Obesity Report

Set-up your environment

Title Page

Obesity

Group: Tutorial 4

Tutorial lecturer: Chantal

Lecturer: Jack Frederick Fitzgerald

Eren Murtezaoglu (2819813) Renata Mutesah (2797482) Jada van Heyningen (2801118) Matilde Maio (2861974) Demir Günbahar (2801828) Duru Ozkan (2757656) Pratyush Sebastian (2789372)

Part 1 - Identify a Social Problem

Obesity is a growing social problem with significant public health, economic, and societal implications. With many factors impacting the rate of obesity. Some including poor diets, mental health issues, low income, and lack of education. In the united states of America, more than 100 millions adults (CDC, 2024) have obesity, causing the USA to rank in the top 10 (Ranking (% obesity by country), n.d.) countries with the highest obesity percentage. It contributes to rising rates of chronic diseases like diabetes, heart disease, and certain cancers, placing immense pressure on healthcare systems. In many communities, limited access to healthy food and safe recreational spaces further deepens health inequalities.
Exploring the correlation between education, income, and obesity is highly relevant because both factors significantly influence health behaviors and access to resources. Individuals with lower levels of education may have limited knowledge about nutrition and the long-term consequences of poor dietary habits. Similarly, those with lower income often face barriers such as food insecurity, reliance on cheap, calorie-dense foods, and reduced access to recreational facilities or healthcare. Especially since we live in a world where eating healthy costs twice as much as eating junk food (Barraclough, 2025) It is assumed that these socioeconomic disadvantages contribute to higher obesity rates in disadvantaged populations- we were interested to explore these rates, specifically in American states.

Part 2 - Data Sourcing

2.1 Load in the data

2.2 Provide a short summary of the dataset(s)

```
head(income2019)
```

```
## # A tibble: 6 x 4
##    ...1 GeoFips GeoName           X2019
##   <dbl> <chr>   <chr>             <dbl>
## 1     1 00000   United States 18349584
## 2     2 01000   Alabama         215152.
## 3     3 02000   Alaska           44460.
## 4     4 04000   Arizona         337257.
## 5     5 05000   Arkansas        131395.
## 6     6 06000   California     2539747.
```

```
head(education2019)
```

```
## # A tibble: 6 x 626
##    ...1 Label..Grouping.           Alabama..Total..Esti~1 Alabama..Total..Marg~2
##   <dbl> <chr>                      <chr>                  <chr>
## 1     1 AGE BY EDUCATIONAL ATTAIN~ <NA>                   <NA>
## 2     2     Population 18 to 24 y~ 457,530                ±5,551
## 3     3         Less than high sc~ 56,454                 ±4,838
## 4     4         High school gradu~ 158,761                ±7,166
## 5     5         Some college or a~ 207,319                ±7,341
## 6     6         Bachelor's degree~ 34,996                 ±3,496
## # i abbreviated names: 1: Alabama..Total..Estimate,
## #   2: Alabama..Total..Margin.of.Error
## # i 622 more variables: Alabama..Percent..Estimate <chr>,
## #   Alabama..Percent..Margin.of.Error <chr>, Alabama..Male..Estimate <chr>,
## #   Alabama..Male..Margin.of.Error <chr>,
## #   Alabama..Percent.Male..Estimate <chr>,
## #   Alabama..Percent.Male..Margin.of.Error <chr>, ...
```

For this report we made use of publicly available datasets to explore the relationship between the variables education and income across the United States from 2019 to 2022. The education datasets were obtained from the U.S. Census Bureau's American Community Survey (ACS). The datasets report educational attainment segmented by state, age group, gender, ethnicity, poverty rate and median earnings. The ACS is a large-scale survey that is conducted annually throughout 3.5 million households across the U.S. Data is collected through different methods, including postal questionnaires, online responses and phone interviews. Although the dataset includes a wide range of age groups, we chose to focus specifically on individuals aged 19 to 24, which is further explained in section 3.1. The income datasets were sourced from the U.S. Bureau of Economic Analysis (BEA). These datasets report the total personal income by state, in millions of U.S. dollars. Unlike the ACS, BEA's data is collected through a combination of administrative rather than surveys. Data is obtained through sources like tax return data from the IRS, as well as employment data from state unemployment insurance programs and federal benefits data. By compiling and modelling this data the BEA produces annual estimates of total personal income for each state. Reflecting earnings from wages, business income, government benefits and investment income.

