Written Report Draft

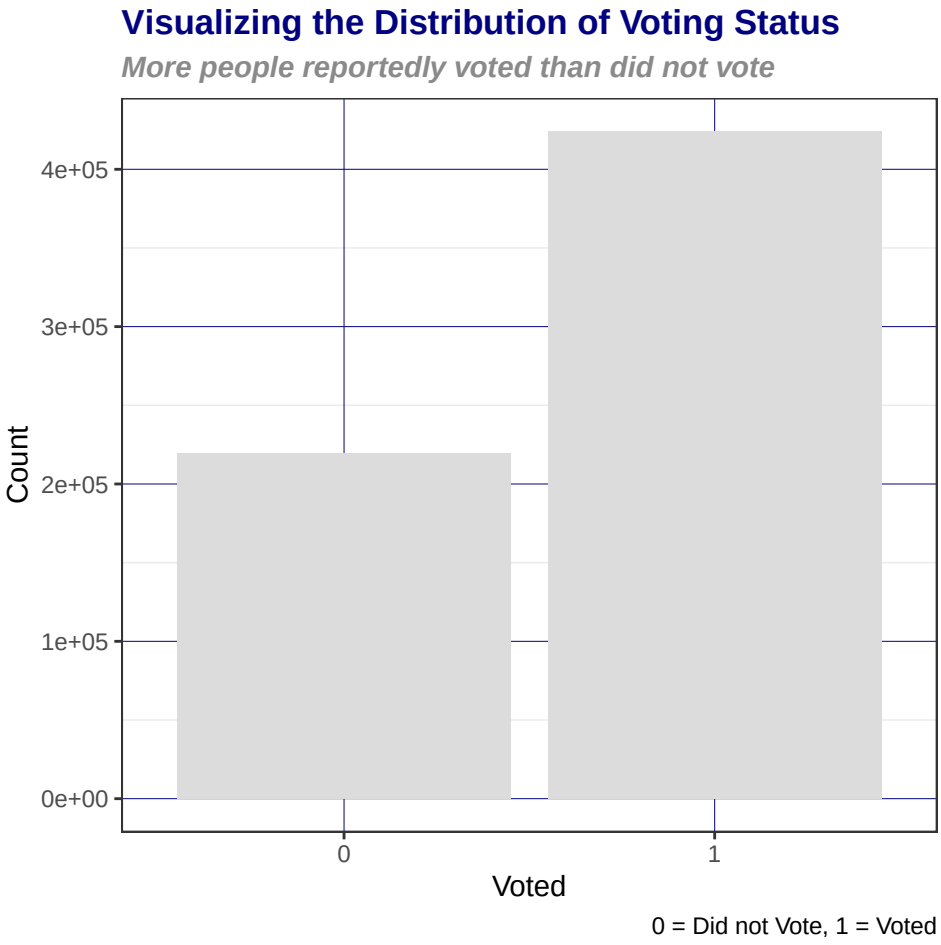
Approximately Normal: Jenny Huang, Bella Larsen, Jessie Bierschenk, Aidan Gildea

28 October 2020

Introduction

We will begin our EDA by visualizing the relationship between the response variable `voted` and several of the other variables of particular interest.

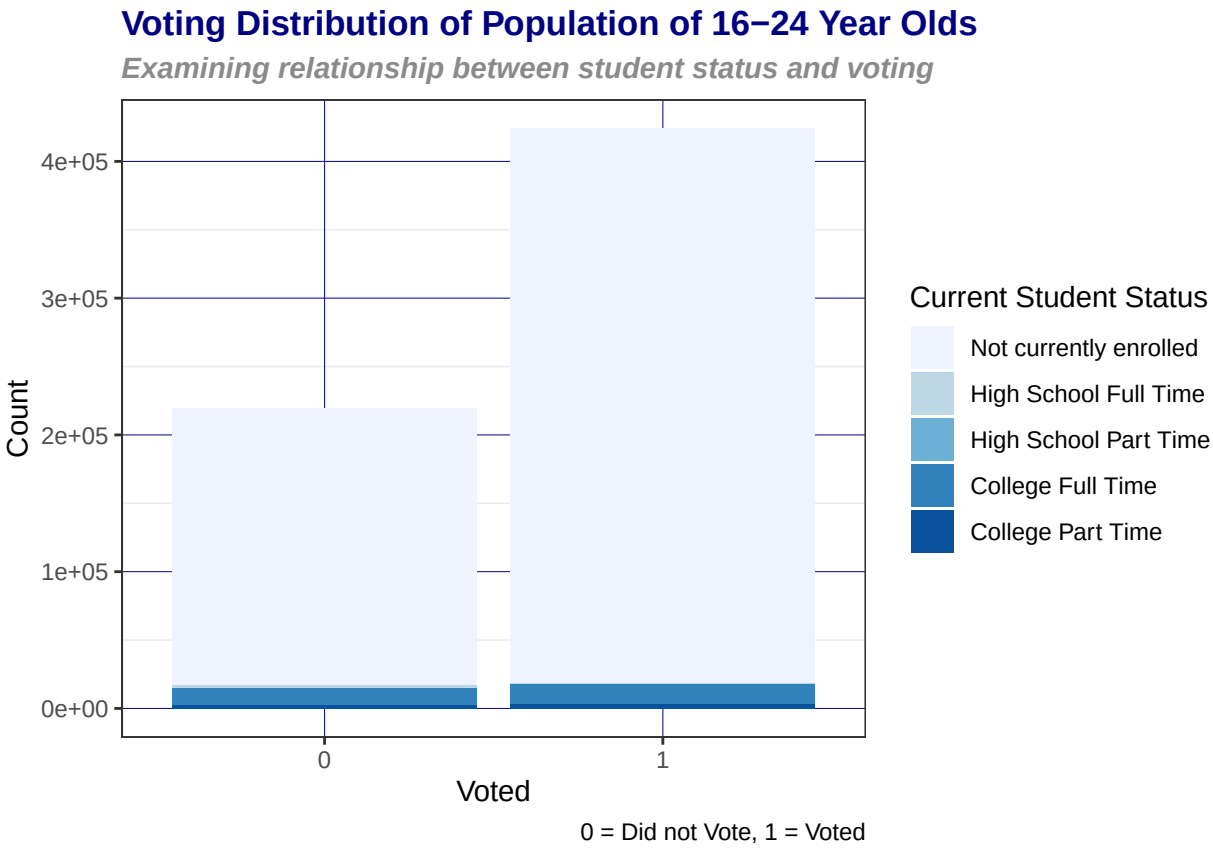
We will begin by simply looking at the distribution of those who voted throughout the last 8 years of elections.



