Data 102 Project Written Report

Kevin Lu, Winston Mok, Saransh Rakshak, Jacob Rodkiewicz

Data Overview

All of our data outside of information on candidate specific contribution amounts came from a dataset *FiveThirtyEight* collected in 2018. They identified all candidates who appeared on the ballot for Democratic and Republican primary races for Senate, House, and governor between January and August of 2018, and collected data about each candidate they deemed to be relevant for holistic analysis. Beyond information relevant to each candidate's election (location, type of election, percentage of votes received, outcome, etc.), the dataset also contains more candidate specific descriptors such as gender, race, sexual orientiation, and endorsements. Most of this data was collected through publicly available, online sources of information, including Ballotpedia, VoteSmart, news reports, and the candidates' own campaign websites. This method of data collection was not systematically exclusive of any groups of candidates because it included general election information on every candidate, however, there were gaps in the data in specific fields such as 'Race' and 'Gender,' which if not found on the candidate's website, were not included.

Our second source of data came from the Federal Election Commission and provided us with candidate specific contribution amounts between the years of 2017-2018 for most of the candidates addressed by the *FiveThirtyEight* dataset. This information is entirely public and is required to be reported to the FEC by every candidate appearing on an official state ballot, and so we can reasonably conclude that there are no groups excluded in the collection of this data. The dataset provides an in-depth breakdown on the specific sources of the contributions made to each candidate, however, we only consider the column of "Total_Contribution" in our causal inference guided analysis. Perhaps in future studies we could use this extra information to compare the effect of each of the sources of contributions with each other. In addition, some candidates may use the assistance of outside groups not directly related to their campaign, such as PACs or political parties, that may externally assist with campaign efforts. Some candidates may also be

hiding other amounts of money that their campaigns use. Therefore, it is possible that the campaign finance data is biased to be lower than what it actually should be.

Some important features we wish were included would be those identified as additional confounders in the context of our causality research question. Limitations in this respect are addressed in more detail within the 'Discussion' section of this study, but a couple features outlined include a candidate's level of fame and the state of the political environment surrounding a candidate's election.

Research Questions

Multiple Hypothesis Testing/Decision Making

Our first research question is to find out if specific qualities of a candidate, including Partisan, candidate, and group support, race, whether or not they were a veteran, etc., influenced whether they won in the primary election. By finding p-values, we can note correlations between a candidate winning their primary election and their qualities and the people that support them. By answering these questions and using this information, we can give more accurate predictions for whether or not a candidate will win their primary election based on their attributes and the different groups supporting them.

Using multiple hypothesis testing is a good fit for our data because we are able to individually test each attribute, by finding their specific p-values and look at them holistically by grouping them into different attributes that a candidate might have. In our analysis, we are able to take advantage of this by grouping together different Candidate Qualities, whether or not they are supported by a specific Presidential candidate, and whether they were supported by a specific group. By doing this and looking at the individual p-values, we are able to use multiple hypothesis testing in order to determine if any of the categorical variables listed above had any significant differences in winning their primary elections.

Causality

Our second research question is if the amount of donations a candidate receives has a causal effect on the percentage of votes a candidate receives in their primary election. In answering this question, we would confirm or disprove the commonly held belief that campaign contributions do in fact have an effect on the outcome of an election. Furthermore, if a causal association is identified, we would more or less be able to quantify the effect that an X amount of dollars of campaign contributions has on the success of that political campaign. This is important information for both political candidates and donors to have in terms of being able use their time and resources efficiently and effectively. Likewise, these findings would be useful in answering broader questions of whether money is playing too big an influence on the state of politics in general, and further analysis could be conducted comparing the effect of campaign donations to that of other variables contributing to a candidate's success.

We will be using the method of causal inference in answering this question, because more so than trying to simply establish an association between donations and political success, we are trying to establish *causation* of the former on the latter. An association between our two variables of interest is relatively easy to establish; however, it isn't very informative. An association tells us nothing about a causal effect, and especially with this being an observational study, there are many confounding variables whose effects must be accounted for via the methods of causal inference in order to isolate the effect that donation amounts have. Specifically, we have chosen the causal inference method of inverse propensity weighting as opposed to other popular techniques such as outcome regression and matching.

In the case of outcome regression, a number of issues arise due to the assumptions which must be made in the context of our problem. For one, outcome regression requires there to be a linear association between our features and the outcome variable, and with the number of confounders being considered, there are likely to be complications in this respect. Furthermore, a linear model assumes that confounders have a comparable effect between our test and control groups, which again is a very difficult assumption to make in good faith. On the other hand, the causal inference technique of matching data points across control and test groups falls short due to limitations in the amount of data available to us. With our model considering X amount of confounders on a

