DATA 608 - Story 1

Kory Martin

2023-09-09

Contents

Overview:

For this assignment we were given a file containing data on the present allocation of the Infrastructure Investment and Jobs Act funding broken out by State and Territory. The goal of this assignment was to develop a story using data visualizations to address the following questions:

1. Is the allocation equitable based on the population of each of the States and Territories, or is bias apparent?
2. Does the allocation favor the political interests of the Biden administration?

In order to help in telling the story, I also included US Census data showing the state populations in 2020, as well as data showing a breakdown of votes for Biden and Trump in the 2020 US Presidential election.

Based on this data, my initial hypothesis is that evidence of bias in funding, would be reflected in one of several ways:

- Differences in how the total funding was apportioned between Biden won states and Trump won states
- Differences in the per capita funding levels between the two groups

The expectation is that if the funding was inequitable, it would be in favor of Biden won states.

Setup

We began by installing the necessary packages and librarry for use in importing and cleaning the data and for visualizing the tables.

```r
options(repos = list(CRAN="http://cran.rstudio.com/"))
knitr::opts_chunk$set(echo = TRUE, error=TRUE)

install.packages("tidyverse")
```

```
##
## The downloaded binary packages are in
##  /var/folders/4l/182nghl547v3mxtj6p_9dgph0000gn/T//RtmpgN8Sxn/downloaded_packages
```

```r
install.packages("rvest")
```

```
##
## The downloaded binary packages are in
##  /var/folders/4l/182nghl547v3mxtj6p_9dgph0000gn/T//RtmpgN8Sxn/downloaded_packages
```

```r
install.packages("kableExtra")
```

```
##
## The downloaded binary packages are in
##  /var/folders/4l/182nghl547v3mxtj6p_9dgph0000gn/T//RtmpgN8Sxn/downloaded_packages
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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 become errors
```

```r
library(rvest)
```

```
##
## Attaching package: 'rvest'
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```r
library(kableExtra)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Import Data

Here we are importing the data from the various sources.

```
funding_data <- readxl::read_excel('/Users/korymartin/Google Drive/My Drive/MS Program/Fall 2023/DATA 6


state_pop <- readxl::read_excel('/Users/korymartin/Google Drive/My Drive/MS Program/Fall 2023/DATA 608/
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
```

```
#Rename the columns for state population table

state_pop <- state_pop %>% rename(
  "state_territory" = "...1",
  "estimate_base_2020" = "...2",
  "estimate_2020" = "...3",
  "estimate_2021" = "...4",
  "estimate_2022" = "...5"
)


# 2020 Election Data
url <- 'https://www.presidency.ucsb.edu/statistics/elections/2020'

web_page <- read_html(url)
table_node <- html_node(web_page, "table")

results_table <- html_table(table_node)
results_filtered <- results_table %>% slice(10:n()) %>% select(1:11)

#Remove Electoral College Votes
results_filtered <- results_filtered %>% select(-c(X4,X5,X7,X8,X10,X11))

results_filtered <- results_filtered %>% rename(
  "state_territory" = "X1",
  "total_votes" = "X2",
  "biden_votes" = "X3",
  "trump_votes" = "X6",
  "other_votes" = "X9"
```

```
)

#Remove unnecessary rows of data
election_results <- results_filtered %>% slice(4:61)
```

Create Combined Dataset

Here we are joining the data from the respective datasets to create a single dataset.

```
## Update column names for funding data table
colnames(funding_data) <- c("state_territory", "total_funding_billions")

## Change case of state_territory data
funding_data <- funding_data %>% mutate(
  state_territory = str_to_title(state_territory)
)

funding_data <- funding_data %>% mutate(
  state_territory = ifelse(state_territory == "Deleware","Delaware",state_territory),
  state_territory = ifelse(state_territory == "District Of Columbia","District of Columbia",state_terri



