Characteristics of Political Endorsements In 2018 Primary Races

Stehn, Jack
jack.stehn@berkeley.edu

Sidlo, Eva
esidlo@berkeley.edu

Kennedy, DK
dpk2010@berkeley.edu

Quiterio, Ashley
ashleyquiterio@berkeley.edu

May 10, 2021

1 Introduction

Political endorsements are ways for politicians and political coalitions to demonstrate their support for candidates. An endorsement from a more established or famous politician or coalition can influence susceptible voters. In 2018, many seats were up for election. We wanted to find out which endorsements are significantly correlated with primary victories and what properties of a race could prompt an endorser to make an endorsement.

2 Data Overview

We used the provided FiveThirtyEight endorsements data. [3] This data set represents a census of all candidates (both democratic and republican in 2018) that ran in their party's primaries for Senate, House and Governor positions excluding races featuring an incumbent. All incumbent races are systematically excluded from this data set, since those seats are not considered open. FiveThirtyEight collected this data by searching candidate websites and reviewing news reports. The information that was collected included each candidate's endorsers and office type. The democratic candidate data had additional features such as candidate race and gender as well as the partisan lean of their state or district (described more below). This information was not included for republican candidates. Each candidate makes up a row in the data for a total of 1585 candidates. There were no significant concerns with selection bias or convenience sampling because this data was a census.

To help understand the importance of partisanship, we brought in 2018 partisan lean data for each state and district from FiveThirtyEight. Partisan lean is "calculated by finding the average difference between how a state or district voted in the past two presidential elections and how the country voted overall, with 2016 results weighted 75 percent and 2012 results weighted 25 percent." [3]

We also collected additional demographic data about each state and congressional district where an election took place in 2018. We queried the American Community Survey (ACS) 1-year estimates from 2018 for state and congressional district demographic data. This includes variables for total population, race, aggregate household income, age, and poverty ratios. This data was collected using Cenpy. [1] [2] The ACS is a sample of the population that estimates a census through an annual survey and is considered to be the best publicly available demographic data at the state and district level, despite the following concerns. The ACS has large margins of errors for highly granular data. This is due to a lack of resources in nationwide surveying for smaller regions. [5] The census and ACS are also prone to not representing individuals without stable housing and other historically marginalized populations. At the state and congressional district granularity we believe the error margins are low enough to use in our analysis.

Finally, we brought in Ballotpedia's list of 2018 battleground states and congressional districts. [6] This data is a strictly qualitative analysis done through Ballotpedia rather than a sampling or a census. Battleground (or swing) locations are those in which there is closely divided support for the major political parties.

We maintained two datasets for this project. The first is the original candidate endorsement data from FiveThirtyEight indexed by individual candidates. Second, we joined the aforementioned data on locations into a table indexed by each race in a state or district with columns corresponding to various demographic features. These two tables allowed us to investigate endorsements at two different scales: individual and regional. Fortunately, we acquired complete data for every state and congressional district for every category so there were no null values.

3 Exploratory Data Analysis (EDA)

First, we wished to understand the relationship between a candidate receiving an endorsement from a political figure and the percentage of the votes that they had achieved in their primary.

The distribution of primary percentage differs by the endorsements each candidate receives as seen in figure 1. Of the 811 candidates in democratic races, those supported by Biden and/or Warren received median primary percentages over 50%, while those endorsed by Justice Democrats or Our Revolution received percentages under 36%. This parallels the differences observed in the republican distribution of primary percentages across various endorsers. In figure 1, candidates supported by Trump had a median percentage over 58%, while those endorsed by Right to Life and Main Street were below 38%. Despite these differences in percentages, the plots do not tell the whole story.

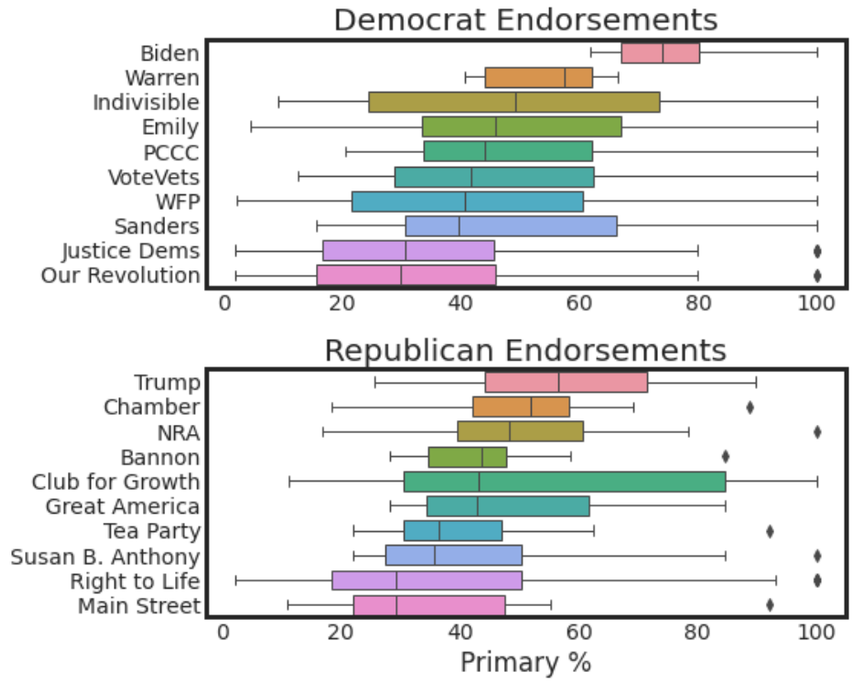

Figure 1: Boxplot summaries of distributions of percent of primary vote achieved for candidates endorsed by specific endorsers.

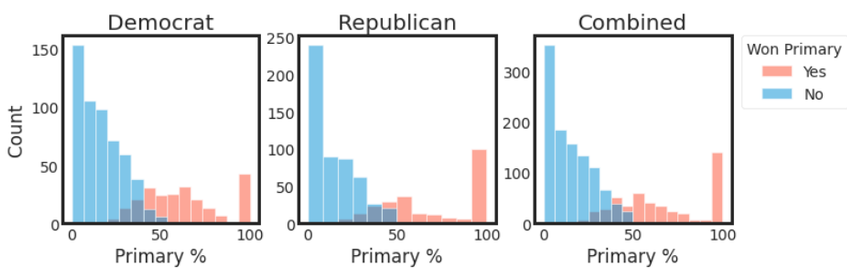

Figure 2: Distributions of percent of primary vote achieved by winning and losing candidates separated by political party.

The primary percent indicates the percent of votes that a candidate received in the first round of primaries. We can see in figure 2 that there is a range of primary percentages that lead to winning a primary. We plotted the distribution of primary percentages shaded by primary outcomes. In the figure, there is no clear difference by political party. It seems that those who won a primary had a range from 20% to 100% in the primary race, while those who lost had a range from 0% to 50%. There is much overlap in these ranges which means this variable is not a clear indicator of race outcomes. Ultimately, this means that win-percent alone is not a productive feature in analyzing endorsement efficacy.

