Walmart Store Distribution and Its Correlation with State GDP in the U.S.

Qianhui Dai, Joseph Easterday

2025-04-30

Table of contents

1 Introduction

This report analyzes the relationship between Walmart store distribution and economic activity (measured by GDP) across U.S. states. Walmart, as the largest retailer in the United States, strategically locates its stores based on demographic and economic factors. By examining

the correlation between the number of Walmart stores per state and state-level GDP, we aim to identify patterns in retail market penetration relative to economic output. Key questions include:

Do states with higher GDP have more Walmart stores?

Which states show the highest/lowest GDP per Walmart store?

Are there regional economic trends reflected in Walmart's distribution?

This analysis provides insights for business strategy, economic research, and retail market analysis.

2 Source and Background of the data

2.1 Walmart Store Distribution Data

This data was found and tidied by Qianhui Dai. Source: The dataset lists the number of Walmart stores (total_stores) per U.S. state (including territories like Puerto Rico and Washington, D.C.). Background: Walmart operates over 4,700 stores nationwide, with density influenced by population, income levels, and urbanization. States like Texas (595 stores) and Florida (386 stores) have the highest counts, reflecting their large populations and consumer demand.

2.2 State GDP Data

This data was found and tidied by Joseph Easterday. Source: The dataset provides 2024 GDP estimates in millions of USD (GDP_In_Millions_2024) for each state, sourced from official economic reports (e.g., U.S. Bureau of Economic Analysis). Background:GDP measures a state's economic output. High-GDP states like California (4.1 trillion) and Texas (2.7 trillion) drive national economic activity, while smaller economies(e.g., Vermont, Wyoming) may have different retail dynamics.

3 FAIR/CARE Principle

1. Cleaning and Structuring the data to meet the FAIR principle.

a. Removed unnecessary header rows and renamed unclear column names.
b. Standardized state names and abbreviations to ensure consistency across datasets.
c. Converted numeric values from text to proper numeric formats for analysis.

2. Merging Datasets to let it be interoperable

a. Merged Walmart store data with state GDP data using a common key, enabling us to analyze relationships across sources.
b. Ensured coordinate and region data from Walmart were preserved to allow for geographic visualization and spatial analysis.

3. Creating Visual and Statistical Outputs to let it be reusable

a. Generated scatter plots and bar plots to illustrate trends and patterns.
b. Performed correlation analysis between GDP and Walmart distribution to uncover relationships.
c. Plan to compile all steps into a reproducible Quarto report, so others can access and reuse our workflow.

4. Respecting CARE Principles

a. Collective Benefit: The analysis may help identify economic development patterns and retail accessibility, benefiting researchers and policymakers.
b. Authority to Control: We used publicly available data from Walmart official Website and official website of United States to respect the data ownership.
c. Responsibility & Ethics: We ensured no personal or sensitive data was included, and we avoided drawing misleading conclusions from preliminary results.