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.



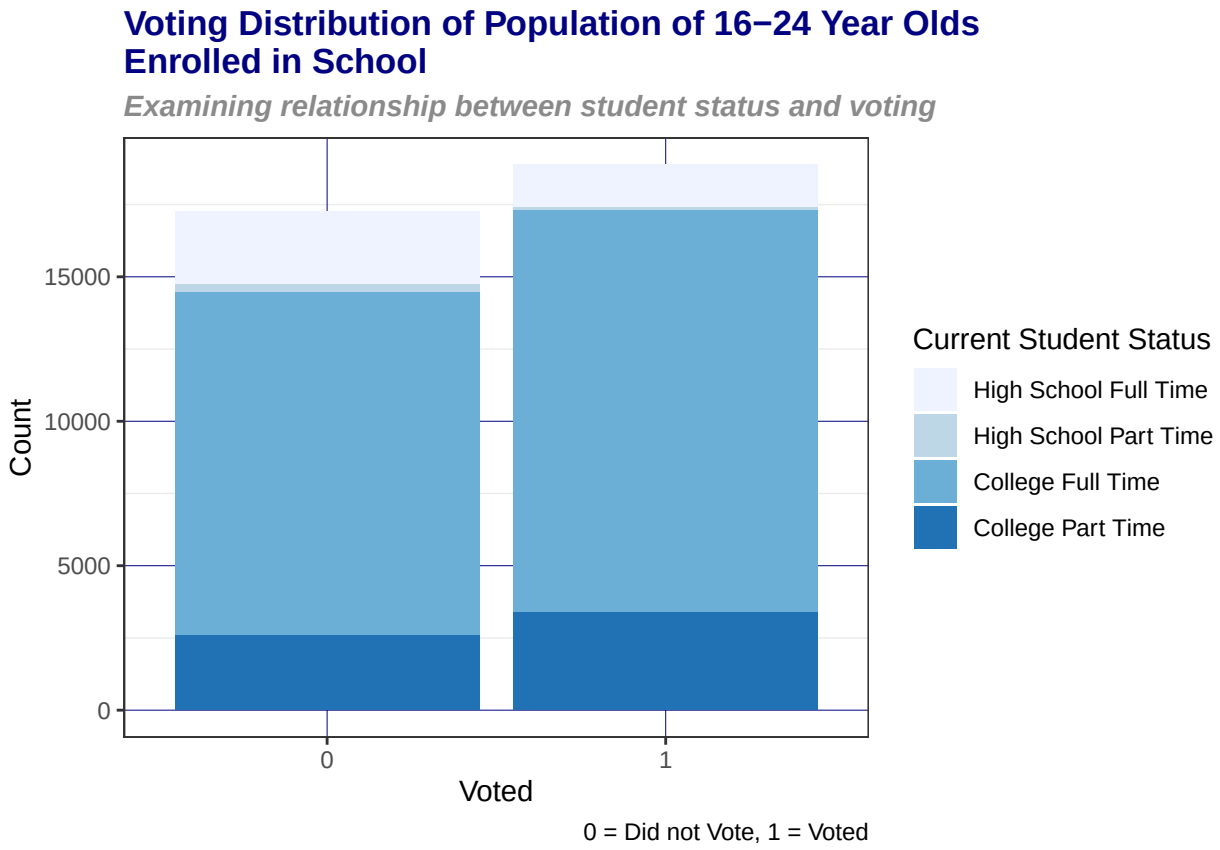
see scale fill brewer code inspiration from reference [1]

