

# **Final Project Written Report**

## **Group Members:**

Brandon Suen,  
Fanyu Cao,  
Shio Huang,  
Wei-Teng, Chang

## **Instructor:**

Eaman Jahani, Ramesh Sridharan

## **Course:**

Data 102 Data, Inference, and Decisions

## **Date:**

May 8th 2023

# Content

<b>1. Data Overview.....</b>	<b>3</b>
<b>2. Research Question 1.....</b>	<b>3</b>
2.1 Introduction.....	4
2.1.1 Purpose.....	4
2.1.2 Model of Choice.....	4
2.1.3 Datasets.....	4
2.2 EDA for Question 1.....	5
2.2.1 Pie Chart and Histogram of Basic Variables.....	5
2.2.2 Heatmap of Basic Variables.....	8
2.3 Method for Question 1.....	9
2.3.1 Logic Construction.....	9
2.3.1.1 DAG.....	9
2.3.1.2 Potential Outcome Framework.....	9
2.3.2 Logistics Regression.....	10
2.3.3 Matching.....	10
2.4 Results.....	11
2.4.1 Results for Logistics Regression.....	11
2.4.2 Results for Matching.....	12
2.5 Discussion.....	13
2.6 Conclusion.....	14
<b>3. Research Question 2.....</b>	<b>16</b>
3.1 Introduction.....	16
3.1.1 Purpose.....	16
3.1.2 Model of Choice.....	16
3.1.3 Datasets.....	16
3.2 EDA for Question 2.....	16
3.2.1 Count of Candidates by Parties.....	16
3.2.2 Count of Candidates by States.....	17
3.2.3 Scatterplot for Total Disbursement and Total Contribution.....	18
3.2.4 Role in the Election.....	18
3.3 Method & Discussion for Question 2.....	19
3.4 Results.....	20
3.4.1 Bayesian GLM Results.....	20
3.4.2 Frequentist GLM Results.....	20
3.4.3 GLM Result Analysis.....	21
3.4.4 Nonparametric Result.....	21
3.4 Conclusion.....	23
<b>Citation.....</b>	<b>24</b>

# 1. Data Overview

## 1.1 Primary Candidates 2018 Dataset

There's both Democrats and Republicans data in the dataset. However, we think the Democrats data has more features we can discuss so we only focus on the `dem_candidates.csv` here. On the other hand, The Democratic Party is closely related to minority issues with freedom and tolerance. We want to see if they really implement their spirit or if it is just a campaign gimmick. `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. It is a census since it contains all the candidates who joined the election for the Democrats. Our data does not have any groups that were systematically excluded from.

## 1.2 Candidate Dataset and Electoral Votes Dataset

The datasets used for this research question are *Candidates* of the year 2018, which was recorded by the Federal Election Commission and published annually. This dataset contains several major benchmarks of a candidate's financial status during election, such as total contributions and total disbursements. This dataset is downloaded from the official website by FEC. We browsed to the page labeled "candidates" and selected the year 2018. In the later construction of the GLM and nonparametric models, we used an additional dataset called Electoral Votes, which contains the votes for each state. We choose to include this dataset in order to take the number of electoral votes into consideration as a predicting feature.

In the *Candidate18* dataset, we notice the abnormality that there are many zeros. After researching, we concluded that this phenomenon is either a result from unreported data or a lack of contribution. To preserve the legitimacy of the dataset, we decided to include all the zeros. Thus, the distribution of the *Candidate18* dataset would be left-tailed for most features. Also, since the dataset only concerns the year 2018, we expect our result to be under representative since the election situation changes due to many variables such as the yearly condition. In conclusion, this dataset may have limitations in generalizability to the population and only represent a sample under specific conditions.

## **2. Research Question 1**

### **2.1 Introduction**

#### **2.1.1 Purpose**

A candidate's background, such as race, veteran status, and LGBTQ identity, can influence the outcome of Democratic primary elections. Voters tend to support candidates who share their identity group, a phenomenon known as "identity politics." For example, candidates of color may face historical barriers due to institutional racism and voter bias, while military service can be viewed as a positive attribute due to appreciation and respect from the people. LGBTQ candidates may mobilize a supportive base but might face resistance from conservative voters. According to the above assumptions, we decide to quantify these effects to see if it is just as we thought. To accomplish this, using logistic regression provides us with estimated coefficients for the associations among the different backgrounds and the result of the primary elections.

#### **2.1.2 Model of Choice**

The method we chose is causal inference. We chose this method because causal inference is a good choice for the impact of a candidate's background on primary election outcomes because it allows us to identify cause-and-effect relationships. Causal inference can help isolate the specific impact of factors like race, veteran status, and LGBTQ identity on voter behavior. This can provide more accurate and reliable insights into the effects of these factors on primary election outcomes and can help evaluate the fairness of the election since equity is one of those ideas we emphasized a lot. Some limitations we found are there are still some factors we can not really isolate from. For example, election strategies will be a complicated factor that will affect the result. However, we think it will be overwhelming to go through every detail of the process. Hence, we will only focus on the backgrounds of candidates that we can collect at the beginning of the election and ignore the other factors that would be found during the election.

– What are the limitations of the method you chose? Under what circumstances might it not do well, and what could go wrong under those circumstances?

#### **2.1.3 Datasets**

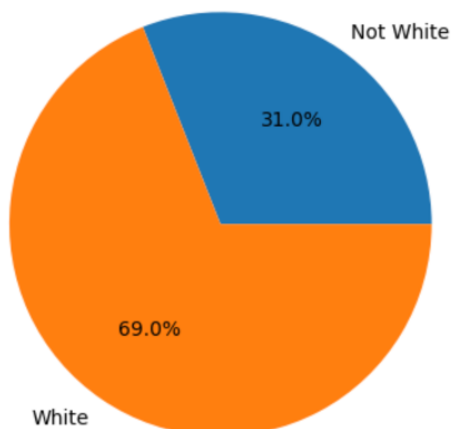
The dataset we used is from the "FiveThirtyEight: 2018 Primary Candidate Endorsements". The "dem\_candidates.csv" provides us the information about the 811 candidates who have appeared on the ballot this year in Democratic primaries for Senate, House and governor in 2018.

## 2.2 EDA for Question 1

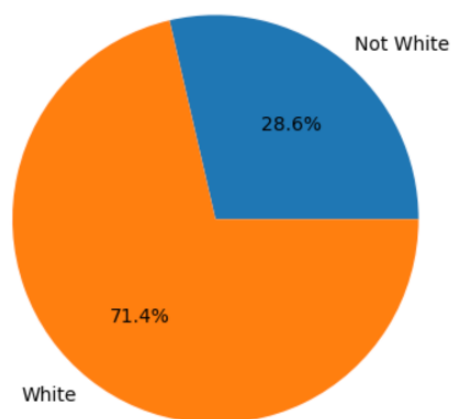
### 2.2.1 Pie Chart and Histogram of Basic Variables

In our study, we examined the factors of race, veteran status, and LGBTQ identity in Democratic primary candidates. These demographic factors were converted into binary variables for analysis. Pie charts were used to visualize the proportions of each group, showing that the majority of candidates are white, non-veteran, and non-LGBTQ. However, these proportions did not significantly shift among primary-victory candidates, hinting that these backgrounds may not provide a distinct advantage. Despite these initial insights, we must further investigate to control for potential biases and confounding variables.