2.3 Describe the type of variables included

- `year`: the calendar year for each data point, indicating when the income and education values were recorded (2019, 2021, or 2022). It is neither a health nor an SES metric.
- `GeoName`: the full name of the U.S. state for each observation, used to merge, label, and map data; not a health or SES metric.
- `education`: the raw count of 18–24 year-old holding a bachelor's degree in a given state and year, extracted from the ACS "total estimate" column; an SES indicator.

- `income`: the total personal income for each state and year, measured in millions of U.S. dollars and derived from BEA administrative records; an SES indicator.
- `avg_education`: the average number of bachelor's-degree holders per state, calculated as the mean of the three annual counts (2019, 2021, 2022); an SES indicator.
- `avg_income`: the average total personal income per state, calculated as the mean of the three annual income figures (in millions of USD); an SES indicator.

Part 3 - Quantifying

3.1 Data cleaning

We selected column 2 (GeoName) and every 12th column from 3 to 604 to capture the "Total Estimate" values for each state in each year.

We appended a year column to each dataframe to label observations by year.

We extracted the 6th row, which corresponds to counts of 18–24 year-olds holding a bachelor's degree or higher.

We combined the three bachelor datasets into one dataframe.

We renamed the grouping column to Variable for clarity.

We removed unneeded columns, pivoted to long format, and cleaned up the state names.

We replaced all periods with spaces in GeoName to get proper state names.

We inspected the unique values of GeoName to ensure formatting was correct.

We removed commas and converted the education column from character to numeric.

For each income dataset (2019, 2021, 2022), we added a new year column to indicate the year the data represents. We also renamed the income column (e.g., X2019) to a standard name income so that we can later combine them easily.

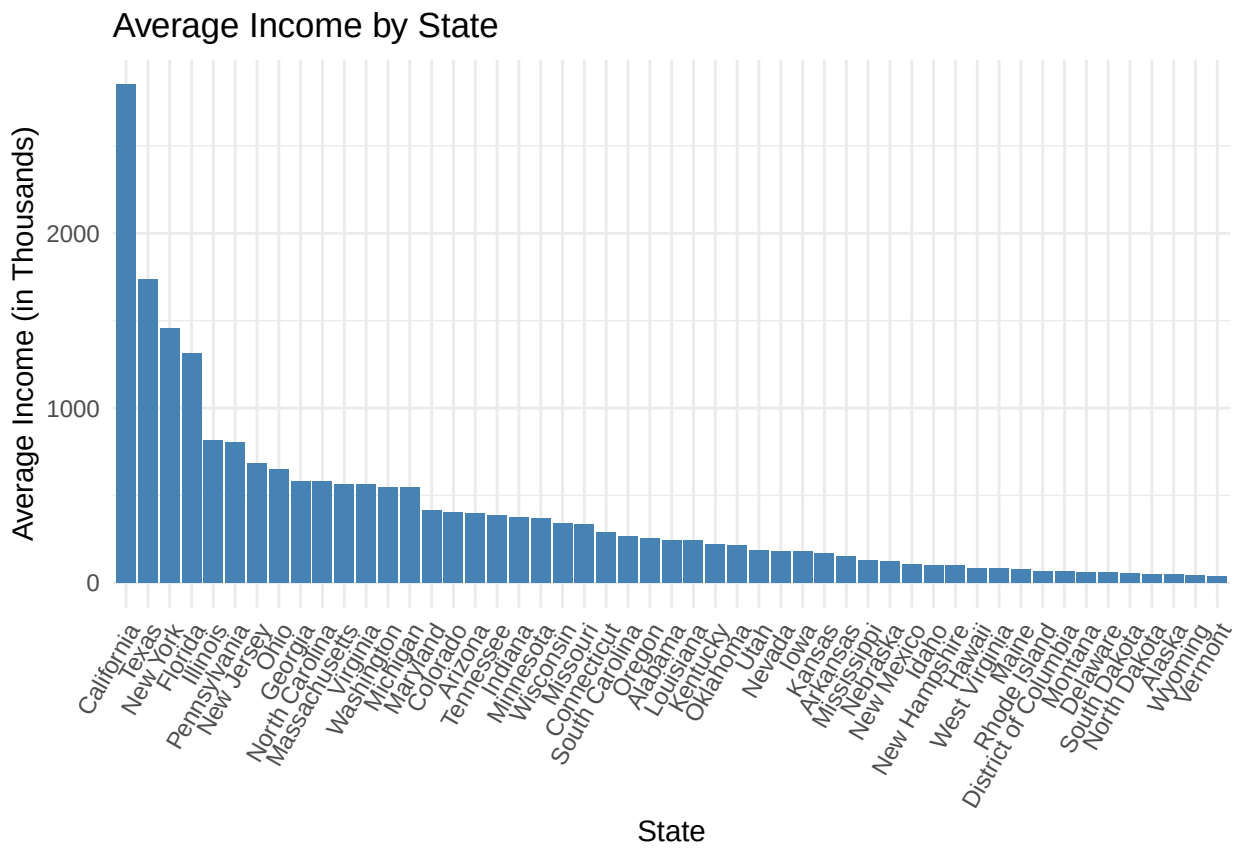
We merged the three cleaned yearly datasets into a single dataframe using bind_rows(). This created one long-format table with income and year labels together.