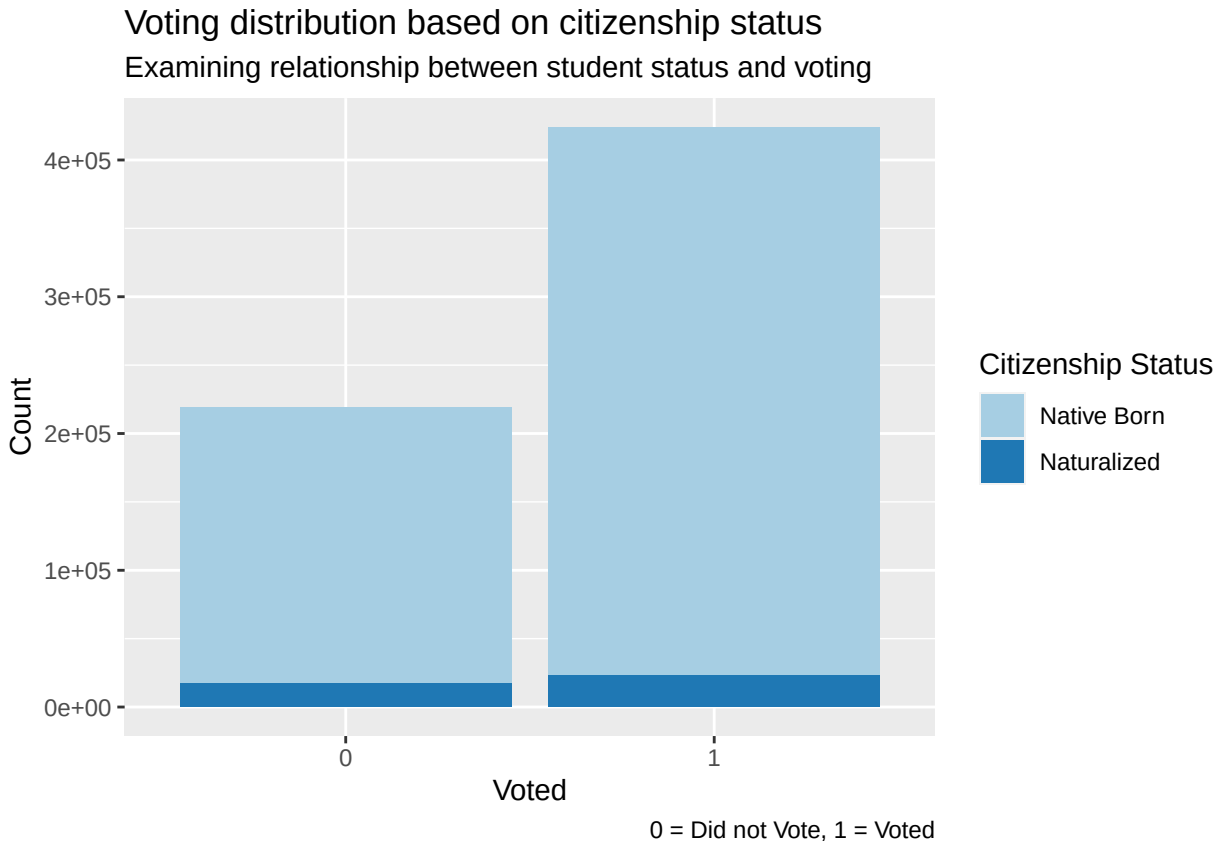
From the bar plot, it is evident that a majority of these individuals were not currently enrolled. This may be a result of a general national trend, but we want to investigate if it is the result of a larger proportion of older individuals within in the range of ages between 16-24. We will investigate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

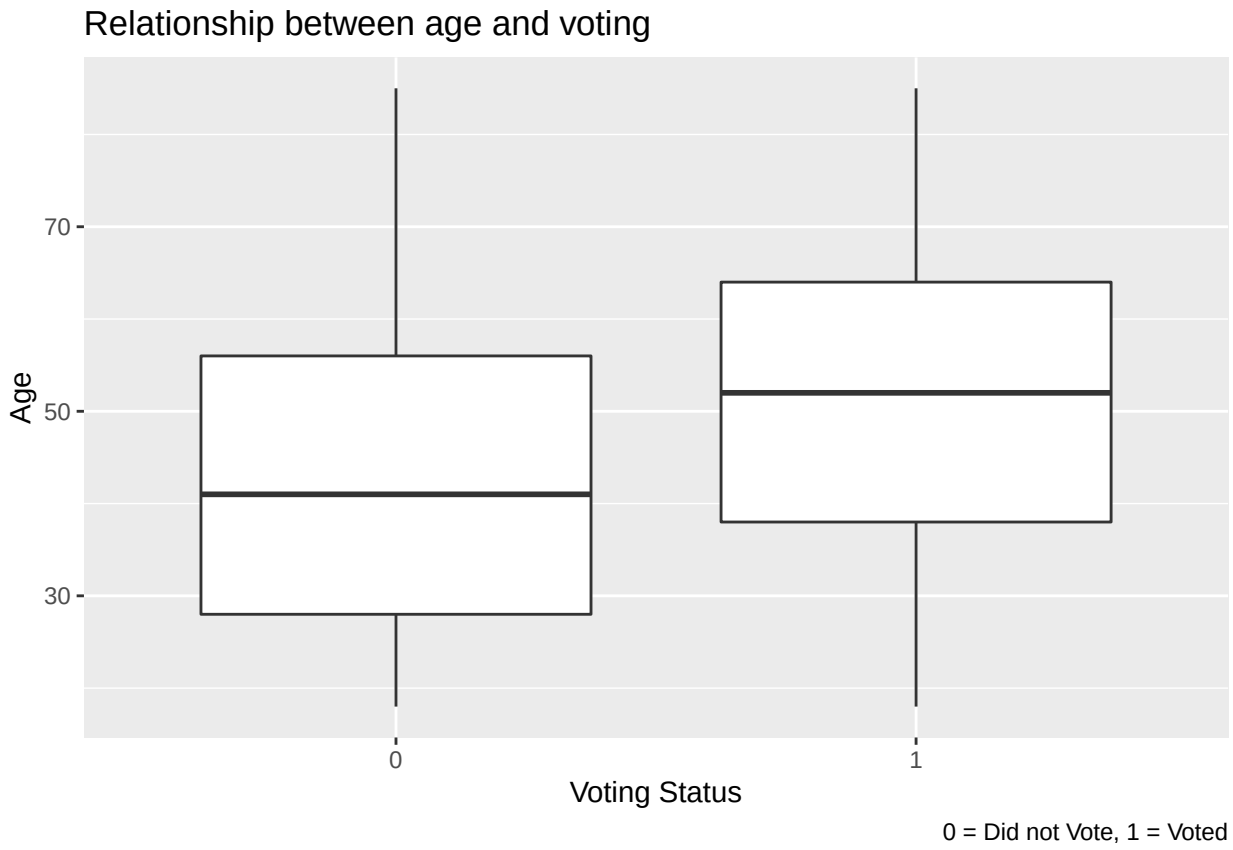
AGE	n	prop
18	2327	0.065
19	3671	0.102
20	4367	0.122
21	4760	0.133
22	5955	0.166
23	6991	0.195
24	7791	0.217

