizations who are hoping that their notoriety or public reputation will influence the outcome of the election. Ultimately, all information being accessed in this project is information that is freely available to the public.

For the scope of the questions we were trying to answer, we didn't seek out any external data sources.

Was Data Modified?

The dataset was not modified for differential privacy. As mentioned earlier, politics is an area where privacy of the units, in this case, candidates, are uniquely denied privacy from their beliefs and actions. Thus, it is not integral for the dataset to be changed to preserve privacy.

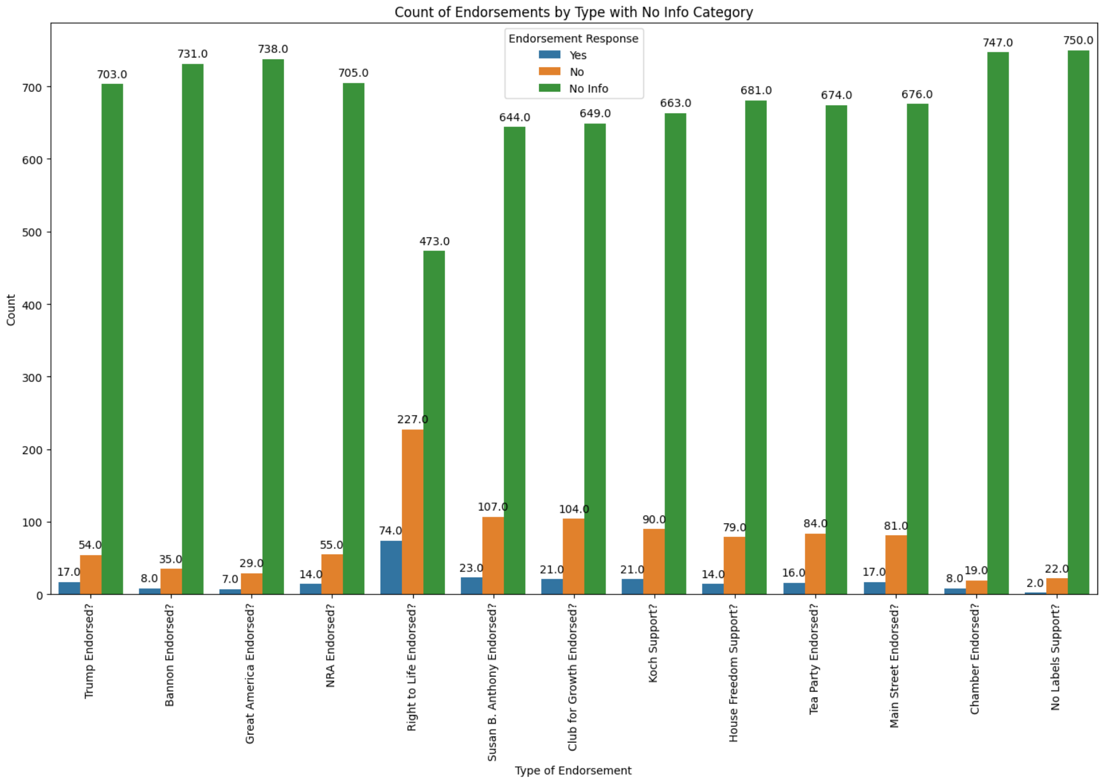
One issue we came across while doing the multiple hypothesis testing is that the republican dataset had a fairly high number of missing values for columns that could be useful. These columns include but are not limited to "Bannon Endorsed?", "Great America Endorsed?", "NRA Endorsed?", "Right to Life Endorsed?". Having so many missing values for these columns did not allow us to test a few hypotheses that we had set at the beginning of the project. Fortunately, we were able to work around this inconvenience but it would have been nice to have fewer missing values. There are a few columns that were not included in the datasets that would have been nice to have access to such as age, ethnicity, and voter turnout. We think these columns could have answered some important and interesting questions regarding the 2018 primaries.

Missing Data

The concept of this dataset was a simple tally. FiveThirtyEight was interested in tracking the kinds of endorsements that were publicly available. This was relayed into the dataset as a boolean variable. Besides attributes like partisan lean and primary percent, the dataset includes boolean values for whether or not a candidate was endorsed by a particular political entity or whether or not they identified as part of a specific group. If a candidate has a false value for a candidate, it's unclear what level of disapproval the political organization has toward a candidate. Some organizations choose to endorse multiple candidates in one race

- Some organizations do not endorse a candidate in every race.
- For some, a lack of endorsement is an admonishment of the other candidates, and for some organizations, they endorse based on a set of preferences.
- There was also an issue where data wasn't entered for particular candidates.

Most of the missing values were seen within the Republican dataset and we worked around it by constructing a binary treatment variable that consolidates all endorsements. By interpreting non-responses as a lack of endorsement, we transformed potentially ambiguous missing data into a clear indicator of whether a candidate received any endorsement. The 'Won Primary' outcome was similarly encoded, treating any missing values as an indication that the candidate did not advance. This pragmatic approach allowed for a straightforward assessment of the influence of endorsements on primary election outcomes while maintaining the integrity of the analysis amidst data gaps.



