Why Americans Don't Vote

FiveThirtyEight Figure Recreation (BST 270 Final Project)

Kevin Kapner

January 19 2024

Introduction

This R notebook walks through all of the necessary steps for figure recreation of the first two figures in the article from FiveThirtyEight titled: *Why Many Americans Don't Vote* which can be found here. Associated data used in this analysis (namely the nonvoters_data.csv) can be found on the public GitHub repository located here.

Please see the sessionInfo section at the end of the notebook for all packages, associated versions, and computer architecture used.

Loading Data

We will begin by loading in data wrangling and visualization packages.

```
library("dplyr")
library("ggplot2")
library("reshape2")
```

The data is located in the **data** directory in the main project directory.

```
no_vote_data <- read.csv(file.path("..", "data", "nonvoters_data.csv"),
                         stringsAsFactors = TRUE)
```

Figure 1 Generation

To recreate the first figure, **Those who almost always vote and those who sometimes vote aren't that different** we will need the following variables (which can be identified using the nonvoters_codebook.pdf in the **data** directory) and their associated interpretations are in parenthesis:

- voter_category (Voter Class)
- educ (Education)
- race (Race)
- income_cat (Income)
- ppage (Age)
- Q33 (Party ID)
- RespId (Respondent ID)

```
req_col <- c("RespId", "voter_category", "educ", "race", "income_cat", "ppage", "Q33")

fig1_data <- no_vote_data %>%
  dplyr::select(all_of(req_col))
```

By looking in the codebook, we see that for Q33 (Party ID), a value of 1 corresponds to the republican party and 2 corresponds to the democratic party. However, we see below that are 2 other data value options: a missing value and a -1. -1 will be assumed to be the independent/neither category. *Missing data (in any category) was removed.*

```
unique(fig1_data$Q33)
```

```
## [1] NA  1 -1  2
```

We also note that the ppage variable needs to be changed from a continuous variable to a categorical variable with the following age breakdowns:

- 26 - 34
- 35 - 49
- 50 - 64
- 65 +

Additionally, all of the categorical variables need to be factorized with the proper leveling such that we can obtain the same order as those in the published figure. Colors were extracted from the original figure using the Image Picker tool on the Coolors website (tool found here).

```
# Using rev to make leveling easier as coord_flip changes the orientation.
# This way we can just write them the same order the appear in the final
# figure.
value_levels <- rev(c("Black", "Hispanic", "Other/Mixed", "White", "Less than $40k",
                      "$40-75k", "$75-125k", "$125k or more", "26-34", "35-49", "50-64",
                      "65+", "High school or less", "Some college", "College",
                      "Democratic", "Independent/Neither", "Republican"))

fig1_data %>%
  # Categorizing age
  dplyr::mutate(ppage = dplyr::case_when(
    (ppage >= 26) & (ppage <= 34) ~ "26-34",
    (ppage >= 35) & (ppage <= 49) ~ "35-49",
    (ppage >= 50) & (ppage <= 64) ~ "50-64",
    ppage >= 65 ~ "65+"
  )) %>%
  # Changing party affiliation
  dplyr::mutate(Q33 = dplyr::case_when(
    Q33 == -1 ~ "Independent/Neither",
    Q33 == 1 ~ "Republican",
    Q33 == 2 ~ "Democratic"
  )) %>%
  # Reformatting table for easier use with ggplot
  reshape2::melt(id.var = c("RespId", "voter_category")) %>%
  dplyr::mutate(value = factor(value,
                               levels = value_levels)) %>%
```

```r
# Adjusting factors for individual plots
dplyr::mutate(variable = factor(variable,
                                levels = c("race", "income_cat", "ppage", "educ", "Q33"),
                                labels = c("Race", "Income", "Age", "Education", "Party ID"))) %>%
# Adjusting factors for voting category
dplyr::mutate(voter_category = factor(voter_category,
                                      levels = c("always", "sporadic", "rarely/never"),
                                      labels = c("Almost Always Vote",
                                                 "Sometimes Vote",
                                                 "Rarely or Never Vote"))) %>%
na.omit() %>%
ggplot(aes(x = value, fill = voter_category)) +
geom_bar(position = "fill", na.rm = TRUE) +
coord_flip() +
facet_wrap(vars(variable), scales = "free", drop = TRUE) +
scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1),
                   labels = c("0%", "25", "50", "75", "100")) +
# Matching colors from published figure
scale_fill_manual(values = c("#AF4CFF", "#E183AF", "#FEC11A"),
                  breaks = c( "Rarely or Never Vote",
                              "Sometimes Vote", "Almost Always Vote")) +
theme_minimal() +
theme(axis.title.y = element_blank(),
      legend.position = "top",
      legend.title = element_blank(),
      axis.title.x = element_blank())
```

