Final Project

Xinyi Zhu

2025-06-26

1.Introduction

In this project, I want to explore how population, GDP, and unemployment rate vary across U.S. states. I'm curious to see which states have large or small populations, which ones have stronger economies, and whether there's any connection between unemployment and either population or GDP.

Here are the three main questions I'll focus on:

1. How does the population size vary across different U.S. states?
2. What are the differences in total GDP and per capita GDP among states?
3. Is there a relationship between unemployment rate and GDP or population?

I think this is worth looking into because it can help show which states are doing better economically and give some insights into how population and jobs might be connected. It's also a good way to practice working with real data and making clear visualizations.

2.Background Information

The U.S. economy varies widely by state. Some states have booming industries and large populations, while others are smaller or more rural. By looking at economic and population data, we can get a clearer picture of how these differences play out across the country.

I'm using data from three Wikipedia pages. All of them give recent stats for U.S. states. I'm only using the 50 states and D.C. to avoid issues with territories that have older or missing data.

The **GDP** data is from the Wikipedia page "List of U.S. states and territories by GDP":
https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_GDP
The original source is the Bureau of Economic Analysis (BEA):

"GDP by State". U.S. Bureau of Economic Analysis. Retrieved April 10, 2022.
https://www.bea.gov/data/gdp/gdp-state

The **population** data is from "List of U.S. states and territories by population":
https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population
The Wikipedia page references multiple sources from the U.S. Census Bureau, including:
"Annual and cumulative estimates of residential population change for the United States, regions, states, District of Columbia, Puerto Rico". U.S. Census Bureau. Retrieved December 20, 2024.
https://www.census.gov
Also includes historical census tables and population estimates from various years.

The **unemployment rate** data is from "List of U.S. states and territories by unemployment rate":
https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_unemployment_rate
The original source is the Bureau of Labor Statistics (BLS):
"Local Area Unemployment Statistics". Bureau of Labor Statistics. Retrieved June 2023.
https://www.bls.gov/lau/

Each dataset is a table where rows are states and columns include things like GDP, population, or unemployment rate. I cleaned the data by filtering out the territories and keeping only states with recent and complete data.

Next, I will organize and visualize these data to examine how the economic size, population, and unemployment rate vary across different states. This will help us understand the economic differences among states and uncover some interesting patterns.

3.Data Summary

I used R to pull the data directly from Wikipedia using the `rvest` package. I had to clean up each dataset before merging.

For the GDP data, things were pretty clean. I only needed to remove the first row (which is a duplicate header) and the last row (which is the U.S. total). That left me with just the 50 states plus D.C., which is what I wanted.

The population and unemployment datasets had more rows that included territories and other info I didn't need. So I decided to use the list of states from the GDP dataset as my base and then joined the other two datasets to it.

In the end, I got a nice clean dataset with each state's GDP, per capita GDP, population, and unemployment rate.

You can see the full code I used in the code appendix.

4.Exploratory Data Analysis

Summary Statistics Table

See Table 1 for an overview of the combined data.

Table 1: Summary statistics for population, GDP, GDP per capita, and unemployment rate

State	GDP	GDP_per_capita	Population	Seasonally_adjusted_rates
California	4103124	104916	39431263	4.6
Texas	2709393	86987	31290831	4.1
New York	2297028	117332	19867248	3.9
Florida	1705565	73784	23372215	2.6
Illinois	1137244	90449	12710158	4.0
Pennsylvania	1024206	78544	13078751	3.8
Ohio	927740	78120	11883304	3.4
Georgia	882535	78754	11180878	3.2
Washington	854683	108468	7958180	3.8
New Jersey	846587	90272	9500851	3.7
North Carolina	839122	75876	11046024	3.3
Massachusetts	780666	110561	7136171	2.6
Virginia	764475	86747	8811195	2.7
Michigan	719392	71083	10140459	3.6
Colorado	553323	93026	5957493	2.8
Arizona	552167	73203	7582384	3.5
Tennessee	549709	75748	7227750	3.2
Maryland	542766	87021	6263220	2.0
Indiana	527381	76004	6924275	3.2
Minnesota	500851	86371	5793151	2.9
Wisconsin	451285	75605	5960975	2.5
Missouri	451201	72108	6245466	2.6
Connecticut	365723	100235	3675069	3.7
South Carolina	349965	63711	5478831	3.1
Oregon	331029	77916	4272371	3.5
Louisiana	327782	71642	4597740	3.6
Alabama	321238	61846	5157699	2.2
Utah	300904	86506	3503613	2.4
Kentucky	293021	64110	4588372	3.8
Oklahoma	265779	64719	4095393	2.7
Nevada	260728	80880	3267467	5.4
Iowa	257021	79631	3241488	2.7
Kansas	234673	79513	2970606	2.8

