Data 608 Assignment 1

Alex Khaykin

2023-09-10

```
library(stringr)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(tidyr)
```

INTODUCTION

The Infrastructure Investment and Jobs Act (IIJA) allocated approximately $350 billion for Federal highway programs over a five-year period (fiscal years 2022 through 2026). US president Joe Biden signed into law in 2021 and its stated purpose is largely for improvement of highway infrastructure.

Data on IIJA allocation for 2023 for US states and territories including tribal communities was used as well as 2020 census data to ask whether funding was allocated equitably according population size. Popular votes for president Biden from the 2020 election were also used to assess whether there is a bias with higher allocations going to the president's supporters.

Data Sources

The 2020 presidential election data was sourced from the UCSB Presidency project.

All of the population data were sourced from the 2020 US Census.

Note

- I chose to use the 2020 actual census data rather than a 2022 estimated population, because we are using 2020 presidential election data. Further, if state population influenced funding, it is more likely that actual census data would be used.
- Population size could not be attained for tribal communities directly. I used US census counts for American Indian-Alaska Native (AIAN) persons as a proxy for this. I used the counts for all persons identifying as AIAN even though it is possible that not all of these individuals are services by tribal communities.

Is the allocation equitably based on the population of each of the States and Territories, or is bias apparent?

Reading in the IIJA data.

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.2.3
```

```
iija <- read_excel("C:\\Users\\akhay\\OneDrive\\Documents\\DATA_SCIENCE\\DATA_608\\Major Assignments\\As
#rename the columns for merging
names(iija) <- c("Geographic_Area", "Funding_in_Billions")
str(iija)
```

```
## tibble [57 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Geographic_Area    : chr [1:57] "ALABAMA" "ALASKA" "AMERICAN SAMOA" "ARIZONA" ...
##  $ Funding_in_Billions: num [1:57] 3 3.7 0.0686 3.5 2.8 18.4 3.2 2.5 0.792 1.1 ...
```

Reading in the population data.

```
pop <- read_excel("C:\\Users\\akhay\\OneDrive\\Documents\\DATA_SCIENCE\\DATA_608\\Major Assignments\\Ass
str(pop)
```

```
## tibble [59 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Geographic_Area: chr [1:59] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ Census_2020    : num [1:59] 5024356 733378 7151507 3011555 39538245 ...
```

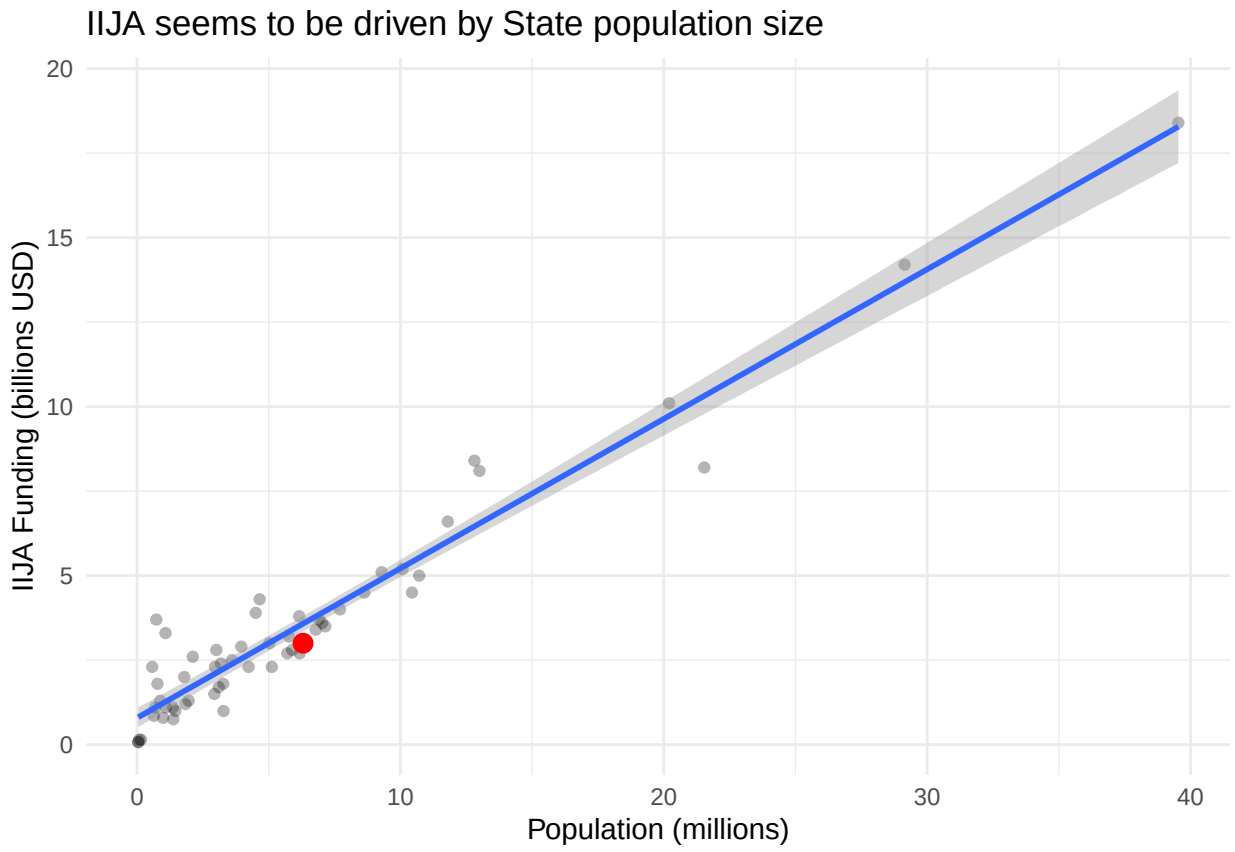
To merge the iija funding data with the population data.

