Data Analysis for Cornell Votes

Adriana Lorena Jimenez Bonilla

2024-12-01

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(viridis)
```

```
Loading required package: viridisLite
```

```
library(scales)
```

```

Attaching package: 'scales'

The following object is masked from 'package:viridis':

    viridis_pal

The following object is masked from 'package:purrr':
```

```
    discard

The following object is masked from 'package:readr':

    col_factor
```

```
library(readxl)
library(ggplot2)
```

Introduction

This project focuses on analyzing and communicating insights from Cornell's NSLVE (National Study of Learning, Voting, and Engagement) reports. These reports provide data on student voter participation, serving as a foundation to strategize and enhance voting efforts on campus. The ultimate goal is to ensure Cornell's voting rate remains at or above the average voting rates of similar institutions in NSLVE. In this case, similar institutions consist of private research institutions.

Data Description

The data analyzed in this project comes from Cornell University's NSLVE (National Study of Learning, Voting, and Engagement) reports for the years 2012-2020. These reports provide detailed insights into student voter participation, including registration rates, voting rates, and voting methods. The data is disaggregated by demographic and academic characteristics such as age, gender, field of study, and enrollment status. Adjustments have been made to exclude ineligible voters, such as non-resident aliens, based on institutional records. The reports compare Cornell's performance to averages across similar institutions, which is where our primary focus lies.

The objective of this project was to process and harmonize voting-related datasets across multiple election years to enable the comparison of demographic statistics between 2012 and 2022. Two datasets containing 2016 and 2018 data contained conflicting but similar values. To resolve these inconsistencies, we averaged the numeric columns by computing the mean of corresponding values.

To prepare for combining all datasets into a single dataframe, we added a new column named `year` to each dataset to indicate the corresponding election year. This ensured the data could be differentiated after merging. The unified dataset is now ready for further analysis, such as comparing voting rates, registration rates, and demographic trends over time.

A description of each variable (column) is as follows:

- corn_reg_rate: the percentage of voting-eligible students who registered to vote.

- corn_vote_rate: the percentage of eligible students who voted on Election Day.
- goal_rate: the voting rate of all research institutions, our goal voting rate.
- method_nonperson: the percentage of people who voted absentee/by mail.
- method_early: the percentage of people who voted early.
- method_person: the percentage of people who voted in person.
- method_unknown: the percentage of people whose voting method is unknown.
- age1: the voting rate of people between ages 18-21.
- age2: the voting rate of people between ages 22-24.
- age3: the voting rate of people between ages 25-29.
- age4: the voting rate of people between ages 30-39.
- age5: the voting rate of people between ages 40-49.
- age6: the voting rate of people aged 50+.
- ugrad: the voting rate of undergraduate students.
- grad: the voting rate of graduate students.
- fresh: the voting rate for first-year undergraduates (freshmen).
- soph: the voting rate for second-year undergraduates (sophomores).
- up: the voting rate for upperclass undergraduates (juniors and seniors).

```
data_2012 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2012",
    na = "null"
  )

data_2014 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2014",
    na = "null"
  )

data_20161 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2016-1",
    na = "null"
  )

data_20162 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2016-2",
```

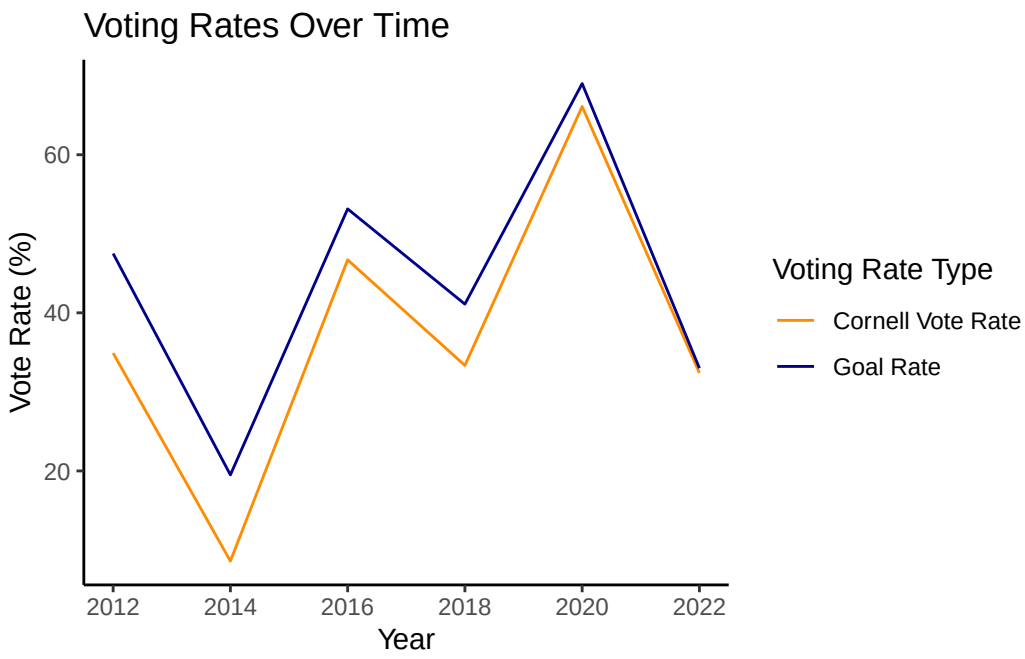
```
    na = "null"
  )