election_results
```

```
## # A tibble: 58 x 5
##    state_territory        total_votes  biden_votes  trump_votes other_votes
##    <chr>                  <chr>        <chr>        <chr>       <chr>
##  1 ""                     ""           ""           ""          ""
##  2 "Alabama"              "2,323,282"  "849,624"    "1,441,170" "32,488"
##  3 "Alaska"               "359,530"    "153,778"    "189,951"   "15,801"
##  4 "Arizona"              "3,387,326"  "1,672,143"  "1,661,686" "53,497"
##  5 "Arkansas"             "1,219,069"  "423,932"    "760,647"   "34,490"
##  6 "California"           "17,500,881" "11,110,250" "6,006,429" "384,202"
##  7 "Colorado"             "3,256,952"  "1,804,352"  "1,364,607" "87,993"
##  8 "Connecticut"          "1,824,280"  "1,080,680"  "715,291"   "28,309"
##  9 "Delaware"             "504,010"    "296,268"    "200,603"   "7,139"
## 10 "District of Columbia" "344,356"    "317,323"    "18,586"    "8,447"
## # i 48 more rows
```

```
state_pop <- state_pop %>%
  mutate(state_territory = str_remove(state_territory,"."))


## Create Merged Table
m1 <- left_join(funding_data, election_results, by="state_territory")
m2 <- left_join(m1, state_pop, by="state_territory")

## Remove NA values (US Territories without voting data) from dataset
m2 <- m2 %>% drop_na()
```

```
## Remove comma from voter counts and convert to numeric
m2 <- m2 %>%
  mutate_at(c("total_votes","biden_votes", "trump_votes", "other_votes"),~(str_replace_all(., ",", "")))
  mutate_at(c("total_votes","biden_votes", "trump_votes", "other_votes"),~as.numeric(.))
```

Clean Combined Data

Finally, we clean the resuling datasets and create additional summary tables.

```
combined_data <- m2 %>% mutate(
  "biden_won" = ifelse(biden_votes > trump_votes, "yes","no")
)

combined_data
```

```
## # A tibble: 51 x 11
##    state_territory    total_funding_billions total_votes biden_votes trump_votes
##    <chr>                               <dbl>       <dbl>       <dbl>       <dbl>
##  1 Alabama                             3         2323282      849624     1441170
##  2 Alaska                              3.7        359530      153778      189951
##  3 Arizona                             3.5       3387326     1672143     1661686
##  4 Arkansas                            2.8       1219069      423932      760647
##  5 California                         18.4      17500881    11110250     6006429
##  6 Colorado                            3.2       3256952     1804352     1364607
##  7 Connecticut                         2.5       1824280     1080680      715291
##  8 Delaware                            0.792      504010      296268      200603
##  9 District of Colum~                  1.1        344356      317323       18586
## 10 Florida                             8.2      11067456     5297045     5668731
## # i 41 more rows
## # i 6 more variables: other_votes <dbl>, estimate_base_2020 <dbl>,
## #   estimate_2020 <dbl>, estimate_2021 <dbl>, estimate_2022 <dbl>,
## #   biden_won <chr>
```

```
combined_data <- combined_data %>%
  mutate(total_funding = total_funding_billions * 1e9,
  per_cap_spending = total_funding/estimate_base_2020,
  biden_pct = (biden_votes/total_votes))

colnames(combined_data)
```

```
##  [1] "state_territory"        "total_funding_billions" "total_votes"
##  [4] "biden_votes"            "trump_votes"            "other_votes"
##  [7] "estimate_base_2020"     "estimate_2020"          "estimate_2021"
## [10] "estimate_2022"          "biden_won"              "total_funding"
## [13] "per_cap_spending"       "biden_pct"
```

