Data Visualization

2025-05-05

Introduction

Obesity remains one of the most pressing public health challenges in the United States, contributing to increased risks of chronic diseases such as diabetes, heart disease, and certain cancers. Understanding the lifestyle behaviors that influence obesity—such as physical activity levels and dietary habits—is essential for developing effective health policies and interventions.

This project explores patterns in nutrition, physical activity, and obesity using two complementary datasets. The first dataset, sourced from the CDC and hosted on Kaggle, provides state-level data on adult weight status and related health behaviors. The second dataset focuses specifically on the American Indian and Alaska Native population, offering insight into disparities in health outcomes across demographic groups.

Through a series of data visualizations, this project aims to tell a story about how lifestyle factors relate to obesity across regions and populations, and to highlight where public health efforts may be most needed.

Description of the Data

This project utilizes two datasets that provide insight into adult health behaviors and outcomes related to nutrition, physical activity, and obesity in the United States.

Dataset 1: Nutrition, Physical Activity, and Obesity – CDC (Kaggle)

This dataset, sourced from the Behavioral Risk Factor Surveillance System (BRFSS) and hosted on Kaggle, includes state-level data for various health indicators among adults. Key variables include:

- Percentage of adults with obesity
- Levels of physical inactivity
- Fruit and vegetable consumption
- Sugar-sweetened beverage intake
- State and regional identifiers

The dataset offers a comprehensive overview of health behaviors across all 50 U.S. states, enabling comparisons and trend analysis.

Dataset 2: 2023 American Indian/Alaska Native BRFSS Subset

This dataset is a subset of the BRFSS focusing specifically on American Indian and Alaska Native populations in 2023. It includes similar variables to the first dataset but is limited to a specific demographic group. This allows for the exploration of health disparities and targeted behavioral trends, which are often underrepresented in broader public health analyses.

Both datasets are in CSV format and are suitable for exploratory visualization and multivariate analysis.

```
# Load the datasets
dataset1 <- read_csv("Nutrition_Physical_Activity_Obesity.csv")
```

```
## Rows: 53392 Columns: 33
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (25): LocationAbbr, LocationDesc, Datasource, Class, Topic, Question, Da...
## dbl  (7): YearStart, YearEnd, Data_Value, Data_Value_Alt, Low_Confidence_Lim...
## lgl  (1): Data_Value_Unit
## 
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dataset2 <- read_csv("2023-AmericanIndian-AlaskaNative.csv")
```

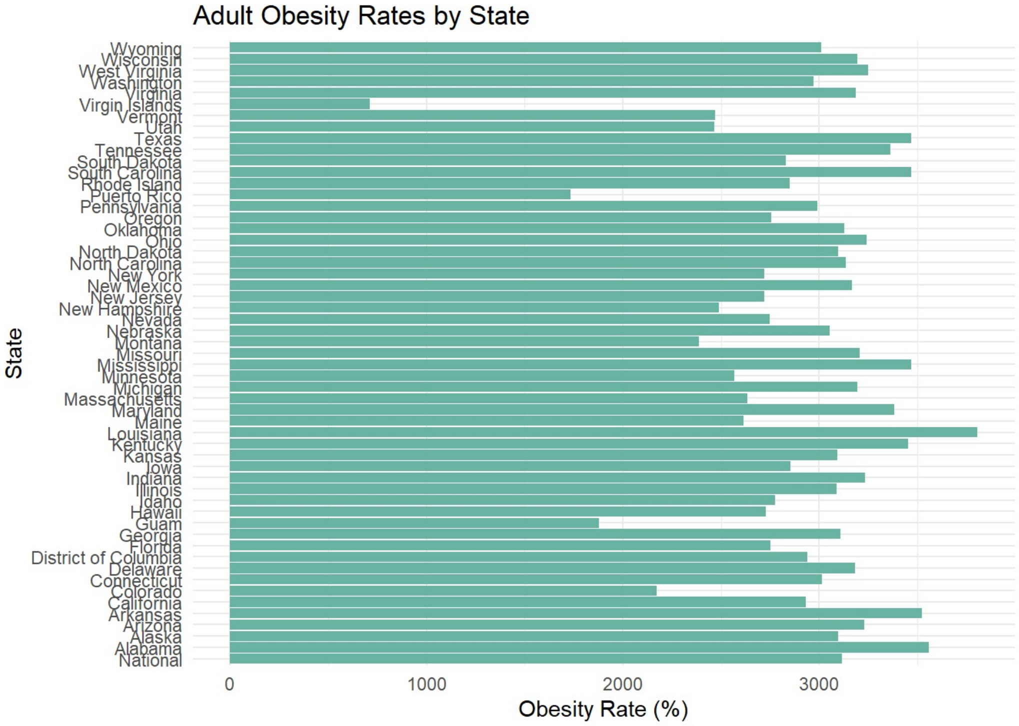
```
## Rows: 54 Columns: 3
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): State, Prevalence, 95% CI
## 
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Adult Obesity Rates by U.S. State

This bar chart shows the percentage of adults with obesity in each U.S. state. It helps illustrate geographic variation in obesity rates and highlights where public health efforts may be most needed.

