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

1. Data Overview

How were your data generated? Is it a sample or census?
• If you chose to use your own data, describe the data source and download process.
• If you chose to add additional data sources, explain why.
• If your data represents a sample:
– Compare the distribution of one of your variables to what is expected in the population. For example, if your data has an age variable, compare it to the age structure of the population.
∗ Do you notice any differences?
∗ How does this affect the generalizability of your results?
• If your data represents a census:
– Are there any groups that were systematically excluded from your data?

2. Research Question 1

2.1 Introduction

3.1.1 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.

3.1.2 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?

3.1.3 Datasets

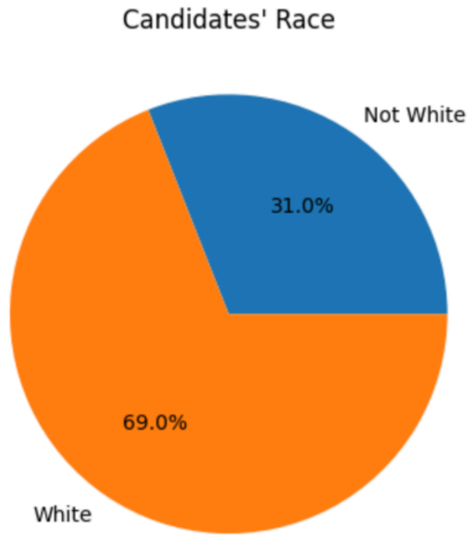
The dataset we used is from the “FiveThirtyEight: 2018 Primary Candidate Endorsements”. The “dem_candidates.csv” provides us the information about the 811 candidates who have appeared on the ballot this year in Democratic primaries for Senate, House and governor in 2018.

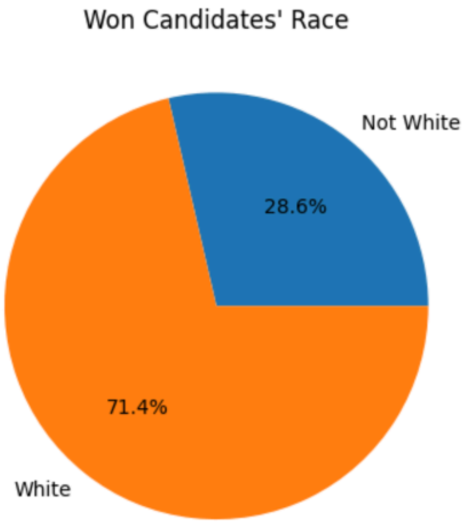
2.2 EDA for Question 1

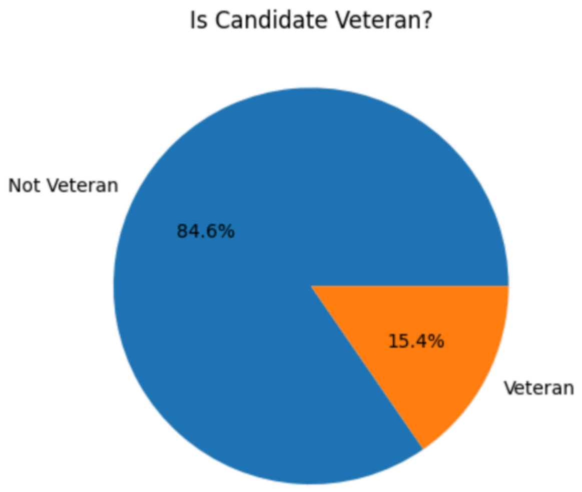
First, we want to isolate the factors we want to focus on, which are the factors. We extract "Race", "Veteran?", "LGBTQ?", "Primary Status", "District", "Justice Dems Endorsed?", "Sanders Endorsed?", "Biden Endorsed?" and "Primary %”. These are the Demographic factors we are going to discuss. For the first three attributes, which are “Race”, “Veteran”, and

