Data 102 Final Project Report

Jiaji Wu, Jiayang Xing, Mackenzie Chen, Wai Yam Gordon Tsai

May 2022

Contents

1 Introduction

In the open Democratic primary election for the U.S. Senate, U.S. House, and governor, various factors influence the general public's choice of candidates. Intuitively, one aspect is whether a candidate's ballot proposition corresponds to a voter's political view or interests. Because of insufficient knowledge of politics or misleading by unproved news, the voter may not differentiate the candidate who genuinely supports their interests. As the voter realizes these potential issues, the voter may consider political endorsements by a well-known figure or organization while making a decision. Our first research question is whether a political endorsement impacts a candidate's result in the primary election in this study. Moreover, if a voter supports a candidate, the voter may financially support the candidate. Subsequently, a candidate's funding may reflect public support and can even predict the candidate's election result. It motivates us to form our second research question: whether the candidates who receive over-median financial support in primary election have a higher probability of winning the primary election.

2 Data Overview

For the first research question, we use one dataset named "Primary Candidates 2018" from FiveThirtyEight which is a polling aggregation website centered on politics, economics, and sports. The data about two parties are separated into two CSV files by FiveThirtyEight. We focus on the dataset that includes the information of democratic candidates, who are our study population. In this study, our population is all democratic candidates who ran the primary election. The dataset that we use gathers information about every individual in that population. Therefore, the data we use in the first research question is a census. The granularity of this dataset is the information of each candidate. Each column represents a candidate's self-identity, election information, and endorsements whether they received, which provides enough details for us to interpret our findings.

For the second research question, in addition to the "Primary Candidates 2018" data, we use one more dataset named the "2018 campaign finance data" from a government agency, Federal Election Commission. We access "2018 campaign finance data" from "all candidates file" during the period of 2017–2018 on the Federal Election Commission website. The "2018 campaign finance data" is a census since it contains summary financial information for each candidate who raised or spent money during 2017-2018, regardless of when they are up for election. The granularity of this dataset is each candidate who raised or spent money during the period. Each row contains the information about the candidate, total receipts, transfers received from authorized committees, total disbursements, transfers given to authorized committees, cash-on-hand totals, loans and debts, and other financial summary information.

We download these two datasets as CSV files directly from the source. According to the Federal Election Commission, the way of collecting data of the "2018 campaign finance data" was through the reports of the funding data from the candidate's committees, measurement error might appear in this dataset because of a tremendous amount of records at various states included thousands of senders and receivers. All candidates were aware of these two data collection since the election was a open primary. We would focus on the Democratic party in this study.

3 Research Question

3.1 First Research Question

Our first research question is that, in the democratic party, whether a candidate's endorsements affected whether he or she won his or her primary and has advanced to November.

To investigate in this question, we use the method of multiple hypothesis testing to test the impact of each endorsement. In this study, we focus on the impact of endorsement by Joe Biden, Elizabeth Warren, Bernie Sanders, Our Revolution, and Justice Democrats. As we develop several hypotheses that might explain the phenomenon we want to study, the method of multiple hypothesis allows us to evaluate the impact of each endorsement.

3.2 Second Research Question

Our second topic is to explore the causal impact of received financial support on candidates' primary election results. Our research question is that whether the total amount of money raised by each candidate exceeds the median of total amount of money raised by all candidates affects the probability of the candidate passing the primary.

Causal inference is a good fit to answer our question because there are more influencing factors between the amount of money collected and whether the candidate can pass the primary, and these influencing factors are confounders. To remove the effect of confounding variables and understand the relationship between the two, we can apply the method of unconfoundedness and visualize how strong the connection between the two is by estimating average treatment effect.

4 EDA

4.1 Data Cleaning

For our first research question, we use one of CSV files from the "Primary Candidates 2018" data which contains the information of Democratic candidates. We only use seven columns of the "Primary Candidates

2018" data for EDA and the model since the first research question is about the correlation between certain endorsements and the candidates' election result. We do not remove any null values since the null values helped display the relationships among the columns used or variables we define in our models.

For the second research question, since the "2018 campaign finance data" contains both Democratic and Republican candidates, we removed all rows related to Republican candidates. To make datasets better fit our goal. To examine the impact of a candidate's funding on his or her election result, we merge the "Primary Candidates 2018" data and the "2018 campaign finance data" to generate a new dataset named "combined dataset". Since these two datasets do not have the same primary column, we normalized candidates' name and chose the column of candidate names for merging. The "combined dataset" has 823 rows and 21 columns.

Since the "combined dataset" contains the summary of funding information for each candidate who raised or spent money during the period, regardless of when they are up for election, after merging, some candidates who did not raise or spend money during the primary election have null values in the total receipt column in the "combined dataset", so we replace these null values by 0. What's more, there are 157 candidates who do not have ethnic group information in the "combined dataset". According to the description in the source website, the value of the race column is “White” if they are identified the candidate as non-Hispanic white; the value is “Nonwhite” if they are identified the candidate as Hispanic and/or any nonwhite race; the values is null if they could not identify the candidate's race or ethnicity. To determine race and ethnicity, they checked each candidate's website to see if he or she identified as a certain race. If not, they spent no more than two minutes searching online news reports for references to the candidate's race. Based on the provided information, we decide to drop the rows with no race values because obviously we could not randomly assign null value to either White or Nonwhite, and searching for the missing information on internet is hard. In addition, we create dummy variables for each columns and fill NaN values in endorse? columns by 0. With all above steps done, we end up having the "combined dataset" with 665 rows and 21 columns.

