DATA 102 - Final Written Report

1. Data Overview:

1.1. We are using the Dataset 3 from the given dataset options. This is an Election Dataset compiled by FiveThirtyEight with information on numerous primary elections (primary elections determine the candidates that each political party nominates for the general election) in 2018. The data contains information about each candidate, including their political leanings, endorsements, and gender, race, and veteran identities. The dataset also reports outcomes for each election.

1.2. In our analysis we are using `dem_candidates.csv` as our main dataset to answer our research questions. `dem_candidates.csv` contains information about the 811 candidates who have appeared on the ballot this year in Democratic primaries for Senate, House and governor, not counting races featuring a Democratic incumbent, as of August 7, 2018. Hence this is a census and democratic incumbents of the final position (as of Aug 7, 2018) are systematically excluded from the data. The participants were aware of this as they stood in the primaries. The granularity of the data is one candidate per row with each candidate contesting a specific role in a given state. There may be multiple candidates for the same role from the same state.

1.3. Some data that we wished we had and that we do try and account for later on include the total number of registered Democratic voters in each state. This is a key driver in determining the percentage of the vote that a single candidate may receive as it would be easier to gain a high percentage of the vote in a smaller, presumably more homogenous, voter base as opposed to a large and more diverse one. Another parameter that was not explicitly listed (but could have been computed) was the number of candidates running for each position by state as this too would play a role in the percentage of the vote that could be amassed by a single candidate. However, given the limited size of our dataset, we felt that including too many distinct features, specifically in our causal model, would result in a model that overfits to our “training” (current) dataset.

1.4. The original dataset was reasonably clean containing all the demographic identifiers for every single candidate except missing the race of a few candidates. There were several missing values in the dataset under the specific endorsement columns which corresponded to no comments being made by the respective endorser which meant there was no endorsement so we encoded these values to "No" meaning they did not receive that specific individual/organization's endorsement.

1.5. `dem_candidates.csv` includes the following columns:

Column	Description
`Candidate`	All candidates who received votes in 2018's Democratic primary elections for U.S. Senate, U.S. House and governor in which no incumbent ran. Supplied by Ballotpedia.
`State`	The state in which the candidate ran. Supplied by Ballotpedia.
`District`	The office and, if applicable, congressional district number for which the candidate ran. Supplied by Ballotpedia.
`Office Type`	The office for which the candidate ran. Supplied by Ballotpedia.
`Race Type`	Whether it was a "regular" or "special" election. Supplied by Ballotpedia.
`Race Primary Election Date`	The date on which the primary was held. Supplied by Ballotpedia.
`Primary Status`	Whether the candidate lost ("Lost") the primary or won/advanced to a runoff ("Advanced"). Supplied by Ballotpedia.
`Primary Runoff Status`	"None" if there was no runoff; "On the Ballot" if the candidate advanced to a runoff but it hasn't been held yet; "Advanced" if the candidate won the runoff; "Lost" if the candidate lost the runoff. Supplied by Ballotpedia.
`General Status`	"On the Ballot" if the candidate won the primary or runoff and has advanced to November; otherwise, "None." Supplied by Ballotpedia.
`Primary %`	The percentage of the vote received by the candidate in his or her primary. In states that hold runoff elections, we looked only at the first round (the regular primary). In states that hold all-party primaries (e.g., California), a candidate's primary percentage is the percentage of the total Democratic vote they received. Unopposed candidates and candidates nominated by convention (not primary) are given a primary percentage of 100 but were excluded from our analysis involving vote share. Numbers come from official results posted by the secretary of state or local elections authority; if those were unavailable, we used unofficial election results from the New York Times.
`Won Primary`	"Yes" if the candidate won his or her primary and has advanced to November; "No" if he or she lost.

Column	Description
`Partisan Lean`	The FiveThirtyEight partisan lean of the district or state in which the election was held. Partisan leans are calculated by finding 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.
`Race`	“White” if we identified the candidate as non-Hispanic white; “Nonwhite” if we identified the candidate as Hispanic and/or any nonwhite race; blank if we could not identify the candidate’s race or ethnicity. To determine race and ethnicity, we checked each candidate’s website to see if he or she identified as a certain race. If not, we spent no more than two minutes searching online news reports for references to the candidate’s race.
`Veteran?`	If the candidate’s website says that he or she served in the armed forces, we put “Yes.” If the website is silent on the subject (or explicitly says he or she didn’t serve), we put “No.” If the field was left blank, no website was available.
`LGBTQ?`	If the candidate’s website says that he or she is LGBTQ (including indirect references like to a same-sex partner), we put “Yes.” If the website is silent on the subject (or explicitly says he or she is straight), we put “No.” If the field was left blank, no website was available.
`Elected Official?`	We used Ballotpedia, VoteSmart and news reports to research whether the candidate had ever held elected office before, at any level. We put “Yes” if the candidate has held elected office before and “No” if not.
`Self-Funder?`	We used Federal Election Committee fundraising data (for federal candidates) and state campaign-finance data (for gubernatorial candidates) to look up how much each candidate had invested in his or her own campaign, through either donations or loans. We put “Yes” if the candidate donated or loaned a cumulative $400,000 or more to his or her own campaign before the primary and “No” for all other candidates.
`STEM?`	If the candidate identifies on his or her website that he or she has a background in the fields of science, technology, engineering or mathematics, we put “Yes.” If not, we put “No.” If the field was left blank, no website was available.
`Obama Alum?`	We put “Yes” if the candidate mentions working for the Obama administration or campaign on his or her website, or if the candidate shows up on this list of Obama administration members and campaign hands running for office. If not, we put “No.”
`Dem Party Support?`	“Yes” if 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. (Note: according to the DGA’s press secretary, the DGA does not get involved in primaries.) “No” if the candidate is running against someone for whom one of the above things is true, or if one of those groups specifically anti-endorsed or spent money to attack the candidate. If those groups simply did not weigh in on the race, we left the cell blank.
`Emily Endorsed?`	“Yes” if the candidate was endorsed by Emily’s List before the primary. “No” if the candidate is running against an Emily-endorsed candidate or if Emily’s List specifically anti-endorsed or spent money to attack the candidate. If Emily’s List simply did not weigh in on the race, we left the cell blank.
`Gun Sense Candidate?`	“Yes” if the candidate received the Gun Sense Candidate Distinction from Moms Demand Action/Everytown for Gun Safety before the primary, according to media reports or the candidate’s website. “No” if the candidate is running against an candidate with the distinction. If Moms Demand Action simply did not weigh in on the race, we left the cell blank.
`Biden Endorsed?`	“Yes” if the candidate was endorsed by Joe Biden before the primary. “No” if the candidate is running against a Biden-endorsed candidate or if Biden specifically anti-endorsed the candidate. If Biden simply did not weigh in on the race, we left the cell blank.
`Warren Endorsed?`	“Yes” if the candidate was endorsed by Elizabeth Warren before the primary. “No” if the candidate is running against a Warren-endorsed candidate or if Warren specifically anti-endorsed the candidate. If Warren simply did not weigh in on the race, we left the cell blank.