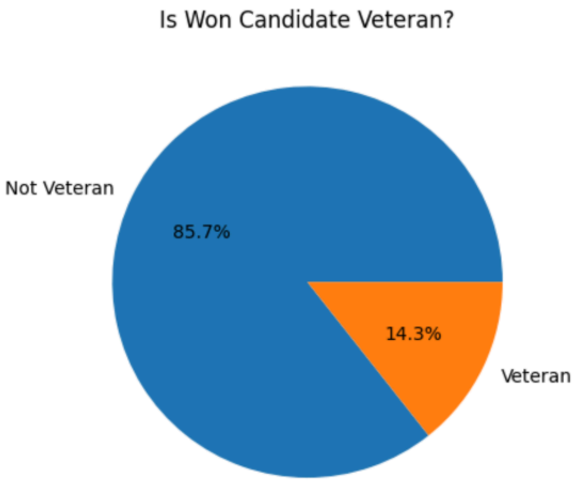
“LGBTQ”, we first feature engineered the categorical variables into 1 and 0. After that we use pie charts to show the proportion in each group, as the graph below.

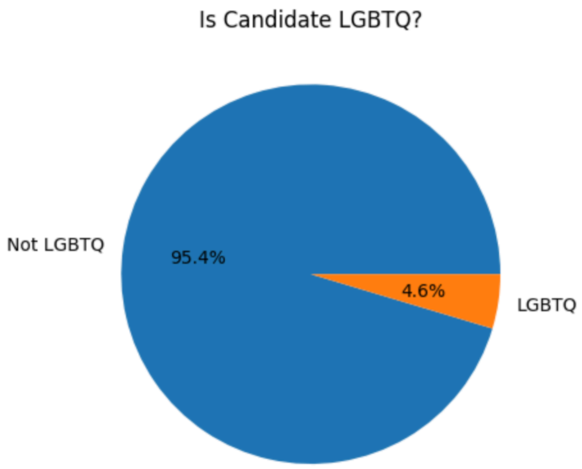
Among the specific backgrounds of the candidates we focused on, we found that white, non-Veteran, and non-LGBTQ groups were in the majority. But that percentage didn't change significantly among primary-victory candidates, perhaps suggesting that certain backgrounds don't give candidates a clear advantage. These **three** variables are the objects of our research as **categorical variables**. This visualization will give an intuitive answer to our research results, but we still need to eliminate the influence of some bias and confounding on the results through follow-up research.