```
# Filter the dataset for relevant rows
obesity_data <- dataset1 %>%
  filter(Question == "Percent of adults aged 18 years and older who have obesity") %>%
  select(LocationDesc, Data_Value) %>%
  distinct()

# Create the bar plot
ggplot(obesity_data, aes(x = reorder(LocationDesc, -Data_Value), y = Data_Value)) +
  geom_bar(stat = "identity", fill = "#69b3a2") +
  coord_flip() +
  labs(title = "Adult Obesity Rates by State",
       x = "State",
       y = "Obesity Rate (%)") +
  theme_minimal()
```



### Relationship Between Obesity and Physical Inactivity

This scatter plot shows the relationship between physical inactivity and obesity across U.S. states. States with higher physical inactivity tend to report higher obesity rates.

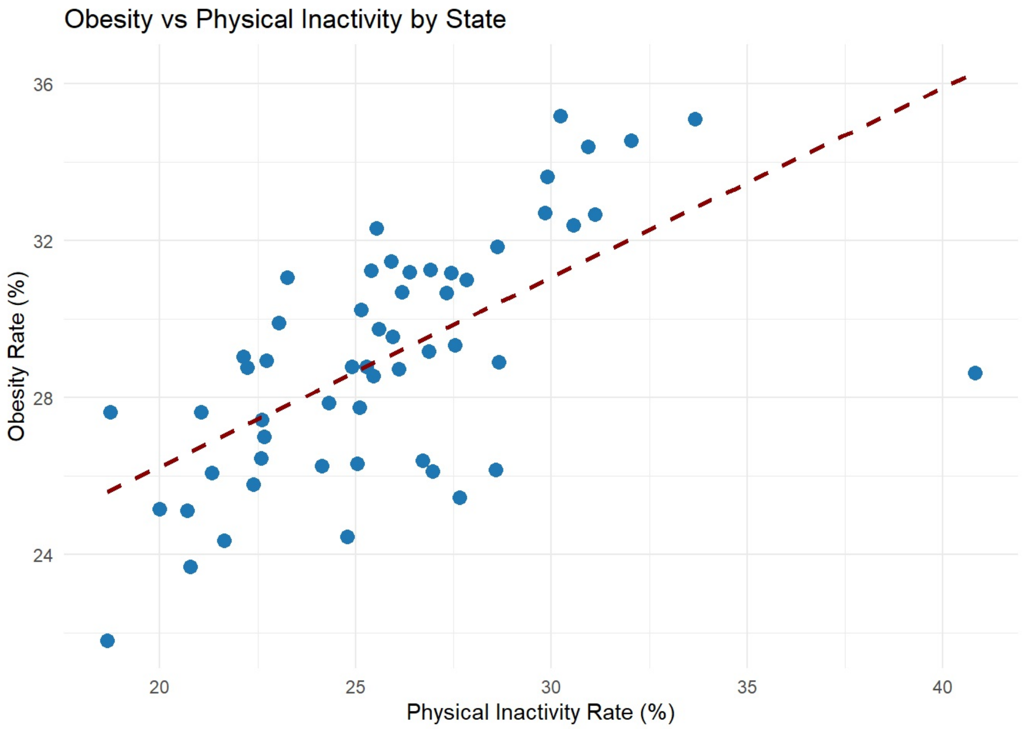
```
library(dplyr)
library(tidyr)
library(ggplot2)