We filtered out the row corresponding to "United States" (a national-level aggregate not needed for state-level analysis). Then, we kept only the year, income, and GeoName columns. Finally, we removed any rows with missing (NA) income values to ensure clean data for further analysis.

3.2 Generate necessary variables

Variable 1:Average income

Here, we are forming a bar chart of average education by state. We ordered it from highest to lowest

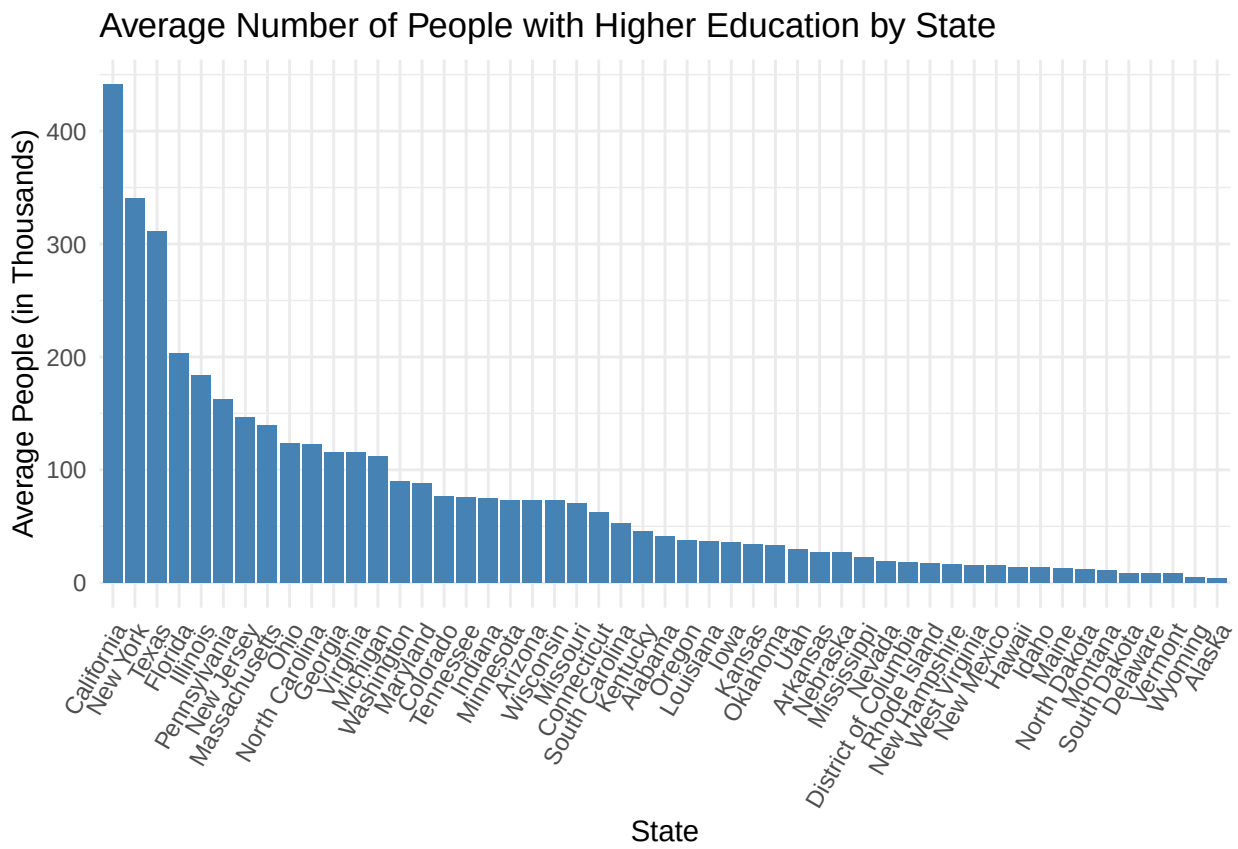


Variable 2:Average education

Here, we cleaned the data first, then we grouped it by states. We calculate the average education.

We are creating a plot here. Sorting the states from highest to lowest average education.

```
#Plot figure