Table 1: Summary statistics for population, GDP, GDP per capita, and unemployment rate

State	GDP	GDP_per_capita	Population	Seasonally_adjusted_rates
Arkansas	188723	60276	3088354	2.6
Nebraska	185411	93145	2005465	1.9
District of Columbia	184916	263220	702250	5.1
Mississippi	157491	53061	2943045	3.1
New Mexico	140542	66229	2130256	3.5
Idaho	128132	63991	2001619	2.7
New Hampshire	121189	85518	1409032	1.8
Hawaii	115627	80325	1446146	3.0
West Virginia	107660	60783	1769979	3.3
Delaware	103253	98055	1051917	4.2
Maine	98606	69803	1405012	2.4
Rhode Island	82493	74594	1112308	2.9
Montana	75999	66379	1137233	2.4
North Dakota	75399	95982	796568	2.0
South Dakota	75179	80685	924669	1.8
Alaska	69969	95147	740133	3.7
Wyoming	52946	90335	587618	3.1
Vermont	45707	70131	648493	1.9

This table shows the basic stats for the main data points, like population, GDP, and unemployment. It helps get a quick idea of the overall data and what to expect.

Table: Top 5 States by GDP

As shown in Table 2, the top five states by GDP are led by California.

Table 2: Top 5 States by GDP

State	GDP
California	4103124
Texas	2709393
New York	2297028
Florida	1705565
Illinois	1137244

Table: Summary Statistics of GDP and GDP per Capita Across U.S. States

As shown in Table 3 & Table 4, the summary statistics for GDP and GDP per capita across U.S. states reveal notable differences. The maximum GDP is heavily influenced by states like California and Texas, while GDP per capita varies significantly among states, indicating economic disparities.

Table 3: Summary Statistics of GDP Across U.S. States

Min_GDP	Max_GDP	Avg_GDP	Median_GDP
45707	4103124	569363.7	327782

Table 4: Summary Statistics of GDP per Capita Across U.S. States

Min_GDPperCapita	Max_GDPperCapita	Avg_GDPperCapita	Median_GDPperCapita
53061	263220	84104.84	78754

Data Visualizations

Visualization 1: Population vs State

As shown in Figure 1, the population size varies widely across states. This bar chart shows how the population varies across states. States like California and Texas clearly stand out with much larger populations compared to others. Knowing population size is important because it affects everything from the economy to resources.

Visualization 2: GDP vs State

As shown in Figure 2, total GDP varies significantly across states. Big states like California, Texas, and New York clearly lead with much higher GDP compared to others. This highlights which states have the strongest economies overall.

Visualization 3: GDP per Capita vs State

As shown in Figure 3, GDP per capita reveals that some smaller states have high average wealth. This means that while their total GDP might be smaller, people there tend to be wealthier on average. It's a useful way to understand economic well-being beyond just total GDP.