1.6. We had all our features and identifiers within our dataset, we did not need to look outward to create newer features.

1.7. All No’s in the endorsement columns were replaced with 0s and all Yes were replaced with 1s. White which is an identifier was encoded as 0 and Nonwhite as 1. All identifiers were also encoded to 0s and 1s with 0s corresponding to No values and 1s corresponding to Yes. For the cases with missing race values, they were assumed to be white (79% of representatives in the US are white according to a 2017 US News article written around the time of the primaries in question which motivated this decision)

1.8. All of the endorsement columns had missing values (NaN, None), which indicated when endorsers did not have an official opinion on the candidates. In

our analysis, we primarily focused on positive endorsements by totaling the number of “Yes” endorsements. All “No” or NaN values were omitted in our analysis.

2. Research Questions

2.1. How do individual identifiers affect a candidate’s percent of the vote?

2.1.1. The individual identifiers we tested were candidates’ race, veteran status, open LGBTQ identity, elected official status, STEM involvement, and Obama Alum status. Our test may be useful in determining how individual candidates and their qualities resonate with voters. Multiple hypothesis testing allowed us to isolate each variable and their distributions.

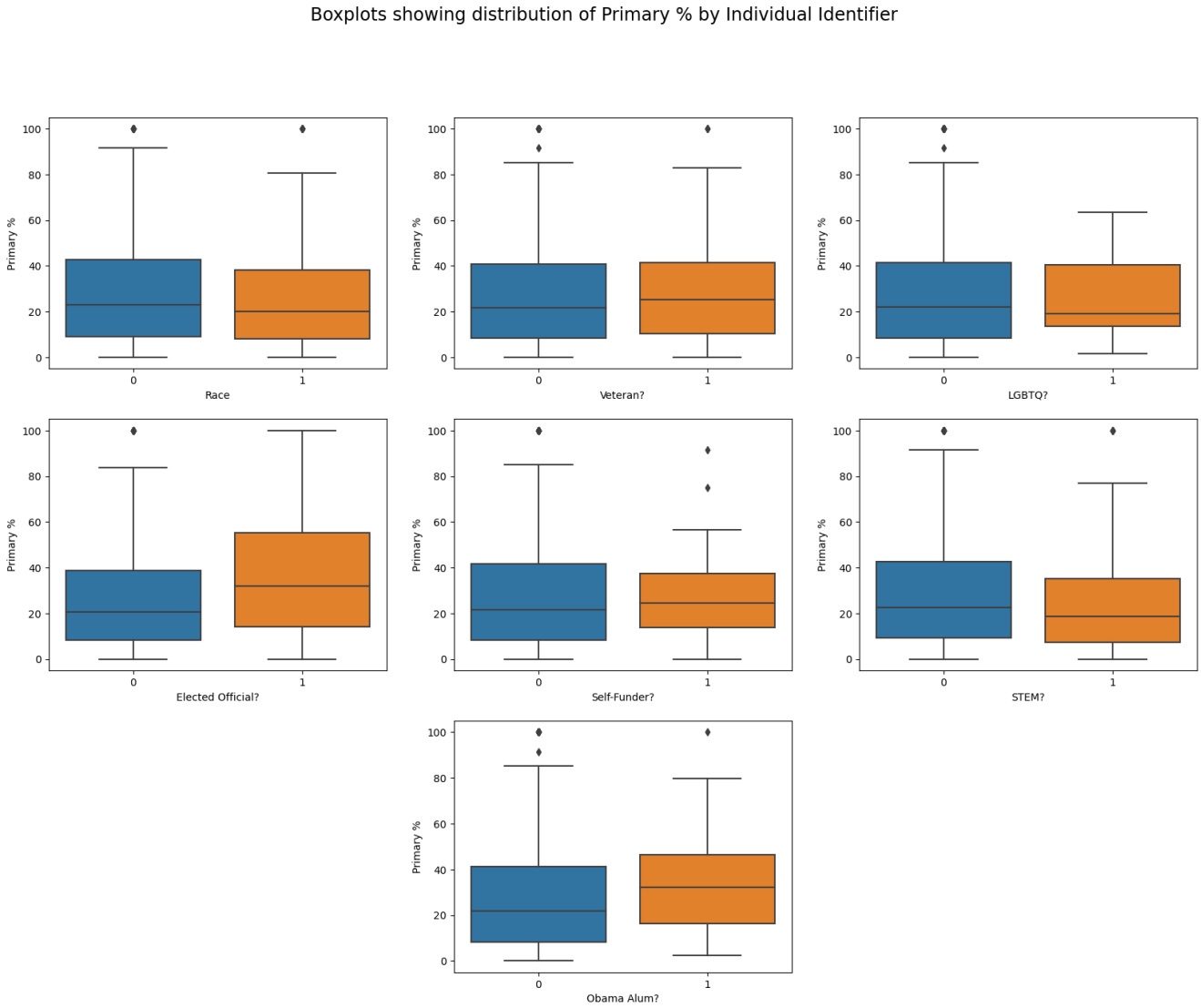
2.2. How does the number of candidate’s endorsements, partisan lean, and office type have an effect on the percent of the vote?