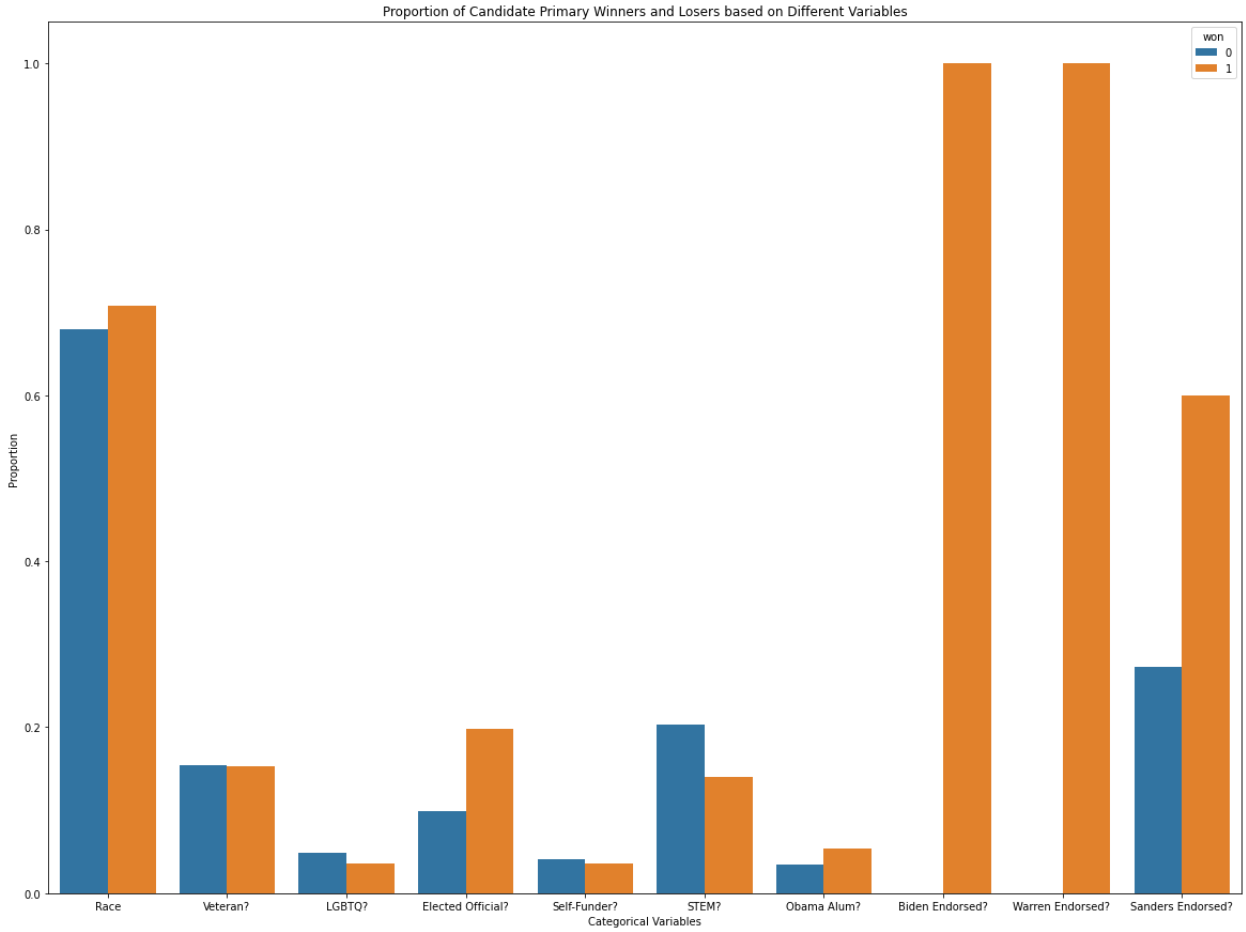
dataset of only 500 or so data points, exact matches between our test and control groups will be few and far between, and thus this method will likewise be ineffective in estimating a causal effect. Inverse propensity weighting, on the other hand, is a technique whose effectiveness is far less affected by violations of the aforementioned assumptions, and as such, is our method of choice for answering the question.

EDA

Below are some of the steps we performed to clean our data:

- Imputed missing values in the "Won Primary" field of our candidates donation data with values from 'Primary Status' field. This did not impact the accuracy of our data.
- Removed rows from FiveThirtyEight table with missing values across multiple fields of variables (e.g. all candidate features were missing). Very little data was lost in this process.
- Binarized most columns by replacing “Yes” and “No” values with 1’s and 0’s so as to enable analysis in the form of outcome regression and inverse propensity weighting. For some features, such as endorsements, there was a difference between a group or politician making no endorsement (not having an opinion on the race) and anti-endorsing (supporting a different candidate in the race). We binarized both of these scenarios as 0. This does affect the quality of our data, but we felt that doing so simplified our analysis.

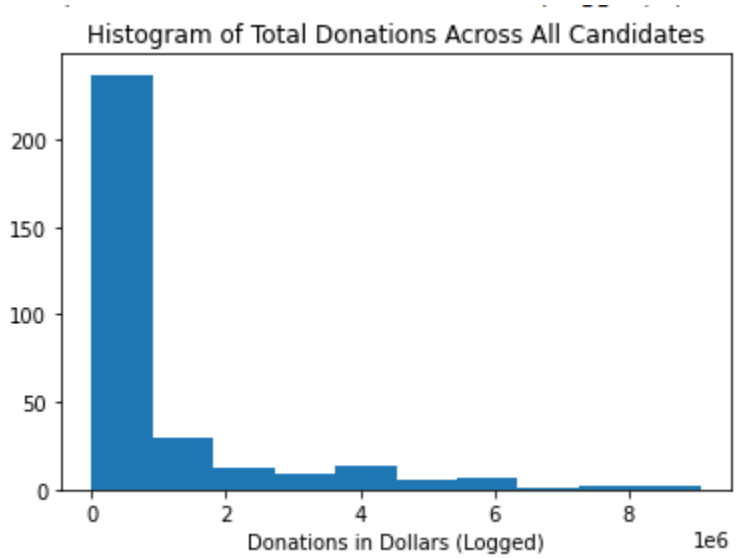
After performing data cleaning, we wanted to examine the distribution of the data. First, looking at candidate features and politician endorsements, we found the relative proportions of winners and losers:



As seen from the data above, we can see the most prominent differences in proportions between winners and losers occurred for the presidential endorsed candidates. This means that when the presidential candidates endorsed someone, they were very likely to win their Primary. Everyone endorsed by Biden and Warren won their primaries and 2/3 of the candidates endorsed by Sanders won their primaries. In terms of the other categorical variables, it seems like there were no major differences. The most notable ones being that if they were an elected official before they won their primary more often and if they were STEM they lost their primary more often.

This visualization is directly related to our research question asking what variables affect if a candidate lost or won their election. We can see a rough estimate that presidential endorsements have a very large effect and should dive into this deeper. We should also look at the effects of the other categorical variables, like Elected Official and STEM, and see what the exact numerical effects on each are.

Next, we chose to combine the FiveThirtyEight dataset with one about candidate contributions in the 2018 primary. Plotting the log of the total contributions to each campaign:

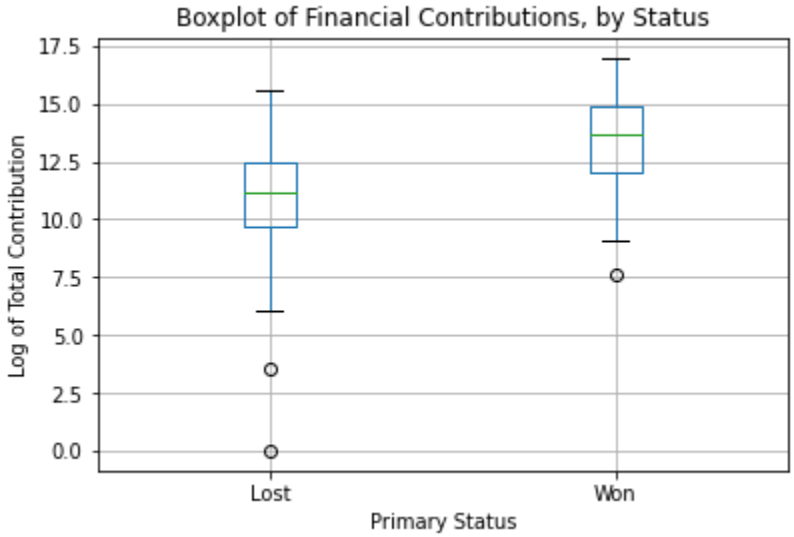


This histogram helps us understand the distribution of total donation amounts across all of our candidates. The visualization informs us that the distribution is skewed, with most donations received being under the magnitude of 1 million dollars. If we decide to analyze donations received as a categorical variable instead of a continuous variable in answering our question of whether donations received have a causal relationship with polling percentages, then we know that 1 million dollars serves as somewhat of an intuitive cutoff point between the two categories of "less funded" candidates and "more funded" candidates.