# Summarize to get a single value per State-Question
obesity_vs_inactivity <- dataset1 %>%
  filter(Question %in% c("Percent of adults aged 18 years and older who have obesity",
                         "Percent of adults who engage in no leisure-time physical activity")) %>%
  group_by(LocationDesc, Question) %>%
  summarise(Data_Value = mean(Data_Value, na.rm = TRUE), .groups = "drop") %>%
  pivot_wider(names_from = Question, values_from = Data_Value)

# Rename columns
colnames(obesity_vs_inactivity) <- c("State", "Obesity", "Inactivity")

# Drop rows with missing values (just in case)
obesity_vs_inactivity <- obesity_vs_inactivity %>%
  filter(!is.na(Obesity), !is.na(Inactivity))

# Plot
ggplot(obesity_vs_inactivity, aes(x = Inactivity, y = Obesity)) +
  geom_point(color = "#1f77b4", size = 3) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "darkred") +
  labs(title = "Obesity vs Physical Inactivity by State",
       x = "Physical Inactivity Rate (%)",
       y = "Obesity Rate (%)") +
  theme_minimal()
```



Relationship Between Obesity and Low Fruit Consumption

This scatter plot shows the relationship between low fruit consumption and obesity across U.S. states. States where more adults report eating fruit less than once daily tend to have higher obesity rates, suggesting a potential dietary link to obesity prevalence.

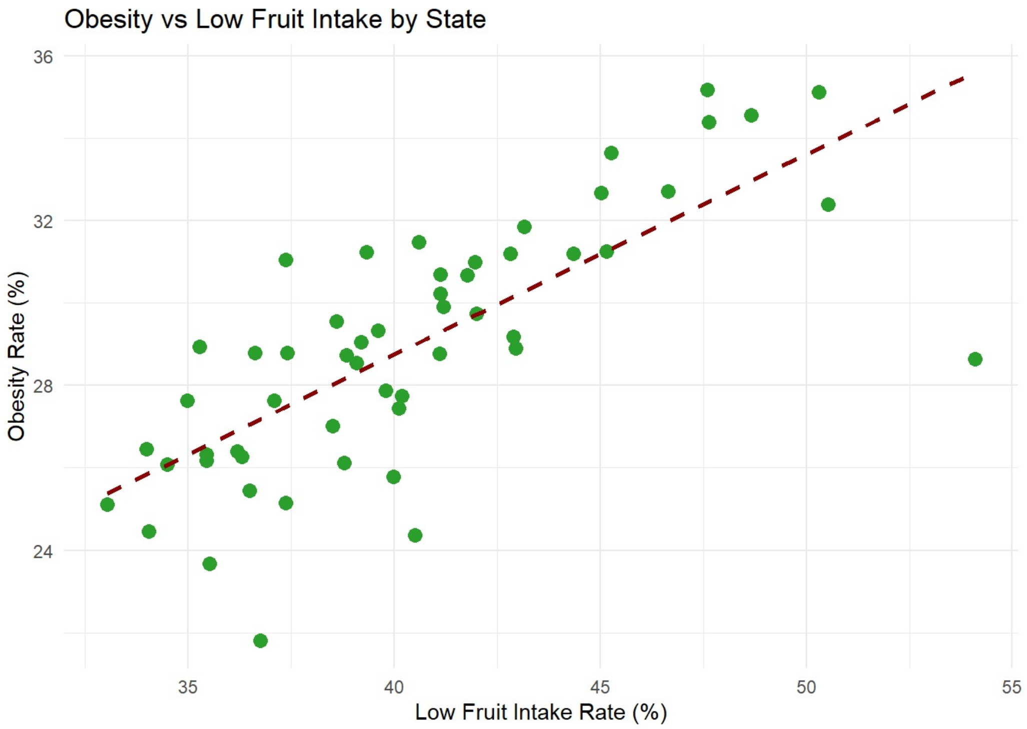
```r
library(dplyr)
library(tidyr)
library(ggplot2)