2.2.1. Causal Inference would be the right technique to answer this question because we can define all the treatment (endorsements), control (percent of vote), confounders (partisan lean), IV (office type) clearly and it is very easy to fit an ordinary OLS model to derive the causal relationship.

3. EDA:

We performed the following exploratory data visualizations to understand the distributions of primary vote percentages corresponding to the individual identifiers, and the distributions of primary vote percentages corresponding to endorsement.

The above side-by-side boxplots show the side-by-side distribution of the percentage of the primary vote won by candidates based on their demographic information in the original dataset. We encoded all "Yes" answers to 1 and all "No" answers to 0 for brevity and this is the data represented above



Note: For race, "White" was set to 0 and "Nonwhite" was set to 1.

Figure 1: From the boxplots above, the median percentage of the primary vote seems to be fairly similar for most of the above identifiers with a fairly even split of 0s and 1s each having a larger median. The two categories with the most variance in medians are the "Elected Official?" and "Obama Alum?" categories.

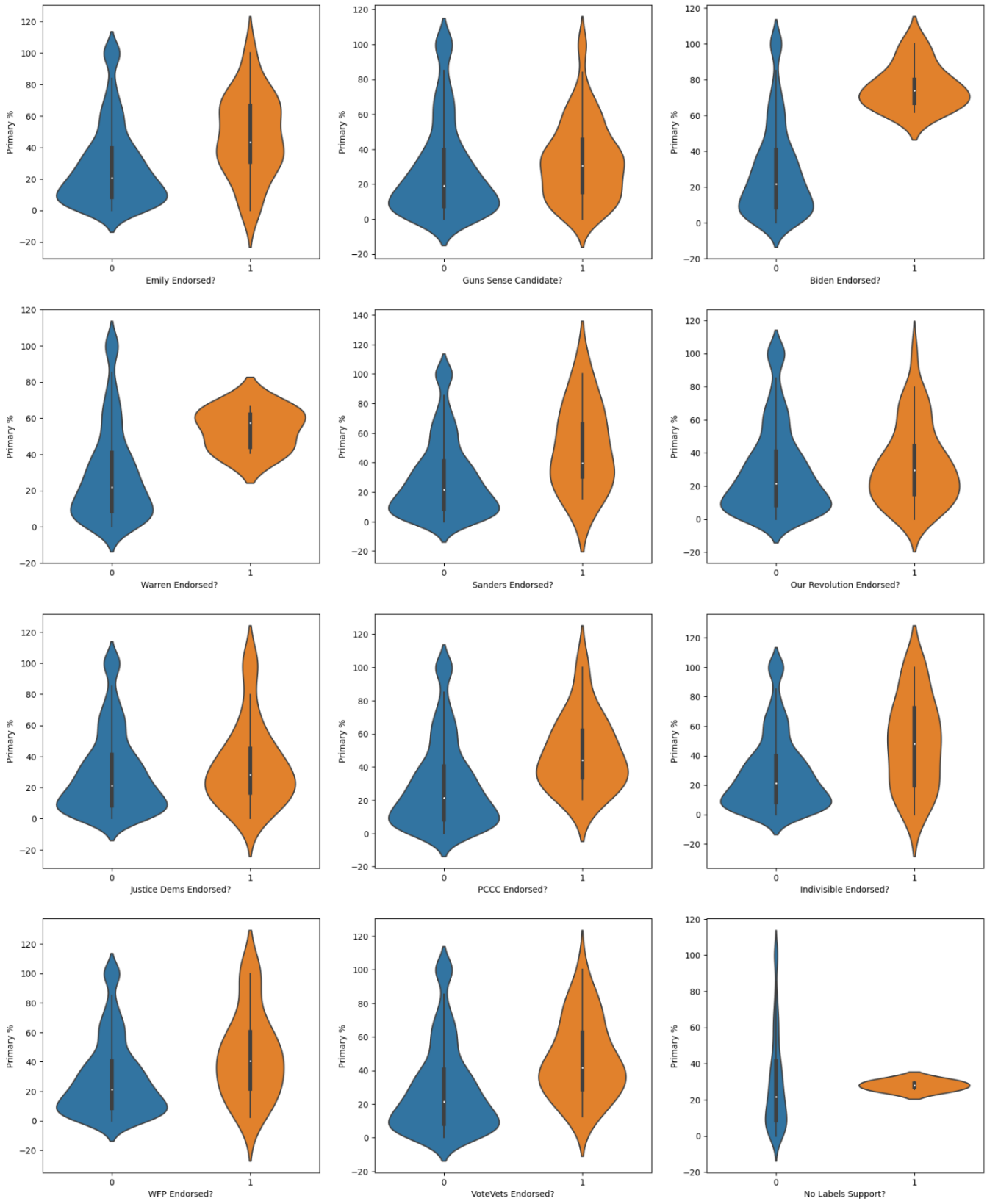
The main differences between 0s and 1s in categories can be seen in the IQRs with 1s generally having lower IQRs except in the case of elected officials. The difference is most noticeable in the "Self-Funder?" categories.

These are relevant to our (multiple hypothesis testing) second research question as this got us interested in seeing if there is a difference in distributions of votes between 0s and 1s to the above categories which we first noticed upon inspection of some of these plots.

Figure 2 (below): The below violin plots show the distribution of votes received by candidates who are each either endorsed (1) or not endorsed (0) by the above endorsers. The most striking figures in this set are definitely the "Biden Endorsed?" and the "No Labels Support?" graphs. The Biden one is of particular interest as it seems that the candidates he endorsed tended to do better than candidates endorsed by other endorsers. However, there doesn't seem to be a significant change between the distributions of "0" endorsements from chart to chart with each one having very comparable general trends and medians.