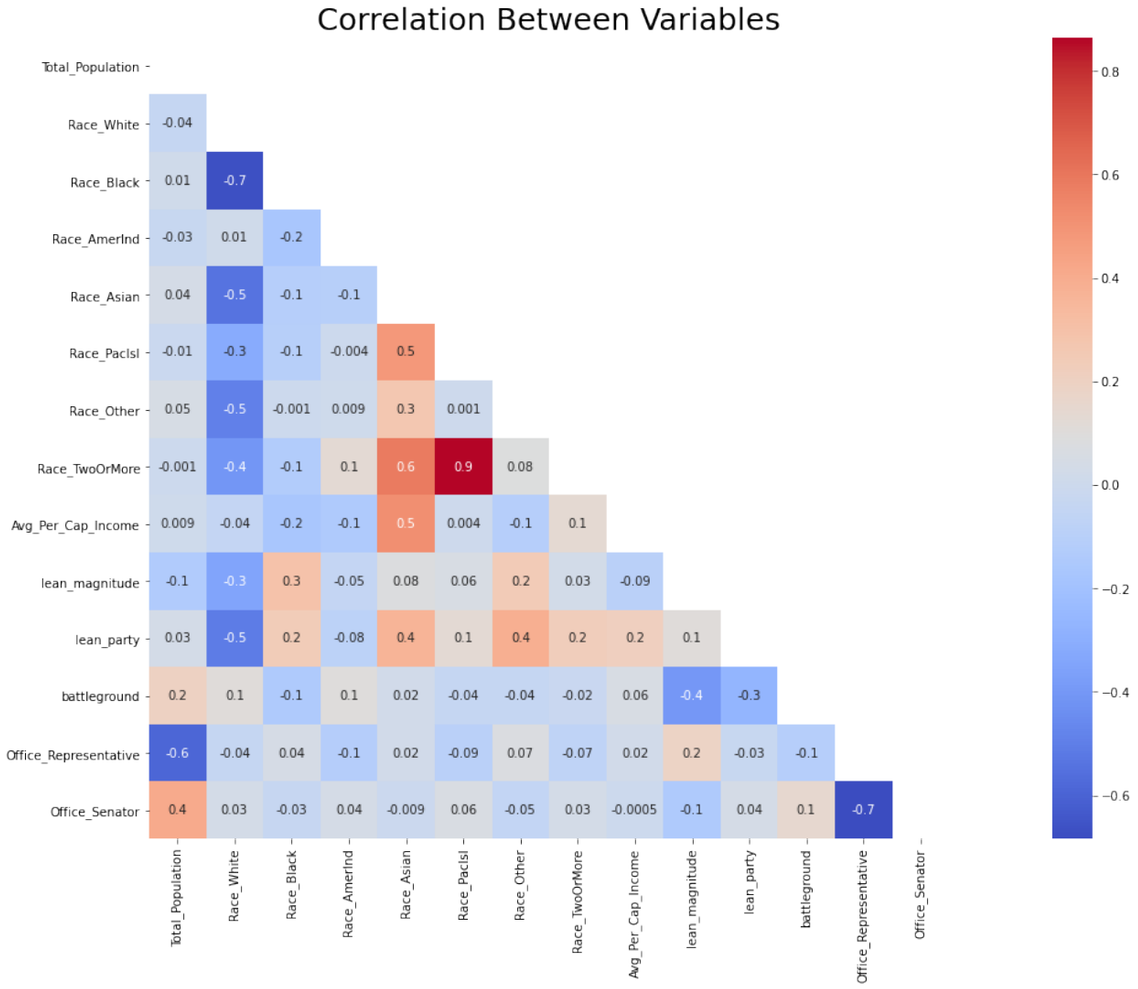

Figure 3: Heatmap of correlations between variables

In figure 3, we show the correlation heatmap between variables in the 2018 races data set. This guided our feature selection and also raised additional interesting questions about the election data. The percentage of respondents who labelled themselves as two or more races correlates strongly with those who labelled themselves as Pacific Islanders. Therefore we removed the two or more races feature from our model. We also see strong negative correlation between percentage of white and black people in a region. This prompted interesting questions about the relationships between racial groups in the United States that we ultimately did not pursue for analysis in this project. Finally, there is a strong negative correlation between senator races and representative races – this makes sense as they are mutually exclusive events.

4 Research Questions

With the endorsements data, we immediately identified some interesting patterns that we wanted to investigate further. We focused on two large questions about endorsement patterns.

4.1 Question 1

Trump has been described repeatedly in the media as a "king-maker"; a person who brings leaders to power through the exercise of political influence. The NRA's approval is seen as a necessity in many parts of the country if you wish to win a primary. Biden and Bernie Sanders threw their weight around in 2018 by endorsing primary candidates in an effort to shape this increasingly polarized nightmare of United States politics.

This leads us to question the outcomes of primary races given candidate endorsements. *Are there any significant differences in outcomes between a candidate who receives an endorsement from a specific political entity and the general population of primary candidates?* We looked to Multiple Hypothesis Testing to answer this question.

There are 20 different endorsers in our dataset. If we were to naively perform tests with 95% confidence for each endorser, there is a high probability that we could get false positives and we wish to control for that

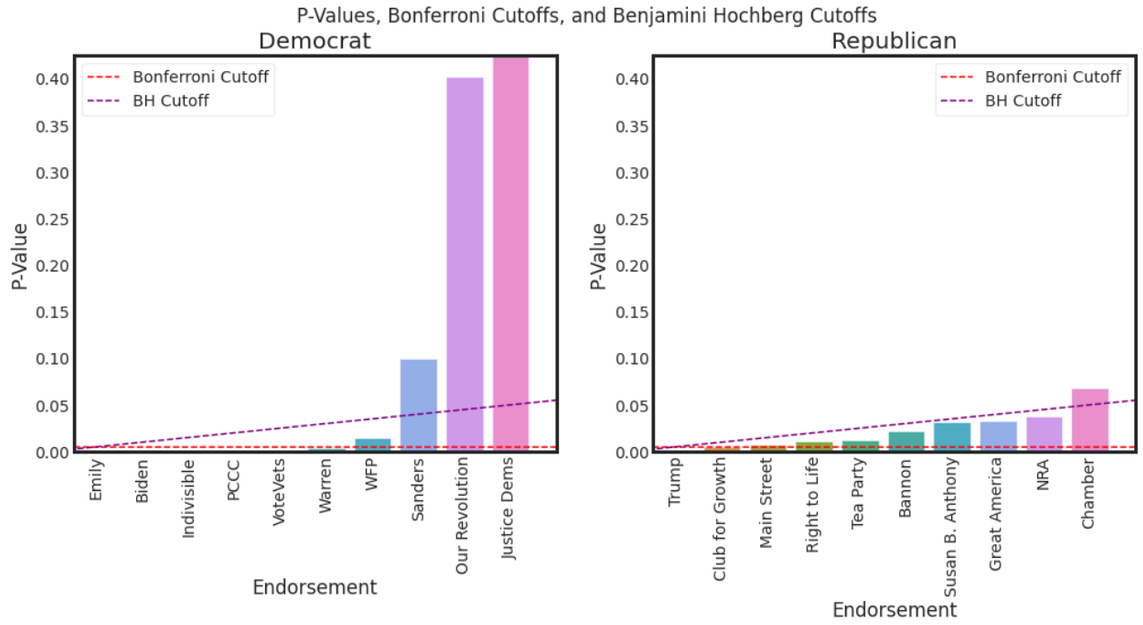

Figure 4: P-Values for each test. The cutoff for the Bonferroni correction is shown by a red line and the cutoff for the Benjamini-Hochberg Process is shown with a purple line.

rate. The results can be used to see which endorsers and the ideas they represent are associated with victory in races. It can be used to alter strategies in these elections. Also, answering this question could benefit the actual endorsers by validating their endorsement strategy.