Candidates' Race

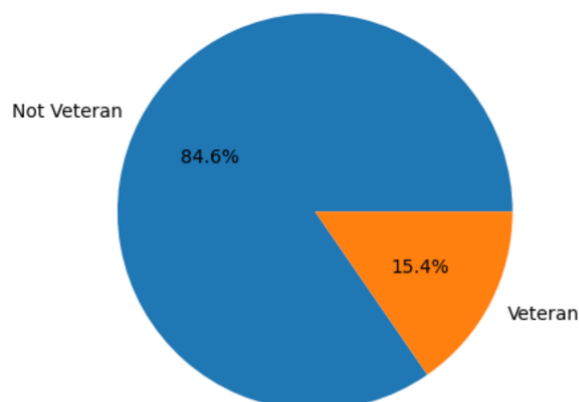


Won Candidates' Race



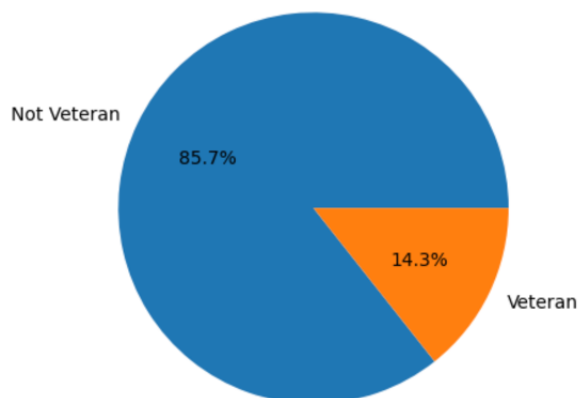
*Distribution of Candidates by Race*

Is Candidate Veteran?



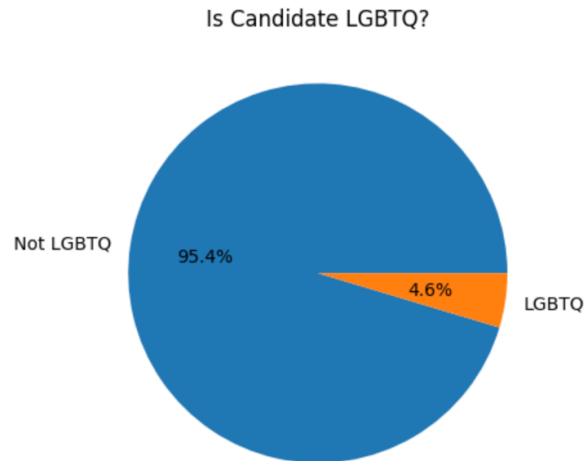
*Distribution of Won Candidates by Race*

Is Won Candidate Veteran?

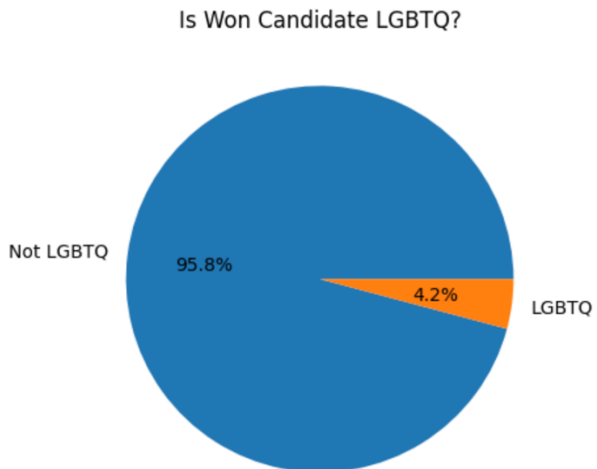


*Distribution of Candidates by Veteran Status*

*Distribution of Won Candidates by Veteran Status*

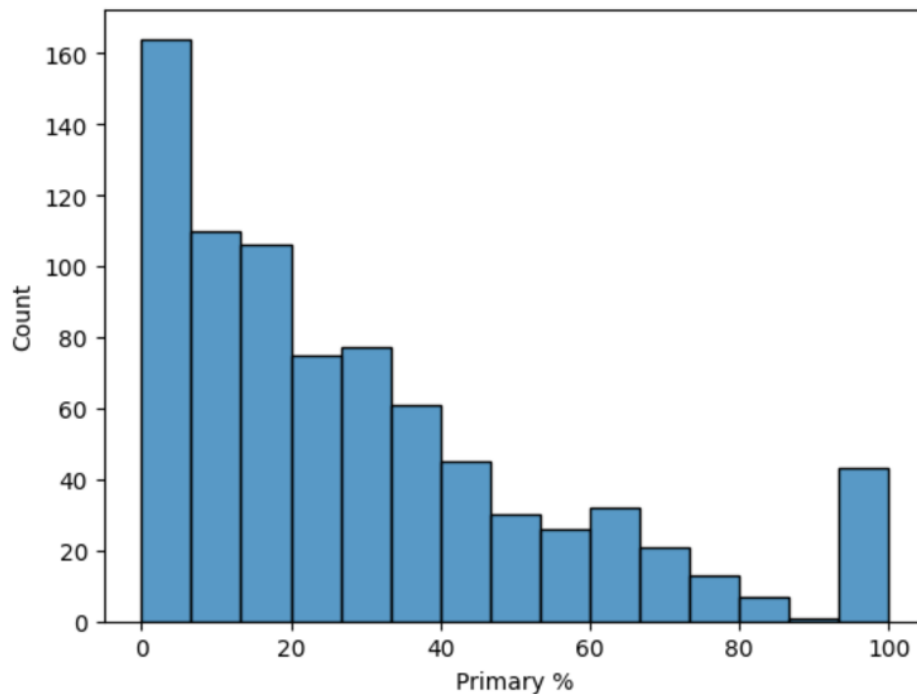


*Distribution of Candidates by LGBTQ Identification*



*Distribution of Won Candidates by LGBTQ*

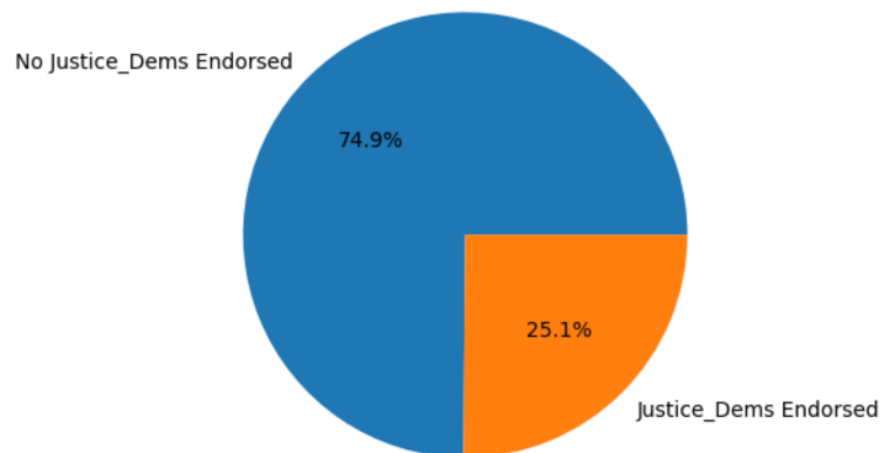
We analyzed the "Primary %" feature, which represents the distribution of primary votes. The data showed a high proportion of votes in the 0-20% and 100% ranges, but a small proportion in the 50-90% range. This suggests that the primary vote rate, influenced by various factors including candidate background, could be a potential confounder in our causal inference as it may also impact the final primary election results, given that some states require primary elections before proceeding to general ones.