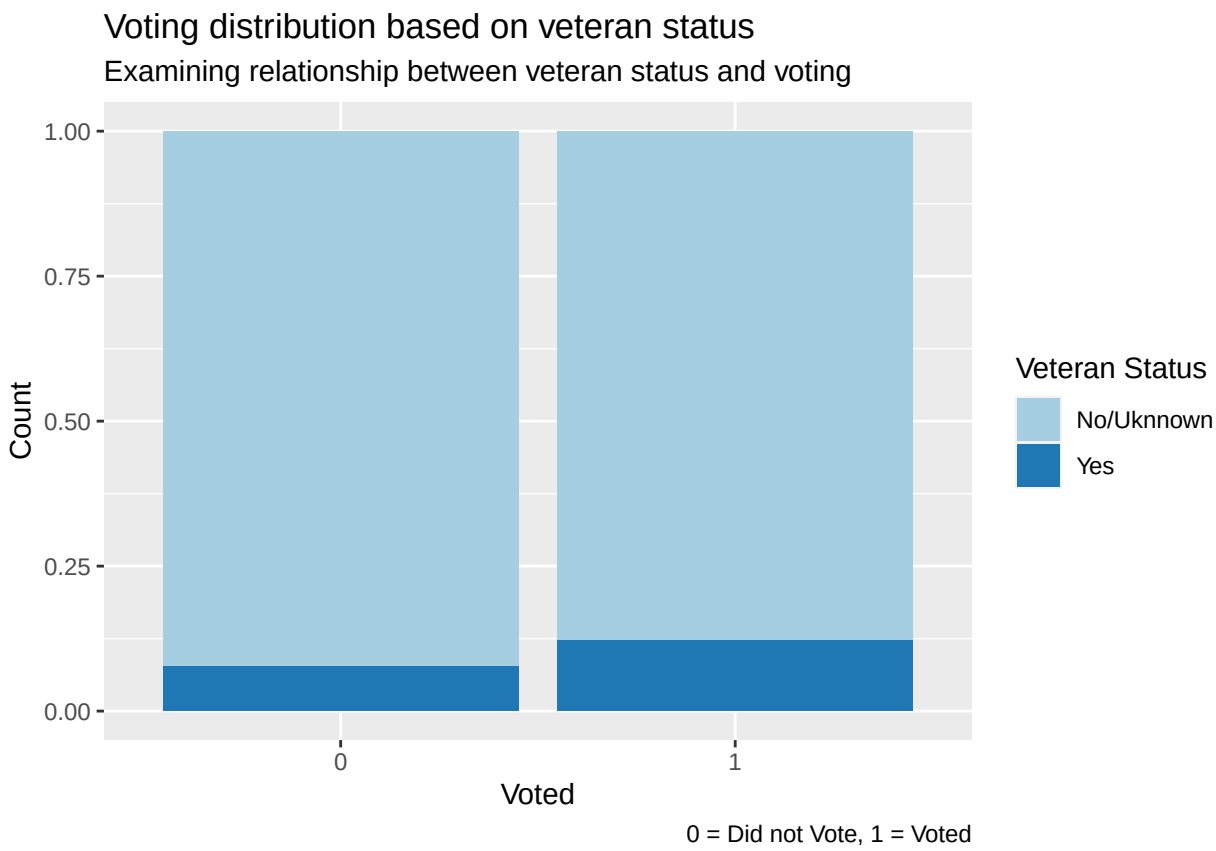
From the table above, it is apparent that more than 40% of those not currently enrolled in school are 23-24 years old. This could be a potential reason for why this age range includes so many who are not currently enrolled as a student.

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.


Voting Distribution of Population of 16−24 Year Olds Enrolled in School
Examining relationship between student status and voting
15000
10000
5000
0
Count
0
1
Voted
0 = Did not Vote, 1 = Voted
Current Student Status
High School Full Time
High School Part Time
College Full Time
College Part Time

