Generational Dynamics, Gender and Politics in the US*

Predicting 2020 US Presidential Voting Preferences

Hritik Shukla

March 16, 2024

This study investigates voting patterns in the 2020 US presidential election, analyzing how political preferences are influenced by a person's gender and the generation they're born in. With the help of exploratory data analysis, logistic regression and data from the 2020 CES Common Consent dataset, it is observed that younger generations tend to favor Biden over Trump. As generations get older, Biden loses more and more support. It is also found that women tend to favor Biden more heavily than the men of the same generation as them.

1 Introduction

'If you're not a liberal when you're 25, you have no heart. If you're not a conservative by the time you're 35, you have no brain.'

Although commonly, but incorrectly attributed to Winston Churchill, this quote summarizes a common belief held in our society - a person generally starts out in life with a left leaning political compass, and as they get older, their beliefs transform into more right leaning ideologies. It is also a popular belief that newer, younger generations tend to be more inclusive, open and liberal in their ideologies (Itkowitz (2019)), and this only intensifies as further generations get old enough to participate in the political landscape of the US.

This paper aims to find out whether these generalizations hold true by observing the 2020 US Presidential Elections. It is important to note that 2020 was a tumultuous year for the entire world, it first year of COVID-19 when not much was known about the disease. President Trump, who was in power before the elections, was greatly criticized for his poor handling of pandemic and lockdown, which lead to massive change in sentimenets amongst the citizens of the country (Jurkowitz (2020)). As a result, with highly charged sentiments, this election

*Code and data are available at: https://github.com/hritikshuklas/vote-preferences-generations

saw the highest voter turnout of any presidential election held in the US in the 21st century (*2020 Presidential Election Voting and Registration Tables Now Available* (2021)). Another interesting factor to consider is that this was the first election where Generation Z could participate. Furthermore, they held highest voter turnout of any generation (Hess (2020)). Therefore, it would be even more interesting to observe if these long accepted norms held up in these extraordinary conditions, under situations which would not be present under any other normal circumstances - the first election of its kind in recent memory.

The remainder of this paper is structured as follows:

- Section **??** explores the dataset and the variables within it used for the study
- Section **??** explains our models setup and our assumptions going into the study
- Section **??** explores our findings in detail
- Section **??** evaluates these findings in the context of our reality

I used R (R Core Team (2023)), along with multiple packages to aid in the data analysis and modelling. More particularly, the packages tidyverse (Wickham et al. (2019)), dplyr (Wickham et al. (2023)), tidyr (Wickham, Vaughan, and Girlich (2023)), dataverse (Kuriwaki, Beasley, and Leeper (2023)) and arrow (Richardson et al. (2024)) were used for data acquisition, testing and cleaning. The package rstanarm (Goodrich et al. (2024)) was used for modelling, and ggplot2 (Wickham (2016)), knitr (Xie (2015)) and modelsummary (Arel-Bundock (2022)) were used for data visualization.

2 Data

The final release of the 2020 CES Common Consent dataset by (Schaffner, Ansolabehere, and Luks (2021a)) was used for this study, acquired through the dataverse package (Kuriwaki, Beasley, and Leeper (2023)). This dataset was created from a survey conducted by YouGov, an internet-based data analytics firm based in the UK. Random sample methodology was used to survey 61,000 adults over the internet from September to October 2020 (Schaffner, Ansolabehere, and Luks (2021b)). For our model, the "gender" (Section **??**) and "CC20_410" (Section **??**) variables were used from this dataset directly, and the "birthyr" variable was used to construct the "generation" variable (Section **??**).

2.1 Data Cleaning

The "votereg", "CC20_410", "gender" and "birthyr" variables from CCES 2020 dataset (as seen in Table **??**) were selected initially.

As the first step, any voters who weren't registered to vote were removed from our data. The "votereg" variable was used for this process, and then was subsequently removed from our

Table 1: Raw Dataset Preview

votereg	CC20_410	gender	birthyr
1	2	1	1966
2	NA	2	1955
1	1	2	1946
1	1	2	1962
1	4	1	1967
1	2	1	1961
2	NA	1	1950
1	2	2	1947
1	2	2	1970
1	1	2	1963

finalized data as it served no purpose for our model. Next, "CC20_410", which records who the respondent voted for for President of the Untied States, was used to filter out respondents who voted for Joe Biden or Donald J. Trump, as we are interested in peoples' preference towards the Democrats or the Republicans. Any respondents who were born outside the years 1928 to 2012 were removed from the survey, as there wouldn't be enough data for those generations to make any statistical inferences from them.

Next, any responses which were missing any of the these variables we needed were removed during cleaning. This omission culled down our total responses from 61,000 to 43,540. However, we still have more than enough data to derive meaningful results.