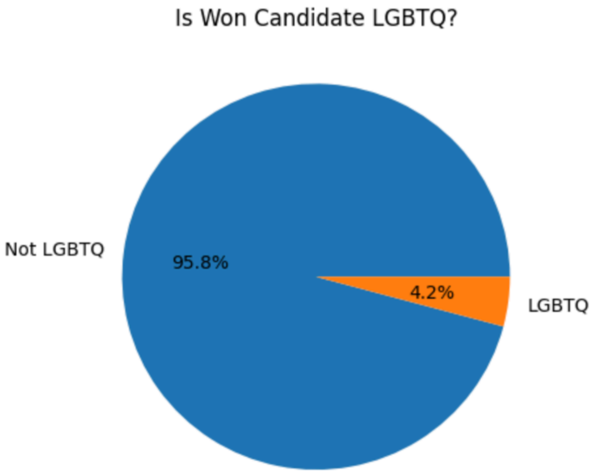




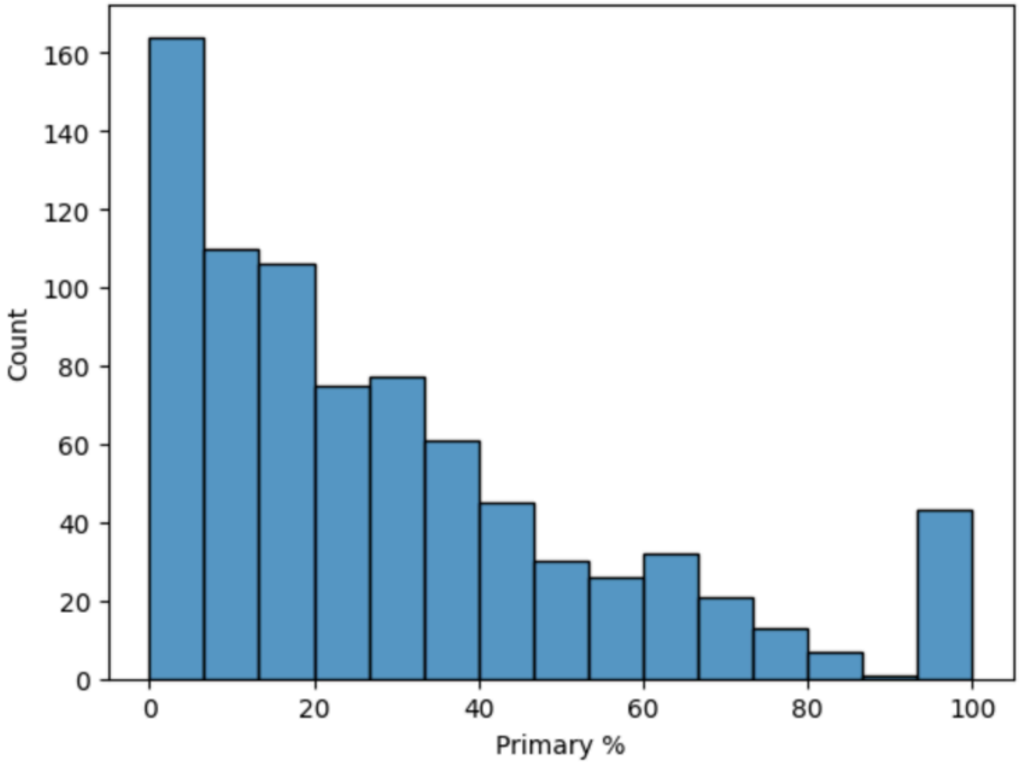




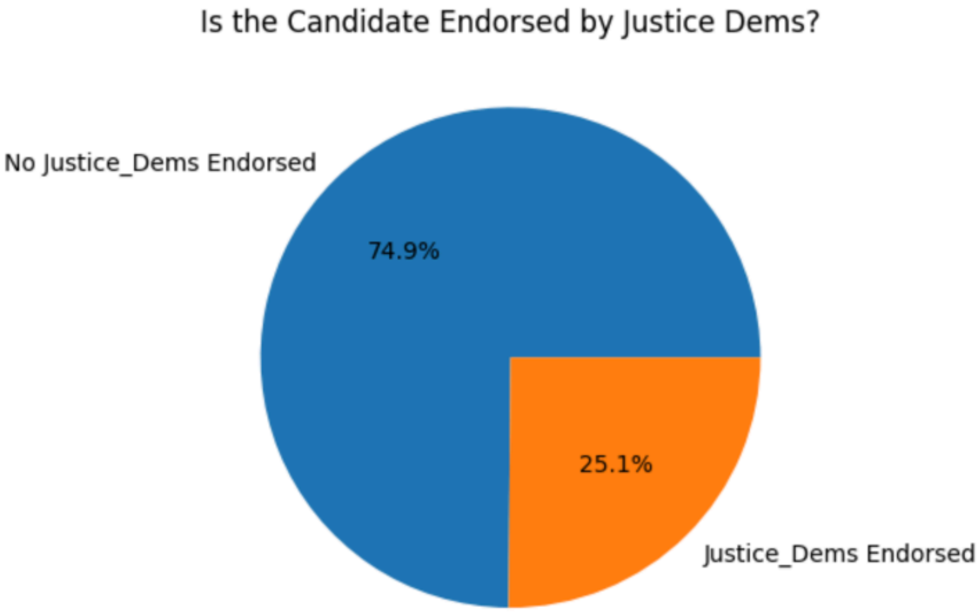




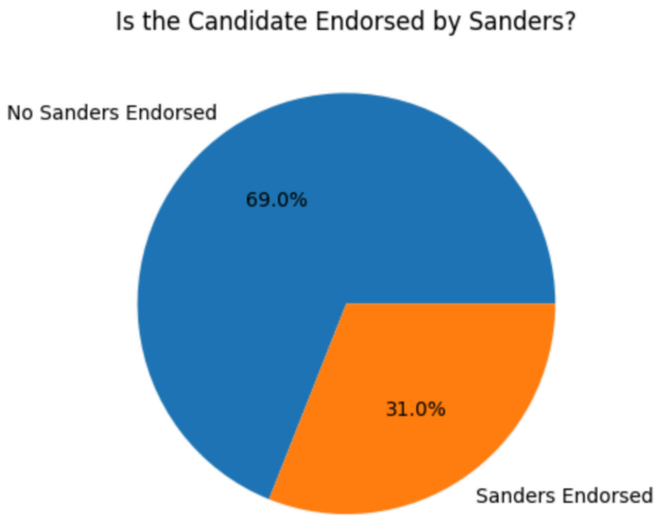
Next, we went on the “Primary %” feature. We looked at the distribution of primary votes. We found that the proportion of votes in the range of 50-90 percent is very small, the proportion of 0-20 percent is high, and the proportion of 100 percent is relatively high. As a **quantitative variable**, the vote rate in the primary election is not only related to the specific background we focus on, but may also affect the final primary election results, because some states need to conduct primary elections in the state before deciding whether to enter the general primary election . So "primary percent" may have an impact on our casual inference as a confounder variable.



We also looked at the distribution of Justice Democrats endorsement. We found that about a quarter of the candidates could get the support of this party. We follow him because, based on their website descriptions and their political advocacy, the specific backgrounds of the candidates we research are likely to be their main considerations. So we visualized it by converting it into a **quantitative variable**. This variable is not only related to race, veteran, and LGBTQ, but also affects the results of the primary election through publicity and sponsorship, so it may be used as a confounder variable to affect our final casual inference.



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.



2.3 Method & Discussion for Question 1

2.3.1 Method

In this study, the effect of a candidate's background on the outcome of Democratic primary elections was analyzed, focusing on the candidate's race, veteran status, and LGBTQ identity. A logistic regression model was used to examine the relationship between these factors and the primary election outcome, while controlling for potential confounding factors such as endorsements from Justice Democrats, Bernie Sanders, and Joe Biden.