We then combined the two datasets in order to find other relationships existing. To manage the number of variables, we limited our analysis to Democratic candidates who had a total contribution greater than zero. We joined the datasets by finding all rows where the state, last initial, and first initial all matched.

We also wanted to see how contribution amounts differed based on whether candidates won or lost. First, we found the log of total contribution amounts, in order to make processing and visualization easier:

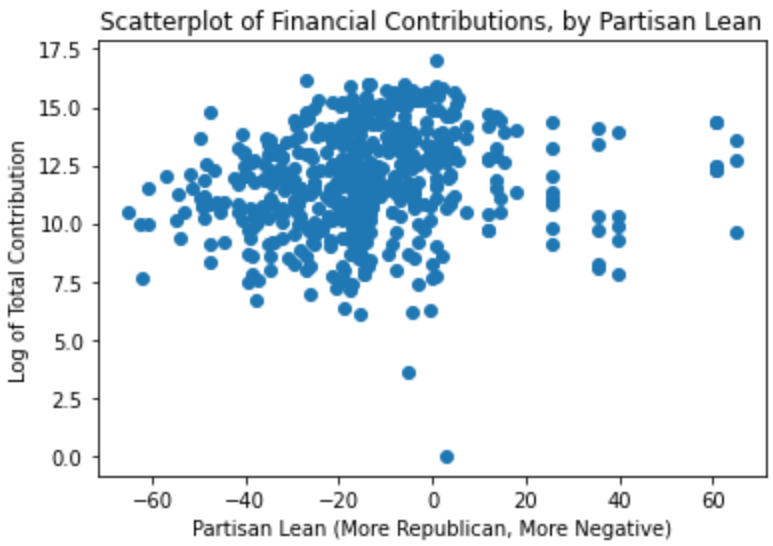
We then generated a boxplot, separated by whether candidates won or lost:



From the plot above, it's clear that, unsurprisingly, candidates who won had on average a larger contribution amount than candidates that lost. It's interesting to note that, amongst candidates that lost, the variation in amounts is much larger than amongst candidates that won. This is also unsurprising; it's likely that some failed candidacies never really got off the ground, while others were serious contenders.

While this plot provides associations between contribution amounts and winning, the relationship between the two is unclear. More contribution amounts likely do cause candidates to be more likely to win, but perhaps there are confounders, such as an endorsement or a feature of the candidates, that would both increase the amount of donations, and increase likelihood to win. To find causality, we need to investigate further.

It would also be interesting to examine financial contributions based on the partisan lean of the district or state the race is in. The plot for this is below:



It seems that there are overall more candidates in races with a negative partisan lean (are Republican). This was somewhat surprising; we expected that, as Democrats have a higher chance in winning races in Democratic areas, it would be more popular for candidates to join, with greater hopes of winning. It seems that the relation between partisan lean and contributions is not clear. It may be that the relationship is quadratic, with low contributions to especially partisan areas, and higher contributions in more competitive areas.

Is partisan lean a variable that influences campaign contributions? While the graph does not show a strong relationship, intuitively there should be greater contributions to more competitive races. With winning races in mind, it is also possible that more moderate districts are more likely to nominate more moderate candidates. Thus, in investigating the relationship between contributions and winning, partisan lean seems to be a confounding variable. Further research is needed to determine this.

Research Question 1: Multiple Hypothesis Testing

Methods

CleanFrame(*pd.DataFrame dataframe****)***

Assigns binary allocations to categorical variable values. In this function, ‘Yes’ is mapped to binary value **1**, and ‘No’ is mapped to **0.**

Candidate Attributes and Primary Wins

Democrat data *dem_cat* was the only data present with categorical variables that described the candidates personal attributes and background. We isolated columns ‘Won Primary’, ‘Obama Alumn’, and ‘Race’- allocated to **1** for ‘non-white’ and **0** if the candidate was ‘white’ or no data was available. *CleanFrame* was run, and the returned DataFrame was stored as a new DataFame *dem_cand_attributes*.

Political Figure Support and Primary Wins

For finding correlations between political figure support and primary election wins, we isolated columns ‘Won Primary’, ‘Biden Endorsed?’, and ‘Sanders Endorsed?’ from *dem_cat*. Similarly, columns ‘Won Primary’, ‘Trump Endorsed?’, and ‘Susan B. Anthony Endorsed’ from *rep_test*. *CleanFrame* was run on both the isolated frames and stored as *dem_people_endorsements* and *rep_people_endorsements*, respectively.

Group Endorsement and Primary Wins

'Won Primary', 'Party Support?', 'PCCC Endorsed?', and 'WFP Endorsed?' are used from *dem_cat*. Similarly from *rep_test* we isolated columns ‘Won Primary’, ‘Rep Party Support’, ‘NRA Endorsed?’, and ‘Right to Life Endorsed’. *CleanFrame* was run on both, and saved as *dem_group_support* and *rep_group_support* respectively.

Hypothesis Tested:

Hypothesis #1: Is there a substantial difference in the proportion of Democratic Primary Winners between those that were listed as White and Nonwhite?

Hypothesis #2: Is there a substantial difference in the proportion of Democratic Primary Winners between those that were Obama Alumni (worked for the Obama Administration) and those who were not?

Hypothesis #3: Is there a substantial difference in the proportion of Democratic Primary Winners between those that were directly endorsed by Joe Biden and those who were not?

Hypothesis #4: Is there a substantial difference in the proportion of Democratic Primary Winners between those that were directly endorsed by Bernie Sanders and those who were not?

Hypothesis #5: Is there a substantial difference in proportion of Republican Primary Winners between those that were directly endorsed by Donald Trump and those who were not?