```
combined_simplified <- combined_data %>% select(
  c("state_territory", "per_cap_spending", "biden_won", "total_funding_billions","biden_pct",
    "estimate_base_2020","total_funding")
)
```

Table 1: Summary of Spending Levels for Biden Won and Lost Stats

Biden Won	Total Funding	Population	Per Capita	Avg Per Capita
no	88.5	141,066,781	627.36	1,155.29
yes	103.1	190,382,739	541.52	674.30

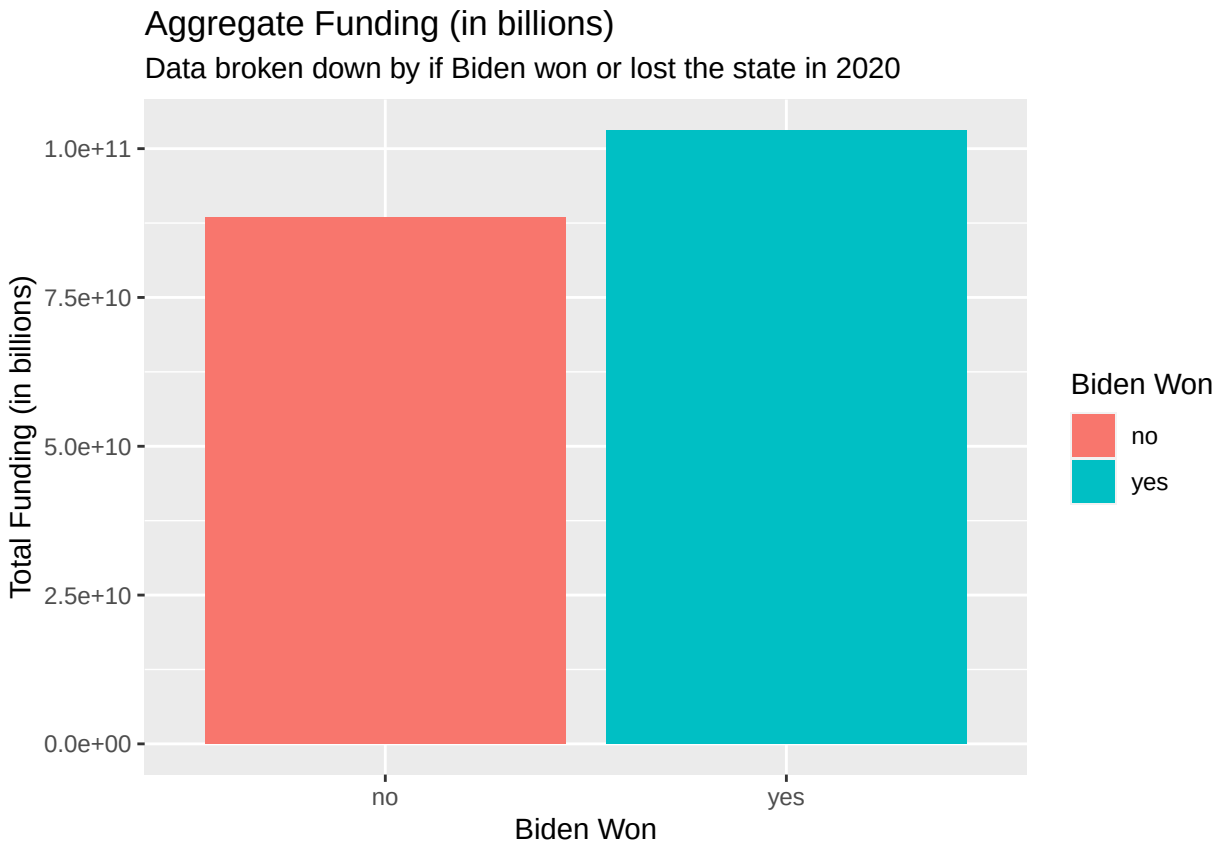
```
summary_df <- combined_simplified %>%
  group_by(biden_won) %>%
  summarize(total_funding_billions = sum(total_funding_billions),
            total_funding = sum(total_funding),
            total_pop = sum(estimate_base_2020),
            avg_per_cap = mean(per_cap_spending),
            med_per_cap = median(per_cap_spending))