4 Data Visualizations and Results

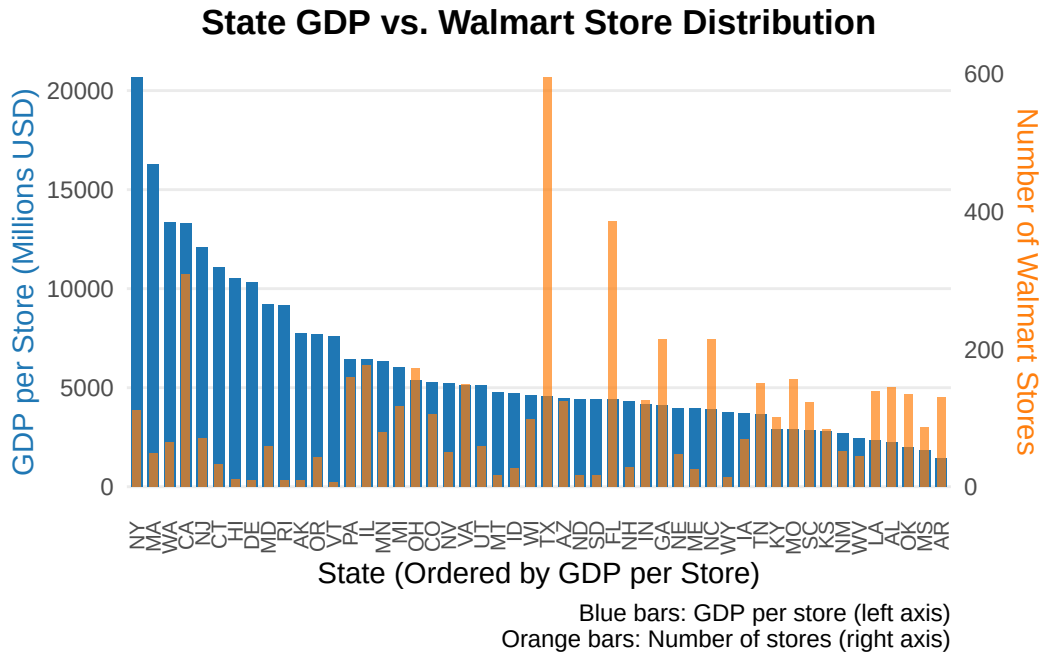
In this project, we examined the relationship between Walmart store distribution across U.S. states and each state's gross domestic product (GDP). Using two datasets — one listing the number of Walmart stores per state, and another detailing state GDP — we created multiple visualizations to explore potential patterns and correlations between store counts and economic output.

```
# A tibble: 52 x 3
   StateName      StateAbbr Stores
   <chr>          <chr>      <dbl>
 1 Texas          TX           595
 2 Florida        FL           386
 3 California     CA           308
 4 Georgia        GA           214
 5 North Carolina NC           214
 6 Illinois       IL           177
 7 Ohio           OH           172
 8 Pennsylvania   PA           159
 9 Missouri       MO           156
```

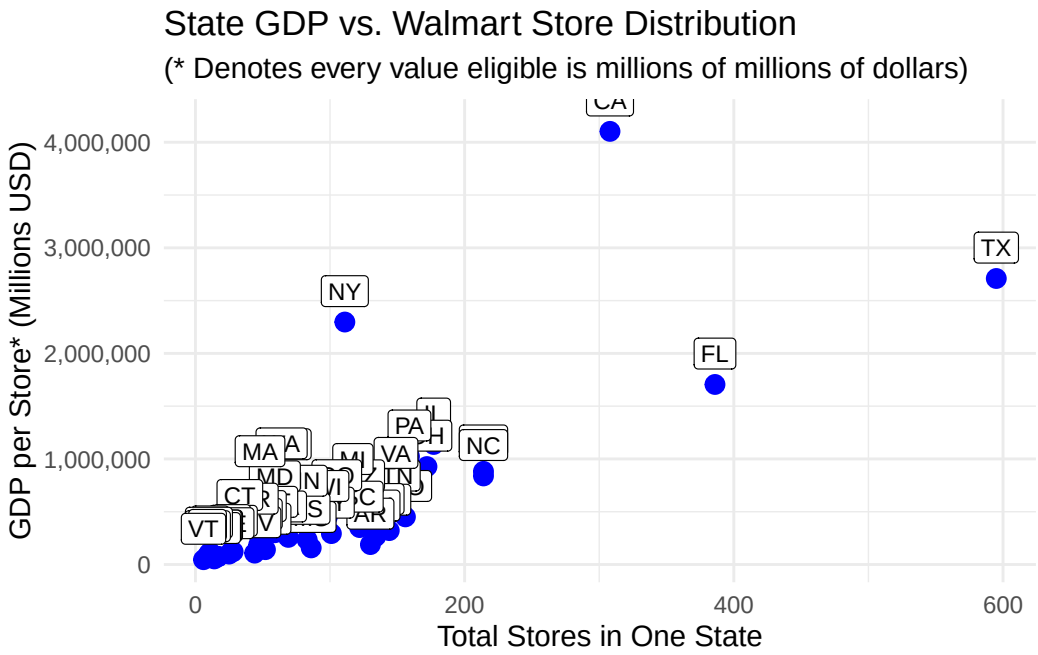
```
10 Tennessee      TN           150
# i 42 more rows
```

```
# A tibble: 51 x 3
   GeoName     StateAbbr GDP_In_Millions_2024
   <chr>       <chr>                    <dbl>
 1 Alabama     AL                     321238.
 2 Alaska      AK                      69969
 3 Arizona     AZ                     552167
 4 Arkansas    AR                     188723.
 5 California  CA                    4103124.
 6 Colorado    CO                     553322.
 7 Connecticut CT                     365723.
 8 Delaware    DE                     103253.
 9 Florida     FL                    1705565.
10 Georgia     GA                     882534.
# i 41 more rows
```

