Data 102 Final Project Written Report

Data Overview

The 2018 Primary Candidate Endorsements dataset collects data from multiple reliable sources: Ballotpedia, Secretaries Of State, Associated Press, Candidate Websites, Votesmart, Federal Election Commission, Obama Alumni Association, Various News Reports. It provides information on the Democratic and Republican primary elections for the U.S. Senate, U.S. House, and governor in 2018. Since the data contains information of candidates who appeared on the ballot, the dataset is a census rather than a sample.

That being said some elections are systematically not included such as senators who were excluded from our data because in 2018 there were no senator elections held. There was also no presidential election held in this year, so the president and presidential candidates were also excluded. Incumbents were excluded from the dataset as well, however this only applies towards candidates running for the same position. If the candidate is in office and running for a different position they were not excluded. Only popular candidates are included in ballots so write-in candidates may have been excluded.

Besides using the dataset from FiveThirtyEight, we also use data from the Center For American Women and Politics to extract all women who joined the primary election in 2018. Because unconfoundedness is one of the techniques used in our research, we need more information to control all of our potential confounders such as gender. Having a list of women in the 2018 election, we can easily classify candidate gender in our dataset.

For the primary election, candidates are aware that they need to register and provide information about themselves and their campaign. Their data was collected by tracking candidate websites. Therefore, they are generally aware that their information will be made public. However, their awareness might not extend to the specifics of how their data is handled post-election and how their data was gathered from different sources other than their main communication channel. Our dataset wasn't modified for differential privacy.

Regarding data granularity, we only use one type of dataset, so we don't have to deal with different granularity. The granularity of our dataset is at the level of individual candidates. Every row represents a candidate, and the columns contain information related demographic information and endorse-related details of each candidate. Analyzing individual candidate data allows for a detailed examination of demographic patterns, voting outcomes, and the influence of various factors on candidates' success.

For our research question, the inclusion of parental socioeconomic status features is important. It serves as a crucial confounder in our causal analysis, aiming to examine the relationship between possessing a STEM degree and succeeding in an election. Achieving unconfoundedness is a critical assumption. However, the absence of parental socioeconomic status may introduce bias, jeopardizing the integrity of our findings.

Moving on to data completeness, it is notable that many columns with missing data which primarily are found in the endorsements section of the dataset. These missing values exist because many of the candidates in our dataset are running in small local or regional elections and are simply not big enough to get endorsements from say Obama or Trump and therefore endorsements for these candidates is rare. For the majority of our analysis we are not looking at specific endorsements and we should be able to remove these columns from our analysis entirely.

In our data preparation process, we initiated the gender information by cross-referencing the women's list from the Center for American Women and Politics with our dataset. As a dataset consisted of categorical data, a crucial step involved converting these categorical values into a binary format for compatibility with our chosen analytical methods. For instance, we transformed variables such as LGBTQ, Race, Veteran status, is_female, and Won Primary into binary representations (0 or 1). We applied one-hot encoding to the “District” variable. This conversion not only facilitates the application of causal inference methods but also enhances the interpretability of the data within the context of the analytical techniques chosen for our research questions.

Research Questions

Our first research question: How well does a region's partisan lean predict the level of support for the winning candidate?

Knowing the margin of victory for a certain candidate could be important in predicting for example, party dominance. If we’re able to predict the margin of victory, or probability of victory, before the primary occurs we could determine a number of political metrics such as the number of democratic house members after the general election. This predictive power could also be used to determine the amount of resources needed in the general election if, say the majority of candidates in a primary have a high margin of victory, less funding may be needed for those candidates.

To answer this question we need a model that’ll produce a continuous output because we’re predicting a variable with continuous outputs (from 0 to 100). This output restriction and the need for predictive power, interpretability and simplicity means we’ll be using a GLM. We will also use a Neural Network due to the complicated nature of feature interactions in this dataset especially surrounding some features impact on election outcomes.

The biggest problem in using a GLM is the simplicity of the model because without good features the GLM could create incorrect results. Also if some features aren’t included or too many features included it could overfit the data. Due to the limited size of our dataset, if the model requires a large number of features a GLM might not do very well due to the constraints of regression analysis and the problems with high dimensional regression.

The problem with the Neural Network t is the lack of transparency or interpretability. The benefactors of this model wouldn't be data scientists as well so an interpretable model is preffered which a Neural Network doesn't do well.

