Final Written Report

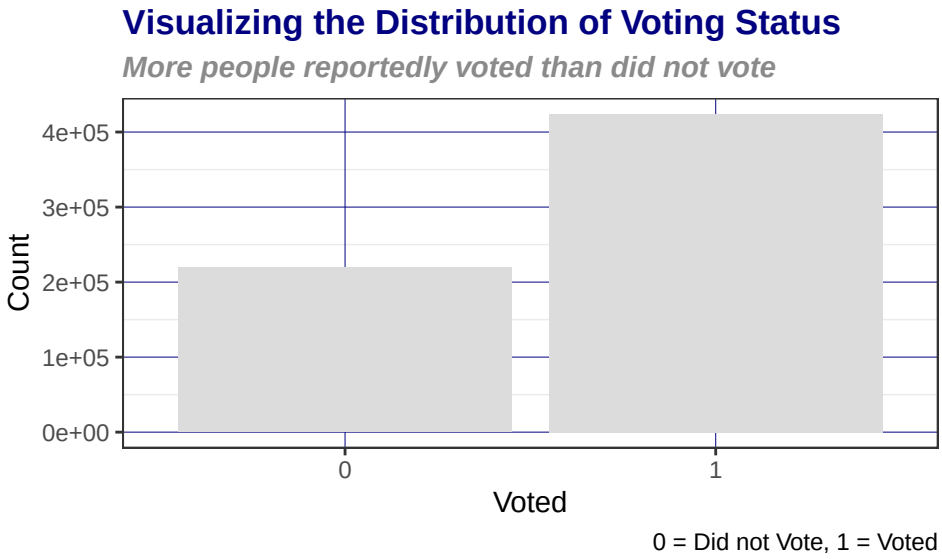
Approximately Normal: Jenny Huang, Bella Larsen, Jessie Bierschenk, Aidan Gildea

17 November 2020

Introduction

We will begin our EDA by visualizing the relationship between the response variable describing whether or not someone voted and several of the other variables of particular interest.

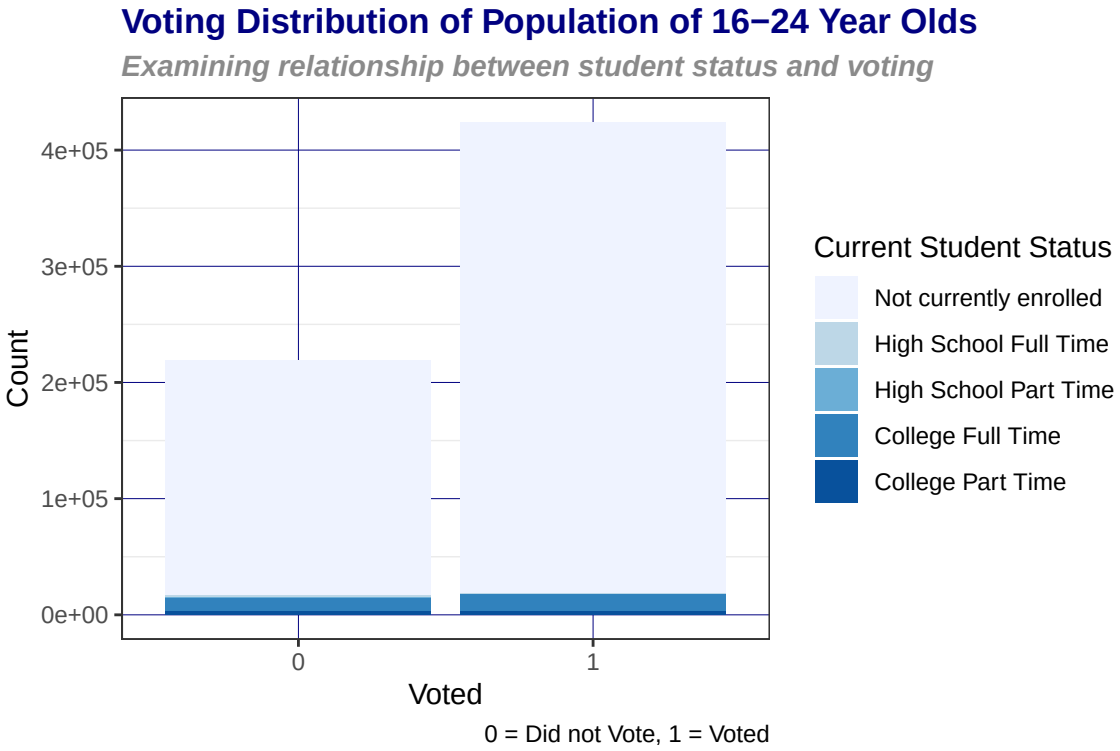
We will begin by simply looking at the distribution of those who voted throughout the last 8 years of elections.



see theme code inspiration at reference [1] see scale fill code inspiration at reference [2]

From the barplot above, it is clear the more individuals in the data set voted (voted = 1) than did not (voted = 0).

As college students ourselves, we want to analyze whether or not being a student influences the frequency of voting. We will explore this preliminarily by visualizing the distribution of if school aged individuals (18-24) voted or not – categorized by their current student level. This is seen in the bar plot below.