Count of Endorsements by Type with No Info Category: This visualization offers a detailed breakdown of the number of candidates who have received specific high-profile political endorsements, those who have not, and those for whom there is no information. Each bar represents one type of endorsement, such as those from Trump, Bannon, and Great America, which are central to the treatment variable in the causal study.

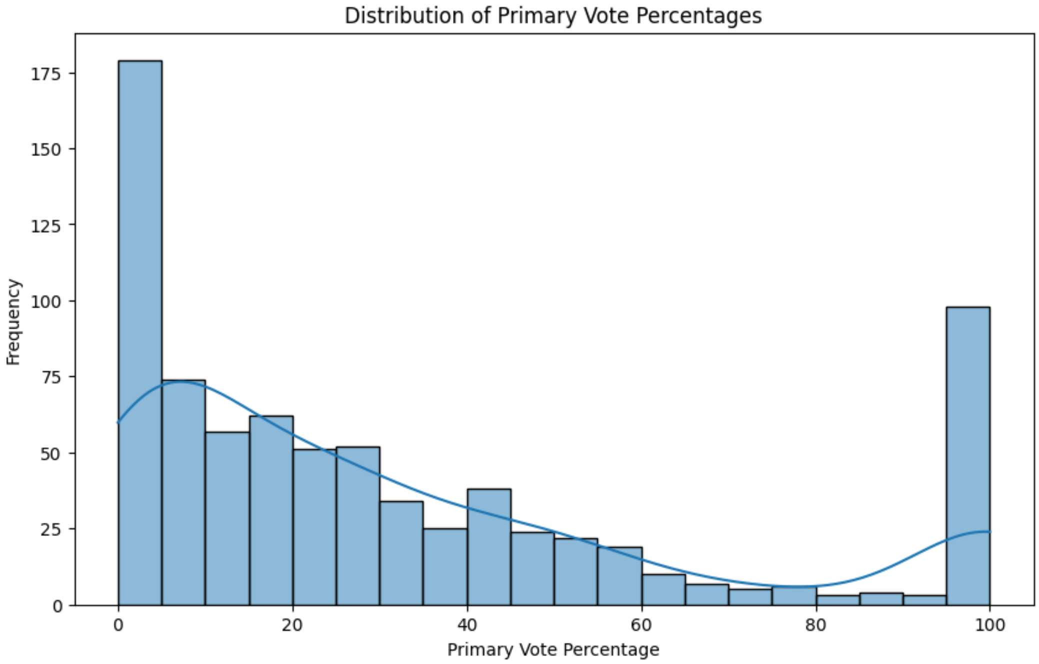
Data Cleaning

In preparing the data for analysis, we applied several cleaning and pre-processing steps. First, we addressed missing values, particularly in the 'Won Primary' outcome variable, ensuring that our models operated on complete cases. We then combined various endorsement-related columns into a single binary treatment variable to simplify the analysis and focus on the presence of any high-profile endorsement. This binarization was crucial to standardize the treatment across diverse endorsement types, facilitating the implementation of propensity score matching. Additionally, we encode categorical variables, such as 'Primary Status', into binary formats suitable for logistic regression. These preprocessing decisions, while streamlining the data for our specific analytical needs, potentially limited our ability to distinguish between the effects of different types of endorsements or to capture more nuanced relationships in the data. However, they were essential for the clarity and focus of the study, targeting the primary research question about the overall impact of political endorsements.