summary_df <- summary_df %>%
  mutate(per_cap = total_funding/total_pop)
```

Data Visualizations

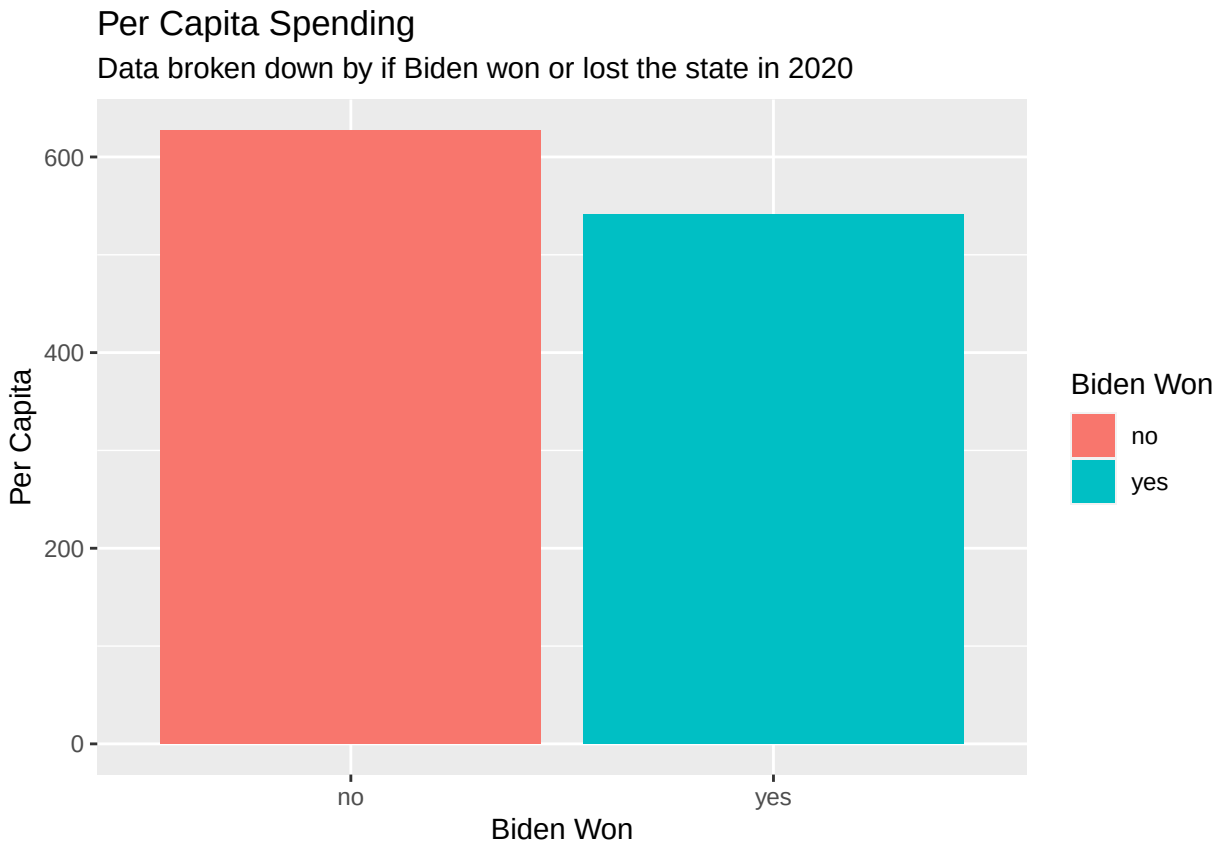
The following datasets were chosen to help to illustrate the relative differences in funding allocation based on whether or not Biden carried the state in the 2020 US Presidential Election.