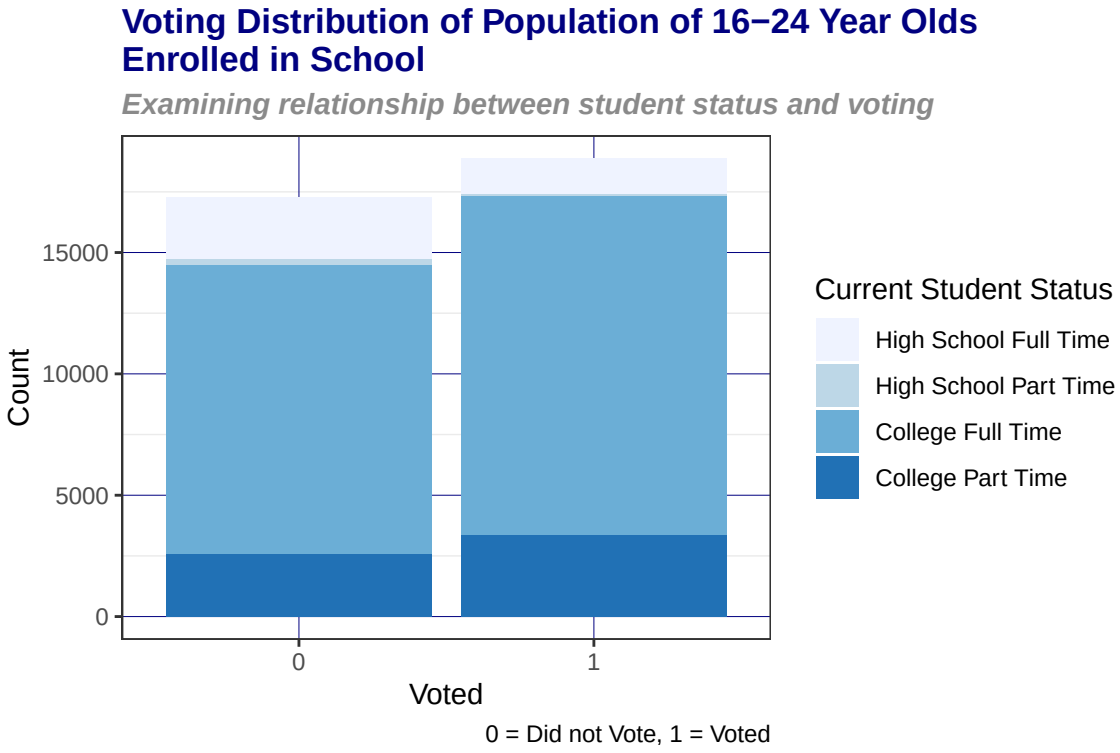
see scale fill brewer code inspiration from reference [1]

From the bar plot, it is evident that a majority of these individuals were not currently enrolled. This may be a result of a general national trend, but we want to investigate if it is the result of a larger proportion of older individuals within in the range of ages between 16-24. We will investigate this by analyzing those who are not currently enrolled in school within this age range.

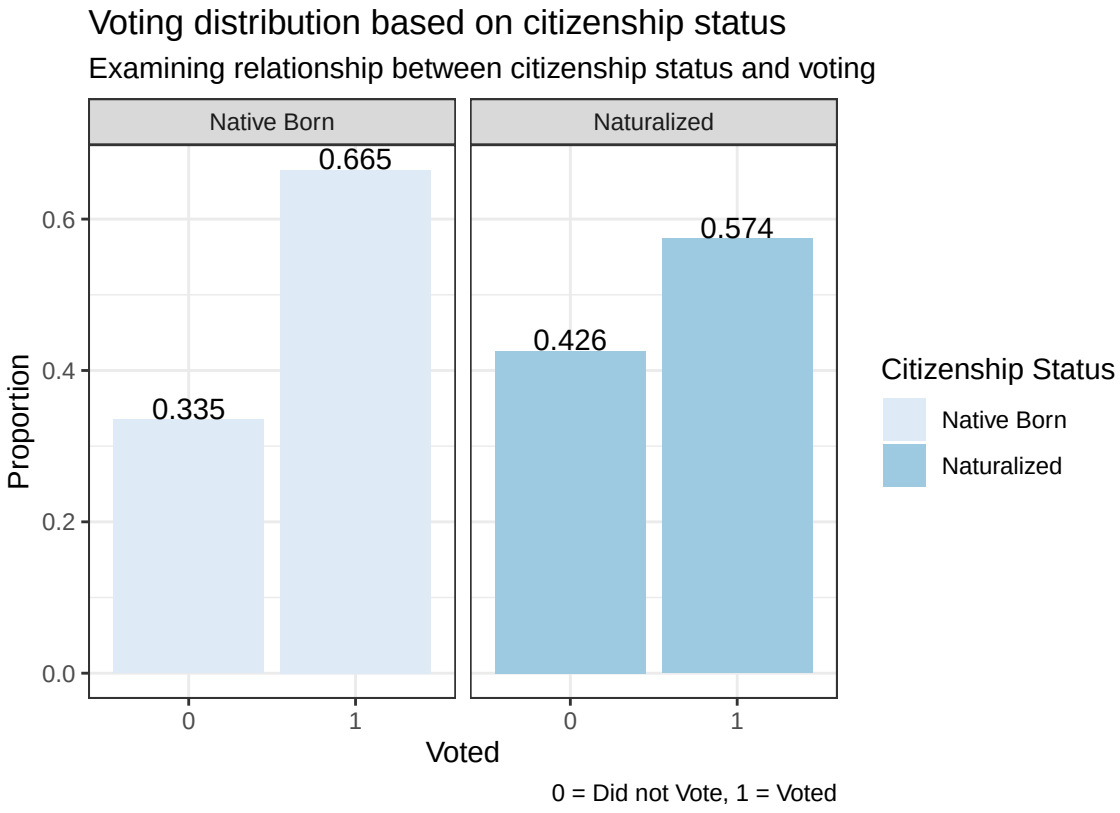
AGE	n	prop
18	2327	0.065
19	3671	0.102
20	4367	0.122
21	4760	0.133
22	5955	0.166
23	6991	0.195
24	7791	0.217

From the table above, it is apparent that more than 40% of those not currently enrolled in school are 23-24 years old. This could be a potential reason for why this age range includes so many who are not currently enrolled as a student.

To more meaningfully analyze the relationship between being a student and if they vote or not, we adjusted our visualization to only include those currently enrolled in some level of education. This is seen in the visualization below.