*The distribution of Primary Votes Among the Candidates*

We examined the endorsement distribution from Justice Democrats and found that about 25% of the candidates received their support. Given their political advocacy and candidate selection considerations, this endorsement is related to race, veteran status, and LGBTQ identity. Furthermore, it can influence primary election outcomes through increased publicity and sponsorship. Hence, it is considered as a potential confounder in our causal inference analysis.

#### Is the Candidate Endorsed by Justice Dems?

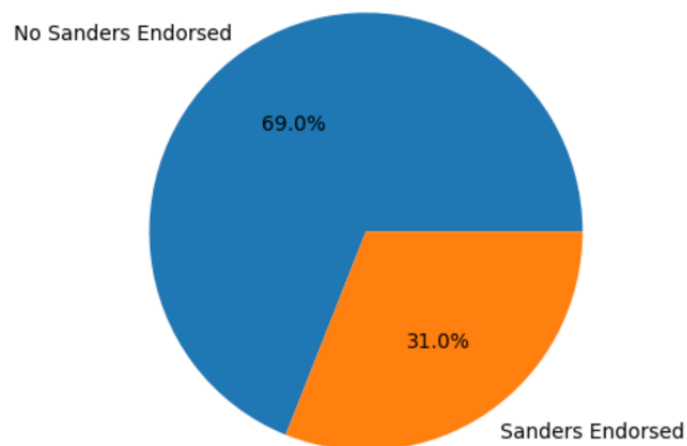


#### *Distribution of Candidates by Endorsement of Justice Democrats*

We're also looking at whether candidates have the endorsement of Bernie Sanders ahead of the primary. We converted it into a **quantitative variable** for visualization. We found that a third of the candidates had his backing. Bernie Sanders' political views pay more attention to some disadvantaged and minority groups, so this is related to our treatment. At the same time, his approval and support will also have an impact on the results of the primary election. So this may be a confounder variable that will have an impact on our final causal inference.

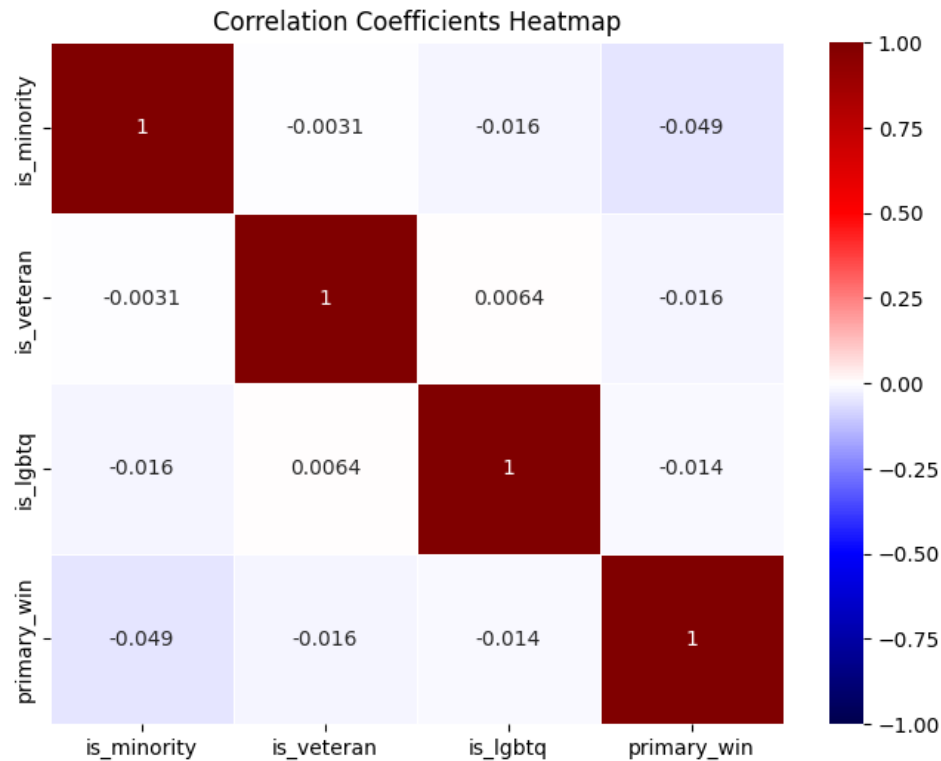
---

#### Is the Candidate Endorsed by Sanders?



#### *Distribution of Candidates by Endorsement of Bernie Sanders*

## 2.2.2 Heatmap of Basic Variables



*Heatmap of the Correlation Coefficients Between Candidate Background Variables and Election Outcomes*

The heatmap of correlation coefficients shows the strength and direction of the linear relationship between the variables `is_minority`, `is_veteran`, `is_lgbtq`, and `primary_win`. The diagonal elements of the matrix, always 1, signify the correlation of each variable with itself. The off-diagonal elements reveal the correlation between different variables.

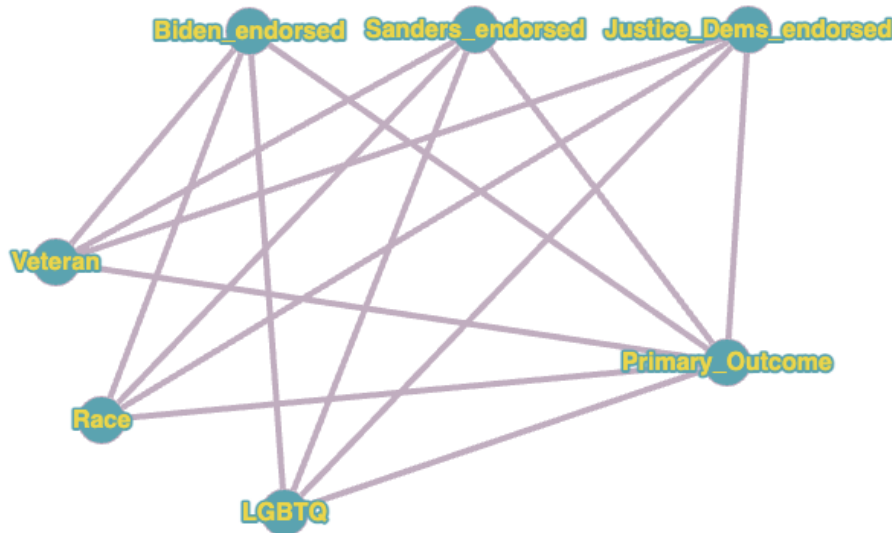
The heatmap of correlation coefficients shows a weak negative association between being a minority and winning primary elections, with a coefficient of -0.049254. Veterans and LGBTQ candidates also show weak negative correlations with winning primaries, with coefficients of -0.016059 and -0.013733, respectively. The independent variables (`is_minority`, `is_veteran`, and `is_lgbtq`) are not strongly related to each other. These findings suggest a need for further research to understand these factors' impact on election outcomes, despite the weak correlations observed.



