My title*

My subtitle if needed

First author Another author

March 13, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

2 Data

2.1 Raw data.

The data used in this paper is derived from CCES in Havard DataVerse. This was the final version of the 2020 Cooperative Election Study Common Content dataset. All the data analysis was done through R (R Core Team 2023) with the aid of the following packages: …..

The raw data is published by Cooperative Election Study, a national stratified sample survey administered by YouGov. The data was gathered through two surveys: Pre-election (September 29 to November 2, 2020) and Post-election (November 8 to December 14, 2020). Each variable in the data set is constructed by the answer to one question inside the surveys. 600+ questions provide information on how Americans view Congress and hold their representatives accountable during elections, how they voted and their electoral experiences, and how their behavior and experiences vary with political geography and social context. The data included a large sample of 60,000+ representative Americans. The unit of observation is ‘ respondent’.

*Code and data are available at: LINK.

2.2 Cleaned data

Since the data constructed contained 600+ variables, only the demographics and social economics variables, which are gender, education level, and household income, were selected to analyze the effect of demographics on voters' preferences. To identify who the voters prefer in the 2020 election, it is important to select only the survey participants who registered to vote. Moreover, since the 2 most popular candidates in the 2020 presidential election are Trump, the representative candidate of the Republican Party, and Biden, the representative of the Democratic Party, only the participants who voted for either candidate will be selected for simplicity. Furthermore, some data points had missing attributes whereby an “NA” was put in place of the true value. Such entries were removed entirely in the data cleaning process as the number of observations was large and removing those entries won't have a significant impact on the outcome. The data now has 43,547 observations and 4 variables: candidates they voted for, gender, education level, and household income level. All 4 variables are factors variables and will further be explored below.

X	voted_for	gender	education
1	Trump	Male	2-year
2	Biden	Female	4-year
3	Biden	Female	4-year
4	Trump	Male	Some college
5	Trump	Female	Some college
6	Trump	Female	High school graduate

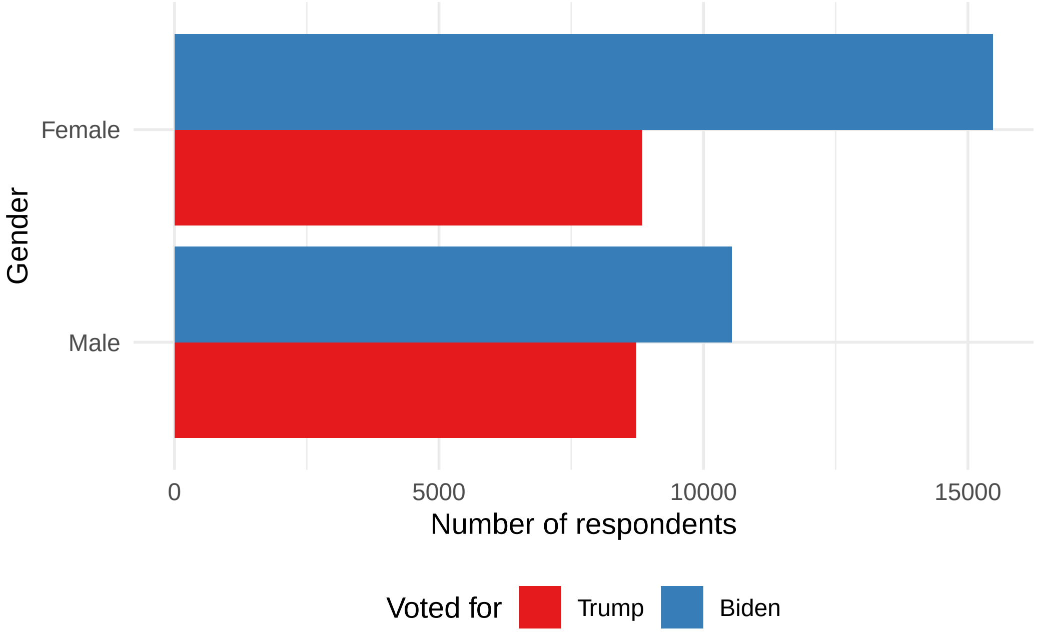
```
 voted_for         gender                         education    
 Trump:17558   Male  :19251   No HS               :  689  
 Biden:25996   Female:24303   High school graduate: 9814  
                              Some college        : 9290  
                              2-year              : 4971  
                              4-year              :11518  
                              Post-grad           : 7272  
```