```
summary_df %>%
  select(c(1,2,4,7,5)) %>%
  kbl(
    col.names = c("Biden Won", "Total Funding", "Population", "Per Capita", "Avg Per Capita"),
    align = "c",
    digits = 2,
    format.args = list(big.mark = ",", scientific=FALSE),
    caption = "Summary of Spending Levels for Biden Won and Lost Stats"
  ) %>%
  kable_material(c("striped"))
```

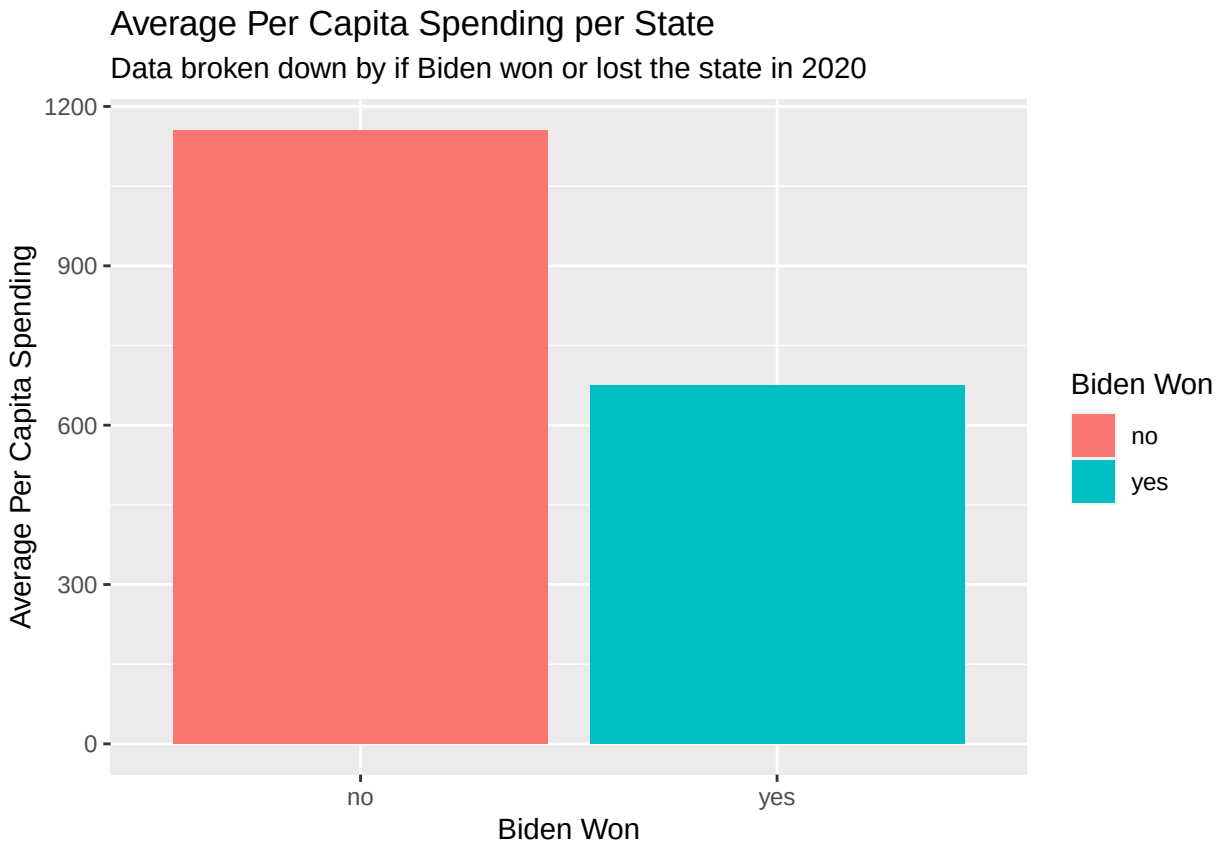
```
summary_df %>%
  ggplot(aes(x=biden_won, y=total_funding, fill=biden_won)) +
  geom_bar(stat="identity") +
  labs(title = "Aggregate Funding (in billions)",
      subtitle = "Data broken down by if Biden won or lost the state in 2020",
      x = "Biden Won",
      y = "Total Funding (in billions)",
      fill = "Biden Won")
```



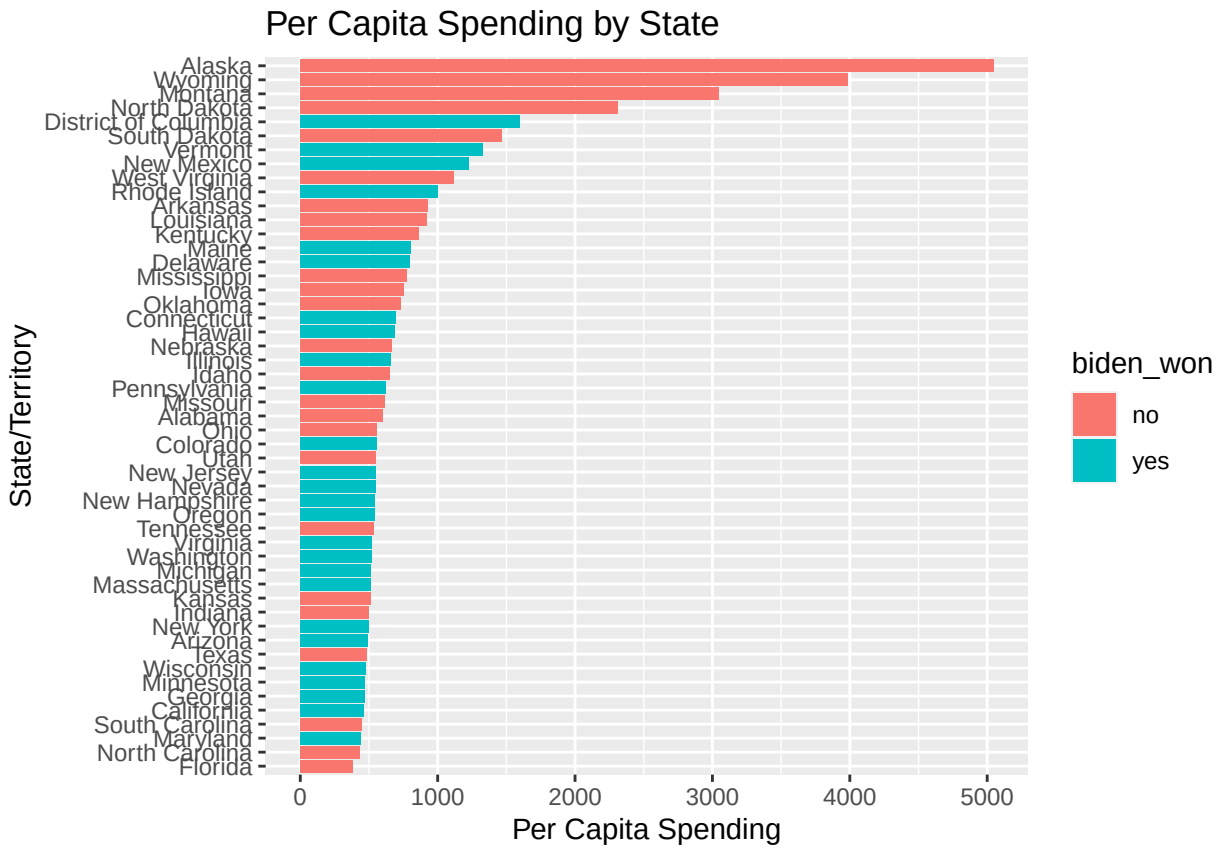