4.2 Democratic Candidates and Endorsements

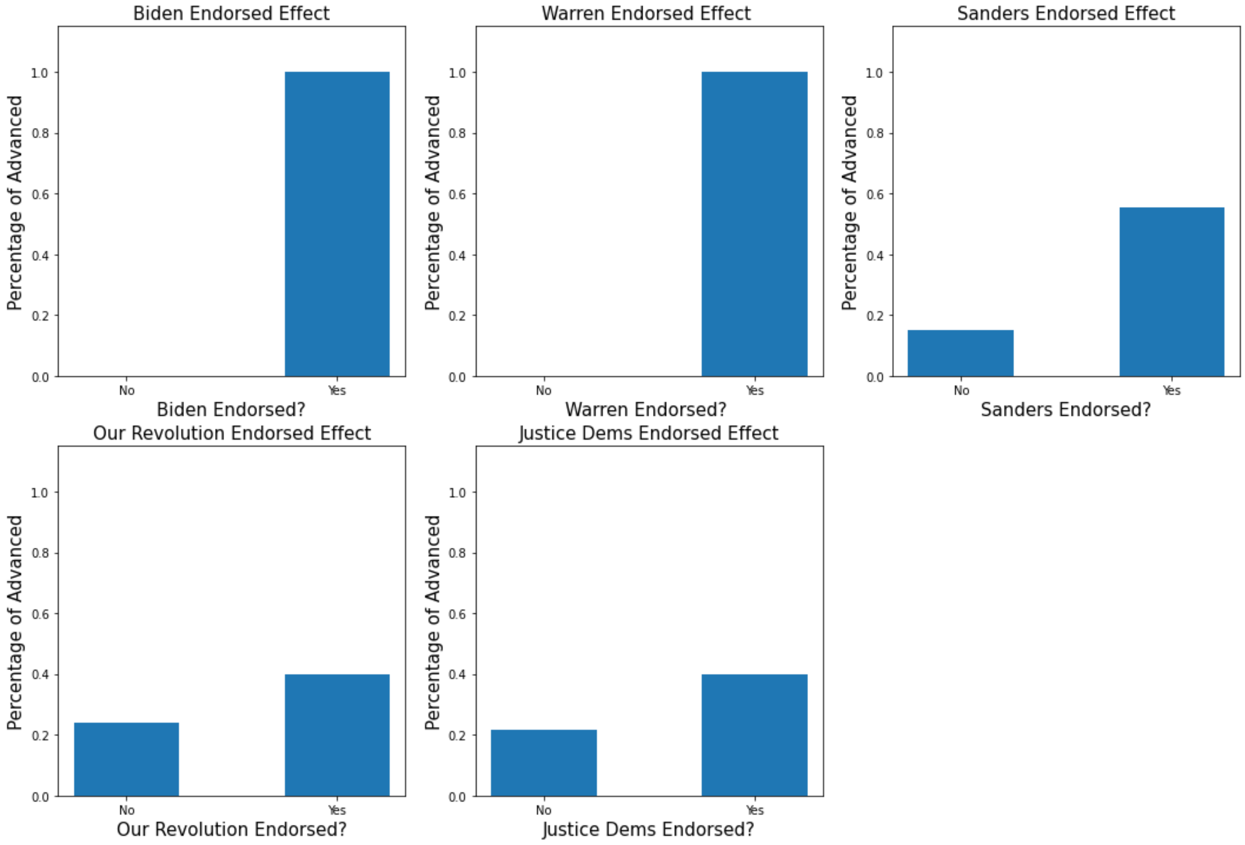

Figure 1: Distribution of Election Status for the Candidates Who Received Certain Endorsements

To make bar charts, we modify the value of the column "Primary Status", which is a categorical variable and has two distinct values, 'Advanced' and 'Lost'. We replace 'Advanced' by 1 and 'Lost' by 1.

From the above graphs, we could conclude that Biden's and Warren's endorsement made a significant impact on the election result of candidates. The candidates who got their endorsements all passed the primary election, and the candidates who they ran against or anti-endorsed all did not pass the primary election. Compared to these two endorsements, Sanders's endorsement makes a less significant impact on the election result but better than Our Revolution and Justice Dems, which we could observe from the rest three graphs. Subsequently, we could make initial guesses for our multiple hypothesis testings based on the above observations: Biden's, Warren's, and Sanders's endorsements would affect the result of candidates, and Our Revolution and Justice Dems's endorsements would not affect.

4.3 The Dependency Among Endorsements

To visualize relationships among endorsement variables, we modify the values of these endorsement columns, which have three unique values: none, yes, and no. We set all none values as -1, 'yes' as 1, and 'no' as 0. We set none value as -1 instead of 0 is because none and 'no' have different meanings (check

