Exploring the Relationship Between Candidate Features and Funding and Examining the Impact of Candidate Party on Primary Election Outcomes

Cameron Gelvezon, Saaketh Gunukula, Lily Lin, Parker Winig

December 12, 2023

1 Data Overview

The dataset used in this analysis, compiled by FiveThirtyEight, focuses on primary elections that took place in 2018, and represents a census. The data encompasses a wide array of information about each primary candidate, including their political leanings, endorsements, and personal characteristics such as race and veteran identity, as well as outcomes for each election. The compilation process involved multiple sources, notably Ballotpedia, ProPublica, and the Federal Election Commission (FEC). We utilized this dataset in conjunction with FEC campaign financing data, which provided information about the financial aspects of each candidate's campaign, including the amount of money raised and spent, as well as details about donors and contributions.

1.1 Granularity and Interpretation

The dataset represents each of the 811 Democratic and 774 Republican primary candidates from the 2018 elections, with each row signifying a distinct candidate in a specific primary. Although our high granularity allows for more precise analysis and identification of patterns, it may also be liable to overfitting or challenges in interpretation. While candidates are likely aware of their participation in primary elections and the public availability of their information, potential biases and limitations should be considered, such as potential measurement error, if there are inaccuracies in the original data sources. However, we circumvented this potential issue through cross-verification of the data used and throughout the data cleaning process to identify outliers and inconsistencies.

1.2 Data Limitations and Considerations

The absence of information regarding media coverage and candidates' likability and personality traits within the dataset represents a significant limitation, as these factors play pivotal roles in shaping public perception and influencing both campaign funding and electoral outcomes. Media coverage serves as a key determinant of a candidate's visibility, with the frequency, tone, and platform of coverage influencing voters' perspectives. Understanding the nuances of media attention could offer insights into campaign and funding dynamics and potentially correlate with success in primary elections. Similarly, likability and personality traits are critical determinants of voter preferences, impacting electoral choices. Candidates perceived as relatable, trustworthy, or possessing positive traits often enjoy advantages in garnering monetary and voter support. Without explicit measures of likability or approval rating, the dataset may lack a nuanced understanding of how voters perceive and connect with candidates on a personal level.

In our analysis of candidate funding for primary elections, it is crucial to acknowledge and discuss the presence of missing data, particularly in variables such as "Race," "Incumbent," and "Self-funder," where information is exclusively available for Democratic candidates. The absence of data for non-Democratic candidates in these variables poses limitations on the comprehensiveness of our investigation. We recognize that this limitation may introduce biases and impact the generalizability of our findings, as conclusions may not be representative of other political affiliations. However, we aim to address the results holistically, and are not aiming to identify specific differences between parties, but we will keep in mind this potential bias, and attempt to minimize its impact through our methods and discussion.

For our causal inference, we would like more information about beliefs/policies that the candidate supports, but that was unavailable to us and the information that was available had more than 50% of the data not filled out. This would help us get more confounders identified, making our causal inference stronger.

There are columns with a lot of missing data. This missing data has to do with a large part of the personal data for the candidates, which is something that would be valuable for causal inference. They don't particularly mean anything, this was just a case of one data set not having the columns filled out whereas another dataset did. We imputed the values without assumptions, but this again leaves a lot of bias which we discuss later in our findings.

1.3 Data Cleaning and Processing

In our data cleaning and pre-processing phase, we made several key decisions to refine the dataset for a more targeted analysis. First, we narrowed the scope of our research question by focusing exclusively on Senate and Representative elections while filtering out Governor and other election types. This decision was driven by the desire to streamline our analysis and ensure relevance to our specific area of interest. We joined the Democratic dataset with the Republican dataset to create a unified dataset that encompasses candidates from both major political parties. Furthermore, we integrated campaign funding data from the Federal Election Commission (FEC) into our dataset. This addition of financial information is essential for our analysis, as it enables a direct exploration of the relationship between candidate characteristics and their funding levels. In addition to these steps, we introduced a new feature, "Swing State," indicating whether a candidate participated in a swing state election. This feature is strategically included to capture the unique dynamics of elections in swing states, which are often characterized by heightened competition and strategic campaigning.

2 EDA

2.1 EDA: Research Question 1

Before diving into our first research question, we wanted to visualize the relationships between our different hypotheses and their funding levels. For all of our hypotheses, we plotted the mean, median, and count to account for outliers and provide a more holistic basis for potential effects in hypothesis tests.

Democratic candidates had greater mean and median funding levels compared to Republican candidates, as seen in Figure 1. However, the mean funding is over ten times that of median funding for both parties, suggesting a wide range of funding for all candidates.

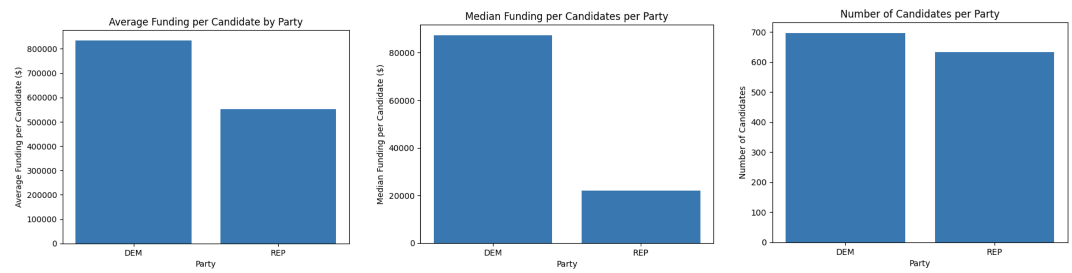

Figure 1: Average, median, and count of candidates by party.