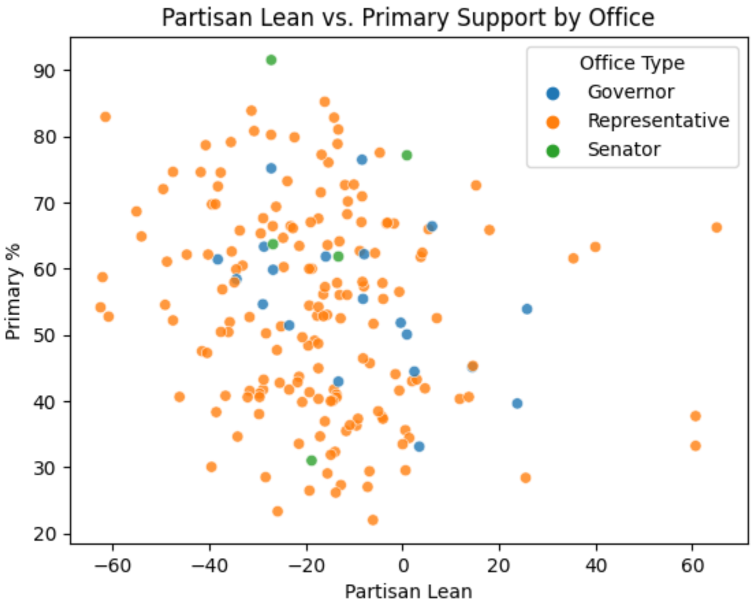
Our second research question: Does having a stem degree cause a candidate to win the primary election within the Democratic Party?

The answer to this question can inform political strategists about the potential impact of a STEM degree on the success of candidates in Democratic primary elections. This information could influence campaign strategies and candidate selection processes regarding the qualifications of political candidates.

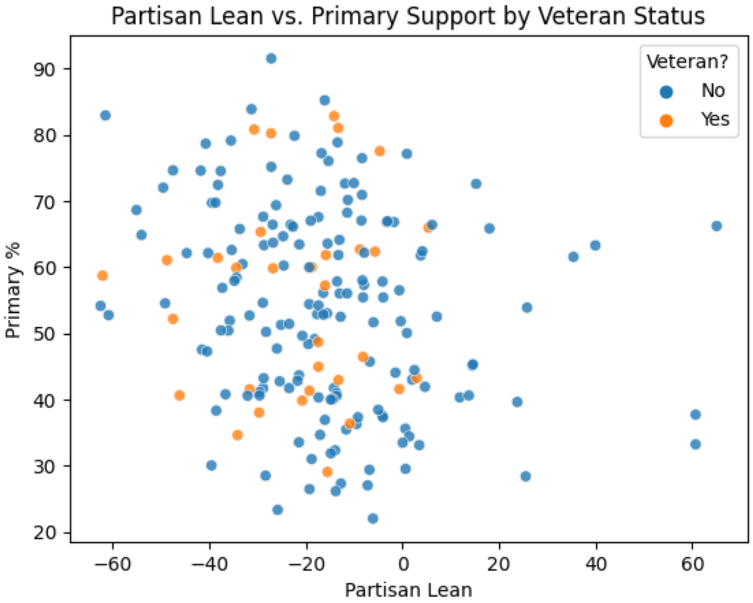
We use causal inference to answer this research question. It is a good fit as it allows us to explore the potential causal relationship between having a STEM degree and winning a Democratic primary election. Causal inference helps to understand whether changes in one variable are responsible for changes in another.

Causal inference is a powerful technique but there are some limitations such as selection bias, technique specification, unobserved confounding, and sensitivity to assumptions. Non-randomized studies may suffer from bias when individuals voluntarily choose to be part of specific treatment groups and these choices are influenced by characteristics that are not directly observed.