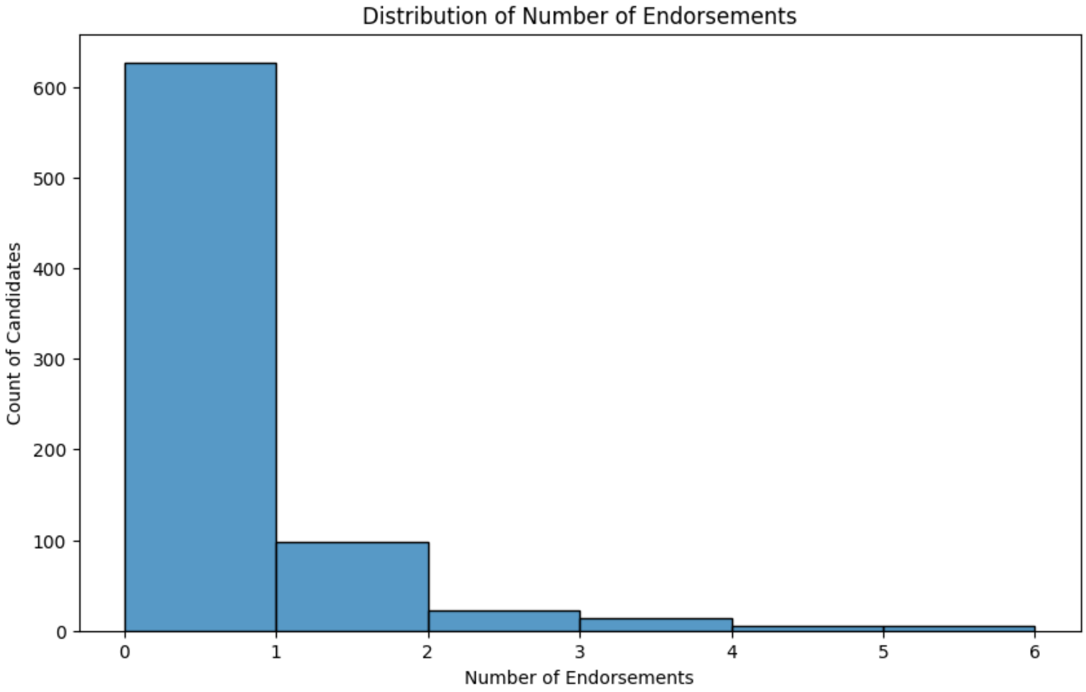
EDA

Visualizations and Their Relevance to Research Questions

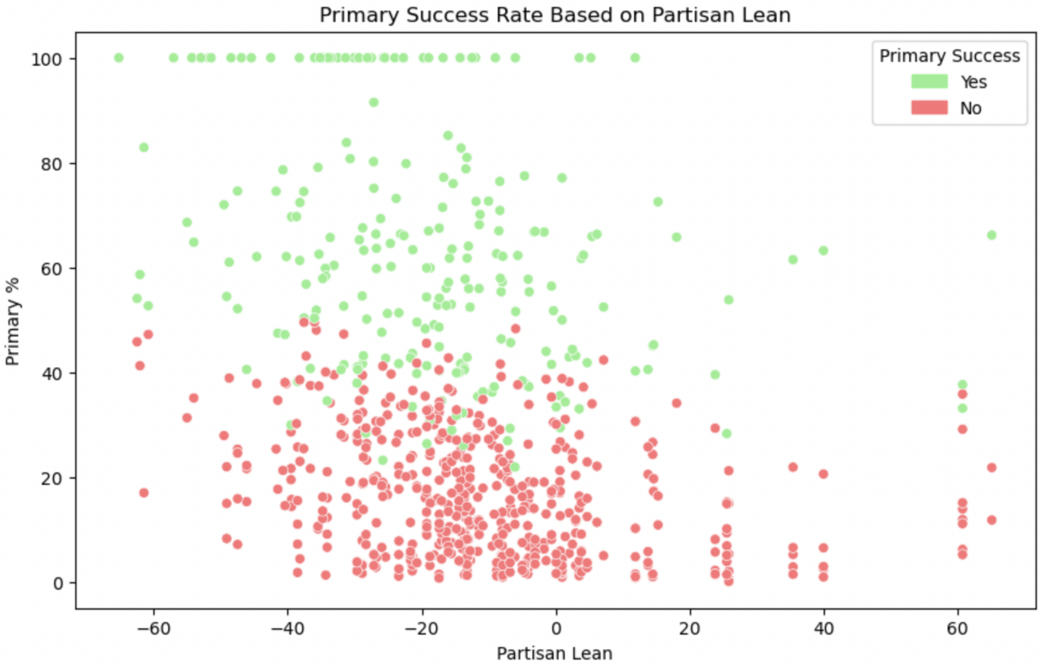
Visualization of Quantitative Variables:



Distribution of Primary Vote Percentages: This histogram reveals the distribution of the percentage of votes received by candidates in primaries. The bimodal distribution, with peaks near 0% and 100%, may indicate a pattern of either very low support or unopposed victories. This visualization is crucial as it contextualizes the outcome variable in our analysis, showing the range of primary vote percentages across candidates.



Distribution of Number of Endorsements: The bar chart illustrates the frequency distribution of candidates based on the number of endorsements received. The steep drop-off after zero indicates that most candidates receive few or no endorsements, which underscores the rarity and potential influence of receiving multiple endorsements.