```
#To make the cases match in Geographic Area
iija$Geographic_Area <- toupper(iija$Geographic_Area)
pop$Geographic_Area <- toupper(pop$Geographic_Area)
#To perform inner join of data sets by geographic area
df <- inner_join(iija, pop, by="Geographic_Area")
```

Scatter Plot of Population and iija Funding.

```
#To divide population by 1 million to make axes more equitable
df <- df %>% mutate(pop_in_millions=Census_2020/10^6)
#To make a df to highlight outliers/points of interest
highlight <- df %>% filter(df$Geographic_Area == "TRIBAL COMMUNITIES")
#To make a scatter plot
df %>% ggplot(aes(x = pop_in_millions, y = Funding_in_Billions)) +
  geom_point(alpha=0.3) +
  geom_smooth(method = "lm") +
  labs(x = "Population (millions)", y = "IIJA Funding (billions USD)", title = "IIJA seems to be driven
  theme_minimal() +
  geom_point(data = highlight, aes(x = pop_in_millions, y = Funding_in_Billions), size=3, color="red")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



CONCLUSION There seems to be a fairly strong positive correlation between a State/Territory's population and level of funding allocated. One concern in choosing the population proxy for Tribal communities was that it would be a clear outlier because it grossly either over or under estimates the funding allocation to those peoples. I have highlighted that data point in red, although it is still a proxy it does not fall out as a strong outlier, and I will keep in the dataset.It is worth noting that for states with smaller populations the relationships seems less clear, and there may be other factors to help explain IIJA funding allocation.

Does the allocation favor the political interests of the Biden administration?

Reading in the election data.

```
elec <- read_excel("C:\\Users\\akhay\\OneDrive\\Documents\\DATA_SCIENCE\\DATA_608\\Major Assignments\\As
#To change the first column name to match other datasets.
names(elec)[1] <- "Geographic_Area"
str(elec)
```

```
## tibble [51 x 11] (S3: tbl_df/tbl/data.frame)
##  $ Geographic_Area      : chr [1:51] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ TOTAL_VOTES          : num [1:51] 2323282 359530 3387326 1219069 17500881 ...
##  $ BIDEN_VOTES          : num [1:51] 849624 153778 1672143 423932 11110250 ...
##  $ BIDEN_PROP           : num [1:51] 0.37 0.43 0.49 0.35 0.63 0.55 0.59 0.59 0.92 0.48 ...
##  $ BIDEN_ELECTORAL_VOTES: num [1:51] NA NA 11 NA 55 9 7 3 3 NA ...
##  $ TRUMP_VOTES          : num [1:51] 1441170 189951 1661686 760647 6006429 ...
##  $ TRUMP_PROP           : num [1:51] 0.62 0.53 0.49 0.62 0.34 0.42 0.39 0.4 0.05 0.51 ...
```

```
##  $ TRUMP_ELECTORAL_VOTES: num [1:51] 9 3 NA 6 NA NA NA NA NA 29 ...
##  $ OTHER_VOTES          : num [1:51] 32488 15801 53497 34490 384202 ...
##  $ OTHER_PROP           : num [1:51] 0.01 0.04 0.02 0.03 0.02 0.03 0.02 0.01 0.02 0.01 ...
##  $ OTHER_ELECTORAL_VOTES: logi [1:51] NA NA NA NA NA NA ...
```

To merge the iija and pop data with the 2020 presidential election data.

