Your project title

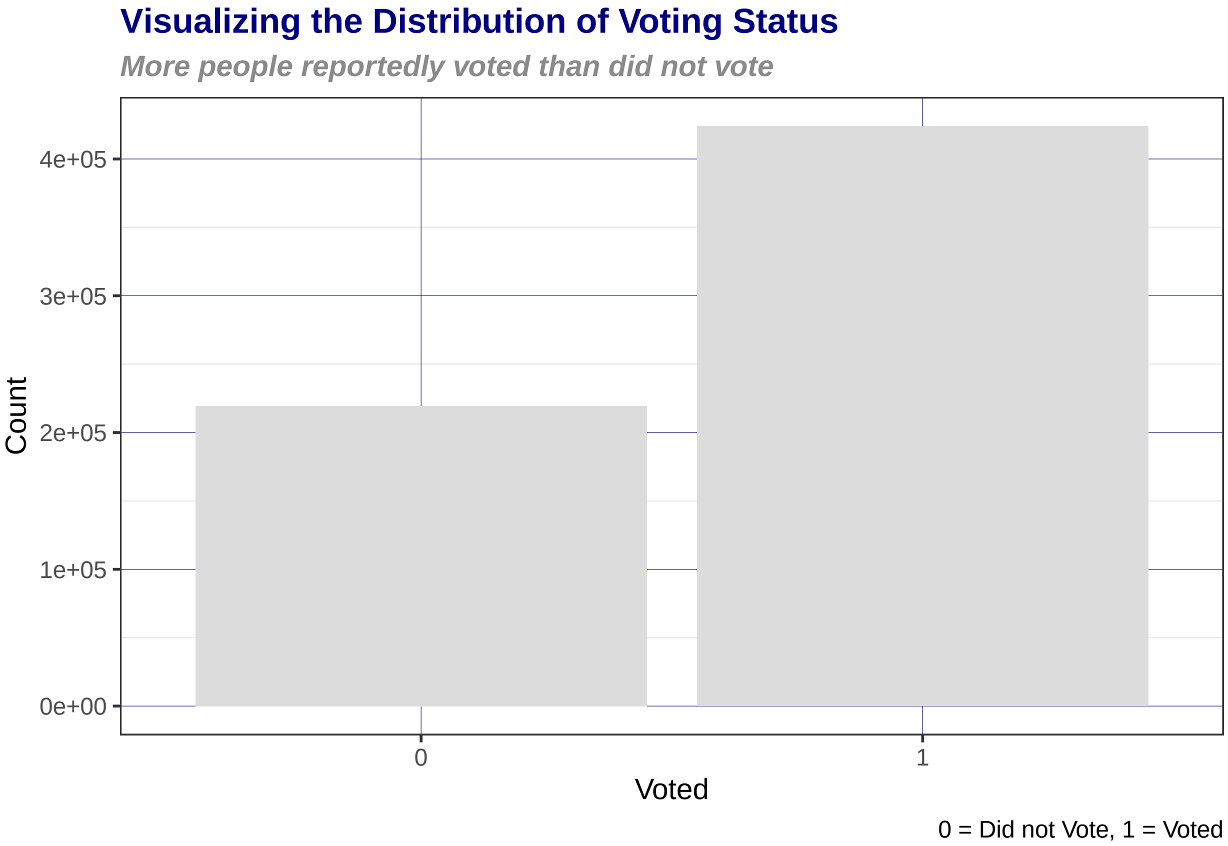
Approximately Normal: Jenny Huang, Bella Larsen, Jessie Bierschenk, Aidan Gildea