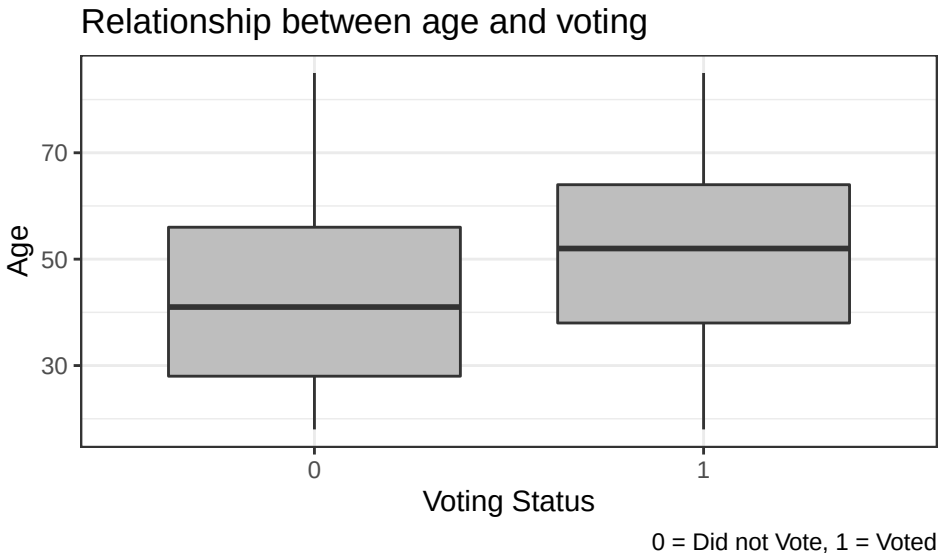


This visualization shows that eligible voters between the age 16-24 enrolled in some education at the time were mainly full time college students. More full time and part time college students that were eligible to vote did vote compared to those that did not vote. The opposite is true for high school students: more full time and part time high school students that were eligible to vote did not vote compared to those that did vote.

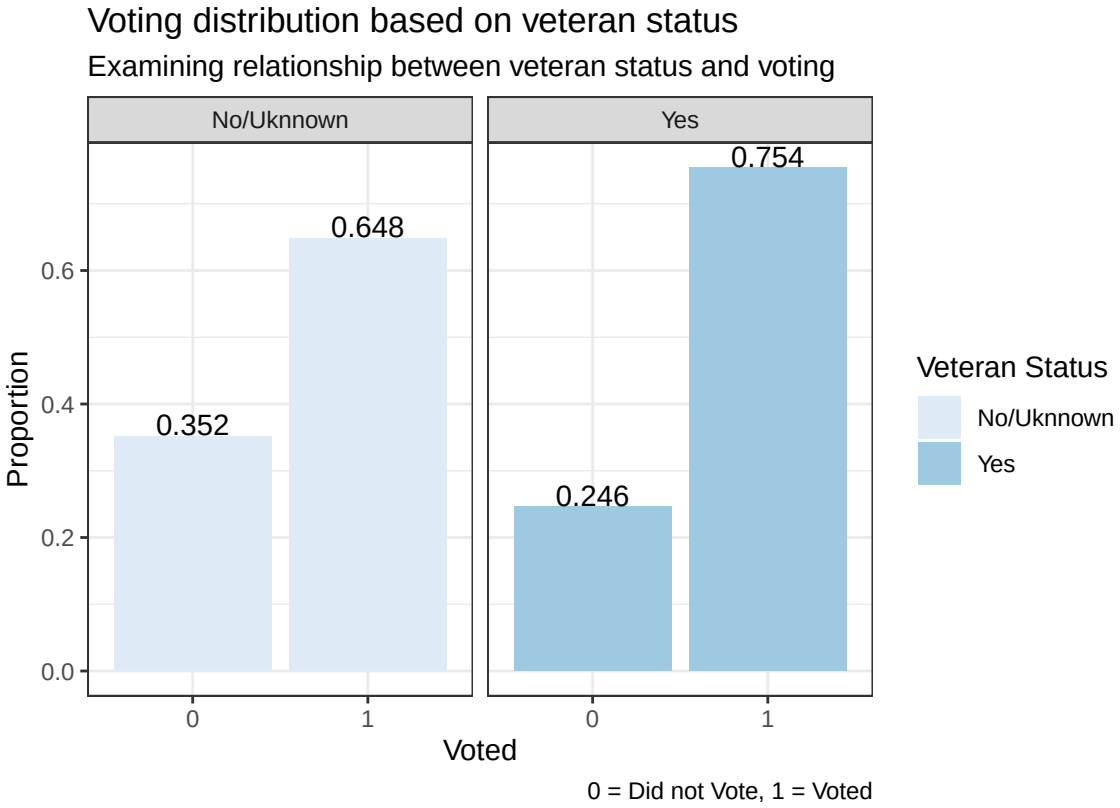


see geom_text() code inspiration in reference [4]

This visualization shows that for both native born and naturalized individuals, more citizens that were eligible to vote did vote compared to those that did not vote; however, the proportion is much greater for native born citizens than for naturalized.



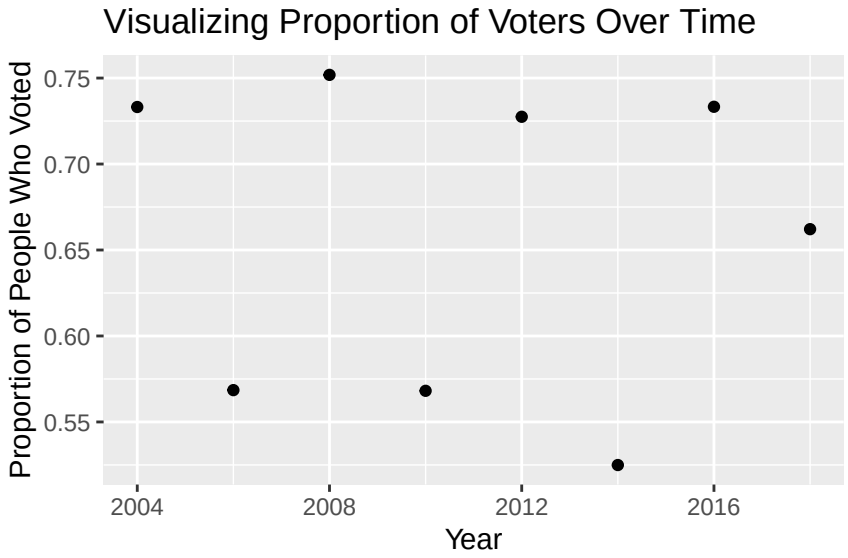
This box plot shows that eligible voters that did vote were generally older than eligible voters that did not vote.



