US General Society Survey Analysis*

Cristina Burca, Yan Mezhiborsky, Sakura Noskor

March 16, 2024

Abstract

This data uses the US 2020 Census results that is in the Harvard Database in order to analyse if there is a correltaion between the election numbers in 2020 and other factors. The factors include immigration status, gender, birth year, and results from the 2016 election. It was found that there does exist a correlation between the outcomes of the 2020 election and the above variables. These are important results as the US is a country that si considered a hub, due to all the external relations they have with multiple countries. These results will also help predict who may win the 2024 elections.

Table of contents

*Code and data are available at: repository

Introduction

The United States is one of the leading countries in export, imports and almost everything economic and socially related. The US contributes to the worlds economy by 20% despite the fact that they contain 5% of the populations(n.d.). This makes the US very relevant to not only national news but also international news. This is why the US presidential elections are broadcaster to worldwide ever election term. The results of the election not only affect American citizens but also external affairs related to the country.

The United States of America is a democratic government which means that they hold election every 4 years("3 u.s. Code § 1 - Time of Appointing Electors," n.d.). The 2 parties that historically run against each other with the most votes are the Republican party and the Democratic Party. The republicans are often associated with conservative beliefs and values such as views opposing abortion and privatization to save their economy. The Democrats, on the other hand, are often associated with liberal views such as social welfare programs and higher taxes to support the government aids provided to citizens(Encyclopedia Britannica n.d.). There are many other parties such as Libertarian Party, Green Party, Constitution Party and other independent candidates but because majority of the votes goes to the 2 parties the others are often over looked. The legal voting age is 18 in the states and you must be a registered voter in order to take part of the election which means non-citizens are not taken into account. Students are not as well. Also there seems to be a 66% voter turnout which means the remaining 34% decided not to vote (DeSilver 2022). The 2024 elections is also coming up this year meaning that analyzing this results may help us predict what the outcome of the election might be. As for the results of the election in 2020, Donald Trump won the precidential title.

This research examines the voting patterns in the 2016 and 2020 US Presidential elections. We will be looking at data collected by the Cooperative Election Study and accessed through the Harvard University Database (Kuriwaki, Beasley, and Leeper 2023). The analysis is based on a representative sample of 61,000 American adults, which provides detailed information about each individuals gender, birth year, race, registered state, employment, education loans, immigration status, dual-citizenship, religion, and 2016 and 2020 Presidential vote. The goal of this study is to use relevant variables from the electoral data to investigate patterns, trend and predictions regrading American electoral preferences from 2016 and 2020.

Data

Data Used

This paper was modeled with the help of R (R Core Team 2023) along with other useful packages like tidyverse (Wickham et al. 2019a) (which includes graphing functions like ggplot2), patchwork (Pedersen 2024). There are parts of the code which were guided by Rohan Alexander's Telling Stories with fire (Wickham et al. 2019b) chapter 13 section 13.2.2. The data was used from the Harvard database (Kuriwaki, Beasley, and Leeper 2023)

Variables inspected

Starting off, we examine the columns 'votereg' and 'voted_for'. They represent the number of persons that registered to vote and which candidate they voted for in 2020, respectively. We filtered out the rows with a 'votereg' value of 2, which indicated unregistered voters, to focus exclusively on individuals who were registered to vote. We then focused on the 'presvote16post' variable, which reveals the candidates Americans voted for in the 2016 United States Presidential Election. This is an important variable as it enables us to assess whether American citizens were satisfied with the service that the previous government provided. Next we look at 'gender' as well as 'employment'. Both 'gender' and 'employment' shows us if there is a correlation between certain parties views versus the demographic they represent. 'Gender' contains 2 values (male and female) while, employment has 9 values; full time worker, part time worker, laid off, unemployed, retired, permanently disabled individual, Homemaker, Student or Other. We also explore the variable 'immstat' which represents the immigration status of the of individual represented by one of the following: immigrant and citizen, immigrant not citizen, born in US, but parent(s) immigrant, parent and I born in US but grandparent(s) immigrant, or all born in US.

The Destination to Reach with the Data

There could have been many other similar data sets that could have been used for this project for example we could have chosen to look at the census and election data for Canada. However, our group decided that because part of the analysis was done in Wickham et al. (2019b), there were still many other variables that we could explore as we dive further into the 2020 presidential election and try to interpret if there are any correlations between the variables and the result. Our team found it interesting to see all the variables that were collected by the US government and the correlations we saw during the analysis process; where there most definitely was a positive correlation between each variable and the outcome of the votes. Although we are analyzing the 2020 election that has already taken place, the analysis we do in the later sections are believed to apply to the 2024 elections happening this year. This is enough reason for us and the reader to dive into the patterns that exist with this large data set.