## 2.3 Method for Question 1

### 2.3.1 Logic Construction

#### 2.3.1.1 DAG



*DAG between Candidate Background Variables, Election Outcomes, and Endorsement Cofounders*

The treatment variables were binary indicators representing the candidate's race (white or non-white), veteran status (yes or no), and LGBTQ identity (yes or no). The outcome variable was also binary, indicating whether a candidate won or lost the primary election. Endorsements from Justice Democrats, Bernie Sanders and Joe Biden were considered as potential confounders in the analysis. There was no direct evidence of colliders in the dataset, so no adjustments were made for them in the analysis. If the unconfoundedness assumption holds in your analysis, this would mean that, conditional on the candidate's endorsements, the potential election outcomes are independent of the candidate's race, veteran status, and LGBTQ identity.

#### 2.3.1.2 Potential Outcome Framework

The causal effect we are interested in is the difference in the potential outcomes under treatment (being a minority, veteran, or LGBTQ) and control (not being a minority, veteran, or LGBTQ) conditions, while adjusting for confounders.

In this framework, for each candidate, we would ideally like to observe two potential outcomes: the outcome if the candidate belongs to the minority/veteran/LGBTQ group (treatment) and the outcome if the candidate does not belong to the minority/veteran/LGBTQ group (control).

However, in reality, we can only observe one of these outcomes for each candidate, as a candidate cannot simultaneously belong and not belong to the same group.

Therefore, the true causal effect is unobservable for each individual candidate. However, under certain assumptions (such as ignorable treatment assignment), we can estimate the average treatment effect (ATE) for the population. The ATE represents the average difference in primary election outcomes between treated and untreated candidates, while adjusting for confounders.

Our goal, then, is to estimate the ATE of each treatment variable on the primary election outcomes, while adjusting for endorsements from Justice Democrats, Bernie Sanders, and Joe Biden. To do this, we will use methods such as logistic regression and matching to control for the confounding effects and provide a more accurate estimate of the causal effects.

### **2.3.2 Logistics Regression**

In our case of predicting Democratic primary election outcomes, logistic regression is an apt choice. It allows us to adjust for multiple confounders, including endorsements from Justice Democrats, Bernie Sanders, and Joe Biden. These confounders, along with the candidate's race, veteran status, and LGBTQ identity, are used as predictors to estimate the log odds of a candidate winning the primary election. The coefficients from this model show the change in these log odds for each unit change in the predictor, while keeping other variables constant. This versatile method can handle interactions between predictors, providing nuanced insights into the relationships between these factors and the election outcomes.

### **2.3.3 Matching**

In our study, matching was used to adjust for confounding variables - specifically, endorsements. The aim was to create a balanced sample where groups (minority, veteran, and LGBTQ candidates vs. their counterparts) were similar in terms of endorsement distribution. This way, any differences in election outcomes could be more credibly attributed to the candidate's background, not the endorsements. This method doesn't rely on assumptions about the relationship between treatment and outcome, making it a robust choice for our analysis.

## 2.4 Results

### 2.4.1 Results for Logistics Regression

Logistic Regression Model Accuracy: 0.68

Logistic Regression Model Coefficients

	Race_Binary	Veteran_Binary	LGBTQ_Binary	Justice_Dems_Endorsed_Binary	Sanders_Endorsed_Binary	Biden_Endorsed_Binary
Coefficient	-0.2498	0.0534	0.2544	0.4238	0.6923	2.0538

*The Coefficients of the Logistic Regression Model*

Based on the results, there seems to be an association between a candidate's race, veteran status, LGBTQ identity, and the outcome of Democratic primary elections, after controlling for potential confounding effects of endorsements.

The coefficients from the logistic regression model suggest that candidates from non-white racial backgrounds have a slightly lower likelihood of winning the primary, given a negative coefficient of -0.2498. Conversely, being an LGBTQ candidate or a veteran shows a positive association with primary outcomes, with coefficients of 0.2544 and 0.0534 respectively. It's important to remember that these are associations, not causal relationships, due to the observational nature of the data.

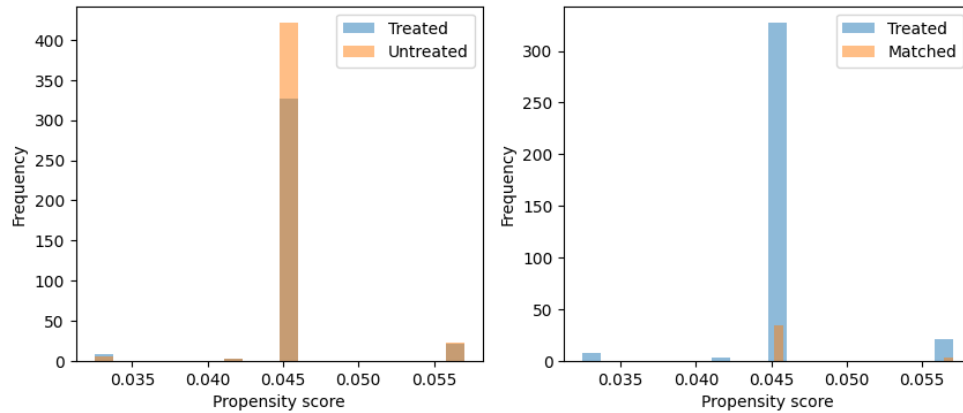
Endorsements appear to have a substantial impact, with those from Biden showing the strongest positive association with a coefficient of 2.0538. This suggests that endorsements could play a vital role in primary election outcomes.

In terms of causality, the assumptions necessary for drawing causal inferences include the absence of omitted variable bias and correct specification of the functional form of the relationship. Although the current study tries to control for confounding effects of endorsements, there might be other unobserved variables that could bias the results, such as campaign funding or public awareness of the candidate. The linear relationship assumption between the log-odds of the outcome and predictor variables is also crucial and needs to be verified.

Regarding the magnitude of the effect, the logistic regression coefficients are in log-odds units. To provide a more intuitive understanding of the effect size, these could be transformed into odds ratios. For instance, the odds ratio for a non-white candidate winning the primary is approximately  $e^{(-0.2498)} = 0.78$ , suggesting that non-white candidates have about 78% of the odds of winning compared to white candidates, controlling for other variables. The uncertainty in these estimates could be expressed through confidence intervals.