28 October 2020

Introduction

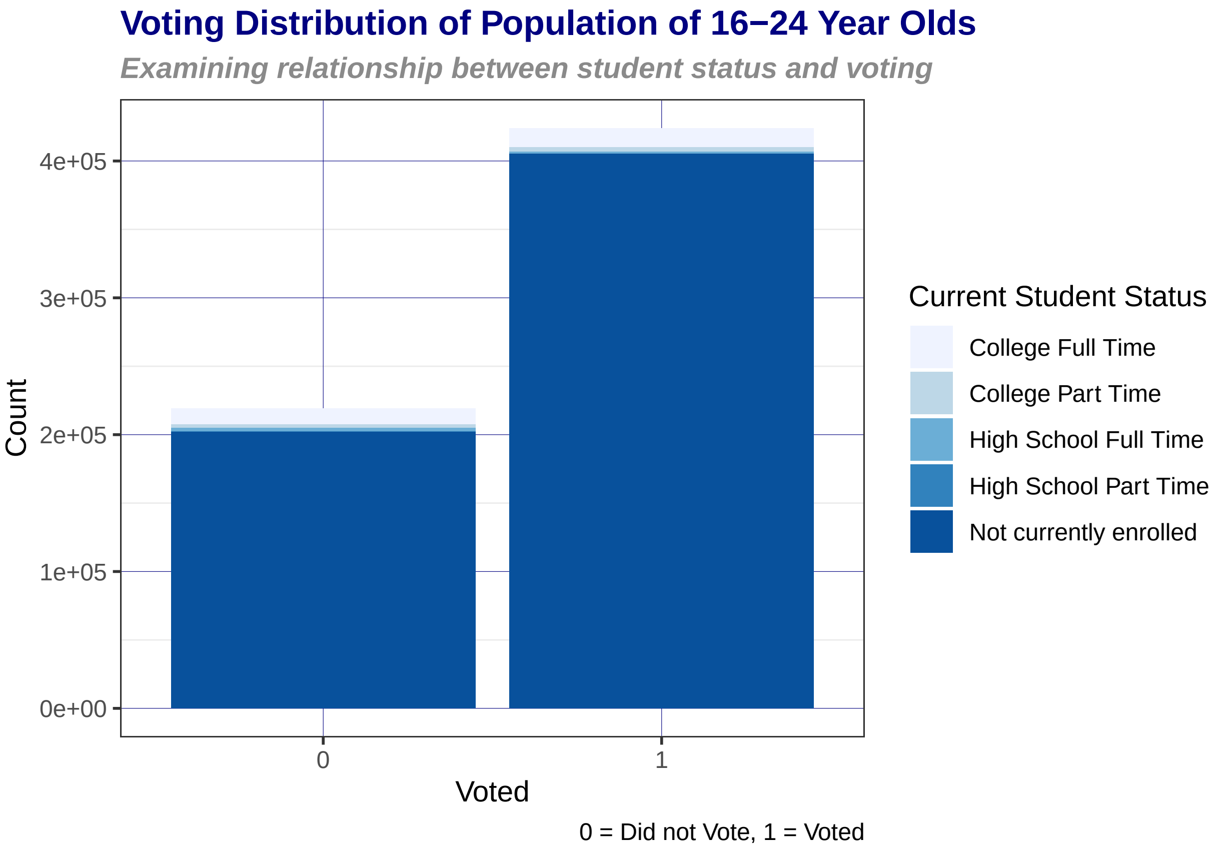
We will begin our EDA by visualizing the relationship between the response variable `voted` and several of the other variables of particular interest.

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



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.



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.

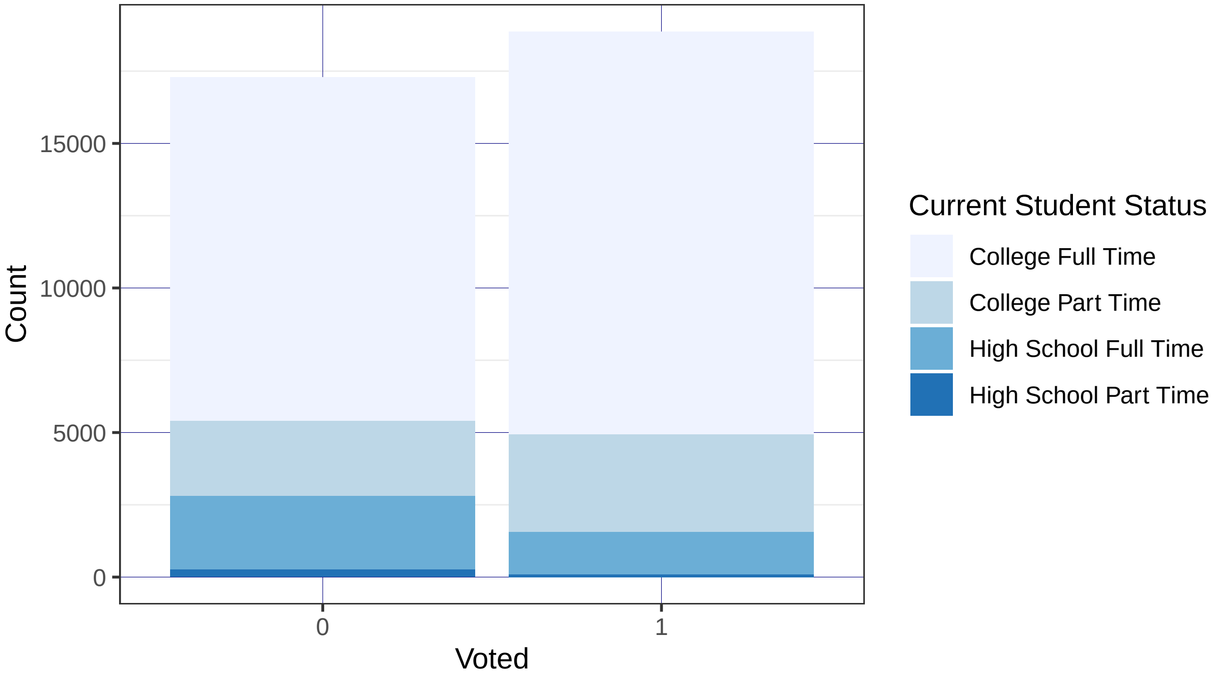
AGE	n	prop
18	2327	0.065
19	3671	0.102
20	4367	0.122
21	4760	0.133
22	5955	0.166
23	6991	0.195
24	7791	0.217