Models/Results

Aging and Its Impact on Voting

In these following models we will of our research paper, we will look at the relationship between the demographic of age within the population and their voting patterns.The focus will primarily be on understanding how different age groups align themselves politically. By investigating the percentage of voters within each age bracket who support particular candidates, we aim to uncover nuanced insights into the demographic underpinnings of political support.Through the integration of descriptive statistics, analysis, and visual data representation we will offer a comprehensive overview on age base voting patterns. This will not only provide empirical evidence on the voting preferences of different ages but also help for a deeper understanding of the potential motivations behind their voting choices.

Voters in 2016

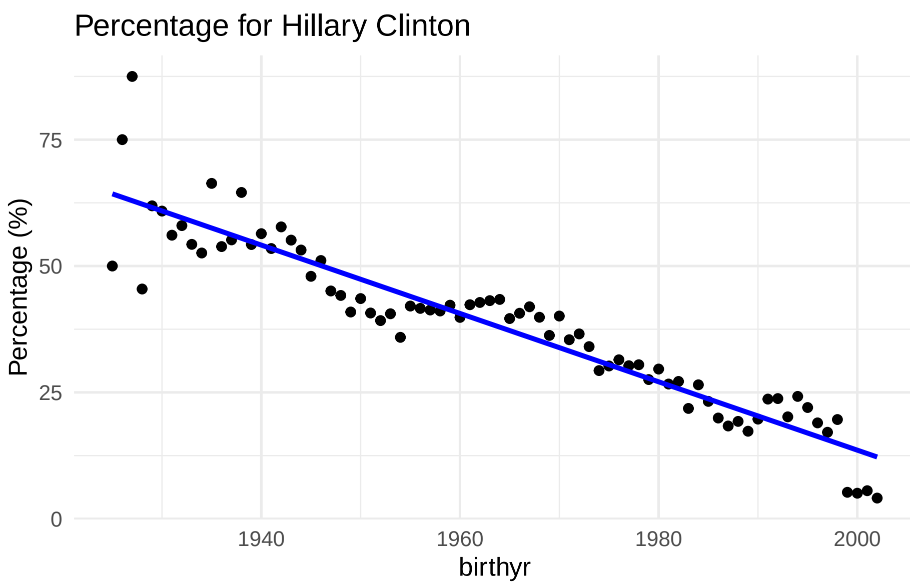


Figure 1: Linear Regression comparing Percentage of Votes For Hillary Clinton and Year of Birth (2016)

Model 2

In this model, we conduct an examination of the relationship between voters' birth year and their gender for the 2016 and 2020 Presidential elections. This analysis is visualized by plotting a histogram that separate female voters on the left and male voters on the right, with voters' birth

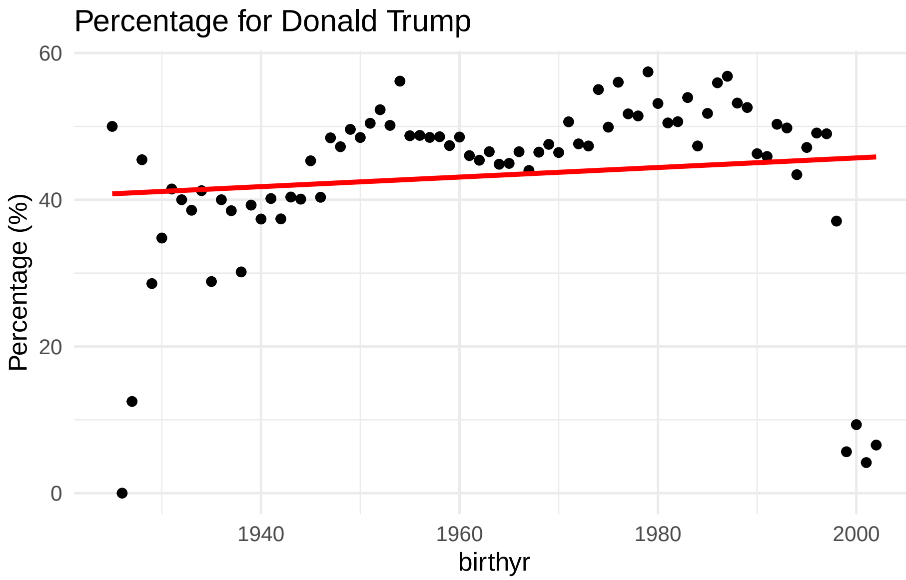


Figure 2: Linear Regression comparing Percentage of Votes For Donald Trump and Year of Birth (2016)