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.

To control for confounding factors, endorsements were considered as potential confounders in the analysis. The unconfoundedness assumption was justified based on the argument that endorsements play an important role in political campaigns, and can be related to both a candidate's background and the likelihood of winning the primary election. Logistic regression was used to adjust for these confounding variables, which were included in the model as predictors alongside the treatment variables.

There was no direct evidence of colliders in the dataset, so no adjustments were made for them in the analysis.

2.3.2 Discussion of Related Issues:

Model assumptions: Logistic regression assumes a linear relationship between the log-odds of the outcome and the predictor variables. It's important to verify whether these assumptions hold in the context of the research question. If not, alternative modeling techniques such as generalized additive models or tree-based models could be explored.

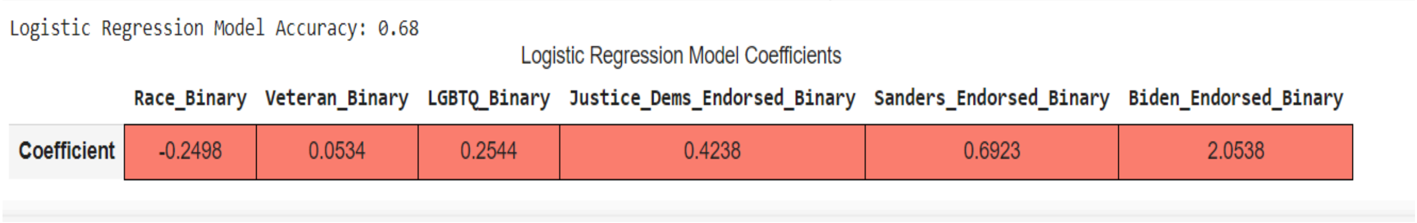
Confounder selection: Although endorsements were considered as potential confounders in this study, there might be other confounding variables not included in the analysis, such as campaign funding, political experience, and public awareness of the candidate. Future research should consider incorporating additional confounders to better understand the relationship between a candidate's background and election outcomes.

Model performance: The accuracy of the logistic regression model was reported, but other performance metrics, such as precision, recall, and the area under the ROC curve, could also be evaluated to provide a more comprehensive understanding of the model's performance.