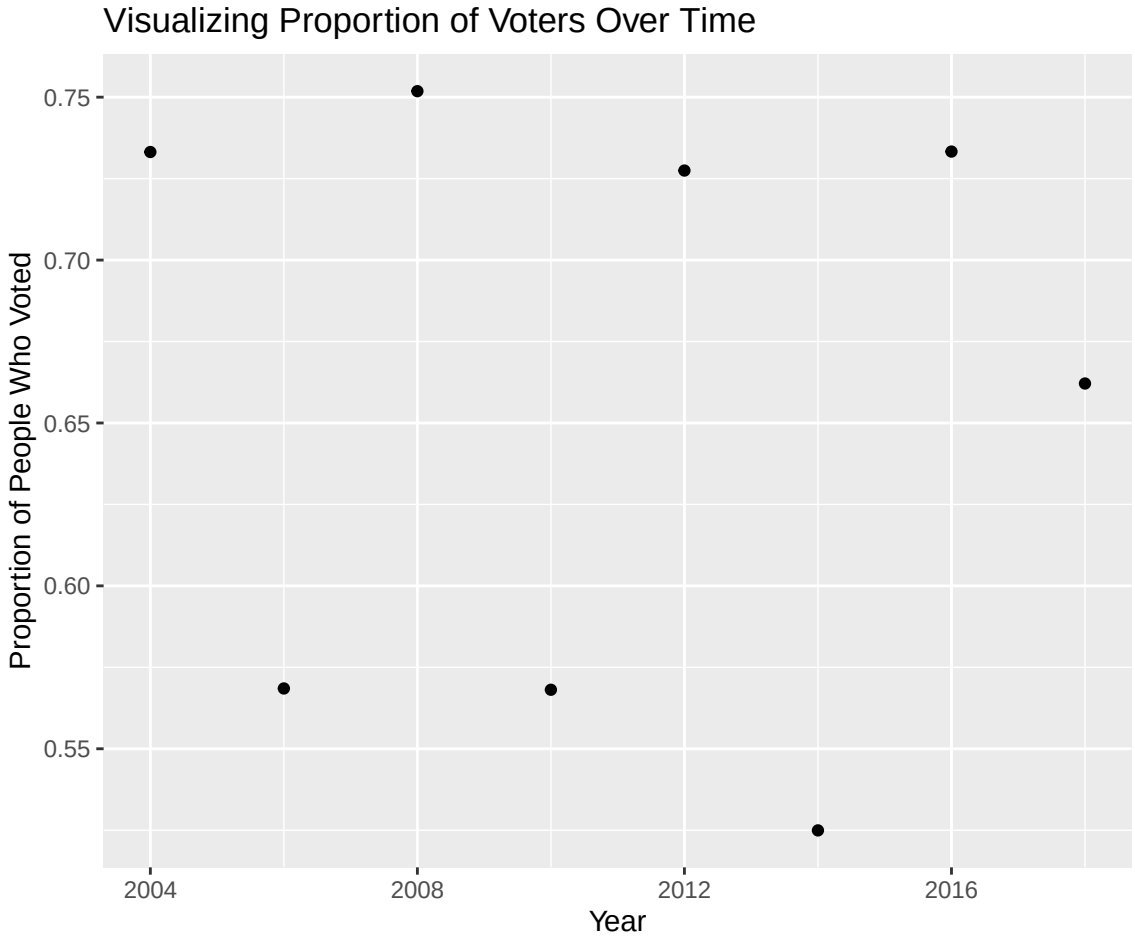

Voting distribution based on citizenship status
Examining relationship between student status and voting
4e+05
3e+05
2e+05
1e+05
0e+00
Count
0
1
Voted
0 = Did not Vote, 1 = Voted
Citizenship Status
Native Born
Naturalized


Relationship between age and voting
70
50
30
Age
0
1
Voting Status
0 = Did not Vote, 1 = Voted

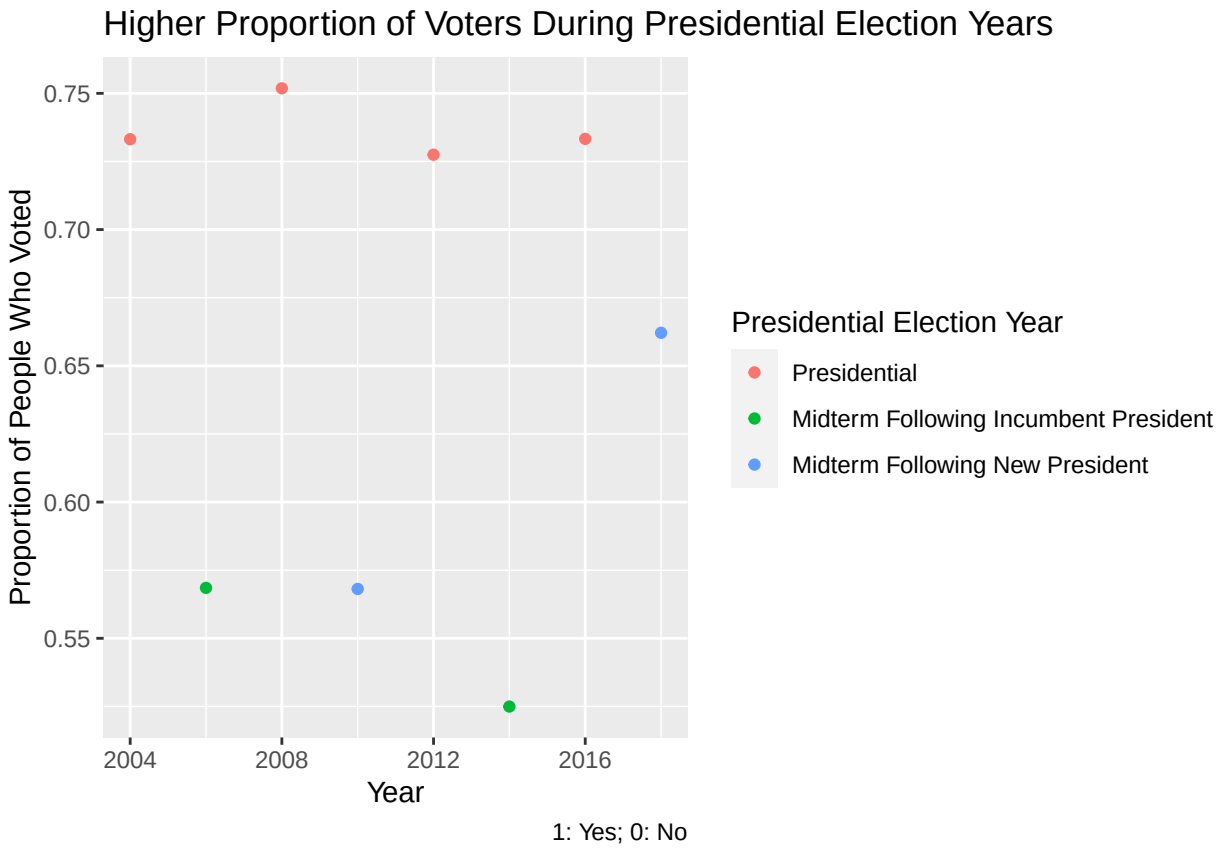

Voting distribution based on veteran status
Examining relationship between veteran status and voting
1.00
0.75
0.50
0.25
0.00
Count
0
1
Voted
0 = Did not Vote, 1 = Voted
Veteran Status
No/Uknnown
Yes

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.