Within Democratic candidates, Figure 2 illustrates that average funding for white candidates notably exceeds that of non-white candidates. Despite this gap, median funding levels remain comparable. With over double the number of white candidates, the funding landscape suggests a concentration of exceptionally high amounts for white candidates, influencing the mean.

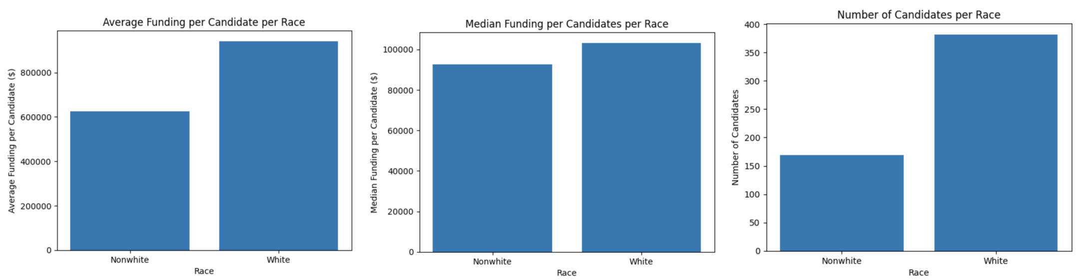

Figure 2: Average, median, and count of candidates by race.

Six of the top 20 states, ranked by average candidate funding, are swing states (Figure 3). Florida stands out with nearly double the average funding of the second-highest state. However, only three swing states make it into the top 20 for median funding, indicating potential disparities in funding distribution. Interestingly, eight swing states are among the top 20 for candidate counts, emphasizing their significant role in the political landscape. From this initial analysis, the effect of swing state status is unclear, so further analysis must be done.

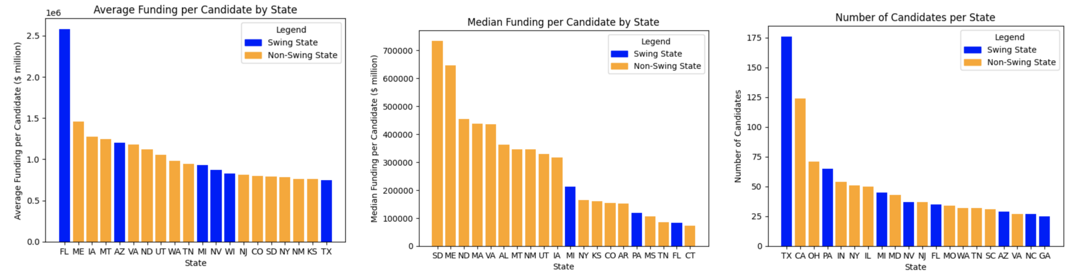

Figure 3: Average, median, and count of candidates by swing state.

As shown in Figure 4, candidates backed by their party enjoy a substantial financial advantage, with both average and median funding approximately tenfold higher than those without support. Despite this, the count of supported candidates is just over half that of unsupported candidates. This financial gap underscores the influential role of party support in shaping campaign financing dynamics.

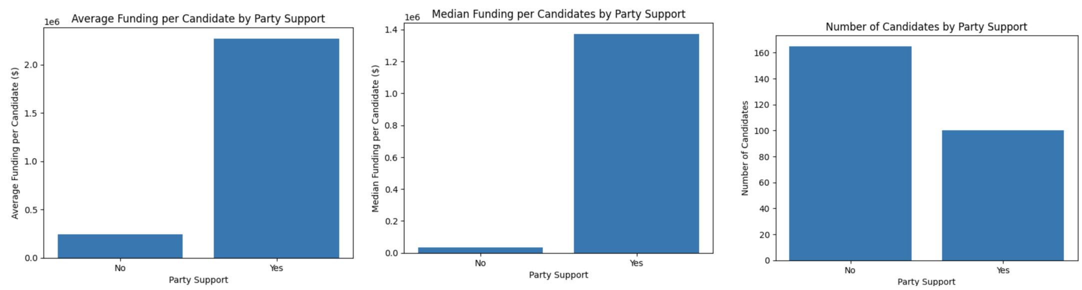

Figure 4: Average, median, and count of candidates by party endorsement.

On average, incumbent candidates receive around four times more funding, while the median funding for incumbents is approximately seven times higher than that of non-incumbents (Figure 5). These significant differences underscore a substantial financial advantage for incumbents in comparison to their challengers.

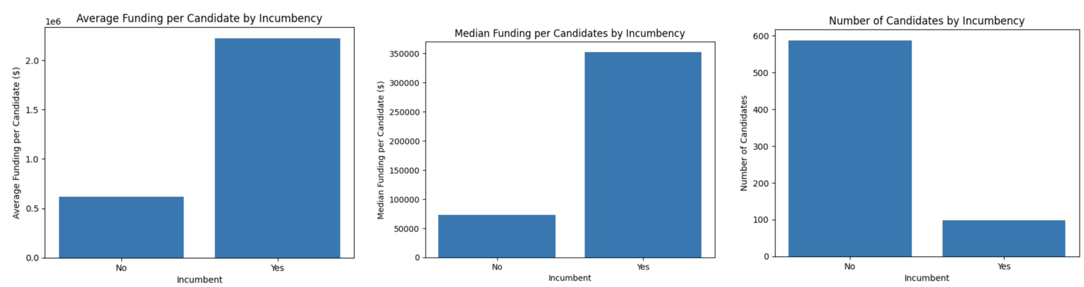

Figure 5: Average, median, and count of candidates by incumbency.