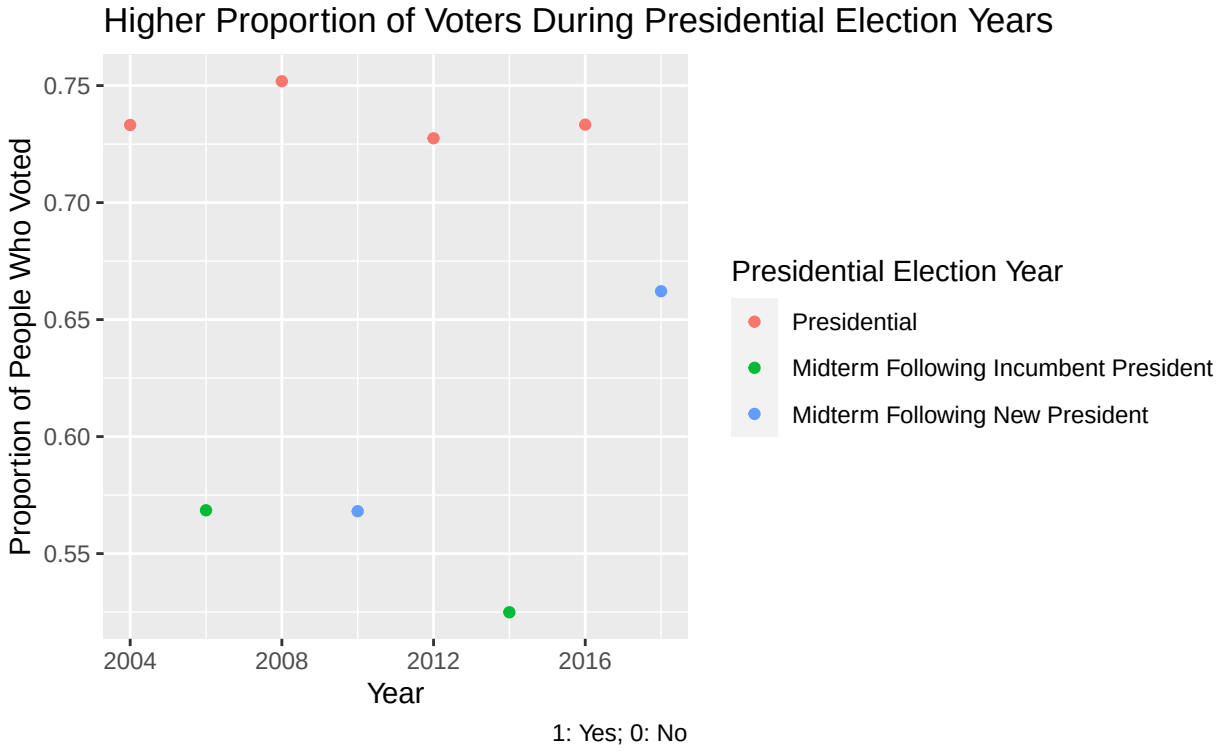
see geom_text() code inspiration in reference [4]

We are also interested in looking at how voter turnout has changed over the years.

We notice that the proportion of people who voted fluctuates depending on whether the year falls on a presidential election. In the trend of the proportion of voting over time, we see a clear divide between the years when there is a presidential elections versus when there is not. In the future, we may decide to add the variable "Election Year" as an interaction term with year as a divide between years that fall on an election.



We decided to look at the scatterplot of voter turnout over time broken down by the status of the election (Presidential election, midterm after incumbent president, or midterm after new president) to see how it may differ depending on the election.



From the above scatterplot, we confirmed that there is higher voter turnout during presidential elections compared to midterm elections. In addition, there is equal or higher voter turnout for midterm elections following the election of a new president compared to midterm elections following the election of an incumbent president, and an especially high voter turnout in the midterm election after Trump's election in 2016.

Finally, it is important to acknowledge that our any missingness in our data, as it may influence the outcome of our regression analysis.

Model Selection

##Select a random subset of the data to create model.

In order to make our model, we have decided to take a random sample of 10,000 to be sure that the model selection is accurate.

term	estimate	std.error	statistic	p.value
(Intercept)	16.369	11.317	1.446	0.148
sexMale	-0.078	0.048	-1.628	0.104
marstDivorced/Separated	-0.702	0.071	-9.924	0.000
marstNot Married/Other	-0.474	0.057	-8.336	0.000
citizenNaturalized	-0.684	0.103	-6.637	0.000
employedYes	0.406	0.060	6.768	0.000
highest_educationHigh School Degree/GED	-1.357	0.074	-18.430	0.000
highest_educationSome College	-0.630	0.081	-7.744	0.000
highest_educationSome High School	-2.220	0.099	-22.398	0.000
highest_educationAssociate Degree	-0.524	0.095	-5.494	0.000
highest_educationMasters Degree	0.287	0.125	2.304	0.021
highest_educationProfessional Degree	0.487	0.282	1.727	0.084
highest_educationDoctorate Degree	0.403	0.275	1.464	0.143
highest_educationNone/Unknown	-3.161	0.653	-4.838	0.000
current_studentHigh School Full Time	1.177	0.305	3.865	0.000
current_studentHigh School Part Time	0.299	1.134	0.264	0.792
current_studentCollege Full Time	0.444	0.123	3.626	0.000
current_studentCollege Part Time	0.573	0.248	2.312	0.021
raceBlack	0.599	0.084	7.142	0.000
raceAsian or Pacific Islander	-0.372	0.133	-2.791	0.005
raceNative American	-0.060	0.202	-0.295	0.768
race2 or more races	0.258	0.205	1.256	0.209
AGE	0.040	0.002	22.706	0.000
Presidential_Election_StatusMidterm Following Incumbent President	-0.989	0.057	-17.347	0.000
Presidential_Election_StatusMidterm Following New President	-0.669	0.063	-10.696	0.000
YEAR	-0.008	0.006	-1.440	0.150

The final model included YEAR + sex + current_student + citizen + employed + race + marst + Presidential_Election_Status + AGE + highest_education.

The final model is:

The backward selection based on AIC took out the variables metro, veteran, and hispanic_status.

Interpretation of coefficients of interest:

Baseline: Female, married, native born, not employed, Bachelors Degree is highest education, not currently enrolled in school if between the age of 16-24, white, and the time of voting is midterm following the election of an incumbent president.