4.2 Question 2

When looking at these endorsements, we were struck by how few endorsements were given out. There were thirty-five seats in the United States Senate and all 435 seats in the United States House of Representatives up for election in 2018. However major actors like Biden, Trump, Warren, and Sanders only endorsed a handful of candidates (as few as 5) during the primary. *What sort of features of a district and election explain whether or not these endorsers get involved in a race?* We looked to a Bayesian generalized linear model (GLM) to identify the strength of various demographic features in association with endorsement outcomes.

4.3 Technique 1: Multiple Hypothesis Testing

To address Question 1 we wanted to see which endorsements were important and indicated a higher probability of winning. For each endorser, the null hypotheses is that endorsed candidates had a win-rate distribution identical to that of the general population. Each null hypothesis says that the probability of winning is independent of the endorsement of each endorser.

4.4 Methods

The p-values for each test were calculated via permutation testing. For each endorser, the variables indicating endorsement from a particular endorser were shuffled and the win-rate was calculated. This process was performed 1,000 times for each endorser. The resulting distribution of win-rates were then compared to the observed win-rate. The p-value is the percentage of this bootstrapped sample that had the same or higher win-rate than the observed win-rate. A visualization of each test is shown in the Appendix Figures A1 and A2.

In one group of multiple hypothesis tests, we used the Bonferroni correction to control the Family Wise Error Rate (FWER). Next, we used the Benjamini-Hochberg Procedure (BH) to control the False Discovery Rate (FDR). For each test, we used a significance level of $\alpha = 0.95$.

4.5 Results

The results of each test and cutoffs for each correction method are shown in Figure 4. For both democrats and republicans, there are p-values that are significant after the Bonferroni correction. The FWER is controlled such that the probability of any rejection, under the null hypothesis, is less than or equal to 5%. This indicates that the results are significant and that the claim that the win-rate of people with endorsements is the same as the win-rate of people without endorsements should be rejected.

The BH procedure controls the FDR, meaning that of all the endorsements that are declared significant we expect 5% or fewer of them to be falsely declared significant. A majority of these endorsements are shown to be significant. We rejected 7 of the 10 null hypotheses for the democrats and 9 of the 10 null hypotheses for the republicans.

For the democrats, candidates with endorsements from Emily's List, Joe Biden, Indivisible, Progressive Change Campaign Committee (PCCC), VoteVets, Elisabeth Warren, and Working Families Party (WFP)

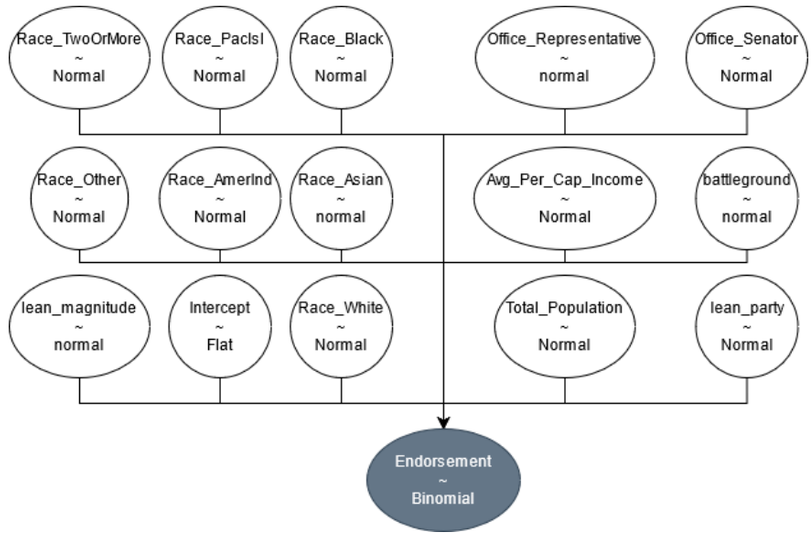

Figure 5: Graphical Model of Bayesian GLM. White indicates a hidden variables (coefficients on features) and gray indicates known value (results).

have a significantly higher win-rate than the general candidate population. For republicans, candidates with endorsements from Trump, Club for Growth, the Tea Party, Republican Main Street Partnership, National Right to Life Committee, Susan B. Anthony List, National Rifle Association (NRA), and Great America Alliance have a significantly higher win-rate than the general candidate population.

4.6 Discussion

After using the Bonferroni correction, endorsements from Warren, Working Families Party, Tea Party, Club for Growth, Right to Life, Bannon, Susan B. Anthony, Great America, and the NRA were all no longer rejected. Ultimately, this is unimportant for this specific error-rate control because we are controlling the FWER. A rejection under Bonferroni indicates to us that there was a maximum 5% probability that there would be a rejection if the null hypothesis were true for all of the tests. This tells us that endorsements do have some significance when determining the outcome of a primary.

In the BH procedure, all the hypotheses that would have been rejected without the correction were also rejected with the correction. This helps us identify which specific endorsements are important with a low number of false discoveries.

These decisions can help us understand a bit more about the candidates, their probability of winning, and the significance of an endorsement from a particular group or person. When looking further into certain single-issue endorsers, like Emily's List or the NRA, we can see and make decisions about the importance of these issues to voters in the primaries.

These tests do not make any claims of causality and should not be viewed as such. Instead, they indicate association between win-rate and endorsements. More data would allow for rigorous causal inference.

5 Technique 2: Bayesian Inference

We looked to Bayesian Inference to understand which races endorsers felt compelled to make an endorsement in. We established a graphical model (shown in figure 5) to explain what factors we initially believe are relevant to whether or not an endorser gets involved in an election. Our hypothesis is, for these endorsers, the decision to get involved in election is based on the properties of the race and governmental seat of power rather than belief in a specific candidate.

5.1 Methods

First, we established the graphical model shown in figure 5. Here, we state that the probability that an endorser gets involved in an election is based on total population, racial makeup, household income, partisanship of the region, status as a battleground, and the office that is up for election. Although more features could certainly be added, the scale of data wrangling and computation required to do so was unreasonable for this project. Therefore we brainstormed the top features we believe weigh heavily in an endorsement decision.

