Is There Growing Polarization?
From Harvard's Dataverse

Chase Mathis

Packages and Data

Data from: Harvard Dataverse

```r
library(tidyverse)
library(tidymodels)
library(knitr)


data <- read_csv('data/countypres_2000-2020.csv') %>%
  mutate(percentage = candidatevotes/totalvotes,
         year = as.factor(year))
```

Exploratory Data Analysis

Data Manipulation

```r
variance <- data %>%
  group_by(county_name, year) %>%
  summarise(variance = sd(percentage)^2) %>%
  ungroup()


simple_difference <- data %>%
  group_by(year,county_name) %>%
  arrange(desc(percentage), .by_group = TRUE) %>%
  slice(1:2) %>% # take top two candidates by county
  summarise(difference = abs(percentage - lag(percentage))) %>%
  drop_na() %>%
  ungroup()#calculate difference between the two candidates
```

What do the Graph's Show?

```
variance %>%
  ggplot(aes(x = year, y = variance)) +
  geom_boxplot() +
  labs(
    x = 'Year',
    y = 'Variance',
    title = 'The Variance Between 1st and 2nd Place Presidential Candidate By Year'
  )
```
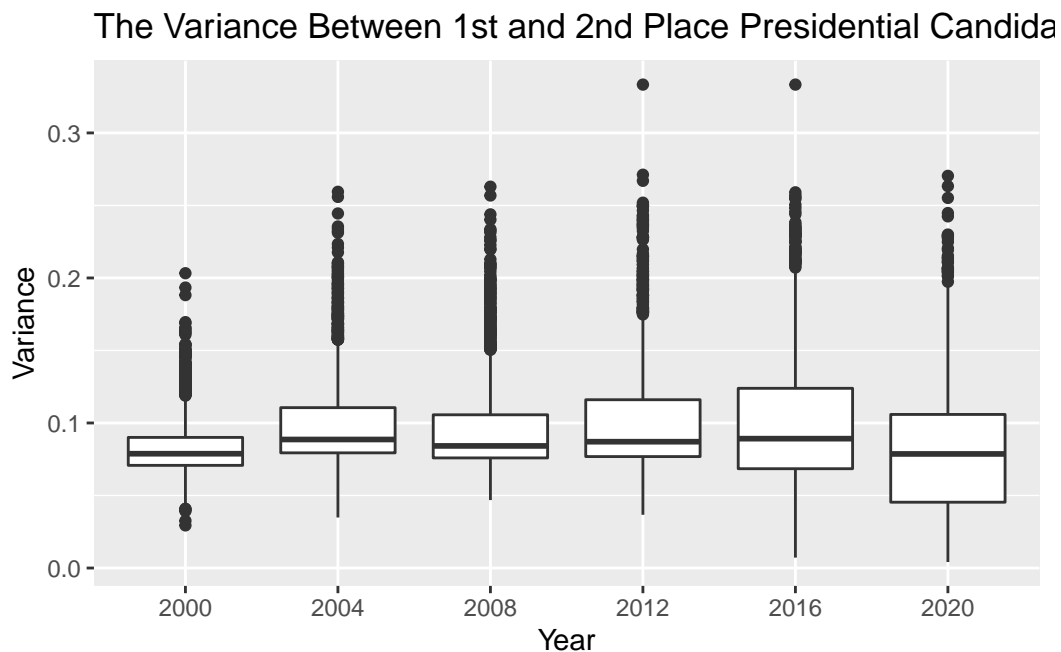

Variance Graph

Variance is a commonly used statistical tool to measure the spread of data. Therefore, I initially thought variance would be a good tool to see if counties are becoming more extreme when voting for president. I did not realize, however, that there are many observations in the dataframe where there a candidate is listed, but 0 people have voted for them. In fact, there are 11227 observations where candidates have received less than 0.5% of the county's vote.

Simple Difference Graph

```
simple_difference %>%
  ggplot(aes(x = year, y = difference)) +
  geom_boxplot() +
  labs(
    x = 'Year',
    y = 'Difference',
    title = 'The Difference Between 1st and 2nd Place Presidential Candidate By Year'
  )
```

The Difference Between 1st and 2nd Place Presidential Candi
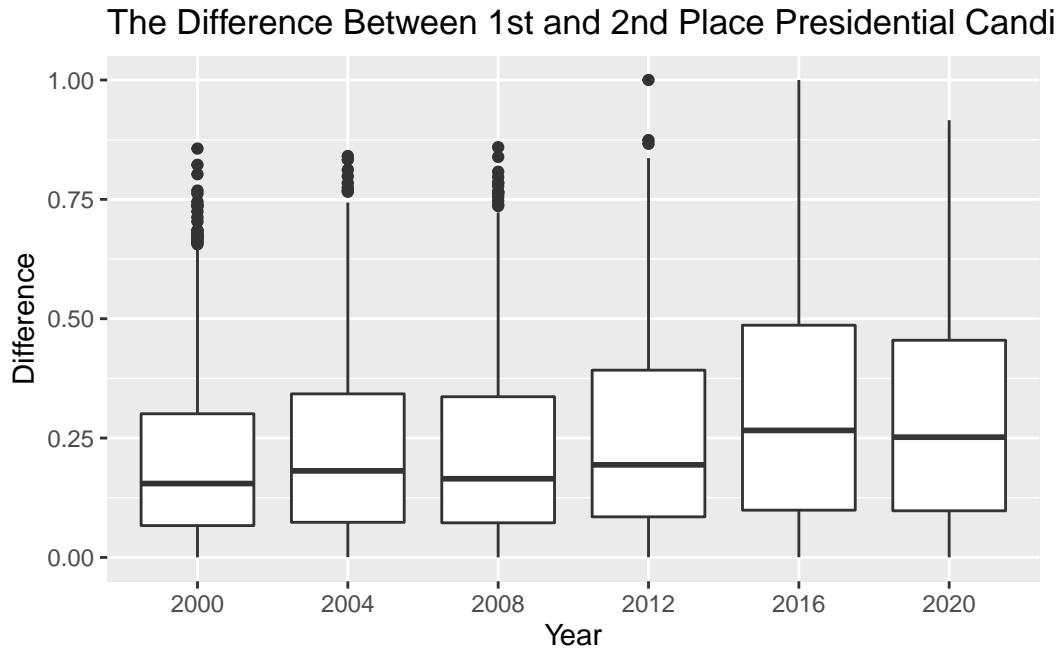
Therefore, I chose to eliminate all but the top two nominee's for each county, to eliminate outliers that can come from third party nominees. This graph shows a pretty clear trend that the average difference between the 1st and the 2nd place candidate by county has grown from 2000 to 2020. In 2000, the average difference between the 1st and 2nd place candidate was: 20.4 percentage points whereas in 2020, the average difference between the 1st and 2nd place candidate was 29.2 percentage points.

Conclusion

Although this quick analysis does not have enough evidence to support growing polarization across America, it is interesting to see at least one point of evidence that supports the fact that: red states are getting more red and blue states are getting more blue.