Effect size interpretation: The logistic regression coefficients provide information about the direction and magnitude of the relationships between the predictor variables and the outcome. However, the interpretation of these coefficients in terms of odds ratios or marginal effects could be more informative for understanding the practical implications of the results.

Generalizability: The results of this study are specific to Democratic primary elections, and may not be generalizable to other election contexts, such as general elections or primaries for other political parties. Future research should explore the impact of candidate backgrounds on different types of elections to better understand the broader implications of these findings.

2.4 Results



```
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 results suggest that a candidate's race, veteran status, and LGBTQ identity are associated with the outcome of Democratic primary elections, after accounting for potential confounding effects of endorsements. Candidates from non-white racial backgrounds (Race_Binary) have a slightly negative association with the primary outcome, as indicated by a negative coefficient (-0.2498). In contrast, candidates with LGBTQ identity (LGBTQ_Binary) show a positive association with primary outcomes, with a coefficient of 0.2544. Veteran status (Veteran_Binary) also has a small positive association with a coefficient of 0.0534.

In terms of endorsements, they appear to have a significant impact on primary outcomes. Candidates endorsed by Justice Democrats, Bernie Sanders, or Joe Biden have higher odds of winning the primary. The Biden endorsement (Biden_Endorsed_Binary) has the strongest association among the variables, with a coefficient of 2.0538.