On average, self-funding candidates receive double the funding compared to non-self-funders (Figure 6). Additionally, the financial gap widens when looking at median values, with self-funders having about seven times more funding than their counterparts. Interestingly, despite this funding discrepancy, the count of non-self-funding candidates exceeds that of self-funders by over ten times, revealing the skewed proportion of funding for self-funders.

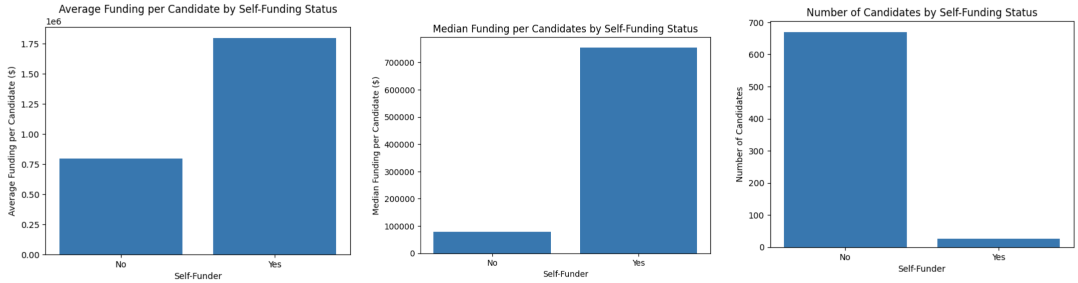

Figure 6: Average, median, and count of candidates by self-funding status.

These visualizations are pivotal in validating our hypotheses regarding primary election funding. They provide a snapshot of funding disparities based on party affiliation, racial influences, the impact of swing state status, party support advantages, funding disparities related to incumbency, and the role of self-funding. The observed patterns in mean, median, and count offer valuable insights into the central tendencies and distribution of funding levels across various candidate characteristics. As we proceed with our hypothesis testing, these visualizations will serve as a crucial reference point, allowing us to interpret the significance of our findings within the context of the observed funding landscapes.

2.2 EDA: Research Question 2

Before we went to answer our second research question, we thought it was important to visualize the relationships between the outcome variable and other variables in our data set.

One relationship that we were particularly interested in was the relationship that being a swing state had on the primary percentage, which was the outcome variable that we wanted to explore. To visualize this relationship, we utilized a histogram that showed the distribution of primary percentages for both swing states and non-swing states.

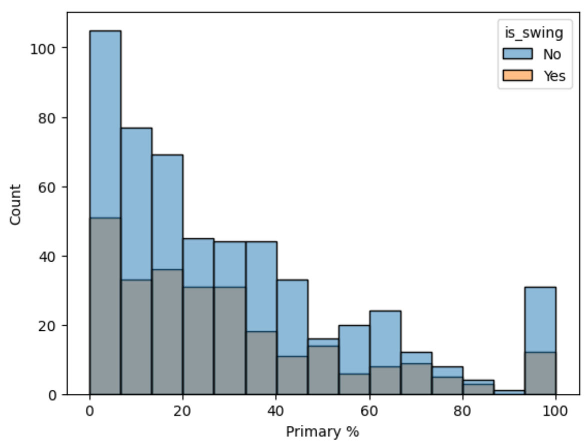

Figure 7: Histogram of Primary Percentage Based on Whether Swing State or Not.

From what we see, the difference between the two distributions is very minimal. They both skew right, having less and less height the more you go left. The similarities between the two distributions suggest that being a swing state or not doesn't really have that much effect on the primary percentage.

Another relationship that we wanted to visualize was the relationship between endorsements and primary vote percentage earned, across a variety of endorsement types and across both the Democrat and Republican parties. These plots provide us with a greater understanding of how endorsements from interest groups and prominent figures impact outcomes of primary elections across all candidates, separated by political party.

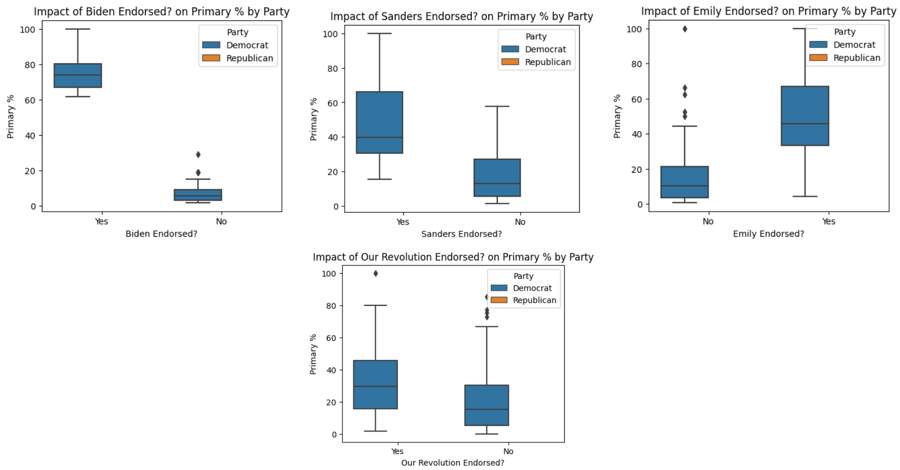

Figure 8: Boxplot of primary percentage based on endorsements (Democrats).