Once we established our graphical model we implemented four GLM's predicting the involvement of four different endorsers: Biden, Trump, Our Revolution, and the Tea Party. Biden and Trump represent key political figures while Our Revolution and the Tea Party represent political coalitions with complex missions and significant win-rates (as established in Method 1). Each GLM is a logistic regression model that attempts to find the probability that a potential endorser endorses any candidate in the race based on the properties of that area. This assumes that the probability that an endorser gets involved in a race can be properly be modeled as a Bernoulli random variable that is dependent on all these other variables.

Mathematically, the model makes the following assertions:

Let p be the probability any candidate in a particular race is endorsed by some political entity x (i.e. Biden, Trump, Tea Party, etc.). β_{xi} indicates the ith coefficient in the logistic model for candidate x.

Feature	Biden Model	Trump Model	Our Revolution Model	Tea Party Model
Intercept	No	No	No	No
Total_Population	**Yes**	No	**Yes**	**Yes**
Race_White	No	No	No	No
Race_Black	No	No	No	No
Race_AmerInd	No	No	No	No
Race_Asian	No	No	No	No
Race_PacIsl	No	No	No	No
Race_Other	No	No	No	No
Avg_Per_Cap_Income	No	No	No	No
lean_magnitude	**Yes**	No	**Yes**	No
lean_party	No	No	**Yes**	**Yes**
battleground	**Yes**	**Yes**	No	No
Office_Representative	**Yes**	**Yes**	**Yes**	No
Office_Senator	No	No	No	**Yes**

Table 1: Table displaying whether or not each coefficient is significant for each model. A coefficient is considered significant if its 94 percent interval of highest density does not contain 0. See Appendix for numerical values of each coefficient in each model.

$$\log\left(\frac{p}{1-p}\right) = \beta_{x0} + \beta_{x1}\text{Total\_Population} + ... + \beta_{x13}\text{Office\_Representative} + \beta_{x14}\text{Office\_Senator} \quad (1)$$

Logistic regression allows us to examine and compare the weights of each feature and their association to the log-likelihood of receiving an endorsement. In the Bayesian setup, the model coefficients are unknown while the data are fixed.

To prepare our data for fitting in a GLM we performed standard processing steps such as one-hot encoding categorical variables as well as converting raw population numbers to proportions. This latter transformation allowed us to control for the disparity in population between regions. We also broke up the partisan lean feature into two measurements: one for which party the lean is toward and one for the magnitude of the lean (a continuous numeric value between 0 and 1). We then normalized all data to fall between 0 and 1 using min-max normalization. Normalizing the data allows us to directly compare the magnitudes of the coefficients of the fitted model.

Although we considered specifying non-default priors on the coefficients, we did not have a strong enough belief in the distribution of coefficients to override PyMC3's default parameters. We fit the four models and adjusted parameters such as number of samples, number of chains, acceptance rate, and tune to ensure a reasonable trade-off between convergence and computing time. [4] Each model took 35 minutes to 2.5 hours to fit.

5.2 Results

The purpose of our models is not so much to make predictions but rather to explain which features of a race are better associated with an endorsement. What characteristics of a state or district warrant an endorsement from Trump, Biden, Our Revolution, and the Tea Party? Table 1 shows significance of each feature we examined for each model. The statistics of the posterior distributions for these coefficients are shown in detail for each model in the appendix in tables 2, 3, 4, and 5. A coefficient is considered significant if its 94% interval of highest density does not contain 0. These significant coefficients have a stronger association with endorsement outcomes in the states and districts of the 2018 races.

The consistently significant coefficients included:

1. Population of a state or district - Higher population correlated to higher probability of endorsement (except for Biden).
2. Degree of partisanship (lean magnitude) - Higher partisanship correlated to lower probability of endorsement.
3. Office - The office that is being pursued has a significant effect on whether they receive an endorsement.

Notably, there is a strong correlation between battleground elections and endorsements from figures like Trump or Biden, but not organizations like Our Revolution or Tea Party.

5.3 Discussion

These four features – the population of a state or district, its partisan lean, its status as a swing state or district, and the type of office – make a lot of sense as important factors an endorser would consider before committing to an endorsement. Population determines representation in the House of Representatives and, in general, more populous areas have a lot of political visibility. While all other models had strong positive correlation between population, Biden's had an extremely negative correlation. A location's partisan lean and its status as a battleground election set the stakes of a particular election and the location's importance to either major political party. Finally, the type of office (Senator or Governor at the state level, or Representative at the district level) also signals the importance of a race to endorsers.

Interestingly, these four features were common to almost all four endorsers regardless of ideology, status as an individual politician, or coalition. This suggests that there may be widespread agreement about which features of a race make it worth endorsing. Fitting GLMs on the remaining endorsers in the dataset would confirm this.

Surprisingly, race and income – highly polarizing topics in U.S. politics – were not significantly associated with whether a location received an endorsement from the four endorsers we selected. The incredibly wide-ranging distributions of the posteriors of these coefficients suggests that they may be important in some elections but not others, or that they associate positively with an endorsement in some locations, but negatively in others.

We originally created a multi-level Bayesian model that attempted to first predict the probability that an endorser would endorse anybody in a race and then tried to use the personal features of each candidate to create probability distributions that each individual candidate would be endorsed. This process proved to be conceptually interesting but computationally expensive so we simplified our model.

We also wished to include other population features such as religion, age distributions, poverty rates, major industries, and other explanatory variables, however such features were often unwieldy and difficult to obtain and would have increased run-time. These would have been added to the logistic regression similarly to the other variables in the set.

We also wished to include individual data on the candidates such as platforms, beliefs, income, primary campaign contributions, and other variables. This would have been useful for the second stage of the hierarchical model mentioned previously.

6 Conclusion and Future Work

We asked the following questions: *Are endorsements associated with winning candidates? What are the characteristics of races in which endorsers choose to make endorsements?*

The results of our multiple hypothesis tests show that a significant number of endorsements are associated with better election results. This allows us further analyze endorsement strategy of those political entities. We are able to see if they are effective in their political aims as well as use them as indicators as to the political feelings of the general population.

A Bayesian logistic GLM shows that several characteristics of states and districts are associated with whether or not particular endorsers decide to get involved in that race by making an endorsement. This allows for more insight into the endorsement process by explaining what sorts of races merit endorsements from various political figures and coalitions.

Many of these results suggest that there may be a lot of power in the words of relatively few people. Additionally, the GLM's suggested that there were relatively few races that actually mattered enough to receive endorsements from key figures. In some cases, locations with low populations had out-sized influence, either due to constitutional law or their coveted position as a swing district/state. This warrants much more investigation and causal analysis. Collaboration with political scientists can help us understand how endorsements fit into elections.