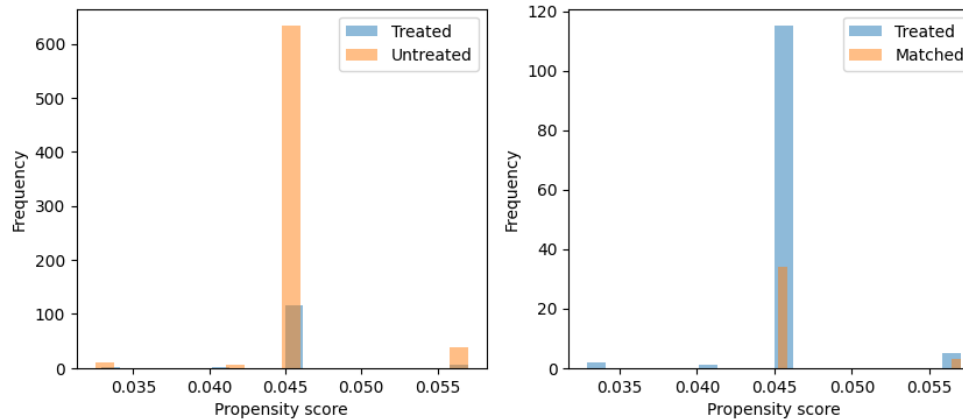
## 2.4.2 Results for Matching

Propensity score distribution before and after matching - Race



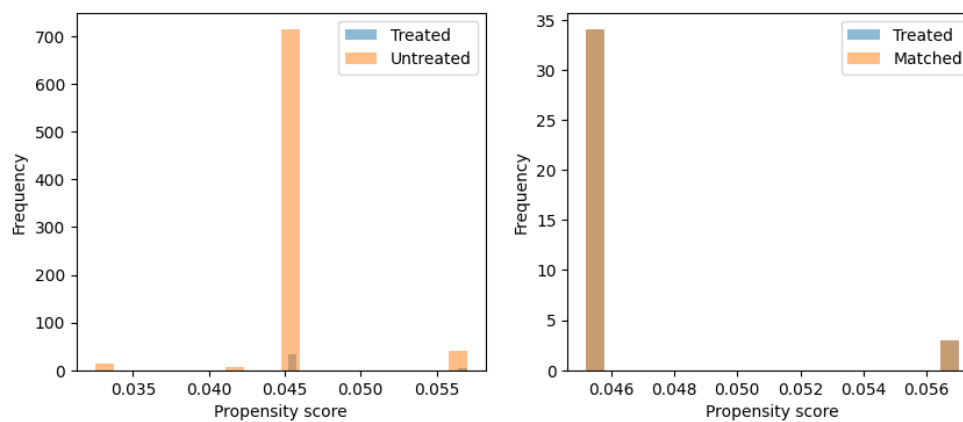
*The Distribution of Propensity Scores Before and After Matching on the Candidates' Races*

Propensity score distribution before and after matching - Veteran

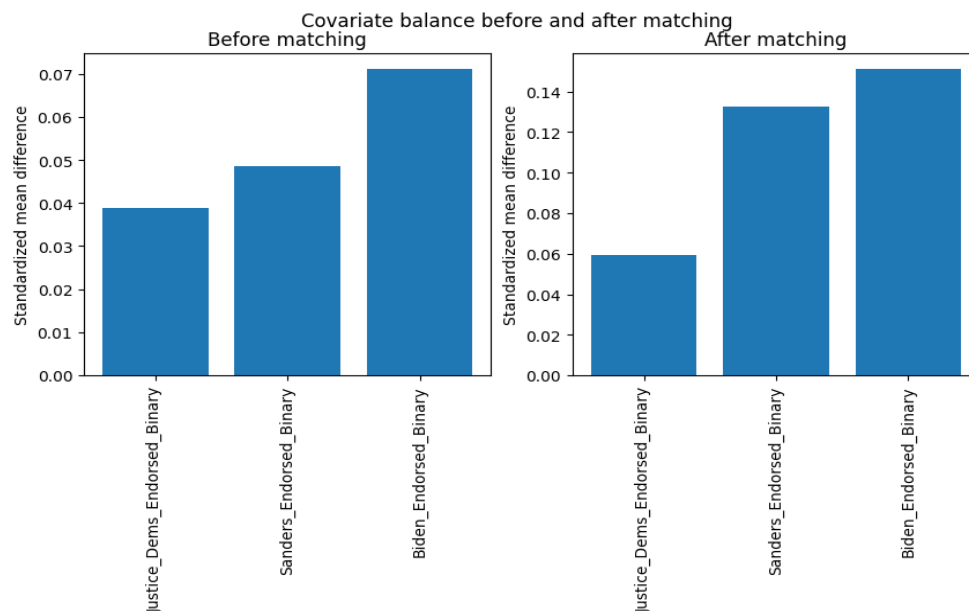


*The Distribution of Propensity Scores Before and After Matching on the Candidates' Veteran Status*

Propensity score distribution before and after matching - LGBTQ



*The Distribution of Propensity Scores Before and After Matching on the Candidates' LGBTQ Identification*



*The Distribution of Propensity Scores Before and After Matching on the Candidates' Races*

The propensity score matching procedure improved the balance of covariates between treated and untreated groups for all three treatment variables, which suggests that the treatment effect estimates derived from this matched sample may be more reliable and less biased. The matching procedure also improved the balance of observed covariates between the groups, which reduces the bias due to confounding variables. However, unobserved confounding variables may still exist and limit the ability to draw causal inferences. Overall, the use of propensity score matching appears to be appropriate for this research question.

## 2.5 Discussion

Now, let us talk about the limitations of the Methods. Logistic regression assumes a linear relationship between the log odds of the outcome and the predictor variables. If this assumption is violated, our estimates may be biased. Moreover, logistic regression estimates are sensitive to model specification. If we omit relevant variables or include irrelevant ones, the estimates could be misleading. While matching can be powerful, it relies on the availability of good matches in the data. If there are not enough similar untreated individuals for each treated individual, the method can struggle. Additionally, matching can only control for observed confounders. If there are unmeasured confounders, estimates may still be biased.

For additional Useful Data, other demographic factors such as age, education level, and socioeconomic status could also influence election outcomes and should ideally be controlled for. Also, a candidate's political experience could influence both their chances of receiving endorsements and winning elections. Data on previous political roles or offices held would be

useful. Local political climates can vary widely, and the same candidate might fare differently in different locations. Data on the location of each election could help control for this.

Our confidence in the causal relationship between our chosen treatment and outcome should ideally be based on the strength and significance of our results, as well as the plausibility and satisfaction of our causal assumptions. Our regression and matching results don't show a strong and statistically significant relationship between the treatments and the outcome, so it would increase our confidence in a causal relationship.

Also, the unconfoundedness assumption can't hold. In our case, the unconfoundedness assumption is untestable, and there's always a risk of unmeasured confounding. This would decrease our confidence in a causal relationship.

