Who's going to vote?

Alex Katopodis

```
library(tidyverse)
library(broom)
library(knitr)
library(yardstick)
```

```
voters_data <- read_csv("data/nonvoters_data.csv")
```

Previewing the Data

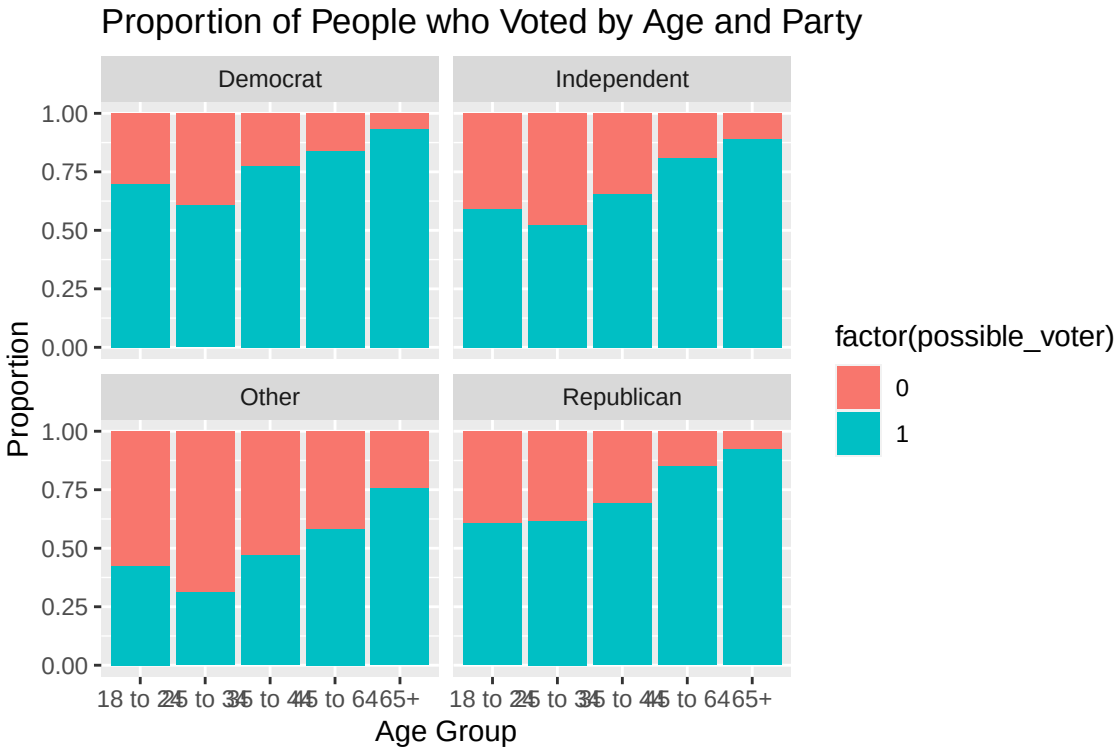
The data collected in 2019 contains information about 5836 voters who were registered to vote in at least 4 elections (defined as presidential or midterm elections).

Before previewing it, a bit of work is done to clean it up. Political affiliations are narrowed down to Republican, Democrat, Independent, and Other, and voter_category is changed to 1 or 0 to reflect if a person votes frequently or not.