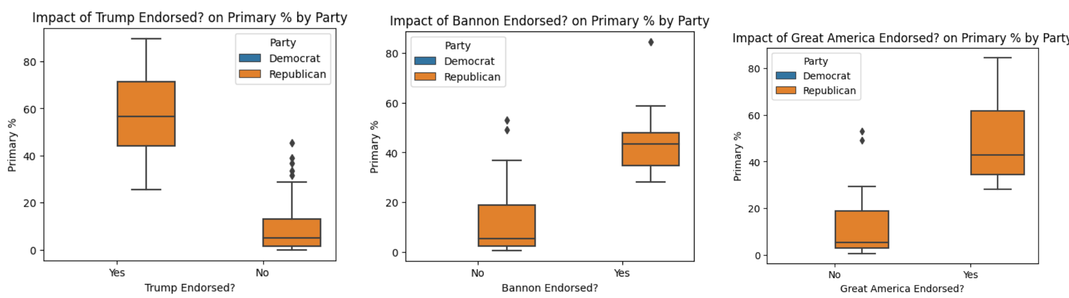

Figure 9: Boxplot of primary percentage based on endorsements (Republicans).

Candidates with endorsements outperform those without endorsements across the variety of politicians and interest groups that we analyzed here, yet the margin of impact differs across each endorsement. We see large divides between endorsements from big-name politicians, such as Trump and Biden, where there are very large differences in the primary percentage earned between endorsed and non-endorsed candidates. For other figures, such as Sanders and Bannon, as well as interest groups, the difference is not as pronounced, which may be influenced by the popularity of each figure and group, as well as by other important factors in each candidacy.

3 Research Questions

We aim to research the following questions:

3.1 Which features are associated with higher funding for candidates in primary elections?

To analyze campaign funding trends in primary elections, we have chosen to leverage Multiple Hypothesis Testing as our analytical method. The primary focus of our investigation encompasses six distinct hypotheses, each focusing on a different feature of a candidate's profile. We aim to discern potential variations in funding concerning party affiliation, racial background, swing state status, party endorsement, prior experience in holding public office, and self-funding. This comprehensive approach allows us to gain a better understanding of the multifaceted nature of funding dynamics in primary elections, offering valuable insights for political strategists, campaign managers, and policymakers. By employing Multiple Hypothesis Testing, we can identify key determinants of campaign funding, inform targeted fundraising strategies, and shape resource allocation and campaign planning decisions. Additionally, this analysis has the potential to shift campaign funding and may point out flaws in current campaign funding schemes. We aim to determine if funding is based on pure qualifications and prowess of the candidate, or merely factors out of the candidate's control.

Multiple Hypothesis Testing is well-suited for this question due to its ability to accommodate the simultaneous analysis of multiple variables. However, it comes with certain limitations, including an increased risk of Type I errors, especially in the context of numerous comparisons. Caution is necessary to avoid overinterpretation of statistically significant results, and the method might not provide causal insights under complex relationships influenced by unobserved variables.

3.2 Is there a causal relationship between a candidate's political party and their 2018 primary percentage?

The main question that we wanted to address is whether the candidate's party has an effect on their primary %. We think that this is a very important question to ask because candidates want to maximize their primary % as much as possible to win, and so knowledge about a party's effect is very relevant to a candidate. By analyzing the effect that a candidate's party has on their primary %, they can modify policies or arguments to increase their chance of winning.

There are limitations to the methodology that we are using, however. The main limitation is our access to confounding variables. While we do have access to a multitude of confounding variables, we do not have access to every single confounding variable possible, so our conclusions are limited and may not fully capture the causal effect that our treatment has on our outcome.

4 Research Question 1: Which features are associated with higher funding for candidates in primary elections?

4.1 Methods

For our hypothesis testing, we used the following hypotheses:

H0: Differences in a specific characteristic of a candidate have no relation to the amount of funding that a candidate receives.

H1: Differences in a specific characteristic of a candidate correlate with the amount of funding a candidate receives.

We are testing six specific candidate characteristics: party, race, swing state, party endorsed, held office previously, and self-funded. From our data overview, it is reasonable to proceed with multiple hypothesis testing, as we can observe the uniqueness of each of the six different characteristics. Furthermore, through our EDA, we have seen trends in funding in association with each of these characteristics, which justifies the usage of multiple hypothesis tests simultaneously. Since we are working with multiple characteristics of candidates, each of which can be binarized, we utilized A/B testing for our hypotheses.

4.1.1 Controlling FWER and FDR

For corrections, we first used naive thresholding with an alpha value of 0.05. Then, we used the Bonferroni correction method to set the family-wise error rate (FWER) to 0.05. The family-wise error rate is the probability of making at least one Type I error among all the hypotheses tested. Finally, we utilized the Benjamini-Hochberg procedure to control the false discovery rate (FDR) to 0.05. The FDR is the proportion of false positives among rejected hypotheses.

In utilizing both Bonferroni and Benjamini-Hochberg to control FWER and FDR, respectively, we found similar discoveries for each method in our testing. For our research question, which was to determine whether candidate characteristics impact the amount of funding received, controlling FDR is a better fit for our use case. We allow for the possibility of some controlled rates of false positives with the tradeoff of gaining increased insights and identifying potential discoveries, with less severe consequences of false positives allowing us to explore further. Therefore, utilizing the Benjamini-Hochberg error correction is more appropriate for our research question and hypothesis testing where we can afford to take a less conservative approach, while still managing the risk of false positives.

4.1.2 Alternative Hypothesis and Power of the Test

Specifically examining the effect of self-funding on funding, we calculated the power of the test to be 0.0469. This power is low, indicating that the test may have limited ability to detect a true effect if it exists. However, the power of the test may be impacted by the relatively small sample size of identified "Self-Funders" and a wide range of funding for both self-funders and non-self-funders, as seen in the graph below.

