Data608 - Story 1:Infrastructure Investment & Jobs Act Funding Allocation

Hazal Gunduz

2024-09-12

Instructions

The attached Excel file contains data on the present allocation of the Infrastructure Investment and Jobs Act funding by State and Territory.

Your story (Data Visualization(s)) should address the following questions:

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

Notes:

1.You will need to source data on the current (estimated) population of each of the States and Territories (accuracy is more important than precision) and on the official election results of the 2020 Presidential election.

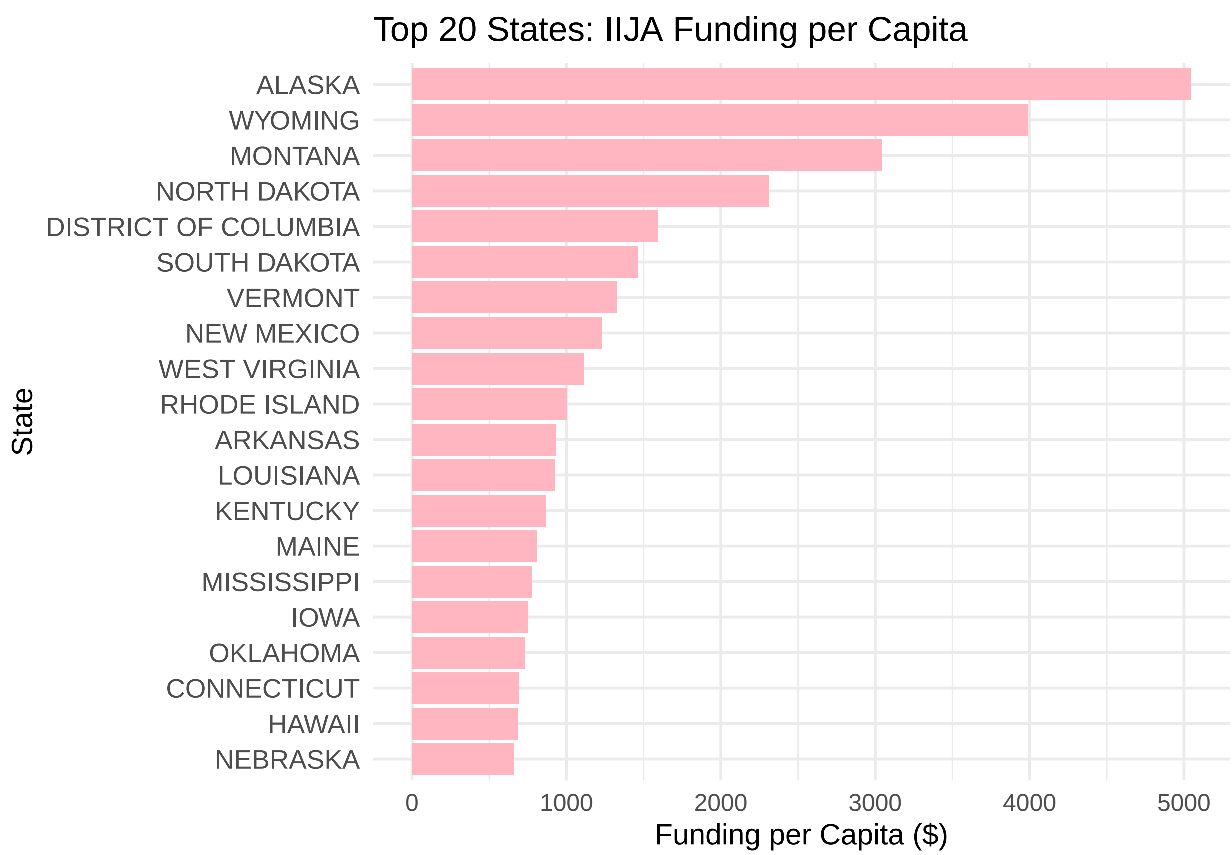
2.You may choose to develop you visualizations using a desktop application or a code library. Your submittal should be in the form of a report (document) or a presentation.