From the kable 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

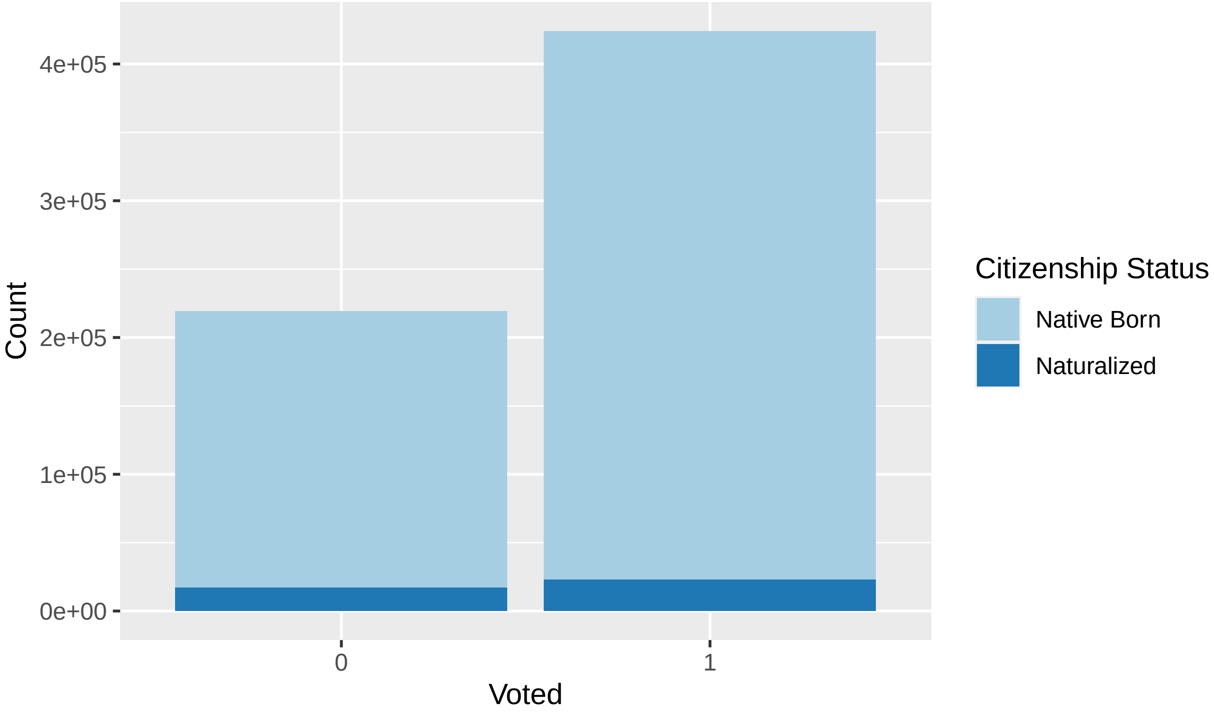
Examining relationship between student status and voting