The bar plot revealed that states with higher GDP (e.g., California, Texas) tend to have a larger number of Walmart stores. However, some states with high GDP (like New York) did not have proportionally high Walmart store counts, suggesting regional market or policy differences.



The scatter plot confirmed a moderate positive correlation between state GDP and the number of Walmart stores. This suggests that while GDP is a general predictor of store presence, it is not the only factor — other considerations like population density, land availability, and corporate strategy may influence store distribution.



These findings can help Walmart and other retailers assess potential markets for expansion based on economic indicators. By visualizing how store presence aligns with economic activity, decision-makers can identify underserved areas or prioritize states with untapped potential. Moreover, the Pearson correlation coefficient between state GDP and the number of Walmart stores is approximately 0.75, indicating a strong positive linear relationship. This suggests that, generally, states with higher economic output tend to host more Walmart locations.

```
$pearson
$pearson$estimate
      cor
0.7496365

$pearson$p_value
[1] 3.750816e-10

$pearson$conf_int
[1] 0.5955575 0.8505159
attr(,"conf.level")
```

```
[1] 0.95


$spearman
$spearman$estimate
      rho
0.8123229

$spearman$p_value
[1] 8.025663e-13


$correlation_matrix
             gdp    stores
gdp    1.0000000 0.7496365
stores 0.7496365 1.0000000
```

5 Conclusion

In this project, we investigated the relationship between the number of Walmart stores in each U.S. state and the corresponding state-level GDP. By merging and analyzing data from Walmart's public store distribution and state GDP statistics, we aimed to explore whether economic strength correlates with retail presence.

Our visualizations — including a bar chart comparing store count and GDP by state, a scatter plot, and a correlation matrix — revealed a clear pattern: states with higher GDPs generally have more Walmart locations. Most notably, our correlation analysis showed a Pearson coefficient of 0.75, indicating a strong positive linear relationship between GDP and Walmart store count. This suggests that Walmart tends to concentrate stores in economically stronger states, likely due to higher consumer demand, infrastructure, and urban development.

While GDP proved to be a strong indicator, we acknowledge that it is not the sole factor influencing store distribution. States like New York and New Jersey, for example, showed lower-than-expected store counts despite having large economies — possibly due to high real estate costs, population density, or regional business strategies. Including population, land area, or urbanization metrics in future analyses could offer a more complete explanation.

Overall, our findings highlight how data visualization and correlation analysis can offer practical insights for business strategy, retail planning, and economic geography. This analysis may assist retailers or policy makers in identifying market opportunities or evaluating how commercial infrastructure aligns with economic activity.

6 References

Bureau of Economic Analysis. "GDP by State (Annual)". U.S. Department of Commerce, https://apps.bea.gov/itable/?ReqID=70&step=1.