We'll need to combine information from two files: "IIJA FUNDING AS OF MARCH 2023.xlsx" for the funding data and "PopulationReport.xlsx" for the population data. Then, we'll calculate the funding per capita and create a bar plot. Let's start by loading and processing the data.

```r
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(scales)

# Read IIJA funding data
iija_funding <- read_excel("IIJA FUNDING AS OF MARCH 2023.xlsx", sheet = "Sheet1")
colnames(iija_funding) <- c("State", "Total_Billions")

# Read the new population data
population_report <- read_excel("PopulationReport.xlsx", sheet = "PopulationReport")

# Clean population data
population_report <- population_report %>%
  select(Name, `Pop. 2020`) %>%
  rename(State = Name, Population = `Pop. 2020`)

# Clean state names
iija_funding$State <- toupper(trimws(iija_funding$State))
population_report$State <- toupper(trimws(population_report$State))

# Merge the datasets
merged_data <- inner_join(iija_funding, population_report, by = "State")

# Calculate funding per capita
merged_data$Funding_Per_Capita <- (merged_data$Total_Billions * 1e9) / merged_data$Population

# Sort by funding per capita
merged_data <- merged_data[order(-merged_data$Funding_Per_Capita), ]

# Select top 20 states for better visualization
top_20 <- head(merged_data, 20)

# Create the bar plot
plot <- ggplot(top_20, aes(x = reorder(State, Funding_Per_Capita), y = Funding_Per_Capita)) +
  geom_bar(stat = "identity", fill = "lightpink") +
  coord_flip() +
  labs(title = "Top 20 States: IIJA Funding per Capita",
       x = "State",
       y = "Funding per Capita ($)") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 10))

# Display the plot
print(plot)
```



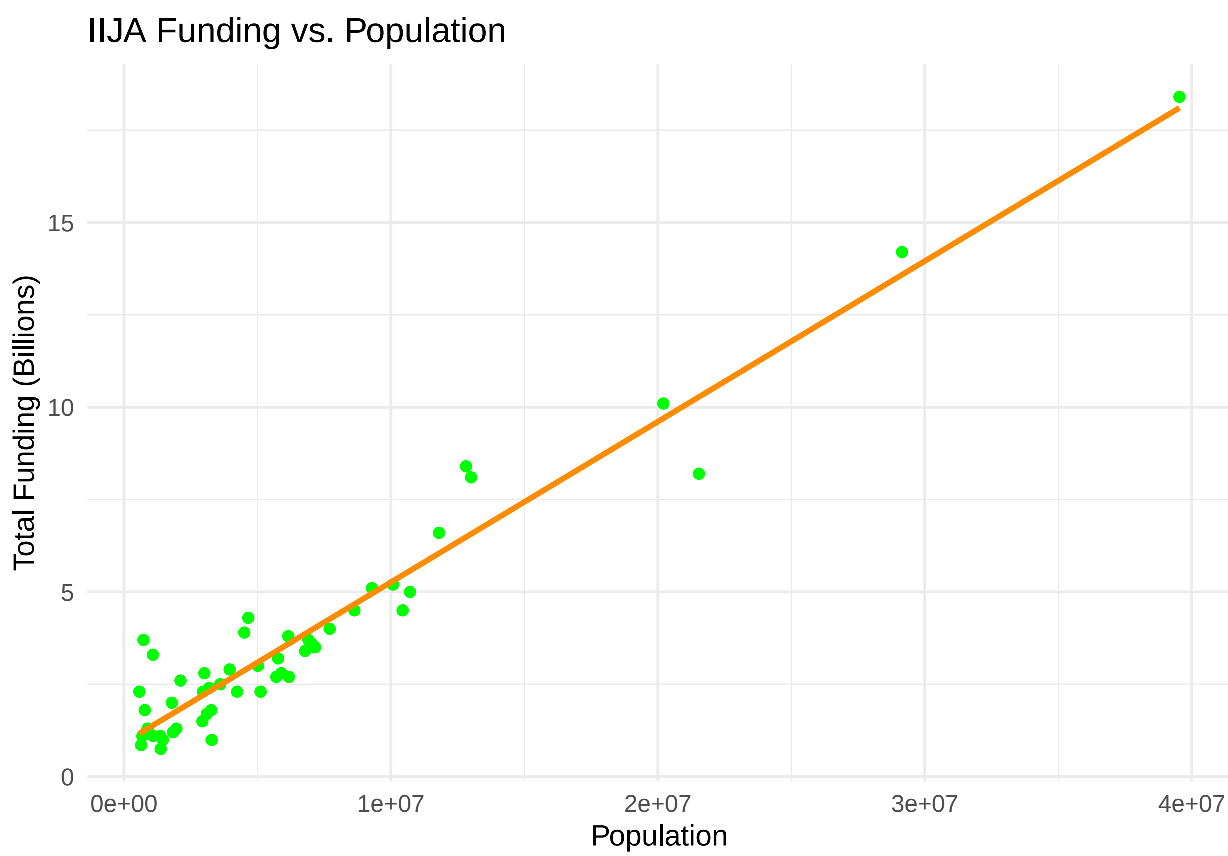

```r
# Show the top 20 data
print(head(top_20[, c("State", "Total_Billions", "Population", "Funding_Per_Capita")], 20))
```

```
## # A tibble: 20 x 4
##    State                Total_Billions Population Funding_Per_Capita
##    <chr>                         <dbl>      <dbl>              <dbl>
##  1 ALASKA                        3.7       733374              5045.
##  2 WYOMING                       2.3       576850              3987.
##  3 MONTANA                       3.3      1084244              3044.
##  4 NORTH DAKOTA                  1.8       779079              2310.
##  5 DISTRICT OF COLUMBIA          1.1       689548              1595.
##  6 SOUTH DAKOTA                  1.3       886668              1466.
##  7 VERMONT                       0.852     643077              1325.
##  8 NEW MEXICO                    2.6      2117525              1228.
##  9 WEST VIRGINIA                 2        1793713              1115.
## 10 RHODE ISLAND                  1.1      1097371              1002.
## 11 ARKANSAS                      2.8      3011490               930.
## 12 LOUISIANA                     4.3      4657785               923.
## 13 KENTUCKY                      3.9      4506297               865.
## 14 MAINE                         1.1      1363177               807.
## 15 MISSISSIPPI                   2.3      2961306               777.
## 16 IOWA                          2.4      3190427               752.
## 17 OKLAHOMA                      2.9      3959411               732.
## 18 CONNECTICUT                   2.5      3605912               693.
## 19 HAWAII                        1        1455274               687.
## 20 NEBRASKA                      1.3      1961965               663.
```

