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

Qianhui Dai, Joseph Easterday

2025-05-02

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. This analysis operates within the paradigm of economic geography, which assumes that patterns in commercial infrastructure—such as store presence—can be understood and predicted through macroeconomic indicators like GDP. We begin with the perspective that retail distribution is influenced by economic output but may also be shaped by additional factors such as population density, land use, and regional strategy.

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

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

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

We structured our data analysis process to align with the FAIR (Findable, Accessible, Interoperable, and Reusable) and CARE (Collective benefit, Authority to control, Responsibility, and Ethics) principles, while also reflecting on the challenges of fully meeting these ideals.

1.Ensuring FAIR Data Practices Findability & Accessibility: We used openly available datasets: Walmart's store location data from the company's open data portal, and state GDP data from the U.S. Bureau of Economic Analysis. Both datasets are publicly hosted by credible institutions, ensuring long-term accessibility and proper metadata documentation.

Cleaning & Structuring for Interoperability: To improve machine and human readability, we:

Removed redundant header rows and renamed unclear or inconsistent columns.

Standardized state names and abbreviations to support merging across datasets.

Converted numeric columns from text to proper numeric format.

Merging for Interoperability: We integrated the Walmart and GDP datasets using standardized state codes, which allowed us to explore cross-dataset relationships. We also preserved geographic identifiers (e.g., state-level names and coordinates) to support downstream visualization and mapping tasks. A challenge we encountered was ensuring that regional codes (e.g., state abbreviations) matched exactly, which required extra cleaning to avoid merge failures.

Reusability through Transparent Workflow: Our entire workflow—from data import to visualization—was built in a Quarto (.qmd) file that can be reproduced, shared, and extended by others. We also used well-supported packages (e.g., dplyr, ggplot2) and avoided hardcoded paths or non-portable dependencies, supporting long-term usability. One limitation is that external users would still need to manually download the source files from their respective websites.

2.Addressing CARE Principles Collective Benefit: Our analysis aims to surface insights that can support public understanding of economic access and inform retail decision-making. By mapping Walmart presence against economic indicators, we contribute a view that may benefit researchers, policymakers, or underserved communities.

Authority to Control: We exclusively used datasets that are open and licensed for public use by the organizations that collected them. This ensures that data ownership and governance are respected, particularly with regard to how commercial data (e.g., store locations) is used.

Responsibility & Ethics: We were careful to avoid misrepresenting preliminary data or drawing conclusions unsupported by the analysis. Our dataset included no personal or sensitive information. However, a challenge in applying CARE is ensuring that interpretations do not reinforce bias or overlook socioeconomic factors not captured in the data (e.g., income inequality, rural access).

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.

```
## Complete Walmart Store Distribution by State
```

Table 1: Sorted by number of stores (descending)

State Name	State Abbr	Number of Stores
Texas	TX	595
Florida	FL	386
California	CA	308
Georgia	GA	214
North Carolina	NC	214
Illinois	IL	177
Ohio	OH	172
Pennsylvania	PA	159
Missouri	MO	156
Tennessee	TN	150
Virginia	VA	149
Alabama	AL	144
Louisiana	LA	139
Oklahoma	OK	134
Arkansas	AR	130
Indiana	IN	126
Arizona	AZ	124
South Carolina	SC	122
Michigan	MI	117
New York	NY	111
Colorado	CO	105

State Name	State Abbr	Number of Stores
Kentucky	KY	101
Wisconsin	WI	98
Mississippi	MS	86
Kansas	KS	83
Minnesota	MN	79
New Jersey	NJ	70
Iowa	IA	69
Washington	WA	64
Maryland	MD	59
Utah	UT	59
New Mexico	NM	52
Nevada	NV	50
Massachusetts	MA	48
Nebraska	NE	47
West Virginia	WV	44
Oregon	OR	43
Connecticut	CT	33
New Hampshire	NH	28
Idaho	ID	27
NA	PR	26
Maine	ME	25
North Dakota	ND	17
South Dakota	SD	17
Montana	MT	16
Wyoming	WY	14
Hawaii	HI	11
Delaware	DE	10
Alaska	AK	9
Rhode Island	RI	9
Vermont	VT	6
NA	DC	2

```
## Descriptive Statistics
```