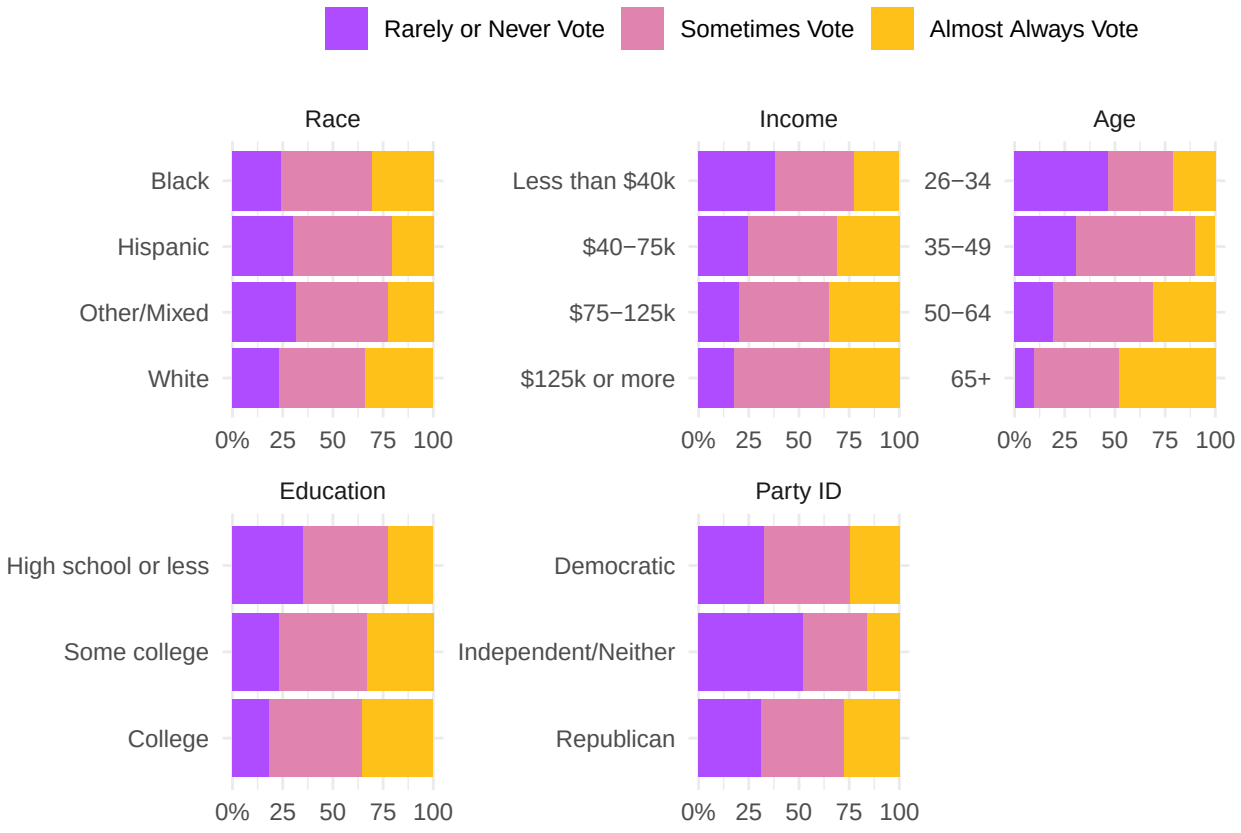


Figure 2 Generation

To recreate the second figure, **All types of voters report experiencing barriers** we will need the following variables (which can be identified using the nonvoters_codebook.pdf in the **data** directory) and their associated interpretations are in parenthesis:

- voter_category (Voter Class)
- RespId (Respondent ID)