2.2.1 Preference of candidates:

The variable “vote_ for” represents the candidate the survey participants voted for during the 2020 presidential elections. The participants either voted for Trump or Biden. (Table 2) shows summary statistics of the clean data. It shows that 17,555 survey participants voted for Trump which was approximately 40.3% of the voters while 25,992 people voted for Biden which was around 59.7% of voters.

2.2.2 Voters Demographics:

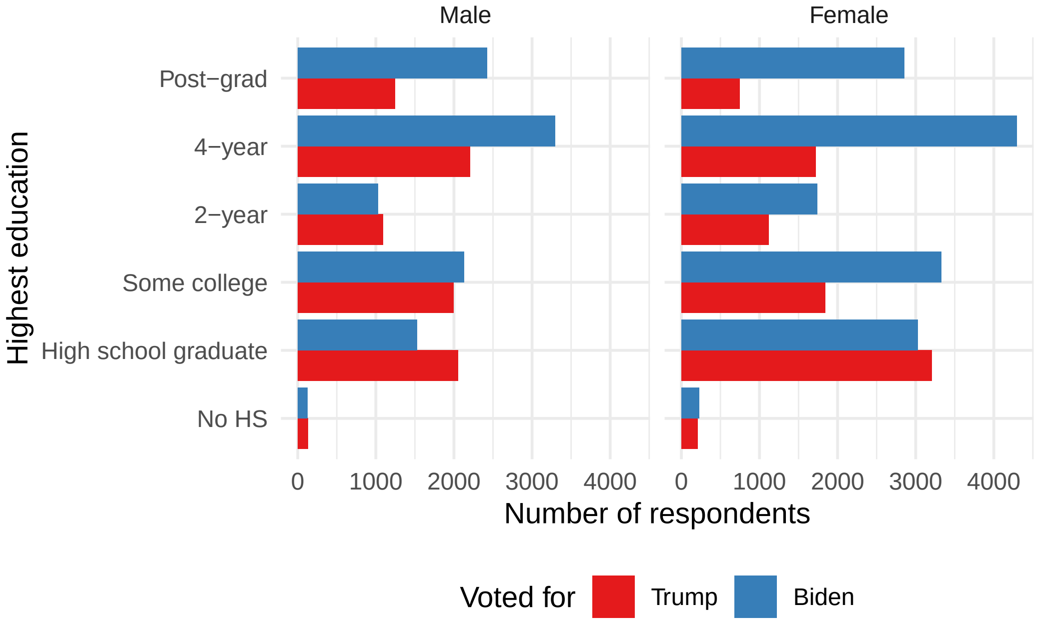
2.2.2.1 Gender.



The gender of survey participants will take values of either “Male” or “ Female”. Out of the 43,547 survey participants, 19,248 were male and 24,299 were female. (Graph 1) shows a bar graph of who respondents vote for, grouped by gender. It is shown that around 36.4% of female participants voted for Trump while 63.6% voted for Biden. For their male counterparts, while 45.3% of male participants voted for Trump, 54.7% of male respondents voted for Biden. These number indicates that female voters are more likely to vote for Biden than male voters. However, this evidence is inconclusive and needs further testing to conclude the relationship between voters’ choice of candidate and their gender.