Hypothesis #6: Is there a substantial difference in proportion of Republican Primary Winners between those that were directly endorsed by Steve Bannon and those who were not?

Hypothesis #7: Is there a substantial difference in proportion of Democratic Primary Winners between those that were directly endorsed by the Democratic Party and those who were not?

Hypothesis #8: Is there a substantial difference in proportion of Republican Primary Winners between those that were directly endorsed by the Republican Party and those who were not?

Hypothesis #9: Is there a substantial difference in proportion of Democratic Primary Winners between those that were directly endorsed by the Progressive Change Campaign Committee (PCCC) and those who were not?

Hypothesis #10: Is there a substantial difference in proportion of Democratic Primary Winners between those that were directly endorsed by the Working Families Party (WFP) and those who were not?

Hypothesis #11: Is there a substantial difference in proportion of Republican Primary Winners between those that were directly endorsed by the National Rifle Association (NRA) and those who were not?

Hypothesis #12: Is there a substantial difference in proportion of Republican Primary Winners between those that were directly endorsed by the National Right to Life Committee and those who were not?

We decided to use multiple hypothesis testing for the data set because of the amount of endorsements and other attributes that may affect if a candidate won or lost is large. There are so many attributes and endorsements that exist, that we wanted to use Multiple Hypothesis testing to get an accurate view if each variable mattered.

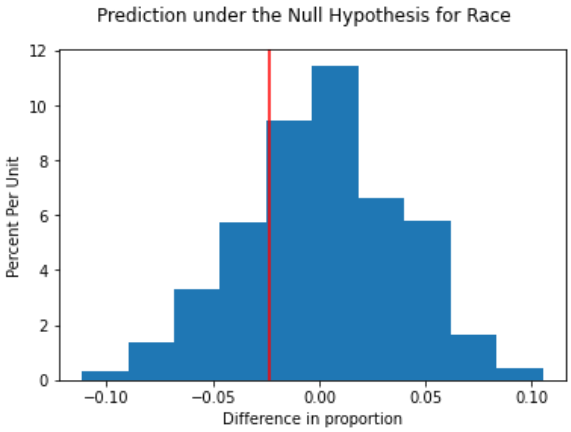
Each hypothesis above was tested by using A/B testing. We decided to use this method to find Empirical P-values, because we wanted to test if the proportion of people that won their primary with a specific attribute originated from the same distribution. By shuffling the labels of wins and losses and using each categorical variable with a new label to create the distribution, we were able to see if the two attributes' (Nonwhite vs. White) proportions of winners was different enough to come from 2 different distributions.

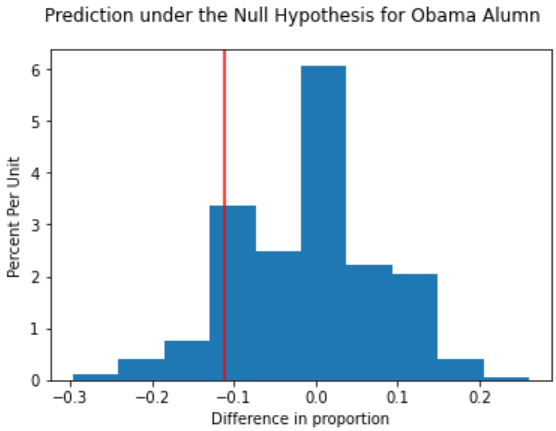
As for controlling the error rates for the multiple hypothesis tests, we decided to use two different methods, the Bonferroni correction and the Benjamin-Holchberg correction. The Bonferroni correction is much more conservative and guarantees a Family Wise Error Rate (FWER). We decided to use this as one of our corrections because when choosing candidates to run for office because of the few chances that parties may have, they may want to only find the qualities that are 100% true positives, or as close as they can be. However, for a less conservative approach, the B-H correction controls the False Discovery Rate. Even though it is not as restrictive, by correcting based on the FDR, we are able to take into account expectation into the randomness of our decisions.

Results

Candidate Attributes and Primary Wins

We first chose to compare democratic candidates Primary Wins with 'Race' (white or non-white), and if the candidate was an 'Obama Alum'. 'Race' produced a p-value of 0.28, and 'Obama Alum' got a non-zero score as well of 0.171. In both cases we do not reject the null hypothesis with a p-value of 0.05. Our predictions under the Null hypothesis are shown as:

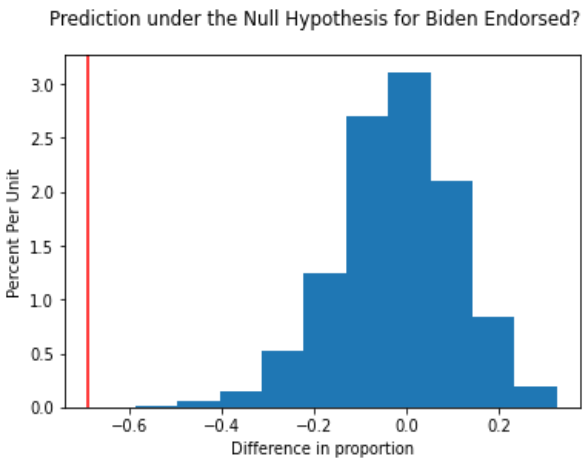


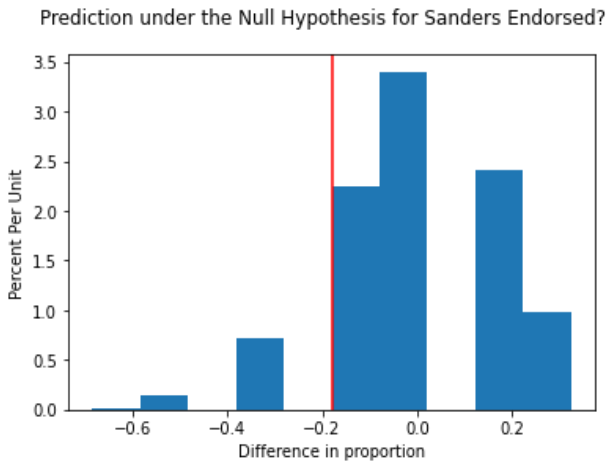


People Endorsements and Primary Win

(Democrat)