# Summarize to get a single value per State-Question
obesity_vs_fruit <- dataset1 %>%
  filter(Question %in% c("Percent of adults aged 18 years and older who have obesity",
                         "Percent of adults who report consuming fruit less than one time daily")) %>%
  group_by(LocationDesc, Question) %>%
  summarise(Data_Value = mean(Data_Value, na.rm = TRUE), .groups = "drop") %>%
  pivot_wider(names_from = Question, values_from = Data_Value)

# Rename columns
colnames(obesity_vs_fruit) <- c("State", "Obesity", "LowFruitIntake")

# Drop rows with missing values
obesity_vs_fruit <- obesity_vs_fruit %>%
  filter(!is.na(Obesity), !is.na(LowFruitIntake))

# Plot
ggplot(obesity_vs_fruit, aes(x = LowFruitIntake, y = Obesity)) +
  geom_point(color = "#2ca02c", size = 3) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "darkred") +
  labs(title = "Obesity vs Low Fruit Intake by State",
       x = "Low Fruit Intake Rate (%)",
       y = "Obesity Rate (%)") +
  theme_minimal()
```



Relationship Between Obesity, Physical Inactivity, and Low Fruit Consumption

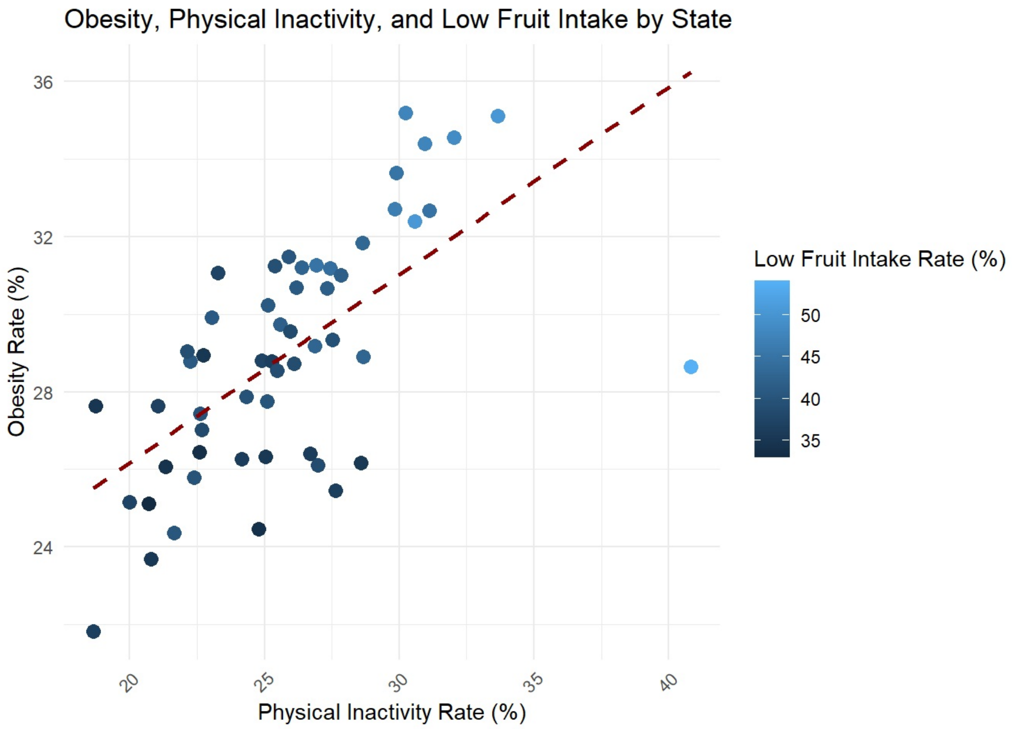
This scatter plot illustrates the relationship between obesity, physical inactivity, and low fruit consumption across U.S. states. States with higher physical inactivity and lower fruit consumption tend to report higher obesity rates, suggesting that both sedentary lifestyles and poor dietary habits contribute to higher obesity prevalence. The color represents the percentage of adults with low fruit intake, while the size of the points reflects the rate of physical inactivity, providing a clear view of how these three factors interact.