data_20181 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2018-1",
    na = "null"
  )

data_20182 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2018-2",
    na = "null"
  )

data_2020 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2020",
    na = "null"
  )

data_2022 <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Yearly Voting Data.xlsx",
    sheet = "voting-2022",
    na = "null"
  )

data_fos <-
  read_excel(
    "~/Desktop/Downloads/Einhorn Center/Voting Data Project/Field of Study Voting Data.xlsx"
    na = "null"
  )


data_2012$year <- 2012
data_2014$year <- 2014
data_20161$year <- 2016
data_20162$year <- 2016
```

```r
data_20181$year <- 2018
data_20182$year <- 2018
data_2020$year <- 2020
data_2022$year <- 2022

data_2016 <- data_20161 %>%
  mutate(across(everything(), ~ (. + data_20162[[cur_column()]]) / 2))

data_2018 <- data_20181 %>%
  mutate(across(everything(), ~ (. + data_20182[[cur_column()]]) / 2))

combined_data <- rbind(data_2012,
                       data_2014,
                       data_2016,
                       data_2018,
                       data_2020,
                       data_2022)
```

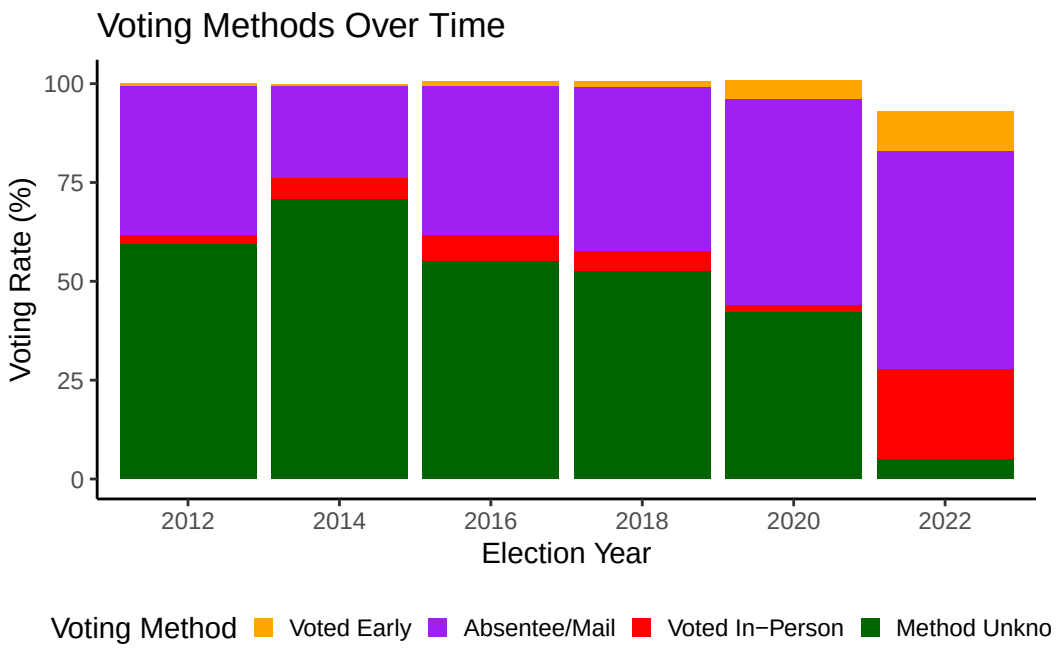
Data Analysis

Research Question: How can Cornell University sustain or improve its voter turnout rates to remain at or above the average voting rates of private research institutions?