This led us to believe that the effect of a single endorsement or lack thereof might not affect a candidates final vote tally, but we were interested to see if looking at the total number of endorsements received by each candidate played a role in the final outcome or not which is the premise of our causal inference research question. This was definitely surprising to us as we assumed there would be some extremely “precious” endorsements that would carry candidates and earn them a large share of the vote and definitely motivated our direction for the second research question.

Violin plot showing distribution of Primary % by Endorsement



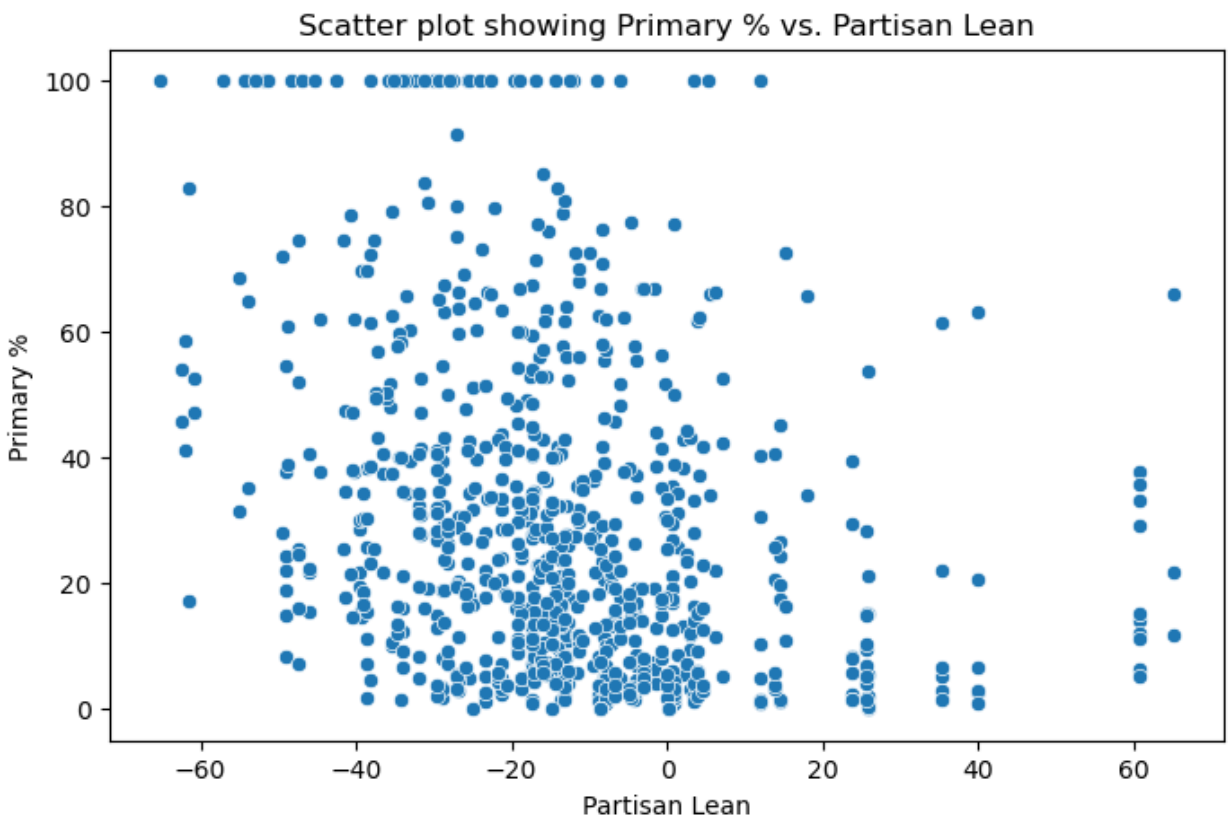


Figure 3: The above visualization looks at the Partisan Lean plotted against the percentage of the votes each candidate received in their primary election. From this, there is a clear trend of states with a very low (very negative) partisan lean (Republican states) having single candidates with an overwhelming majority share of the primary vote. This could be due to a number of reasons including fewer competing candidates for positions (fewer points in each vertical line) in Red states as well as lower voter turnout in those states for primaries. The trend visible shows a general negative correlation which is significantly strengthened when ignoring outliers.

We were initially confused about the implications of the partisan lean on the vote of a (closed) primary election which led us to thinking this might be an instrumental variable we could use in our analysis framework for our first research question which focuses on the computing the causal effect of the number of total endorsements received by each candidate. However, this graph clearly shows us that there seems to be a noticeable

negative correlation of around -0.3 between points in the scatterplot above, which is especially noteworthy when we consider that the numerous outliers were included in the calculations. This caused us to reevaluate our analysis and work in the partisan lean in our problem statement. One interesting thing we postulate is that partisan lean might in some way help us account for a major missing data point in this analysis, namely the number of voters by state. This is because there is now some distinction between states with high voter turnouts (blue states - high and positive partisan lean) and low voter turnouts (red states - very negative partisan lean) being introduced to our model.