We will additionally need all Q18_X columns, where X is a number 1-10. The mapping from number to question is as follows:

1. Was told they did not have the correct identification
2. Could not find the polling place
3. Missed the voter registration deadline
4. Was unable to physically access the polling place
5. Could not obtain necessary assistance to fill out a ballot
6. Had to cast a provisional ballot
7. Couldn't get off work to vote when polls were open
8. Waited in line to vote for more than an hour
9. Was told name was not on the list even though they were registered
10. Did not receive absentee or mail-in ballot in time

```
req_col_2 <- c("RespId", "voter_category", paste0("Q18_", 1:10))

fig2_data <- no_vote_data %>%
  dplyr::select(all_of(req_col_2))
```

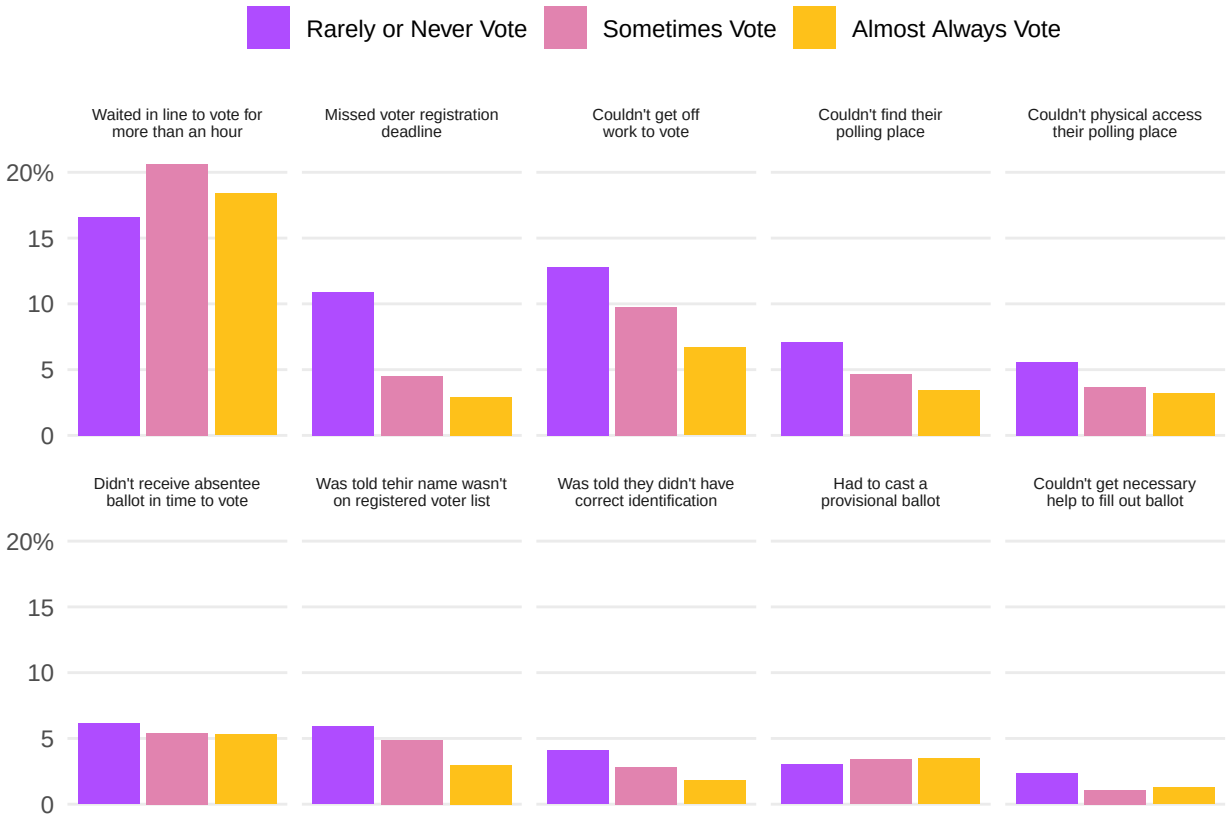
Using the codebook, we see that for all of the Q18 data values, a 1 corresponds to "Yes" and a 2 corresponds to "No". In the published figure, the reported values are the percentages of respondents who faced the obstacle (which would be a response of "Yes" to the question).