Of course there is always the wish to include more observations as well as include more complexity to our models. In this case, we would like to extend our inference by adding more demographic information about the states and districts such as religion, age distributions, major industries, and more. Additionally, we believe the results of our location-based model could themselves be fed into another Bayesian logistic model along with other features about specific candidates to explain characteristics of endorsements for individuals.

Finally, we would like to be able to make causal inferences about the relationships between endorsements and electoral outcomes. This would require even more data on the characteristics of these candidates and races.

7 Academic Honesty Statement

As members of the UC Berkeley community, we act with honesty, integrity, and respect for others. All of the work in this project is our own and we have cited all of the sources we used to the best of our abilities.

8 Appendix

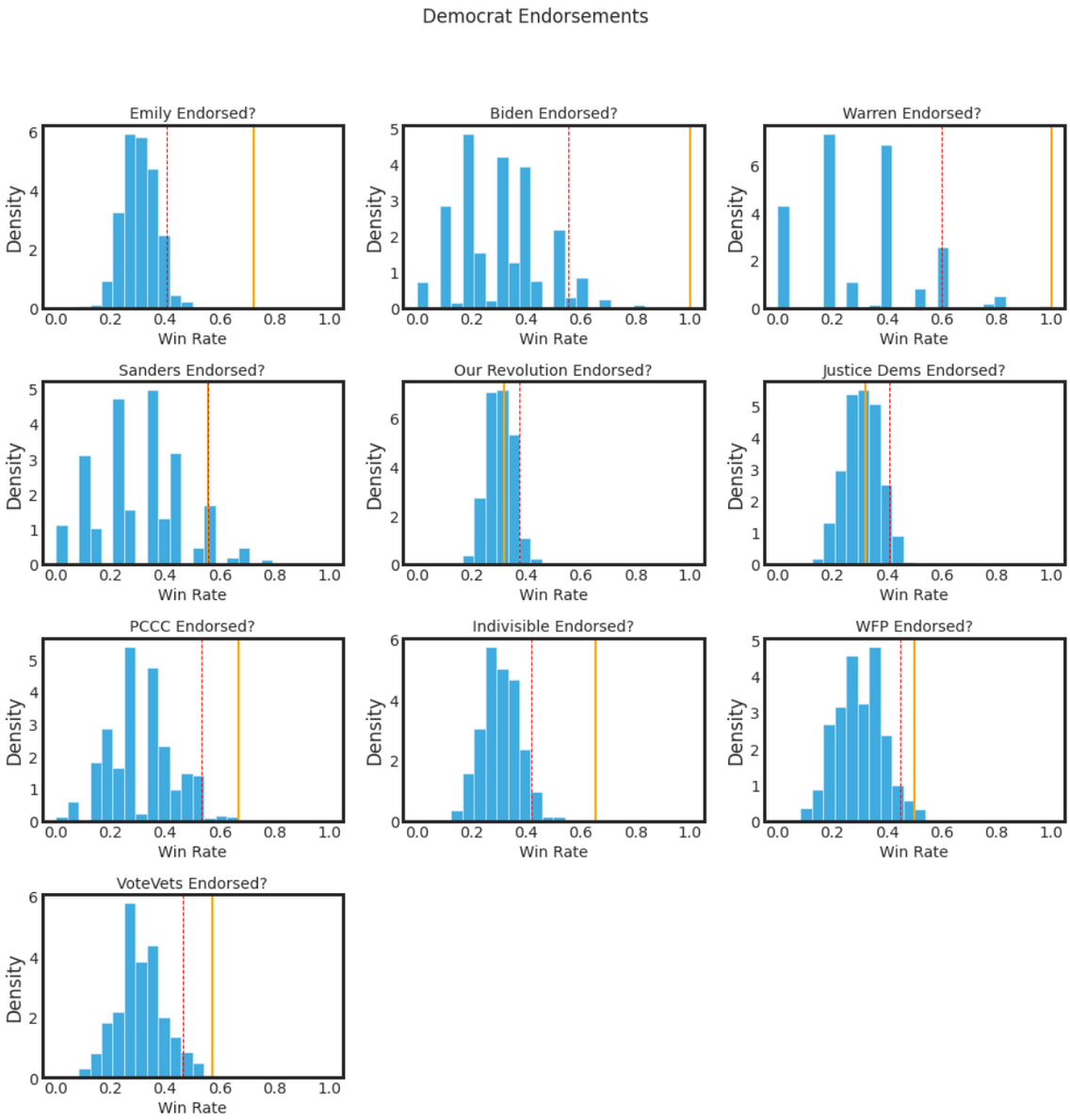

Figure A1: Individual 95% confident hypothesis tests for democratic endorsers. The red line indicates the 95 percentile of win rates and orange line indicates the observed win rate.