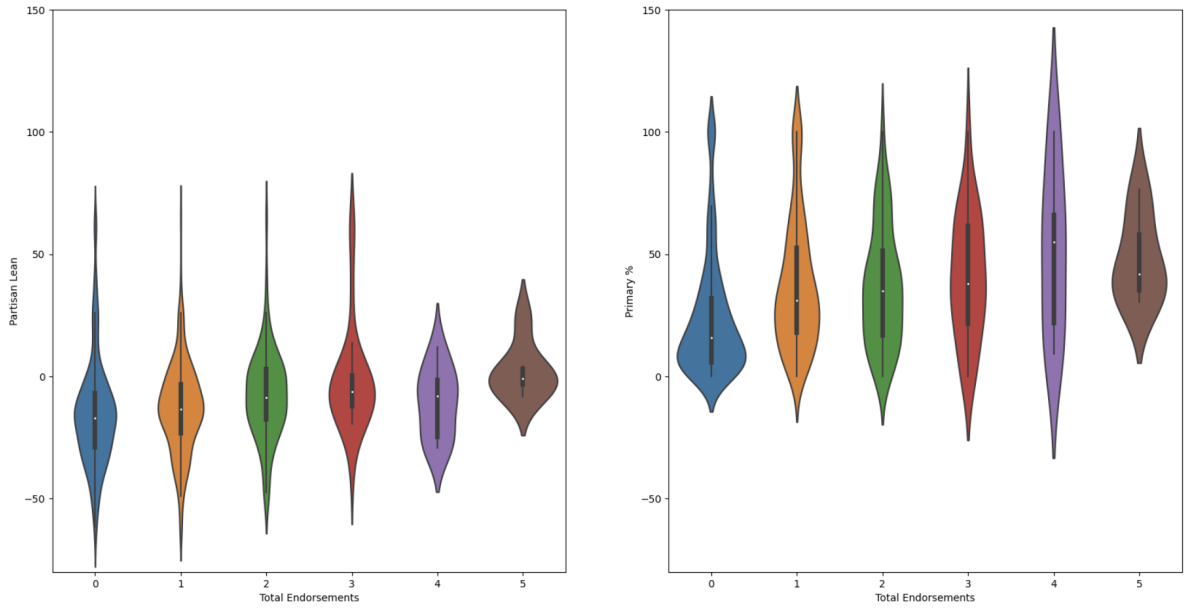


Figure 4: This plot shows side by side the relationship between the number of endorsements for a candidate along with the percentage of vote they received as well as the partisan lean of their states. Candidates from the states with the lowest absolute value of partisan lean seemed to attract the greatest number of endorsements. This would correspond with greater involvement and activity in "Swing States" that might explain this trend. This shows how partisan lean is a confounder by affecting both the treatment (total endorsements received by a candidate) as well as the outcome (percentage of vote received) as explained by the scatter plot above.

The right hand plot forms the basis for our first research question. There seems to be a sharp rise in median percentage of vote received between 0 and 1 total endorsements but endorsements past that point do not seem to substantially increase the mean. There is a strange drop off between the median at 4 (much higher than others) and 5 total endorsements which further increased our curiosity in the matter and led to us choosing our research question focusing on the total endorsements along with the partisan lean in our framework for causal inference.

4. Multiple Hypothesis Testing:

4.1. Methods

4.1.1. We conducted an A/B test on the 7 indicators in the dataset to determine if the distribution of primary vote percentage differed for indicator values. It was necessary to test many hypotheses in order to test each of the different demographic indicators.

4.1.2. In such situations (with our research question), if we were to perform a single hypothesis test, the probability of incorrectly rejecting a true null hypothesis (Type I error) increases with the number of tests performed. This is the multiple comparison problem, and it can lead to false positive results. To account for the multiple comparison problem, we used multiple hypothesis testing procedures. These procedures, such as the Bonferroni correction, control the overall false positive rate by adjusting the p-value threshold for each individual test. We thought it is important to do multiple hypothesis testing when we are testing multiple hypotheses simultaneously, so that we can control the overall false positive rate and avoid making incorrect conclusions.

4.1.3. One indicator was being an elected official, an alternative hypothesis in this test was “The average proportion of the primary vote was higher for candidates listed as elected officials compared to candidates who were not.”

4.1.4. To correct our multiple hypotheses, we used the Bonferroni correction to control for FWER and Benjamini-Hochberg to control FDR. FWER control limits the probability that our test yields a false positive. Comparatively Benjamini-Hochberg controls our expectation of making false discoveries. For the Benjamini-Hochberg correction, we controlled FDR by setting our p-value cutoff to the largest p-value under the $k(\alpha/n)$ line. Both corrections came to the same conclusions, however Benjamini-Hochberg is more appropriate for our research question because it still allows for some false positives. We do not necessarily need to eliminate all false positives in our research question. In this case, the number of discoveries made by each control procedure is the same.

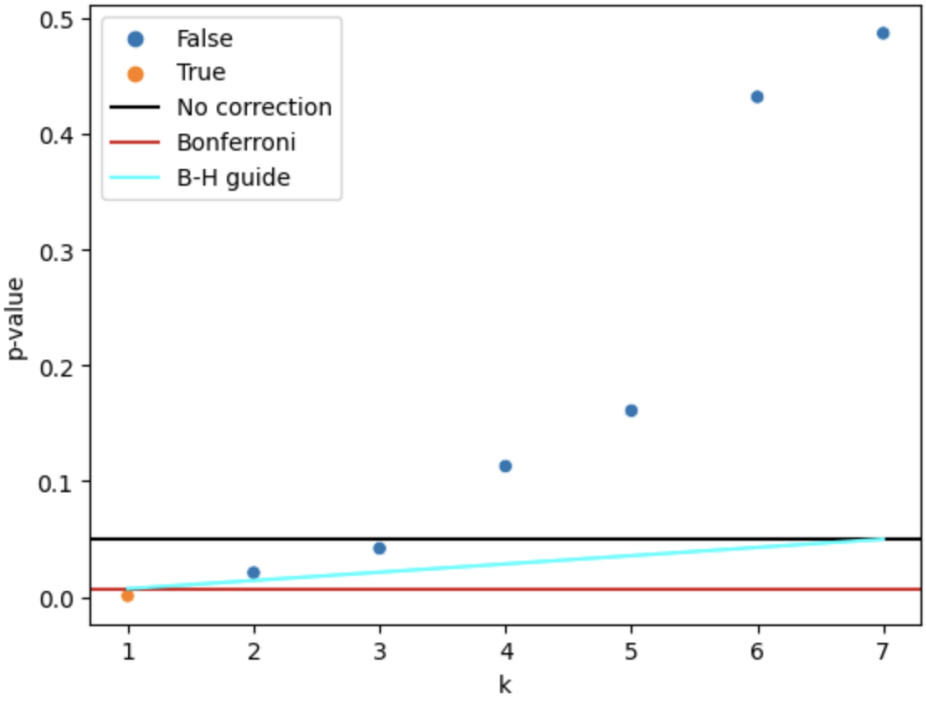
4.1.5. Assumptions we are making in the model is that there is an underlying linear relationship in the population between each of the demographic identifiers as well as the corresponding proportion of votes received by respective candidates. There is potential for collinearity amongst these features, specifically with the “Elected Official?” and “Obama Alum?” features as these are most likely going to pose a significant overlap.