ggplot(average_education, aes(x = reorder(GeoName, -avg_education), y = avg_education/1e3)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(
    title = "Average Number of People with Higher Education by State",
    x = "State",
    y = "Average People (in Thousands)"
  ) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

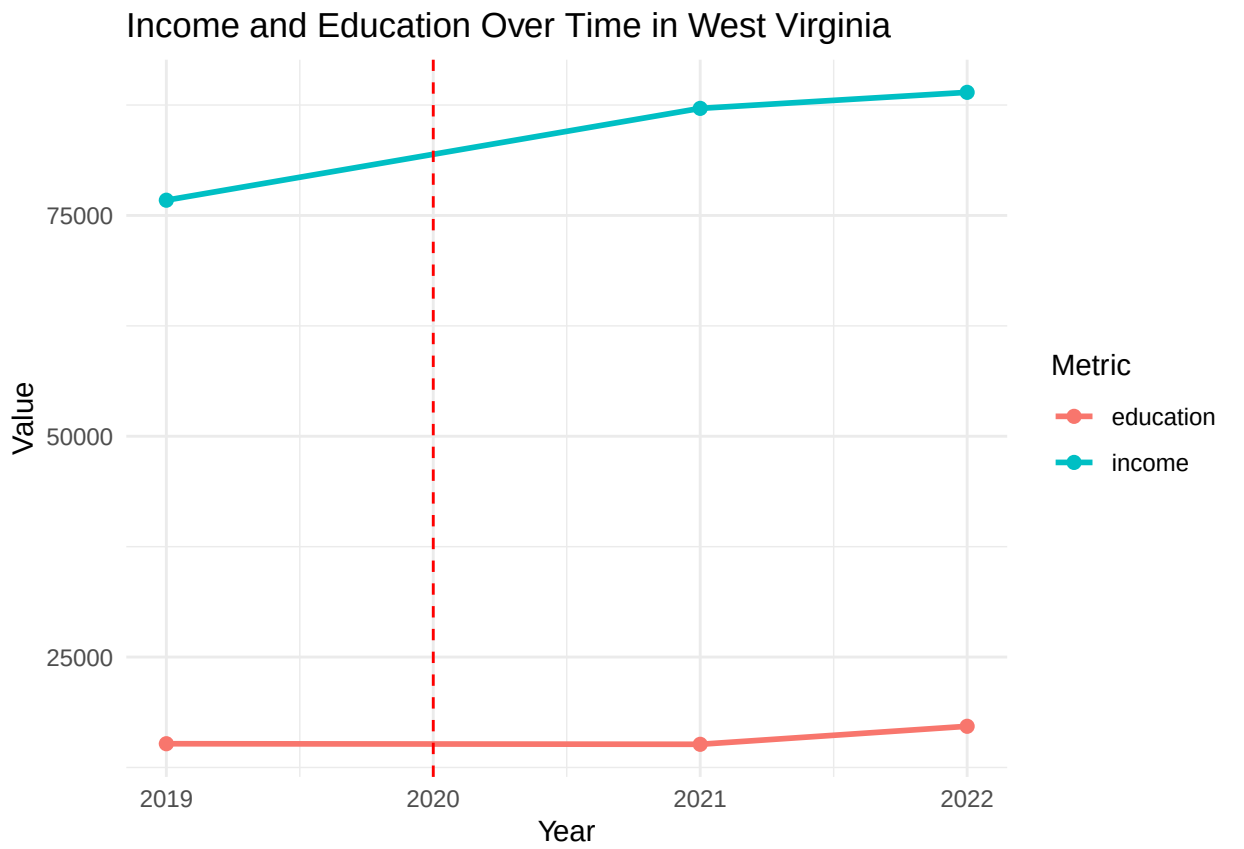


We created a new variable, average income and average education per state, to gain a better understanding of income and education trends across the United States during the period analysed. Creating an average per state helps to smooth out year-to-year fluctuations and provides a more representative measure of income and education levels.

3.3 Visualize temporal variation