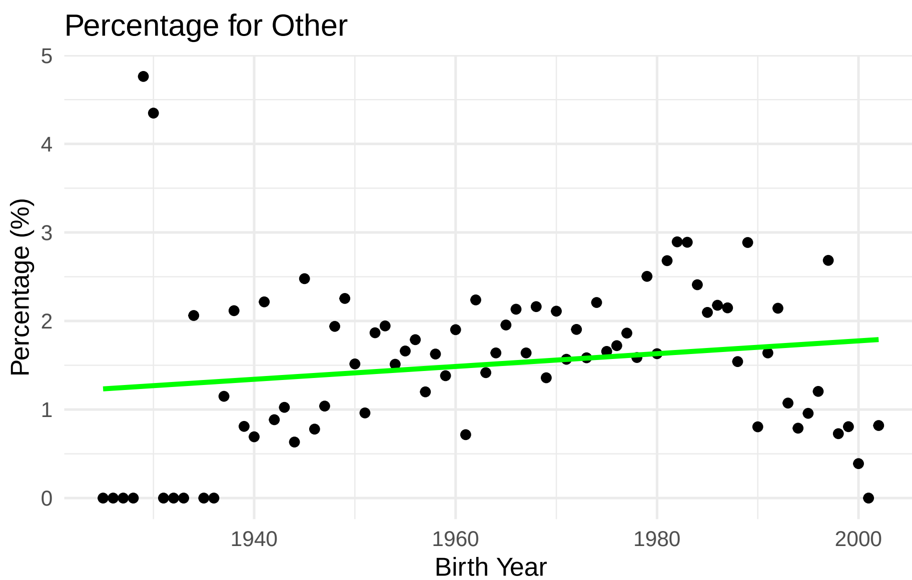


Figure 3: Linear Regression comparing Percentage of Votes For Other Caniddates and Year of Birth (2016)

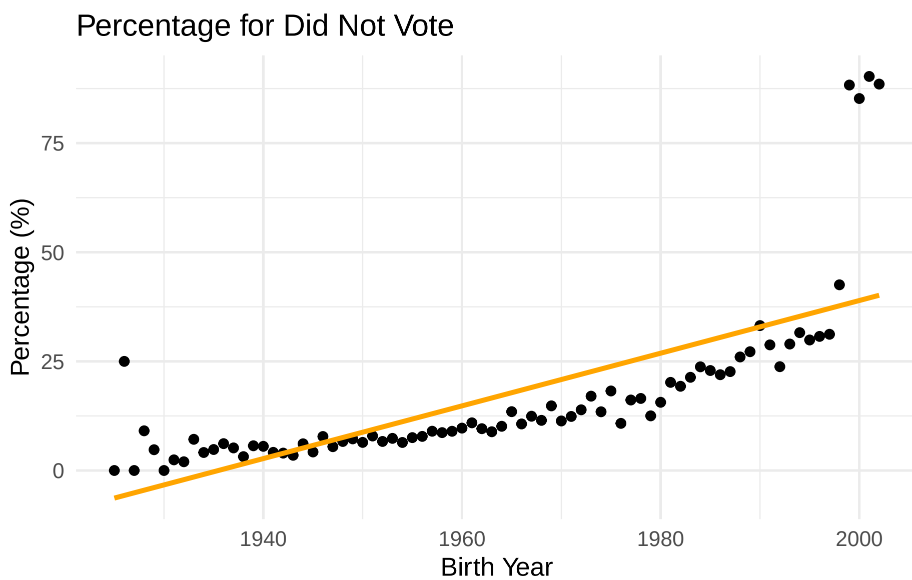


Figure 4: Linear Regression comparing Percentage of Non-Voters and Year of Birth (2016)