4.2 Results

Name	Value	Naive	Bonferroni	Benjamini_Hochberg
party_pval	0.155	False	False	False
race_pval	0.409	False	False	False
swing_pval	0.233	False	False	False
party_support_pval	0.0	True	True	True
incumbent_pval	0.0	True	True	True
self_fund_pval	0.044	True	False	False

Table 1: P-values for each test and their corrections.

Following our A/B tests we achieved p-values of 0.153 for party, 0.3261 for race, 0.2424 for swing state, 0.0 for party endorsement, 0.0025 for incumbency, and 0.048 if they were self-funded (Table 1). Using a naive $\alpha = 0.05$, we rejected three tests (party endorsement, incumbency, and self-funded) signaling that there is an effect of those variables on candidate funding. Controlling for multiple hypothesis tests, we used Bonferroni and Benjamini-Hochberg corrections to control for FWER and FDR respectively. Setting the FWER = 0.05, the significance level, α is adjusted to 0.0083 (α / number of tests). With this adjusted significance threshold, we rejected two tests (party endorsement and incumbency), indicating that there is still statistically significant evidence of an effect for these variables on candidate funding, even after adjusting for multiple comparisons. Additionally, we applied the Benjamini-Hochberg correction to control the False Discovery Rate at a desired level (e.g., 0.05). This method allows for a more lenient adjustment compared to Bonferroni, potentially increasing power. After applying Benjamini-Hochberg, we found that party endorsement remained statistically significant, providing further support for its influence on candidate funding. These results suggest that party endorsement and incumbency are influential factors affecting candidate funding (Figure 7), as their effects withstand correction for the increased risk associated with multiple hypothesis testing.

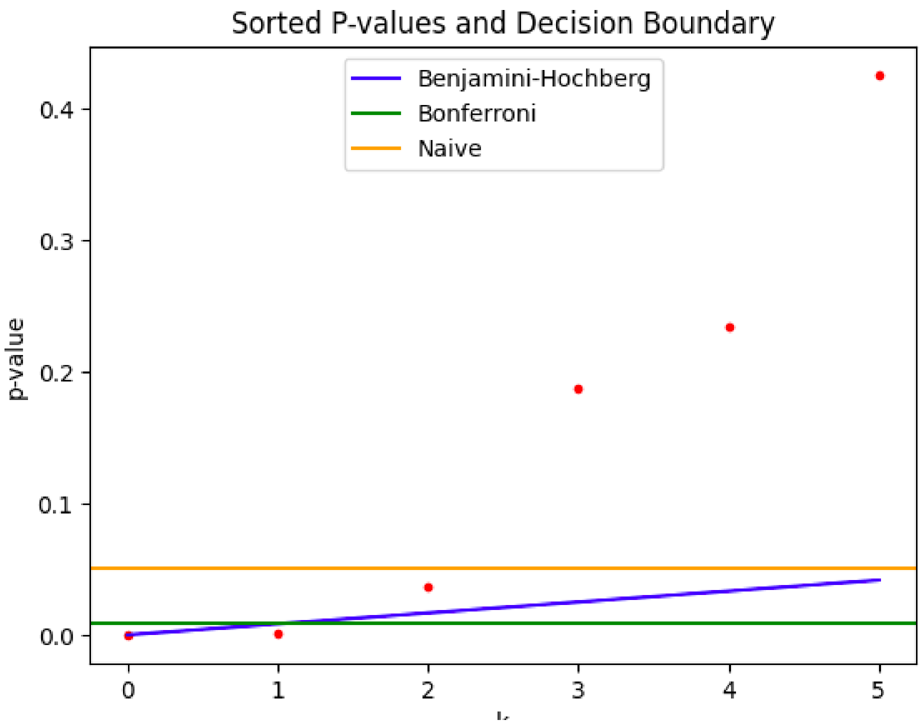

Figure 10: Plot of sorted p-values and decision boundaries for error correction.

4.3 Discussion

In our testing, we conducted multiple hypothesis tests to investigate the influence of several candidate factors on the amount of funding received. After applying correction procedures to account for multiple tests, the significance of some discoveries shifted. However, the corrected results revealed notable findings for certain factors. Party affiliation (p_val = 0.144), race (p_val = 0.302), and swing state (p_val = 0.259) all did not demonstrate a significant association with candidate funding after applying correction procedures, suggesting that these factors do not affect funding outcomes. In contrast, party support (p_val = 0.0) and incumbency (p_val = 0.004) exhibited a highly significant association with candidate funding even after FWER and FDR corrections, leading to two discoveries being made. This implies that candidates with strong party support and incumbent status are likely to secure increased funding, pointing to the influential role of party backing in fundraising efforts.

Alternatively, the association between self-funding and candidate funding approached marginal significance after corrections, as it only had a significant association with funding under naive thresholding. While not as robust as party support or incumbency, this result implies a potential effect where candidates who contribute significantly to their own campaigns may experience a boost in funding. Nonetheless, the observed marginal significance suggests caution in making definitive conclusions about the impact of self-funding, especially considering the low power of the test. The relatively small sample size of identified "Self-Funders" and the wide range of funding for both self-funders and non-self-funders may have contributed to the low power, and a larger sample size would be beneficial for a more robust analysis.

4.3.1 Limitations

These results, however, are contingent on the assumptions and limitations of the analysis. The power and generalizability of the findings may be influenced by the relatively small sample size and specific characteristics of the dataset, potentially limiting the broader applicability of these conclusions. Additionally, the associations observed may be context-dependent, and caution should be exercised in extrapolating these results to different political landscapes or periods. Our analysis also lacks unquantifiable information about each candidate, such as personality, which could be a significant limitation. Candidate personalities, including traits, communication styles, and public perception, may play a crucial role in influencing fundraising outcomes, and the absence of this information in our study restricts a comprehensive understanding of the dynamics at play.