I'll create a scatter plot to visualize the relationship between total IIJA funding and population for each state. This will help us observe if there's any trend or pattern in how funding is distributed by population size.

```
# Scatter plot of total funding vs. population
scatter_plot <- ggplot(merged_data, aes(x = Population, y = Total_Billions)) +
  geom_point(color = "green") +
  geom_smooth(method = "lm", color = "darkorange", se = FALSE) +
  labs(title = "IIJA Funding vs. Population",
       x = "Population",
       y = "Total Funding (Billions)") +
  theme_minimal()

# Display the scatter plot
print(scatter_plot)
```

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



This scatter plot shows the relationship between a state's population and the amount of IIJA funding it receives. As the population increases, the amount of funding generally increases as well. This visualization helps us understand that although population is a strong determinant of IIJA funding allocation, other factors also play a role. States that deviate significantly from the trend line may have unique infrastructure needs, geographic challenges, or other factors affecting funding allocation.

- Is the allocation equitable based on the population of each of the States and Territories, or is bias apparent? IIJA funding allocation is equitable based on the population of each state and territory, we should consider the trend in the scatter plot and the variation around that trend. Here's a more detailed examination: Correlation Coefficient, Residual Analysis, Funding/Population Ratio..

Let's calculate the correlation coefficient to see if there is a strong linear relationship. And also examine residuals from the linear regression to identify any patterns or systematic deviations indicating bias.

```r
library(ggplot2)
library(dplyr)
library(broom)
```

Here are the insights and analyses based on the IIJA funding and population data:

Correlation & Regression Analysis Correlation Coefficient: The correlation coefficient between population and total funding is very high at 0.9667, indicating a strong positive linear relationship.

Linear Regression: The regression analysis shows a statistically significant relationship between population and funding, with both the intercept and the slope coefficients being highly significant ($p < 0.001$).

```r
library(ggplot2)
library(dplyr)
library(broom)

# Calculate correlation coefficient
correlation <- cor(merged_data$Population, merged_data$Total_Billions)

# Perform linear regression
lm_model <- lm(Total_Billions ~ Population, data = merged_data)

# Add predicted values and residuals to the data
merged_data <- merged_data %>%
  mutate(Predicted = predict(lm_model),
         Residuals = residuals(lm_model),
         Funding_Per_Capita = (Total_Billions * 1e9) / Population)

# Calculate summary statistics
summary_stats <- merged_data %>%
  summarise(
    Mean_Funding_Per_Capita = mean(Funding_Per_Capita),
    Median_Funding_Per_Capita = median(Funding_Per_Capita),
    SD_Funding_Per_Capita = sd(Funding_Per_Capita))

# Print results
print(paste("Correlation coefficient:", round(correlation, 4)))
```

```
## [1] "Correlation coefficient: 0.9667"
```

```r
print(summary(lm_model))
```

```
##
## Call:
## lm(formula = Total_Billions ~ Population, data = merged_data)
##
```