```r
combined_data_long <- combined_data %>%
  gather(key = "rate_type", value = "vote_rate", corn_vote_rate, goal_rate)

ggplot(data = combined_data_long, aes(x = year, y = vote_rate, color = rate_type)) +
  geom_line() +
  scale_color_manual(
    values = c("corn_vote_rate" = "darkorange",
               "goal_rate" = "darkblue"),
    labels = c("Cornell Vote Rate", "Goal Rate")  # Use clear English for legend labels
  ) +
  labs(title = "Voting Rates Over Time",
       x = "Year",
       y = "Vote Rate (%)",
       color = "Voting Rate Type") +  # Title for the legend
  scale_x_continuous(
    breaks = combined_data$year
  ) + theme_classic()
```




```r
voting_method_long <- pivot_longer(
  combined_data,
  cols = c(method_person, method_early, method_nonperson, method_unknown),
                                  names_to = "voting_method", values_to = "vote_rate")

label_mapping <- c(
  "method_early" = "Voted Early",
  "method_nonperson" = "Absentee/Mail",
  "method_person" = "Voted In-Person",
  "method_unknown" = "Method Unknown"
)

ggplot(voting_method_long, aes(x = factor(year), y = vote_rate, fill = voting_method)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("method_early" = "orange",
                               "method_nonperson" = "purple",
                               "method_person" = "red",
                               "method_unknown" = "darkgreen"),
                    labels = label_mapping) +
  labs(title = "Voting Methods Over Time",
       x = "Election Year", y = "Voting Rate (%)", fill = "Voting Method") +
  theme_classic() +
  theme(legend.position = "bottom", legend.key.size = unit(0.30, "cm"))
```



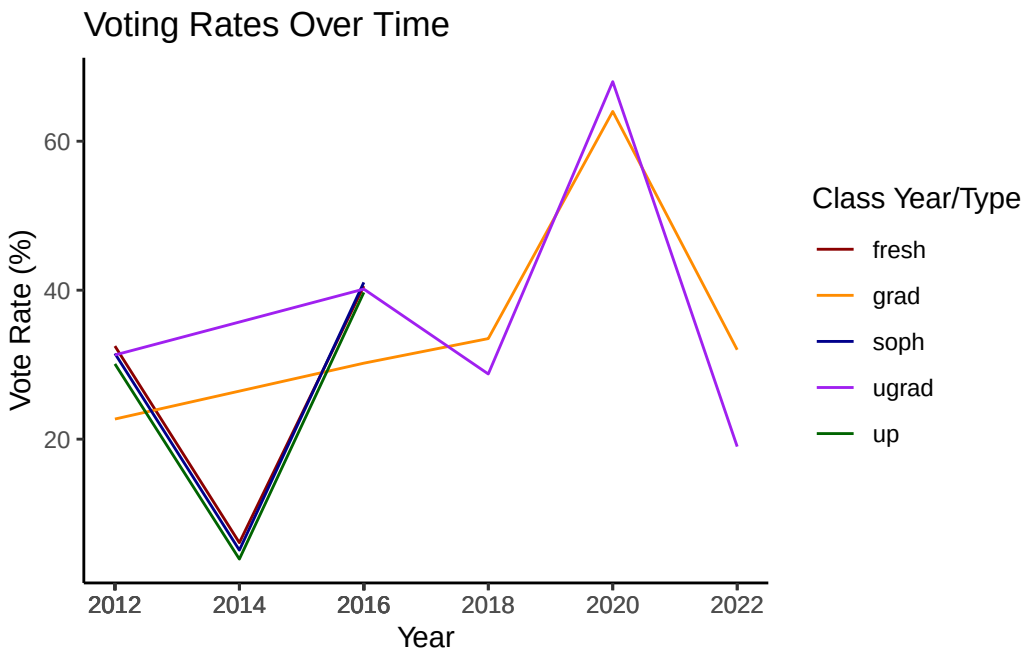
While Cornell Votes makes an active effort not to persuade prospective voters toward any one voting method, it may be beneficial to consider trends among students so we can best support whatever methods they choose.

```
combined_data_long <- combined_data %>%
  gather(key = "age_group", value = "vote_rate", age1, age2, age3, age4, age5, age6)

ggplot(data = combined_data_long, aes(x = year, y = vote_rate, color = age_group)) +
  geom_line() +
  scale_color_manual(
    values = c("age1" = "darkred",
               "age2" = "darkblue", "age3" = "darkgreen", "age4" = "purple", "age5" = "hotpi
    labels = c("18-21", "22-24", "25-29", "30-39", "40-49", "50+")
  ) +
  labs(title = "Voting Rates Over Time",
       x = "Year",
       y = "Vote Rate (%)",
       color = "Age Group") +
  scale_x_continuous(
    breaks = combined_data_long$year
  ) + theme_classic()
```

Let's focus on two very general trends. The youngest and oldest age groups, ages 18-21 and 50+, see a sharp decrease in voting rate during midterm elections. Meanwhile, age groups in between were consistently increasing (although less rapidly), until a uniform drop from the 2020 election to the 2022 election. Deeper insights may be available when 2024 data is accessible.