```
# Plot
ggplot(plot_temporal, aes(x = year, y = value, color = Metric)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  geom_vline(xintercept = 2020, linetype = "dashed", color = "red") +
  labs(
    title = "Income and Education Over Time in West Virginia",
    x = "Year",
    y = "Value",
    color = "Metric"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

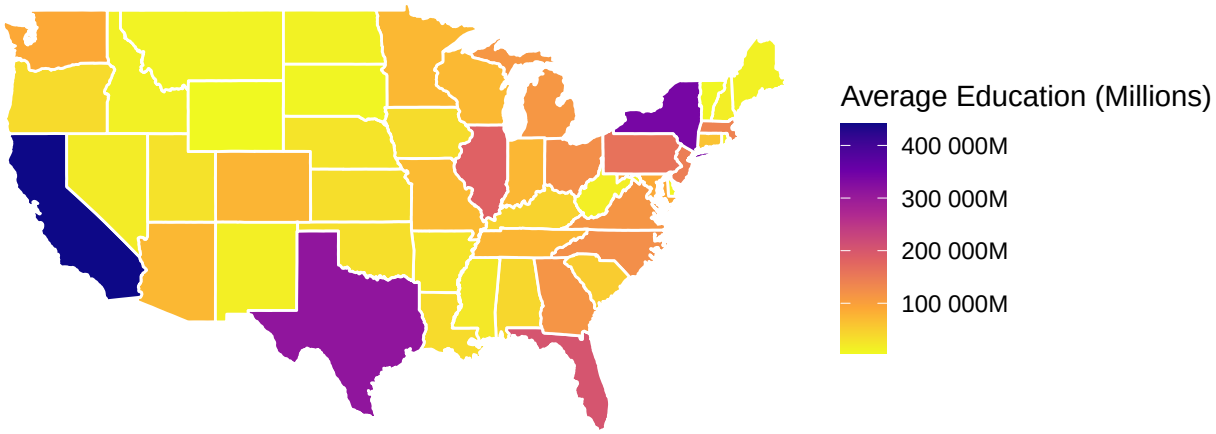


In this plot we try to show the relationship and the linkage between education, and income throughout 2019 to 2022. We selected West Virginia because it is the state with the highest obesity rate. The plot also tries to show if covid(2020) had a causal relationship with the increase in education and income. We did not use the 2020 data since during the covid there was no data collected from our source.

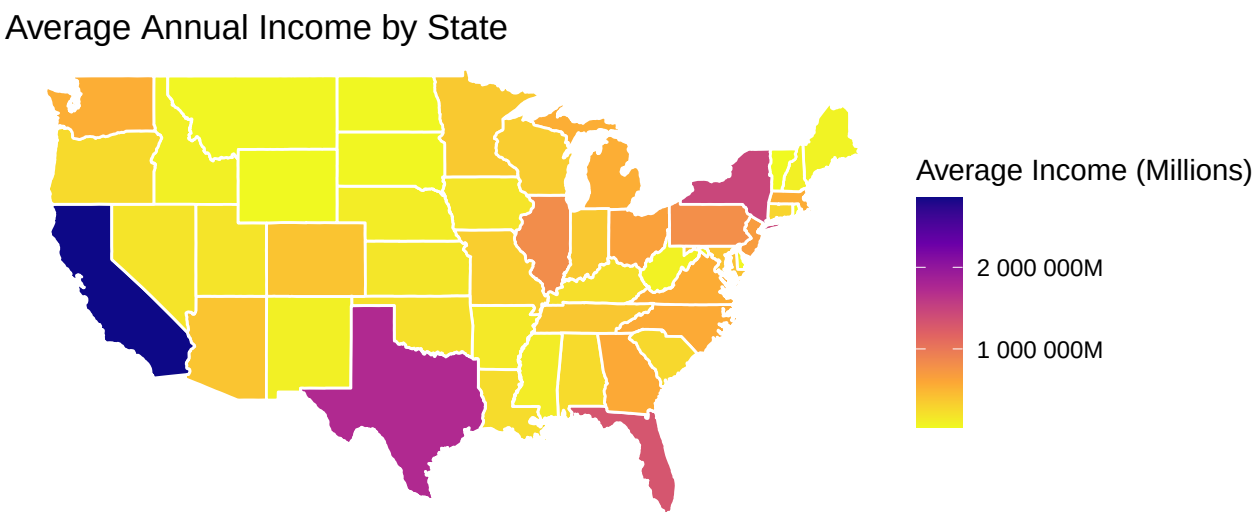
3.4 Visualize spatial variation

```
# Plot average education per state with color scale in millions

ggplot(education_map, aes(x = long, y = lat, group = group, fill = avg_education)) +
  geom_polygon(color = "white") +
  coord_fixed(1.3) +
  labs(
    title = "Average Number of People with Higher Education by State",
    fill = "Average Education (Millions)"
  ) +
  scale_fill_viridis_c(option = "plasma", direction = -1, labels = scales::label_number(suffix = "M", a
  theme_void()
```