2.2.2.2 Education level.

Participants’ education level is divided into 6 groups: No High School, High school graduates, some colleges, 2 years of college, 4 years of college, and post-grad. As shown in the summary statistics, there are 687 out of 43,547 respondents who have not finished high school, and 9810 participants are high school graduates. Additionally, 9290 respondents have some college education, and 4970 have finished 2 years of college. Out of 43,547 participants, the largest group consists of individuals with a 4-year college degree, totaling 11,518 respondents. Finally, the group with the highest level of education is post-grads with 7272 respondents.



Graph 2 is a bar graph illustrating the candidates respondents vote for grouped by gender and education level. Overall, it is found that for the most part, females with higher education levels are more likely to vote for Biden while there is no clear trend in male counterparts. It shows that 51.8% of male participants with no high school experience voted for Trump while only 48.2% voted for Biden. However, 47.5% of female respondents with no high school experience voted for Trump while 52.5% voted for Biden. The graph also demonstrates that 51.4% of female high school graduate participants voted for Trump while 57.3% of male high school graduate respondents voted for Trump. Interestingly, it is found that high school graduates, regardless of gender, are more likely to vote for Trump than Biden. Furthermore, Graph 2 shows that 51.5% of male participants with 2 years of college voted for Trump and 48.9% voted for Biden. However, 39.1% of female participants coming from the same educational background voted for Trump while 60.9% voted for Biden. Additionally, 35.6% of female respondents with some college experience voted for Trump while 48.3% of their male counterparts voted for Trump. The bar graphs also show that 28.6% of female respondents with 4-year college graduates voted for Trump while 40.1% of the male participants with the same educational level voted for Trump. Lastly, 34% of post-grads male participants voted for Trump while only 20.7% of the female counterparts voted for Trump. These statistics not only show that females are less likely to vote for Trump than males but also show that in general, but it also demonstrates that people with higher education levels are more likely to vote for Biden.

3 Model

The model used to analyze the relationship between the candidate respondents vote for and voters' demographics is logistics regression. Logistics regression provides a framework to analyze categorical outcome variables. Furthermore, logistics regression shows the probability of the occurrence of an event which is suitable for analyzing binary variables such as ones in our research. The model that we are interested in is (something similar but b3 will be the coefficient of the interaction terms)

The dependent variable y represents the political preference of the respondent and is equal to 1 if Biden and 0 if Trump, Gender_iis the gender of the respondent with 1 equal to female and 0 if male and education_i is the education of the respondent. Additionally, Gender_i*education_i, which is the interaction term of gender and education, was included to enhance the accuracy of the model and account for the conditional relationship between the two independent variables.

i is the probability that the ith respondent voted for Biden. Moreover, it is assumed that the distribution of coefficients B0, B1, and B2 are normal distributions with a mean of 0 and standard deviations of 2.5. This assumption is made as the model follows a Bayesian framework which allows us to incorporate prior information into the model. Assuming that the distribution of coefficients B0, B1, and B2 follows a normal distribution allows for a weakly informative prior which implies a neutrality on the possible value of the coefficients. Mean centering around 0 implies that there is no bias in the direction of all coefficients and a standard deviation of 2.5 allows for a moderate level of variation in the predictor variable. Furthermore, this assumption also prevents overfitting of the model by constraining the coefficients to reasonable values.

4 Result

	Support Biden
(Intercept)	−0.157
	(0.234)
genderMale	0.319
	(0.375)
education4-year	−1.072
	(0.309)
educationHigh school graduate	0.485
	(0.305)
educationNo HS	0.704
	(0.584)
educationPost-grad	−0.766
	(0.341)
educationSome college	−0.672
	(0.312)
genderMale × education4-year	0.452
	(0.457)
genderMale × educationHigh school graduate	−0.393
	(0.488)
genderMale × educationNo HS	30.129
	(25.941)
genderMale × educationPost-grad	0.191
	(0.503)
genderMale × educationSome college	0.309
	(0.475)
Num.Obs.	1000
R2	0.072
Log.Lik.	−642.912
ELPD	−654.0
ELPD s.e.	9.7
LOOIC	1308.0
LOOIC s.e.	19.3
WAIC	1307.9
RMSE	0.48