In comparing democratic candidates' Primary Wins with endorsements by other significant democratic politicians, we chose ‘Biden Endorsed?’ and ‘Sanders Endorsed?’ to see if endorsement by either Joe Biden or Bernie Sanders played a role in Primary Wins. Biden Endorsement found a score of 0.0, while endorsement by Bernie Sanders got a non-zero score of 0.298. With a p-value of 0.05, we would not reject the null for Biden Endorsement and would reject the null for Sanders endorsement. Group Support for democratic candidates gives the following predictions:

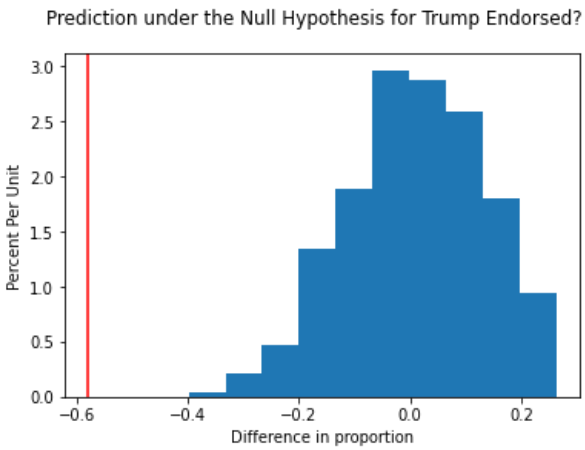


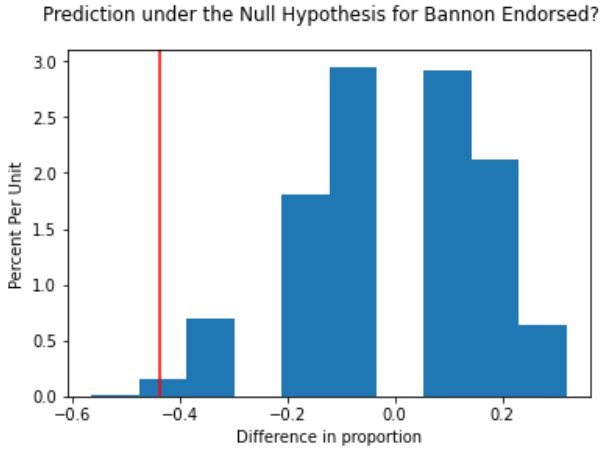


(Republican)

We chose to focus on endorsements by Donald Trump and Steve Bannon, the White House Chief Strategist during Trump’s Administration. ‘Trump Endorsement?’ provided a score of 0.0, which makes sense since we expect a republican candidate who is endorsed by the president to win over

another republican candidate competing for the same seat but is not endorsed by Trump. Our score for 'Bannon Endorsed?' is a non-zero value, 0.009, but is still extremely low, for reasons similar to that of Trump endorsements. With a p-value of 0.05, we would reject both null hypotheses.

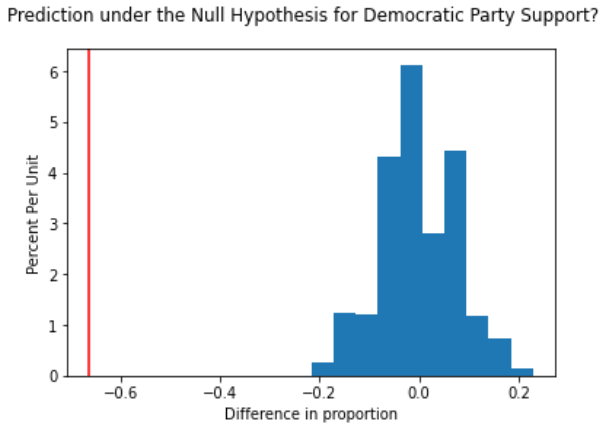


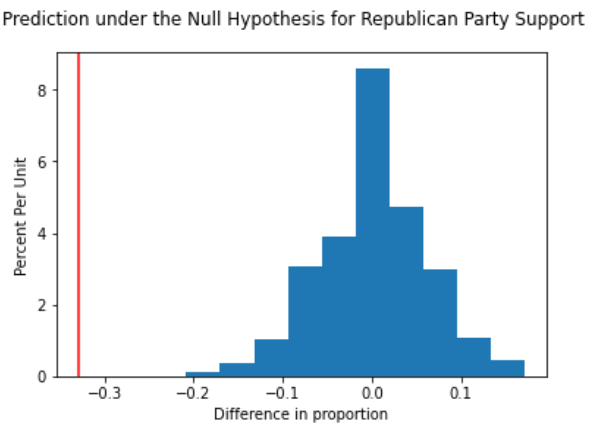


Group Endorsements and Primary Wins

'Party Support'

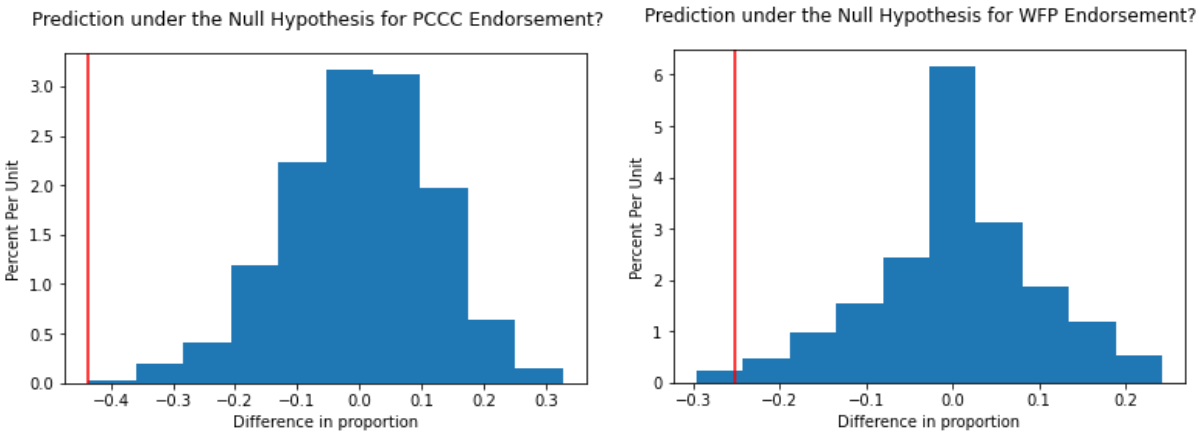
When comparing 'Party Support' for the democratic candidates- we are able to find a p-value of 0.0. This makes sense, as we can expect that a democrat endorsed by their party is more likely to win for a particular position than another democratic competing for the same position but is not officially endorsed by the party. Similarly, for the republican candidates, we get an extremely low p-value of 0.009, for the same reasons as before. With a p-value of 0.05, we would reject the null for both hypotheses. Party Support for democratic and republican candidates yielded a prediction as follows:





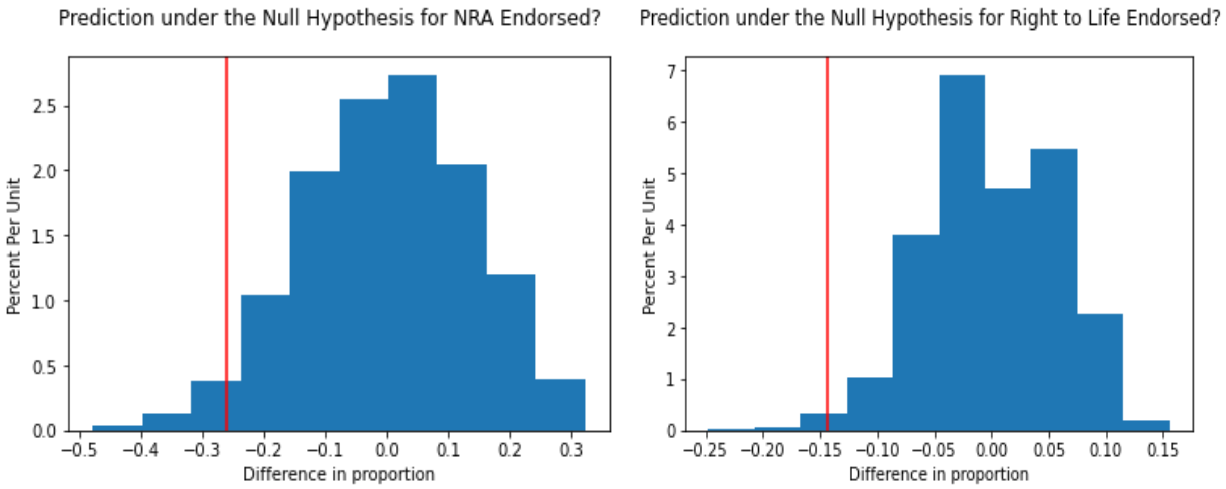
(Democrat)

We wanted to also see if 'PCCC Endorsement?' and 'WFP Endorsement?' had an effect on primary wins. We found a p-value of 0.003 with respect to PCCC support and WFP support yielded a p-value of 0.009. With a p-value of 0.05, we would reject the null for both hypotheses. The predictions under our Null Hypothesis are plotted as following:



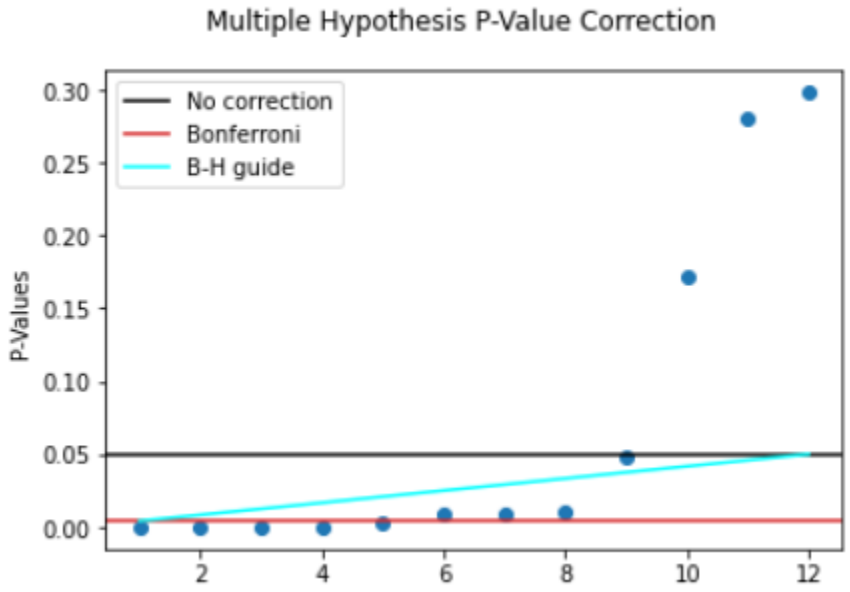
(Republican)

For our republican candidates we sought to find NRA endorsement and Right to Life endorsement on a candidate's Primary Win. We chose these two because laws regarding gun ownership and abortion are currently very controversial in our country. As the republican stancepoint can be seen to lean pro-firearms and anti-abortion, we gave a score of **1** to those endorsed by either group. We found a p-value of 0.048 for NRA and a score of 0.01 for Right to Life. With a p-value of 0.05, we would reject the null for both hypotheses. Our predictions are seen as follows:



Correction Procedures

As stated above, we decided to use two different correction techniques: the Bonferroni Correction and the Benjamin Hochberg Correction, both of which have their advantages and disadvantages. The Bonferroni Correction came out with a p-value threshold of 0.004. The Bonferroni Correction guarantees a FWER of 0.05 with this threshold. On the other hand, the Benjamin-Hochberg correction came out with a p-value threshold of 0.01 (the largest p-value under B-H guide). The Benjamin-Hochberg correction controls the False Discovery Rate, making it a less conservative approach than the Bonferroni.



Discussion

After applying the correction procedures, for the Bonferroni Correction, the remaining discoveries for Biden Support, Trump Support, both Party Supports, PCCC Endorsements. For the B-H Procedure, the remaining discoveries are the same as above, including Bannon Endorsed, WFP Endorsed, and Right to Life Endorsed. As seen from the more conservative correction with the Bonferroni, people that were endorsed by both presidential candidates or parties won their election. This means that we can assume that people that are endorsed by a party or presidential candidate may have a higher chance of winning in their Primary Elections. Expanding to the significant Discoveries under the B-H correction, we can see there was a significant difference in those that were endorsed by specific groups that had similar interests, such as the PCCC for democrats and Right to Life for Republicans, in terms of Winning their Primary Election. However, we can see that candidates supported by the NRA, with a p-value of 0.05, failed to reject the null. Our group hypothesized that maybe the NRA spent a significant amount supporting all types of candidates rather than strategically supporting candidates like the other groups and presidential candidates.As for both being an Obama Alumni and Race, for the Democratic Party, we were unable to find any significant differences in the win proportion for the Primary Elections, between the respective options.