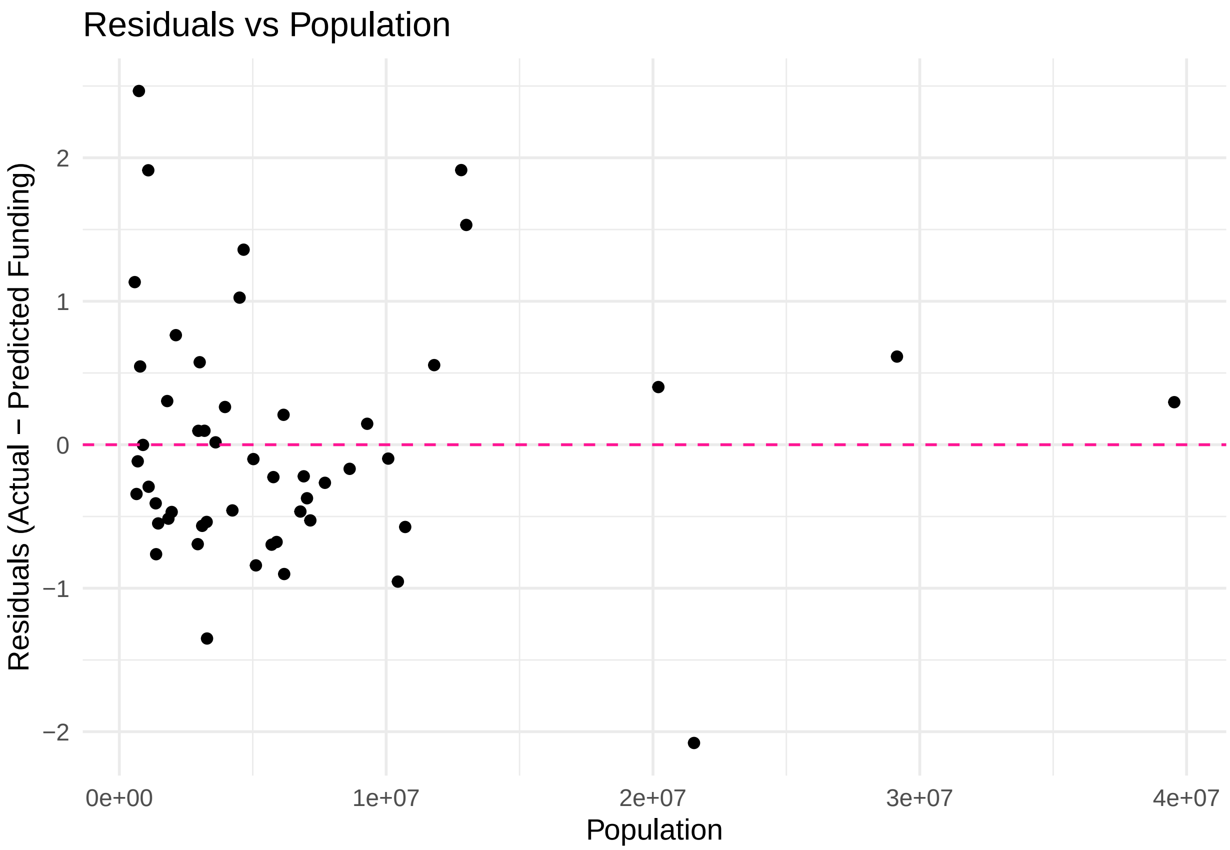
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0787 -0.5328 -0.1680  0.3532  2.4653
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.159e-01  1.613e-01   5.679 7.28e-07 ***
## Population  4.347e-07  1.644e-08  26.442  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.858 on 49 degrees of freedom
## Multiple R-squared:  0.9345, Adjusted R-squared:  0.9332
## F-statistic: 699.2 on 1 and 49 DF,  p-value: < 2.2e-16
```