```
# Summarize to get a single value per State-Question
obesity_inactivity_fruit <- dataset1 %>%
  filter(Question %in% c("Percent of adults aged 18 years and older who have obesity",
                         "Percent of adults who engage in no leisure-time physical activity",
                         "Percent of adults who report consuming fruit less than one time daily")) %>%
  group_by(LocationDesc, Question) %>%
  summarise(Data_Value = mean(Data_Value, na.rm = TRUE), .groups = "drop") %>%
  pivot_wider(names_from = Question, values_from = Data_Value)

# Rename columns
colnames(obesity_inactivity_fruit) <- c("State", "Obesity", "Inactivity", "LowFruitIntake")

# Drop rows with missing values
obesity_inactivity_fruit <- obesity_inactivity_fruit %>%
  filter(!is.na(Obesity), !is.na(Inactivity), !is.na(LowFruitIntake))

# Plot: 3-variable relationship
ggplot(obesity_inactivity_fruit, aes(x = Inactivity, y = Obesity, color = LowFruitIntake)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "darkred") +
  labs(title = "Obesity, Physical Inactivity, and Low Fruit Intake by State",
       x = "Physical Inactivity Rate (%)",
       y = "Obesity Rate (%)",
       color = "Low Fruit Intake Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Details of Visualization Choices

In this project, a consistent minimalist style was used across all visualizations to promote clarity and maintain visual coherence. The purpose of each visualization was to convey specific relationships between health behaviors and obesity prevalence across U.S. states, and design decisions were made to highlight these relationships effectively and accessibly.

1. Adult Obesity Rates by State (Bar Chart)

A horizontal bar chart was chosen to visualize adult obesity rates by state because it is ideal for comparing a single quantitative variable across multiple categorical groups (in this case, states). By ordering the bars from highest to lowest obesity rates, the visualization emphasizes regional disparities and draws attention to states with the most pressing public health challenges.

- Orientation: Horizontal orientation allows longer state names to be displayed clearly and makes the chart easier to scan visually.
- Color: The bars are filled with a soft teal green color (#69b3a2), chosen for its neutral, non-distracting tone that maintains focus on the values.
- Theme: A minimal theme ensures readability and eliminates unnecessary visual clutter.
- Axis Labels and Title: Clear and concise labeling was used to ensure viewers immediately understand what the chart represents.

This chart serves as a foundational visualization, setting the stage for deeper analyses of contributing behavioral factors.

2. Obesity vs Physical Inactivity (Scatter Plot)

To explore the relationship between physical inactivity and obesity, a scatter plot was used. This type of plot is well-suited for showing correlations between two continuous variables and helps identify trends and outliers across states.

- X and Y Variables: The x-axis shows the percentage of adults reporting no leisure-time physical activity, while the y-axis shows obesity rates. This directly examines the hypothesis that inactivity is linked to obesity.
- Color and Points: Data points are plotted in a clean blue color (#1f77b4) to stand out without overwhelming the viewer.
- Trend Line: A dashed dark red linear regression line was included to highlight the general direction of the relationship. The use of a contrasting color for the line draws attention to the trend without being overly bold.
- Theme: The minimalist theme keeps the focus on the relationship between the two variables.

This visualization helps support the narrative that physical inactivity is positively associated with higher obesity prevalence, emphasizing the importance of promoting active lifestyles.

3. Obesity vs Low Fruit Consumption (Scatter Plot)

This scatter plot investigates whether there is an association between low fruit intake and obesity rates. Again, a scatter plot was used for its ability to represent relationships between two continuous variables.