These results help address the research question by providing evidence for associations between a candidate's background and the outcome of Democratic primary elections. Specifically, they suggest that race and LGBTQ identity are factors that may influence primary outcomes, while veteran status has a smaller effect. The results also demonstrate that endorsements, particularly those from prominent political figures, play a significant role in primary election outcomes.

However, it is important to emphasize that these findings do not establish causal relationships due to the observational nature of the data. While they provide valuable insights into the associations between candidate background and election outcomes, more robust study designs, such as randomized controlled trials or natural experiments, would be necessary to draw stronger causal conclusions. Nevertheless, this analysis offers a valuable starting point for understanding the factors that can influence Democratic primary elections and can guide future research in this area.

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**. In specifics, we are interested in identifying whether there is a routine of spending the Election Funding for candidates belonging to the same political party. To ahieve this goal, we would use both **GLM** and **Nonparametric** methods, predicting **Total Contribution** as the target variable and use 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 sample of this dataset, which we assume to be non-normal because parties with more popularity is 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 *DescisionTreeRegressor* model from *Scikitlearn* package, which makes no assumption on the distributions and would return an interpretable result.

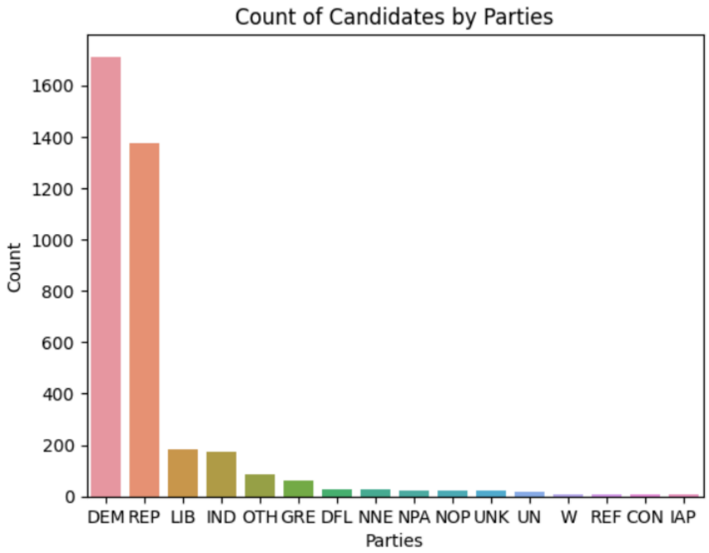
3.1.3 Datasets

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 complement to include the effect of different states to the financial spending patterns.

3.2 EDA for Question 2

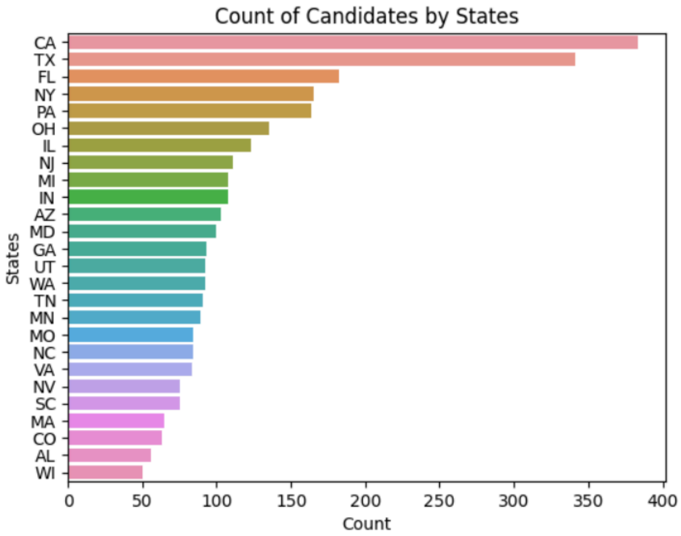
3.2.1 Count of Candidates by Parties

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.



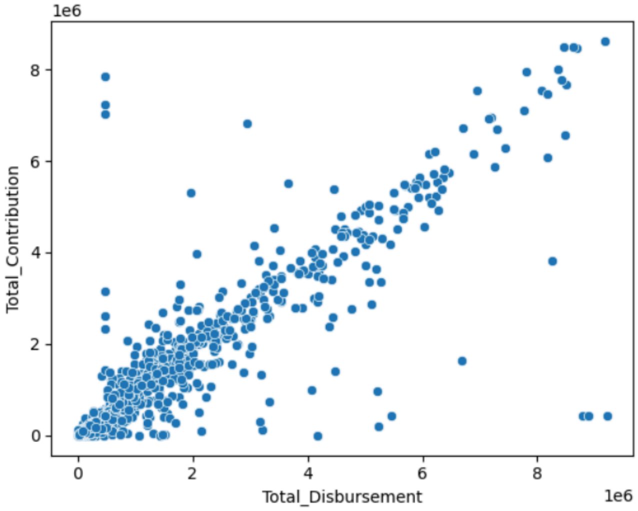
3.2.3 Count of Candidates by States

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 has the highest number of electoral votes. Potentially, spending in different regions usually possess different traits, as it might be higher in area with higher living cost. As a result, the size of the state one candidate belongs to might also be a potential influencer to his/her spending pattern.



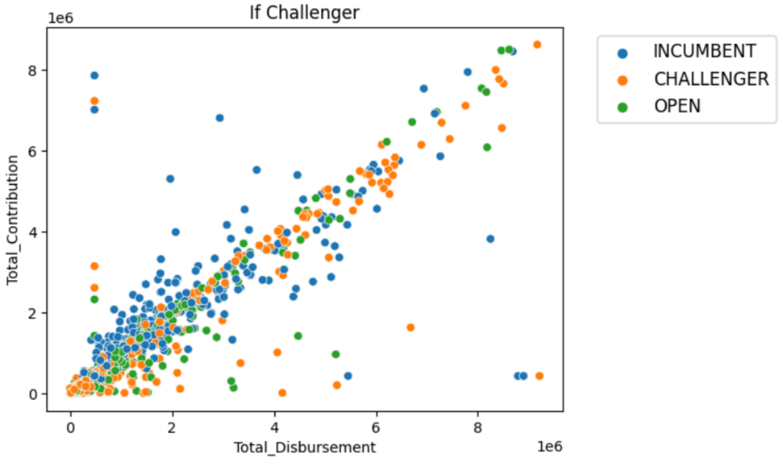
3.2.3 Scatterplot for Total Disbursement and Total Contribution

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 the 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***.



3.2.3 Role in the Election

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



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.

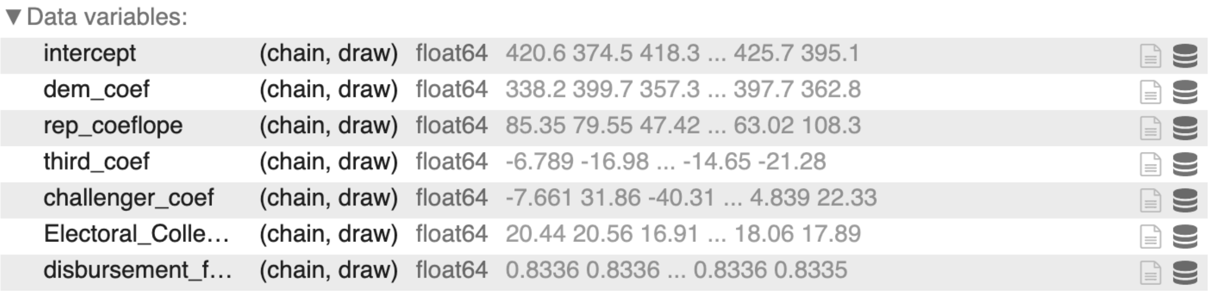
The GLM we used was a linear regression model. This choice seemed fairly straightforward solely based on the fact that we wanted the domain of our output variable to be the set of all real numbers. The model assumes that the input features are quantitative, so we had to preprocess the data to implement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

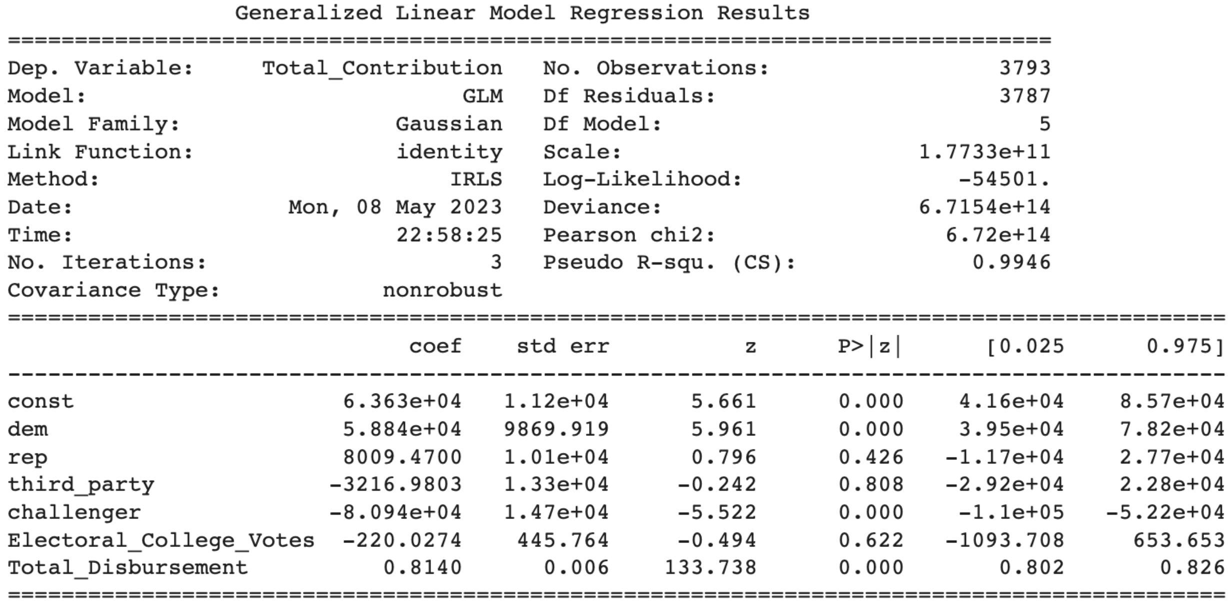
3.4 Results

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
```