Additonally, the questions need to be reordered to match the same order as the figure.

```
question_levels <- c("Q18_8", "Q18_3", "Q18_7", "Q18_2", "Q18_4",
                     "Q18_10","Q18_9", "Q18_1", "Q18_6", "Q18_5")
question_labels <- c("Waited in line to vote for\nmore than an hour",
                     "Missed voter registration\ndeadline",
                     "Couldn't get off\nwork to vote",
                     "Couldn't find their\npolling place",
                     "Couldn't physical access\ntheir polling place",
                     "Didn't receive absentee\nballot in time to vote",
                     "Was told tehir name wasn't\non registered voter list",
                     "Was told they didn't have\ncorrect identification",
                     "Had to cast a\nprovisional ballot",
                     "Couldn't get necessary\nhelp to fill out ballot")

fig2_data %>%
  reshape2::melt(id.var = c("RespId", "voter_category")) %>%
  # Need to calculate percentage of responsdents per group with a Yes (value of 1)
  dplyr::group_by(voter_category, variable) %>%
  dplyr::mutate(percentageYes = sum(value == 1)/n()) %>%
  # Can safely discard No responses now and don't need individual level data
```

```
  dplyr::filter(value == 1) %>%
  dplyr::select(-c("RespId")) %>%
  unique() %>% # Collapsing down for gropu level
  dplyr::mutate(voter_category = factor(voter_category,
                                        levels = c("rarely/never", "sporadic", "always"),
                                        labels = c("Rarely or Never Vote",
                                                   "Sometimes Vote",
                                                   "Almost Always Vote"))) %>%
  dplyr::mutate(variable = factor(variable,
                                  levels = question_levels,
                                  labels = question_labels)) %>%
  na.omit() %>%
  ggplot(aes(x = voter_category, y = percentageYes, fill = voter_category)) +
  # Matching colors from publisehd figure
  scale_fill_manual(values = c("#AF4CFF", "#E183AF", "#FEC11A"),
                    breaks = c( "Rarely or Never Vote",
                                "Sometimes Vote", "Almost Always Vote")) +
  geom_bar(stat = "identity") +
  scale_y_continuous(breaks = c(0, 0.05, 0.1, 0.15, 0.2),
                     labels = c("0", "5", "10", "15", "20%")) +
  facet_wrap(vars(variable), ncol = 5) +
  theme_minimal() +
  theme(axis.title.y = element_blank(),
        legend.position = "top",
        legend.title = element_blank(),
        axis.title.x = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.text.x = element_blank(),
        strip.text = element_text(size = 6))
```



Conclusion

Overall the figures were reproducible with the provided data. There appear to be some minor discrepancies in the percentage values for Figure 1. I believe this is due to the way in which missing values for handled (I dropped rows with missing values), whereas they may have had some other procedure for handling them. Even with these differences, the trends are all identical and thus the analysis (for these two figures) was reproducible in my hands.

sessionInfo

```
## R version 4.3.1 (2023-06-16)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.2.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
## [1] reshape2_1.4.4 ggplot2_3.4.3  dplyr_1.1.2
##
## loaded via a namespace (and not attached):
##  [1] vctrs_0.6.3       cli_3.6.1         knitr_1.43        rlang_1.1.1
##  [5] xfun_0.39         highr_0.10        stringi_1.7.12    generics_0.1.3
##  [9] glue_1.6.2        colorspace_2.1-0  plyr_1.8.8        htmltools_0.5.6
## [13] scales_1.2.1      fansi_1.0.4       rmarkdown_2.24    grid_4.3.1
## [17] munsell_0.5.0     evaluate_0.21     tibble_3.2.1      fastmap_1.1.1
## [21] yaml_2.3.7        lifecycle_1.0.3   stringr_1.5.0     compiler_4.3.1
## [25] Rcpp_1.0.11       pkgconfig_2.0.3   rstudioapi_0.15.0 farver_2.1.1
## [29] digest_0.6.33     R6_2.5.1          tidyselect_1.2.0  utf8_1.2.3
## [33] pillar_1.9.0      magrittr_2.0.3    withr_2.5.0       tools_4.3.1
## [37] gtable_0.3.3
```