Table 2: Summary statistics of Walmart stores distribution

Statistic	Value
Mean	100.7

Statistic	Value
Median	74.5
Std_Dev	104.4
Minimum	2
Maximum	595
Total_Stores	5234
Number_of_States	52

State GDP Data for 2024 (in Millions)

Table 3: Sorted by GDP (descending order)

State Name	State Abbr	GDP (Millions)
California	CA	4103123.6
Texas	TX	2709392.9
New York	NY	2297028.0
Florida	FL	1705564.9
Illinois	IL	1137243.6
Pennsylvania	PA	1024206.3
Ohio	OH	927740.1
Georgia	GA	882534.5
Washington	WA	854683.3
New Jersey	NJ	846587.5
North Carolina	NC	839122.2
Massachusetts	MA	780666.2
Virginia	VA	764474.8
Michigan	MI	706615.8
Colorado	CO	553322.5
Arizona	AZ	552167.0
Tennessee	TN	549708.5
Maryland	MD	542765.8
Indiana	IN	527381.1
Minnesota	MN	500851.4
Wisconsin	WI	451285.3
Missouri	MO	451201.4
Connecticut	CT	365723.1
South Carolina	SC	349965.4
Oregon	OR	331028.6
Louisiana	LA	327782.1
Alabama	AL	321237.6

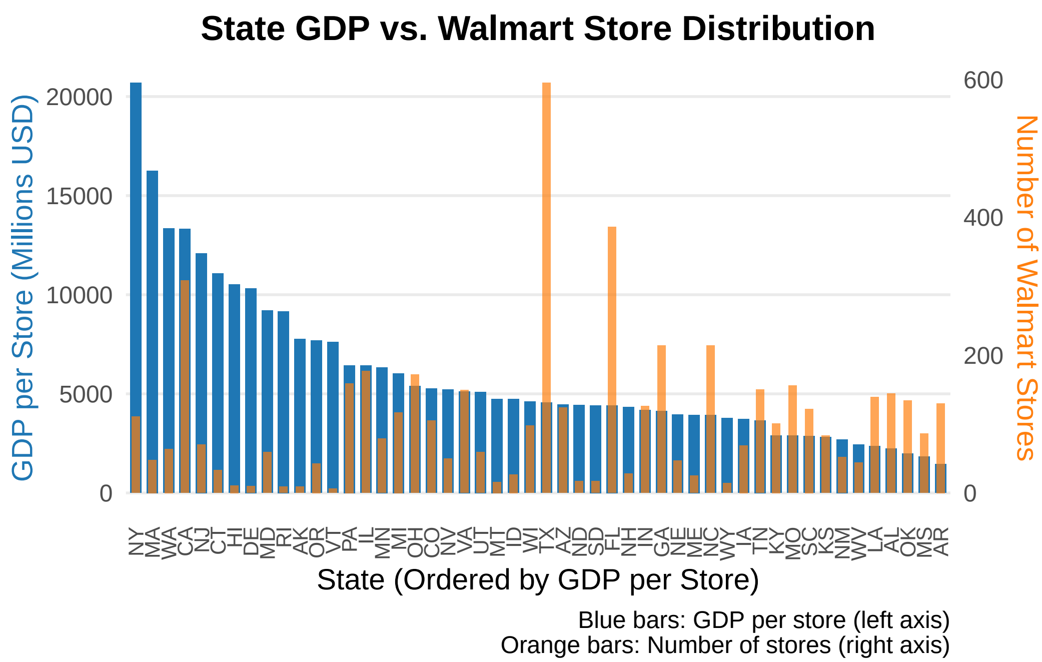
State Name	State Abbr	GDP (Millions)
Utah	UT	300903.8
Kentucky	KY	293021.0
Oklahoma	OK	265779.1
Nevada	NV	260728.4
Iowa	IA	257020.8
Kansas	KS	234673.2
Arkansas	AR	188723.1
Nebraska	NE	185411.0
Mississippi	MS	157491.0
New Mexico	NM	140542.0
Idaho	ID	128132.1
New Hampshire	NH	121189.4
Hawaii	HI	115627.2
West Virginia	WV	107660.1
Delaware	DE	103253.3
Maine	ME	98606.0
Rhode Island	RI	82492.5
Montana	MT	75999.2
North Dakota	ND	75399.4
South Dakota	SD	75179.5
Alaska	AK	69969.0
Wyoming	WY	52946.1
Vermont	VT	45707.2
NA	NA	0.0

```
## GDP Descriptive Statistics
```