```
summary_df %>%
  ggplot(aes(x=biden_won, y=per_cap, fill=biden_won)) +
  geom_bar(stat="identity") +
  labs(title = "Per Capita Spending",
       subtitle = "Data broken down by if Biden won or lost the state in 2020",
       x = "Biden Won",
       y = "Per Capita",
       fill = "Biden Won")
```



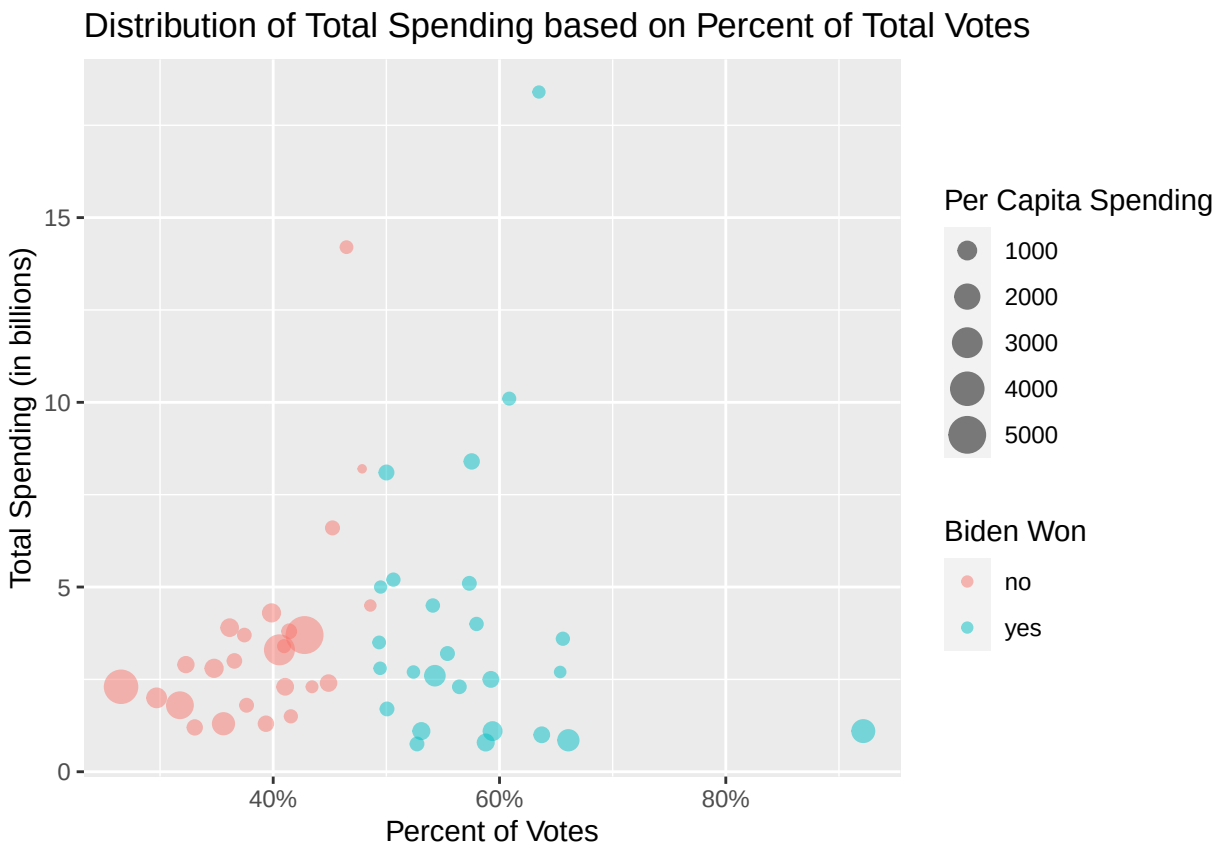