Some variables were renamed to be more human-readable - "CC20_410" was renamed to "voted_for" and "birthyr" was renamed to "birthyear". The values for categorical variables (such as "gender" and "voted_for") within the dataset were cleaned to represent their more meaningful, intended values. The questions corresponding to these variables in the survey have a select number of options for the subject to choose from. These responses are recorded in the dataset as a number, where each number corresponds to a relevant response - for example, for the gender variable, a response of 1 means female and 2 means male. During cleaning, these numbered responses were replaced by their true values.

The end result of the cleaning process leaves the dataset shown in Table **??**.

2.2 Gender

The gender variable stores the responses of the respondent to a question which asks them to choose between "Male" and "Female" as their only two options to describe their gender. This question would have been better phrased if it were asking for biological sex assigned at birth rather than gender based on the options provided to the respondents. Even though the survey

Table 2: Cleaned Dataset preview

Gender	Birthyear	Generation	Voted For
Male	1966	Generation X	Trump
Female	1946	Baby Boomer	Biden
Female	1962	Baby Boomer	Biden
Male	1961	Baby Boomer	Trump
Female	1947	Baby Boomer	Trump
Female	1970	Generation X	Trump
Female	1963	Baby Boomer	Biden
Female	1966	Generation X	Biden
Female	1961	Baby Boomer	Biden
Male	1959	Baby Boomer	Biden

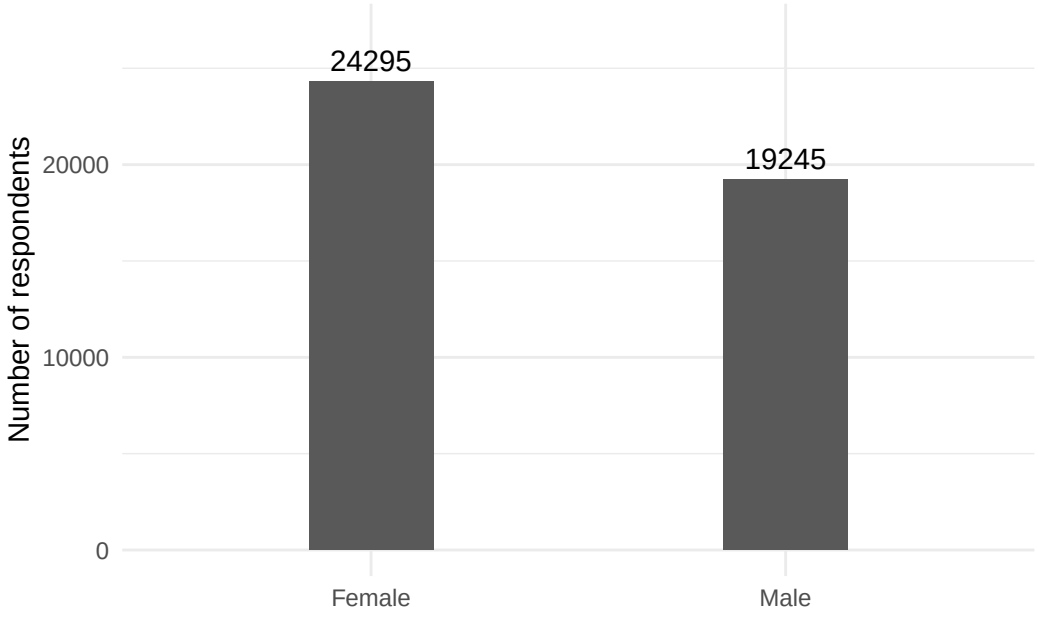

Figure 1: Gender Distribution

asks for the respondent's sexual orientation later, the phrasing of this question alone could be a reason for people to not submit their survey responses as they don't identify with the given options which may lead to under-representation of people belonging to this demographic in the dataset.

In Figure **??**, we can see the number of respondents who identified as male and female. According to "The Gender Ratio of United States of America (2020 - 2028, Males Per 100 Females)" (2024), the United States had a gender ratio of 97.14 males to 100 females in 2020. Our data consists of 24,295 females and 19,245 male - the gender ratio represented here is considerably below the national average at the time of the survey - this might introduce some bias into our data.