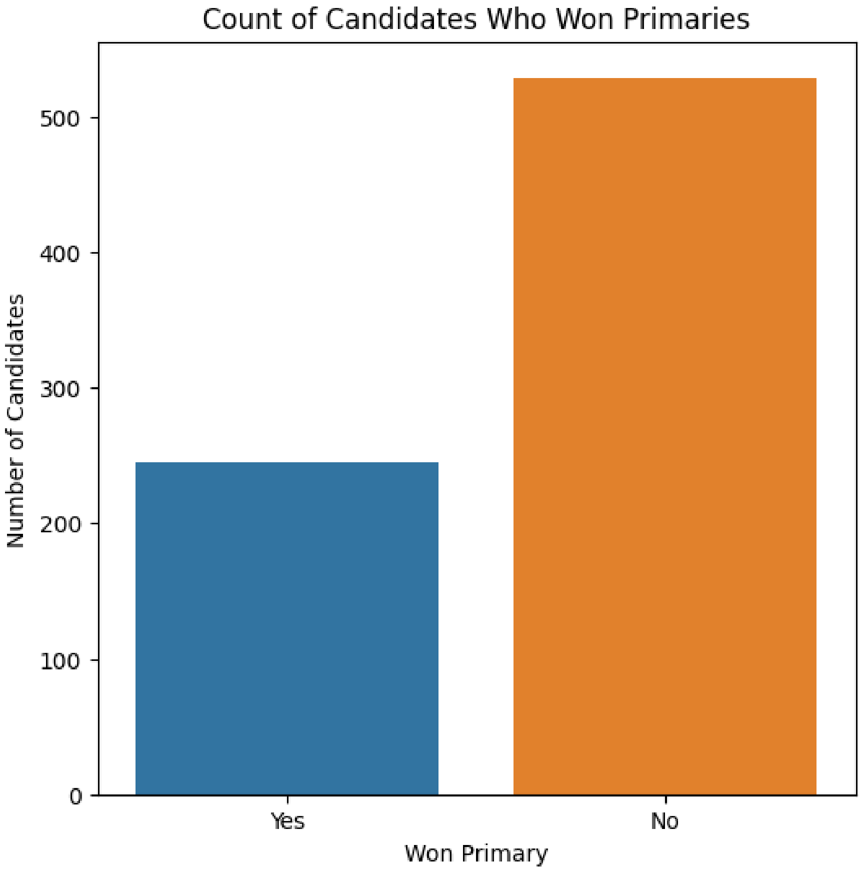
Correlation between Partisan Lean and Primary %: -0.32064837995428497

Correlation between Partisan Lean and Primary Success: -0.19121074902138763

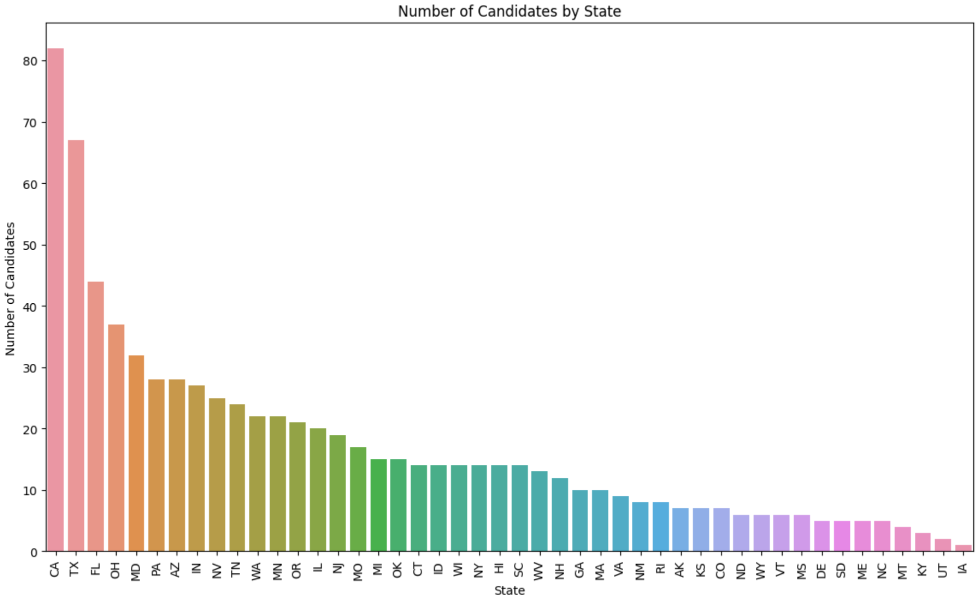
Takeaways: Pearson's correlation between Partisan Lean and Primary % of -0.32 indicates that as Partisan Lean increases, the primary % tends to decrease. This is interesting because this shows that as the district or state becomes more Democratic-leaning, the percentage of the primary vote received tends to decrease. This shows a negative association between partisan lean and primary %. However, it is important to know this is not a strong correlation; this is a medium correlation, as it is greater than 0.3 but less than 0.7 away from 0.

Pearson's correlation between Partisan Lean and Primary Success suggests that on average as the district or state becomes more democratic leaning, the likelihood of winning tends to decrease. This might suggest that winning primaries in a more democratic-leaning area is a bit more challenging. This shows a negative association between partisan lean and primary success. However, this is a weak correlation as it is less than 0.29 away from 0.