Additionally, we generalized each hypothesis, grouping them into specific categories like Attributes, Presidential Candidate support, and Group support. Looking at the p-values for each, reinforces our idea that people that were endorsed by a Presidential candidate had a high correlation with winning their primaries. We also noticed similar patterns in those endorsed by specific groups, but not to the extent that of the presidential candidates endorsements.
One limitation of our analysis is that we can not conclude direct causality from the hypothesis tests. We do not know whether the candidates won because they were supported by the Presidential Candidates or Groups (whether that through public opinion or money) or the groups and Presidential Candidates especially, chose to support those that they already knew would win their primary elections. Most likely it is a combination of both of these options, but with multiple hypothesis testing we are unable to find out. Another limitation of our model was the lack of categorical variables describing background, race, or any other personal attributes for the republican candidates. If we were able, these would be something we would explore further. This

way we can see the differences between parties and see what people within each party value more.

Research Question 2: Causality

Methods

We set out to identify a causal effect of contribution amount (donations) on a candidate's percentage of votes received, and we've initially approached the question with the two techniques of outcome regression and inverse propensity weighting. With both methods we considered the same confounding variables with those being mostly binary features such as 'Veteran?,' 'LGBTQ?', and 'Elected Official?', to name a few. Our outcome variable was the percentage of votes a candidate received in their primary election, and the treatment variable studied was whether a candidate received a ‘large’ contribution (>$1M) during or before their campaign.

We did not know which other features we had access to were confounders, so we decided to perform our analysis with the idea that every other factor was a potential confounding variable. This included partisan lean of the state or district, candidate features, and candidate endorsements. Since the variable of interest was the primary percentage of the vote, it cannot be susceptible to other variables being colliders- the primary will occur after all other variables have been set, so the primary percentage cannot ‘cause’ any other variable.

We will perform a causality analysis through two different methods: multiple regression and inverse propensity weighting. With regression, if we have adequately captured confounders, then the coefficient for the contribution term should be the causal effect. With inverse propensity weighting, the methodology should theoretically account for confounders by assigning weights to data points that should be considered more heavily.

Results

Having established that percentage of votes has a more linear correlation with contribution amount logged than the raw contribution amount, we regressed percentage votes on contributions

logged and confounding variables, and we estimated the treatment effect to be 4.58, implying that every 1% increase in contribution amount causes a 0.458% increase in percentage votes received.

OLS Regression Results

Dep. Variable:	Primary %	R-squared:	0.459
Model:	OLS	Adj. R-squared:	0.435
Method:	Least Squares	F-statistic:	19.90
Date:	Mon, 09 May 2022	Prob (F-statistic):	2.61e-55
Time:	02:33:36	Log-Likelihood:	-2390.3
No. Observations:	540	AIC:	4827.
Df Residuals:	517	BIC:	4925.
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-31.3524	5.877	-5.335	0.000	-42.897	-19.807
Veteran?	-2.7704	3.015	-0.919	0.359	-8.693	3.152
LGBTQ?	-7.0284	4.566	-1.539	0.124	-16.000	1.943
White?	-5.7785	2.174	-2.658	0.008	-10.049	-1.508
Elected Official?	9.0160	2.822	3.195	0.001	3.472	14.560
Self-Funder?	-9.1942	4.466	-2.059	0.040	-17.968	-0.421
STEM?	-1.3922	2.307	-0.603	0.546	-5.925	3.140
Obama Alum?	-2.6644	4.319	-0.617	0.538	-11.149	5.821
Party Support?	18.4988	4.652	3.977	0.000	9.360	27.638
Emily Endorsed?	5.2783	4.125	1.280	0.201	-2.825	13.382
Guns Sense Candidate?	-1.4120	2.102	-0.672	0.502	-5.542	2.718
Biden Endorsed?	11.9074	7.837	1.519	0.129	-3.488	27.303
Warren Endorsed?	-7.0800	11.662	-0.607	0.544	-29.992	15.832
Sanders Endorsed?	14.9379	9.069	1.647	0.100	-2.879	32.754
Our Revolution Endorsed?	0.4771	3.000	0.159	0.874	-5.417	6.372
Justice Dems Endorsed?	2.0621	3.632	0.568	0.570	-5.073	9.197
PCCC Endorsed?	5.4585	7.076	0.771	0.441	-8.442	19.359
Indivisible Endorsed?	8.3859	3.678	2.280	0.023	1.160	15.611
WFP Endorsed?	1.2847	4.918	0.261	0.794	-8.376	10.946
VoteVets Endorsed?	-1.8270	5.451	-0.335	0.738	-12.535	8.881
No Labels Support?	1.3959	14.900	0.094	0.925	-27.876	30.668
Partisan Lean	-0.6880	0.050	-13.888	0.000	-0.785	-0.591
Log_Contribution	4.9272	0.511	9.649	0.000	3.924	5.930

Omnibus:	78.070	Durbin-Watson:	1.363
Prob(Omnibus):	0.000	Jarque-Bera (JB):	110.252
Skew:	1.018	Prob(JB):	1.15e-24
Kurtosis:	3.871	Cond. No.	429.

This measure of causal effect of contributions has some uncertainty- as seen in the table above, this effect ranges from 3.92 to 5.93.

Next, we chose to binarize the contribution amount to be above or below $1 million. The regression table for this is below:

OLS Regression Results

Dep. Variable:	Primary %	R-squared:	0.442
Model:	OLS	Adj. R-squared:	0.418
Method:	Least Squares	F-statistic:	18.59
Date:	Mon, 09 May 2022	Prob (F-statistic):	4.86e-52
Time:	03:19:50	Log-Likelihood:	-2398.5
No. Observations:	540	AIC:	4843.
Df Residuals:	517	BIC:	4942.
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	13.9923	2.318	6.035	0.000	9.438	18.547
Veteran?	-1.7723	3.086	-0.574	0.566	-7.835	4.290
LGBTQ?	-5.2331	4.637	-1.129	0.260	-14.343	3.877
White?	-4.7919	2.199	-2.180	0.030	-9.111	-0.473
Elected Official?	10.6351	2.842	3.742	0.000	5.052	16.219
Self-Funder?	-9.5355	4.553	-2.095	0.037	-18.479	-0.592
STEM?	-1.5662	2.344	-0.668	0.504	-6.172	3.039
Obama Alum?	-0.2986	4.357	-0.069	0.945	-8.859	8.262
Party Support?	25.0863	4.625	5.424	0.000	16.000	34.172
Emily Endorsed?	7.5975	4.154	1.829	0.068	-0.564	15.759
Guns Sense Candidate?	-0.0310	2.118	-0.015	0.988	-4.191	4.129
Biden Endorsed?	15.4618	7.950	1.945	0.052	-0.156	31.080
Warren Endorsed?	-6.4552	11.842	-0.545	0.586	-29.719	16.809
Sanders Endorsed?	14.8207	9.215	1.608	0.108	-3.283	32.924
Our Revolution Endorsed?	1.0824	3.051	0.355	0.723	-4.912	7.077
Justice Dems Endorsed?	1.5082	3.698	0.408	0.684	-5.757	8.773
PCCC Endorsed?	7.8275	7.179	1.090	0.276	-6.276	21.931
Indivisible Endorsed?	9.4153	3.725	2.527	0.012	2.096	16.734
WFP Endorsed?	3.0858	4.981	0.620	0.536	-6.699	12.870
VoteVets Endorsed?	-1.2299	5.544	-0.222	0.825	-12.122	9.662
No Labels Support?	-1.1518	15.139	-0.076	0.939	-30.894	28.590
Partisan Lean	-0.6928	0.050	-13.767	0.000	-0.792	-0.594
Large Contribution?	18.5130	2.141	8.646	0.000	14.306	22.720