We expect the odds of an eligible voter voting for a divorced/separated eligible voter to be 0.50 (exp(-0.702)) times the odds of individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, white, and the election

is a midterm election following the election of an incumbent president, after factoring in all other voter characteristics/information.

For each additional year in age, we expect the odds of an eligible voter voting to be 1.04 (exp(0.040)) times the odds of individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, white, and the election is a midterm election following the election of an incumbent president, after factoring in all other voter characteristics/information.

We expect the odds of an eligible voter voting when it is a presidential election year to be 2.689 (exp(0.989)) times the odds of individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, white, and the election is a midterm election following the election of an incumbent president, after factoring in all other voter characteristics/information.

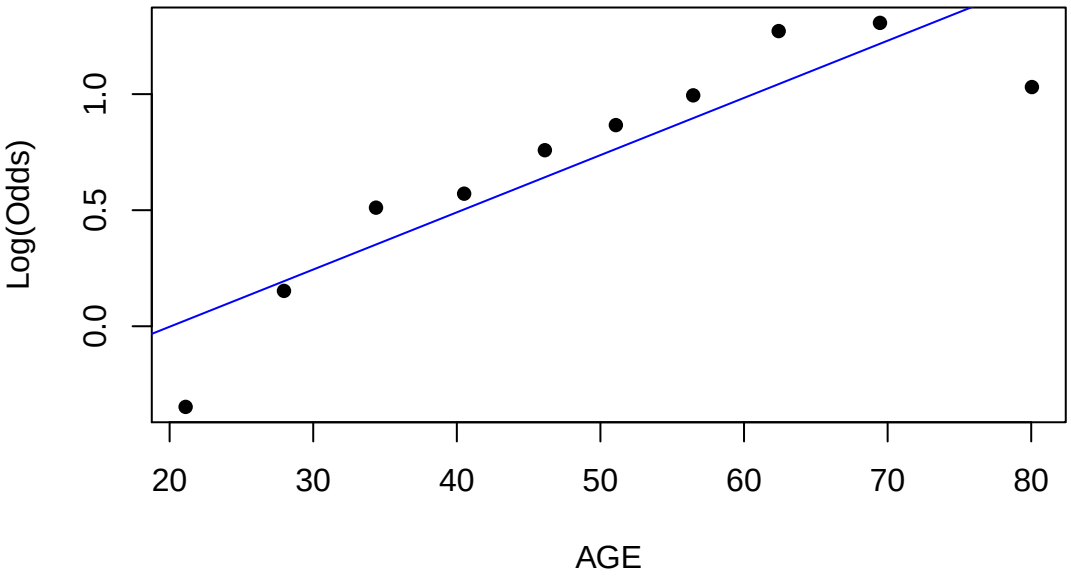
Checking Model Conditions

#Linearity

sex	voted	n	prop	emp_logit
Female	1	3582	0.6704099	0.7100395
Male	1	3029	0.6504187	0.6208803

marst	voted	n	prop	emp_logit
Married	1	4145	0.7346686	1.0184397
Divorced/Separated	1	783	0.5878378	0.3550343
Not Married/Other	1	1683	0.5561798	0.2256720

According to the plot below, there is a linear relationship between the empirical logit and the predictor variable age. Hence linearity is satisfied for AGE.



Randomness

It is possible that randomness is not satisfied because our data is from the census survey, which may not be random (ie might select for people who have time to fill it out). However, there is no reason to believe that this will not generalize to the US population as a whole in a significant way, particularly due to the large sample size.

Independence

Independence may be violated because geographic location may influence voting due to factors such as (residuals by state ID). Hence, we will look at misclassification rate by region.

Here, we create a confusion matrix with a misclassification rate of 0.5.

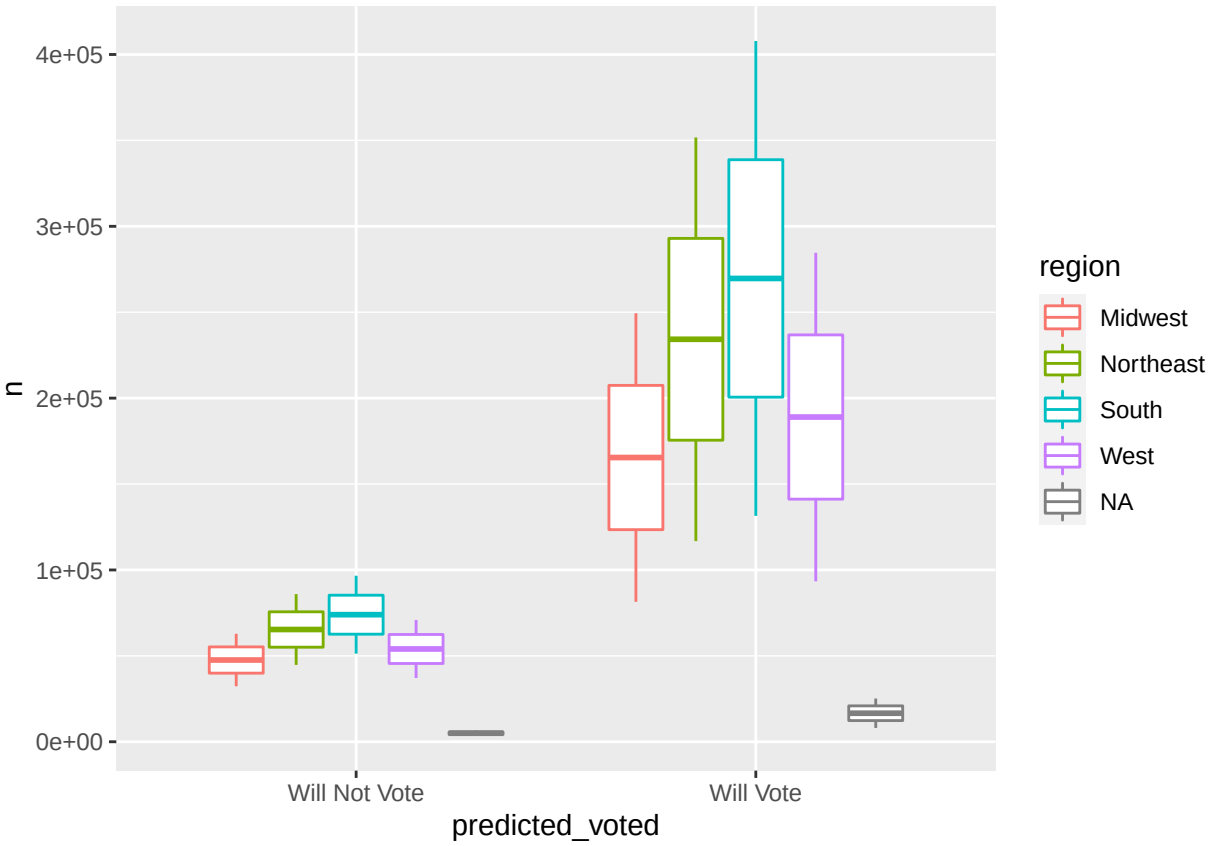
We will left join with region to look at misclassification rates by region.