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
===========================================================================================
```

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.

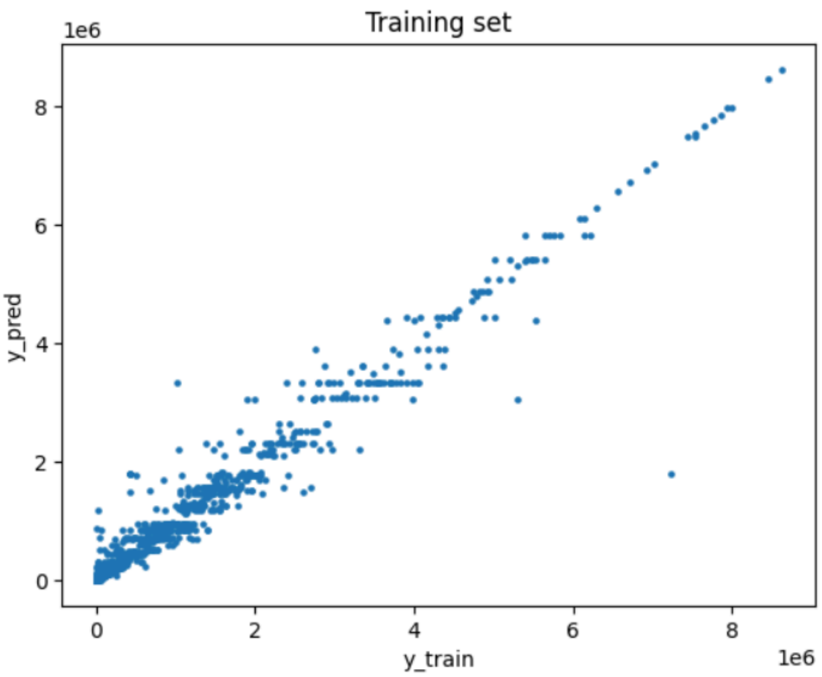
3.4.4 Nonparametric Result

The results using major metrics for DescisionTreeRegressor are as follows:

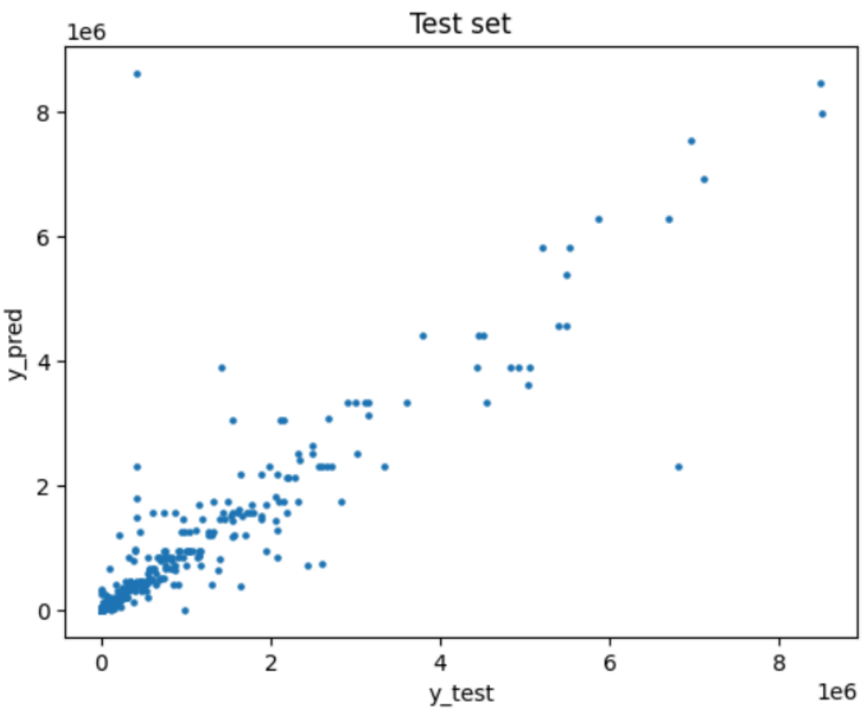
Metrics	**Test set**	**Training set**
Mean squared error	189276336710.31	33875816233.46
Root mean squared error	435059.01	184053.84
Mean absolute error	113344.21	52996.37
R-squared	0.84	0.96

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 tuning 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 training set is isslustrated 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*** an ***y_train***.



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.



As a result, the nonparametric strategy perform well 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 has most data points concentrated around (0, 0). Compared to GLM for this dataset, the nonparametric method performs better in accuracy. 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 choosed wrong priors. In 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.fec.gov/data/candidates/?election_year=2018&election_full=true&is_active_candidate=true&sort=-total_receipts