```
combined_data_long <- combined_data %>%
  gather(key = "class", value = "vote_rate", fresh, soph, up, ugrad, grad) |> filter(!is.na(v

ggplot(data = combined_data_long, aes(x = year, y = vote_rate, color = class, group = class))
  geom_line() +
  labs(title = "Voting Rates Over Time",
       x = "Year",
       y = "Vote Rate (%)",
       color = "Class Year/Type") +
  scale_color_manual(
    values = c("fresh" = "darkred",
               "soph" = "darkblue",
               "up" = "darkgreen",
               "ugrad" = "purple",
               "grad" = "darkorange")
  ) +
  scale_x_continuous(
```

```
    breaks = combined_data_long$year
  ) + theme_classic()
```

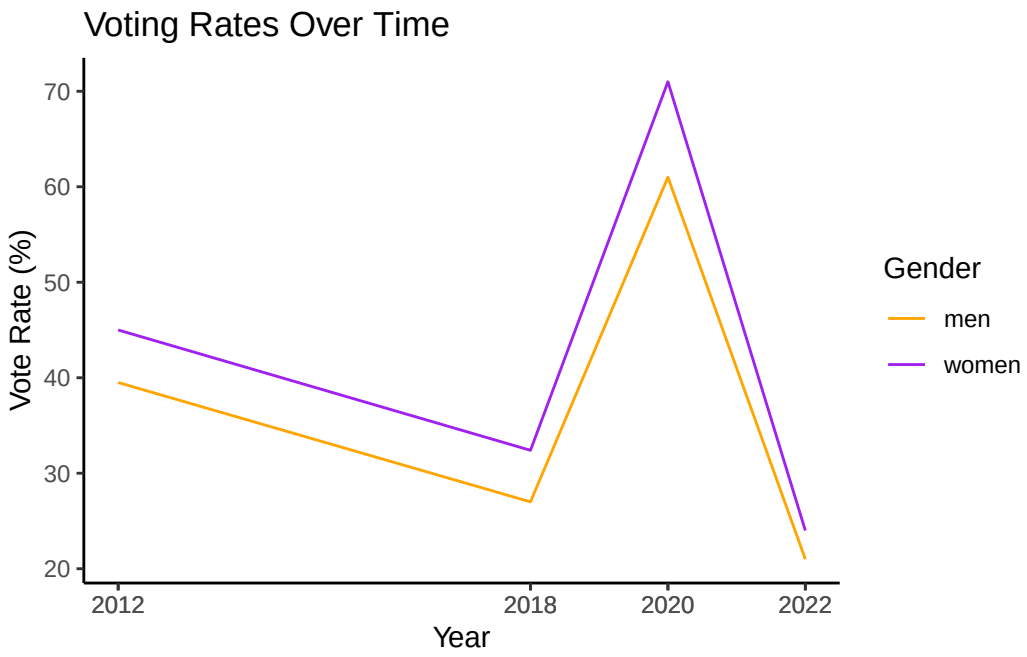


While some data was unavailable, making the graph incomplete, we can still note a few things. Notice that, from 2012-2016, freshmen, sophomores, and upperclassmen moving in nearly identical trends. Then, from 2016-2022, we see that changes in undergraduate student voting rates are more drastic from year-to-year than that of graduate students. Also, notice that from 2016-2018 undergraduate voting rates dropped, while that of graduate students increased. Why might that be?

The two previous graphs reflect two key things: 1) young voters are more impressionable (i.e. more susceptible to changes in Cornell Votes' curriculum) and 2) young voters are less involved in midterm elections. Since voting is habit-forming, it may be important to consider why the spike in 2020 was followed by a consistent plummet across all ages in 2022. This presents the question, how can we retain high voting rates from year to year (i.e. how can we make people care about midterm elections?)