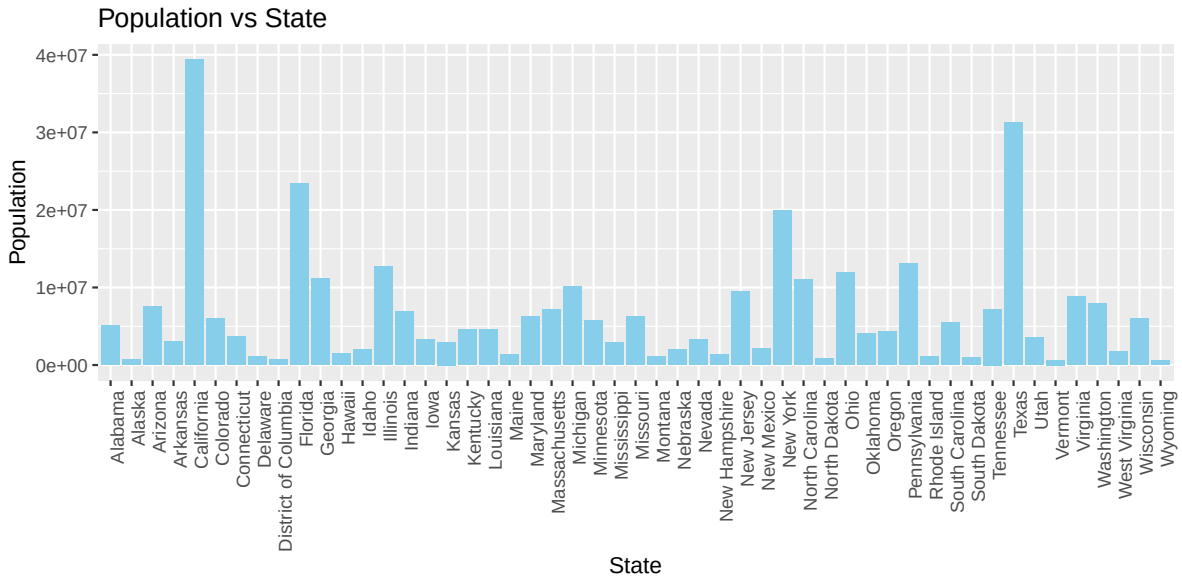

Figure 1: Figure 1: Population size varies across U.S. states, with California and Texas the largest.

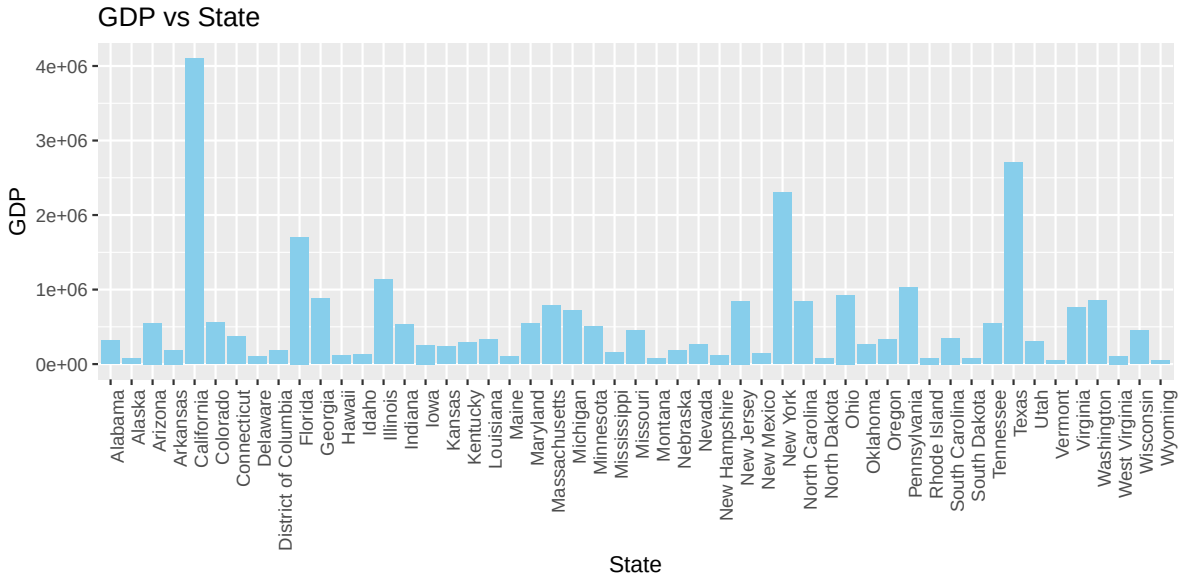

Figure 2: Figure 2: Total GDP shows California and Texas leading the economy.