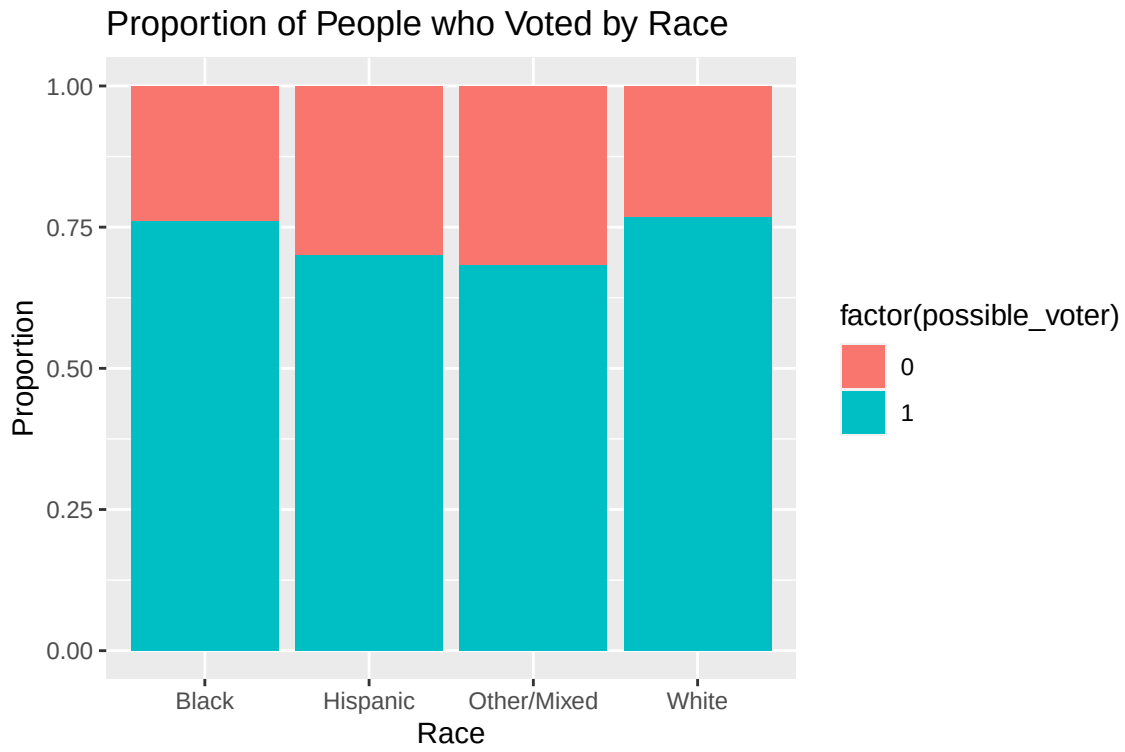
```
voters_data <- voters_data %>%
  mutate(party = case_when(
    Q30 == 1 ~ "Republican",
    Q30 == 2 ~ "Democrat",
    Q30 == 3 ~ "Independent",
    TRUE ~ "Other"
  ), possible_voter = if_else(
    voter_category == "always" | voter_category == "sporadic", 1,
    0
    ), mean_cent_age = ppage - mean(ppage),
  age_group = case_when(
    ppage <= 24 ~ "18 to 24",
    ppage <= 34 ~ "25 to 34",
    ppage <= 44 ~ "35 to 44",
    ppage <= 64 ~ "45 to 64",
    ppage >= 65 ~ "65+"
  ))
```

Now, we can begin to review the data. We will compare some attributes to voter category to make an early assessment of what categories would be important to include in a model.

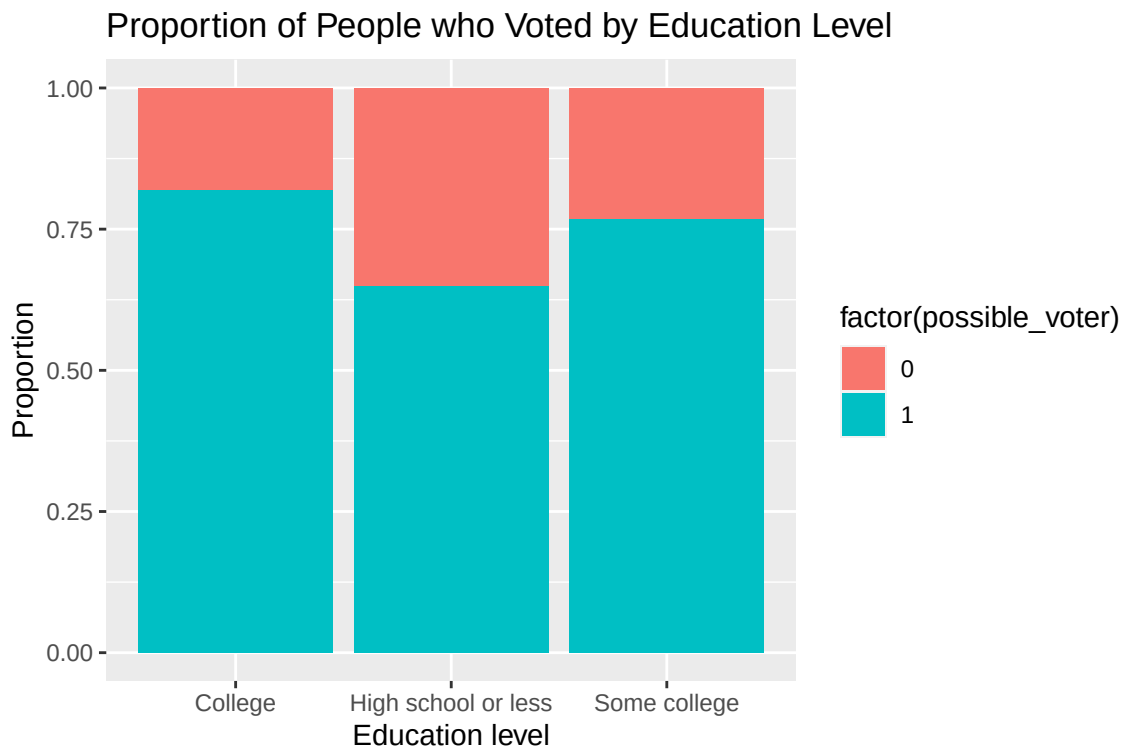
```
ggplot(data = voters_data, aes(x = age_group, fill = factor(possible_voter))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of People who Voted by Age and Party",
       y = "Proportion",
       x = "Age Group") +
  facet_wrap(~ party)
```