Visualization of Categorical Variables:



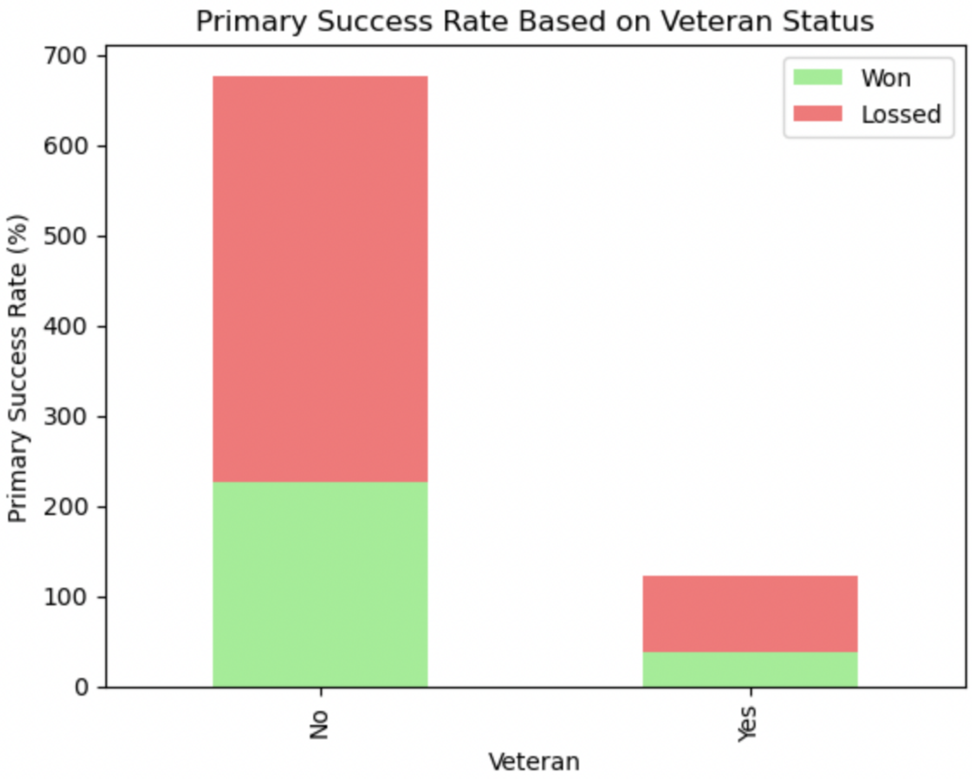
Count of Candidates Who Won Primaries: This bar chart provides a direct comparison between candidates who won their primaries and those who did not. The noticeable difference in counts is indicative of the competitive nature of primaries and sets the stage for analyzing the impact of endorsements on these outcomes.



Number of Candidates by State: Displaying candidates across states highlights the geographical spread and potential regional differences in the data. This is relevant in considering how state-specific factors might interact with endorsements and election outcomes.

Explanation of Relevance:

Each visualization is tailored to provide insight into the variables and relationships at the heart of our research question — the effect of political endorsements on election outcomes. The chosen visualizations not only describe the data but also reveal patterns and trends that are directly relevant to understanding the dynamics of political endorsements in primary elections. They support the narrative by visualizing the underlying data that motivates the research question or hints at potential answers.



```
Percent of success among candidates who are Veterans:  44.70588235294118
Percent of success among candidates who are not Veterans:  50.44444444444445
Difference between percentages:  5.738562091503269
```

Takeaways: The is a 5.738 % difference in success between veteran and non-veteran candidates. Although this isn't a big difference, there is potential for veteran status to impact primary success. We will need to do testing to see if this is the case.

Multiple Hypothesis Testing

Research Question

Is there a significant difference in the probability of winning a primary election between candidates with and without specific endorsements, affiliations, or personal characteristics in the 2018 primaries?

For this question, it makes sense for us to use multiple hypothesis testing because we are comparing the probability of winning a primary election across various categories. Each category will correspond to at least one different hypothesis. We will also compute a few of the hypothesis tests for each party (dataset) Democratic & Republican. Testing multiple hypotheses will allow us to determine whether there is a significant difference in the probability of winning a primary election based on specific endorsements, affiliations, or personal characteristics. For the Democratic dataset, we will be focusing our analysis on whether or not a candidate was self-funded, a veteran, part of the LGBTQ community, and partisan lean. For the Republican dataset we will test whether or not having a higher primary percentage or whether they were endorsed by Trump increases the likelihood of them winning the primaries.

Test Used

We used a range of tests based on the kinds of variables and how many we were dealing with. We used the Chi-Square Test, Pearson Correlation Test, and T-testing to evaluate the correlation and association and output P-values for each hypothesis. The Chi-Squared Test was used when we had two binary categorical variables we were testing as this test requires two categorical variables. The Pearson Correlation Test was only used for one hypothesis because this hypothesis dealt with quantitative variables as opposed to categorical variables. Last but not least, we used T-testing for one of our hypotheses because we are dealing with primary % and a categorical variable. We could have used logistic regression but we chose to do a T-test.

Methods for Corrections

The correction methods we used were the Bonferroni Correction method for the Familywise Error Rate (FWER) and the Benjamini - Yekutieli method for the False Discovery Rate (FDR). The Bonferroni correction method does not mind if our tests are dependent or not, and only wants to make sure the probability of making a Type I error is lower. However, in regards to the FDR, we had to use the Benjamini - Yekutieli method, which is concerned with general dependence. This method is different from Benjamini-Hochberg in that it uses the harmonic number when used under arbitrary dependence. Both B-H and B-Y are valid under independence, but when tasked with dependence as our tests are, B-Y is a better fit. Seeing as we are factoring independence, then the B-Y could be stated as a bit more “conservative.”