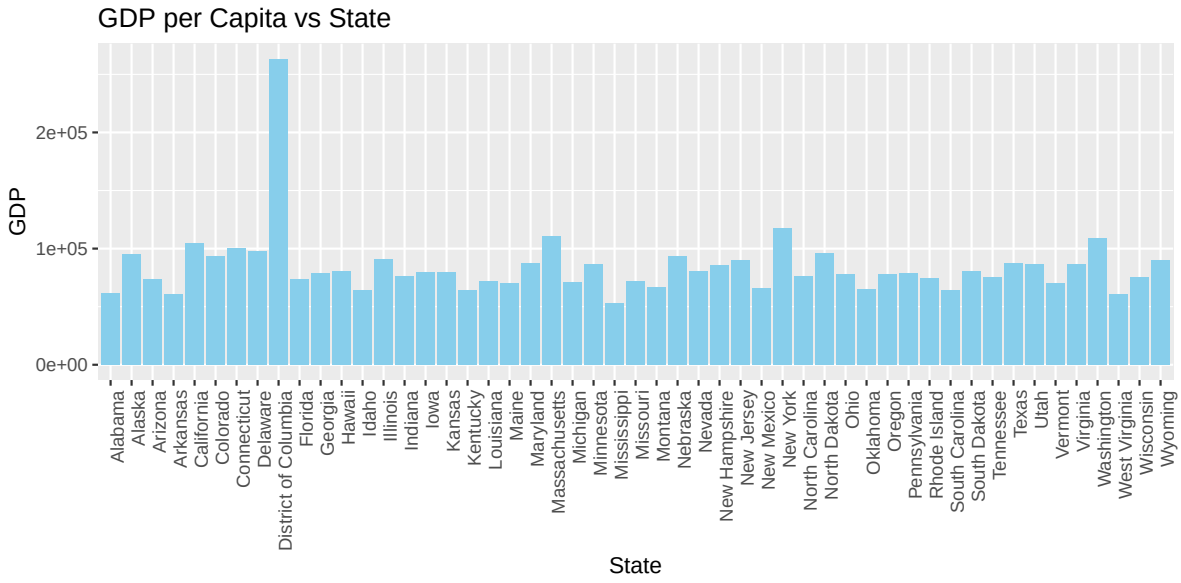

Figure 3: Figure 3: GDP per capita shows smaller states can have high average wealth.

Visualization 4: Unemployment Rate vs GDP

As shown in Figure 4, the scatter plot illustrates a general trend where states with higher GDP tend to have lower unemployment rates. This suggests that stronger economies usually correspond with more job opportunities. However, it's not a perfect correlation, as other factors also influence unemployment. This visualization helps us understand the relationship between economic size and the job market.

5.Conclusion

Here's what I found from the data:

- **Some states have way more people than others.** California and Texas are the biggest. States like Wyoming and Vermont have much smaller populations.
- **Big states also have higher total GDP.** California, Texas, and New York lead here — not surprising, since they have more people and business.
- **GDP per person tells a different story.** States like Delaware and Alaska have fewer people but still high GDP per person. So small states can still be rich.
- **Unemployment doesn't fully follow GDP.** States with more GDP tend to have lower unemployment, but not always. Other stuff matters too.

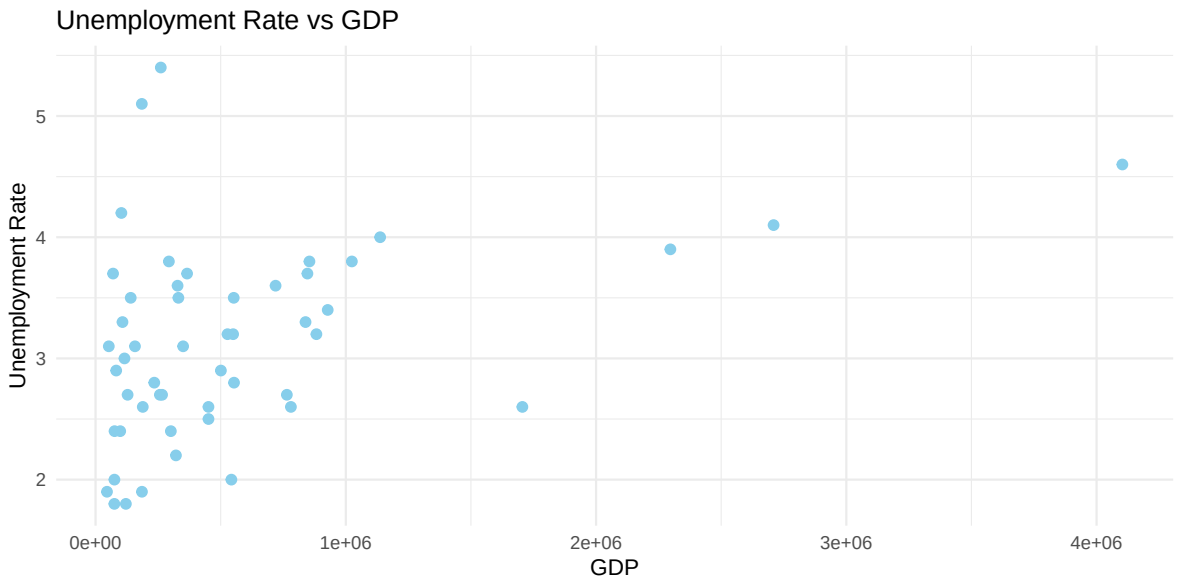

Figure 4: Figure 4: Scatter plot shows a general trend of lower unemployment in states with higher GDP.

So overall, population, total GDP, and GDP per person all show different things. Big states make more money overall, but smaller ones can still be strong when you look at GDP per person.