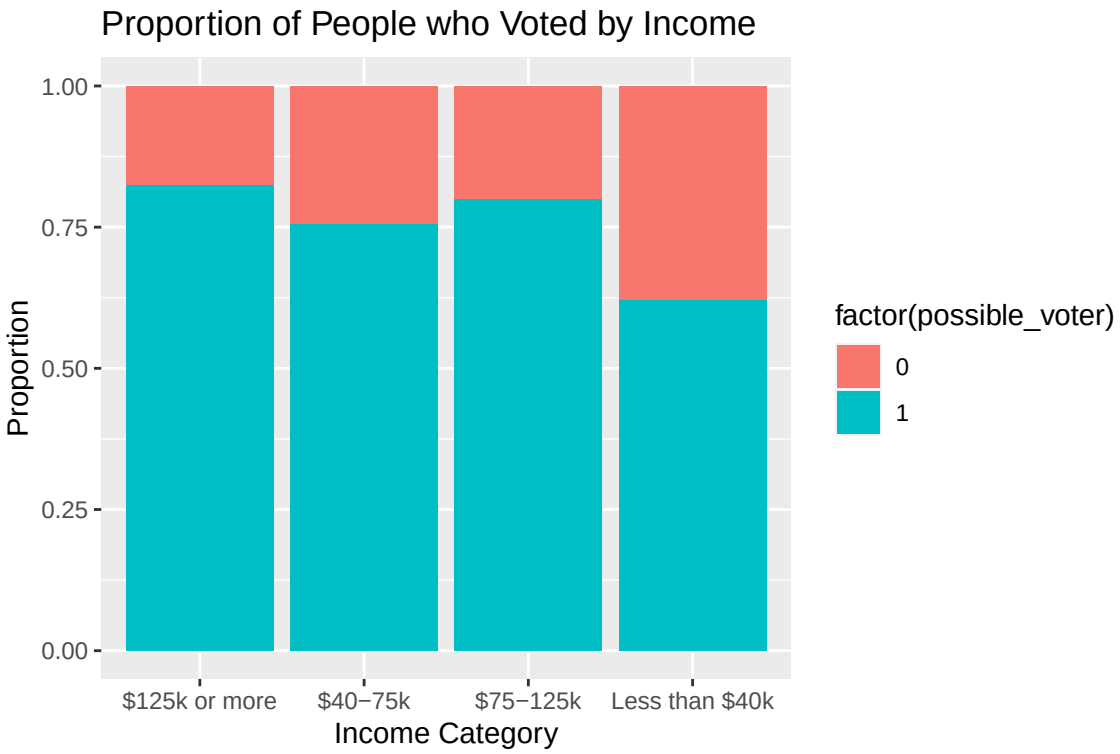

```
ggplot(data = voters_data, aes(x = race, fill = factor(possible_voter))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of People who Voted by Race",
       y = "Proportion",
       x = "Race")
```




```
ggplot(data = voters_data, aes(x = educ, fill = factor(possible_voter))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of People who Voted by Education Level",
       y = "Proportion",
       x = "Education level")
```




```
ggplot(data = voters_data, aes(x = income_cat, fill = factor(possible_voter))) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of People who Voted by Income",
       y = "Proportion",
       x = "Income Category")
```



We can see that as people get older, they generally become more inclined to vote. Also, people affiliated with any party vote in higher proportions than people unaffiliated with a party.