4.2. Results

4.2.1.1. With the naive alpha threshold, we rejected the null hypothesis for the "Race," "STEM?," and “Elected Official?” identifiers. However with both the Benjamini-Hochberg and Bonferroni corrections, we reject the null that the Yes and No responses to the "Elected Official?" personal identifier have the same distribution of percentage vote received in the primaries. We failed to reject the null hypothesis when working with all other personal identifiers with the corrections.

4.2.1.2. We used the Bonferroni correction to control for FWER and Benjamini-Hochberg correction to control for FDR. Since we have a relatively small number of tests (7) the choice of using a specific correction does not stand to make a significant difference in the discoveries made across the tests and we see that this is indeed the

case as we reject the same hypothesis when using both the Bonferroni and Benjamini-Hochberg corrections and their respective thresholds.



4.3. Discussion

4.3.1.1. After applying our correction procedures, we rejected the null hypothesis on the Elected Official identifier's distribution of primary vote percentage. Compared to all other identifiers, a candidate's status as an elected official was associated with vote percent.

4.3.1.2. From all identifiers besides Elected Official, we failed to reject the null hypothesis that there was no significant difference in primary vote percentages. In aggregate, our results demonstrate that candidates listed as elected officials have an advantage in obtaining votes.

4.3.1.3. With more data we would want to further study the race identifier. The Race column only listed race as White or Nonwhite. Because

we rejected the null hypothesis for the Race identifier in our naive implementation, we are interested in how different races compare in primary percentage.

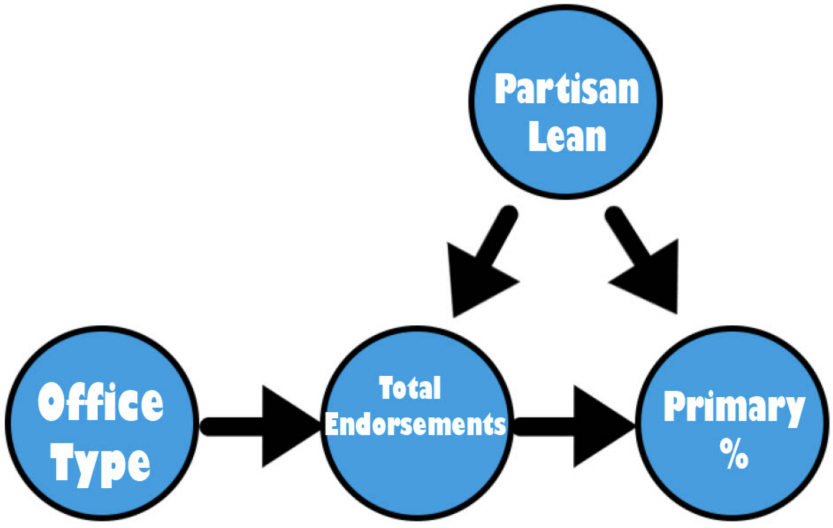
5. Causal Inference:

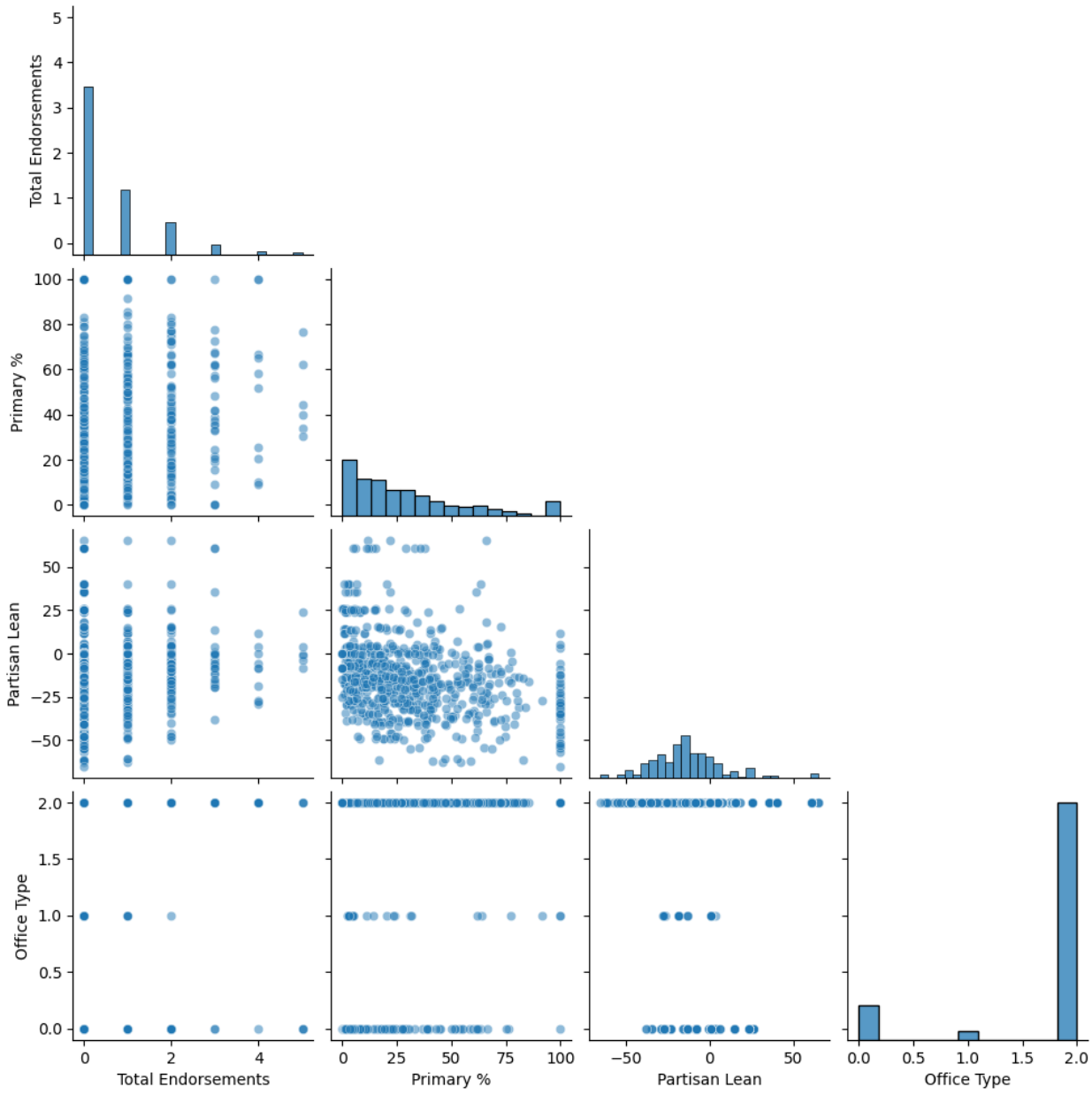
5.1. Methods

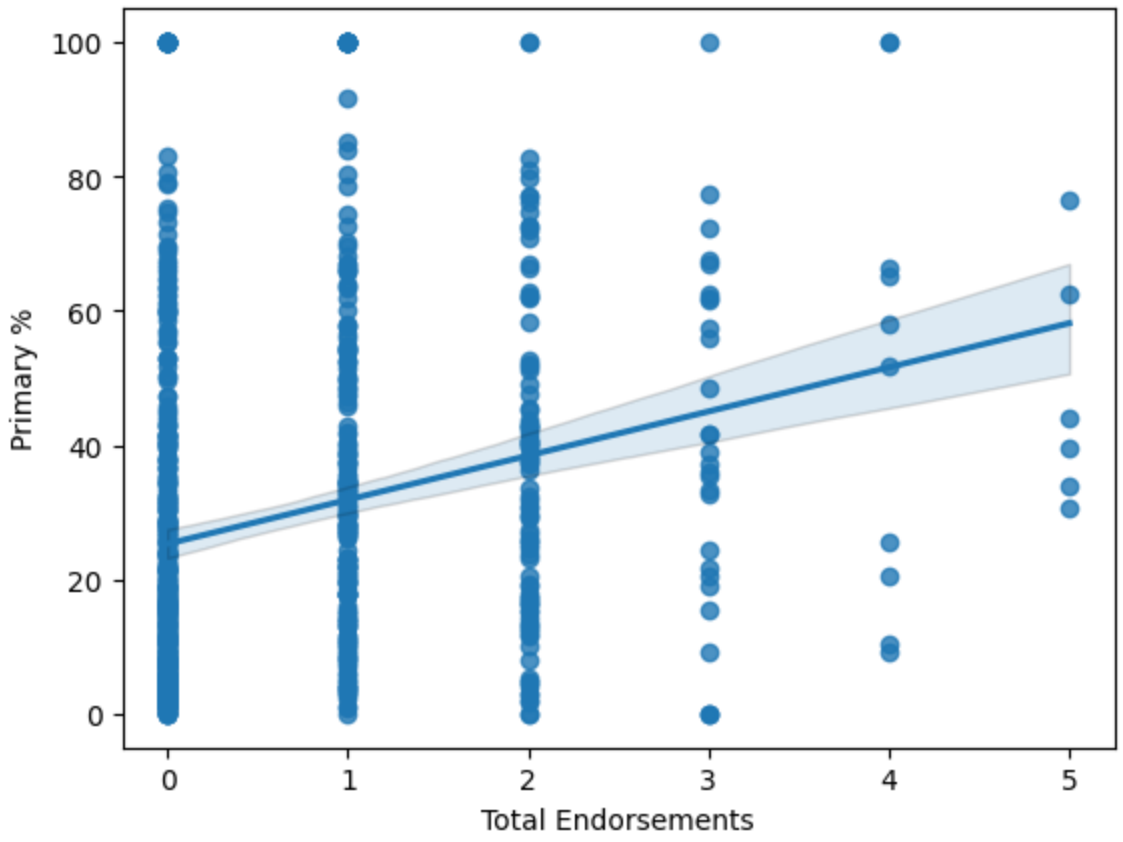
5.1.1. The treatment variable was the Total Endorsements received by the candidate and the outcome variable was the Primary Vote Percentage. A compounding variable that was factored into the approach was the Partisan Lean of the district or state in which the election was held.

5.1.2. Partisan leans are calculated by finding 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. We adjusted for this by including it in our OLS model and we also hoped to account for by including this (refer Section 3's Scatter Plot description) is the number of voters that participated in the election as it creates a distinction between Red and Blue states which will have different number of registered Democratic voters.

5.1.3. Causal DAG:




Total Endorsements
Primary %
Partisan Lean
Office Type
0
1
2
3
4
5
0
20
40
60
80
100
−50
−25
0
25
50
0.0
0.5
1.0
1.5
2.0
0
2
4
0
25
50
75
100
−50
0
50
0.0
0.5
1.0
1.5
2.0



