Voter Prediction Model - Data Exploration

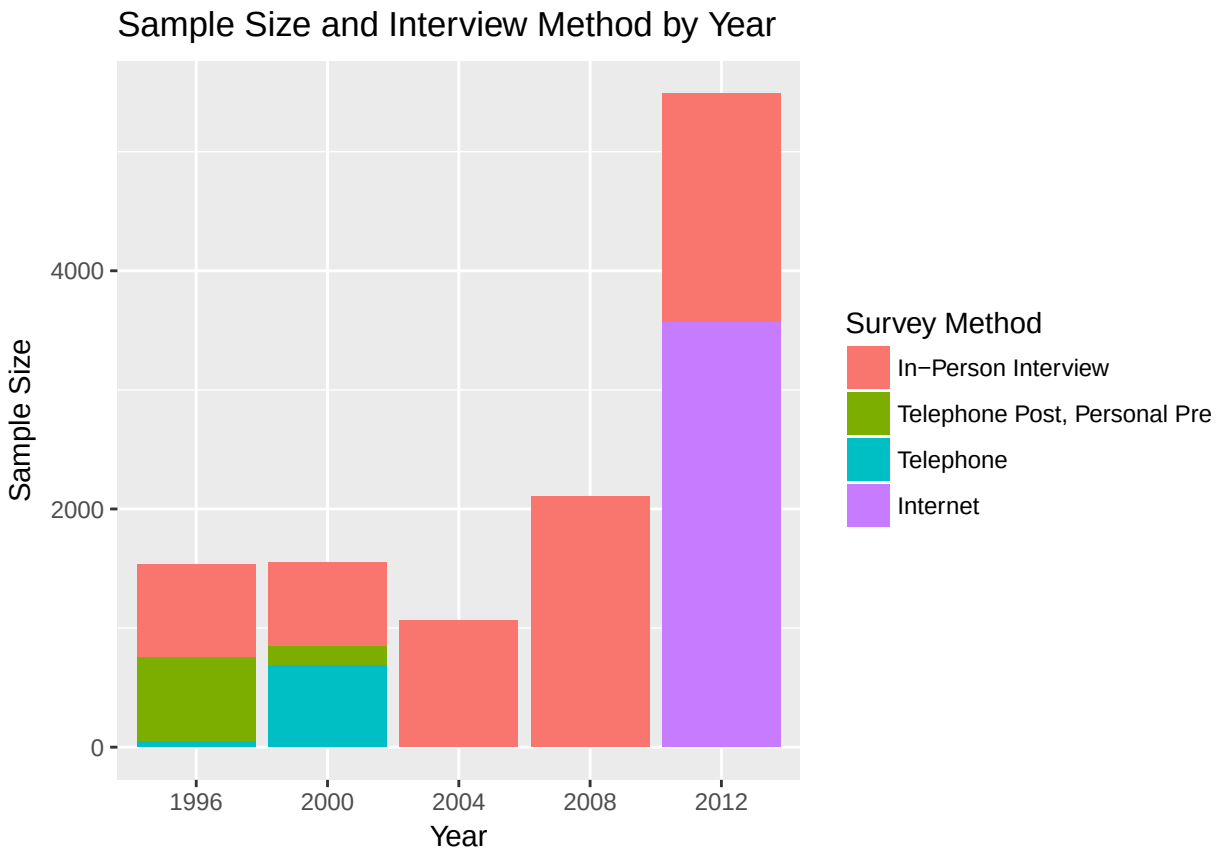
Keyan Halperin

```
setwd("C:/Users/Keyan/Google Drive/Projects")
data = read.csv('ANES Final Data.csv')
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
data$year = as.factor(data$year)
#Graph of sample size by year
ggplot(data, aes(x = year)) +
  geom_bar(aes(fill = factor(method, labels = c('In-Person Interview', 'Telephone Post, Personal Pre', 'Telephone', 'Internet'))))  +
  labs(title = 'Sample Size and Interview Method by Year', y = 'Sample Size', x = 'Year', fill = 'Survey Method')
```



As we can see, the sample size in 2012 is much bigger than it was in previous years. This is because in 2012, the ANES started doing some of their polling online, which allowed them to survey significantly more people. We will now examine the responses of individuals who were surveyed online to see if they significantly differ from the responses of those who were surveyed over the phone or in person.

```
didvote.by.method = xtabs(~ method + did.vote, data = data)
didvote.by.method
```

```
##                                     did.vote
## method                               1. No, did not vote 2. Yes, voted
##   0. All personal                                  1627          4931
##   2. Telephone post (personal pre)                  200           665
##   3. All telephone                                  151           597
```

```
##   4. All internet (2012: pre and post)                 587          2986
```

```
prop.table(didvote.by.method, margin = 1)
```

```
##                                     did.vote
## method                               1. No, did not vote 2. Yes, voted
##   0. All personal                                0.2480939     0.7519061
##   2. Telephone post (personal pre)               0.2312139     0.7687861
##   3. All telephone                               0.2018717     0.7981283
##   4. All internet (2012: pre and post)           0.1642877     0.8357123
```

It appears that those who were interviewed online online were more likely to say that they voted. But in order to examine if the difference is significant, we can perform a Chi-Square test for association.

```
chisq.test(didvote.by.method)
```

```
##
##  Pearson's Chi-squared test
##
## data:  didvote.by.method
## X-squared = 97.19, df = 3, p-value < 2.2e-16
```