We also see that, in general, race does not immediately appear to drastically impact one's chances of voting. Higher incomes generally vote more often, and people with some college education vote more than their counterparts with no college education.

Making a model

```
pv_model <- glm(possible_voter ~ mean_cent_age + race + gender + income_cat +
                  educ, data = voters_data, family = binomial)
tidy(pv_model) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.130	0.125	17.017	0.000
mean_cent_age	0.052	0.002	24.087	0.000
raceHispanic	-0.134	0.120	-1.115	0.265
raceOther/Mixed	-0.480	0.149	-3.218	0.001
raceWhite	-0.066	0.097	-0.680	0.497
genderMale	-0.138	0.067	-2.070	0.038
income_cat$40-75k	-0.105	0.105	-1.005	0.315
income_cat$75-125k	0.054	0.101	0.541	0.589
income_catLess than $40k	-0.693	0.106	-6.535	0.000
educHigh school or less	-1.078	0.091	-11.893	0.000
educSome college	-0.378	0.089	-4.252	0.000

$H_0 : \beta_{Republican} = \beta_{Democrat} = \beta_{Independent} = \beta_{Other} = 0$ $H_a :$ at least one β_j is not equal to 0 $\alpha = 0.05$

```
pvm_full <- glm(possible_voter ~ mean_cent_age + race + gender + income_cat +
                  educ + party, data = voters_data, family = binomial)

# Reduced then Full
tidy(anova(pv_model, pvm_full, test = "Chisq")) %>%
  kable(digits = 3)
```

term	Resid..Df	Resid..Dev	df	Deviance	p.value
possible_voter ~ mean_cent_age + race + gender + income_cat + educ	5825	5574.125	NA	NA	NA
possible_voter ~ mean_cent_age + race + gender + income_cat + educ + party	5822	5436.443	3	137.682	0
	5825	5574.125	NA	NA	NA
	5822	5436.443	3	137.682	0

The drop in deviance test yields a p-value of 0. This is lower than the pre-established α of 0.05, so we can reject the null hypothesis. There is sufficient evidence to suggest that at least one coefficient of Party is NOT equal to 0.

Therefore, I will include Party as a predictor in my model going forward.

```
# New Model
tidy(pvm_full) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.308	0.130	17.743	0.000
mean_cent_age	0.050	0.002	22.680	0.000
raceHispanic	-0.075	0.123	-0.606	0.544
raceOther/Mixed	-0.424	0.152	-2.800	0.005
raceWhite	-0.004	0.103	-0.034	0.973
genderMale	-0.128	0.069	-1.866	0.062
income_cat$40-75k	-0.047	0.106	-0.445	0.656
income_cat$75-125k	0.069	0.102	0.677	0.499
income_catLess than $40k	-0.611	0.108	-5.665	0.000
educHigh school or less	-1.001	0.093	-10.807	0.000
educSome college	-0.308	0.090	-3.402	0.001
partyIndependent	-0.454	0.093	-4.901	0.000
partyOther	-1.093	0.100	-10.926	0.000
partyRepublican	-0.111	0.098	-1.136	0.256

Interpreting the Model

Some interpretations of the coefficents are given below:

The intercept of 2.308 represents the log odds that a NA year old, black, female, democrat with a college education and an income >$125k will vote. In terms of the odds, the intercept represents the $e^{2.308}$ (10.0543) odds that they will vote in an election.

The coefficient for age of 0.050 represents the change in the log odds that a person will vote for every 1 year increase in age. In terms of the odds, this coefficient represents the $e^{0.050}$ (1.05127) increase in odds that they will vote for a 1 year increase in age, holding all other factors constant.

Political party ID impacts the odds that a registered voter will vote. The baseline party is Democrat. Republicans, Independents, and "Others" all were LESS likely to vote than their democratic counterparts. More specifically, Independents had 0.63508 times the odds of voting compared to democrats, Republicans 0.89494 times the odds, and Others 0.33521 times the odds, holding all other factors constant.