OLS Model:

```
model = fit_OLS_model(endorsed, 'Primary %', ['Total Endorsements', 'Partisan Lean', 'Office Type'])
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Primary %   R-squared (uncentered):                   0.629
Model:                            OLS   Adj. R-squared (uncentered):              0.628
Method:                 Least Squares   F-statistic:                              456.9
Date:                Mon, 08 May 2023   Prob (F-statistic):                   1.66e-173
Time:                        22:13:22   Log-Likelihood:                         -3730.9
No. Observations:                 811   AIC:                                      7468.
Df Residuals:                     808   BIC:                                      7482.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Total Endorsements     9.3143      0.853     10.916      0.000       7.639      10.989
Partisan Lean         -0.4808      0.042    -11.563      0.000      -0.562      -0.399
Office Type            8.8318      0.643     13.739      0.000       7.570      10.094
==============================================================================
Omnibus:                      119.777   Durbin-Watson:                   1.518
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              173.994
Skew:                           1.048   Prob(JB):                     1.65e-38
Kurtosis:                       3.871   Cond. No.                         27.7
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

5.2. Results

5.2.1. The estimated causal effect of primary vote percentage of an additional endorsement is about 9.3% of the vote. The Partisan Lean has a very small estimated causal effect of a -0.5% of the vote. Finally the office type has a smaller estimated causal effect than the total endorsements but still adds 8.8% primary vote percentage depending on the office. All the coefficients in the OLS have a p-value of 0 indicating that these are all significant when computing the effect of endorsements on the percentage of votes received in the 2018 Democratic primary elections.

5.3. Discussion

5.3.1. Our method was limited in that we determined causal effect by simply taking the sum of the endorsements candidates received. Some endorsements, such as from Gun Sense, may have had a negative effect on primary vote, or alternatively a negative endorsement may have had a positive effect of the primary vote for Democrats. Similarly an endorsement from Joe Biden may have a different effect than an endorsement from Elizabeth Warren, especially if they contradict. One way to expand on our existing work would be to study the overall effect that different endorsements had on primary vote percent. For example, we could have determined weights for different endorsers rather than conducting our test with the total number of endorsements to see which individual endorsements were the most impactful in determining candidate success. That said, that is a slightly different lens of analyzing the causal impact of political endorsements on elections and is not exactly the focus of our analysis though it would be interesting to try and see if there is an interactive relationship between the number of endorsements for a candidate as well as the specific endorsements they receive as people may just be swayed by the sheer number of fancy endorsement badges in the campaign materials they come across.

5.3.2. An extremely valuable additional data point that would be highly useful for this analysis would be the total number of registered Democratic voters in each state as well as the turnouts to the primary elections. This is because the response variable we are measuring here is the **percentage** of the total vote received by each candidate which could potentially be affected by the number of voters. States with high voter turnout could see a greater variation in the percentages received by candidates and states with lower voter turnout could see greater homogeneity in preferences and could see more extreme percentage values. Another data point that could prove to be valuable for a similar reason would be the total number of candidates by state and by office. A higher number of candidates would stand to split the vote which once again could lead to an incorrect estimate of the treatment effect of the number of endorsements on the percentage of vote received by a candidate. And finally, perhaps the most straightforward omission from our model is the individual candidates' policies and general voter-friendliness looked at together (could consider adding an interaction term) with voting trends in the general populous of voters by state. While we do consider partisan lean, this does not account for the preferences of the specific subset of Democratic voters in these states and does not compare these to the specific policies of each candidate. Naturally, one can safely assume the candidate whose ideas line up most closely with that of the voting populace will receive more of the vote and hence this is another important thing to consider though will be extremely difficult to numerically quantify.

5.3.3. There stands to be a causal relationship between our treatment and outcome variables because of the "lazy voter" paradigm, a case that can be broken down into an semi-economic analysis of "cost" and "benefit". The modern voter, particularly in the case of primaries in Blue states (since we are looking at data for Democratic candidates), are faced with a multitude of options with respect to candidates for positions whose scope and responsibilities most voters do not fully understand in the first place.

Additionally, unlike in the case of highly publicized and televised presidential election coverage, a voter would have to make a conscious effort to consume local news and media to fully educate themselves towards making an informed decision when voting in such an election. The reality is that there are a large number of voters out there who may not see the value in this time investment (comparable to looking at it from a cost/benefit perspective) and hence may either make an uninformed decision or abstain from making a decision at all (election day no-shows). This is where political endorsements play a major role in affecting the outcome of elections such as these. Endorsements, particularly from high profile Presidential candidates like Warren, Sanders, and Biden could sway their respective supporters to be reassured and confident of voting for their endorsees without complete knowledge of the endorsee's own platforms. This has historically alleviated voters. burden of actively seeking out information about all available candidates and stands to encourage greater voter turnout overall. While the high-profile endorsements are usually the most valued, we were interested to see if this relationship holds even when taking a raw count of the number of endorsements a candidate has which stands to sway voters who might assume a list of endorsements enhances a candidate's credibility.

6. Conclusion:

6.1. In studying our first research question, we determined that out of all candidate identifiers, candidates who are already elected officials tended to have a higher proportion of the vote. For our second research question, we determined that additional endorsements have an effect on primary vote percentage.

6.2. Our findings are limited to the data on Democratic candidates from 2018. The political landscape is constantly changing, and voter trends are often enforced by the preceding years of voting.

6.3. We did not merge data sources because the dataset contained all features we needed to conduct our analysis.

6.4. Future studies could build on our work by further examining the effect of state partisan lean and historic trends of the Democratic party primaries across states. Additionally, further study of partisan politics in states would inform the effect of different endorsers on candidate success.

6.5. Political campaigns should prioritize establishing credibility for candidates. In our research we determined that previous experience as an elected official and endorsements are associated with greater voter support. Other factors such as race and state partisan lean were found to be less indicative of vote percentage and indeed in some cases detrimental to the same. Keeping this in mind, building political campaigns highlighting a candidate's demographic information might not be the optimal strategy. Furthermore, obtaining endorsements from politicians in office and Democratic groups can be beneficial in increasing votes.

6.6. During the course of our work on the project, perhaps the most challenging portion came in the causal inference design portion when we couldn't work out exactly how the Partisan Lean might play a role in a primary election since these are restricted only to registered partisan (in this case Democratic) voters. This made our choice of instrumental variable extremely difficult, though once we recognized the lean's role as a confounding variable, we were able to account for it and use the office type as an instrumental variable as we discovered in our analysis that this is the feature that had a significant causal impact on the number of endorsements a candidate receives as positions with greater spheres of influence attract greater publicity and attention encouraging more support and endorsements. We also learnt, unsurprisingly, that previously elected officials were more likely to get a higher proportion of the vote in their respective races. An interesting thing we took time to conduct independent research on was the exorbitantly high (90%+) win-rate across primaries in the US which seemed to validate the results of our hypothesis tests as well, and it would be interesting to see the stark contrast in these results compared to those that were excluded due to a Democratic incumbent.