Average Number of People with Higher Education by State



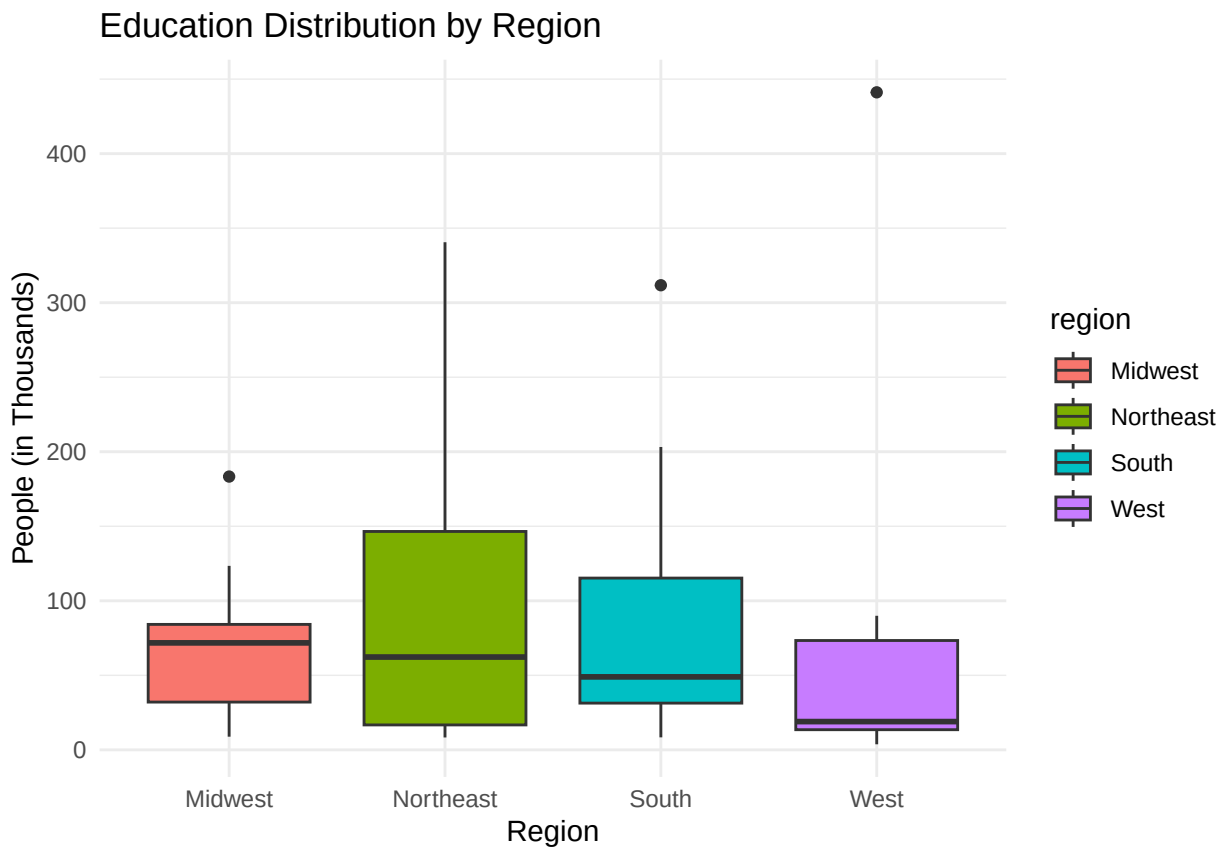

```
# Plot average income per state with color scale in millions

ggplot(income_map, aes(x = long, y = lat, group = group, fill = avg_income)) +
  geom_polygon(color = "white") +
  coord_fixed(1.3) +
  labs(
    title = "Average Annual Income by State",
    fill = "Average Income (Millions)"
  ) +
  scale_fill_viridis_c(option = "plasma", direction = -1, labels = scales::label_number(suffix = "M", a
  theme_void()
```