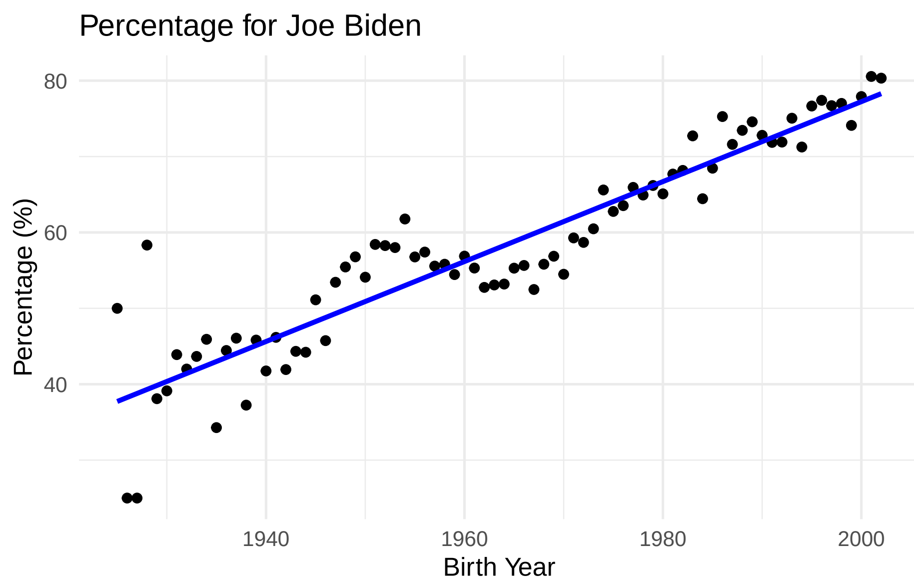


Figure 5: Linear Regression comparing Percentage of Votes For Joe Biden and Year of Birth (2020)

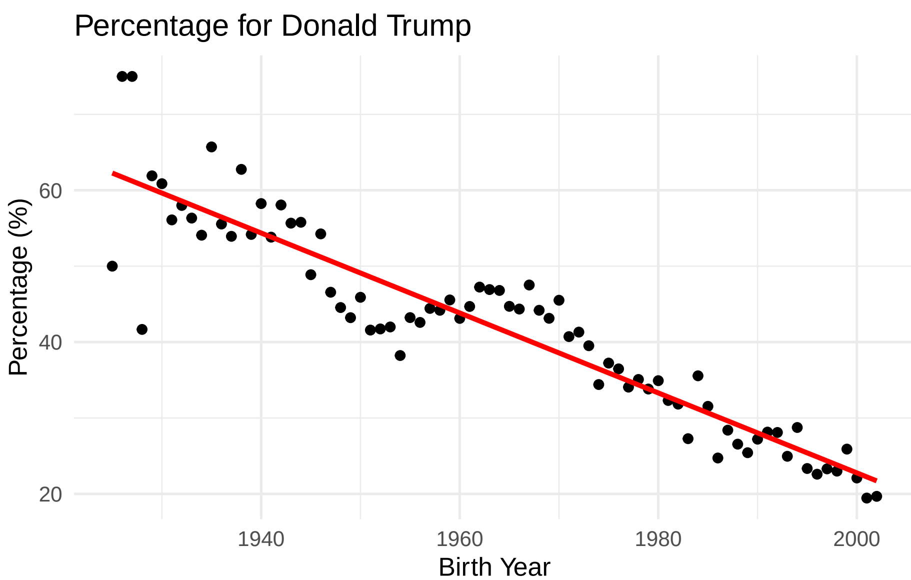


Figure 6: Linear Regression comparing Percentage of Votes For Donald Trump Year of Birth (2020)

years measured along the x-axis, which ranges from 1925 to 2002. The y-axis quantifies the voter turnout for the year. For clarity ad symbolic representation, the colour blue was chosen to represent the Democratic candidates –Hillary Clinton for the 2016 election, and Joe Biden for the 2020 election, while red was chosen to the Represent the Republican candidate, Donald Trump, who sought the presidency in both terms. Figure 7 present the distribution of votes in 2016, and Figure 8 presents the data from the 2020 election.