2.3 Voted For (CC20_410)

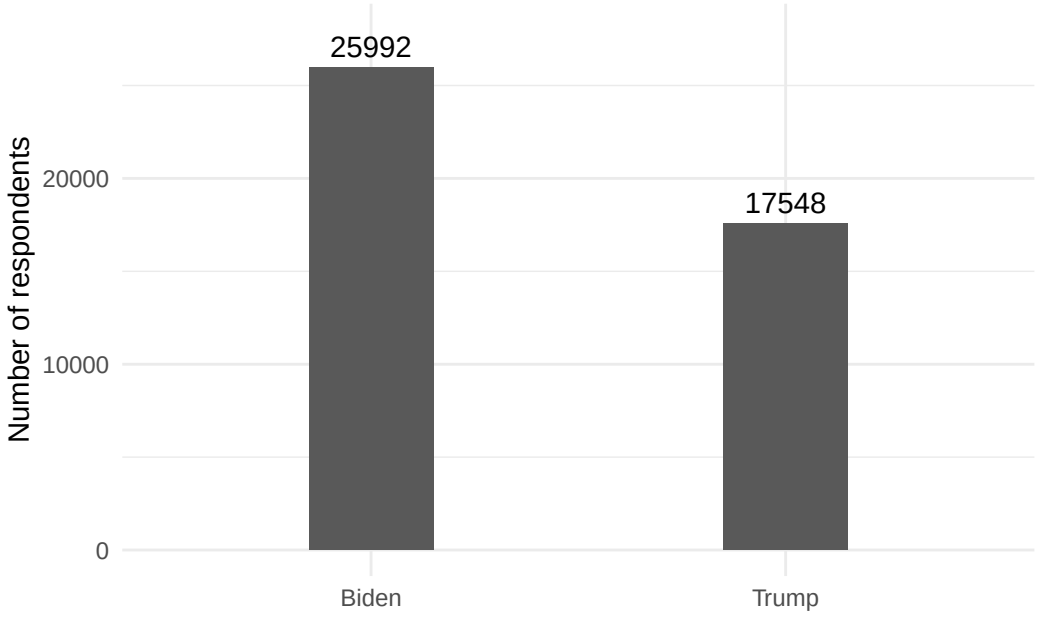

Figure 2: Vote Distribution

The CC20_410 variable, renamed to "voted_for", records who the respondent voted for in the 2020 Presidential Elections. The respondents were given options other than Joe Biden and Donald Trump for this question, such as "Other", "I dd not vote", "Not Sure", etc, but these options were removed as they weren't needed for our purposes.

In Figure **??**, we can see that Biden has an overwhelming majority over Trump in terms of popularity, with around a 6,000 vote difference between the two.

2.4 Generation (and birthyear)

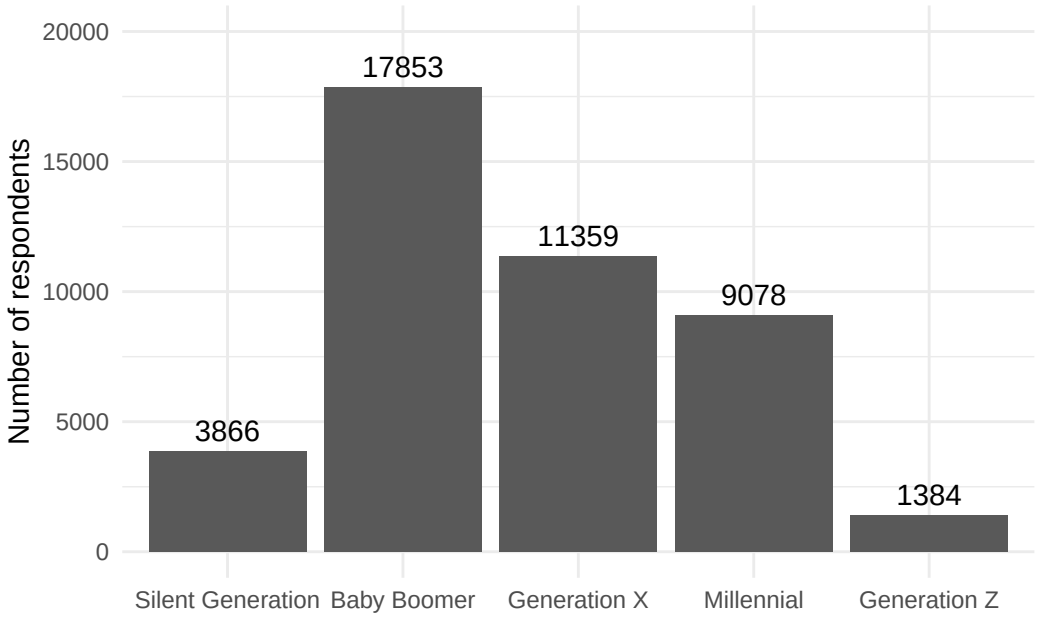

Figure 3: Generation Distribution

The generation variable was created from the "birthyr" variable in the dataset, where respondents recorded their year or birth. Each respondent was assigned a "generation" according to their year of birth. The generation breakpoints used for this paper are defined by Pew Research Center (Dimock (2019)) as follows:

- Silent Generation: 1928 - 1945
- Baby Boomer: 1946 - 1964
- Generation X: 1965 - 1980
- Millennial: 1981 - 1996
- Generation Z: 1997 - 2012

Any respondents whose year of birth falls outside of these year ranges were removed from the dataset. These numbers fall in line with the voter turnout

In Figure **??**, we can see that Baby Boomers show the largest representation of all groups. Generation X and Millennial have similar number of respondents, and Silent Generation and Generation Z have the lowest number of respondents. This distribution falls in line with the voter turnout by age statistic reported by Pew Research Center (Gramlich (2020)). Note that

even though Generation Z had the highest voter turnout for all generations (Dimock (2019)) and their numbers look low compared to others, it falls in line with expected numbers as the oldest Generation Z respondent who would've been able to vote would've been born in 2002, leaving a considerable population of this generation unable to vote.

3 Model

We will be using logistic regression to model our data, where our outcome variable would be whether a respondent prefers Biden as the presidential candidate. Gender and generation will be used as predictors our outcome variable.

3.1 Model set-up

Define:

- y_i is the political preference of the respondent and equal to 1 if Biden, and 0 if Trump
- gender_i is the gender of the respondent
- generation_i is the generation of the respondent

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{generation}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

Logistic regression is employed for this model since our variable of interest can be constructed as a binary outcome variable - respondent prefers Biden, respondent doesn't prefer Biden (prefers Trump). Logistic regression is well suited for situations where the outcome variable represents two mutually exclusive categories and its probability is based on a set of predictor variables (here, gender and generation).

According to the common beliefs in our society as established in Section **??**, we expect to see a positive relationship in the younger generations, who we expect to be more left-leaning and therefore be more favorable towards Biden. And vice versa, we expect to see a negative relationship in the older generations, who we expect to be more right leaning, and therefore, more conservative. However, due to the large dissatisfaction in the majority of US citizens in 2020 due to how President Trump handled the COVID pandemic (Jurkowitz (2020)), this skew might not be as great as it could be for other elections. Regardless, we expect to see a linear relationship between a person's generation and preference for Biden - younger generational cohorts leaning more towards voting for Biden.

The relationship between gender and political preference is a bit more complicated. According to articles from the Conversation (Rosie Campbell (2023)) and the Gallup (Saad (2024)), women tend to vote more conservative than men before 2017, after which this trend flipped on its head and women are began to vote more liberally.

4 Results

Our results are summarized in.

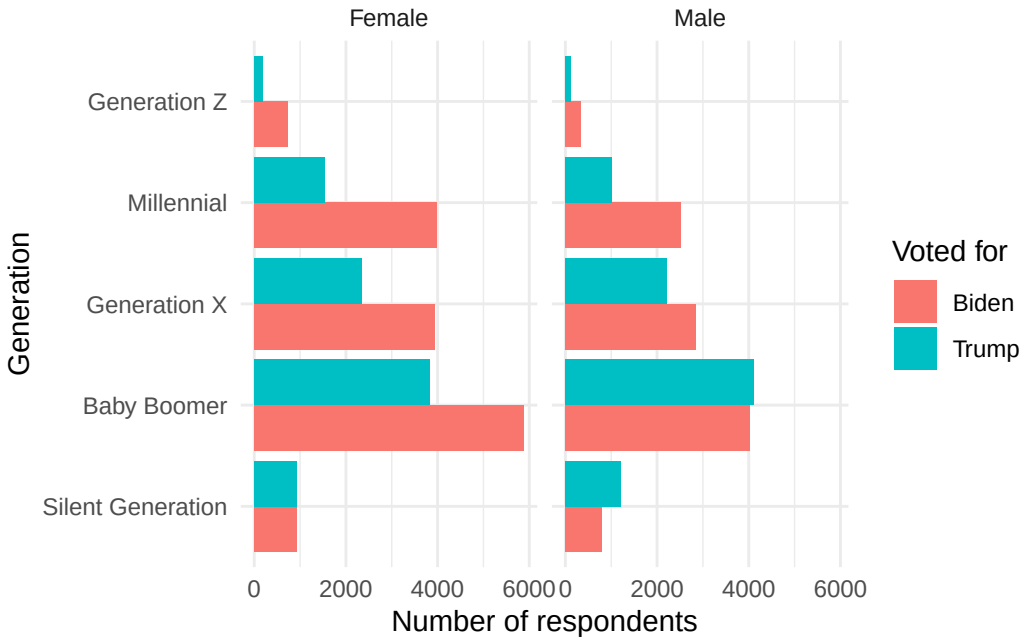

Figure 4: Distribution of presidential preferences, by gender, and by generation

Figure **??** Compares the political preferences between respondents of different generations, separated by gender. The voting behaviour on this histogram closely resembles most of our