- X and Y Variables: The x-axis shows the percentage of adults consuming fruit less than once daily, while the y-axis shows the obesity rate.
- Color and Points: Points are colored green (#2ca02c) to symbolically align with the concept of fruit/vegetables and healthy eating.
- Trend Line: A dashed dark red regression line is included to indicate the trend, helping viewers identify a potential linear relationship.
- Design Choices: The use of subtle color and minimal design helps viewers concentrate on the key message without distraction.

This plot visually reinforces the idea that poor dietary habits—specifically low fruit consumption—may contribute to higher obesity rates, providing support for nutritional interventions.

4. Obesity, Inactivity, and Low Fruit Intake (Three-Variable Scatter Plot)

To simultaneously explore the relationships among obesity, physical inactivity, and low fruit intake, a three-variable scatter plot was created. This is one of the most informative visualizations in the project, showing how multiple behavioral factors interplay.

- X and Y Variables: The x-axis represents low fruit intake, and the y-axis represents obesity rates—consistent with the earlier bivariate plot

for continuity.
- Size of Points: The size of each point reflects physical inactivity levels. Larger points indicate higher inactivity, adding a third variable visually without overloading the graph.
- Color Gradient: A color scale is used to further emphasize the values of fruit intake, helping differentiate states more clearly and visually link the severity of low intake to obesity.
- Regression Line: A linear trend line was included to guide interpretation of the overall relationship.
- Consistency: The minimalist design, similar axis labels, and calm visual palette were retained to match previous charts.

This plot brings together multiple behavioral risk factors and presents them in one integrated visualization, supporting the idea that multiple lifestyle behaviors contribute collectively to obesity trends.

Consistency Across Visualizations

All visualizations maintain the same minimalist aesthetic, font styles, and axis formatting to ensure consistency. A limited color palette was used across the project—teal green, blue, and green tones for individual behaviors, and dark red for regression lines—chosen for their visual harmony and symbolic relevance (e.g., green for healthy eating, blue for inactivity). This consistency helps reinforce the story being told without requiring the viewer to constantly adjust to new styles.

Each visualization was crafted to highlight key findings, maintain readability, and allow for intuitive interpretation by the viewer.

References

Data Sources Centers for Disease Control and Prevention. (n.d.). Behavioral Risk Factor Surveillance System (BRFSS). U.S. Department of Health and Human Services. Retrieved from https://www.cdc.gov/brfss/index.html (https://www.cdc.gov/brfss/index.html)

Centers for Disease Control and Prevention. (n.d.). Nutrition, Physical Activity, and Obesity – BRFSS Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/cdc/nutrition-physical-activity-and-obesity (https://www.kaggle.com/datasets/cdc/nutrition-physical-activity-and-obesity)

Indian Health Service. (n.d.). 2023 American Indian and Alaska Native BRFSS Data Subset. U.S. Department of Health and Human Services. Retrieved from https://www.ihs.gov (https://www.ihs.gov)

R Packages Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., … Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686 (https://doi.org/10.21105/joss.01686)

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag. https://ggplot2.tidyverse.org/ (https://ggplot2.tidyverse.org/)

Wickham, H., & Grolemund, G. (2017). R for data science. O'Reilly Media. https://r4ds.had.co.nz/ (https://r4ds.had.co.nz/)

Coding References R Graph Gallery. (n.d.). R graph examples using ggplot2. Retrieved from https://r-graph-gallery.com (https://r-graph-gallery.com)

Wickham, H. (n.d.). ggplot2 reference manual. Tidyverse. Retrieved from https://ggplot2.tidyverse.org/reference/ (https://ggplot2.tidyverse.org/reference/)

Stack Overflow. (n.d.). Questions tagged [ggplot2]. Retrieved from https://stackoverflow.com/questions/tagged/ggplot2 (https://stackoverflow.com/questions/tagged/ggplot2)

RStudio. (n.d.). Cheat sheets. Retrieved from https://www.rstudio.com/resources/cheatsheets/ (https://www.rstudio.com/resources/cheatsheets/)