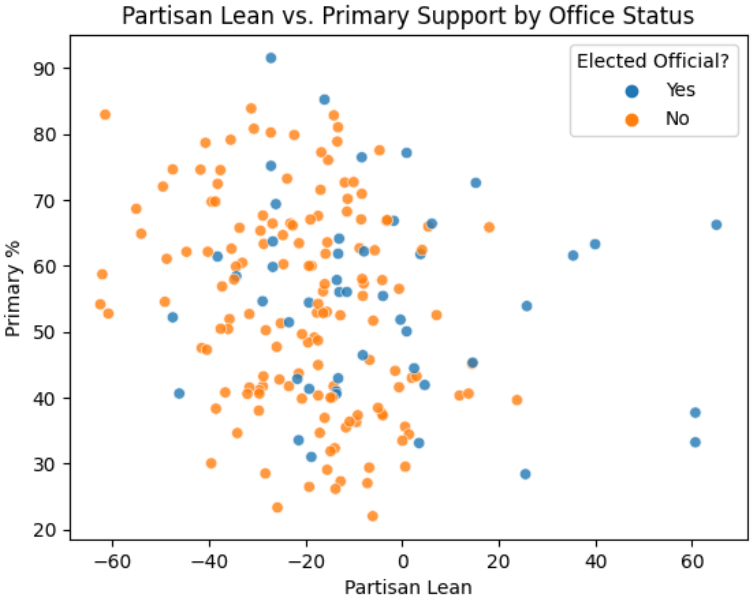
EDA



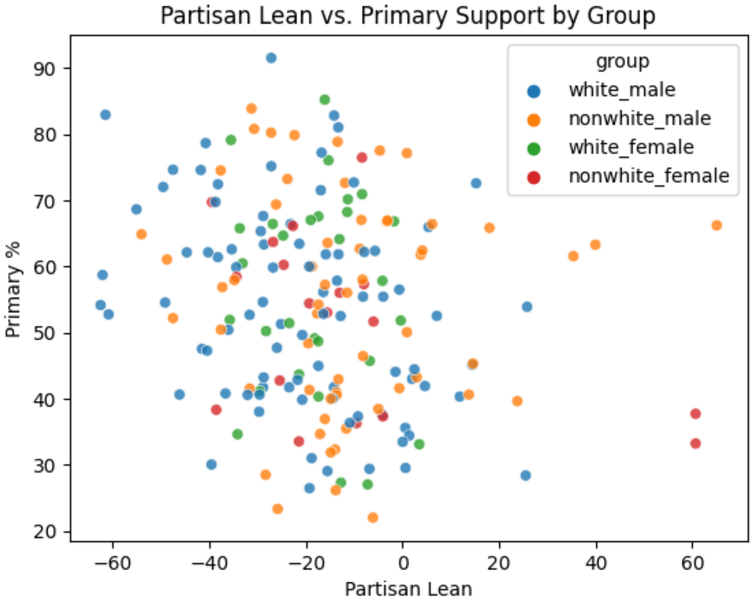
Shown above is the relationship between the regional Partisan Lean and the primary election support percentage of the winning candidate. Due to complications in the data set only Democratic candidates are shown. In the figure there exists a somewhat negative linear relationship between these variables indicating a relationship between a region's partisan lean and the winning candidate's primary support, which is the subject in question. This plot also highlights the various offices in the dataset and shows a similar relationship between all the offices in the dataset.



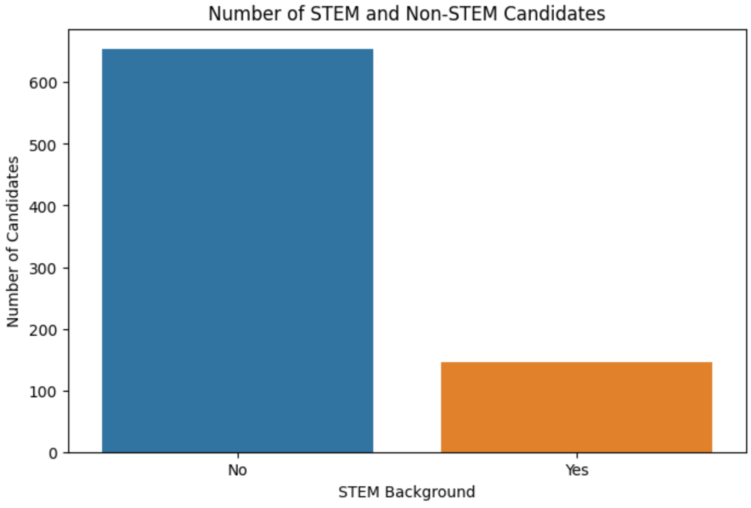
Shown in the figure above is the same plot as the previous one but instead plots which candidates are veterans. This plot tells us that being a veteran really does not have any discernible difference in either the level of primary support or its relationship with partisan lean and seems to be somewhat random in the plot. This shows us that being a veteran might not be a critical factor in our analysis and that in order to maintain the simplicity of the model setup we will not be including this feature in our analysis.



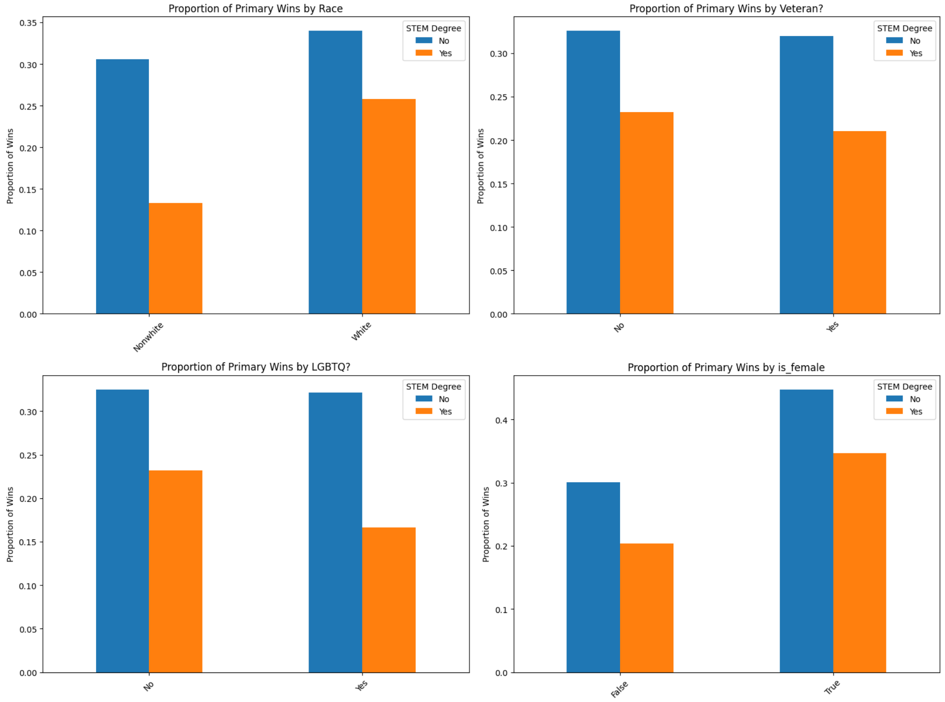
Above is the relationship between the two features in the prior plot but this plot highlights the difference that being an elected official makes in primary election outcomes. More specifically there seems to be a very different relationship for elected officials whereas private citizens show a much stronger negative relationship. This indicates that this is an incredibly important feature to include in our analysis. Shown below are two additional categorical features and their respective combinations.