For three of our key features, race, self-funder, and incumbent, we also only have data from Democratic candidates. This limitation raises concerns about the potential bias in our findings, as the dataset may not accurately capture the diversity of these features across different political affiliations. The absence of information for Republican candidates in these categories compromises the comprehensiveness of our analysis and hinders our ability to draw robust conclusions that are representative of the entire political spectrum. This bias emphasizes the need for cautious interpretation and acknowledgement of the specific focus and constraints inherent in our dataset.

Candidate	State	Cand_Party	Net_Contribution	Race	Party Support?	Self-Funder?	Swing?	Elected Official?
John James	MI	REP	12229389.45	NaN	NaN	NaN	Yes	NaN
Marsha Blackburn	TN	REP	12249043.31	NaN	NaN	NaN	No	NaN
Martha McSally	AZ	REP	17907003.18	NaN	NaN	NaN	Yes	NaN
Rick Scott	FL	REP	82702798.34	NaN	NaN	NaN	Yes	NaN
Jacky Rosen	NV	DEM	23791657.75	NaN	Yes	No	Yes	Yes
Beto O'Rourke	TX	DEM	78904203.61	White	NaN	No	Yes	Yes

Table 2: Candidates with funding three standard deviations above the mean.

Funding outliers, as shown in Table 2, may also affect the tests since they have the same characters or are the only outliers for most of the features. These features include race, party support, self-funder, and elected officials (incumbent), most of which have low p-values. The low p-values may be due to the disproportionately high funding associated with these outliers, potentially skewing the statistical significance of these features.

To address the limitations and dive deeper into the factors influencing candidate funding, additional tests could be considered. Sentiment analysis on candidates' public statements, speeches, or media coverage could unveil correlations between specific personality characteristics and funding. Additionally, examining social media engagement metrics, such as sentiment expressed by followers and the overall tone of interactions, could offer valuable insights into the candidate's public image and its relation to funding. Furthermore, a more nuanced analysis of candidate demographics, such as age, gender, or educational background, could be conducted to identify any additional variables that might contribute to funding outcomes.

In conclusion, while our current findings provide valuable insights into the impact of certain candidate characteristics on funding, addressing the outlined limitations and conducting additional tests could further our understanding of the relationship between candidate attributes and campaign funding.

5 Research Question 2: Is there a causal relationship between a candidate's political party and their 2018 primary percentage?

5.1 Methods

The method we will use to answer this question is causal inference, as it is the perfect way to figure out whether a candidate's party causes an effect on the primary % because it is used to see if the treatment (candidate's party) has an effect on the outcome (primary %).

5.1.1 Confounding Factors

In our causal inference analysis, we have opted not to use instrumental variables, as there are limited variables that exclusively influence a candidate's party affiliation (the treatment) without also impacting the primary percentage (the outcome). Furthermore, finding variables that don't affect other confounders is equally challenging. Therefore, our focus is primarily on confounding variables, of which there are several significant ones. We have identified 'Net_Contribution', 'Net_Operating_Expenditure', and 'swing' as primary confounders for our study.

5.1.2 Rationale Behind Selected Confounders

Net_Contribution: This factor is crucial as it directly influences a candidate's campaign capabilities, thereby affecting their primary percentage. Additionally, the level of contributions received can sway a candidate's decision to affiliate with a particular party.

Net_Operating_Expenditure: Similar to Net_Contribution, this variable impact a candidate's campaign operations and, by extension, their performance in primaries. It also has a potential indirect influence on party affiliation.

Swing States: The 'swing' variable is pivotal due to the typically tighter election margins in swing states. This factor not only influences primary percentages but also potentially affects a candidate's party choice, given the variable political climate in these regions.

While these are our primary confounders, we acknowledge the existence of additional variables that could potentially influence the causal relationship. Variables such as 'Race', 'Veteran?', 'LGBTQ?', 'STEM?', and 'Obama Alum?' are potentially significant. However, a notable challenge with these variables is the substantial amount of missing data, exceeding 40% in some cases. This missing data is predominantly from the Republican candidates, which could introduce a notable bias if these variables are used in their current state. Imputing values for these missing data points would involve making considerable assumptions that may not be entirely accurate or reflective of reality.

5.1.3 Bias Introduced by Imputation

By choosing to impute values for these additional confounders, we must be cognizant of the significant bias this introduces into our analysis. The assumptions made during imputation are not guaranteed to be accurate and could skew our results. This consideration is critical in our secondary analysis, where these variables will be included.

It's important to recognize that this inclusion while offering a broader perspective, may alter the causal inference landscape.

5.1.4 Unconfoundedness and SUTVA

Our analysis assumes unconfoundedness, underpinned by the belief that using all these variables creates independence between the treatment and potential outcomes. This is key to estimating the causal effect accurately.

We also maintain that the Stable Unit Treatment Value Assumption (SUTVA) is upheld. We posit that there is no interference between units, as voters' decisions in primaries are based solely on their preferences for a specific candidate, independent of others. Additionally, the consistency of party affiliation, which generally remains unchanged, reinforces the treatment's stability across individuals.

In conclusion, our selection of confounders and the decision to include additional variables in a secondary analysis are driven by a commitment to a comprehensive understanding of the causal dynamics. However, we must also acknowledge and explicitly state the limitations and potential biases these decisions introduce, particularly in the realm of data imputation.

Because there is unconfoundedness, we can use outcome regression based on the confounding variables that we have set out. We run a regression on the outcome, with the treatment and the confounding variables as the explanatory variables. For this question, we will run two outcome regressions. One with the confounding variables which we don't have to impute values for and another where we include all the confounding variables. We also run a bootstrap to see if our regression is statistically significant.