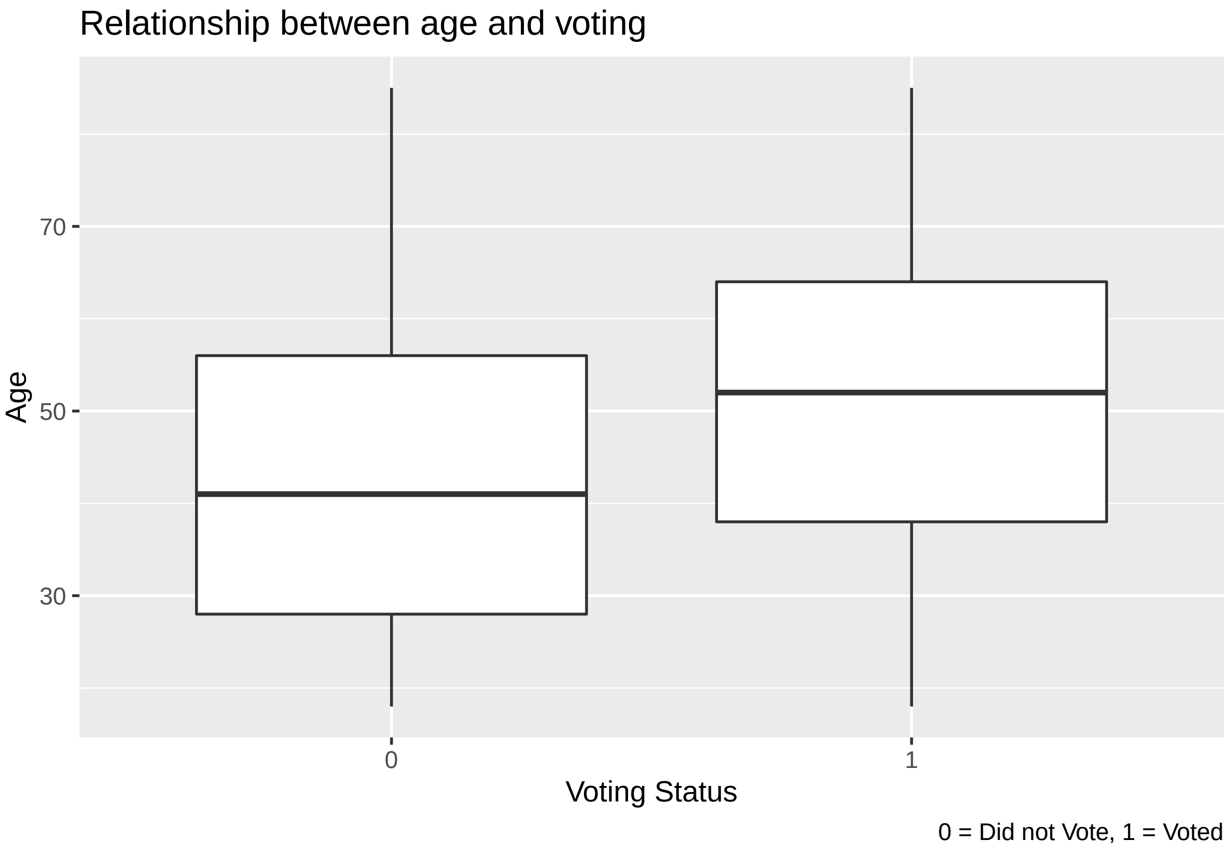
0 = Did not Vote, 1 = Voted

Voting distribution based on citizenship status

Examining relationship between student status and voting

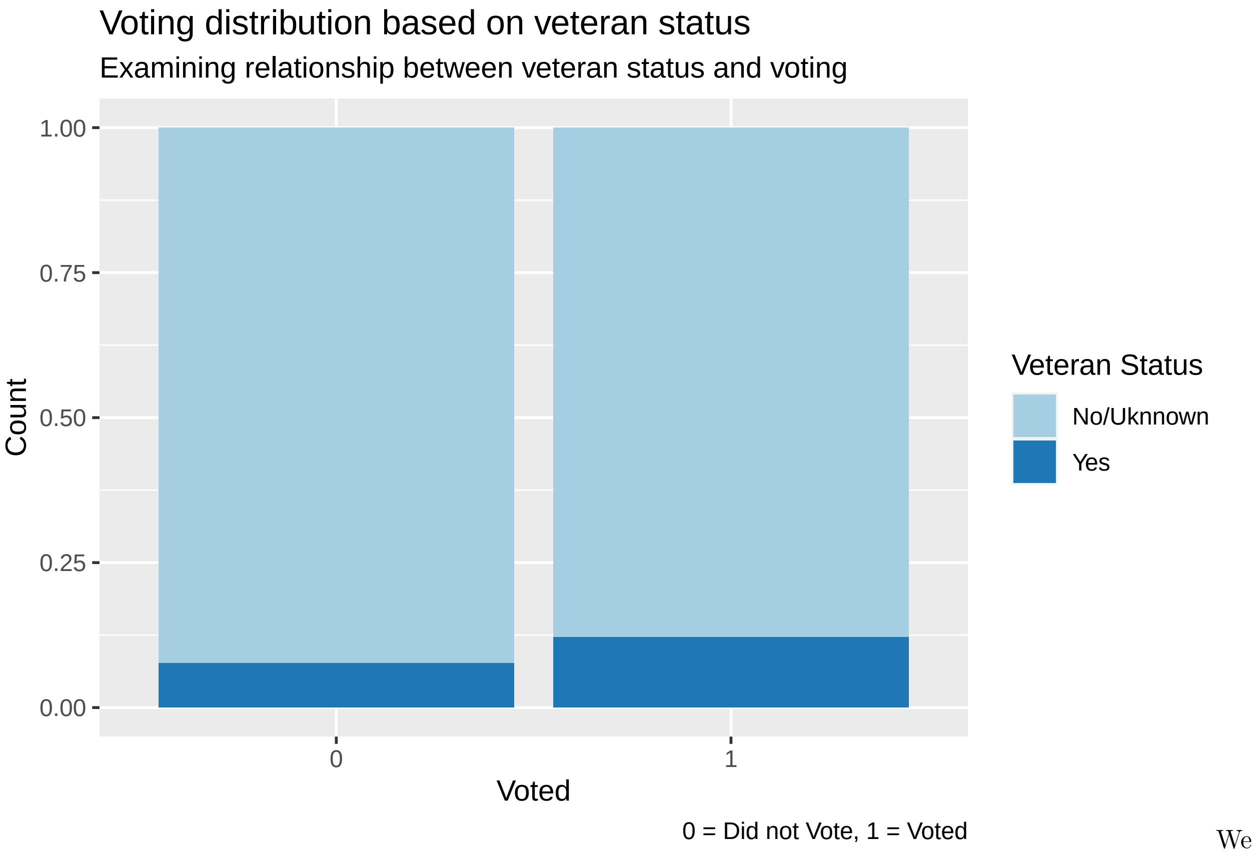


0 = Did not Vote, 1 = Voted



Relationship between age and voting