```
summary_df %>%
  ggplot(aes(x=biden_won, y=avg_per_cap, fill=biden_won)) +
  geom_bar(stat="identity") +
  labs(title = "Average Per Capita Spending per State ",
      subtitle = "Data broken down by if Biden won or lost the state in 2020",
      x = "Biden Won",
      y = "Average Per Capita Spending",
      fill = "Biden Won")
```




```
combined_simplified %>%
  ggplot(aes(x=reorder(state_territory, per_cap_spending), y=per_cap_spending, fill=biden_won)) +
  geom_col() +
  coord_flip() +
  labs(title = "Per Capita Spending by State",
       y="Per Capita Spending",
       x = "State/Territory")
```




```
combined_simplified %>%
  ggplot(aes(x=biden_pct, y=total_funding_billions)) +
  geom_point(aes(size=per_cap_spending, color=biden_won), alpha=.5) +
  labs(title = "Distribution of Total Spending based on Percent of Total Votes",
       x = "Percent of Votes",
       y = "Total Spending (in billions)",
       color = "Biden Won",
       size = "Per Capita Spending") +
  scale_x_continuous(labels = scales::percent_format(accuracy = 1))
```



Conclusion

In looking at the data, it appears that overall spending levels were higher in states that Biden did not carry in the 2020 election based on both aggregate per-capita spending as well as average per-capita spending per state. Based on the data, we find that the average per-capita spending was X% higher for Non-Biden states compared to Biden states, while the average per-state per-capita spend was X% higher in Non-Biden states compared to Biden states.

While my initial thinking was that if there was political bias that it would be in favor of those states that Biden won; it's possible that the political bias may be more strategic and that the administration would be motivated to invest in places where there's a benefit for future elections. Thus the rationale for the apportionment of funds could have actually been based on providing more spending to those states that could potentially be in play in future election cycles in order to curry favor.