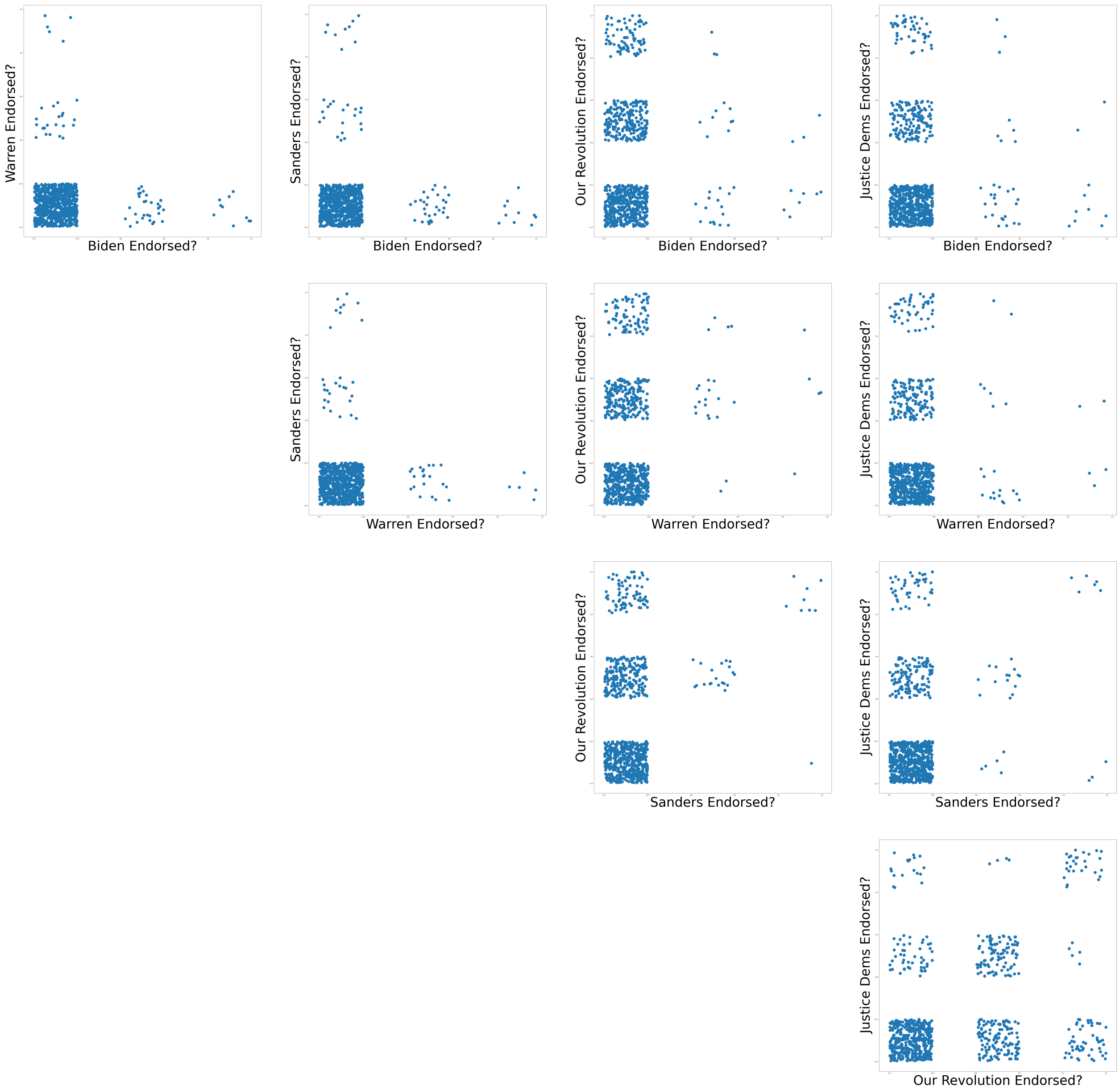

Figure 2: One-to-One Dependency Among Endorsements

the column description below). Thus, to clearly display the relationships, we used two numerical values to represent 'none' and 'no'.

Here is the description of one endorsement column - "Biden Endorsed?" from source data:

1. 'Yes' if the candidate was endorsed by Joe Biden before the primary.
2. 'No' if the candidate is running against a Biden-endorsed candidate or Biden specifically anti-endorsed the candidate.
3. If Biden simply did not weigh in on the race, we left the cell blank.

(The values of each endorsement column have same meaning as one in the above description.)

Moreover, since there are nine possible coordinates of two endorsement variables: (-1, -1), (-1, 0), (-1, 1), (0, -1), (0, 0), (0, 1), (1, -1), (1, 0), (1, 1), to avoid a graph only has nine dots from which we can observe little things, we added random noise ranged from 0.01 to 0.5 to each value. Then, we used scatter plots to visualize the relationships between two endorsement variables on each graph, and 10 graphs were generated. On each graph, we could see the region is divided into nine parts which represent nine possible coordinates. If a graph is patternless, we can assume variables on this graph are independent. Otherwise, we cannot. For example, on the graph(3, 3) - "Our Revolution Endorsed?" vs "Justice Dems Endorses?", there are some dots on all nine parts. We may assume "Our Revolution Endorsed?" is independent from "Justice Dems Endorses?". Another example is the graph of "Our Revolution Endorsed?" vs "Sanders Dems Endorses?". On this graph, there are not some dots on every parts; instead, 5 parts out of 9 parts have dots; we treated the only dot on one part as an outlier at this time. In particular, there are dots on (0, 0) and (1, 1) but no dots on (0, 1) and (1, 0), by which we thus deduced there exists some dependency between "Our Revolution Endorsed?" and "Sanders Dems Endorses?". And, there are more patternless graphs, such as the graph of "Biden Endorsed?" vs "Warren Endorsed?", the graph of "Biden Endorsed?" vs "Sanders Endorsed?", and graph of "Warren Endorsed?" vs "Sanders Endorsed?".

Through the analysis above, we could not assume there exists no dependency among these endorsement variables, which would affect our choice of the method to control error rate. We cannot use Benjamini-Hochberg since it assumes independence among variables. Instead, we would use the Bonferroni Correction and the Benjamini-Yekutieli procedure which do not require independence to decrease false discovery rate.

4.4 Democratic Candidates and Financial Data

After finishing the data cleaning process, we computed the median total receipts for democratic candidates who have appeared on the ballot this year in Democratic primaries, which is $26,460.01, and the maximum total receipts is $18,468,367.89.