0 = Did not Vote, 1 = Voted



Voting distribution based on veteran status

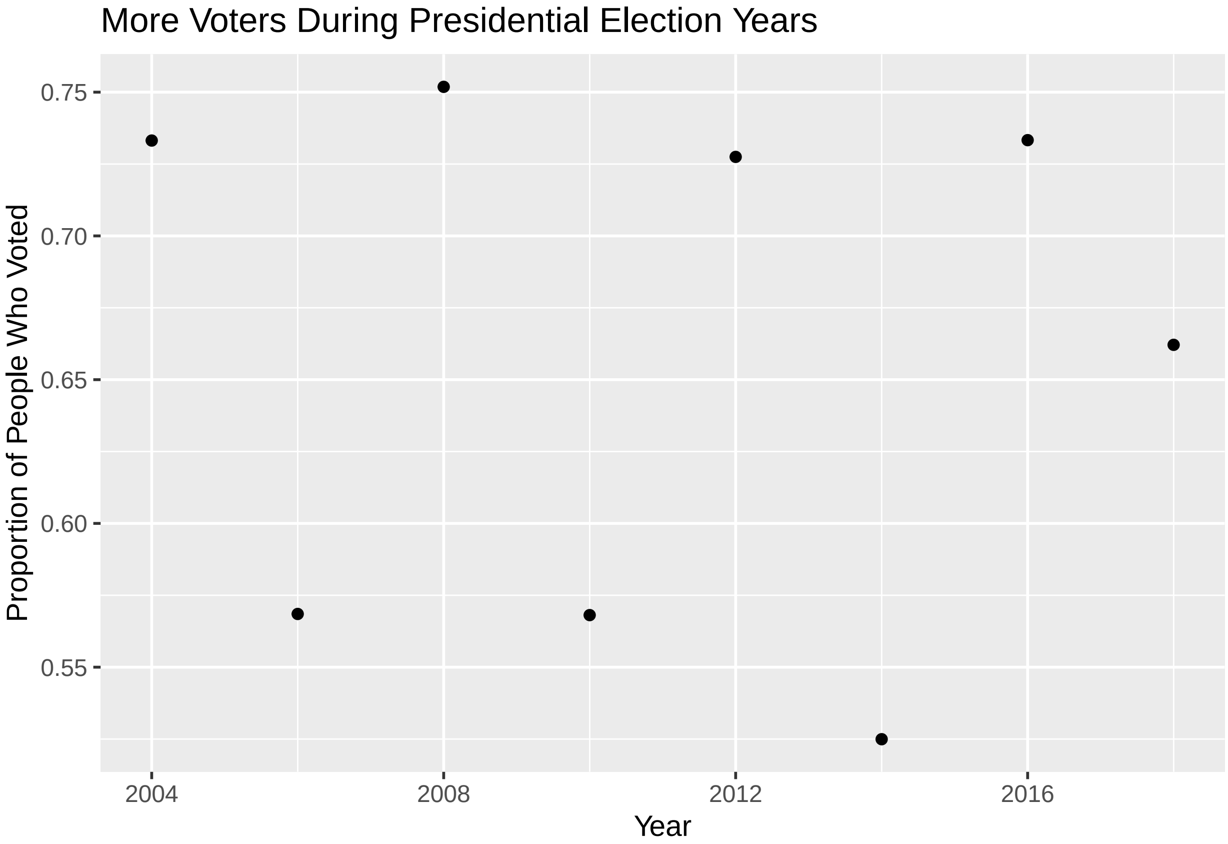
Examining relationship between veteran status and voting

0 = Did not Vote, 1 = Voted

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.