## 2.6 Conclusion

Our analysis suggests an association between a candidate's race, veteran status, LGBTQ identity, and the likelihood of winning Democratic primary elections. Non-white candidates appeared slightly less likely to win, while veterans and LGBTQ candidates showed a slight positive association. Endorsements, especially Those from Biden, showed a substantial positive impact. But we cannot conclude that there is a causal relationship.

The results are specific to Democratic primary elections and may not apply to general elections or primaries of other parties. It's also geographically and temporarily bound by the data.

Given the influence of endorsements, political organizations might focus on supporting diverse candidates in gaining prominent endorsements. A concerted effort to support non-white, veteran, and LGBTQ candidates might also be beneficial to counterbalance any potential biases.

We don't merge different data sources. If different data sources were merged, the benefits would include a more comprehensive view of the factors impacting election outcomes.

The analysis may not have accounted for all confounding variables, such as campaign funding or public awareness of the candidate. The assumed linear relationship between the log odds of the outcome and predictors also needs verification.

Research could explore the impact of other potential confounders like political experience, local political climates, and additional demographic factors. Studies could also investigate these relationships in different contexts, such as general elections or primaries of other parties .

Finally, this project underscored the importance of considering potential confounders when investigating causal relationships and reinforced the need to scrutinize the assumptions of the chosen analysis methods. It also demonstrated the value of robust statistical techniques like logistic regression and matching in observational studies.

## 3. Research Question 2

### 3.1 Introduction

#### 3.1.1 Purpose

For the research question 2, we seek to investigate the influence of Party Affiliation to one candidate's pattern of **Financial Spending**. Specifically, we are interested in identifying whether there is a routine of spending Election Funding for candidates belonging to the same political party. To achieve this goal, we would use both **GLM** and **Nonparametric** methods, predicting **Total Contribution** as the target variable and use multiple features concerning party affiliation along with **Total Disbursement** as independent variables. The regression on **Total Contribution** based on **Total Disbursement** could imply a pattern of election spending for a certain candidate, adjusted by the other variables that indicate the influence of party affiliation.

#### 3.1.2 Model of Choice

For this regression model, GLM would be a good choice to implement since we don't know about the population distribution of Total Contribution but only a sample of this dataset, which we assume to be non-normal because parties with more popularity are usually more financially stable. For nonparametric method, since both the feature **Total Disbursement** and target **Total Contribution** are continuous, we would use the pre-trained *DecisionTreeRegressor* model from *Scikitlearn* package, which makes no assumption on the distributions and would return an interpretable result.

#### 3.1.3 Datasets

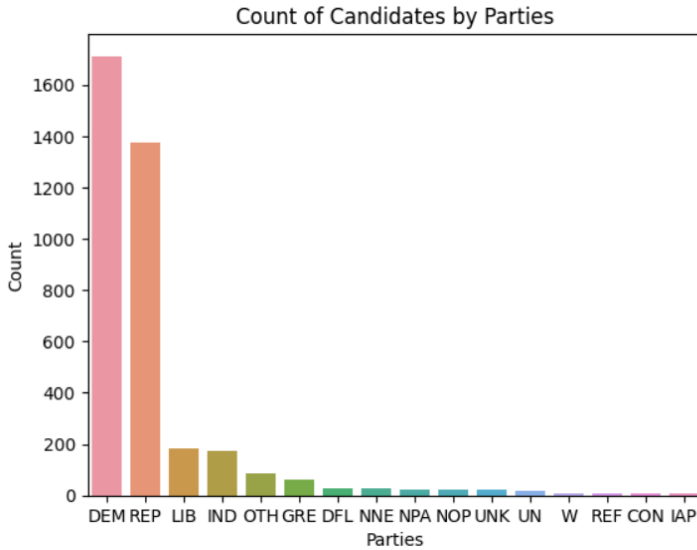
The datasets used for this research question are *Candidates* of year 2018 (*Candidate18* in later text) from the Federal Election Commission, which contains the main financial benchmarks of individual candidates per year. We also used an outside dataset *Electoral Votes* as a complement to include the effect of different states to the financial spending patterns.

## 3.2 EDA for Question 2

### 3.2.1 Count of Candidates by Parties

To explore the dataset *Candidate18*, we firstly plot the distribution of candidates to identify potential bias. As the plot below shows, the candidates in this dataset are mostly affiliated to Democrats or Republicans. Thus, this dataset is potentially biased since the sample for smaller parties are scarce, which might cause the regression result to be subject to undercoverage bias. To cope with this, we may combine the smaller parties together in the model and conduct cumulative pattern analysis, but that would be less representative for each party. Thus, the candidates affiliated to Democrats and Republicans will be the focus for this research question.

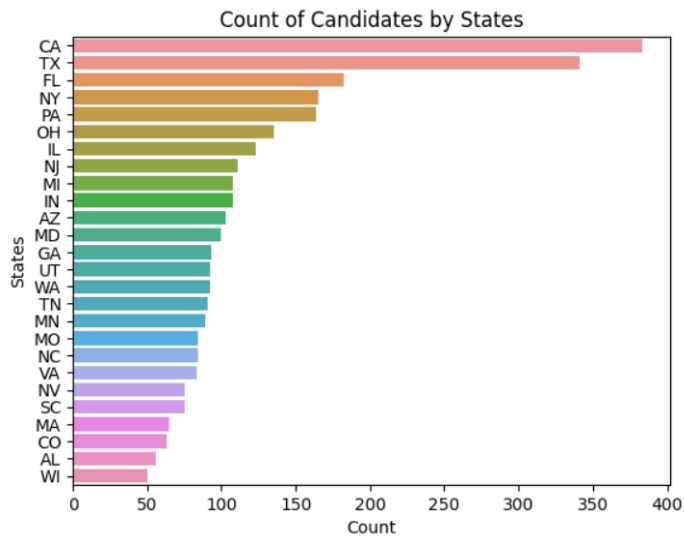




*Count of Candidates by Parties*

### 3.2.2 Count of Candidates by States

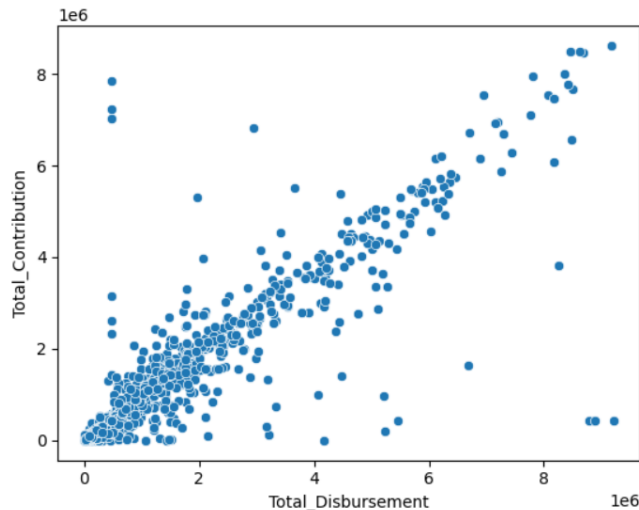
Then we visualized the distribution of candidates by states. As the plot below shows, the number of candidates varies among states with most are in the states with larger populations. Among the states, California and Texas seem to have the most candidates, while those two states also have the highest number of electoral votes. Potentially, spending in different regions usually possess different traits, as it might be higher in areas with higher living costs. As a result, the size of the state one candidate belongs to might also be a potential influencer to his/her spending pattern.