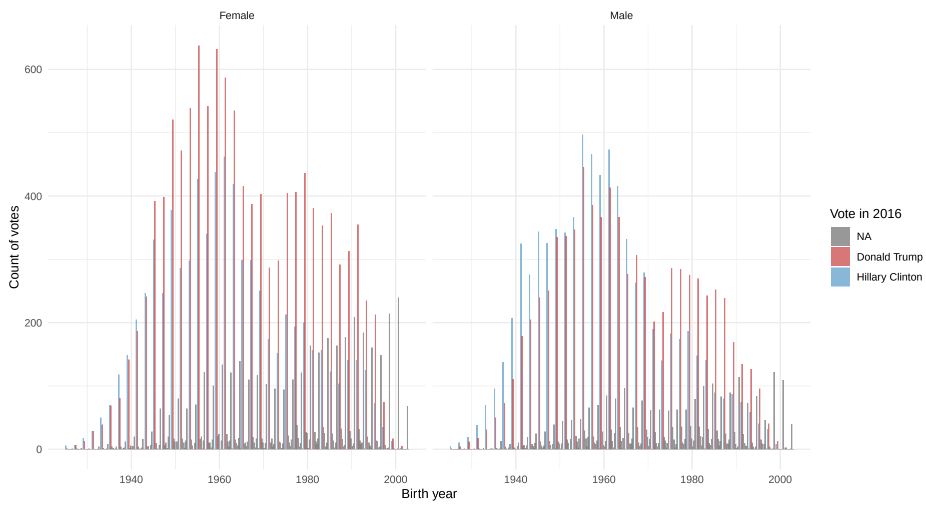


Figure 7: Logistic regression of 2016 US Presidential votes comparing parameters of gender and immigration status

We notice in Figure 7 the graph displays two distinct high points for the number of female Republican voters, with the peak around 1960 being the most pronounced, followed by another around 1980. There are similar peaks in the graph for female Democratic voters, but with the overall count being considerably lower. We also notice that the Democratic party received slightly more votes from the older demographic, where as the younger demographic greatly preferred the Republican party. In comparison, the graph displaying the male votes has a more balanced distribution, and similar to the female voters, the older and younger demographic preferred the Democratic Party and the Republican Party, respectively. It is worthy to mention that there is a higher count of women than men.

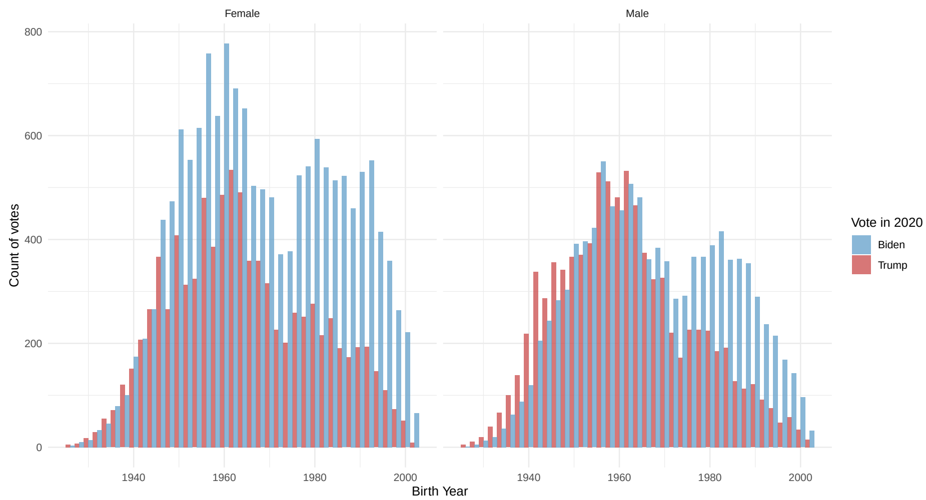


Figure 8: Logistic regression of 2020 US Presidential votes comparing parameters of gender and immigration status

We notice in Figure 8 that the graphs bear a strong resemblance to the distributions of the 2016 votes depicted in Figure 7, but with the parties reversed on the graph. We are interested in the ratios of these graphs.

We analyze the ratio of Democratic to Republic votes in 2016 in Table 1 and in 2020 in Table 2. We divide respondents by birth decade, and we list the total number of votes for the Democratic Party and thee Republican Party, then list the ratio of the two. The decades 1920 and 2000 were omitted since the number of respondents in those Birth Decades were very low. The value of the ratio in centered at 1, where if the values is greater than 1, then within that group, the Democratic Party has more votes than the Republican Party. Where the values are less than 1, the

Table 1: **?(caption)**

(a)

Table: Ratio of 2016 Votes for Clinton to Trump by Birth Decade (Females)

Birth Decade	Total Clinton Votes	Total Trump Votes	Ratio
1930	285	231	1.23
1940	1179	1360	0.87
1950	1730	2712	0.64
1960	NA	NA	NA
1970	NA	NA	NA
1980	741	1835	0.40
1990	NA	NA	NA

Table: Ratio of 2016 Votes for Clinton to Trump by Birth Decade (Males)

Birth Decade	Total Clinton Votes	Total Trump Votes	Ratio
1930	361	184	1.96
1940	NA	NA	NA
1950	NA	NA	NA
1960	1917	1731	1.11
1970	966	1262	0.77
1980	NA	NA	NA
1990	NA	NA	NA

Table 2: **?(caption)**

(a)