Model Selection

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

```
## Start:  AIC=10749.78
## voted ~ metro + sex + marst + veteran + citizen + hispanic_status +
##     employed + highest_education + current_student + race + AGE +
##     Presidential_Election_Year + YEAR
##
##                              Df Deviance   AIC
## - metro                       1    10694 10748
## - YEAR                        1    10694 10748
## - veteran                     1    10694 10748
## - hispanic_status             1    10695 10749
## <none>                             10694 10750
## - sex                         1    10696 10750
## - current_student             4    10721 10769
## - citizen                     1    10731 10785
## - employed                    1    10740 10794
## - race                        4    10759 10807
## - marst                       2    10824 10876
## - Presidential_Election_Year  1    11008 11062
## - AGE                         1    11216 11270
## - highest_education           8    11572 11612
##
## Step:  AIC=10747.78
## voted ~ sex + marst + veteran + citizen + hispanic_status + employed +
##     highest_education + current_student + race + AGE + Presidential_Election_Year +
##     YEAR
##
##                              Df Deviance   AIC
## - YEAR                        1    10694 10746
## - veteran                     1    10694 10746
## - hispanic_status             1    10695 10747
## <none>                             10694 10748
## - sex                         1    10696 10748
## - current_student             4    10721 10767
## - citizen                     1    10731 10783
## - employed                    1    10740 10792
## - race                        4    10760 10806
## - marst                       2    10825 10875
## - Presidential_Election_Year  1    11008 11060
## - AGE                         1    11216 11268
## - highest_education           8    11579 11617
##
## Step:  AIC=10745.79
## voted ~ sex + marst + veteran + citizen + hispanic_status + employed +
##     highest_education + current_student + race + AGE + Presidential_Election_Year
##
##                              Df Deviance   AIC
## - veteran                     1    10694 10744
## - hispanic_status             1    10695 10745
## <none>                             10694 10746
## - sex                         1    10696 10746
## - current_student             4    10722 10766
## - citizen                     1    10731 10781
## - employed                    1    10740 10790
## - race                        4    10760 10804
## - marst                       2    10825 10873
## - Presidential_Election_Year  1    11018 11068
## - AGE                         1    11218 11268
## - highest_education           8    11584 11620
##
## Step:  AIC=10743.83
## voted ~ sex + marst + citizen + hispanic_status + employed +
##     highest_education + current_student + race + AGE + Presidential_Election_Year
##
##                              Df Deviance   AIC
## - hispanic_status             1    10695 10743
## <none>                             10694 10744
## - sex                         1    10696 10744
## - current_student             4    10722 10764
## - citizen                     1    10731 10779
## - employed                    1    10740 10788
## - race                        4    10760 10802
## - marst                      2    10826 10872
## - Presidential_Election_Year  1    11018 11066
## - AGE                        1    11238 11286
## - highest_education          8    11585 11619
##
## Step:  AIC=10743.2
## voted ~ sex + marst + citizen + employed + highest_education +
##     current_student + race + AGE + Presidential_Election_Year
##
##                              Df Deviance   AIC
## <none>                             10695 10743
## - sex                        1    10698 10744
## - current_student            4    10723 10763
## - citizen                    1    10739 10785
## - employed                   1    10741 10787
## - race                       4    10762 10802
## - marst                      2    10828 10872
## - Presidential_Election_Year  1    11020 11066
## - AGE                        1    11253 11299
## - highest_education          8    11609 11641
```

```
## # A tibble: 24 x 5
##    term                                    estimate std.error statistic  p.value
##    <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                              -1.30      0.258      -5.03 4.81e- 7
##  2 sexMale                                  -0.0755    0.0479     -1.58 1.15e- 1
##  3 marstMarried                              0.703     0.0706      9.95 2.51e-23
##  4 marstNot Married/Other                    0.228     0.0773      2.95 3.21e- 3
##  5 citizenNaturalized                       -0.684     0.103      -6.64 3.05e-11
##  6 employedYes                               0.405     0.0599      6.75 1.44e-11
##  7 highest_educationBachelors Degree         0.532     0.0952      5.59 2.31e- 8
##  8 highest_educationDoctorate Degree         0.927     0.278       3.33 8.53e- 4
##  9 highest_educationHigh School Degree/GED  -0.833     0.0841     -9.90 4.13e-23
## 10 highest_educationMasters Degree           0.812     0.132       6.16 7.31e-10
## # ... with 14 more rows
```

The final model included sex + marst + citizen + employed + highest_education + current_student + race + AGE + Presidential_Election_Year . The model took out the variables metro, veteran, and hispanic_status.

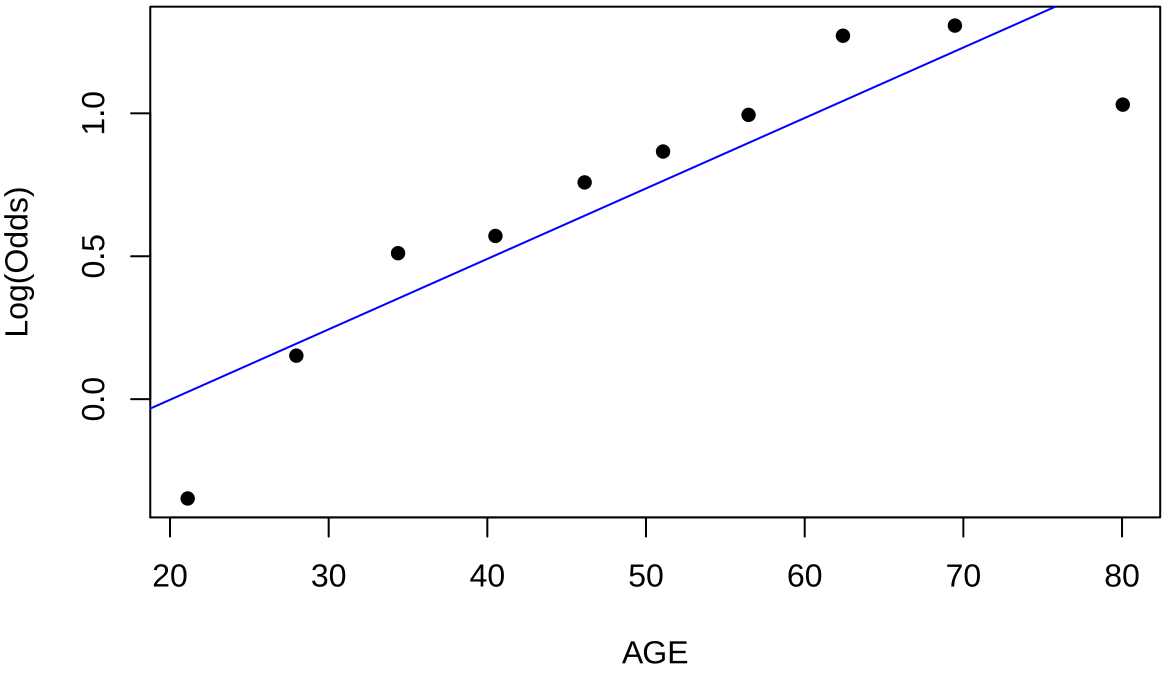
(variable: midterm after presidential election vs. not, presidential, midterm after incumbant, midterm after new pres)