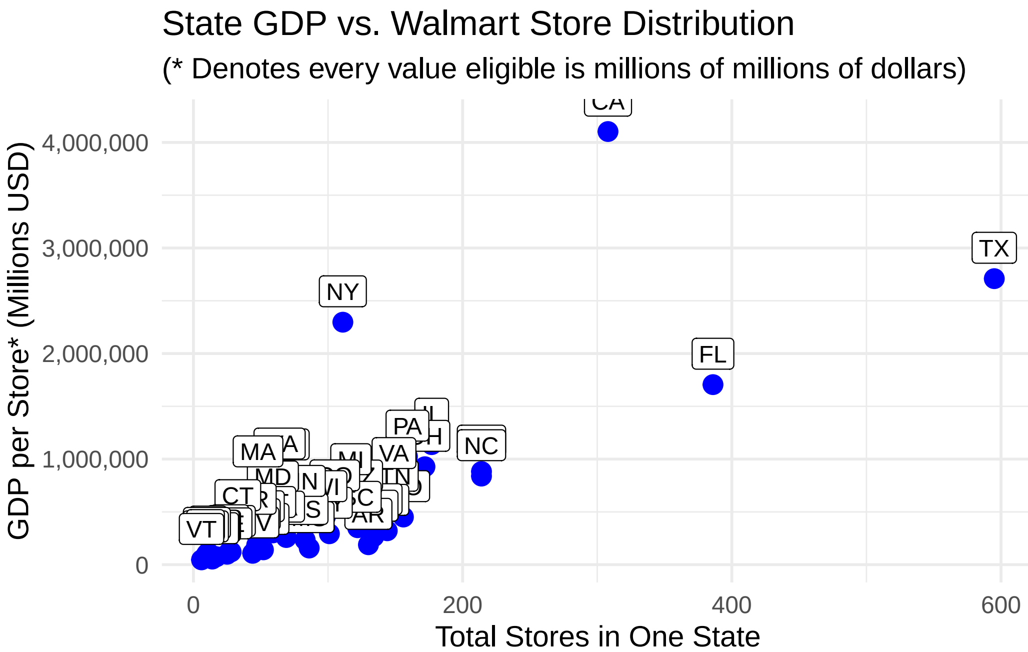
Table 4: Summary statistics of 2024 State GDP Data

Statistic	Value (Millions)
Mean_GDP	565487.4
Median_GDP	327782.1
Std_Dev	738388.5
Minimum	0.0
Maximum	4103123.6
Total_US_GDP	28839857.9
Number_of_States	51

The bar plot uses dual-axis encoding to simultaneously visualize GDP per Walmart store (left axis, blue bars) and total number of Walmart stores per state (right axis, orange bars). This visual comparison immediately draws attention to states like New York (NY), Massachusetts (MA), and California (CA), which have extremely high GDP per store, but relatively few Walmart locations. This discrepancy suggests that GDP alone does not determine store presence and invites interpretation around real estate prices, market saturation, or urban planning constraints. In contrast, states like Texas (TX) and Florida (FL) combine high total GDP with high store counts, reinforcing the hypothesis that strong economies in large, low-density states are correlated with high retail penetration.



The scatter plot reinforces this trend by spatially encoding the relationship between total store count and GDP per store. The clustering of most states in the lower-left quadrant illustrates that many have modest GDP per store and moderate Walmart presence. Outliers such as California (CA) and New York (NY) appear far above the trend, confirming earlier observations and suggesting a strategic restraint in these markets. The plot also reveals a broad positive correlation (supported by a Pearson coefficient of 0.75, calculated separately), but the dispersion and labeled anomalies highlight that other variables — such as population density, zoning restrictions, or consumer demographics — may drive deviations from the expected trend.



Together, these two visualizations use comparison, hierarchy, and emphasis to guide interpretation and suggest that GDP is a strong, but not exclusive, factor in Walmart's store distribution. Future exploration could test hypotheses involving population size, geographic spread, or cost of commercial space to explain the residual variance.