Walmart Tech. "Walmart Store Status Public Dataset". Walmart Open Data Hub, https://walmart-open-data-walmarttech.opendata.arcgis.com/datasets/39ce1c357bd2424ca481db84aed29464_0

7 Code Appendix

7.1 Store distribution table

```
library(dplyr)
library(readxl)
library(writexl)

walmart_raw <- read_excel("~/Desktop/walmart_store_distribution.xlsx")

state_abbr_df <- tibble(
  StateName = c(
    "Alabama", "Alaska", "Arizona", "Arkansas",
    "California", "Colorado", "Connecticut", "Delaware",
    "Florida", "Georgia", "Hawaii", "Idaho",
    "Illinois", "Indiana", "Iowa", "Kansas",
    "Kentucky", "Louisiana", "Maine", "Maryland",
    "Massachusetts", "Michigan", "Minnesota", "Mississippi",
    "Missouri", "Montana", "Nebraska", "Nevada",
    "New Hampshire", "New Jersey", "New Mexico", "New York",
    "North Carolina", "North Dakota", "Ohio", "Oklahoma",
    "Oregon", "Pennsylvania", "Rhode Island", "South Carolina",
    "South Dakota", "Tennessee", "Texas", "Utah",
    "Vermont", "Virginia", "Washington", "West Virginia",
    "Wisconsin", "Wyoming"
  ),
  StateAbbr = c(
    "AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE",
    "FL", "GA", "HI", "ID", "IL", "IN", "IA", "KS",
    "KY", "LA", "ME", "MD", "MA", "MI", "MN", "MS",
    "MO", "MT", "NE", "NV", "NH", "NJ", "NM", "NY",
    "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC",
```

```
    "SD", "TN", "TX", "UT", "VT", "VA", "WA", "WV",
    "WI", "WY"
  )
)

walmart_tidy <- walmart_raw %>%
  rename(Stores = total_stores) %>%
  left_join(state_abbr_df, by = c("state" = "StateAbbr")) %>%
  select(StateName, StateAbbr = state, Stores) %>%
  arrange(desc(Stores))

print(walmart_tidy)

write_xlsx(walmart_tidy, "~/Desktop/walmart_distribution_tidy.xlsx")
```

7.2 State GDP table

```
library(dplyr)
library(readxl)
library(writexl)

raw_data <- read_excel("~/Downloads/Gross_Domestic_Product.xlsx", skip = 5)

cleaned_data <- raw_data %>%
  select(GeoFips, GeoName, `2024`) %>%
  filter(!GeoName %in% c("District of Columbia", "United States"))

state_abbr_df <- tibble(
  GeoName = c(
    "Alabama", "Alaska", "Arizona", "Arkansas",
    "California", "Colorado", "Connecticut", "Delaware",
    "Florida", "Georgia", "Hawaii", "Idaho",
    "Illinois", "Indiana", "Iowa", "Kansas",
    "Kentucky", "Louisiana", "Maine", "Maryland",
    "Massachusetts", "Michigan", "Minnesota", "Mississippi",
    "Missouri", "Montana", "Nebraska", "Nevada",
    "New Hampshire", "New Jersey", "New Mexico", "New York",
    "North Carolina", "North Dakota", "Ohio", "Oklahoma",
    "Oregon", "Pennsylvania", "Rhode Island", "South Carolina",
    "South Dakota", "Tennessee", "Texas", "Utah",
```

```
    "Vermont", "Virginia", "Washington", "West Virginia",
    "Wisconsin", "Wyoming"
  ),
  StateAbbr = c(
    "AL", "AK", "AZ", "AR", "CA", "CO", "CT", "DE",
    "FL", "GA", "HI", "ID", "IL", "IN", "IA", "KS",
    "KY", "LA", "ME", "MD", "MA", "MI", "MN", "MS",
    "MO", "MT", "NE", "NV", "NH", "NJ", "NM", "NY",
    "NC", "ND", "OH", "OK", "OR", "PA", "RI", "SC",
    "SD", "TN", "TX", "UT", "VT", "VA", "WA", "WV",
    "WI", "WY"
  )
)

tidy_data <- cleaned_data %>%
  left_join(state_abbr_df, by = "GeoName") %>%
  group_by(GeoName, StateAbbr) %>%
  summarise(
    GDP_In_Millions_2024 = sum(`2024`, na.rm = TRUE),
    .groups = 'drop'
  )

print(tidy_data)

write_xlsx(tidy_data, "~/Downloads/state_statistics.xlsx")
```