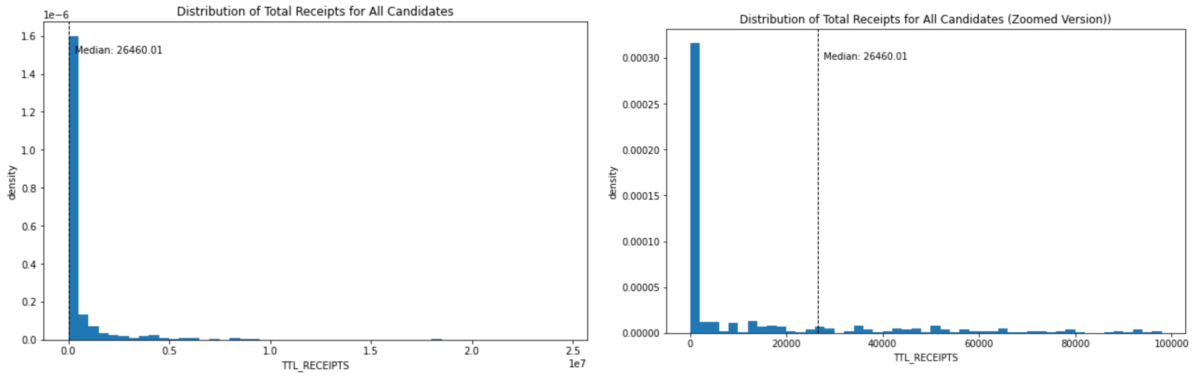

Figure 3: Distribution of Total Receipts for All Candidates

From figure 3, the distribution of total receipts for all candidates, we can observe that the histogram is right skewed, which means a lot of candidates who have appeared on the ballot in 2018 in Democratic primaries for Senate, House and governor did not raise money or only raised a small amount of money for their election.

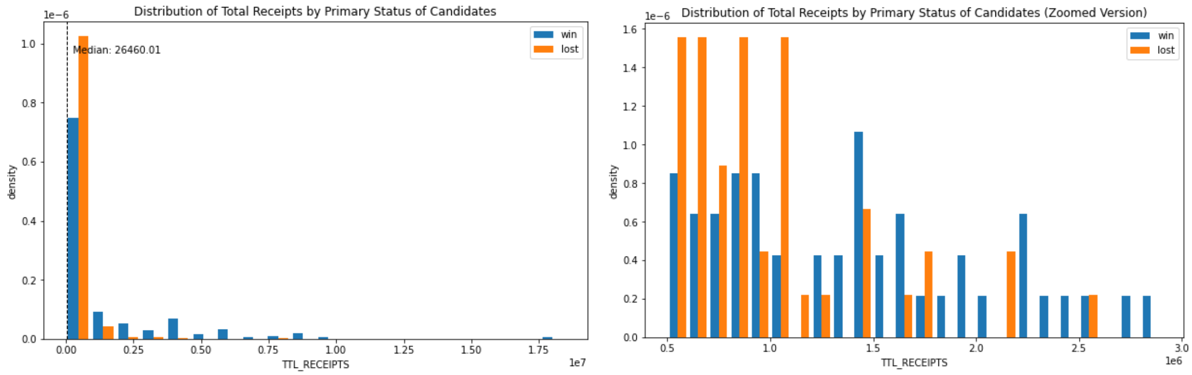

Figure 4: Distribution of Total Receipts by Primary Status of Candidates

When we look at feature 4, the distribution of total receipts by the primary status of candidates, we observe the number of lost candidates starts being greater than the number of advanced candidates when the total receipts amount exceeds $1,200,000.

In addition, feature 5, the boxplot, also shows that the advanced candidates has a wider range of total receipts amount for the lost candidates.

Primary Status	Median Total Receipts	Percentage of Over-Median (($26,460.01))
Lost	8688.830	40.82
Won	171501.225	68.30

Table 1: Median Total Receipts and the Percentage of Candidates Having Over-Median Receipts ($26,460.01) for Advanced Candidates and Lost Candidates

Moreover, table 1 shows the difference of median total receipts between advanced candidates and lost candidates. Apparently, the median receipts of advanced candidates is larger than the median receipts of lost candidates. It motivates us to investigate more the relationship between total receipts of a democratic candidate and his/her primary status. In addition, table 1 shows the percentage of Advanced and Lost candidates who had over-median receipts. We can observe that higher proportion of Advanced candidates had over-median receipts than Lost candidates. It suggests that whether a candidate has over-median receipts might affect his/her primary status.

Race (Is White)	Number of Candidate Received Over-Median Receipts	Percentage of Candidate Received Over-Median Receipts	Number of Candidate Won Primary	Percentage of Candidate Won Primary
Non-White	106	51.21	66	31.88
White	227	49.56	158	34.50

Table 2: Number and Percentage of Candidate Received Over-Median Receipts and Number and Percentage of Candidate Won Primary by Whether a Candidate is Identified as White