```r
print(summary_stats)
```

```
## # A tibble: 1 x 3
##   Mean_Funding_Per_Capita Median_Funding_Per_Capita SD_Funding_Per_Capita
##                     <dbl>                     <dbl>                 <dbl>
## 1                    900.                      559.                  888.
```

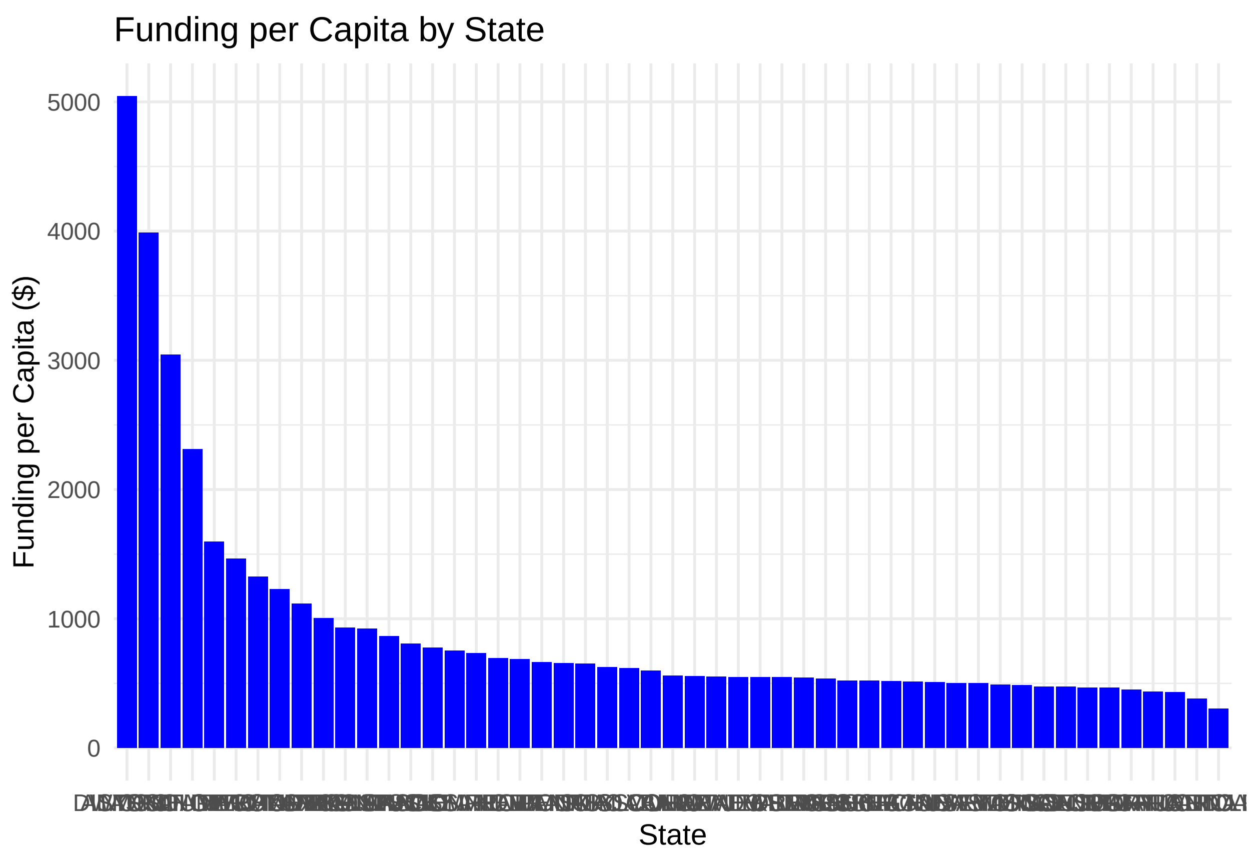
Residuals Analysis The residual plot does'nt show a systematic pattern, suggesting that there is no apparent bias in the allocation based purely on population. The residuals are spread randomly around the horizontal axis (zero line).

```r
library(ggplot2)
library(dplyr)
library(broom)

# Plot residuals
residual_plot <- ggplot(merged_data, aes(x = Population, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "deeppink") +
  labs(title = "Residuals vs Population",
       x = "Population",
       y = "Residuals (Actual - Predicted Funding)") +
  theme_minimal()

print(residual_plot)
```



The variability in funding per capita likely reflects specific state circumstances, such as economic conditions, geographic challenges, or policy priorities.

```
library(ggplot2)
library(dplyr)
library(broom)

# Plot funding per capita
funding_per_capita_plot <- ggplot(merged_data, aes(x = reorder(State, -Funding_Per_Capita), y = Funding_
  geom_bar(stat = "identity", fill = "blue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)) +
  labs(title = "Funding per Capita by State",
       x = "State",
       y = "Funding per Capita ($)") +
  theme_minimal()

print(funding_per_capita_plot)
```



States with funding per capita more than 2 standard deviations away from the mean are considered outliers. These include Alaska, Wyoming, and Montana, all of which have significantly higher funding per capita than the average.