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. Supporting this, the Spearman correlation coefficient is even higher at 0.81, reinforcing a strong monotonic relationship between the two variables—even when the relationship may not be perfectly linear. Both correlations are statistically significant, with p-values near zero, and the 95% confidence interval for the Pearson correlation (0.60 to 0.85) confirms the robustness of this positive association. These findings suggest that economic activity is a meaningful driver of store distribution across states.

```
$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 official state GDP statistics, we aimed to explore whether economic strength correlates with retail presence, following the economic geography paradigm that commercial development reflects broader economic conditions.

Our visualizations — including a bar chart comparing store count and GDP by state, a scatter plot with labeled states, and a correlation matrix — revealed a consistent pattern: states with higher GDPs generally have more Walmart locations. Most notably, our correlation analysis returned a Pearson coefficient of "0.75", indicating a strong positive linear relationship between GDP and store count. This supports our initial perspective that economic output is a key driver of store distribution, suggesting that Walmart tends to concentrate stores in economically stronger states due to greater consumer demand, infrastructure capacity, and business opportunity.

However, our perspective evolved through the analysis. GDP alone does not fully explain store distribution. States like "New York" and "New Jersey", despite their large economies, have fewer Walmart stores than expected — potentially due to high real estate costs, urban density, zoning restrictions, or strategic business decisions. These outliers prompted us to reconsider the limitations of using GDP as a standalone predictor. Variables such as "population size, land availability, income level, or urbanization" may be equally or more important in certain regions.

Overall, this analysis demonstrates how data visualization and correlation methods can yield valuable insights for business strategy, economic development, and retail planning. The findings may assist Walmart and similar corporations in identifying underserved markets, prioritizing expansion efforts, or aligning infrastructure investment with economic indicators. Additionally, this approach can inform policymakers interested in equitable commercial access and evidence-based economic planning. Our perspective is that further research incorporating additional variables would offer a more nuanced and accurate model of retail distribution patterns in the U.S.

6 Contributor Roles and Documentation of Sources

6.1 Authors' Contributions

This project was a collaborative effort: - Qianhui Dai cleaned and visualized the Walmart store distribution data, created the bar plot and performed the correlation analysis, and wrote the full reproducible Quarto report. - Joseph Easterday collected and cleaned the state GDP dataset did the presentation and created the labeled scatter plot visualization.

6.2 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/explore.

7 Code Appendix

7.1 Store distribution table

```
library(dplyr)
library(readxl)
library(knitr)

# STEP 1: Read in Walmart store distribution data
```

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

# STEP 2: Create a lookup table to match state abbreviations with full names
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"
  )
)

# STEP 3: Clean and join data with state names
walmart_tidy <- walmart_raw %>%
  # Rename for clarity
  rename(Stores = total_stores) %>%
  # Add full state names
  left_join(state_abbr_df, by = c("state" = "StateAbbr")) %>%
  select(StateName, StateAbbr = state, Stores) %>%
  # Sort from most to fewest stores
  arrange(desc(Stores))

# STEP 4: Display the cleaned and sorted data as a formatted table
cat("## Complete Walmart Store Distribution by State\n\n")
```

```r
kable(walmart_tidy,
      align = c('l', 'c', 'r'),
      col.names = c("State Name", "State Abbr", "Number of Stores"),
      caption = "Sorted by number of stores (descending)")

# STEP 5: Generate descriptive statistics for the distribution
desc_stats <- walmart_tidy %>%
  summarize(
    # average number of stores
    Mean = sprintf("%.1f", mean(Stores, na.rm = TRUE)),
    # middle value
    Median = median(Stores, na.rm = TRUE),
    # standard deviation
    Std_Dev = sprintf("%.1f", sd(Stores, na.rm = TRUE)),
    # fewest stores
    Minimum = min(Stores, na.rm = TRUE),
    # most stores
    Maximum = max(Stores, na.rm = TRUE),
    # total across all states
    Total_Stores = sum(Stores, na.rm = TRUE),
    # number of states in the dataset
    Number_of_States = n()
  ) %>%
  t() %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Statistic") %>%
  rename(Value = V1)

# STEP 6: Display the summary statistics as a table
cat("\n## Descriptive Statistics\n\n")
kable(desc_stats,
      align = c('l', 'r'),
      caption = "Summary statistics of Walmart stores distribution",
      col.names = c("Statistic", "Value"))
```