```
combined_data_long <- combined_data %>%
  gather(key = "gender", value = "vote_rate", women, men) |> filter(!is.na(vote_rate))

ggplot(data = combined_data_long, aes(x = year, y = vote_rate, color = gender, group = gende
  geom_line() +
  labs(title = "Voting Rates Over Time",
```

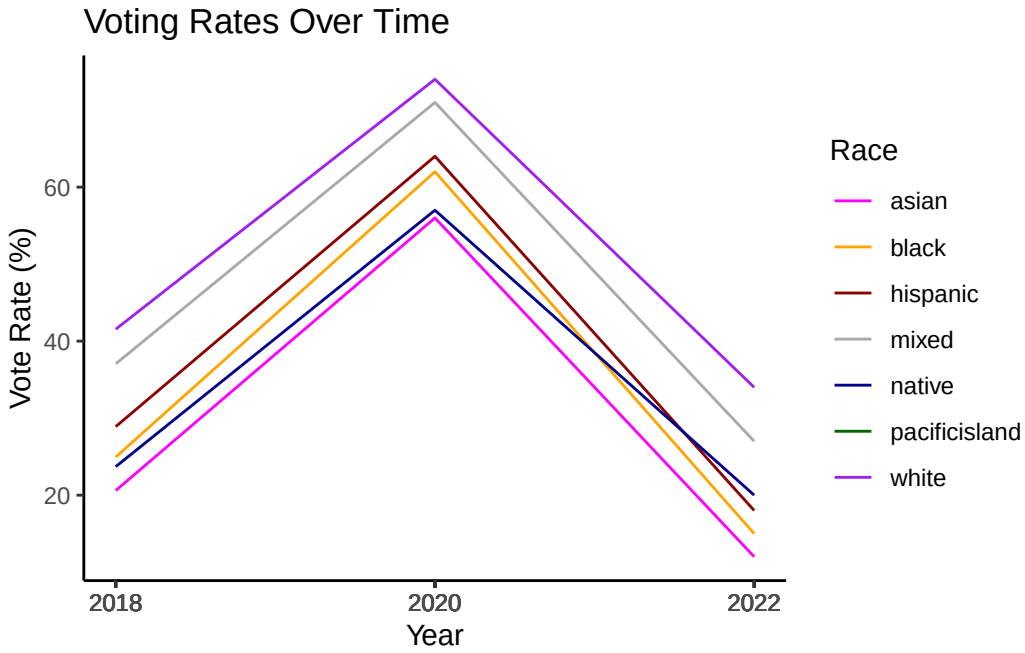
```
       x = "Year",
       y = "Vote Rate (%)",
       color = "Gender") +
  scale_color_manual(
    values = c("women" = "purple",
               "men" = "orange")
  ) +
  scale_x_continuous(
    breaks = combined_data_long$year
  ) + theme_classic()
```



While the trends are indentical, women consistently present a higher voting rate than men.

```
combined_data_long <- combined_data %>%
  gather(key = "race", value = "vote_rate", asian, black, native, hispanic, pacificisland, wh

ggplot(data = combined_data_long, aes(x = year, y = vote_rate, color = race, group = race)) +
  geom_line() +
  labs(title = "Voting Rates Over Time",
       x = "Year",
       y = "Vote Rate (%)",
       color = "Race") +
  scale_color_manual(
```

```
    values = c("asian" = "magenta",
               "black" = "orange",
               "hispanic" = "darkred",
               "native" = "darkblue",
               "pacificisland" = "darkgreen",
               "white" = "purple",
               "mixed" = "darkgrey")
  ) +
  scale_x_continuous(
    breaks = combined_data_long$year
  ) + theme_classic()
```



While the trends are nearly identical, white students show consistently higher voting rates than that of minority races.

```
fos_long <- data_fos |>
  gather(key = "fos", value = "vote_rate", Agriculture, Architecture, Area_Ethnic_Cultural_ar
  filter(!is.na(vote_rate))

ggplot(data = fos_long, aes(x = year, y = vote_rate, color = fos, group = fos)) +
  geom_line() +
  labs(title = "Voting Rates Over Time by Field of Study",
       x = "Year",
```

```
      y = "Vote Rate (%)",
      color = "Field of Study") +
scale_x_continuous(
  breaks = unique(fos_long$year)  # Adjust x-axis to only include unique years
) +
theme_classic() +
theme(
  strip.text = element_text(size = 12),  # Optional: adjust facet label size
  legend.position = "right",  # Position the legend to the right
  legend.key.size = unit(0.5, "cm"),  # Adjust size of legend keys
  legend.text = element_text(size = 8),  # Adjust size of legend text
  legend.title = element_text(size = 10)  # Adjust size of legend title
) +
guides(
  color = guide_legend(
    ncol = 2,  # Set the number of columns for the legend
    byrow = TRUE  # Arrange the legend items by row
  )
)
```