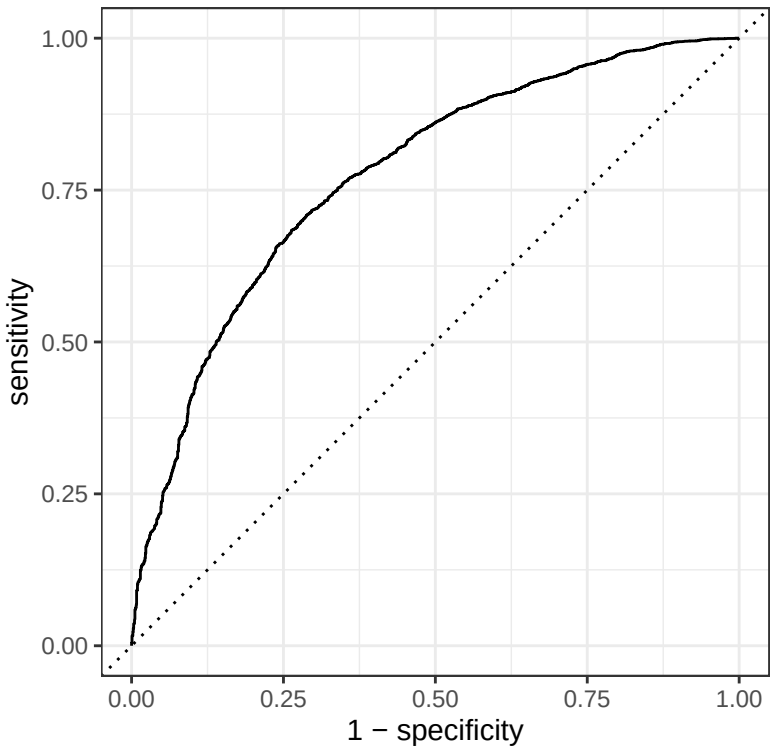
Overall, the model suggests that the people most likely to vote are older black males with a high income and level of education.

Assessing Model Performance

```
model_aug <- augment(pvm_full, type.predict = "response") %>%
  mutate(possible_voter = factor(possible_voter, levels = c("1", "0")))

roc_curve_data <- model_aug %>%
  roc_curve(possible_voter, .fitted)

autoplot(roc_curve_data)
```




```
model_aug %>%
  roc_auc(possible_voter, .fitted) %>%
  pull(.estimate)
```

```
## [1] 0.7738068
```

Based off of the ROC curve, it appears that my model is decently fit to the data. It has an AUC of 0.7738, which is not close enough to 0.5 that the model is poorly fit, but also not close enough to 1 that we'd say it's highly fit to the data. It could likely be improved upon, but it's definitely a solid model considering it is using so few predictors.

Making Predictions and Other Thoughts

We will test the model on data using 0.6 as the cutoff threshold for classifying a voter. The 0.6 coming from the results of the ROC curve above.

```
model_aug <- model_aug %>%
  mutate(pred_status = if_else(.fitted > .6, "Possible Voter", "Not a Possible Voter"))

model_aug %>%
  count(possible_voter, pred_status)
```

```
## # A tibble: 4 x 3
##   possible_voter pred_status              n
##   <fct>          <chr>                <int>
## 1 1              Not a Possible Voter   477
## 2 1              Possible Voter        3908
## 3 0              Not a Possible Voter   636
## 4 0              Possible Voter         815
```

```
tp <- 3908 / (3908 + 477)
tn <- 636 / (636 + 815)

tp
```

```
## [1] 0.8912201
```

```
1 - tn
```

```
## [1] 0.5616816
```

True Positive Rate: 89%

False Positive Rate: 56%

My model has a true positive rate of 0.89. In the context of the data, this means that if a person truly is a possible voter, there is a 89% chance that my model will classify them CORRECTLY.

My model has a false positive rate of 0.56. In the context of the data, this means that if a person is NOT truly a possible voter, there is 56% chance my model will classify them INCORRECTLY.

The false positive rate is very high, implying the model is not specific This could be for a number of reasons, but changes to make the model in the future could be including more predictors and using AIC/BIC to decide which ones are useful and gathering more data to train on. The data is quite limited at only a few thousand entries, so more data could lead to a more accurate model. Also, deciding to vote is a complex decision. Perhaps there are far more factors involved, such as how accessible a polling place is, a voters thoughts on the current administration, etc. Using such few predictors likely massively oversimplifies the decision.