We are not using any colliders as well.

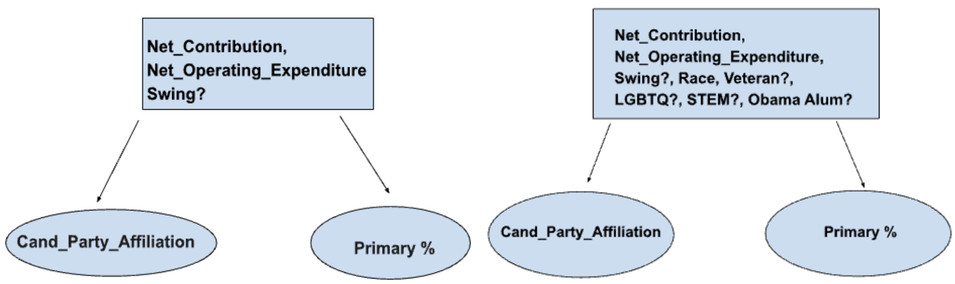

Figure 11: DAG with the confounders that we don't have to impute a significant amount into (left) and DAG with all confounders (right).

5.2 Results

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:              Primary %   R-squared (uncentered):                   0.364
Model:                            OLS   Adj. R-squared (uncentered):              0.362
Method:                 Least Squares   F-statistic:                              187.8
Date:                Mon, 11 Dec 2023   Prob (F-statistic):                   2.54e-127
Time:                        21:37:41   Log-Likelihood:                         -6560.8
No. Observations:                1317   AIC:                                  1.313e+04
Df Residuals:                    1313   BIC:                                  1.315e+04
Df Model:                           4
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Net_Contribution          1.434e-06   5.02e-07      2.856      0.004    4.49e-07    2.42e-06
Net_Operating_Expenditure  4.52e-07   4.14e-07      1.093      0.275   -3.59e-07    1.26e-06
is_dem                      22.4036      1.510     14.835      0.000      19.441      25.366
swing                       21.1339      1.787     11.824      0.000      17.628      24.640
==============================================================================
Omnibus:                       80.955   Durbin-Watson:                   1.671
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               94.876
Skew:                           0.653   Prob(JB):                     2.50e-21
Kurtosis:                       3.157   Cond. No.                     1.10e+07
==============================================================================
```

Figure 12: OLS regression results for non-imputed confounders.

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:              Primary %   R-squared (uncentered):                   0.524
Model:                            OLS   Adj. R-squared (uncentered):              0.520
Method:                 Least Squares   F-statistic:                              159.8
Date:                Mon, 11 Dec 2023   Prob (F-statistic):                   1.54e-203
Time:                        21:37:44   Log-Likelihood:                         -6370.3
No. Observations:                1317   AIC:                                  1.276e+04
Df Residuals:                    1308   BIC:                                  1.281e+04
Df Model:                           9
Covariance Type:            nonrobust
=============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Net_Contribution         9.773e-07   4.36e-07      2.240      0.025    1.21e-07    1.83e-06
Net_Operating_Expenditure 4.196e-07  3.59e-07      1.169      0.243   -2.85e-07    1.12e-06
is_dem                      8.2753      1.687      4.906      0.000       4.966      11.585
swing                       8.0924      1.679      4.820      0.000       4.799      11.386
white                      25.1056      1.214     20.688      0.000      22.725      27.486
veteran                     6.9087      3.332      2.073      0.038       0.371      13.446
lgbtq                      -5.9794      5.919     -1.010      0.313     -17.592       5.633
stem                       -3.1563      3.057     -1.032      0.302      -9.154       2.842
obama                       3.7876      6.059      0.625      0.532      -8.099      15.675
==============================================================================
Omnibus:                      149.447   Durbin-Watson:                   1.794
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              202.205
Skew:                           0.955   Prob(JB):                     1.24e-44
Kurtosis:                       3.184   Cond. No.                     3.93e+07
==============================================================================
```

Figure 13: OLS regression results for all confounders.

After conducting an outcome regression analysis, we observed intriguing results concerning the Average Treatment Effect (ATE). The estimated ATE, using confounders that did not require significant value imputation, was 22.4036. This figure notably changed when we included all confounders, dropping to 8.2753. Initially, these results seemed to indicate that a candidate's party affiliation might significantly impact their primary percentage. This assumption was made under the conditions of SUTVA and unconfoundedness. However, a deeper dive into the bootstrap estimates for both sets of causal inferences suggests a different narrative.

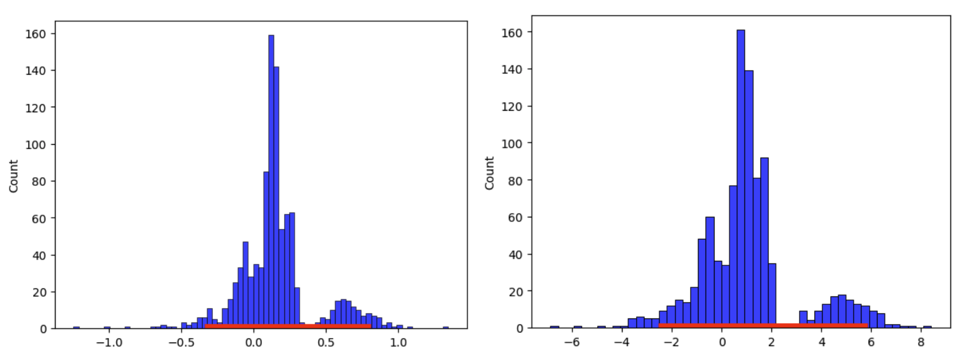

Figure 14: Bootstrapped estimates of the ATE for non-imputed confounders (left) and bootstrapped estimates of the ATE for all confounders (right).