Based on our test, we have strong evidence that there is an association between interview method and whether or not an individual said that they voted. However, this may be because interview method is confounded with year of survey. We will see if this relationship still holds after controlling for year.

```
data.2012 = data[data$year == '2012', ]
data.2012$method = factor(data.2012$method)
didvote.by.method.2012 = xtabs(~ method + did.vote, data = data.2012)
chisq.test(didvote.by.method.2012)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  didvote.by.method.2012
## X-squared = 87.598, df = 1, p-value < 2.2e-16
```

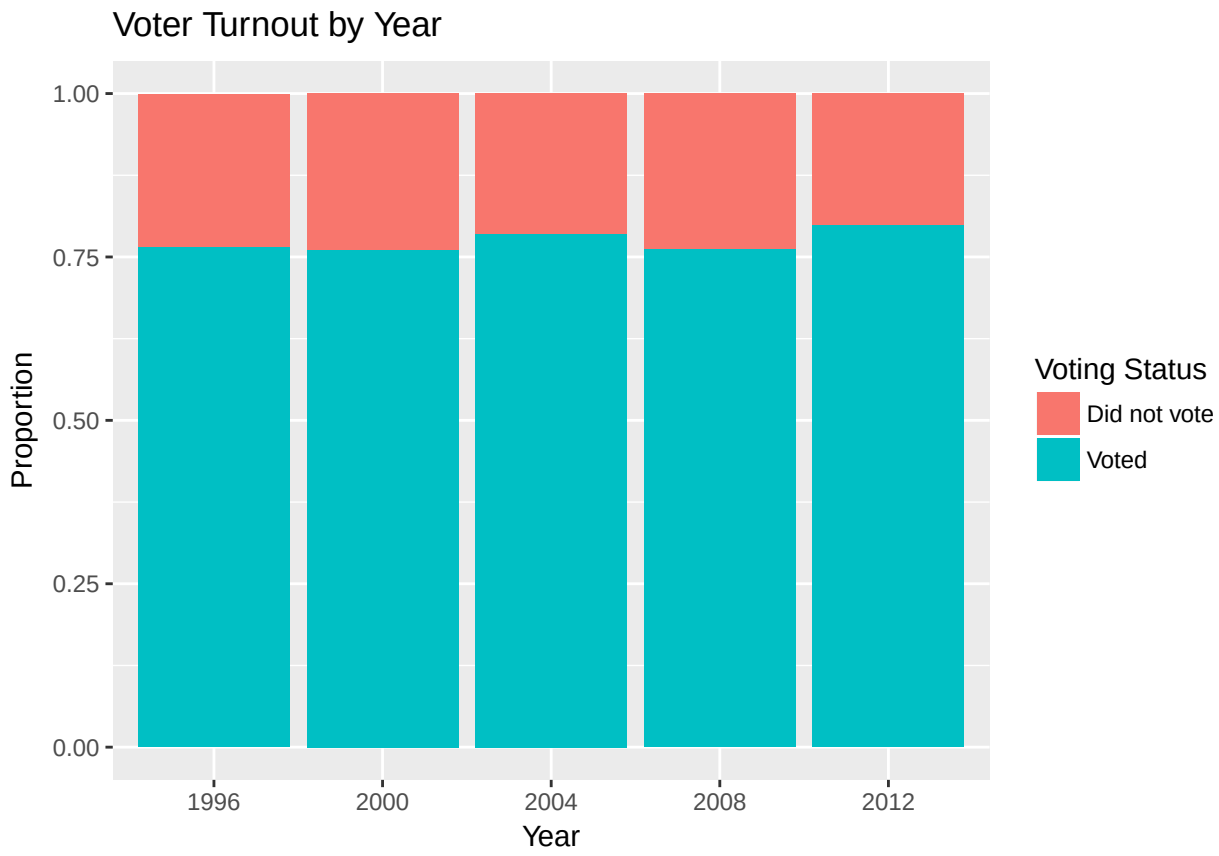
```
data.2000 = data[data$year == '2000', ]
data.2000$method = factor(data.2000$method)
didvote.by.method.2000 = xtabs(~ method + did.vote, data = data.2000)
chisq.test(didvote.by.method.2000)
```

```
##
##  Pearson's Chi-squared test
```

```
##
## data:  didvote.by.method.2000
## X-squared = 6.9419, df = 2, p-value = 0.03109
```

Even when look within a particular year, there is a significant relationship between interview method and how an individual voted. Consequently, we will proceed with caution.