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.

```
## Start:  AIC=10729.96
## voted ~ metro + sex + marst + veteran + citizen + hispanic_status +
##     employed + highest_education + current_student + race + AGE +
##     Presidential_Election_Status + YEAR
##
##                               Df Deviance   AIC
## - metro                        1    10672 10728
## - veteran                      1    10672 10728
## - hispanic_status              1    10673 10729
## - YEAR                         1    10674 10730
## <none>                              10672 10730
## - sex                          1    10675 10731
## - current_student              4    10700 10750
## - citizen                      1    10709 10765
## - employed                     1    10718 10774
## - race                         4    10735 10785
```

```
## - marst                          2    10801 10855
## - Presidential_Election_Status  2    11008 11062
## - AGE                            1    11197 11253
## - highest_education              8    11547 11589
##
## Step:  AIC=10727.96
## voted ~ sex + marst + veteran + citizen + hispanic_status + employed +
##     highest_education + current_student + race + AGE + Presidential_Election_Status +
##     YEAR
##
##                                 Df Deviance   AIC
## - veteran                        1    10672 10726
## - hispanic_status                1    10673 10727
## - YEAR                           1    10674 10728
## <none>                                10672 10728
## - sex                            1    10675 10729
## - current_student                4    10700 10748
## - citizen                        1    10709 10763
## - employed                       1    10718 10772
## - race                           4    10736 10784
## - marst                          2    10803 10855
## - Presidential_Election_Status  2    11008 11060
## - AGE                            1    11197 11251
## - highest_education              8    11554 11594
##
## Step:  AIC=10726.03
## voted ~ sex + marst + citizen + hispanic_status + employed +
##     highest_education + current_student + race + AGE + Presidential_Election_Status +
##     YEAR
##
##                                 Df Deviance   AIC
## - hispanic_status                1    10673 10725
## - YEAR                           1    10674 10726
## <none>                                10672 10726
## - sex                            1    10675 10727
## - current_student                4    10700 10746
## - citizen                        1    10709 10761
## - employed                       1    10718 10770
## - race                           4    10737 10783
## - marst                          2    10803 10853
## - Presidential_Election_Status  2    11008 11058
## - AGE                            1    11219 11271
## - highest_education              8    11556 11594
##
## Step:  AIC=10725.41
## voted ~ sex + marst + citizen + employed + highest_education +
##     current_student + race + AGE + Presidential_Election_Status +
##     YEAR
##
##                                 Df Deviance   AIC
## <none>                                10673 10725
## - YEAR                           1    10676 10726
## - sex                            1    10676 10726
## - current_student                4    10702 10746
```

```
## - citizen                       1    10717 10767
## - employed                      1    10719 10769
## - race                          4    10739 10783
## - marst                         2    10805 10853
## - Presidential_Election_Status  2    11009 11057
## - AGE                           1    11233 11283
## - highest_education             8    11580 11616
```

term	estimate	std.error	statistic	p.value
(Intercept)	15.380	11.315	1.359	0.174
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 New President	0.319	0.069	4.660	0.000
Presidential_Election_StatusPresidential	0.989	0.057	17.347	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 backward selection based on AIC took out the variables metro, veteran, and hispanic_status.

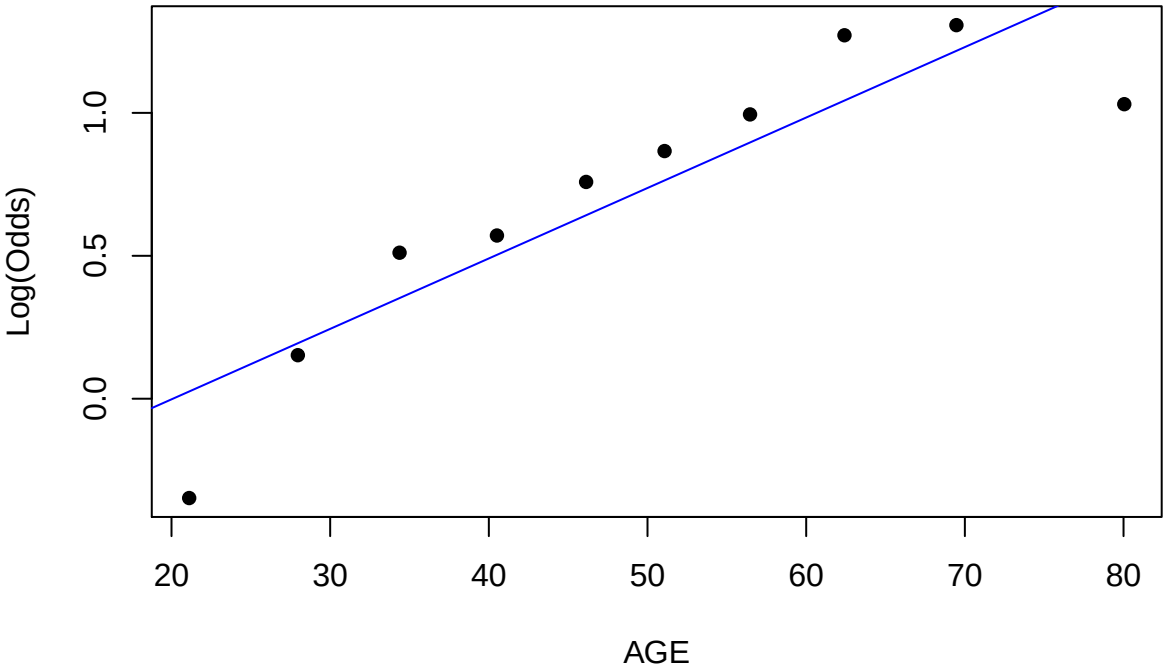
Checking Model Conditions

#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.