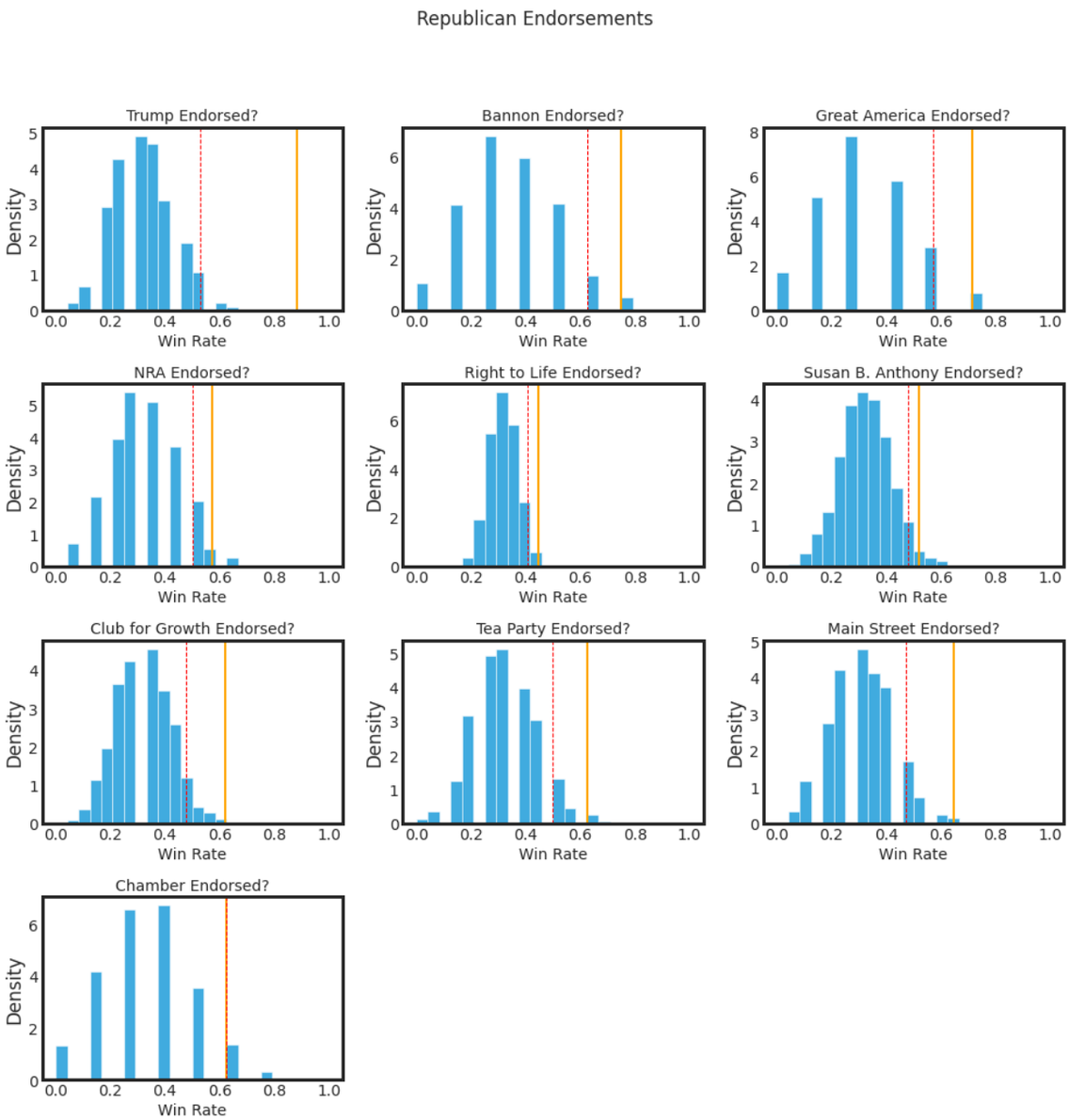

Figure A2: Individual 95% confident hypothesis tests for republican endorsers. The red line indicates the 95 percentile of win rates and orange line indicates the observed win rate.

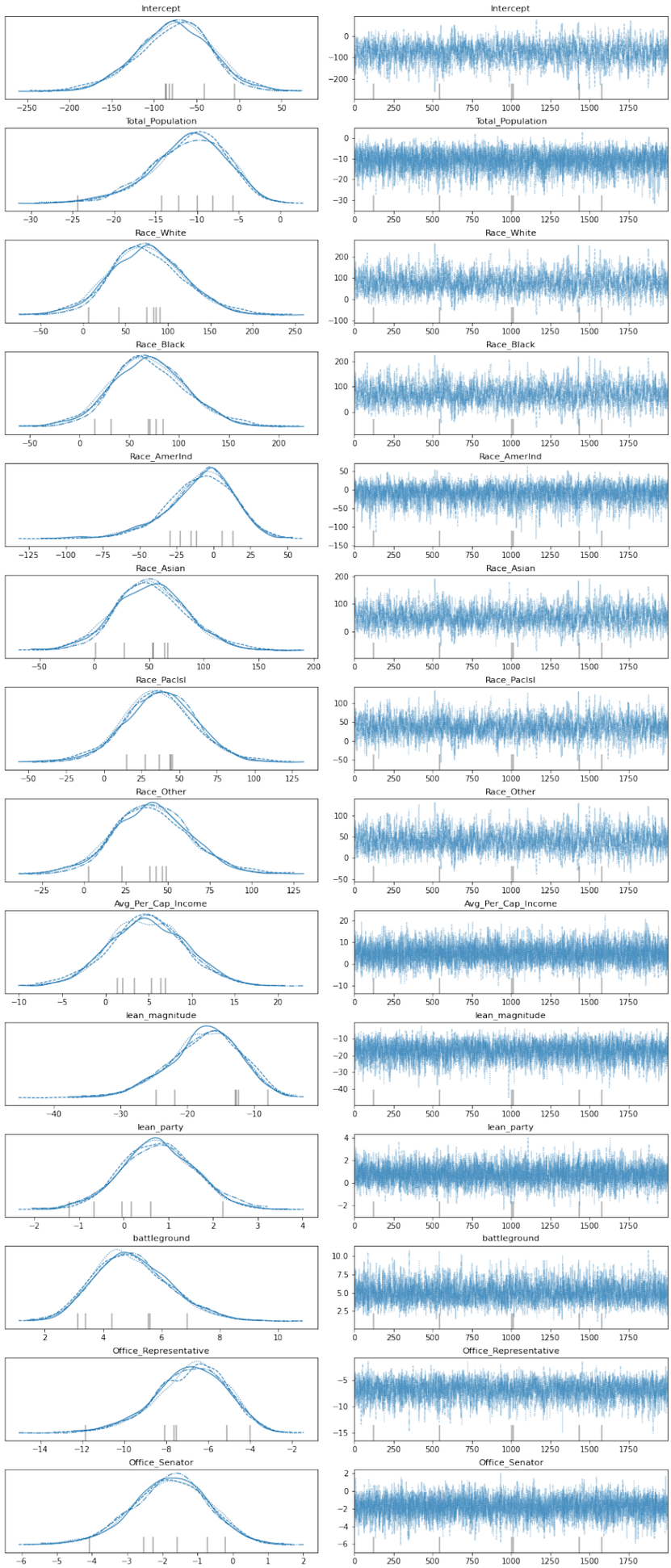

Figure A3: Distribution of Coefficients in Biden's GLM model.