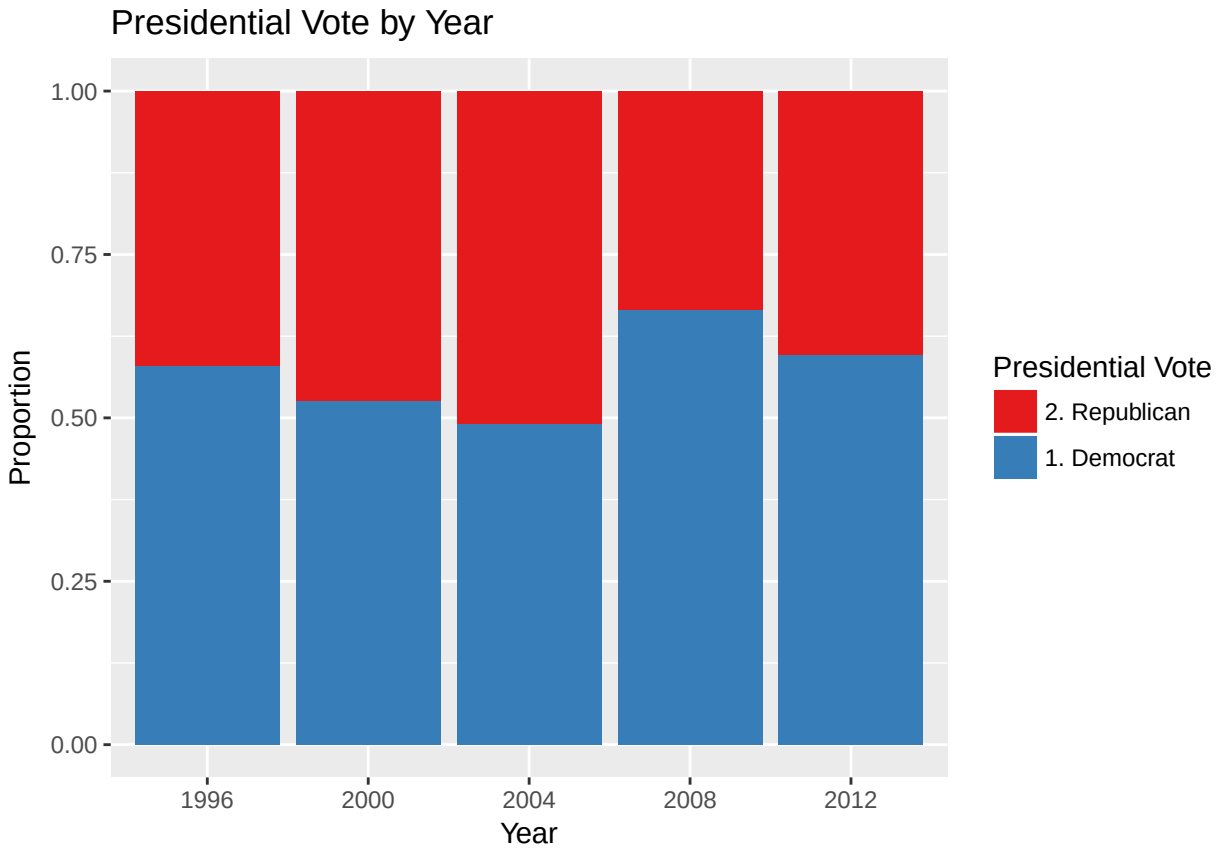
```
#Graph of voter turnout by year
levels(data$did.vote) = c('Did not vote', 'Voted')

ggplot(data, aes(x = year)) +
  geom_bar(aes(fill = did.vote), position = 'fill') + labs(title = 'Voter Turnout by Year',
  y = 'Proportion', x = 'Year', fill = 'Voting Status')
```



There doesn't appear to be a big difference between year and voter turnout. However, one noteworthy thing is that voter turnout is significantly higher among the individuals in this data than it is for the general public.

```
#Presidential vote by year

#Take out people who did not vote from the data
voted.data = data[data$pres.vote != '0. Did not vote; DK/NA if voted; refused to say if',]
voted.data$pres.vote = factor(voted.data$pres.vote, levels = rev(levels(voted.data$pres.vote)))

ggplot(voted.data, aes(x = year, fill = pres.vote)) +
  geom_bar(aes(fill = pres.vote), position = 'fill') + labs(title = 'Presidential Vote by Year',
  y = 'Proportion', x = 'Year', fill = 'Presidential Vote') +
```

```
  scale_fill_brewer(palette = 'Set1', breaks=levels(voted.data$pres.vote))
```