This plot highlights not only the differences between the range of values each demographic group can take but also the differences between relationships for each group. This means that, first of all, each of these demographic groups have different election outcomes and this should be included in the analysis. Secondly because of the difference in relationships the prediction for a single point will be different for different groups and therefore the difference in relationships must be controlled for in the analysis.



We start with a simple visual of the proportion of candidates with and without a stem degree. There are significantly more candidates without a STEM degree (654) than with one (146). This indicates that candidates with STEM backgrounds are less common in this dataset, which could be reflective of broader trends in the educational background of political candidates.



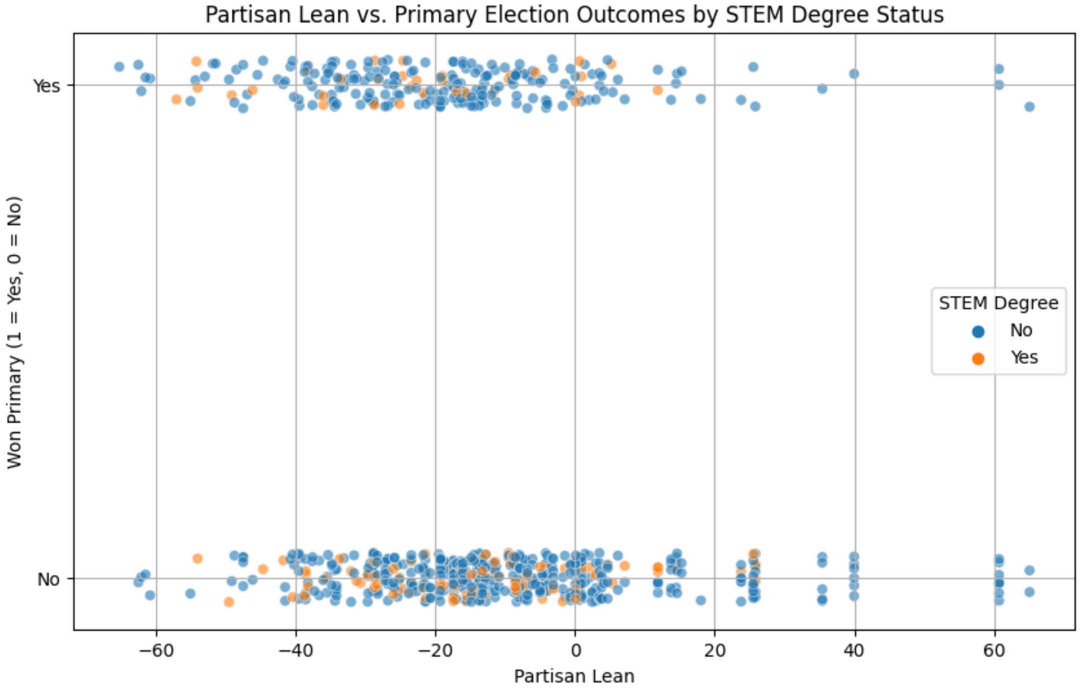
Veteran Status: Candidates with a veteran status who do not have a STEM degree have a higher proportion of primary wins compared to non-veterans with STEM degrees. This could suggest that veteran status may have a more substantial influence on primary election outcomes than having a STEM degree.

Gender: Female candidates with STEM degrees have a lower proportion of wins compared to male candidates with STEM degrees. This might indicate that gender plays a role in primary wins, potentially intersecting with the impact of a STEM background.