Checking Model Conditions

```
## # A tibble: 2 x 5
## # Groups:   sex [2]
##   sex    voted     n  prop emp_logit
##   <fct>  <fct> <int> <dbl>     <dbl>
## 1 Female 1      3582 0.670     0.710
## 2 Male   1      3029 0.650     0.621
```

```
## # A tibble: 3 x 5
## # Groups:   marst [3]
##   marst              voted     n  prop emp_logit
##   <fct>              <fct> <int> <dbl>     <dbl>
## 1 Divorced/Separated 1       783 0.588     0.355
## 2 Married            1      4145 0.735     1.02
## 3 Not Married/Other  1      1683 0.556     0.226
```

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).

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.

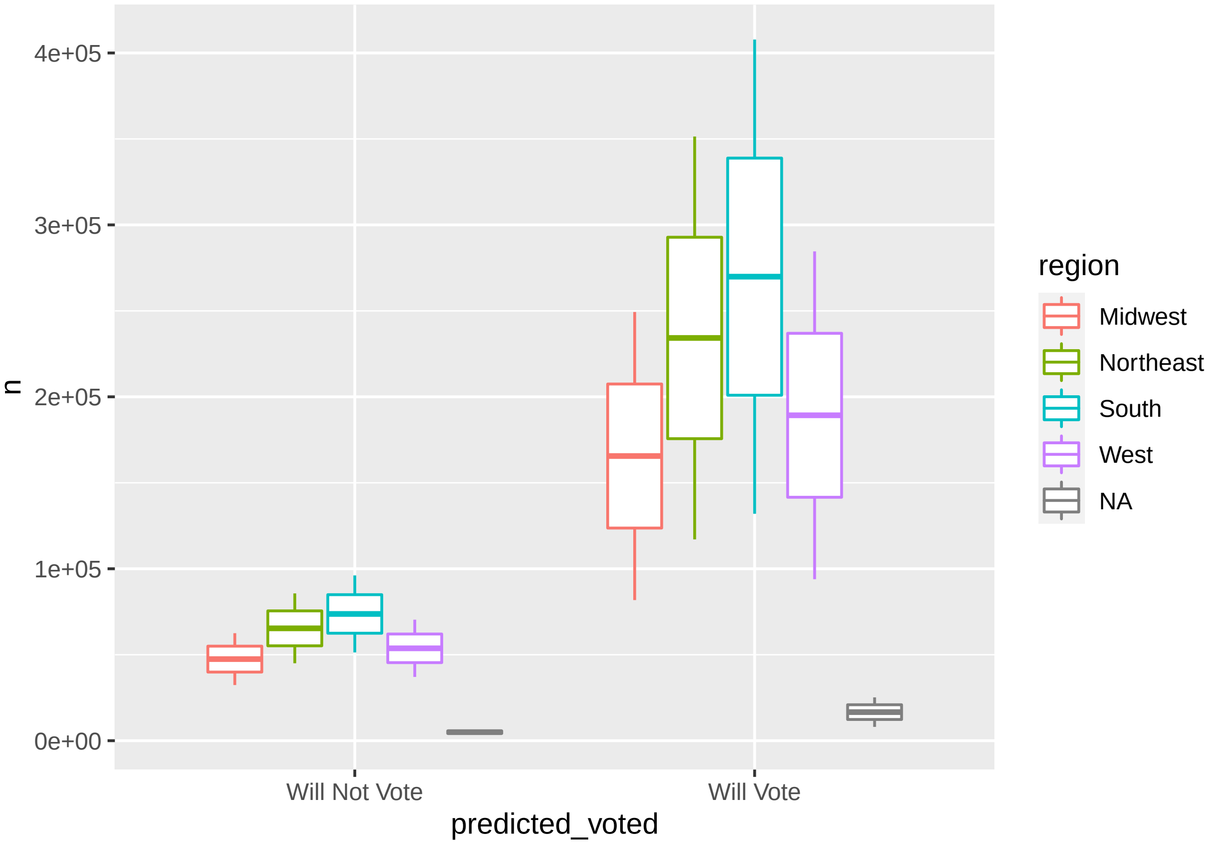
Here, we create a confusion matrix with a misclassification rate of 0.5.