Below are the misclassification rates by region.

region	voted	predicted_voted	n	prop
Midwest	0	Will Vote	81460	0.1912058
Midwest	1	Will Not Vote	32314	0.0758486
Northeast	0	Will Vote	116798	0.1949320
Northeast	1	Will Not Vote	44736	0.0746629
South	0	Will Vote	131479	0.1913145
South	1	Will Not Vote	51284	0.0746231
West	0	Will Vote	93400	0.1921834
West	1	Will Not Vote	37109	0.0763569
NA	0	Will Vote	8040	0.1862793
NA	1	Will Not Vote	3433	0.0795394

Consulted Census data for the region fips number corresponding to region name [3]

We plan to create a plot of misclassification rate by region to determine if independence is satisfied.

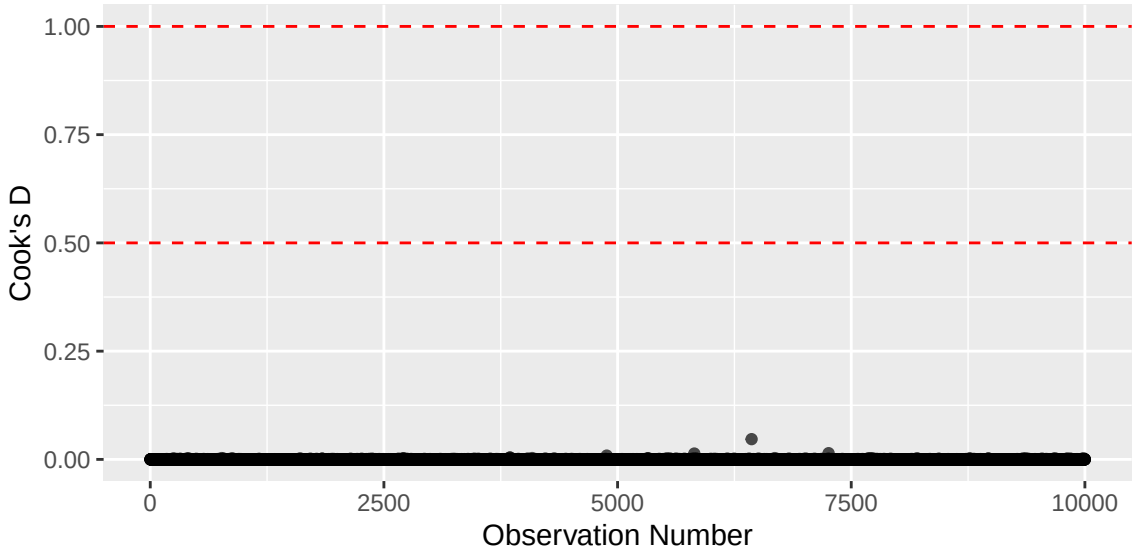


Based on the misclassification rates, we have no reason to believe that the independence condition is not satisfied. The misclassification rates across regions are relatively similar, which suggests that there is not an issue of spatial correlation and that the independence condition is satisfied.

Checking for influential points

Cook's distance

We will also look for influential points using Cook's Distance.



According to Cook's Distance, there are no influential points, so all points can be left in the model.

Multicollinearity

VIF:

names	x
current_studentHigh School Part Time	1.006
raceNative American	1.009
race2 or more races	1.009
highest_educationNone/Unknown	1.016
sexMale	1.023
current_studentCollege Part Time	1.038
highest_educationProfessional Degree	1.043
raceBlack	1.049
highest_educationDoctorate Degree	1.050
current_studentHigh School Full Time	1.086
marstDivorced/Separated	1.116
YEAR	1.168
Presidential_Election_StatusMidterm Following Incumbent President	1.176
raceAsian or Pacific Islander	1.206
citizenNaturalized	1.214
highest_educationMasters Degree	1.246
marstNot Married/Other	1.286
current_studentCollege Full Time	1.303
Presidential_Election_StatusMidterm Following New President	1.311
employedYes	1.433
highest_educationAssociate Degree	1.508
AGE	1.709
highest_educationSome High School	1.744
highest_educationSome College	1.950
highest_educationHigh School Degree/GED	2.216

All of the VIF values are under the threshold of 10, indicating that there is no evidence of multicollinearity in our data.

Interaction Terms

We will add in several interaction terms of interest to us and use a drop-in-deviance test to see if they are meaningful predictors of the odds of someone voting.