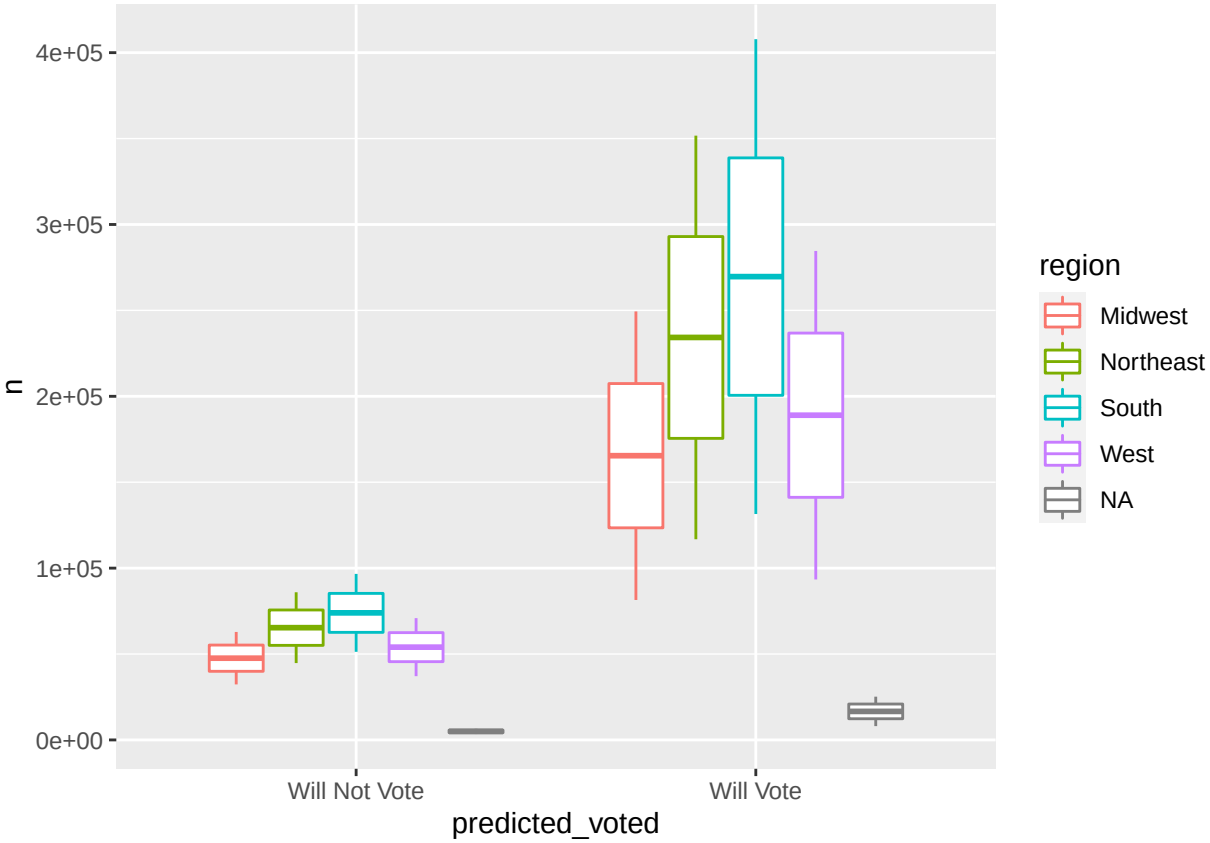
Below are the misclassification rates by region.

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

region	voted	predicted_voted	n	prop
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.

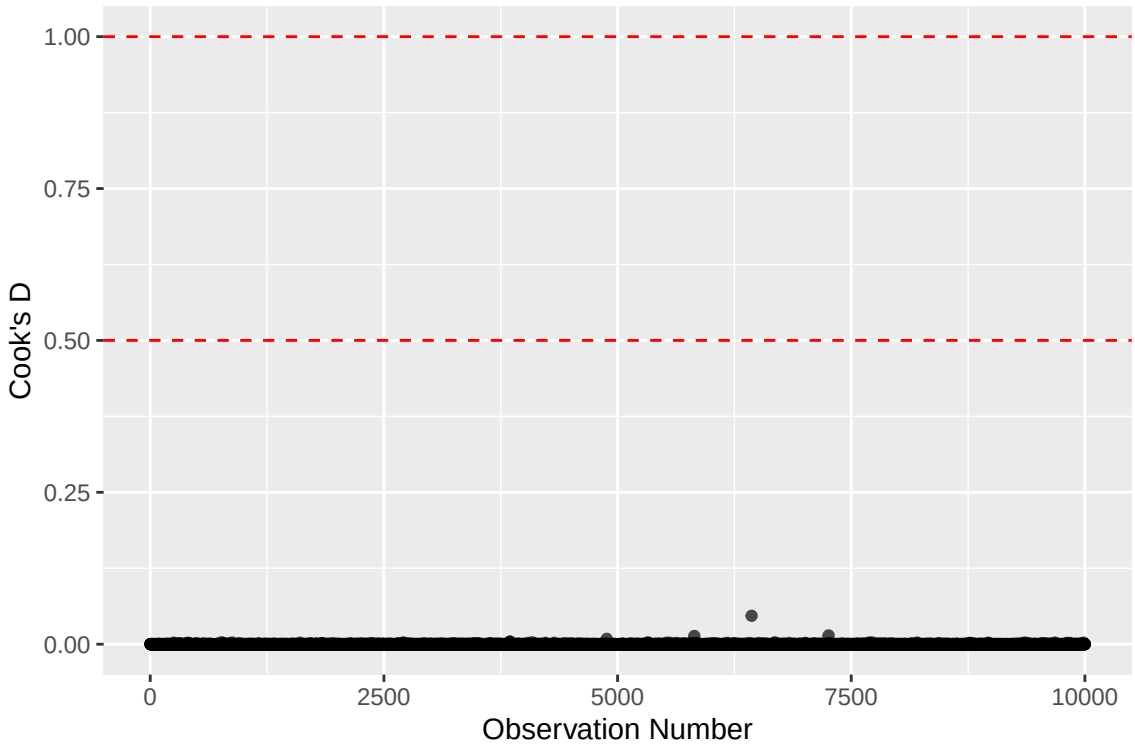


Based on the misclassification rates, we have no reason to believe that the independence condition is not satisfied. The misclassification rates across regions are relatively similar, which suggests that there is not an issue of spatial correlation and that the independence condition is satisfied.