Upon analyzing the confidence intervals derived from these estimates, we find that they both include zero. This inclusion points to the results not being statistically significant enough to confirm a causal relationship. Consequently, **we are unable to assert with certainty that a candidate's party affiliation is a decisive factor in altering their primary percentage.**

5.3 Discussion

Our analysis, though extensive, falls short of confidently establishing a causal relationship between a candidate's party affiliation and their primary percentage. The uncertainties inherent in our methodological approach, coupled with the concerns raised by the structure of our confounder variables and the inclusivity of our confidence intervals, point towards an absence of a definitive causal link. Our findings, therefore, underscore the complexity of political causal analysis and highlight the need for well-rounded data and a meticulous approach to accounting for potential biases and data limitations.

In summary, while the initial results of our analysis hinted at a possible causal relationship, the additional scrutiny and the challenges posed by our dataset lead us to conclude that the evidence does not robustly support this causal connection. This outcome emphasizes the critical nature of comprehensive data and balanced analysis to understand complex political phenomena.

5.3.1 Limitations

Our analysis encounters several notable limitations. One such limitation concerns the adherence to the Stable Unit Treatment Value Assumption (SUTVA). Despite our justification for SUTVA based on the premise that primary percentages are influenced more by voter preference than candidate actions, counterarguments exist. These arguments suggest that the campaigning strategies and actions of certain candidates might adversely impact the primary percentages of others, thus challenging the SUTVA assumption.

Furthermore, the dataset we utilized posed significant challenges, particularly in its scope and completeness. Our dataset predominantly covered Democratic candidates, leading to a substantial imbalance, especially regarding pivotal confounders like race and LGBTQ status. To address this, we imputed missing values for Republican candidates, a step that inevitably introduced bias into our analysis. This bias could have been mitigated with a more comprehensive dataset.

The absence of certain variables also restricted our analysis. Variables like candidate likeability or specific policy beliefs, which could have offered a more nuanced understanding of the electoral process, were not included. Their inclusion could have potentially strengthened our causal inference, providing a more holistic view of the factors influencing primary percentages.

6 Conclusions

6.1 Key Findings

Our analysis reveals several key findings regarding the impact of candidate characteristics on primary election funding. Party support and incumbency emerge as influential factors, with statistically significant associations even after correction procedures. Candidates endorsed by their party or with prior incumbency experience are more likely to secure increased funding. Conversely, factors such as race, swing state status, and self-funding show less robust associations, with nuanced significance levels and potential limitations.

We also found no statistical evidence to point to a candidate's party having a causal relationship with a candidate's primary %. Even after controlling for possible confounders, we were unable to show any statistical significance. The lack of a causal relationship between a candidate's political party and their 2018 primary percentage indicates the possibility of some other relationship that drives primary elections.

6.2 Generalizability

The generalizability of our results is subject to certain limitations. The dataset focuses exclusively on the 2018 primary elections, potentially limiting the broader applicability of our conclusions to different political landscapes or time periods. Additionally, the absence of data for Republican candidates in key features introduces bias, emphasizing the need for cautious interpretation specific to the dataset's focus on Democratic candidates.

6.3 Call to Action

Considering our findings, policymakers and campaign strategists should recognize the pronounced impact of party endorsement and incumbency on funding outcomes. Tailoring campaign strategies to leverage party support and highlighting incumbency advantages may enhance fundraising efforts. Further studies exploring additional variables, such as sentiment analysis on public statements or social media engagement metrics, could enrich our understanding and inform more targeted interventions.

The lack of statistical significance for a causal relationship between political party and primary percentage leads us to believe that political party does not play a significant role in primary elections. With it being difficult to make other assumptions due to the inconclusiveness of the data, utilizing the role of a candidate's political party in order to determine the primary percentage will yield lackluster results at best.

6.4 Data Sources

We merged data from FiveThirtyEight, Ballotpedia, ProPublica, and the Federal Election Commission. This combination allowed for a comprehensive examination of candidate characteristics and their financial implications. However, the absence of information on candidates' likability and media coverage limits a holistic understanding of funding dynamics.

6.5 Limitations

The dataset's exclusivity to Democratic candidates in certain features and the absence of unquantifiable information about each candidate, such as personality traits, pose limitations. The potential bias introduced by focusing on Democratic candidates underscores the need for careful interpretation and acknowledgment of the dataset's specific constraints.

6.6 Future Studies

Future studies could explore additional variables, including sentiment analysis, social media metrics, and more nuanced demographic characteristics. A broader dataset encompassing diverse political affiliations would enhance the generalizability of findings. Exploring the influence of media coverage and likability on funding outcomes could provide a more comprehensive understanding of the dynamics at play.

6.7 Take Aways

Through this project, we learned the importance of addressing dataset limitations and the necessity of cautious interpretation in the context of specific dataset constraints. The incorporation of financial data allowed for a direct exploration of funding patterns, while the visualization of key features provided valuable insights for hypothesis testing.

Additionally, we were able to gather a deeper understanding of the procedures required to determine confoundedness. Moreover, when looking at the political parties and their relationship with the 2018 primary percentage, we were surprised to find no significance. While the data we currently have does not show any causal relationship, the presence of one is still a possibility.

Overall, our analysis contributes valuable insights into primary election funding dynamics, but we acknowledge the need for further exploration and consideration of the study's limitations in future research and policymaking efforts.