7.2 State GDP table

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

```
# STEP 1: Read in raw GDP Excel file (skip metadata rows)
raw_data <- read_excel("Download/Gross_Domestic_Product.xlsx", skip = 5)

# STEP 2: Create a state abbreviation lookup table
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 = state.abb)

# STEP 3: Clean, filter, and join state data
tidy_data <- raw_data %>%
  # Select only relevant columns
  select(GeoFips, GeoName, `2024`) %>%
  # Remove non-state entries
  filter(!GeoName %in% c("District of Columbia", "United States")) %>%
  # Add abbreviations
  left_join(state_abbr_df, by = "GeoName") %>%
  group_by(GeoName, StateAbbr) %>%
  summarise(GDP_2024_Millions = sum(`2024`, na.rm = TRUE), .groups='drop')%>%
  # Sort by GDP
  arrange(desc(GDP_2024_Millions))

# STEP 4: Output clean table
cat("## State GDP Data for 2024 (in Millions)\n\n")
kable(tidy_data,
      align = c('l', 'c', 'r'),
      col.names = c("State Name", "State Abbr", "GDP (Millions)"),
      caption = "Sorted by GDP (descending order)")

# STEP 5: Compute and display summary statistics
desc_stats <- tidy_data %>%
  summarize(
    Mean_GDP = sprintf("%.1f", mean(GDP_2024_Millions, na.rm = TRUE)),
```

```r
    Median_GDP = sprintf("%.1f", median(GDP_2024_Millions, na.rm = TRUE)),
    Std_Dev = sprintf("%.1f", sd(GDP_2024_Millions, na.rm = TRUE)),
    Minimum = sprintf("%.1f", min(GDP_2024_Millions, na.rm = TRUE)),
    Maximum = sprintf("%.1f", max(GDP_2024_Millions, na.rm = TRUE)),
    Total_US_GDP = sprintf("%.1f", sum(GDP_2024_Millions, na.rm = TRUE)),
    Number_of_States = n()
  ) %>%
  t() %>%
  as.data.frame() %>%
  tibble::rownames_to_column("Statistic") %>%
  rename(Value = V1)

cat("\n## GDP Descriptive Statistics\n\n")
kable(desc_stats,
      align = c('l', 'r'),
      caption = "Summary statistics of 2024 State GDP Data",
      col.names = c("Statistic", "Value (Millions)"))
```

7.3 Bar plot of GDP vs. number of Walmart store

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

# STEP 1: Load cleaned GDP and Walmart data
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)

# STEP 2: Merge datasets and calculate GDP per store
combined_data <- inner_join(gdp_data, walmart_data, by = "State") %>%
  mutate(GDP_per_store = GDP / total_stores) %>%
  arrange(desc(GDP_per_store))

# STEP 3: Create bar plot with dual y-axes
ggplot(combined_data, aes(x = reorder(State, -GDP_per_store))) +
  # Left axis
```

```
  geom_bar(aes(y = GDP_per_store), stat = "identity", fill = "#1f77b4", width=
  0.7) +
  # Right axis
  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_store),
                        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 axis)") +
  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"))
```

7.4 Scatter plot of GDP vs. number of Walmart store

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

# STEP 1: Load datasets and merge on state abbreviation
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)

# STEP 2: Plot GDP vs. store count with labeled states
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 = "(* All values in millions of USD)",
       x = "Total Stores in One State",
       y = "GDP (Millions USD)") +
  scale_y_continuous(limits = c(40000, 4200000), labels = comma) +
  theme_minimal()

print(Walplot)
```

7.5 A correlation test

```
library(readxl)
library(dplyr)

# STEP 1: Load cleaned datasets
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)

# STEP 2: Merge and filter out non-state regions
combined_data <- inner_join(gdp_data, store_data, by = "state") %>%
  filter(!state %in% c("PR", "DC"))  # Remove territories

# STEP 3: Run correlation tests
pearson_test <- cor.test(combined_data$gdp, combined_data$stores, method =
"pearson")
spearman_test <- cor.test(combined_data$gdp, combined_data$stores, method =
"spearman")

# STEP 4: Output both correlation estimates and full matrix
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
```