Table: Ratio of 2020 Votes for Biden to Trump by Birth Decade (Females)

Birth Decade	Total Biden Votes	Total Trump Votes	Ratio
1930	271	293	0.92
1940	1560	1257	1.24
1950	3176	1911	1.66
1960	3119	2229	1.40
1970	2293	1253	1.83
1980	2629	1104	2.38
1990	2120	716	2.96

Table: Ratio of 2020 Votes for Biden to Trump by Birth Decade (Males)

Birth Decade	Total Biden Votes	Total Trump Votes	Ratio
1930	219	365	0.60
1940	1154	1542	0.75
1950	2224	2171	1.02
1960	2190	2176	1.01
1970	1670	1170	1.43
1980	1883	841	2.24
1990	1054	393	2.68

Republican Party has more votes. When the ratio is around 1, the number of votes is approximately equal.

We notice that in 2016, the ratio of votes decreases as birth decade increases for both women and men voters. The opposite happens in 2020, where the ratio of votes increases as birth decade increases.

Model 3

We have modeled the following logistic regression in the graphs:

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{immigration status}_i + \beta_3 \times \text{previous 2016 vote}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

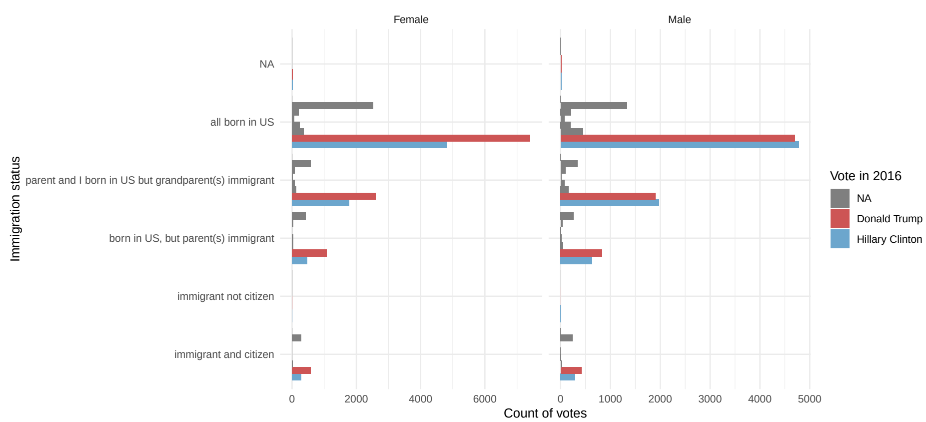
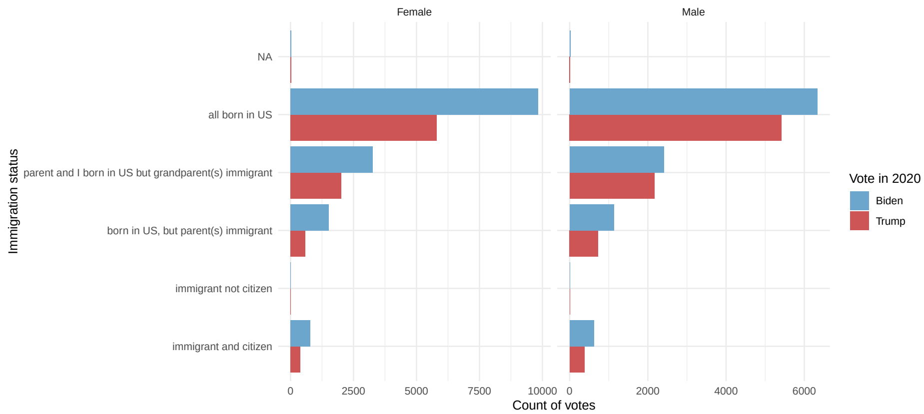


Table 3: Explanatory models of presidential election in 2020 based on gender and immigration status

	Support Biden
(Intercept)	−2.652
	(0.557)
genderMale	−0.249
	(0.247)
immstatimmigrant not citizen	54.378
	(47.512)
immstatborn in US, but parent(s) immigrant	−0.336
	(0.625)
immstatparent and I born in US but grandparent(s) immigrant	0.039
	(0.553)
immstatall born in US	0.186
	(0.504)
presvote16postDonald Trump	6.631
	(0.409)
presvote16postGary Johnson	3.818
	(0.533)
presvote16postJill Stein	5.020
	(1.183)
presvote16postEvan McMullin	33.118
	(25.435)
presvote16postOther	3.589
	(0.595)
presvote16postDid not vote	3.452
	(0.284)
Num.Obs.	994
R2	0.709
Log.Lik.	−249.986
ELPD	−261.6
ELPD s.e.	19.6
LOOIC	523.2
LOOIC s.e.	39.3
WAIC	521.8
RMSE	0.27