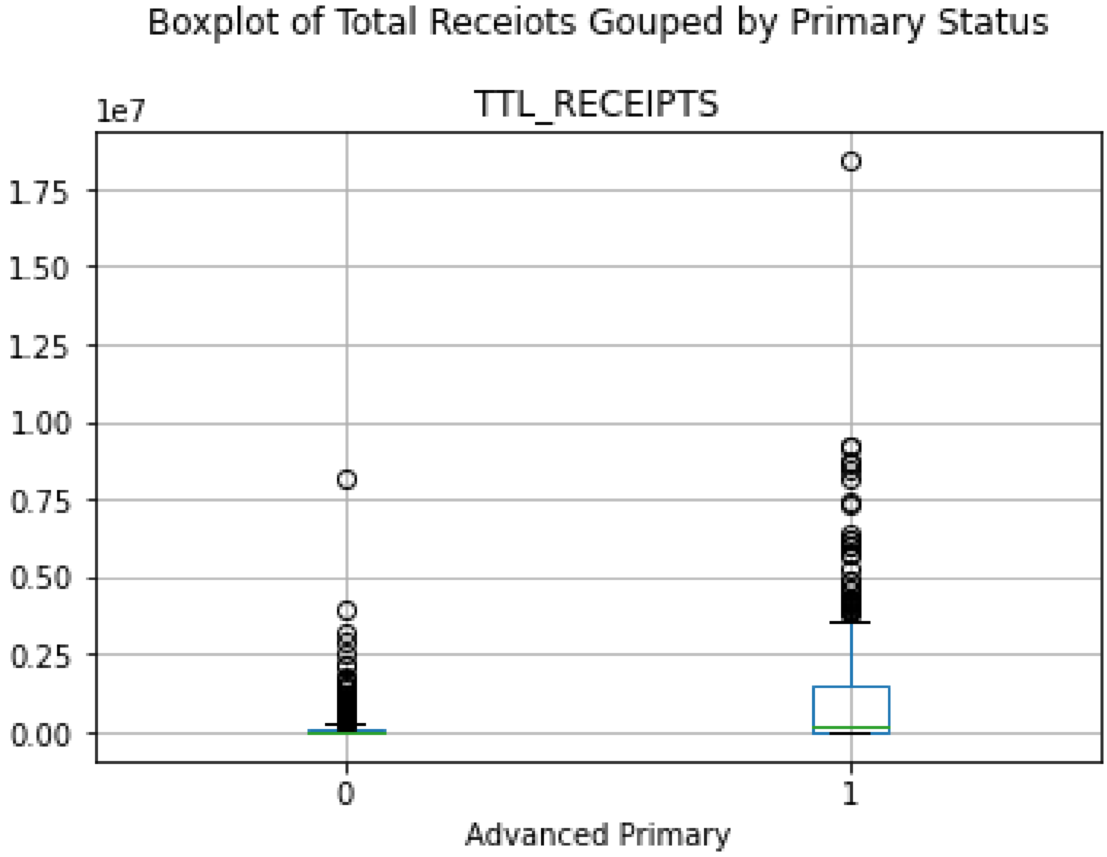

Figure 5: Boxplot of Total Receipts Grouped by Primary Status

Base on table 2 we can see that there are more white had receipts over median, and more white advanced from Primary. However, there is a higher proportion of nonwhite candidates had over median receipt. In addition, higher proportion of white candidates pass Primary. This finding suggests that race as a confounding variable will affects both our treatment, which is whether a candidate receives more than the median amount of money, and our outcome, which is whether a candidate advanced primary. Therefore, in our research question modeling, we will do unconfoundedness by using inverse propensity weighting.

5 Multiple Hypothesis Testing

5.1 Method

For the hypotheses, we use the endorsement columns which represent the most famous politicians and the top organization that has largest percentage of endorsement in the dataset, which are Biden, Warren, Sander, Our Revolution, and Justice endorsement. If we find someone is with low percentage of endorsement and has low reputation, it might be a drawback on our findings, which is hard to reject the null hypothesis since his endorsement was not helpful at all on the candidate's election result. Choosing the most well-known figures, we can set up an more accurate policy and advance for each candidate, which is also a reason why we chose multiple hypotheses.

For the testing of each hypothesis, we use Simulating the Statistic Under the Model since its assumption is not strict and its result is easy to interpret. We set the same null hypothesis for five hypotheses testings: current endorsement does not affect a candidate's election result, and the alternative hypothesis is current endorsement is helpful for a candidate's election result. In every hypothesis testing, we use a binomial model with p = 0.33 , which is the overall win rate in the ”Primary Candidates 2018” data, namely, the proportion of Democratic candidates who won in the election, as well as n = the number of candidates who received current endorsement in the ”Primary Candidates 2018” data. We would stimulate the binomial model 10000 times. Finally, we decide to reject or not based on the expect value which is the number of candidates who received current endorsement and won. For example, in the Biden's case, the number of candidates who had Biden's endorsement is 10, so we stimulated a binomial model (n=10, p=0.33) 10000 times and got an array of 10000 stimulated results. Then, we calculate the proportion of stimulated results are greater than 10. The proportion in our actual testing is 0.0, and since we use a alpha 0.05, we would reject the null hypothesis.

In the EDA section, we have showed that there exist some dependency among endorsement variables. Thus, we should select the error-checking methods that do not require all variables are independent, which are the Bonferroni Correction and the Benjamini-Yekutieli procedure. The Bonferroni Correction can control family-wise error rate (FWER), and the Benjamini-Yekutieli procedure can control false discovery rate (FDR).