The following hypotheses will be used:

H_0: coefficients for the interaction between sex and employment, presidential election status and highest education level, age and presidential election status, sex and presidential election status, and race and presidential election status are all zero

$H_0 : \beta_{sex*employed} = \beta_{Presidential_Election_Status*highest_education} = \beta_{AGE*Presidential_Election_Status} = \beta_{sex*Presidential_Election_Status}$

H_a: at least one of these coefficients for the interaction terms $\neq$ zero

$\alpha = 0.05$

term	estimate	std.error	statistic	p.value
(Intercept)	15.949	11.375	1.402	0.161
YEAR	-0.008	0.006	-1.357	0.175
sexMale	-0.022	0.071	-0.308	0.758
current_studentHigh School Full Time	1.122	0.307	3.654	0.000
current_studentHigh School Part Time	0.367	1.140	0.322	0.747
current_studentCollege Full Time	0.464	0.124	3.738	0.000
current_studentCollege Part Time	0.647	0.251	2.580	0.010
citizenNaturalized	-0.693	0.103	-6.701	0.000
employedYes	0.406	0.060	6.731	0.000
raceBlack	0.876	0.133	6.614	0.000
raceAsian or Pacific Islander	-0.719	0.190	-3.785	0.000
raceNative American	-0.323	0.284	-1.136	0.256
race2 or more races	-0.186	0.278	-0.670	0.503
marstDivorced/Separated	-0.715	0.071	-10.071	0.000
marstNot Married/Other	-0.489	0.057	-8.555	0.000
Presidential_Election_StatusMidterm Following Incumbent President	-1.787	0.216	-8.276	0.000
Presidential_Election_StatusMidterm Following New President	-1.478	0.221	-6.692	0.000
AGE	0.033	0.002	14.696	0.000
highest_educationHigh School Degree/GED	-1.588	0.118	-13.491	0.000
highest_educationSome College	-0.836	0.129	-6.491	0.000
highest_educationSome High School	-2.444	0.144	-16.972	0.000
highest_educationAssociate Degree	-0.603	0.153	-3.930	0.000
highest_educationMasters Degree	0.648	0.237	2.737	0.006
highest_educationProfessional Degree	1.228	0.613	2.003	0.045
highest_educationDoctorate Degree	0.107	0.458	0.233	0.815
highest_educationNone/Unknown	-3.324	0.806	-4.125	0.000
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationHigh School Degree/GED	0.450	0.178	2.525	0.012
Presidential_Election_StatusMidterm Following New President:highest_educationHigh School Degree/GED	0.316	0.182	1.735	0.083
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationSome College	0.292	0.193	1.513	0.130
Presidential_Election_StatusMidterm Following New President:highest_educationSome College	0.384	0.198	1.944	0.052
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationSome High School	0.474	0.235	2.019	0.043

term	estimate	std.error	statistic	p.value
Presidential_Election_StatusMidterm Following New President:highest_educationSome High School	0.341	0.244	1.400	0.162
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationAssociate Degree	0.077	0.229	0.335	0.738
Presidential_Election_StatusMidterm Following New President:highest_educationAssociate Degree	0.171	0.238	0.721	0.471
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationMasters Degree	-0.638	0.316	-2.022	0.043
Presidential_Election_StatusMidterm Following New President:highest_educationMasters Degree	-0.389	0.327	-1.190	0.234
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationProfessional Degree	-1.086	0.752	-1.445	0.148
Presidential_Election_StatusMidterm Following New President:highest_educationProfessional Degree	-0.972	0.800	-1.214	0.225
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationDoctorate Degree	1.014	0.721	1.406	0.160
Presidential_Election_StatusMidterm Following New President:highest_educationDoctorate Degree	0.024	0.633	0.038	0.970
Presidential_Election_StatusMidterm Following Incumbent President:highest_educationNone/Unknown	-10.012	179.645	-0.056	0.956
Presidential_Election_StatusMidterm Following New President:highest_educationNone/Unknown	1.187	1.436	0.827	0.408
Presidential_Election_StatusMidterm Following Incumbent President:AGE	0.013	0.003	3.833	0.000
Presidential_Election_StatusMidterm Following New President:AGE	0.013	0.004	3.612	0.000
sexMale:Presidential_Election_StatusMidterm Following Incumbent President	-0.121	0.114	-1.065	0.287
sexMale:Presidential_Election_StatusMidterm Following New President	-0.089	0.119	-0.749	0.454
raceBlack:Presidential_Election_StatusMidterm Following Incumbent President	-0.505	0.204	-2.475	0.013
raceAsian or Pacific Islander:Presidential_Election_StatusMidterm Following Incumbent President	0.346	0.302	1.148	0.251
raceNative American:Presidential_Election_StatusMidterm Following Incumbent President	0.286	0.497	0.576	0.565
race2 or more races:Presidential_Election_StatusMidterm Following Incumbent President	0.745	0.507	1.471	0.141
raceBlack:Presidential_Election_StatusMidterm Following New President	-0.431	0.203	-2.120	0.034
raceAsian or Pacific Islander:Presidential_Election_StatusMidterm Following New President	0.883	0.301	2.930	0.003
raceNative American:Presidential_Election_StatusMidterm Following New President	0.698	0.492	1.418	0.156
race2 or more races:Presidential_Election_StatusMidterm Following New President	1.105	0.513	2.155	0.031

Resid..Df	Resid..Dev	df	Deviance	p.value
9974	10673.41	NA	NA	NA
9946	10592.36	28	81.054	0

The p-value (7.840e-07) is very small (less than the alpha level 0.05), so we can reject the null hypothesis. Thus, we conclude that the data provide sufficient evidence that the coefficients associated with the additional interaction terms are not equal to 0. Therefore, we should add them to the model.

Significant interaction terms: The effect of the election being a presidential election for an individual with the highest level of education as a High School Degree/GED is significant with a p-value of 0.011 assuming an alpha level of 0.05. The coefficient is negative, indicating that this effect would mean that during a presidential election for an individual with the highest level of education as a High School Degree/GED, they are less likely to vote than individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, white, and the election is a midterm election following the election of an incumbent president.

The effect of the election being a presidential election for an individual that is Black is significant with a p-value of 0.013 assuming an alpha level of 0.05. The coefficient is positive, indicating that this effect would mean that during a presidential election for an individual that is Black, they are more likely to vote than individuals that are female, married, native born, not employed, have Bachelors Degree as their highest education, not currently enrolled in school if between the age of 16-24, white, and the election is a midterm election following the election of an incumbent president

We will use a drop-in-deviance test to determine whether or not sex is a meaningful variable in the model.