*Count of Candidates by States*

### 3.2.3 Scatterplot for Total Disbursement and Total Contribution

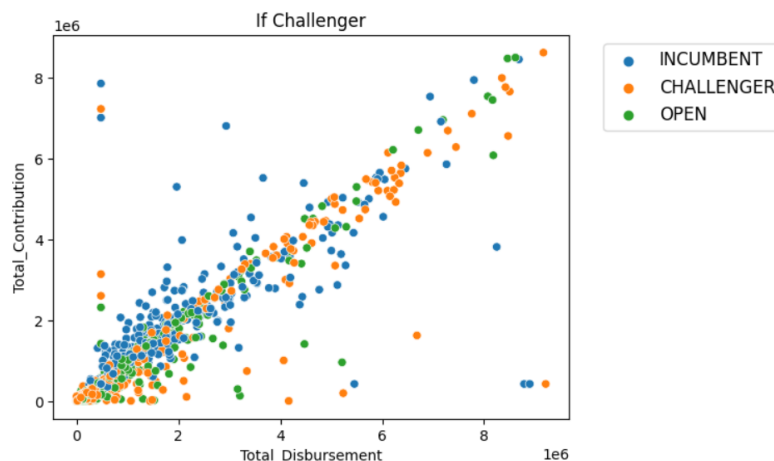
This plot shows the relationship between **Total Disbursement** and **Total Contribution** before manipulations. As the plot below shows, a somehow linear relationship could be noticed but there are still many outliers. Those outliers could either be from smaller parties or abnormality in the two major parties. Hence, we will try to exclude those by adding the party affiliation factors as adjustments. This plot also shows that many data points concentrated at around (0, 0), which indicates a possible left-tailed distribution for both **Total Disbursement** and **Total Contribution**.



*Total Contribution v.s. Total Disbursement*

### 3.2.4 Role in the Election

The graph *If Challenger* below plots the relationship between **Total Disbursement** and **Total Contribution** divided by whether a candidate is a Challenger. From the plot we can notice that there is no clear boundary among different states, as all three roles have points distributed in different sections of the linear pattern. However, intuitively we think whether a candidate is a challenger could influence the likelihood of spending more on increasing popularity.



*Role in the Election divided by If Challenger*

### 3.3 Method & Discussion for Question 2

For this research question, we tried to predict the dollar value for the total contributions given to a campaign for a candidate running for the House of Representatives, Senate, or Presidency. The data source defines this total contribution number to be the sum of “total contributions from individuals,” “contributions from other committees,” “contributions from party committees,” and “contributions from the candidate.” To predict this, we used features representing total disbursement for the campaign, the office being run for (House/Senate/Presidency), the candidate’s political party (Democratic/Republican/Third party), whether the candidate is an incumbent or challenger, and the number of electoral college votes for the state for state-wide elections. We thought these features would all be particularly relevant and had the potential to provide meaningful insights and findings for our research question.

The GLM we used was a linear regression model. This choice seemed fairly straightforward solely based on the fact that we wanted the domain of our output variable to be the set of all real numbers. The model assumes that the input features are quantitative, so we had to preprocess the data to implement one-hot encoding for our categorical features.














For our nonparametric model, we elected to use a decision tree regression model. The primary reason we chose to use decision trees was because of the interpretability insights the final model output can provide, with the nodes of the trees making clear indications as to how the model is making its decisions. We also thought decision trees would be beneficial to use for the flexibility and lack of assumptions we had to make about our data to use them, as well as their decent performance with predictions. We could’ve opted for a potentially higher performance model such as random forests, but we were more interested in the interpretability of the model.

To evaluate performance for our models, we’ll use standard quantitative variable error metrics for our model predictions for both training and test sets, including root mean squared error, mean absolute error, and the r-squared metric. When analyzing our final models, though, pure performance and metrics aren’t the only thing we’ll be looking at. Explainability and interpretability of the models are highly emphasized for this research question since we’re particularly interested in why the model would predict a certain number or another. We also are specifically interested in the relationships between each of the features and the target variable of total contributions.

3.4 Results

3.4.1 Bayesian GLM Results

▼ Data variables:

intercept	(chain, draw)	float64	420.6 374.5 418.3 ... 425.7 395.1	 
dem_coef	(chain, draw)	float64	338.2 399.7 357.3 ... 397.7 362.8	 
rep_coeflope	(chain, draw)	float64	85.35 79.55 47.42 ... 63.02 108.3	 
third_coef	(chain, draw)	float64	-6.789 -16.98 ... -14.65 -21.28	 
challenger_coef	(chain, draw)	float64	-7.661 31.86 -40.31 ... 4.839 22.33	 
Electoral_Colle...	(chain, draw)	float64	20.44 20.56 16.91 ... 18.06 17.89	 
disbursement_f...	(chain, draw)	float64	0.8336 0.8336 ... 0.8336 0.8335	 

Bayesian GLM Results

3.4.2 Frequentist GLM Results

Generalized Linear Model Regression Results						
=====						
Dep. Variable:	Total_Contribution	No. Observations:	3793			
Model:	GLM	Df Residuals:	3787			
Model Family:	Gaussian	Df Model:	5			
Link Function:	identity	Scale:	1.7733e+11			
Method:	IRLS	Log-Likelihood:	-54501.			
Date:	Mon, 08 May 2023	Deviance:	6.7154e+14			
Time:	22:58:25	Pearson chi2:	6.72e+14			
No. Iterations:	3	Pseudo R-squ. (CS):	0.9946			
Covariance Type:	nonrobust					
=====						
	coef	std err	z	P> z	[0.025	0.975]
-----						
const	6.363e+04	1.12e+04	5.661	0.000	4.16e+04	8.57e+04
dem	5.884e+04	9869.919	5.961	0.000	3.95e+04	7.82e+04
rep	8009.4700	1.01e+04	0.796	0.426	-1.17e+04	2.77e+04
third_party	-3216.9803	1.33e+04	-0.242	0.808	-2.92e+04	2.28e+04
challenger	-8.094e+04	1.47e+04	-5.522	0.000	-1.1e+05	-5.22e+04
Electoral_College_Votes	-220.0274	445.764	-0.494	0.622	-1093.708	653.653
Total_Disbursement	0.8140	0.006	133.738	0.000	0.802	0.826
=====						

Statistical Summary Table

Metrics	Frequentist Interpretation Results
Mean squared error	177048005231.89
Root mean squared error	420770.73
Mean absolute error	129190.32
R-squared	0.84

### 3.4.3 GLM Result Analysis

Now getting to the results of the GLM model, both the Bayesian and frequentist interpretation implementations agree that when it comes to party affiliation, being a third party candidate is negatively correlated with total contribution. Between the two primary parties, both are positively correlated with total contribution, but the coefficient corresponding to being a candidate for the Democratic party is much greater. The two implementations also agree that being a challenger as opposed to an incumbent has a negative correlation with total contribution. The one feature the implementations disagree on having a positive or negative correlation is the electoral college votes the state has. While the implementations generally agree over whether variables have a positive or negative correlation with the outcome, there's significant difference in the scaling of the coefficients and some disagreement with the relative scaling of the coefficients compared to each other. The Bayesian implementation coefficients are generally smaller, and the coefficient corresponding to the candidate being a Democrat is the largest. The frequentist implementation coefficients are generally larger, and the coefficient corresponding to the candidate being a challenger or incumbent is the largest. If we'd used different prior distributions for our Bayesian implementation, perhaps this difference could be reduced or exaggerated.