The numbers on the first half of the table Table 3, right next to each variable name, shows us the coefficient related to each of the predictor variables. It tells us whether the respondent is expected to vote for Biden or not based on their immigration status, gender and who they voted for in the 2016 elections. The bracketed values are the errors associated with the coefficients so the smaller the number the less likely the model inaccurately came up with the coefficients. Notice that one of the variables is named intercept which essentially represents gender Female. We see that in "parent and I born in US but grandparent(s) immigrant", "gender male" and "gender female", the coefficients are negative meaning that they are not inclined to vote for Biden in the 2020 election. All other values are positive and we see for people who voted for Trump in the 2016 election, they have a higher coefficient with a relatively small inaccuracy rate.

Discussion

First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

Second discussion point

Modeling immigration status, gender and 2016 voting results

Table 3 was a very interesting plot to examine as they let us see the correlation between the election held in 2020 and other factors such as gender, immigration status, and the election results in 2016. Starting off we look at the negative values. We see that both males and females are less likely to vote for Biden as their coefficient is negative. The female has more of a negative value meaning that their priorities do not align with the view of the democratic party. This makes sense as we look at the news in the past years when we see the Biden government push for anti-abortion laws, the very law that a lot of women are opposed to as they argue they deserve to make that choice. As for the negative value for the males, there does not seem to be a direct answer to that. From here we see that it makes more sense to look at the large coefficients as the smaller values may be a result of many different things. A person is just not male but they may also be from a different ethnicity or religion whose priorities are different from the Biden government. America is a very diverse country so it makes sense that many people have different reasons to support and not support the republicans. The next largest value we notice is for people who were born in the US but have immigrant parents and people who are immigrants but not citizens. This makes sense as the Biden government actually reversed some of the immigration plans implemented by the previous government that harmed immigrants. Some of these laws were the travel ban put on Muslim-majority countries that resulted in many immigrants being unable to come to America to find a better future. It makes sense that children of people who immigrated and people who are immigrants themselves would prioritize such issues. Lastly, we look at the another coefficient found in the table which is the people who voted for Trump in the 2016 election. We see that it is very likely that people who voted for Trump in the last election, voted for Biden in this one. This may be due to the fact that there were many protests of the Trump administration and how they ran the country during their term. Some of the backlash that happened while Trump was president was the travel ban on Muslim-majority countries, reduced government regulation across various industries, and tax cuts from private corporations. These decisions may have led many people to not re-elect Trump for another 4 years as their president. The same is seen for people who voted for Evan McMullin (independent candidate) and other smaller parties. The reasons are hard to infer but it may be due to the fact that their candidate did not run again or their priorities and views changed from the party/person they voted for.

Weaknesses and next steps

It is nearly impossible to have data sets that are completely error-free and the same is true for this data set. The data set we used as analysis does not take into account the opinions of non-citizens and who they would have voted for. A lot of marginalized opinions are being left out when we clean the data and we only look for people that could vote. For example, refugees may have a different opinion on who got elected in the 2020 presidential candidate as they come from a different aspect of life and have different priorities that need to be addressed such as housing and or health care. Another weakness in our data set is what occurs with a lot of data sets, which is missing values. This data set also contained them and those needed to be cleaned so that our analysis would not sway.

Speaking of missing values, insignificant values were ignored in some of the models as well. For, example some of the presidential candidates got so few votes that including them would have actually hindered our ability to see the trends so we had to group them into others. This was beneficial for this paper but if we do indeed continue to analyze this paper for other purposes, this might cause problems.

In terms of the next steps for this paper, we look forward to predicting the 2024 election results by comparing the general priorities of the population. We could look at the population by immigration status, gender, age, or their previous 2020 voting results. If we are able to correctly predict the results based on the model continuously, it would help us understand the priorities of American citizens; knowledge that would in turn help us help them.