```
library(ggplot2)
library(dplyr)
library(broom)

# Identify outliers (states with funding per capita more than 2 SD from the mean)
outliers <- merged_data %>%
  filter(abs(Funding_Per_Capita - mean(Funding_Per_Capita)) > 2 * sd(Funding_Per_Capita)) %>%
  select(State, Population, Total_Billions, Funding_Per_Capita) %>%
  arrange(desc(Funding_Per_Capita))

print("Outliers (states with funding per capita more than 2 SD from the mean):")
```

```
## [1] "Outliers (states with funding per capita more than 2 SD from the mean):"
```

```
print(outliers)
```

```
## # A tibble: 3 x 4
##   State   Population Total_Billions Funding_Per_Capita
##   <chr>        <dbl>          <dbl>              <dbl>
## 1 ALASKA      733374            3.7              5045.
## 2 WYOMING     576850            2.3              3987.
## 3 MONTANA    1084244            3.3              3044.
```

Conclusion: *Equitability:* While there is a strong correlation between population size and allocated funding, there are considerable variations in funding per capita. This suggests that additional factors influence funding decisions.

Bias: The high per capita funding seen in states like Alaska and Wyoming could reflect special circumstances necessitating higher investment, like higher infrastructure costs due to geographic and climatic challenges.

- Does the allocation favor the political interests of the Biden administration? We'll need to analyze the IIJA funding allocation based on each state's political leanings. We'll use the 2020 Presidential Election results as a proxy for political alignment. Let's compare this data to the IIJA funding allocation.

```r
library(readxl)
library(dplyr)

# Read IIJA funding data
iija_funding <- read_excel("IIJA FUNDING AS OF MARCH 2023.xlsx", sheet = "Sheet1")
colnames(iija_funding) <- c("State", "Total_Billions")

# Read population data
population_report <- read_excel("PopulationReport.xlsx", sheet = "PopulationReport")
population_report <- population_report %>%
  select(Name, `Pop. 2020`) %>%
  rename(State = Name, Population = `Pop. 2020`)

# Read 2020 election results
election_results <- read_excel("2020USPresidentialElectionResult.xlsx")

# Display the first few rows and structure of each dataset
print("IIJA Funding Data:")
```

```
## [1] "IIJA Funding Data:"
```

```r
print(head(iija_funding))
```

```
## # A tibble: 6 x 2
##   State          Total_Billions
##   <chr>                   <dbl>
## 1 ALABAMA                3
## 2 ALASKA                 3.7
## 3 AMERICAN SAMOA         0.0686
## 4 ARIZONA                3.5
## 5 ARKANSAS               2.8
## 6 CALIFORNIA            18.4
```

```r
print(str(iija_funding))
```

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

```r
print("\
Population Data:")
```

```
## [1] "\nPopulation Data:"
```

```r
print(head(population_report))
```

```
## # A tibble: 6 x 2
##   State         Population
##   <chr>              <dbl>
## 1 United States  331464948
## 2 Alabama          5024294
## 3 Alaska            733374
## 4 Arizona          7157902
## 5 Arkansas         3011490
## 6 California      39538212
```

```r
print(str(population_report))
```

```
## tibble [55 x 2] (S3: tbl_df/tbl/data.frame)
##  $ State     : chr [1:55] "United States" "Alabama" "Alaska" "Arizona" ...
##  $ Population: num [1:55] 3.31e+08 5.02e+06 7.33e+05 7.16e+06 3.01e+06 ...
## NULL
```

```r
print("\
Election Results Data:")
```

```
## [1] "\nElection Results Data:"
```

```r
print(head(election_results))
```

```
## # A tibble: 6 x 8
##   state  state_abr trump_pct biden_pct trump_vote biden_vote trump_win biden_win
##   <chr>  <chr>         <dbl>     <dbl>      <dbl>      <dbl>     <dbl>     <dbl>
## 1 Alaska AK             53.1      43       189543     153502         1         0
## 2 Hawaii HI             34.3      63.7     196864     366130         0         1
## 3 Washi~ WA             39        58.4    1584651    2369612         0         1
## 4 Oregon OR             40.7      56.9     958448    1340383         0         1
## 5 Calif~ CA             34.3      63.5    5982194   11082293         0         1
## 6 Idaho  ID             63.9      33.1     554128     287031         1         0
```

```r
print(str(election_results))
```

```
## tibble [51 x 8] (S3: tbl_df/tbl/data.frame)
##  $ state     : chr [1:51] "Alaska" "Hawaii" "Washington" "Oregon" ...
##  $ state_abr : chr [1:51] "AK" "HI" "WA" "OR" ...
##  $ trump_pct : num [1:51] 53.1 34.3 39 40.7 34.3 63.9 56.9 47.7 70.4 58.2 ...
##  $ biden_pct : num [1:51] 43 63.7 58.4 56.9 63.5 33.1 40.6 50.1 26.7 37.7 ...
##  $ trump_vote: num [1:51] 189543 196864 1584651 958448 5982194 ...
```