5.2 Result

In our actual hypothesis testings, we get the following p-values: Biden: 0.0, Warren: 0.0044, Sanders: 0.1415, Our Revolution: 0.0935, and Justice Dems: 0.171. If we use a alpha 0.05, we would reject the null hypotheses of Biden and Warren, which means their endorsements statistically significantly help a candidate

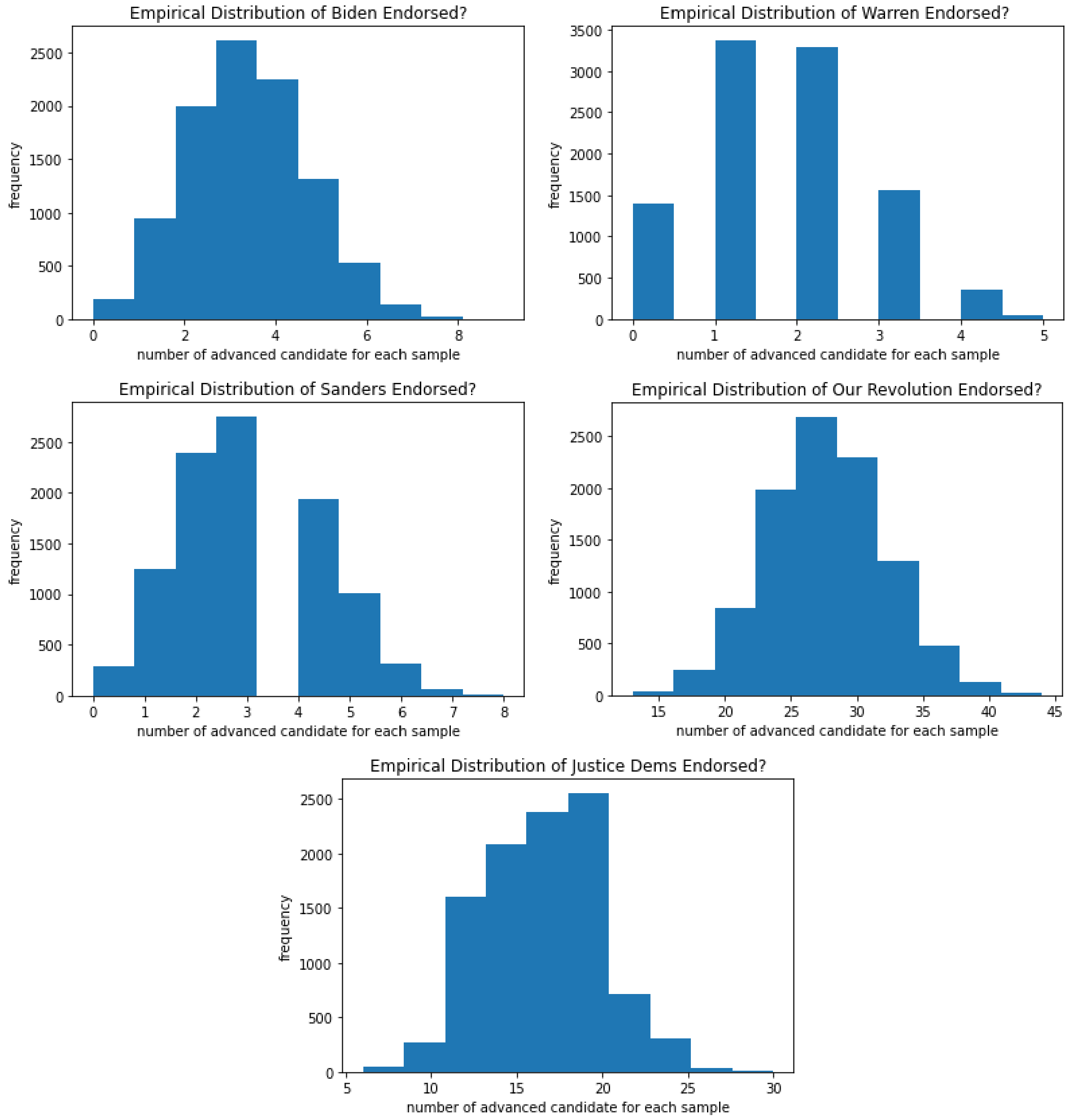

Figure 6: Distribution of Stimulated Model for Selected Endorsements

Bonferroni	Benjamini Yekutieli procedure
True	True
True	True
False	False
False	False
False	False

Table 3: Error Check for Endorsements' Hypothesis Testing

win in the election. And we would fail to reject the null hypotheses of Sanders, Our Revolution, and Justice Dems, which means their endorsement do not affect a candidate's election status.

We first use the Bonferroni correction which controls FWER and ensures not many false positives to check our hypothesis testings' result. We then use the Benjamini–Yekutieli procedure which controls FDR and can strike a balance between controlling false positive and true positive. The results of both error-checking are shown on table 3, which indicates that our decisions on rejecting null hypothesis or not for every hypothesis test are accurate.

5.3 Discussion

As mentioned in the last section, if setting a alpha as 0.05, we reject the null hypotheses of Biden and Warren and fail to reject those of Sanders, Our Revolution, and Justice Dems. After applying two correction procedures, the discoveries of Biden and Warren remained statistically significant to reject. All decisions can be made from the individual tests since the p-values significantly differ from the alpha 0.05. The only condition that we cannot use individual tests is that p-value is closed to the alpha, such as p-value is 0.044 and the alpha is 0.05.

There is a limitation in our model and analysis, which happens due to different meanings of ”no” and ”NaN” value in the endorsement columns (check the column description above in the EDA section). For example, for the Biden's endorsement column, ”no” means the candidate is running against a Biden-endorsed candidate or Biden specifically anti-endorsed the candidate, and ”NaN” value means Biden did not weigh in on the race. Yet since our hypothesis testing requires binary values, we replaced both ”no” and ”NaN” as 0. Subsequently, our hypothesis testing would examine the effect of endorsement of 1 on a candidate’s election status but not the effect of every endorsement value on the election status. The effect of every endorsement value may weigh differently on the election result. For instance, Biden not weighing in on the race may simply means Biden does not know this candidate. Intuitively, its weighs of effect may be lighter than that of ”no” which means this candidate competes against the candidate who Biden supports. Since votes may cross over parties, Democratic candidates who run against each other naturally can have slightly different or even opposite political views or ballot propositions. In non-swing states, this may determine one candidate can win another candidate. However, since our hypothesis test requires binary variable, changing both ”no” and ”NaN” as 0 best fits our analysis.