Resid..Df	Resid..Dev	df	Deviance	p.value
9975	10676.06	NA	NA	NA
9974	10673.41	1	2.649	0.104

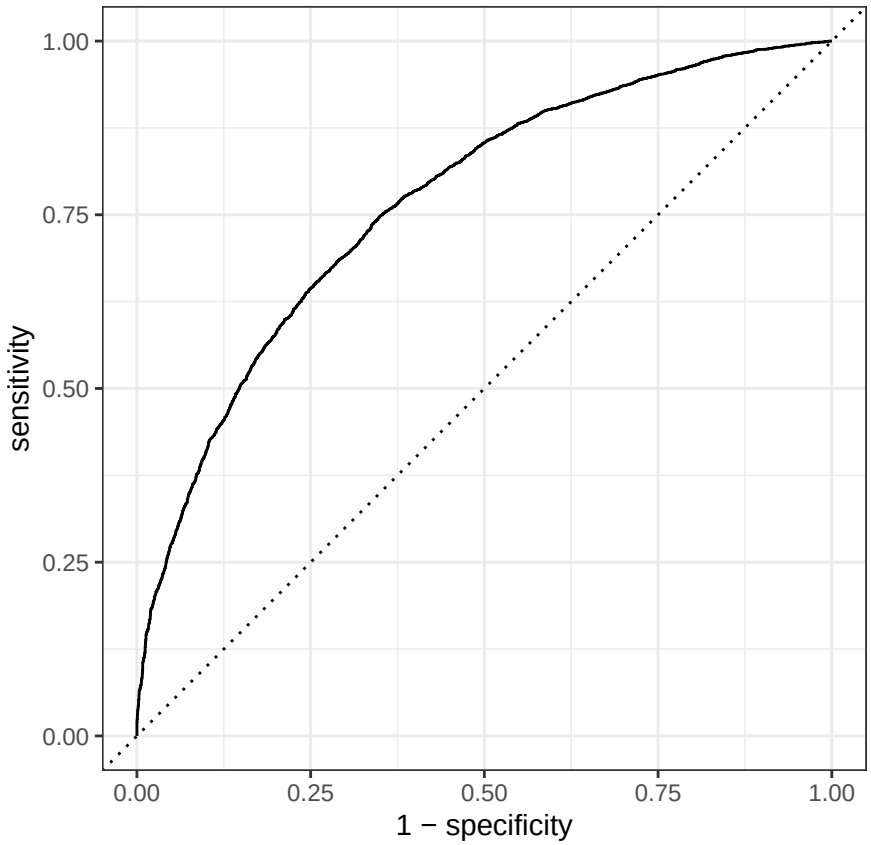
From the drop-in-deviance test to include variable sex, the p-value is greater than an alpha-level of 0.10, so we will exclude sex from the model. This contradicts the model output from the backward selection, which is most likely due their different criteria for statistical significance. Backward selection makes decisions based on the AIC, while the drop-in-deviance test uses the p-value. A potential discrepancy between the two values may have resulted in the different determinations of the significance of sex. We have chosen to use the results of the drop-in-deviance test, and therefore, will be excluding sex from the final model.

term	estimate	std.error	statistic	p.value
(Intercept)	16.528	11.313	1.461	0.144
YEAR	-0.008	0.006	-1.457	0.145
current_studentHigh School Full Time	1.174	0.305	3.854	0.000
current_studentHigh School Part Time	0.325	1.133	0.287	0.774
current_studentCollege Full Time	0.440	0.122	3.595	0.000
current_studentCollege Part Time	0.584	0.247	2.359	0.018
citizenNaturalized	-0.682	0.103	-6.613	0.000
employedYes	0.396	0.060	6.639	0.000
raceBlack	0.601	0.084	7.167	0.000
raceAsian or Pacific Islander	-0.375	0.133	-2.812	0.005
raceNative American	-0.056	0.201	-0.280	0.780
race2 or more races	0.256	0.205	1.247	0.212
marstDivorced/Separated	-0.698	0.071	-9.869	0.000
marstNot Married/Other	-0.469	0.057	-8.270	0.000
Presidential_Election_StatusMidterm Following Incumbent President	-0.988	0.057	-17.338	0.000
Presidential_Election_StatusMidterm Following New President	-0.670	0.063	-10.706	0.000
AGE	0.040	0.002	22.726	0.000
highest_educationHigh School Degree/GED	-1.361	0.074	-18.494	0.000
highest_educationSome College	-0.632	0.081	-7.776	0.000

term	estimate	std.error	statistic	p.value
highest_educationSome High School	-2.229	0.099	-22.513	0.000
highest_educationAssociate Degree	-0.523	0.095	-5.489	0.000
highest_educationMasters Degree	0.292	0.125	2.339	0.019
highest_educationProfessional Degree	0.487	0.282	1.727	0.084
highest_educationDoctorate Degree	0.395	0.275	1.433	0.152
highest_educationNone/Unknown	-3.165	0.653	-4.849	0.000

##Creating Classifier

We'll fit an ROC curve to help us determine a decision-making threshold.



Below we look at values within a range of thresholds in order to choose the threshold that max specificity and and sensitivity.

```
## # A tibble: 790 x 4
##    .threshold specificity sensitivity pred_prob
##         <dbl>       <dbl>       <dbl>     <dbl>
##  1      0.300       0.177       0.972     0.574
##  2      0.300       0.177       0.972     0.574
##  3      0.300       0.178       0.972     0.575
##  4      0.301       0.178       0.972     0.575
##  5      0.301       0.178       0.972     0.575
##  6      0.301       0.178       0.972     0.575
##  7      0.301       0.178       0.971     0.575
##  8      0.302       0.179       0.971     0.575
##  9      0.302       0.179       0.971     0.575
## 10      0.302       0.179       0.971     0.575
```

```
## # ... with 780 more rows
```

According to our ROC curve and modeling objectives, we will choose a threshold of .42 because we want to lean towards having a higher false negative rate (type II error) than a false positive rate as it does not hurt to mail a few extra ballots to people who may already be planning to vote.

Any data point with a probability over 0.60 will be predicted to be in the "voted" category.

Discussion

[1] http://www.sthda.com/english/articles/32-r-graphics-essentials/125-ggplot-cheat-sheet-for-great-customization/#use-themes-in-ggplot2-package [2] https://www.datanovia.com/en/blog/the-a-z-of-rcolorbrewer-palette/ [3] https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf [4] https://stackoverflow.com/questions/12018499/how-to-put-labels-over-geom-bar-for-each-bar-in-r-with-ggplot2

keep in mind: citizenship and registration exclusion for the model