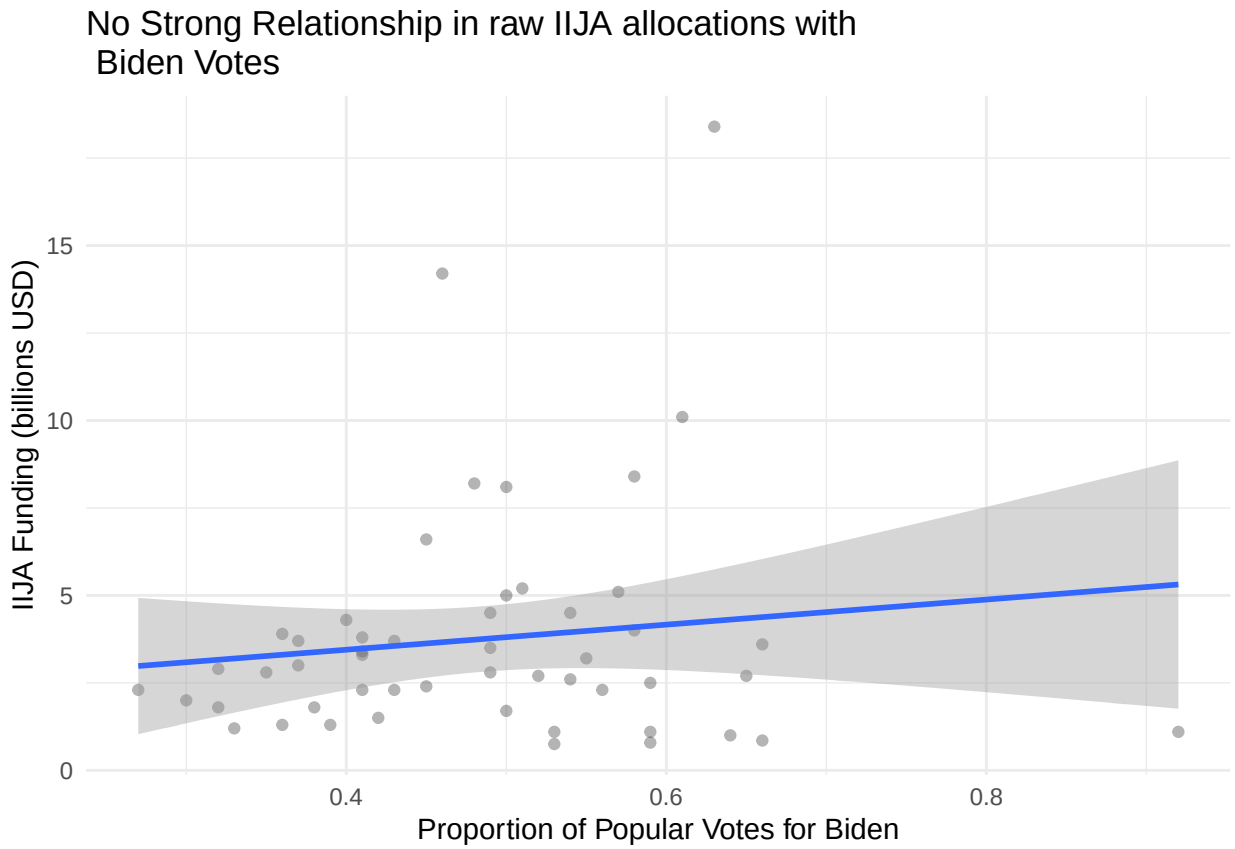
```
#To make the cases match in Geographic Area
elec$Geographic_Area <- toupper(elec$Geographic_Area)
#To perform inner join of data sets by geographic area
df <- inner_join(df, elec, by="Geographic_Area")
```

Note Because US Territories do not have presidential voting power they are excluded from this list. Further, because tribal communities vote as members of their state rather than sovereign entities, that category has also been dropped from the dataset.