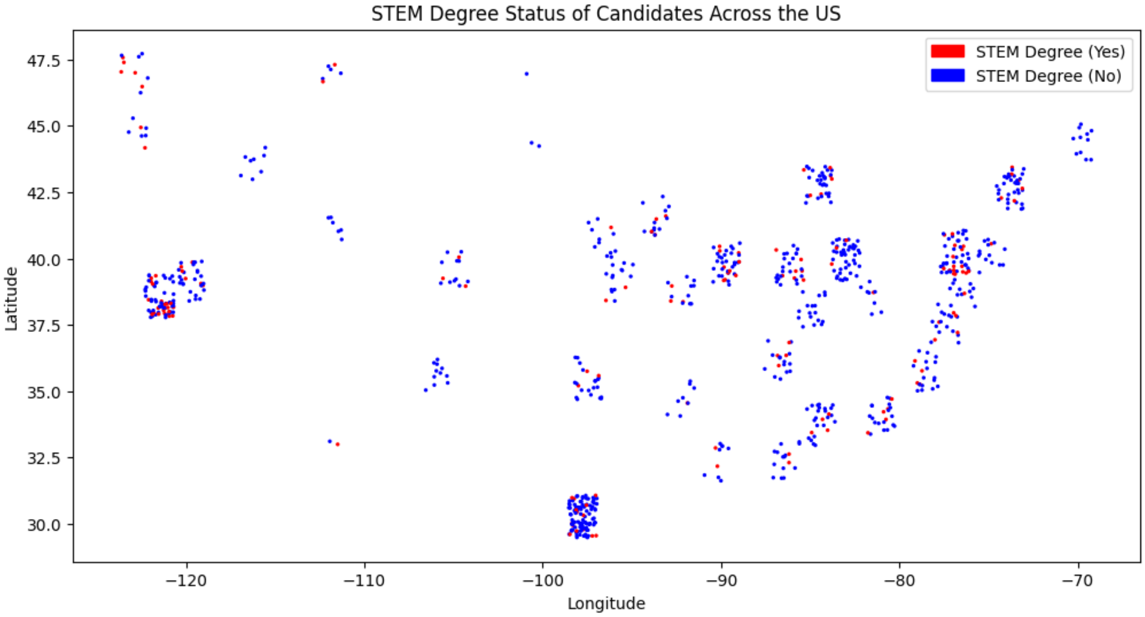
LGBTQ Status: Non-LGBTQ candidates with STEM degrees show a lower win rate compared to their LGBTQ counterparts with STEM degrees. This indicates that within the LGBTQ community, having a STEM degree may not significantly affect the chances of winning a primary.

Race: Nonwhite candidates with STEM degrees have a lower win rate compared to white candidates with STEM degrees. Race seems to be a significant factor, possibly more so than having a STEM degree.

These visualizations are relevant to the research question as they highlight the intersectionality of demographic factors with STEM qualifications in primary elections. They suggest that while having a STEM degree may influence primary outcomes, it is not the only determining factor; veteran status, gender, LGBTQ status, and race all appear to have an impact.



The scatter plot visualizes the relationship between partisan lean and primary election outcomes by STEM degree status. There's a visible trend where candidates without STEM degrees tend to win primaries across a wider range of partisan leans compared to those with STEM degrees. The visualization is relevant to the research question as it explores whether the partisan lean of a district interacts with the STEM background of candidates to affect primary outcomes. It suggests that while STEM status may have some impact, the partisan lean of the district does not show a distinct pattern that favors STEM or non-STEM candidates consistently.



The map visualization showcases the geographic spread of candidates with and without STEM degrees across the mainland United States. There is no immediately apparent geographic clustering of STEM candidates, suggesting that they are distributed throughout the country rather than concentrated in specific regions known for technology or science. This visualization underlines the research question's premise by displaying the geographic diversity of candidates' educational backgrounds. It suggests that STEM qualifications are not confined to specific areas, potentially widening the scope of analysis to consider local district factors that may influence the success of STEM candidates in primaries.

Methods Question 1

In our question we're predicting a candidate's primary support based on the candidate's regional partisan lean and other categorical features that pertain to the candidate. Each of these features is determined prior to the primary election and therefore can be used as features in our models with partisan lean being the main feature in addition to gender, race, and if they are in an elected office. As shown above gender, race, and being an elected official all highlight important aspects of a candidates' ability to gain primary support and therefore should be used in our models.

For our GLM we will be using a Linear Regression with several interaction terms and indicator coefficients for each of the categorical variables and their respective relationships with partisan lean and primary support. The regression formula is as follows:

$$Primary \approx \beta_0 + \beta_1 PL + \beta_2 F + \beta_3 (F * PL) + \beta_4 NW + \beta_5 (NW * PL) + \beta_6 EO + \beta_7 (EO * PL)$$

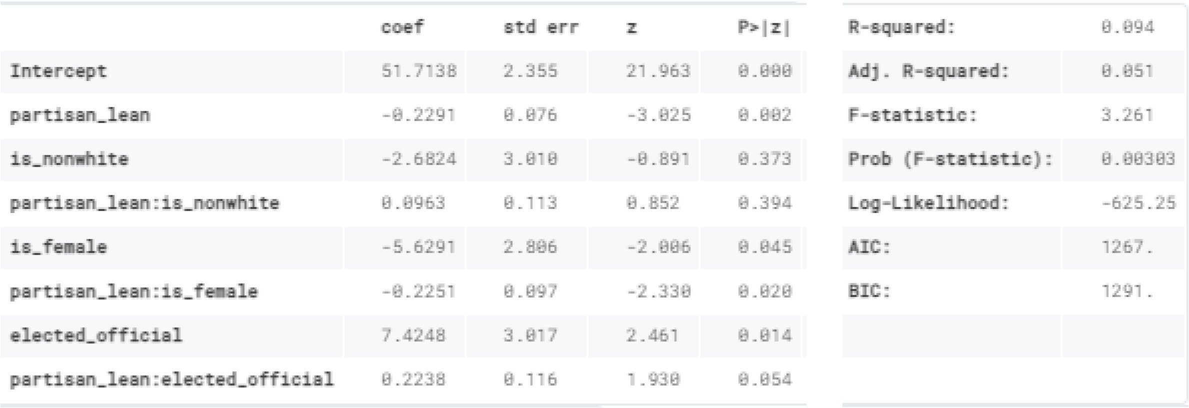
Where PL is Partisan Lean, F is an indicator for female candidates, NW is an indicator for nonwhite candidates, and EO is an indicator for candidates who are elected officials. This regression will not only be able to account for differences in the overall level of primary support which was shown in the EDA to be the case for each categorical variable. These differences will be captured by coefficients β_2, β_4, β_6. This formula will also account for differences between categorical variables at different levels highlighted by the interaction terms with coefficients β_3, $\beta_{5,}$ β_7. Using this model setup we will be assuming first of all that the other variables do not have any additional information or predictive power not already included in these features which is a somewhat strong assumption given the exploration done above. Another assumption this model makes is that there is no relationship between features and that they are independent of each other which is a much weaker assumption. Due to personal preference we decided to choose a frequentist approach because the dataset is large enough that a bayesian approach might be unnecessary.