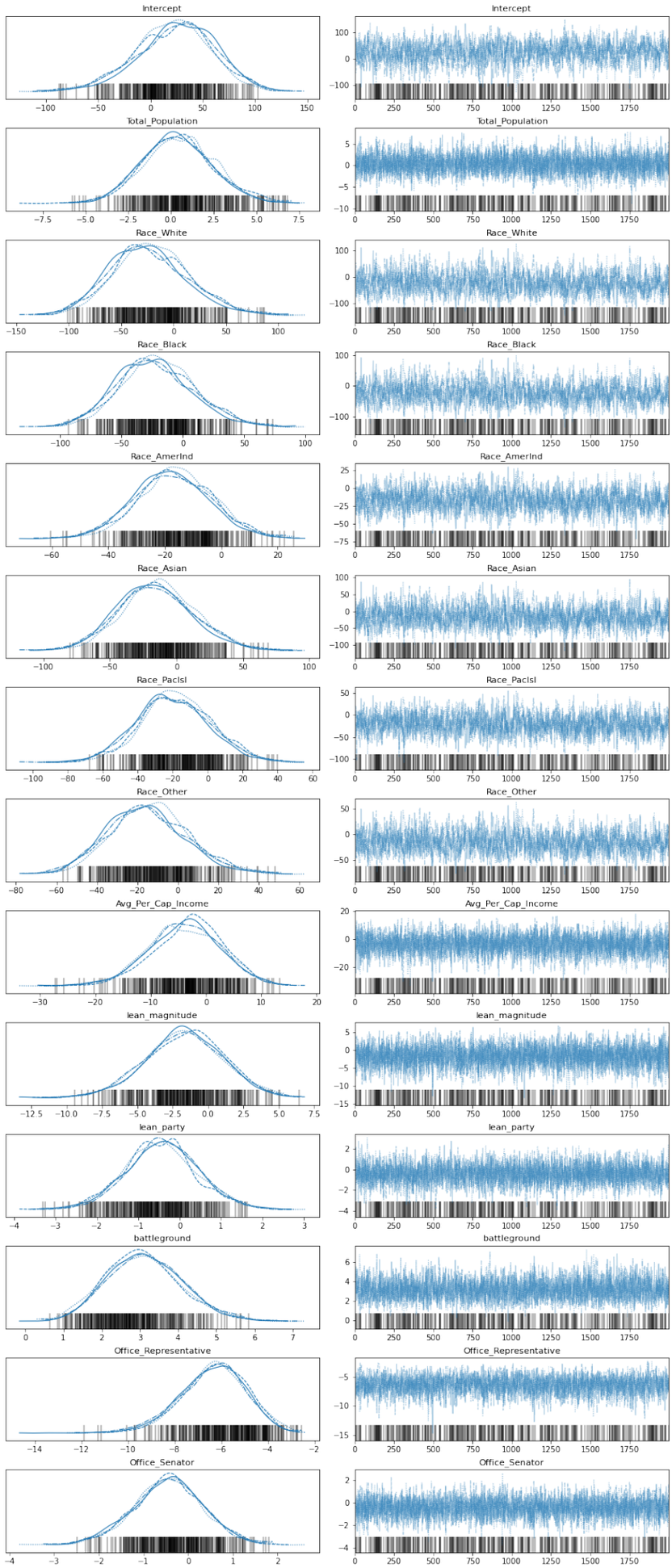

Figure A4: Distribution of Coefficients in Trump's GLM model.

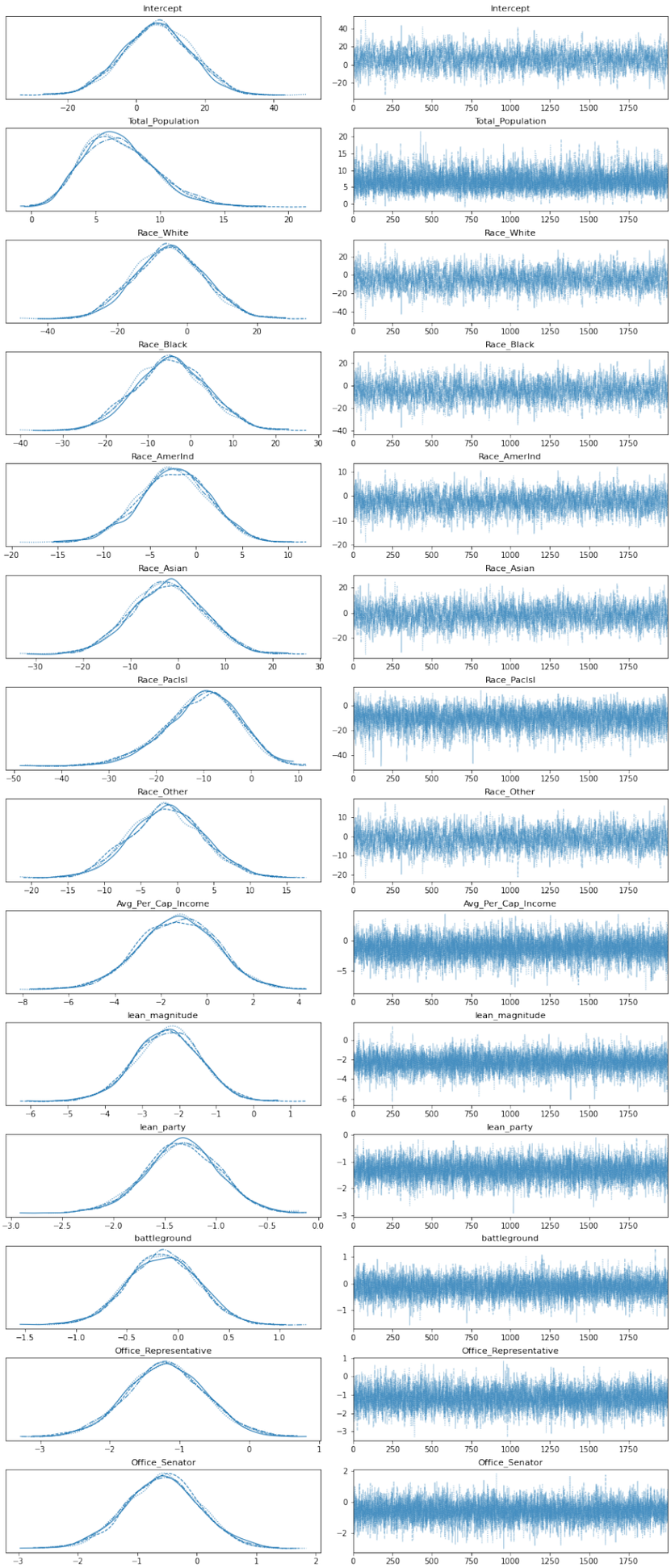

Figure A5: Distribution of Coefficients in Our Revolution's GLM model.