P-hacking is avoidable since we used the original data during our analysis. And, the same result generate by both corrections can confirm the accuracy and address the p-hacking.

If we gather more data, we may discover more interesting aspects to explore. For example, if additional endorsement information about why certain politicians endorse a candidate are available, we may produce a more comprehensive policy to increase a candidate’s win chance in primary election.

6 Causal Inference

6.1 Method

Since we cannot have a clear judgment on the amount of money received by the candidate without a criterion, we decided to use the median of total receipt form all candidates as our criterion. In other words, our treatment is receiving the amount of funds that is higher than the median financing amount, and our outcome is the candidate’ primary status (whether a candidate won the primary election). The confounders that we could observe from the dataset are ”Is White’, ’Veteran?’, ’LGBTQ?’, ’Elected Official?’, ’Self-Funder?’, ’STEM?’, ’Emily Endorsed?’, ’Guns Sense Candidate?’, ’Biden Endorsed?’, ’Warren Endorsed? ’, ’Sanders Endorsed?’, ’Our Revolution Endorsed?’, ’Justice Dems Endorsed?’, ’PCCC Endorsed?’, ’Indivisible Endorsed?’, ’WFP Endorsed?’, ’VoteVets Endorsed?’, ’No Labels Support?’t. Since there are many confounding factors, we will use the inverse propensity weights which first calculate the propensity scores and then get the average treatment effect (ATE).

6.1.1 Improvement of Our Method

Since in multiple hypothesis, we fail to reject that the endorsement of Our Revolution and Justice Dems would not affect a candidate’s primary status, we decide to remove Our Revolution Endorsed? Justice Dems Endorsed? in our confounding variables. Additionally, to address the uncertainty in our ATE estimation, we bootstrap 2,000 times and construct an approximate 95% confidence interval for the ATE.

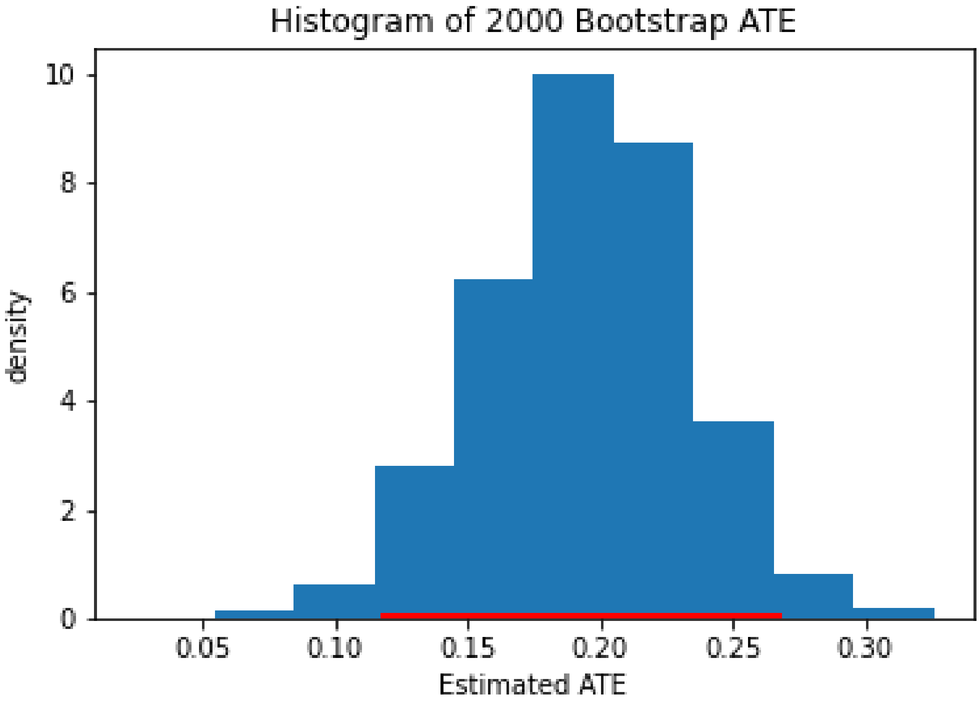

Figure 7: Distribution of 2000 Bootstrap ATE

6.2 Result

After applying the method of inverse propensity weights, our original calculation gives us an average treatment effect of about 0.1712, and after improving our model by deleting two confounding variables, we get the ATE equal to 0.1919. Moreover, based on the bootstrap confidence interval, the approximate 95% CI for ATE is [0.1175, 0.2686], which does not include 0. Therefore, we are 95% confident to reject to null hypothesis that there is no causal effect of the treatment on the outcome. There is a positive causal link between whether the total amount of funds collected by a candidate is over median and whether the candidate passed the primary election in democratic party. Receiving over-median funds would cause a democratic candidate a higher probability to win the primary election.

6.3 Discussion

First of all, from the above coding and analysis we did above, we do have confidence in our conclusion, and there are two reasons. First, the result is highly statistically significant, since the we assume there are no unmeasured confounders; no interference and consistency assumption are also needed, and based on simulation under a model, we get a p value which is smaller than 0.05 so we can reject the null hypothesis that there is no effect for total receipt over the median. Secondly, it is also economically and socially significant because a candidate receiving more money implies that the candidate is more popular or the candidate and their community are richer. In the former case, it means that the candidate has a large number of supporters, which makes the candidate more likely to pass the primary; in the latter case, the candidate's community is more likely to increase promotion efforts to make the candidate more likely to pass the primary due to entangled interests.