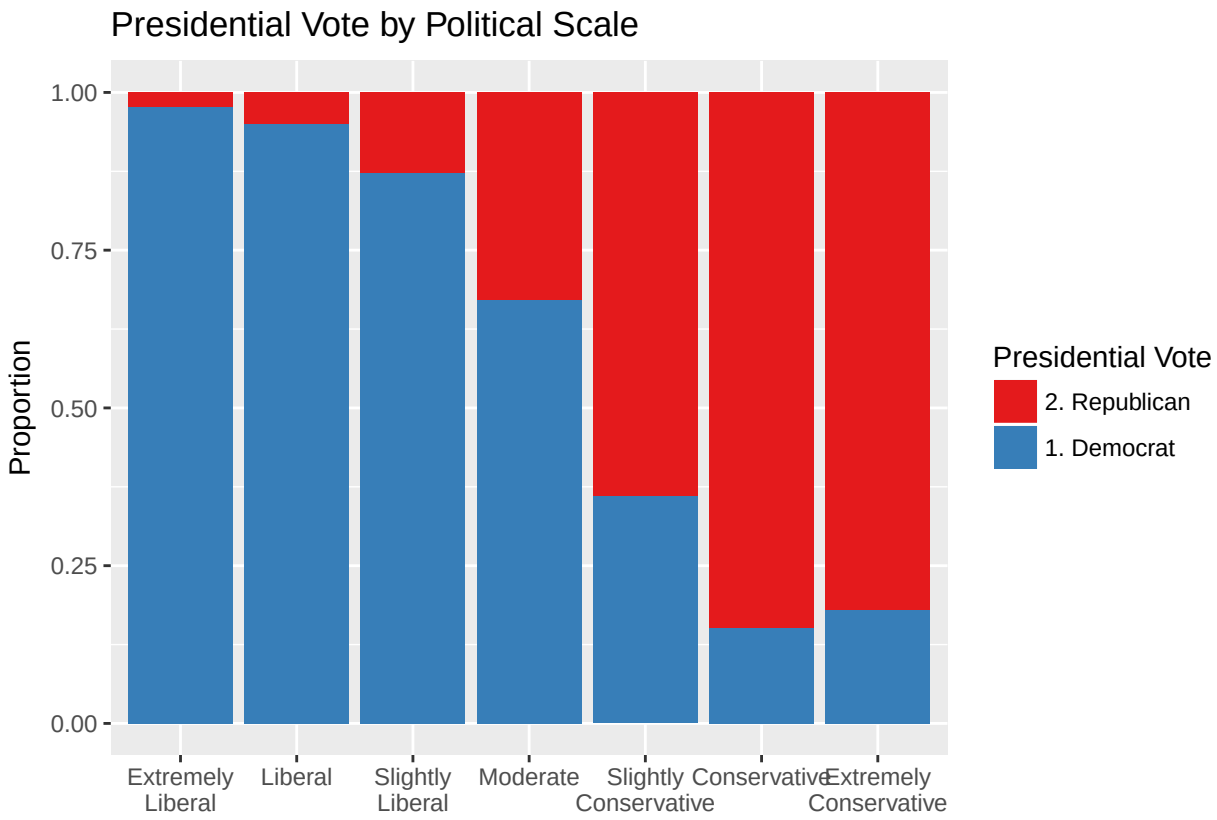


Another noteworthy observation is the Democratic candidate won the popular vote in all elections in our dataset. Consequently, our model could potentially have a Democratic bias,

There is a variable (`political scale`) I want to incorporate into one of my visualizations, but it is in another data frame. I will attempt to merge the data.

```
library(dplyr)
library(readstata13)
anes.data = read.dta13('anes_timeseries_cdf.dta', generate.factors = T, nonint.factors = T)
names(anes.data)[names(anes.data) == 'VCF0803'] <- 'political.scale'
names(anes.data)[names(anes.data) == 'VCF0006a'] <- 'id'
```

```r
temp.data = data.frame(id = anes.data$id, political.scale = anes.data$political.scale)

join.data = inner_join(voted.data, temp.data, by = 'id')
join.data$political.scale[as.numeric(join.data$political.scale) %in% c(1, 9)] <- NA
join.data$political.scale = factor(join.data$political.scale)
join.data$pres.vote = factor(join.data$pres.vote)

join.data = join.data[!is.na(join.data$political.scale), ]

#Presidential vote by political scale

levels(join.data$political.scale) =  c('Extremely\nLiberal', 'Liberal', 'Slightly\nLiberal', 'Moderate',
                                       'Slightly\nConservative', 'Conservative', 'Extremely\nConservative')

ggplot(join.data, aes(political.scale)) +
  geom_bar(aes(fill = pres.vote), position = 'fill') + labs(title = 'Presidential Vote by Political Scale',
  y = 'Proportion', x = '', fill = 'Presidential Vote') + scale_fill_brewer(palette = 'Set1', breaks = levels(join.data$pres.vote))
```



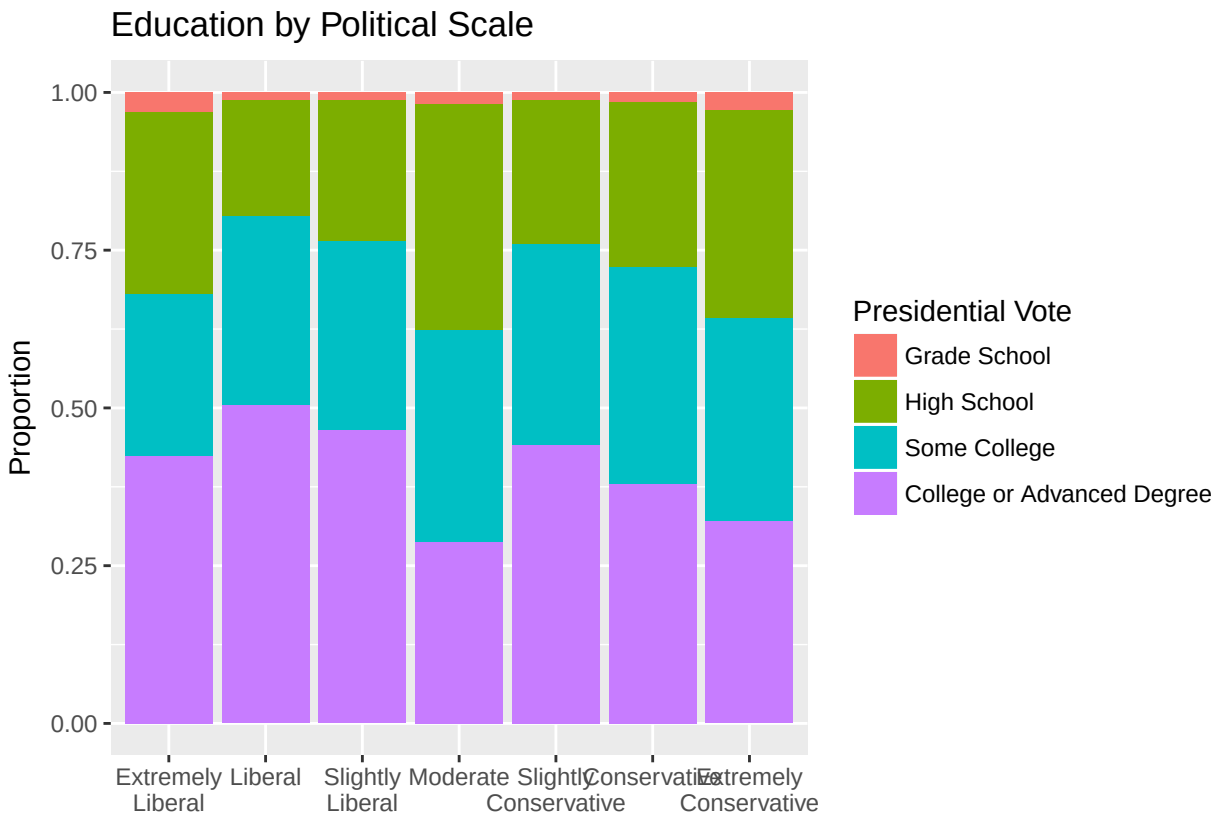

```
presvote.by.polscale = xtabs(~ political.scale + pres.vote, data = join.data)
prop.table(presvote.by.polscale, margin = 1)
```

```
##                          pres.vote
## political.scale           2. Republican 1. Democrat
##   Extremely\nLiberal         0.02316602  0.97683398
##   Liberal                    0.04946996  0.95053004
##   Slightly\nLiberal          0.12788462  0.87211538
##   Moderate                   0.32894737  0.67105263
##   Slightly\nConservative     0.63882784  0.36117216
##   Conservative               0.84934277  0.15065723
```

```
##   Extremely\nConservative    0.81989247  0.18010753
```