The nonparametric method we're using is a Neural Network. We know Neural Networks can learn relevant features from the data during training which we thought would be important because there were a lot of features that we thought could have relationships with each other that would be important, specifically partisan lean and our demographic data (gender, race). Some assumptions we made were that all our features were relevant to the prediction, our input features relationship with the output would be complex and the architecture for the Neural Network would not cause overfitting to the data. The Neural Networks architecture has an input layer of our initial variables, a hidden layer of 64 nodes, and another hidden layer of 8 nodes with ReLU activation so that the model can sufficiently learn underlying patterns. We chose ReLU because it helps the model learn underlying patterns and ensures the output is non-negative for all our inputs because we can't have negative primary percentages for democrats that won an election.

We'll RMSE on a testing set to evaluate each model's performance which will yield us a comparable metric for both models given that they were trained and tested on the same data which is the case for our models. Using a test RMSE, whichever model has a lower value should be the better method due to the fact that it has effectively lower average error for the test dataset.

Results Question 1

Our Neural Network gave an RMSE of 17.88, and our GLM gave an RMSE of about 15.84. Giving a difference of 2.04. This means that for each model trained on the same training set, predictions for the winning democrats primary % on the same test set on average were higher by about 2.04% for our GLM. This level of error for each prediction on average shows there's a significant deviation from the predicted and actual values meaning our uncertainty is high for this model.



	coef	std err	z	P>|z|
Intercept	51.7138	2.355	21.963	0.000
partisan_lean	-0.2291	0.076	-3.025	0.002
is_nonwhite	-2.6824	3.010	-0.891	0.373
partisan_lean:is_nonwhite	0.0963	0.113	0.852	0.394
is_female	-5.6291	2.806	-2.006	0.045
partisan_lean:is_female	-0.2251	0.097	-2.330	0.020
elected_official	7.4248	3.017	2.461	0.014
partisan_lean:elected_official	0.2238	0.116	1.930	0.054

R-squared:	0.094
Adj. R-squared:	0.051
F-statistic:	3.261
Prob (F-statistic):	0.00303
Log-Likelihood:	-625.25
AIC:	1267.
BIC:	1291.

Shown above is the regression output for the GLM. This illustrates several uncertainties this model contains specifically regarding the significance of certain coefficients and how the model fits the data overall. Firstly three of the coefficients are not significant, or have insignificant p-values, at a 95% significance level meaning that the value calculated has a 95% confidence interval which contains zero or in other words the value is not significant from zero. In addition to this, the R-squared and Adj. R-squared are very very low which indicates that this model setup is not a good fit for the data. That being said, the majority of coefficients yielded significant results indicating that there appears to be some sort of predictive power for these features and that there are some interpretable results that could answer our question

Discussion Question 1

Our models gave an RMSE of about 15% for the GLM and 17% for the Neural Network this is important because in context our primary winning percentage can only be between 0 and 100 (it's a percentage). Therefore we believe the uncertainty in our data to be qualitatively high because of this relatively high error in our predictions and our scatterplots not showing a clear and consistent trend within the data.

The Neural Network struggles with interpretability because of the architectures complexity. This also means the model might be overfitting. The GLM likely doesn't represent all the possible relationships between variables meaning, and due to the constraints of high dimensional regression adding these relationships would likely detract from the accuracy and interpretability of the model.

Additional useful data would be more demographic data (ethnicity, age, education), incumbents, regional data (considering partisan lean represents that region more than it represents the candidate).

Methods Question 2

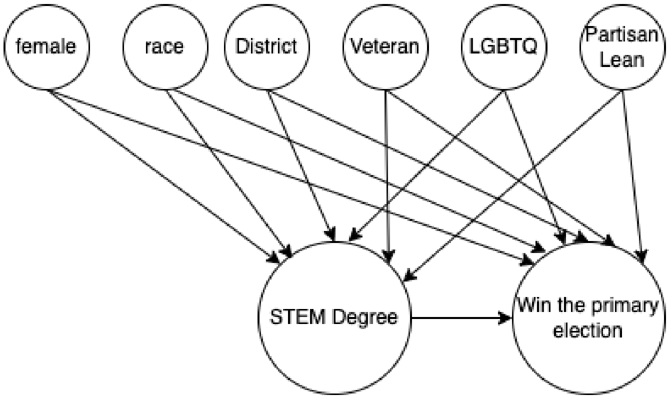
We are trying to define whether having a STEM degree will affect the chance that people will win the primary election within the Democratic Party. Hence, in this study, the treatment variable is the possession of a STEM degree, and the outcome variable is the success of individuals in winning the election.