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
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
employedYes	1.433
Presidential_Election_StatusPresidential	1.442
highest_educationAssociate Degree	1.508
Presidential_Election_StatusMidterm Following New President	1.575
AGE	1.709
highest_educationSome High School	1.744

names	x
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)	14.136	11.374	1.243	0.214
YEAR	-0.008	0.006	-1.354	0.176
sexMale	-0.166	0.115	-1.448	0.148
current_studentHigh School Full Time	1.126	0.307	3.664	0.000
current_studentHigh School Part Time	0.371	1.141	0.326	0.745
current_studentCollege Full Time	0.467	0.125	3.750	0.000
current_studentCollege Part Time	0.650	0.251	2.589	0.010
citizenNaturalized	-0.694	0.103	-6.707	0.000
employedYes	0.392	0.074	5.288	0.000
raceBlack	0.372	0.156	2.388	0.017
raceAsian or Pacific Islander	-0.373	0.245	-1.525	0.127
raceNative American	-0.038	0.408	-0.092	0.927
race2 or more races	0.558	0.424	1.317	0.188
marstDivorced/Separated	-0.714	0.071	-10.046	0.000
marstNot Married/Other	-0.490	0.057	-8.560	0.000
Presidential_Election_StatusMidterm Following New President	0.308	0.241	1.278	0.201
Presidential_Election_StatusPresidential	1.786	0.216	8.272	0.000
AGE	0.046	0.003	15.730	0.000
highest_educationHigh School Degree/GED	-1.137	0.134	-8.474	0.000
highest_educationSome College	-0.544	0.146	-3.734	0.000
highest_educationSome High School	-1.969	0.188	-10.473	0.000
highest_educationAssociate Degree	-0.525	0.171	-3.071	0.002
highest_educationMasters Degree	0.011	0.209	0.051	0.960
highest_educationProfessional Degree	0.143	0.435	0.328	0.743
highest_educationDoctorate Degree	1.122	0.558	2.012	0.044

term	estimate	std.error	statistic	p.value
highest_educationNone/Unknown	-13.331	179.345	-0.074	0.941
sexMale:employedYes	0.033	0.103	0.318	0.751
Presidential_Election_StatusMidterm Following New President:highest_educationHigh School Degree/GED	-0.134	0.193	-0.695	0.487
Presidential_Election_StatusPresidential:highest_educationHigh School Degree/GED	-0.451	0.178	-2.529	0.011
Presidential_Election_StatusMidterm Following New President:highest_educationSome College	0.093	0.209	0.444	0.657
Presidential_Election_StatusPresidential:highest_educationSome College	-0.293	0.193	-1.517	0.129
Presidential_Election_StatusMidterm Following New President:highest_educationSome High School	-0.133	0.272	-0.489	0.625
Presidential_Election_StatusPresidential:highest_educationSome High School	-0.476	0.235	-2.025	0.043
Presidential_Election_StatusMidterm Following New President:highest_educationAssociate Degree	0.094	0.249	0.377	0.706
Presidential_Election_StatusPresidential:highest_educationAssociate Degree	-0.078	0.230	-0.342	0.733
Presidential_Election_StatusMidterm Following New President:highest_educationMasters Degree	0.250	0.307	0.813	0.416
Presidential_Election_StatusPresidential:highest_educationMasters Degree	0.637	0.316	2.020	0.043
Presidential_Election_StatusMidterm Following New President:highest_educationProfessional Degree	0.114	0.673	0.169	0.866
Presidential_Election_StatusPresidential:highest_educationProfessional Degree	1.085	0.752	1.444	0.149
Presidential_Election_StatusMidterm Following New President:highest_educationDoctorate Degree	-0.989	0.708	-1.397	0.162
Presidential_Election_StatusPresidential:highest_educationDoctorate Degree	-1.015	0.721	-1.407	0.160
Presidential_Election_StatusMidterm Following New President:highest_educationNone/Unknown	11.198	179.349	0.062	0.950
Presidential_Election_StatusPresidential:highest_educationNone/Unknown	10.002	179.347	0.056	0.956
Presidential_Election_StatusMidterm Following New President:AGE	0.000	0.004	-0.086	0.932
Presidential_Election_StatusPresidential:AGE	-0.013	0.003	-3.828	0.000
sexMale:Presidential_Election_StatusMidterm Following New President	0.033	0.131	0.253	0.800
sexMale:Presidential_Election_StatusPresidential	0.123	0.114	1.076	0.282
raceBlack:Presidential_Election_StatusMidterm Following New President	0.074	0.220	0.338	0.735
raceAsian or Pacific Islander:Presidential_Election_StatusMidterm Following New President	0.537	0.339	1.585	0.113
raceNative American:Presidential_Election_StatusMidterm Following New President	0.415	0.572	0.724	0.469
race2 or more races:Presidential_Election_StatusMidterm Following New President	0.362	0.604	0.598	0.550
raceBlack:Presidential_Election_StatusPresidential	0.504	0.204	2.471	0.013
raceAsian or Pacific Islander:Presidential_Election_StatusPresidential	-0.345	0.302	-1.145	0.252
raceNative American:Presidential_Election_StatusPresidential	-0.284	0.497	-0.572	0.567

term	estimate	std.error	statistic	p.value
race2 or more races:Presidential_Election_StatusPresidential	-0.743	0.507	-1.466	0.143

Resid..Df	Resid..Dev	df	Deviance	p.value
9974	10673.41	NA	NA	NA
9945	10592.26	29	81.155	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.

[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

keep in mind: citizenship and registration exclusion for the model