```
##  $ biden_vote: num [1:51] 153502 366130 2369612 1340383 11082293 ...
##  $ trump_win : num [1:51] 1 0 0 0 0 1 1 0 1 1 ...
##  $ biden_win : num [1:51] 0 1 1 1 1 0 0 1 0 0 ...
## NULL
```

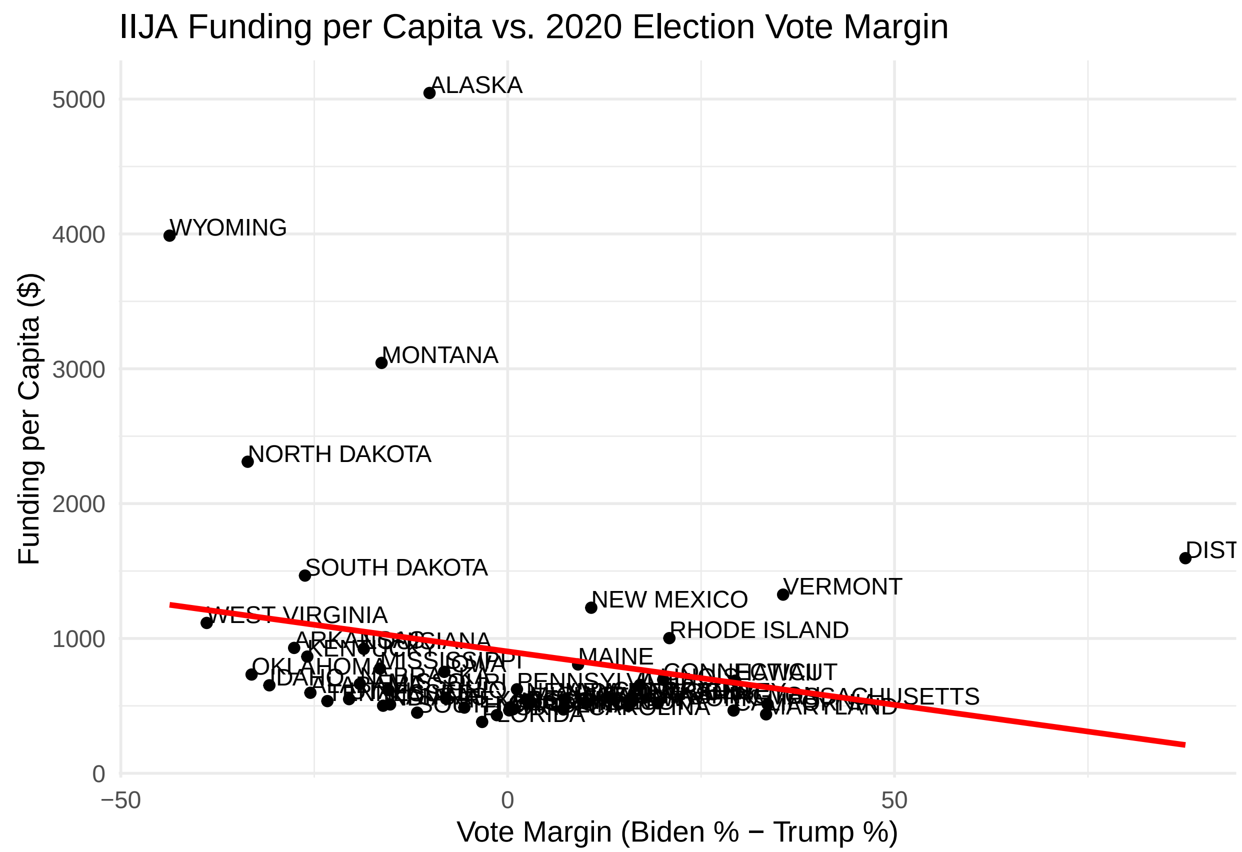
We'll merge the datasets and perform some analysis to look for potential political bias.