Omnibus:	61.942	Durbin-Watson:	1.405
Prob(Omnibus):	0.000	Jarque-Bera (JB):	80.585
Skew:	0.895	Prob(JB):	3.17e-18
Kurtosis:	3.613	Cond. No.	413.

Here, the regression suggests that having more than $1 million in contributions causes an increase of 18.5% of the primary percentage. This also has uncertainty- ranging from between 14.3% to 22.7%. However, the effect is statistically significant.

We also wanted to use another method to confirm the accuracy of this estimation. Using inverse propensity weighting (including removing data points with propensity scores above 0.9 or below 0.1), we estimated the average treatment effect to be 19.44. This means that, all else being equal, the effect of a candidate receiving $1 million or more in contributions is expected to cause an increase in 19.44% of the vote in their expected primary. This indication of causality is based on

the assumption that we have identified all of the confounding variables. While we did account for various candidate features, their endorsements by specific candidates and groups, and the partisan lean of the district they were running in, this may not be all of the confounders.

Discussion

Both methods that we used imply that there is a causal relationship between campaign contributions and primary results. However, we are uncertain about the validity of these results. As mentioned above, there are confounders that may not be accounted for. For example, how well-known the candidate is before the race could affect both primary results and contributions, and is not directly accounted for in our current factors. Another potential confounder is the state political environment; some races may occur in environments that are more anti-establishment, which could mean that political contributions from PACS or corporations may actually negatively affect primary performance. Confounders like these may mean that our hypothesis that there is a positive causal relationship between campaign contributions and primary performance may be incorrect.

Interestingly, from the regressions, it seems that some variables that we did identify as potential confounders did not end up being statistically significant. Of the large list of endorsements, only support from the Democratic Party and from the group Indivisible had an impact at the 5% significance level. In addition, of the list of features, only being white, an elected official, and a self-funder seemed to be significant. Altogether, this means that we could have selected potential confounding variables better; it also probably means that there are other candidate features that we did not include that may have a significant impact on primary performance.

While using regressions and inverse propensity models do give some measure of causality, they both have large limitations. With the regression model, if there are any large confounders that were missed, that would affect our measurement of causality to a very large extent. In addition, with this particular model, we did not implement cross terms, which means that the effects may be more granular than we discovered. With inverse propensity weighting, by necessity values with very low or very high propensity scores were dropped (otherwise some values would have a

nearly infinite weight), but this does mean that information is being lost. However, the fact that both regression and inverse propensity weighting resulted in fairly similar causality estimations (18.51 and 19.44, respectively) is good- it suggests that a relationship between contributions and performance does exist, or at the very least that both models have the same problem of, for instance, missing confounders. Until this is shown, however, we are fairly certain that a causal relationship between contributions and primary performance does exist.

We could more effectively prove causation through a more rigorous method, such as instrumental variable analysis. If we had a variable that both affected campaign contributions, and did not affect primary performance except through campaign contributions, then we would be able to perform this analysis. An imperfect example of this may be the unemployment rate in the district or state the candidate is running in- districts with higher employment would likely be more willing to contribute, and there unemployment would not substantially affect the other variables that we have chosen (although this is imperfect, because contributions can come from outside the district, and because higher unemployment could, for example, increase support for more radical candidates, in turn increasing the likelihood of a Sanders endorsement). If we had access to an instrumental variable, we could be more sure of our analysis.

Conclusions

With our multiple hypothesis testing, we were able to infer that there is a significant difference between candidates that were endorsed by most Presidential candidates and most other groups. It is difficult to generalize our results much further because even in our data there were groups like the NRA that did not have a low enough p-value to reject the null hypothesis, meaning that there was not a large enough difference between those endorsed and those that were not endorsed. Most of the groups analyzed in our multiple hypothesis test, however, did have a significant enough difference, where we can say that the support from these groups matter. As mentioned before, we are unable to conclude if the groups are endorsing the candidates because they believe that the candidate will already win the election (boosting the group or Presidential candidates' brand) or the endorsements they are giving to the candidate is actually helping them win. Most

likely it is a combination of the two ideas. In the future, we can use this data and conclusion to give insight to better predict candidates that may win their primary elections.

We established that monetary contributions to candidates have a substantial effect on those candidates' election outcomes. While our conclusions in respect to this research question are made only on the basis of primary election outcomes for the year 2018, and only used Democratic candidates, these results are still quite generalizable to politics as a whole. Certainly similar patterns would arise within the results of any of these candidate's elections in any of the years they ran for office. The bigger question is whether our conclusions are applicable to elections of any level of politics; it's likely that donations exert a different degree of influence on local elections than they do on statewide senatorial and representative elections.

Given these findings, however, it's clear that donations have not just a statistically significant benefit for candidates, but that donations are by far and large the most deterministic variable in a candidate's success. This is a problematic finding depending on if you believe that money should play a limited role in politics. Government's and constituents may want to strongly consider limiting the amount of donations allowed in a political campaign so as to level the playing field and limit the role that corporations have on policy making.

One limitation in the dataset was a lack of categorical variables describing the attributes for Republican candidates. With this present, we could have compared with the Democratic candidates' rate of primary election wins, even if the variables were not shared, we could have gotten some standpoint as to if individual attributes really did play a role. Undergoing a similar analysis for Republican candidates could be the topic of a future study.