Based on the domain expertise, the potential confounders we think may correlated with treatment and outcome are Demographic Factors, Geographic Location, and financial standing. Demographic Factors, including age, gender, ethnicity, and other demographic variables may be associated with having a STEM degree and also influence electoral outcomes. Geographic location is considered, as the region a candidate is running in could impact their chances of success and may be associated with STEM education, with states like California having the highest proportion of students pursuing STEM degree. Financial standing plays a role in supporting individuals pursuing STEM degrees, particularly when they come from families with high financial status. The backing of family resources can potentially contribute to electoral success by providing financial support during election campaigns.

Due to the limitation of the dataset, we are able to collect and include confounders information regarding the gender, race, Veteran, LGBTQ, district, Partisan Lean of each candidate.

In this observational study, we assume the unconfoundedness and employ the Inverse Propensity Score weighting to adjust confounders. The methodology involves using Logistic regression to estimate the propensity scores, which represent the probability of having a STEM degree given the observed confounding variables. Subsequently, inverse propensity score weighting is applied to assign weights to each observation based on the inverse of its propensity score. IPW gives more weight to individuals whose observed characteristics are less common among individuals with the same propensity for having a STEM degree. This approach aims to enhance the robustness of our analysis and mitigate potential biases arising from unobserved confounding variables.

For the outcome regression, we conducted an Ordinary Least Squares (OLS) regression analysis to assess the relationship between a candidate's possession of a STEM degree and their success in primary elections. The model was constructed to predict the binary outcome of winning a primary election based on the presence of a STEM degree, while controlling for a set of confounding variables as mentioned above. Additionally, to account for the potential influence of geographical variables, we initially included dummy variables for each district. However, due to concerns about data sparsity and multicollinearity—which could compromise the model's interpretability and stability—we subsequently examined these variables through correlation analyses and considered alternative modeling strategies such as grouping districts, regularization, and hierarchical modeling. The final model choice was made to balance the need for a parsimonious model against the theoretical importance of district-level effects. This approach aimed to provide a clear and robust estimation of the effect of having a STEM degree on election outcomes while accounting for a comprehensive set of relevant covariates.



Casual DAG

Results Question 2

Our analysis aimed to estimate the causal effect of holding a STEM degree on the likelihood of winning a primary election within the Democratic Party. We employed Inverse Propensity Score Weighting (IPW) to address confounding variables. The estimated propensity scores ranged mostly between 0.1 and 0.2 after excluding state and district as confounders due to overfitting concerns. Our initial IPW estimate suggests that having a STEM degree is associated with an 11 percentage point decrease in the probability of winning the primary. This estimate points to a negative causal relationship between having a STEM degree and primary election success within the scope of the observed confounders.

It is important to note that this result is subject to our assumption of unconfoundedness, which holds that all the confounders influencing both the treatment and the outcome have been accounted for in our model. Given the limitations of our current dataset, we have included gender, race, veteran status, LGBTQ status, and partisan lean as confounders. Geographic location was initially considered but led to model overfitting and hence was excluded from this analysis. We acknowledge that important financial variables such as individual financial resources are not included and could be an important confounder. Their absence poses a limitation to the current analysis and might be addressed before our final submission.

The results are tentative and may be updated with the inclusion of additional data. The effect size of -0.11 is substantial, considering the scale of election outcomes. However, the uncertainty surrounding our estimate remains high due to potential unobserved confounding. Therefore, we interpret these findings cautiously, recognizing the need for further data collection and analysis to confirm these preliminary conclusions.

Our outcome regression analysis, designed to complement the Inverse Propensity Score Weighting (IPW) approach, revealed nuanced insights into the effect of having a STEM degree on winning a primary election within the Democratic Party. The model including district variables indicated a negative impact of holding a STEM degree on primary election success, with the 'STEM Numeric' variable having a coefficient of -0.1467 (p-value 0.017). In contrast, the simpler model, excluding district variables, showed that the 'STEM Numeric' coefficient was -0.0391 (p-value 0.408), suggesting no statistically significant impact.

This discrepancy between the two models underlines the complexity inherent in modeling electoral outcomes. The R-squared value of the model without district-level variables was 0.264, suggesting that it explains approximately 26.4% of the variance in the outcome. The gap between R-squared and adjusted R-squared (0.258) was smaller in this simpler model, indicating a reduced risk of overfitting compared to the model with district variables.

In terms of other significant predictors, 'Race_bin' emerged as a strong predictor with a coefficient of 0.2674 (p-value close to zero). Similarly, 'Veteran_bin' (coefficient 0.1109, p-value 0.032) and 'female_bin' (coefficient 0.2244, p-value < 0.001) were also statistically significant.