Appendix

Additional data details

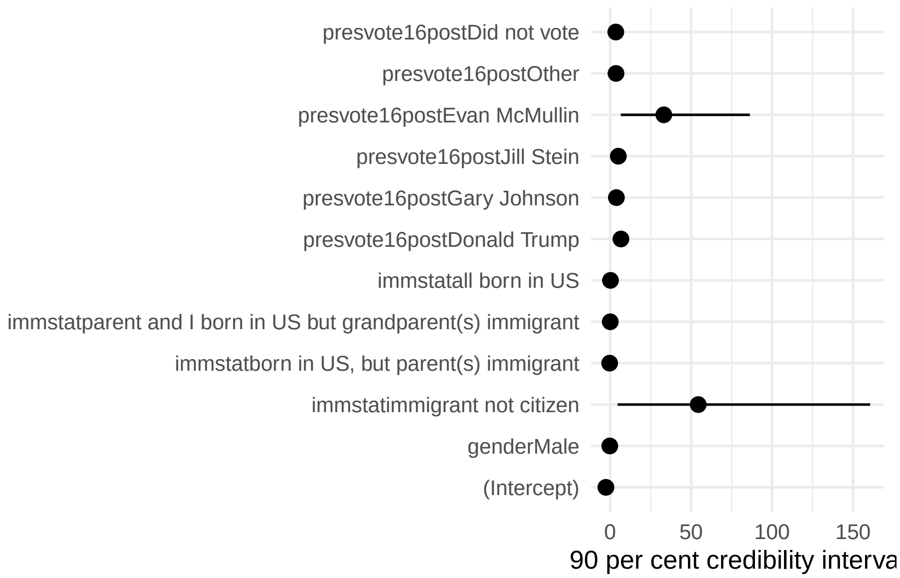


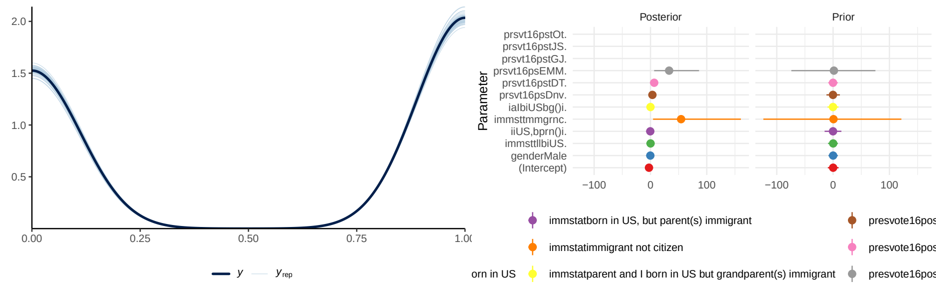
Figure 9: Credible intervals for predictors of support for Biden

Model details

Posterior predictive check

In Figure 10a we implement a posterior predictive check.

In Figure 10b we compare the posterior with the prior.

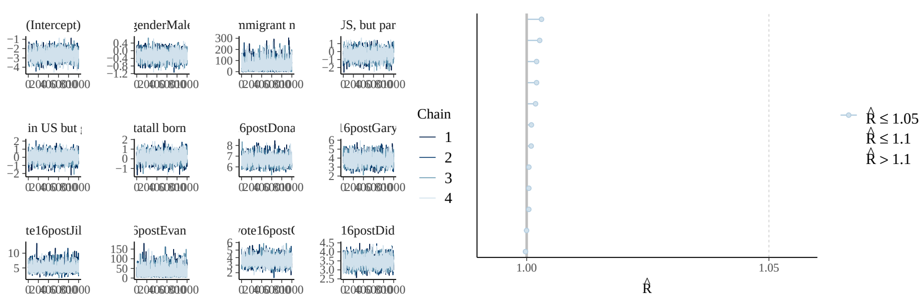


(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 10: Examining how the model fits, and is affected by, the data

Diagnostics



(a) Trace plot

(b) Rhat plot

Figure 11: Checking the convergence of the MCMC algorithm

References

n.d. *United States Trade Representative*. https://ustr.gov/issue-areas/economy-trade#:~:text=Constituting%20less%20than%205%20percent,economy%20and%20leading%20global%20trader.

"3 u.s. Code § 1 - Time of Appointing Electors." n.d. https://www.law.cornell.edu/uscode/text/3/1.

DeSilver, Drew. 2022. "Turnout in u.s. Has Soared in Recent Elections but by Some Measures Still Trails That of Many Other Countries." *Pew Research Center*. Pew Research Center. https://www.pewresearch.org/short-reads/2022/11/01/turnout-in-u-s-has-soared-in-recent-elections-but-by-some-measures-still-trails-that-of-many-other-countries/.

Encyclopedia Britannica. n.d. "How Is the Democratic Party Different from the Republican Party?" https://www.britannica.com/story/how-is-the-democratic-party-different-from-the-republican-party.

Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *Dataverse: R Client for Dataverse 4+ Repositories*.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019a. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.