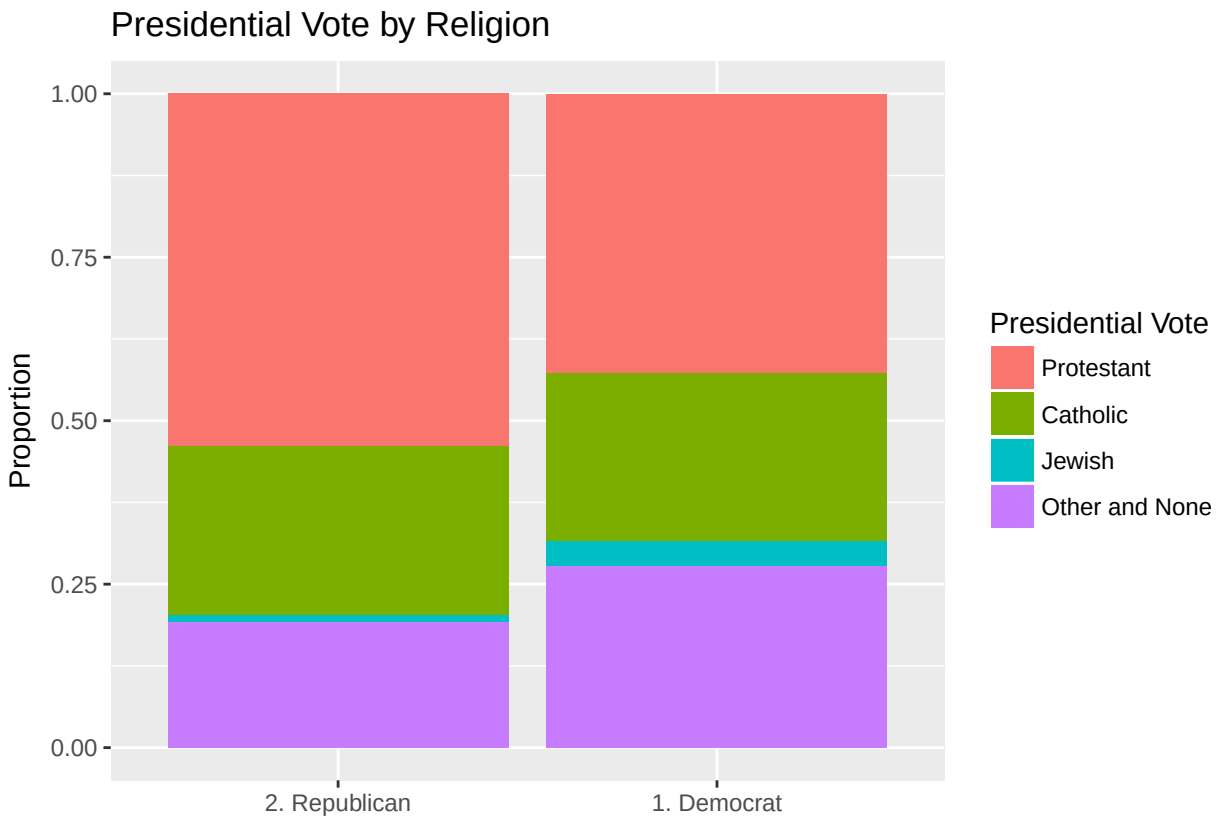
Unsurprisingly, there is a strong relationship between political scale and how an individual voted. But as mentioned earlier, there does appear to be a noticable democratic bias in the data. We can see this because among extreme liberals, 97.7% voted for the Democratic candidate while only 82% of the extreme conservatives voted for the Republican candidate.

```
#Education by political scale

levels(join.data$educ) =  c('Grade School', 'High School', 'Some College', 'College or Advanced Degree')
join.data = join.data[!is.na(join.data$educ), ]

ggplot(join.data, aes(political.scale)) +
  geom_bar(aes(fill = educ), position = 'fill') + labs(title = 'Education by Political Scale',
  y = 'Proportion', x = '', fill = 'Presidential Vote')
```