7.3 Bar plot of GDP vs. number of Walmart store

```
library(readxl)
library(dplyr)
library(ggplot2)
library(scales)

gdp_data <- read_excel("desktop/state_statistics.xlsx") %>%
  select(State = StateAbbr, GDP = GDP_In_Millions_2024)

walmart_data <- read_excel("desktop/walmart_store_distribution.xlsx") %>%
  select(State = state, total_stores)

combined_data <- inner_join(gdp_data, walmart_data, by = "State")
```

```
combined_data <- combined_data %>%
  mutate(GDP_per_store = GDP / total_stores)

combined_data <- combined_data %>%
  arrange(desc(GDP_per_store))

ggplot(combined_data, aes(x = reorder(State, -GDP_per_store))) +
  geom_bar(aes(y = GDP_per_store), stat = "identity", fill = "#1f77b4", width = 0.7) +
  geom_bar(aes(y = total_stores * max(GDP_per_store) / max(total_stores)),
           stat = "identity", fill = "#ff7f0e", alpha = 0.7, width = 0.5) +
  scale_y_continuous(
    name = "GDP per Store (Millions USD)",
    sec.axis = sec_axis(~ . * max(combined_data$total_stores) / max(combined_data$GDP_per_st
                        name = "Number of Walmart Stores")
  ) +
  labs(title = "State GDP vs. Walmart Store Distribution",
       x = "State (Ordered by GDP per Store)",
       caption = "Blue bars: GDP per store (left axis)\nOrange bars: Number of stores (right
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 8),
        plot.title = element_text(hjust = 0.5, face = "bold"),
        axis.title.y = element_text(color = "#1f77b4"),
        axis.title.y.right = element_text(color = "#ff7f0e"),
        panel.grid.major.x = element_blank(),
        panel.grid.minor = element_blank())
```

7.4 Scattered plot of GDP vs. number of Walmart store

```
library(readxl)
library(dplyr)
library(ggplot2)
library(scales)

walmart <- read_excel("desktop/walmart_store_distribution.xlsx")
state_stats <- read_excel("desktop/state_statistics.xlsx")

merged_data <- inner_join(walmart, state_stats, by = c("state" = "StateAbbr")) %>%
  select(GeoName, state, total_stores, GDP_In_Millions_2024)

Walplot <- ggplot(merged_data, aes(x = total_stores, y = GDP_In_Millions_2024, label = state
```

```
  geom_point(color = "blue", size = 3) +
  geom_label(vjust = -0.5, size = 3, fill = "white", label.size = 0.2) +
  labs(title = "State GDP vs. Walmart Store Distribution",
       subtitle = "(* Denotes every value eligible is millions of millions of dollars)",
       x = "Total Stores in One State",
       y = "GDP per Store* (Millions USD)") +
  scale_y_continuous(limits = c(40000, 4200000), labels = comma) +
  theme_minimal()

print(Walplot)
```

7.5 A correlation test

```
library(readxl)
library(dplyr)

gdp_data <- read_excel("desktop/state_statistics.xlsx") %>%
  select(state = StateAbbr, gdp = GDP_In_Millions_2024)

store_data <- read_excel("desktop/walmart_store_distribution.xlsx") %>%
  select(state, stores = total_stores)

combined_data <- inner_join(gdp_data, store_data, by = "state") %>%
  filter(!state %in% c("PR", "DC"))

pearson_test <- cor.test(combined_data$gdp, combined_data$stores,
                         method = "pearson")

spearman_test <- cor.test(combined_data$gdp, combined_data$stores,
                          method = "spearman")

cor_matrix <- cor(combined_data[, c("gdp", "stores")])

results <- list(
  pearson = list(
    estimate = pearson_test$estimate,
    p_value = pearson_test$p.value,
    conf_int = pearson_test$conf.int
  ),
  spearman = list(
```

```
    estimate = spearman_test$estimate,
    p_value = spearman_test$p.value
  ),
  correlation_matrix = cor_matrix
)

results
```