In conclusion, our analysis suggests that while demographic factors like race and gender show consistent and significant associations with primary election success, the impact of having a STEM degree is less clear. The substantial difference in the STEM coefficient between the two models – from a significant negative impact to no significant effect – raises critical questions about the robustness of these findings and calls for further investigation. The exclusion of district variables simplifies the model and reduces multicollinearity risk but also lessens the explanatory power.

Discussion Question 2

Our investigation into the causal relationship between holding a STEM degree and the likelihood of winning a Democratic primary election has highlighted several methodological challenges and directions for future research. One key limitation in our approach was the reliance on the assumption of unconfoundedness, particularly crucial in the Inverse Propensity Score Weighting (IPW) method. This assumption presumes that all significant variables impacting both the possession of a STEM degree and election success are included in our model. However, our dataset lacked some potentially influential variables, leading to concerns about omitted variable bias.

The complexity of our initial model, which included district variables, contrasted with a simpler model that omitted these variables, revealed critical disparities in our findings. The more complex model indicated a significant negative impact of a STEM degree on election success, a result not replicated in the simpler model. This discrepancy raises questions about potential overfitting in the complex model, as suggested by the significant difference between the R-squared and adjusted R-squared values. Conversely, the simpler model, while potentially more robust against overfitting, had reduced explanatory power and did not show a significant impact of the STEM variable. This variance underscores the challenges in balancing model complexity with explanatory depth.

Our model diagnostics indicated issues with the normality of residuals, affecting the reliability of standard errors and p-values. This necessitates a cautious interpretation of our results.

Looking ahead, enhancing our model with additional data and refining our methodologies could offer more insightful results. For instance, integrating more comprehensive demographic data, including age and educational background, could help control for unobserved confounding. Additionally, addressing the sparsity of the district data by grouping districts could reduce the risk of multicollinearity and provide a clearer understanding of regional differences. These modifications could potentially reconcile the differences observed between our two models and offer a more nuanced view of the impact of a STEM degree on electoral success.

Given these considerations, our confidence in establishing a definitive causal relationship between possessing a STEM degree and winning a primary election is tentative. The variation in the STEM variable's effect size between the two models – from significant to non-significant – suggests the need for further research. Our current findings should be viewed as preliminary, pending further investigation with more comprehensive data and refined modeling approaches.

Conclusions Question 1

Overall, we found that there was a weak relationship between partisan lean and the winning democratic primary percentage. In addition to this it's possible that our results are not very generalizable because we only look at data for one year in specific and the models are not very accurate. Our findings are pretty narrow due to the fact that we only had proper data for primary candidates that won and are democratic which for a single year is not enough data points to show any clear relationships. One factor in the lack of accuracy that we couldn’t account for was the intangible qualities that each politician had like charisma and agreeableness. Other factors that we couldn’t account for were voter demographic data, the year of the election which could have implications during a presidential election, and the candidate’s specific public policies .

If however, future models are perfected, or better data is collected, political parties can use these predictions to make decisions based on how they want to allocate their resources. If they can predict that a certain candidate will win their primary they can better allocate their resources to give support to other candidates. We also believe this can help candidates and help them structure or restructure their campaigns. Future research could also include exploring how the year affects the winning candidates' primary percentage. For example, looking at the years in which a presidential election or senator elections took place in comparison to years where there wasn’t any may be useful. Also looking at how the state or district of where you are affects the predictive power can also be useful in addition to the extremity of the winning candidates' views could also be helpful in predicting their winning support percentage.

Conclusions Question 2

The estimated Inverse Propensity Score Weighting (IPW) and the outcome regression for the impact of having a STEM degree on winning the primary election within the Democratic Party is negative. This negative estimate suggests that, on average, individuals with a STEM degree may have a lower likelihood of winning the primary election compared to those without a STEM degree. Therefore, the policy maker and election strategists should consider delving deeper into the role of STEM background in political success. Additionally, efforts to encourage/decourage individuals with STEM expertise to participate in politics could be explored, potentially fostering a more diverse and informed political landscape.

The generalizability of the results is limited by several factors, primarily the narrow scope of the dataset and the assumptions underlying the causal inference methodology. The analysis relies on data from the 2018 Democratic primary elections, restricting the findings to that specific election cycle. Moreover, the use of unconfoundedness as an assumption in the causal inference model introduces a potential source of uncertainty. Therefore, the findings are more narrow than broad. To enhance generalizability, future studies could incorporate data from multiple election cycles, including a broader set of confounding variables and employ robust sensitivity analyses to assess the impact of unobserved confounders.

While our analysis draws upon a combined dataset from various sources, it’s important to note that certain information gaps still persist as the scarcity of information on potential confounders, posing challenges for our analysis. To build upon our current study, future research should aim to address these information gaps by either refining data collection methods or incorporating additional sources. Besides, conducting a study to analyze election outcomes over multiple years or investigate the campaign strategies employed by candidates with STEM backgrounds could provide valuable insights.

Learning point

We learned about how to develop research questions around topics we are interested in, refactoring the question depending on what we have available to us. We also learned a lot about doing preprocessing where we cleaned the data and how to search for data online and merge it with our datasets. Also, we learned how to collaborate within a team and use deepnote to work together on data science projects.