FWER & FDR

FWER control looks at each test and tries to lower the probability of getting that result incorrect closer to 0. The FDR control tries to lower the overall proportion of getting an incorrect result. FWER control tries to be strict with our mistakes, whereas FDR is a bit more lenient with making mistakes. Looking at our tests, seeing as these people are the leaders of the country and are making important decisions, we would want to make sure that the probability of any of our results being wrong is closer to 0, so FWER control would be better. This would also be important if we were the candidates, as we would want to be able to increase our chances of winning.

Results

For our tests, we made discoveries for hypotheses 2, 5, and 6. This means that for Democrats, we concluded that there is a significant association between primary outcomes and a candidate's calculated partisan lean. For Republicans, we have discovered that there are associations between a candidate's Primary Percentage (how many votes they got in their primary election) and whether they were endorsed by Trump or not.

For Democrats, it means that if the candidate has a higher partisan lean, they will also have a higher chance of winning. This can be logically explained with an example of California's voting tendencies. Given that California tends to vote Democratic and has voted Democratic for the past few elections, then a Democratic candidate will have a higher partisan lean and because of this will have a better chance of winning.

For Republican candidates, having a higher Primary Percentage and being endorsed by Trump both have a positive impact on winning the primaries. This may be caused due to the idea that Republicans feeling extreme loyalty to their ideals and candidates, and supporting whoever holds the same values as they do.

There are significant associations between the variables in hypotheses 2,5 and 6.

Bonferroni Method & B-Y Procedure

The Bonferroni method tries to control the Familywise Error Rate, which is the probability of coming to at least one incorrect discovery in a group of hypotheses. This means we want to lower the probability of making a wrong conclusion at all closer to 0. We would use this when we are worried about the results of each specific hypothesis. For the B-Y procedure, we want to lower the False Discovery Rate, which is the proportion of false conclusions overall. This is different from the FWER in that when we look at the bigger picture, we are okay with making a few errors, as we care about the number of specific results, rather than looking at each one individually.

Multiple Hypothesis Conclusion

Key Discoveries

When applying the Bonferroni correction method, we made discoveries for the same hypothesis, hypotheses 2,5, and 6. This is also the same for the B - Y correction method. This means that, after applying our correction methods, the number of discoveries was unchanged. The association between these hypotheses was so significant that they were not caused by chance or error.

Decisions

For Hypothesis 1, 3, and 4 we failed to reject the null hypothesis which means that there were no significant associations between the variables in each hypothesis.

For Hypotheses 2, 5, and 6 we reject the null and favor the alternative hypothesis. This means that for these hypotheses, there were significant associations between the variables in these hypotheses.

Limitations & P-hacking

One limitation we had throughout our analysis was the number of missing values throughout the Republican dataset. Our initial intent was to do at least 3 hypotheses for each dataset but due to the number of missing values in the columns of this dataset, we were only able to compute two. In most of the columns, we were missing about 80% of the values which we felt would alter and hurt our results and accuracy. To avoid P-hacking, we used only columns for which we had at least 90% of the values filled in. We also made sure we did not change any of the data to favor our test and results.

Several of our columns had blank or missing values, especially in the endorsements section. This led us to think about what we should do with these rows - impute values, create a new column for a new metric, etc. In addition, we are only examining 1 year of elections. If we had more years, then we could make more accurate predictions and realistically determine what factors affect winning elections. This is because some years may have events that affect elections or may have some sort of “surprise” result. Finally, if we had data from other types of elections, rather than just primaries, then we could see if these factors also affect elections on a wider scale.

Addition Test & Data

If we had more data, it would be nice to conduct some more tests for hypotheses that consider columns like age, ethnicity, and voter turnout. We believe these columns would have interesting and useful findings, as they could lead to more insight regarding voter behavior and possibly more features that could be significant in predicting success in primaries. Unfortunately, these datasets did not contain this information.

Call To Action

Our results should emphasize the importance of endorsements and primary percentages, and how securing these important factors could be the difference between winning and losing a primary. By doing more research on their respective party and respective voter base. In addition, they could study features that are similar to the ones we used, such as past partisan leanings. By being thorough with their data collection, then that would also allow those trying to study election results to have more accurate results and could pinpoint what aspects of a candidate's campaign needs to be improved, and what seems to be working well.

Merge datasets

We did not merge any datasets. We simply analyzed two sets of data.

Future Studies

This type of project could provide useful insights regarding studying trends for political elections. Candidates who want to increase their chances of winning can learn and expand on what types of features they would want to emphasize over others to increase their chances of winning their primaries and securing their positions. By allowing us to see the limitations of this type of study, it could also help those who are collecting such data in the future to be more thorough and understand what kind of data would be useful when doing this type of work.