```r
library(dplyr)
library(ggplot2)

# Clean and merge data
iija_funding$State <- toupper(trimws(iija_funding$State))
population_report$State <- toupper(trimws(population_report$State))
election_results$state <- toupper(trimws(election_results$state))

merged_data <- iija_funding %>%
  inner_join(population_report, by = "State") %>%
  inner_join(election_results, by = c("State" = "state"))

# Calculate funding per capita and vote margin
merged_data <- merged_data %>%
  mutate(Funding_Per_Capita = (as.numeric(Total_Billions) * 1e9) / Population,
         Vote_Margin = biden_pct - trump_pct)

# Correlation between funding per capita and vote margin
correlation <- cor(merged_data$Funding_Per_Capita, merged_data$Vote_Margin)

# Create scatter plot
plot <- ggplot(merged_data, aes(x = Vote_Margin, y = Funding_Per_Capita, label = State)) +
  geom_point() +
  geom_text(aes(label=State), hjust=0, vjust=0, size=3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "IIJA Funding per Capita vs. 2020 Election Vote Margin",
       x = "Vote Margin (Biden % - Trump %)",
       y = "Funding per Capita ($)") +
  theme_minimal()

# Display results
print(paste("Correlation between Funding per Capita and Vote Margin:", round(correlation, 4)))
```

```
## [1] "Correlation between Funding per Capita and Vote Margin: -0.2136"
```

```r
print(plot)
```

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

## Warning: The following aesthetics were dropped during statistical transformation: label.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```r
# Top 10 states by funding per capita
top_10 <- merged_data %>%
  arrange(desc(Funding_Per_Capita)) %>%
  select(State, Funding_Per_Capita, Vote_Margin) %>%
  head(10)

print("Top 10 states by funding per capita:")
```

```
## [1] "Top 10 states by funding per capita:"
```

```r
print(top_10)
```

```
## # A tibble: 10 x 3
##    State                Funding_Per_Capita Vote_Margin
##    <chr>                             <dbl>       <dbl>
##  1 ALASKA                            5045.       -10.1
##  2 WYOMING                           3987.       -43.7
##  3 MONTANA                           3044.       -16.3
##  4 NORTH DAKOTA                      2310.       -33.6
##  5 DISTRICT OF COLUMBIA              1595.        87.6
##  6 SOUTH DAKOTA                      1466.       -26.2
##  7 VERMONT                           1325.        35.6
##  8 NEW MEXICO                        1228.        10.8
##  9 WEST VIRGINIA                     1115.       -38.9
## 10 RHODE ISLAND                      1002.        20.9
```

```r
# Bottom 10 states by funding per capita
bottom_10 <- merged_data %>%
  arrange(Funding_Per_Capita) %>%
  select(State, Funding_Per_Capita, Vote_Margin) %>%
  head(10)

print("Bottom 10 states by funding per capita:")
```

```
## [1] "Bottom 10 states by funding per capita:"
```

```r
print(bottom_10)
```

```
## # A tibble: 10 x 3
##    State          Funding_Per_Capita Vote_Margin
##    <chr>                       <dbl>       <dbl>
##  1 FLORIDA                      381.      -3.30
##  2 NORTH CAROLINA               431.      -1.40
##  3 MARYLAND                     437.      33.4
##  4 SOUTH CAROLINA               449.     -11.7
##  5 CALIFORNIA                   465.      29.2
##  6 GEORGIA                      467.       0.200
##  7 MINNESOTA                    473.       7.2
##  8 WISCONSIN                    475.       0.700
##  9 TEXAS                        487.      -5.6
## 10 ARIZONA                      489.       0.300
```

```r
# Average funding per capita for Biden-won vs Trump-won states
avg_funding <- merged_data %>%
  mutate(Winner = ifelse(Vote_Margin > 0, "Biden", "Trump")) %>%
  group_by(Winner) %>%
  summarise(Avg_Funding_Per_Capita = mean(Funding_Per_Capita))

print("Average funding per capita by 2020 election winner:")
```

```
## [1] "Average funding per capita by 2020 election winner:"
```

```r
print(avg_funding)
```

```
## # A tibble: 2 x 2
##   Winner Avg_Funding_Per_Capita
##   <chr>                   <dbl>
## 1 Biden                    669.
## 2 Trump                   1155.
```

Conclusion The correlation between funding per capita and vote margin is -0.2136. This indicates a very weak negative correlation, suggesting no significant relationship between the political leaning of the states (as measured by vote margin) and the funding allocation.

States like Alaska, Wyoming, and Montana have high funding per capita with varying vote margins, while states like Florida and California have lower funding per capita.

States won by Trump have an average funding per capita of $1,155.27, while states won by Biden have an average of $669.22.

Bias Analysis: The weak correlation (-0.2136) combined with the similar average funding per capita across Biden and Trump states suggests there is no substantial bias in the funding allocation favoring the political interests of the Biden administration. States received funding varying greatly in alignment, and the distribution appears to be more influenced by other factors.