In terms of the limitations of the IPW method, it is difficult to assure each confounding variables that we set in this case indeed affect both treatment and outcome. To determine each confounding variable, we need to do the hypothesis test on their effect on both treatment and outcome. Therefore, we decide confounding variables intuitively. In other words, it is possible that some variables are not confounding variables. If we include variables that are very good in predicting the treatment but not good in predicting the outcome, this will actually increase the variance of the propensity score estimator. Moreover, although we exclude many confounding variables, the factors we exclude are only for all the data given in the dataset, and there are still many potential influences between our treatment and outcome, such as the candidate's social popularity, the candidate's achievements as a government official, etc. In addition, when we were handling Nan values in columns representing endorsement, we set Nan values as 0, which set them in the same category as those who have "No" in endorsement columns. If we set Nan values to a different value, it might give us a slightly different result of ATE. Furthermore, when we apply the method of IPW to achieve unconfoundedness, one assumption needed to use the Potential Outcomes Framework is consistency. In our research, we defined the treatment to be whether a candidate receive over-median funds, in which case treated units could receive different version of the treatment. For example, for candidates who receive treatment, their received funds could be either $1 over median or $1,000,000 over median. This setting might violate the assumption of the Potential Outcomes Framework.

In addition, there are some additional data that will be very helpful in answering this causal question. For example, people's social media following of candidates is a data that we can increase, and by collecting the number of candidates on Instagram and Facebook follower, we can continue to reduce the bias of our data. Candidate's social media influence is a confounding variable that can largely affect the ate results. If a candidate has a lot of followers on social media, it means that their campaign messages on social media will get a lot of attention and more people will donate money to the candidate. At the same time, more social media followers will also let voters know more about the candidate, thus increasing the probability of the candidate passing the primary election. What's more, we can also add the candidate's personal income and total household income as another data, because the candidate's economic level is also a confounding variable that we can consider. A higher economic level means that the candidate's social circle is also more wealthy, so the candidate is also more likely to get higher financial support. At the same time a higher economic level represents a candidate with more attention and thus more likely to get votes, which can be seen in the wealth of Trump and George Bush.

7 Conclusion

For the first research question summarise, From the above observation, if we assume the alpha is 0.05, after both Benjamini-Yekutieli procedure and Bonferroni test, we can observe that Biden's Endorsement and Warren's Endorsement have a significant effect, and we do not have enough evidence to claim that the Endorsement of Our Revolution, Sander and Justice Democrat is helpful for candidate primary election. To summarize the findings of the second question, a candidate's total receipt over the median of total amount receipt from all candidates will have a positive causal effect on whether the candidate passes the primary or not.

In our study, we focus on the primary election of democratic candidate, so our results can only be applied to the specific party and election phase, which is narrow. To generalize our finding, we may also need to investigate Republican candidates. For multiple hypothesis testing, the result is about specific endorsement by politicians and organizations. Our finding is narrow that precisely focus on the treatment of endorsement and can not be extended to other situations. The design of causal inference study also causes that our results may not be generalizable. As we set the threshold of treatment to be the median receipts, the result of causal inference is not broadly applicable.

Based on the result, we can have a call to action that tries to help that candidate by providing a solid connection to Biden and Warren that our candidate show try to participate in more political activities that Biden and Warren also in there to have a strong connection with them. It would be helpful to pass the primary election if we could have their Endorsement or support. Then we have lower priority to have an excellent connection to other organization or politician since they are not that helpful to the election. Based on the result of causal inference, we can suggest candidates raise as much money as possible to increase probability to win the primary election.

In our research, we combine the two csv's together, the dem-candidates dataset from the polling aggregation website and the finance dataset from the Federal Election Commission website. Both website are authoritative websites, but they have a different focus, so we have reason to believe that combining the two datasets by candidate name will help us to better understand and analyze the question we exported in a comprehensive way.

One limitation of our data is that from our data, we cannot make our ”No” group in our data separate to ”NaN” and ”No” Group. Since binary decision cannot have more than two group since we already has ”yes” group. As a result, there is some limitations in that we can only focus on the Endorsement whether help the primary election instead of having different decisions like ”Democrat important people endorsement or hate will affect the election, etc. However, since the Hypothesis test requires a binary decision. As a result, we can only go deep into the cases about having an endorsement vs. not having an endorsement. In terms of the limitation of financial data specifically, the candidates’ total receipts we mentioned above include the funds raised during the general election time period as it includes all the receipts a candidate receive in 2017-2018. In addition, as we can see in the figure 3, where are some individual candidates’ total receipts are too high. However, we cannot decide whether they are outliers because we do not have a good domain knowledge in political background. Moreover, because candidates may have multiple committees authorized to raise and spend money on their behalf, we have double-counted financial activity in the total receipts.

Future studies trying to find the group’s Endorsement are ”No” and make the same hypothesis that if Democrat important people hate the candidate will they affect the result of a primary election. After that, we would have a bigger picture that the endorsement perspective of Democrat essential people and the Democrat critical perspective hate candidate result of binary decision so that we know how those Democrat important people affect the election result differently. In addition, the false negative rate is much more important than FDP and FDR in this case: the consequence of false positive is that candidates waste their time and resource to build connections with people and organizations that are not useful whereas the consequence of false negative is that candidates miss useful endorsements. Therefore, in future studies we want to study the false negative or finding some algorithm that control the FNR in our future work. Moreover, future studies could look for a more detailed data with fewer missing values and more confounders. Furthermore, future studies could investigate the candidates from other parties.