Causal Inference

Research Question

The study aimed to determine whether receiving high-profile political endorsements increases a candidate's likelihood of winning primary elections, both for Republican and Democratic candidates.

Treatment and Outcome Variables

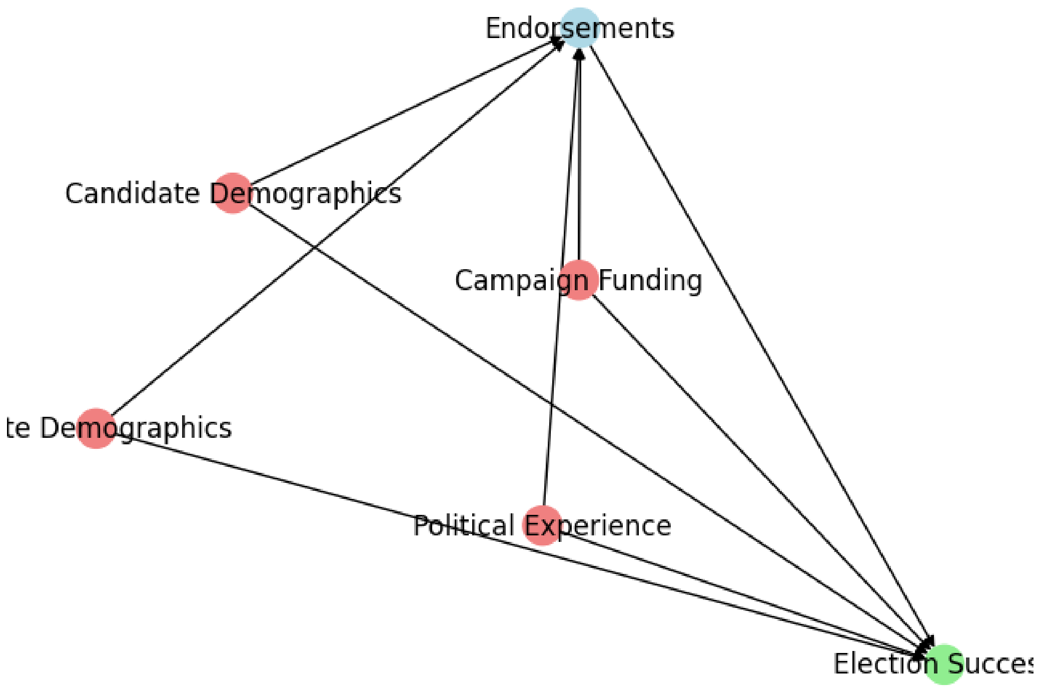
The treatment variable in both the Republican and Democratic datasets is the receipt of high-profile political endorsements. This was operationalized as a binary variable indicating whether a candidate received any endorsement from a specified list of prominent figures or organizations. The outcome variable is the success in primary elections, measured as a binary indicator denoting whether the candidate won their primary election.

The confounders that we considered are variables such as candidate demographics, political experience, campaign funding, and state of election could act as confounders. These factors might influence both the likelihood of receiving endorsements and the election outcomes. In this project we used the unconfoundedness assumption by assuming that all confounding variables that could influence both the receipt of endorsements and primary election success have been accounted for either through the dataset or the matching process. This assumption holds under the belief that our dataset captures a comprehensive set of variables influencing both the treatment and the outcome.

Methods to Adjust for Confounders

To overcome our confounders we used propensity score matching and evaluated potential colliders. Propensity score matching was employed to control for confounders. This method involves matching candidates who received endorsements with those who did not based on a set of observable characteristics, thereby aiming to isolate the effect of endorsements. Variables such as media coverage or public opinion might be colliders if they are influenced by both the endorsements and the election outcomes. We took care to avoid controlling for such variables to prevent introducing bias.

Causal DAG



In the provided Directed Acyclic Graph (DAG), the central hypothesis posits that high-profile political endorsements directly influence a candidate's success in primary elections. The DAG also incorporates four confounders—candidate demographics, campaign funding, state demographics, and political experience—which are theorized to affect both the probability of receiving endorsements and the election outcomes. This conceptual framework underpins the causal analysis, outlining the necessity to control for these confounders to accurately estimate the effect of endorsements on electoral success.

Justification for Methods

We believed that propensity score matching was appropriate due to its effectiveness in reducing selection bias in observational studies. By matching candidates on observable characteristics, we aimed to create comparable groups, thus mimicking a randomized experiment. The choice of caliper width in the matching process was a crucial design decision. Different widths were tested to find the most effective balance between having a sufficient number of matches and ensuring similarity between matched pairs.

Summary and Interpretation of Results

The analysis conducted on both Republican and Democratic datasets aimed to assess the impact of high-profile political endorsements on primary election outcomes. The methodology involved propensity score matching to estimate the Average Treatment Effect on the Treated (ATT) and statistical significance testing using Z-tests.