Unemployment is a bit more random — higher GDP doesn't always mean more jobs. There's probably more going on, like different industries or job types.

If I had more time, I'd look at things like spending, jobs by sector, or education levels. That would give an even better picture.

6.References

Here are the data sources I used:

- Wikipedia contributors. *List of U.S. states and territories by GDP.* Wikipedia. https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_GDP (Original source: U.S. Bureau of Economic Analysis, "GDP by State")
- Wikipedia contributors. *List of U.S. states and territories by population.* Wikipedia. https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population (Original source: U.S. Census Bureau)

- Wikipedia contributors. *List of U.S. states and territories by unemployment rate.* Wikipedia. https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_unemployment_rate (Original source: U.S. Bureau of Labor Statistics)

7.Code Appendix

```
# Style Guide: Using Tidyverse Style Guide https://style.tidyverse.org/

library(tidyverse)
library(rvest)
library(knitr)

# GDP data
gdp_url <- "https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_GDP"
gdp_table <- gdp_url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table(fill = TRUE)
gdp_data <- gdp_table[[1]]
gdp_clean <- gdp_data %>%
  select(State = 1, GDP = 3, GDP_per_capita = 8) %>%
  slice(-1) %>%
  slice(-52) %>%
  mutate(
    GDP = as.numeric(gsub(",", "", GDP)),
    GDP_per_capita = as.numeric(gsub("[$,]", "", GDP_per_capita))
  )

# Population data
pop_url <- "https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population"
pop_tables <- pop_url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table(fill = TRUE)
pop_data <- pop_tables[[1]]
pop_clean <- pop_data %>%
  select(State = 1, Population = 2)%>%
  mutate(Population = as.numeric(gsub(",", "", Population)))
```

```r
# Unemployment data
unemp_url <- "https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_unemployme
unemp_tables <- unemp_url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table(fill = TRUE)
unemp_data <- unemp_tables[[1]]
unemp_clean <- unemp_data %>%
  select(State = 2,Seasonally_adjusted_rates = 3)%>%
  mutate(Seasonally_adjusted_rates = as.numeric(gsub("%", "", Seasonally_adjusted_rates)))

# Summary all three data sets
sum_data <- gdp_clean %>%
  left_join(pop_clean, by = "State") %>%
  left_join(unemp_clean, by = "State")

# Top 5 States by GDP
sum_data %>%
  arrange(desc(GDP)) %>%
  slice_head(n = 5) %>%
  select(State, GDP)

# Table: Summary Statistics of GDP and GDP per Capita Across U.S. States
summary_stats1 <- sum_data %>%
  summarise(
    Min_GDP = min(GDP, na.rm = TRUE),
    Max_GDP = max(GDP, na.rm = TRUE),
    Avg_GDP = mean(GDP, na.rm = TRUE),
    Median_GDP = median(GDP, na.rm = TRUE),
  )
summary_stats2 <- sum_data %>%
  summarise(
    Min_GDPperCapita = min(GDP_per_capita, na.rm = TRUE),
    Max_GDPperCapita = max(GDP_per_capita, na.rm = TRUE),
    Avg_GDPperCapita = mean(GDP_per_capita, na.rm = TRUE),
    Median_GDPperCapita = median(GDP_per_capita, na.rm = TRUE)
  )

# Visualization 1: Population vs State
sum_data %>%
  ggplot(aes(x = State, y = Population)) +
  geom_col(fill = "skyblue") +
```

```r
  labs(title = "Population vs State", x = "State", y = "Population") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Visualization 2: GDP vs State
sum_data %>%
  ggplot(aes(x = State, y = GDP)) +
  geom_col(fill = "skyblue") +
  labs(title = "GDP vs State", x = "State", y = "GDP") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Visualization 3: GDP per Capita vs State
sum_data %>%
  ggplot(aes(x = State, y = GDP_per_capita)) +
  geom_col(fill = "skyblue") +
  labs(title = "GDP per Capita vs State", x = "State", y = "GDP") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Visualization 4: Unemployment Rate vs GDP
sum_data %>%
  ggplot(aes(x = GDP, y = Seasonally_adjusted_rates)) +
  geom_point(color = "skyblue", size = 2) +
  labs(title = "Unemployment Rate vs GDP", x = "GDP", y = "Unemployment Rate") +
  theme_minimal()

write.csv(gdp_clean, "gdp_clean.csv", row.names = FALSE)
write.csv(pop_clean, "pop_clean.csv", row.names = FALSE)
write.csv(unemp_clean, "unemp_clean.csv", row.names = FALSE)
```