Interestingly, the moderates in our data appear to be the least educated, and liberals appear to be a bit more educated than conservatives.

```
#Presidential Vote by Religion

levels(join.data$religion) =  c('Protestant', 'Catholic', 'Jewish', 'Other and None')
join.data = join.data[!is.na(join.data$religion), ]

ggplot(join.data, aes(pres.vote)) +
  geom_bar(aes(fill = religion), position = 'fill') + labs(title = 'Presidential Vote by Religion',
  y = 'Proportion', x = '', fill = 'Presidential Vote')
```

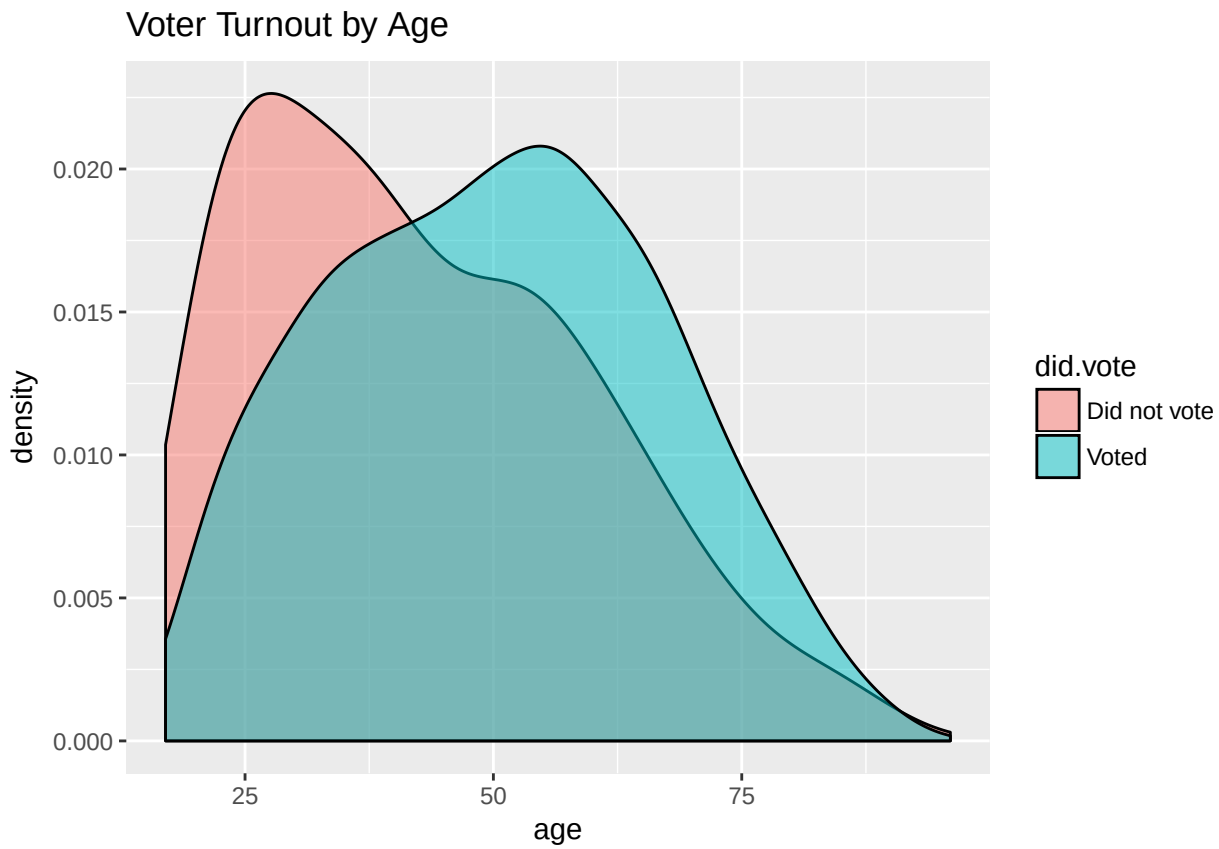


While there is a noticable difference between how individuals voted with respect to their religion, it's not as substantial as one might expect.

I'm tired of all of these barplots! Are there any other plots we can make?! Although we are fairly limited with regards to the kinds of plots we can make due to the fact that almost all of the data is categorical, we do have one noteworthy numerical variable we can examine.