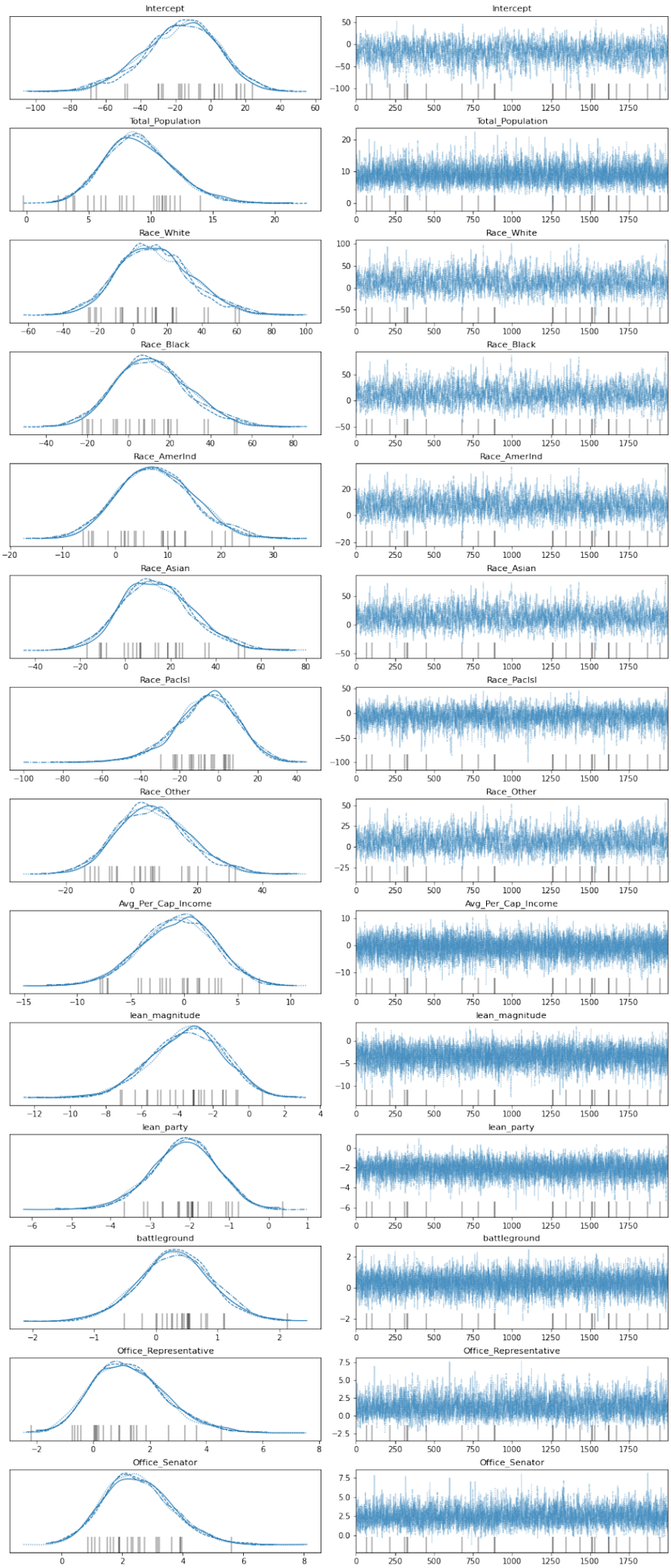

Figure A6: Distribution of Coefficients in Tea Party's GLM model.

	mean	sd	hdi_3%	hdi_97%	Significant?
Intercept	-76.491	44.196	-162.005	4.869	No
Total_Population	-11.033	4.587	-19.896	-3.135	Yes
Race_White	76.183	44.244	-5.765	161.407	No
Race_Black	70.356	38.125	-1.702	142.005	No
Race_AmerInd	-9.076	22.943	-53.147	32.364	No
Race_Asian	51.901	33.216	-10.699	115.621	No
Race_PacIsl	37.184	24.012	-6.555	84.563	No
Race_Other	40.780	22.950	-2.461	84.169	No
Avg_Per_Cap_Income	4.848	4.410	-3.645	12.979	No
lean_magnitude	-17.413	5.378	-27.709	-7.796	Yes
lean_party	0.765	0.834	-0.821	2.338	No
battleground	5.010	1.310	2.675	7.508	Yes
Office_Representative	-6.841	1.731	-10.141	-3.754	Yes
Office_Senator	-1.778	1.090	-3.801	0.259	No

Table 2: Table of coefficients for logistic model for prediction if Biden will endorse someone in the election.

	mean	sd	hdi_3%	hdi_97%	Significant?
Intercept	23.127	37.147	-47.865	92.456	No
Total_Population	0.410	2.085	-3.454	4.360	No
Race_White	-22.668	37.119	-92.186	47.658	No
Race_Black	-21.884	31.276	-81.308	37.102	No
Race_AmerInd	-16.776	13.752	-42.056	9.923	No
Race_Asian	-15.022	28.393	-67.994	38.284	No
Race_PacIsl	-18.905	20.813	-57.487	20.867	No
Race_Other	-14.694	19.577	-51.359	22.258	No
Avg_Per_Cap_Income	-3.705	6.591	-15.974	8.444	No
lean_magnitude	-1.674	2.759	-6.717	3.717	No
lean_party	-0.441	0.878	-2.015	1.325	No
battleground	3.118	0.989	1.279	4.976	Yes
Office_Representative	-6.343	1.426	-9.143	-3.872	Yes
Office_Senator	-0.447	0.791	-1.908	1.110	No

Table 3: Table of coefficients for logistic model for prediction if the Trump will endorse someone in the election.

	mean	sd	hdi_3%	hdi_97%	Significant?
Intercept	-17.099	21.447	-57.951	22.980	No
Total_Population	9.103	2.786	4.146	14.486	Yes
Race_White	12.952	21.416	-27.734	53.042	No
Race_Black	11.726	17.945	-19.422	47.831	No
Race_AmerInd	7.669	7.150	-5.305	21.451	No
Race_Asian	14.042	15.924	-14.387	45.275	No
Race_PacIsl	-6.618	16.517	-39.171	22.665	No
Race_Other	6.652	11.044	-14.250	27.379	No
Avg_Per_Cap_Income	-0.616	3.460	-7.038	5.830	No
lean_magnitude	-3.450	2.033	-7.359	0.138	No
lean_party	-2.145	0.825	-3.694	-0.639	Yes
battleground	0.306	0.593	-0.830	1.401	No
Office_Representative	1.276	1.313	-1.033	3.860	No
Office_Senator	2.521	1.135	0.604	4.827	Yes

Table 4: Table of coefficients for logistic model for prediction if the Tea Party will endorse someone in the election.

	mean	sd	hdi_3%	hdi_97%	Significant?
Intercept	6.105	10.028	-13.326	24.094	No
Total_Population	6.873	2.827	1.812	12.159	Yes
Race_White	-5.177	9.995	-23.422	13.824	No
Race_Black	-4.685	8.458	-20.096	11.561	No
Race_AmerInd	-2.427	3.809	-9.762	4.411	No
Race_Asian	-2.159	7.619	-16.112	12.287	No
Race_PacIsl	-10.445	7.851	-25.818	3.379	No
Race_Other	-1.474	5.044	-10.710	8.249	No
Avg_Per_Cap_Income	-1.270	1.695	-4.387	1.972	No
lean_magnitude	-2.318	0.886	-3.930	-0.617	Yes
lean_party	-1.325	0.359	-1.991	-0.626	Yes
battleground	-0.150	0.343	-0.762	0.518	No
Office_Representative	-1.195	0.560	-2.302	-0.186	Yes
Office_Senator	-0.574	0.625	-1.734	0.616	No

Table 5: Table of coefficients for logistic model for prediction if the Our Revolution will endorse someone in the election.

References

[1] *2018 ACS 1-year Estimates*. URL: https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2018/1-year.html (visited on 05/08/2021).

[2] *Cenpy — cenpy v1.0.0post2 Manual*. URL: https://cenpy-devs.github.io/cenpy/ (visited on 05/08/2021).

[3] *fivethirtyeight/data*. GitHub. URL: https://github.com/fivethirtyeight/data (visited on 05/08/2021).

[4] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: *PeerJ Computer Science* 2 (Apr. 6, 2016), e55. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.55. URL: https://peerj.com/articles/cs-55 (visited on 05/08/2021).

[5] Seth E. Spielman, David Folch, and Nicholas Nagle. "Patterns and causes of uncertainty in the American Community Survey". In: *Applied Geography* 46 (Jan. 1, 2014), pp. 147–157. ISSN: 0143-6228. DOI: 10.1016/j.apgeog.2013.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0143622813002518 (visited on 05/08/2021).

[6] *U.S. Senate battlegrounds, 2018 - Ballotpedia*. URL: https://ballotpedia.org/U.S._Senate_battlegrounds,_2018 (visited on 05/08/2021).