In the spatial variation we graphs, we used the map of the United States of America. There are two graphs(One showing the average education levels, the other one showing the average income levels). Darker the shade of the color of the state, higher the variable becomes. The plots above are relevant to obesity because it helps us visualize which states show low results in the indicators we chose(income and education) for obesity.

3.5 Visualize sub-population variation

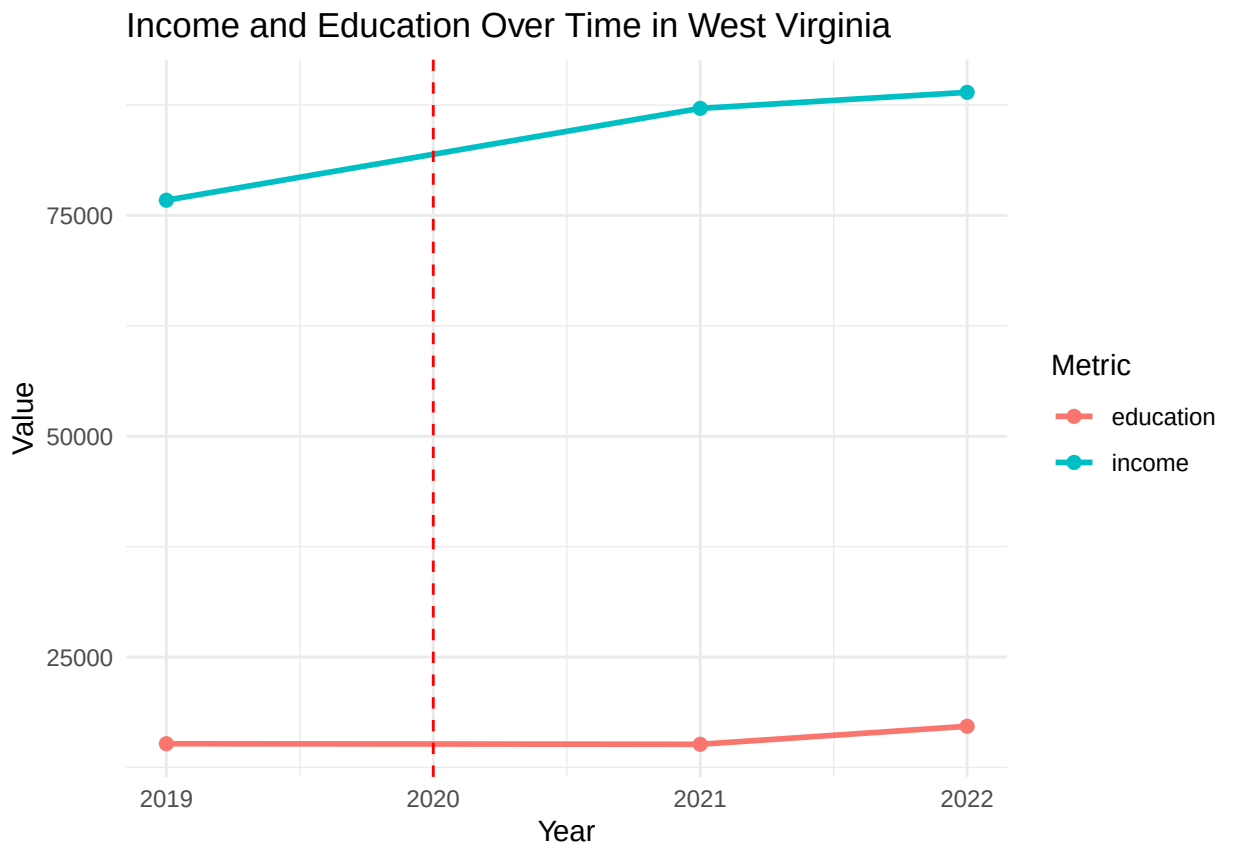
```
#Plot
ggplot(education_states, aes(x = region, y = avg_education / 1e3, fill = region)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Education Distribution by Region",
    x = "Region",
    y = "People (in Thousands)"
  )
```



This plot visualizes the distribution of the number of people by region who had attained a bachelor's degree or higher. It is relevant to our research because education is one of the most important variables we are investigating in relation to obesity across the United States. By comparing regions such as the Midwest, Northeast, South, and West, we can observe how education levels changed geographically.