For both datasets, the ATT values suggest a causal relationship between receiving endorsements and an increased likelihood of winning primary elections. The Republican dataset showed an ATT of approximately 23.81%, and the Democratic dataset showed an ATT of approximately 29.23%. These figures indicate that candidates receiving endorsements have a significantly higher chance of winning their primaries compared to non-endorsed candidates.

The p-values obtained from the Z-tests (3.503374651950894e-05 for both datasets) confirm that these effects are statistically significant and not due to random chance.

The magnitude of the effect is substantial in both cases. The likelihood of winning primaries increases by around 24% to 29% with endorsements, indicating that endorsements play a crucial role in a candidate's success.

Uncertainty and Counter-Evidence

While the results are statistically significant, they are subject to the usual limitations of observational studies. The primary assumption is the absence of unobserved confounders. If relevant variables were omitted, the estimated ATT might be biased. Therefore, conclusions drawn should be seen in light of this potential limitation.Additionally, there was no significant evidence found against the hypothesis that endorsements positively impact election outcomes. However, this does not rule out the possibility of other factors playing a role in these outcomes. For instance, the inherent qualities of the candidates or other aspects of their campaigns might also contribute to their success.

Discussion

Limitations of the Methods

While we ultimately believe that the strategies we used are sound, we are also aware of some of the limitations of propensity score matching, our assumption of unconfoundedness, and the matching process. While propensity score matching is a powerful tool in observational studies to control for observed confounders, it does not account for unobserved confounders. Any variables not included in the dataset but potentially influencing both the likelihood of receiving endorsements and election outcomes could bias the results. The analysis assumes that all relevant confounders were measured and included. If this assumption does not hold, the causal inference might be compromised. The matching process, particularly the selection of the caliper width, involves a degree of subjectivity and could influence the ATT results. Additionally, the effectiveness of matching is limited by the overlap in propensity scores between the treatment and control groups.

Confidence in the Causal Relationship

Our level of confidence is moderate to high in the causal relationship between endorsements and primary election outcomes. The robustness of the propensity score matching and the consistency of results across both political parties bolster this confidence. The statistical significance of the results, as indicated by the very low p-values, suggests that the relationship is not due to random chance. However, the confidence is tempered by the limitations inherent in any observational study, particularly concerning unobserved confounders.

Additionally, we found that the results were similar across the datasets between Republicans and Democrats. The fact that similar patterns were observed in both Republican and Democratic datasets adds to the confidence that the relationship is not spurious or specific to a single political context.

Causal Inference Conclusions

Key Findings Summary

The analysis revealed a significant causal relationship between receiving high-profile political endorsements and the likelihood of winning primary elections. This effect was consistent across both Republican and Democratic datasets. For Republicans, the Average Treatment Effect on the Treated (ATT) was approximately 23.81%, while for Democrats, it was around 29.23%. These figures indicate a substantial influence of endorsements on electoral success.

Merging Different Data Sources

Merging datasets from both Republican and Democratic primaries allowed for a comprehensive analysis and comparison, strengthening the conclusions drawn. While beneficial for broader insights, merging different sources may have introduced complexities, especially in terms of data consistency and variable definitions. For example, while both datasets were compiled by the same author, the Democratic dataset had attributes related to the partisan lean of different locals, while the Republican dataset didn't include this information.

Limitations in Data

The main limitation is the potential presence of unobserved confounders that could bias the results. This includes factors like candidate charisma or campaign strategy nuances. The analysis was limited to the data available, which may not cover all relevant aspects affecting election outcomes.

Call to Action

Ultimately this data and analysis is primarily relevant to political campaigns and their strategists. Our results tell us that political campaigns should prioritize securing valuable endorsements as a key strategy to swaying the electorate to support their candidate. This information is critical to strategists as resources like a candidate's time and the underlying organization of their on-the-ground advocates is finite, but receiving endorsements allows for outside parties to assist in the campaigns with minimal expenditure of resources within the campaign.

Future Studies

The next questions that this research has led to are answering the questions of the impacts of endorsements on general elections, or how the trends change from election to election. This kind of longitudinal data would provide insights on how endorsements affect election results in different political environments. Longitudinal data would also help us distinguish common voting behaviors of a particular district against the endorsements that a particular candidate received, potentially bolstering the impact and validity of our research. It would also be interesting to dive into the types of endorsements political candidates receive, e.g. from a celebrity, politician, or a political organization.

More data on the candidates themselves, especially from the Republican side of the dataset would be interesting as well. The lack of identity signifiers for the Republican candidates meant direct comparison on identity was virtually impossible with the accumulated dataset.

Learnings from the Project

Besides the insights of our results, we also took away a few key points from doing the project itself. Including the difficulty and care that causal inference requires. Our selected methods also emphasized the difficulty of drawing concrete conclusions from observational studies. The project also gave our team an insight into the difficulty of working with political data.

Citations

FiveThirtyEight. (2018). Primary Candidates 2018. GitHub.
https://github.com/fivethirtyeight/data/tree/master/primary-candidates-2018