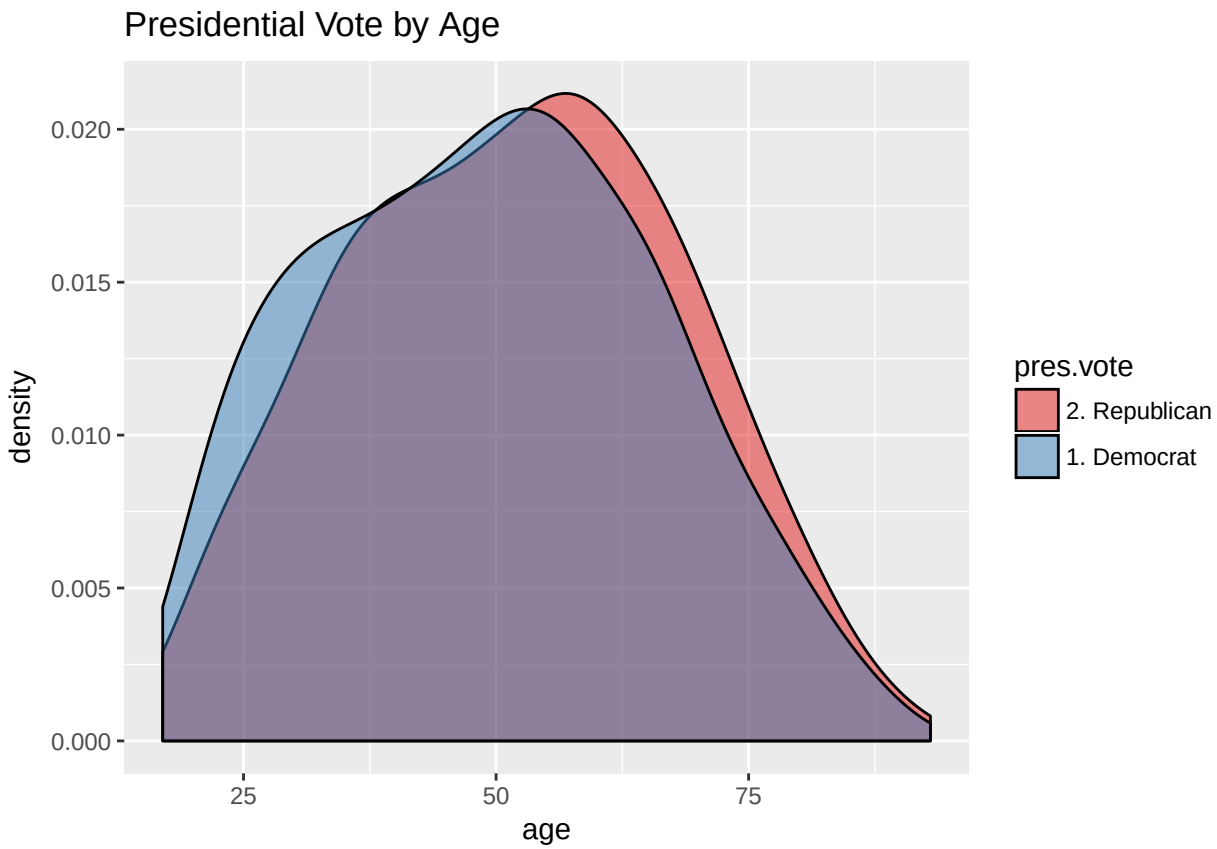
```
#Voter Turnout by age
ggplot(data, aes(age)) +
  geom_density(aes(fill = did.vote), col = 'black', alpha = .5, adjust = 1.5, position = 'identity') +
  ggtitle('Voter Turnout by Age')
```

```
## Warning: Removed 103 rows containing non-finite values (stat_density).
```




```
#Presidential Vote by age
ggplot(voted.data, aes(age)) +
  geom_density(aes(fill = pres.vote), col = 'black', alpha = .5, adjust = 1.5, position = 'identity') +
  ggtitle('Presidential Vote by Age') + scale_fill_brewer(palette = 'Set1')
```

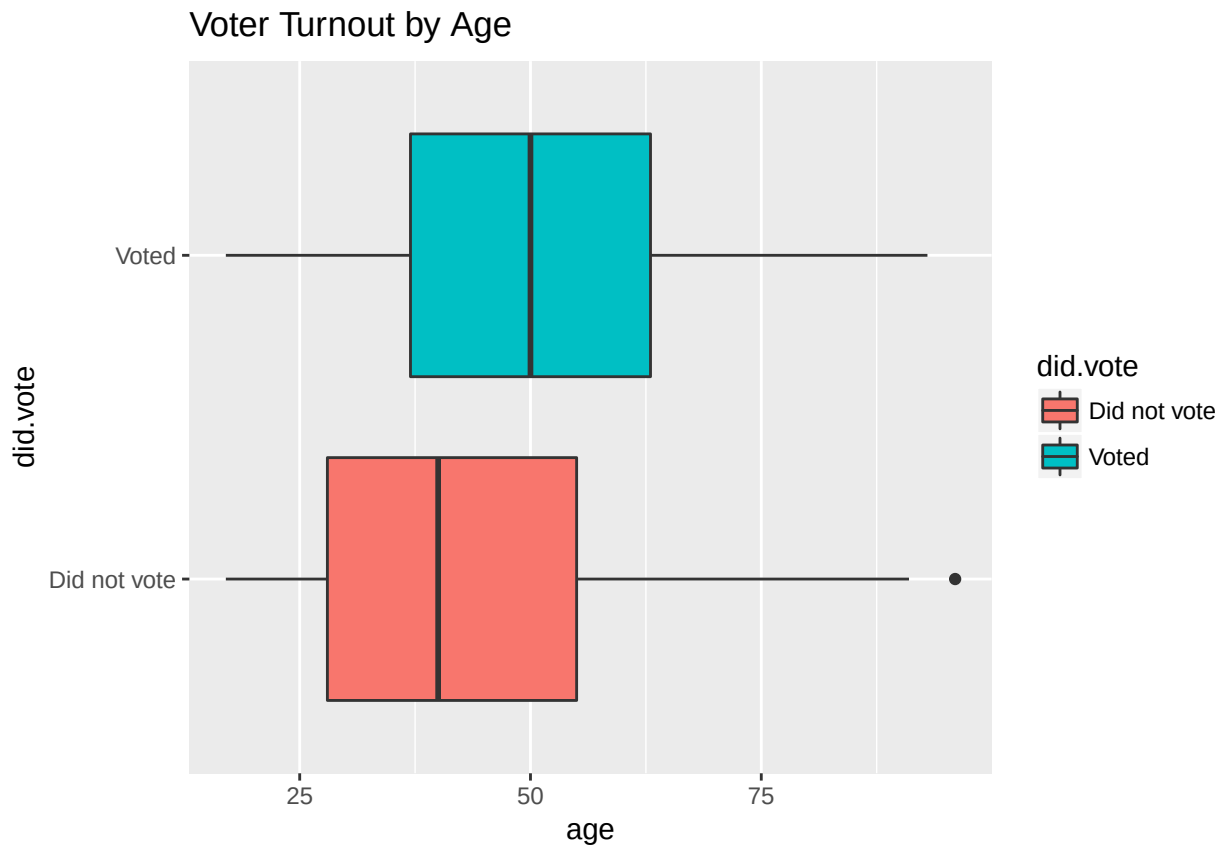
```
## Warning: Removed 65 rows containing non-finite values (stat_density).
```