When it comes to the uncertainty of the predictions, the frequentist implementation had a poor root mean squared error of ~\$420k but a much better mean absolute error of ~\$130k (meaning our predictions were on average ~\$130k off of the actual dollar amount). This is because our predictions had some significant outliers, which will balloon the root mean squared error metric. Inclusion of additional data could've improved these metrics, such as data on the opposing candidate(s). The models are also limited by a low amount of training data with lots of variability. Because the GLM models have high uncertainty, even by the mean absolute error metric, I'd hesitate with applying this model to future datasets without further improvements.

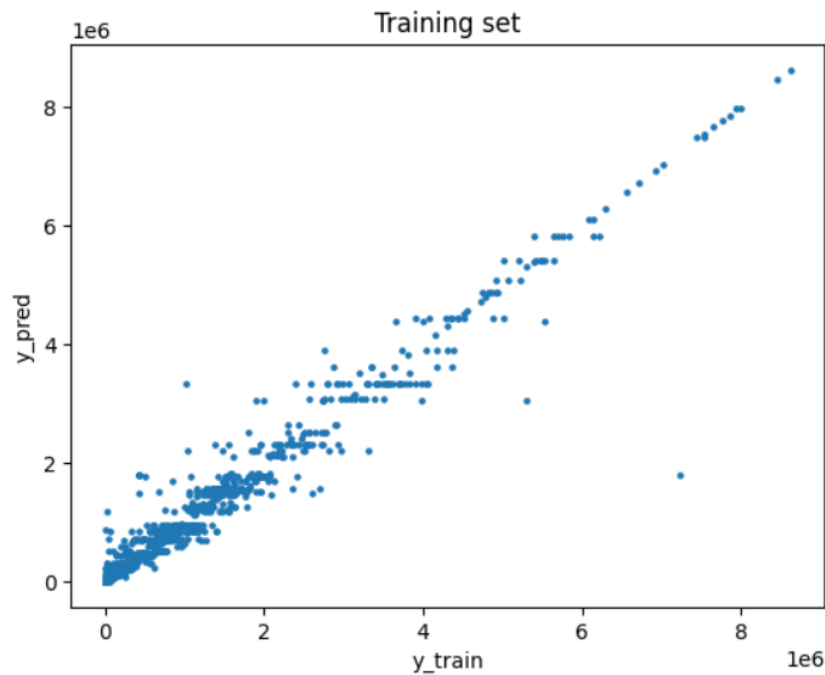
### 3.4.4 Nonparametric Result

The results using major metrics for DecisionTreeRegressor are as follows:

Metrics	Test set	Training set
Mean squared error	189276336710.31	33875816233.46
Root mean squared error	435059.01	184053.84
Mean absolute error	113344.21	52996.37
R-squared	0.84	0.96

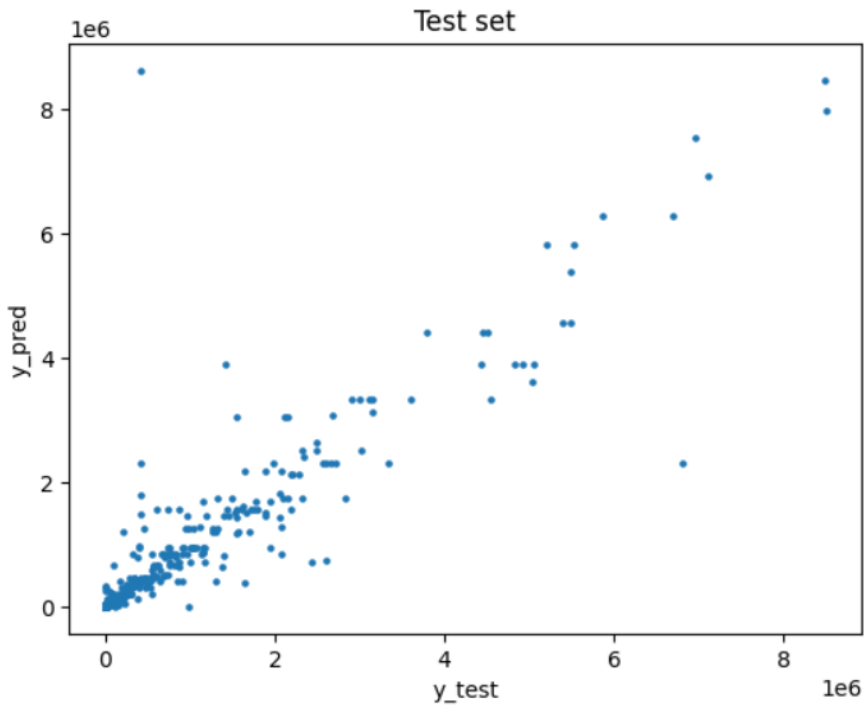
As the chart illustrated, the nonparametric model has achieved excellent accuracy on the training set, with a ***R-squared*** of close to 1. For the test set, the R-squared was much smaller. This could be derived from an overfitting nature of the decision tree model on the training set. To adjust such overfitting, we tried to change the `max_depth` parameter to achieve the best performance of the model, and obtained the best results of a ***max\_depth = 8***.

The result for the training set is illustrated by the scatterplot below. As the plot tells, we successfully get rid of the influence from abnormality by training a strong linear relationship between the ***y\_pred*** and ***y\_train***.



*Prediction v.s. Truth for Training Set*

For the test set, however, the linear relationship is less outstanding than that for the training set. On the other hand, we still exclude much of the abnormalities by adding the party affiliation, electoral, challenger status features as adjustments.



*Prediction v.s. Truth for Test Set*

### 3.4 Conclusion

In conclusion, the nonparametric strategy performs a little better in modeling the relationship by adding our selected features as adjustment. However, it performs much better on the training set than on the test set due to the overfitting nature of the decision tree model. Additionally, this method successfully modeled the left-tailed distribution of the data points since both the training and test graph had most data points concentrated around (0, 0). This might be a result of the limited sample size since this dataset contains only candidates of a single year. It is also because of us not knowing the true distribution of the features we use and hence choosing less reliable prior distributions. On the contrary, due to the maximum depth of 8, the nonparametric model is not as interpretable as the GLM since there are as many as 8 layers of decision to make until we reach the final results.

## Citation

<https://www.archives.gov/electoral-college/2020>