3.6 Event analysis

```
# Plot
ggplot(plot_temporal, aes(x = year, y = value, color = Metric)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  geom_vline(xintercept = 2020, linetype = "dashed", color = "red") +
  labs(
    title = "Income and Education Over Time in West Virginia",
    x = "Year",
    y = "Value",
    color = "Metric"
  ) +
  theme_minimal()
```



Our analysis found a strong positive correlation between income and education across U.S. states. States with higher average income seem to generally have higher levels of education . This pattern was especially visible in both the spatial variation maps and the temporal trends following COVID-19. We chose the COVID-19 pandemic as our main event because it represents one of the most significant social and economic events in recent history. We examined how education and income levels changed through the years of 2019,2021, and 2022 in order to see how these indicators for obesity changed before, during, and after the pandemic crisis. When we examined these indicators, we wanted to understand how the pandemic may have contributed in terms of long-term effects to public health.

Part 4 - Discussion

4.1 Discuss your findings

This study tries to show how education levels, paychecks, and obesity rates line up in every U.S. state over a span of years. Our study used data from education, and income levels(2019, 2021, and 2022) to spot shifts that appeared before and after COVID-19 stepped in. In the COVID-19 period, income and education levels were inflated which also caused the obesity levels to differentiate between states a lot more. In our research, not only did we conclude that income and education levels affect obesity levels, but we also found that income and education are also directly affected by one another. Consequently, we also should not take education levels and income levels in the states as separate entities but as one because they are so closely correlated with each other. For the spatial variation visuals, the interesting fact to look at in both of these graphs is that they look very familiar in terms of the shading of the states. We can see that especially in California and some of the states in the East both become darker in their shadings in both the average income and average education maps. Hence we can understand that there is a causal link between the two of them. We can also arrive at the point that the lightest shaded states in both these graphs are the states that would be

having the hardest time dealing with obesity because both the indicators of obesity(income and education) are the lowest in these states. When we inspect the temporal variation plot, we see that both are affected by covid hence the rapid increase in both of them. During the 2021-2022 period, the line for education becomes steeper, on the other hand, income seems to stop its rapid increase. Gathering information from this graph, we can conclude that covid may have helped increase both education and income during this period. Consequently, since both income and education levels changed in this period, obesity levels would be very different compared to before. The clear pattern shows that, as education and earnings climb,they follow each other very closely hence their causal relationship. When these indicators increase, obesity drops. COVID-19 widened those holes. All in all, to decrease obesity levels throughout the US, the government should focus on increasing income and education.

Part 5 - Reproducibility

5.1 Github repository link

https://github.com/matildemaio/Obesity

5.2 Reference list

Barraclough, A. (2025, January 29). Healthy eating now twice as expensive as junk food, study finds. Women's Health. https://www.womenshealthmag.com/uk/food/healthy-eating/a63598807/healthy-eating-twice-as-expensive-as-junk-food-study/

BEA. (2025, March 28). BEA Interactive Data Application. Apps.bea.gov. https://apps.bea.gov/itable/?ReqID=70&step=1&_gl=1 CDC. (2025, January 16). Adult Obesity Facts. Obesity. https://www.cdc.gov/obesity/adult-obesity-facts/index.html#print States, U. (2025). Explore Census Data. Census.gov. https://data.census.gov/table/ACSST5Y2022.S1501?q=education+by+state&t=Populations+and+People World Obesity. (2024). Ranking (% Obesity by country). World Obesity Federation Global Obesity Observatory. https://data.worldobesity.org/rankings/

Side Note: -BEA does not conduct original surveys; they compile data from federal agencies, administrative reports, and NIPA reports from the government. -Data.census.gov comes from the U.S. Census Bureau's American Community Survey (ACS)