It appears that voters are noticably older than non-voters, and Republican voters are a little older than Democratic ones, but the difference is not as significant. We can also visualize the same information in a side by side boxplot!

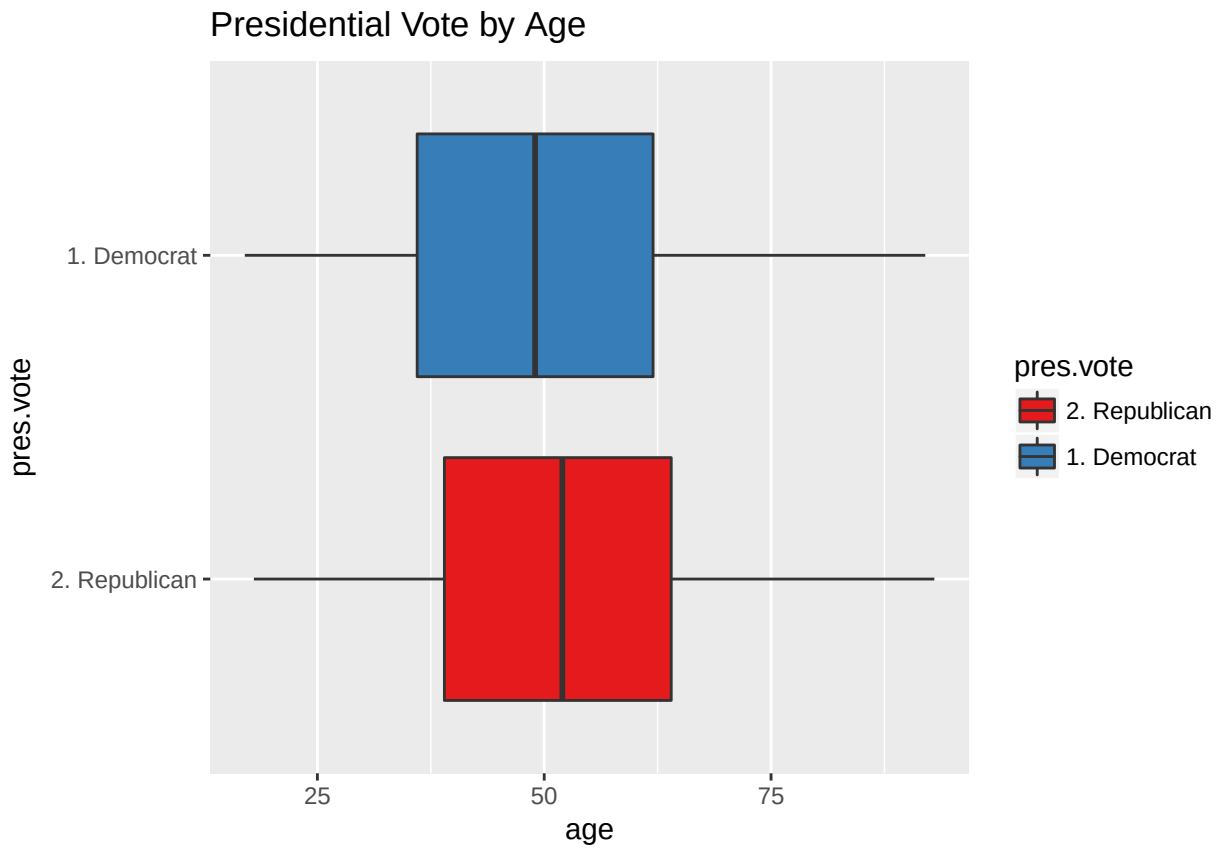
```
ggplot(data, aes(x = did.vote, y = age)) + geom_boxplot(aes(fill = did.vote)) +
  coord_flip() + ggtitle('Voter Turnout by Age')
```

```
## Warning: Removed 103 rows containing non-finite values (stat_boxplot).
```




```
ggplot(voted.data, aes(x = pres.vote, y = age)) + geom_boxplot(aes(fill = pres.vote)) +
  coord_flip() + ggtitle('Presidential Vote by Age') + scale_fill_brewer(palette = 'Set1')
```

```
## Warning: Removed 65 rows containing non-finite values (stat_boxplot).
```



Alright, that's enough data exploration for now!