Scatter Plot of Presidential elections and iija Funding.

```
#To make a scatter plot
df %>% ggplot(aes(x = BIDEN_PROP, y = Funding_in_Billions)) +
  geom_point(alpha=0.3) +
  geom_smooth(method = "lm") +
  labs(x = "Proportion of Popular Votes for Biden", y = "IIJA Funding (billions USD)", title = "No Stro
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



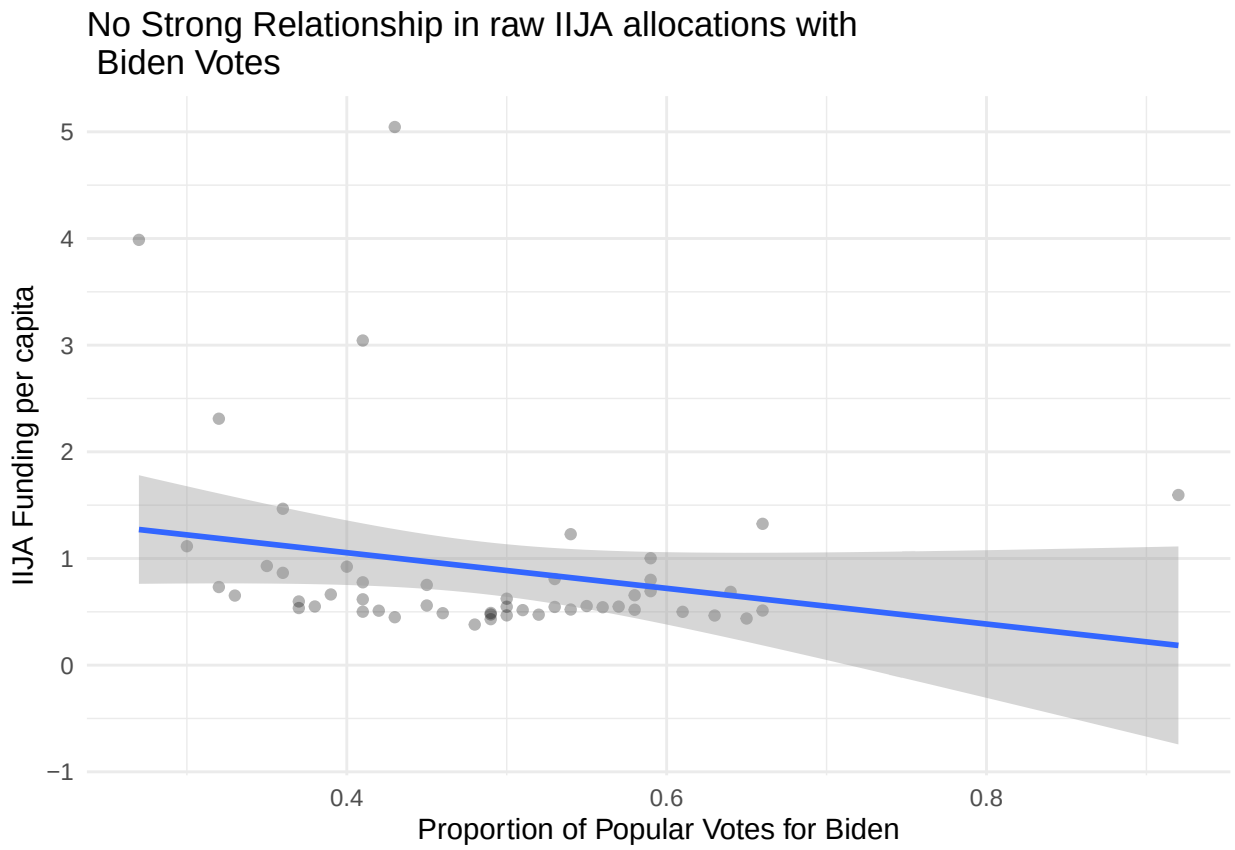
CONCLUSION Although the trend line is slightly positive, there is a strong scatter, especially in the middle of the plot which suggest clear outliers could be influential data points. A more accurate assessment would be to account for population size by caclulating IIJA funding per capita given that a strong positve relationship was althready shown.

Do Presidential votes explain per capita IIJA funding?

Scatter Plot of Presidential elections and per capita IIJA Funding.

```
#To calculate IIJA funding per capita.
df <- df %>% mutate(iija_per_cap = Funding_in_Billions/pop_in_millions)
#To make a scatter plot
df %>% ggplot(aes(x = BIDEN_PROP, y = iija_per_cap)) +
  geom_point(alpha=0.3) +
  geom_smooth(method = "lm") +
  labs(x = "Proportion of Popular Votes for Biden", y = "IIJA Funding per capita", title = "No Strong R
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



CONCLUSION A very weak negative relationship is seen between the proportion of Biden popular votes and IIJA funding per capita.There does not seem to be sufficient support for a relationship between popular votes for the President in 2020 and allocation of IIJA funds.

OVERALL CONCLUSION

Based on IIJA allocation for 2023 for US states and territories including tribal communities, I found a reasonably strong positive correlation with increasing population size per the 2020 US census. This suggests IIJA funding has been allocated equitably according population size. Even though population size could not be attained for tribal communities directly US census counts for AIAN person did not create an extreme outlier.

To test whether there is a bias with higher IIJA allocations going to the president Biden's supporters, I first plotted the proportion of Biden's popular votes from the 2020 election against IIJA dollars. I saw a weak positive relationship that was not convincing due high error around the best fit line, I also plotted IIJA dollars per capita against the proportion of Biden's popular votes from the 2020 election as saw a very weak negative relationship. This seems to suggest politics did not play a significant roll in determining IIJA allocation, at least as measured by poplar votes for president.

Future directions for this project might include looking other sources of political bias that could be tested against IIJA funding allocations. Further, for states receiving low levels of IIJA funding population size was not as strong a predictor as for larger population states.Thus future work could also examine other variable to help explain the allocations.