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

Below are the misclassification rates by region.

```
## # A tibble: 10 x 5
## # Groups:   region [5]
##    region    voted predicted_voted      n   prop
##    <chr>     <fct> <chr>            <int>  <dbl>
##  1 Midwest   0     Will Vote        81790 0.192
##  2 Midwest   1     Will Not Vote    32378 0.0760
##  3 Northeast 0     Will Vote       117077 0.195
##  4 Northeast 1     Will Not Vote    45035 0.0752
##  5 South     0     Will Vote       131989 0.192
##  6 South     1     Will Not Vote    51318 0.0747
##  7 West      0     Will Vote        93919 0.193
##  8 West      1     Will Not Vote    37085 0.0763
##  9 <NA>      0     Will Vote         8042 0.186
## 10 <NA>      1     Will Not Vote     3424 0.0793
```

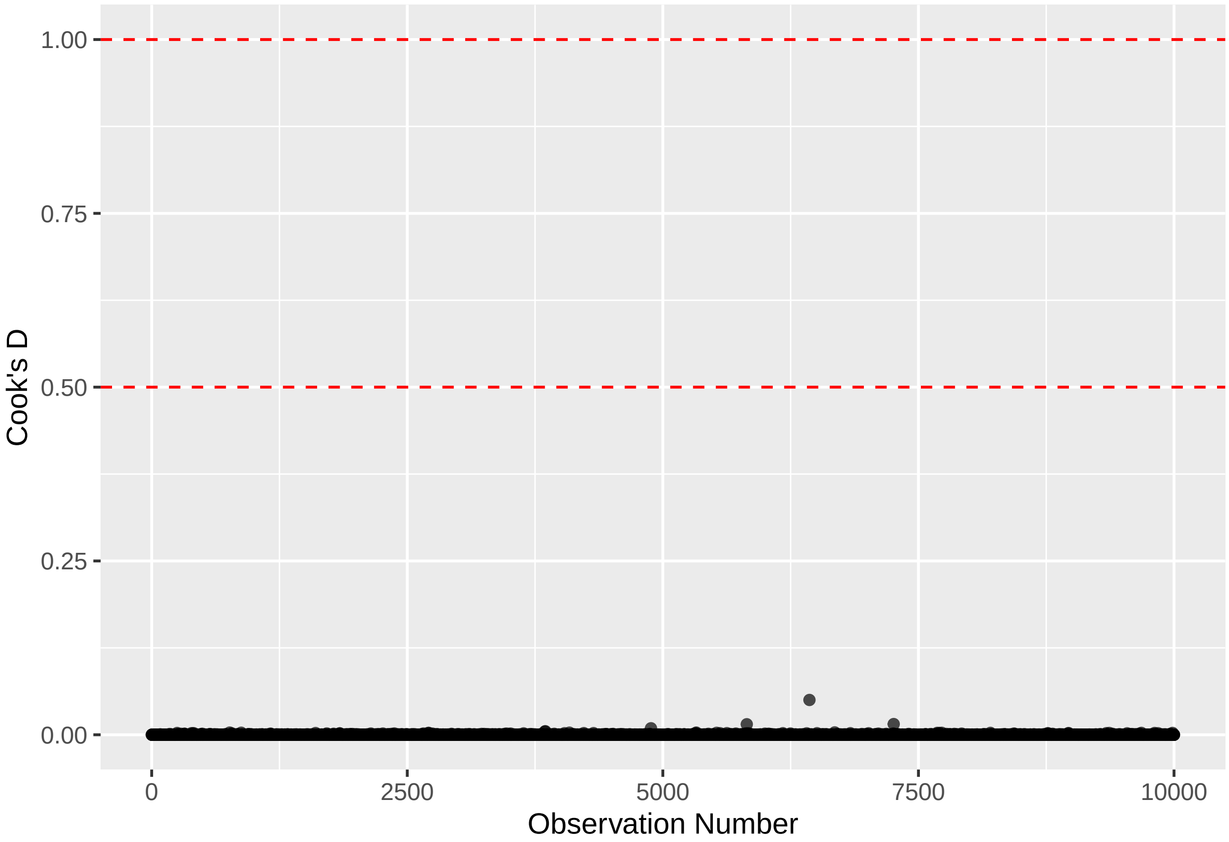
We plan to create a plot of misclassification rate by region to determine if independence is satisfied.



##Checking for influential points

Cook's distance

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



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

Below, we check for multicolinearity.

```
## # A tibble: 23 x 2
##    names                                    x
##    <chr>                                <dbl>
##  1 current_studentHigh School Part Time  1.01
##  2 highest_educationNone/Unknown         1.02
##  3 sexMale                               1.02
##  4 Presidential_Election_Year1           1.03
##  5 highest_educationProfessional Degree  1.07
##  6 highest_educationDoctorate Degree     1.08
##  7 current_studentHigh School Full Time  1.17
##  8 current_studentCollege Part Time      1.19
##  9 citizenNaturalized                    1.21
## 10 highest_educationMasters Degree       1.40
## # ... with 13 more rows
```

We observe that raceWhite has a VIF value that is above 10 (10.072951). This means that colinearity might be present between raceWhite and raceBlack (7.074850).

[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