Analyzing How Physical Inactivity and Income Influence Obesity in U.S. Adults (2021-2025) Using Exploratory Data Analysis and Linear Regression

Leila Shafizadeh
Johns Hopkins University Whiting School of Engineering

Biomedical Data Science
Fall 2025

I. Project Overview

This project explores how income level and physical inactivity relate to obesity rates in adults across U.S. states using the CDC's Nutrition, Physical Activity, and Obesity – Behavioral Risk Factor Surveillance System (BRFSS) dataset. The analysis focuses on data from 2021-2025 to ensure it is relevant to current public health trends and conditions. The dataset was cleaned and reshaped to associate obesity and inactivity percentages by year, state, and demographic group (which was then narrowed to income). Exploratory data analysis and linear regression modeling were chosen to examine the trends in the data, with the ultimate goal to understand how socioeconomic factors and lifestyle behaviors contribute to obesity prevalence in U.S. adults and to demonstrate data-driven insight generation using public health datasets.

II. Background and Motivation

Often referred to as the "Obesity Epidemic," obesity is an urgent growing public health concern in the U.S. It can be closely tied to chronic conditions like diabetes and cardiovascular disease, and many behavioral and socioeconomic factors–especially income level and physical activity–are recognized as key determinants of obesity. Up-to-date population level data can reveal new insights, especially when analyzed using modern tools; this dataset was last updated September 12th, 2025. The BRFSS dataset in particular provides extensive survey data across states and territories allowing for examination of how health behaviors and outcomes vary across regions and demographics. By using an exploratory and linear regression-based approach to compare inactivity and income against obesity rates directly, this report shows how intersecting behavioral and socioeconomic factors can explain public health disparities.

III. Key Project Components

a. Dataset selection: used the CDC *Nutrition, Physical Activity, and Obesity – BRFSS* dataset focusing on questions for obesity (Q036) and physical inactivity (Q047) and filtered for information on only the past five years (2021-2025).
b. Cleaning: standardized column names, removed missing or unreliable entries, filtered timing and location, and limited to "Total," "Income," and "Education" categories.
c. Reshaping: pivoted the dataset such that each individual record contains both obesity and inactivity rates for the same group, allowing for direct comparison.
d. Analysis: performed exploratory data analysis to visualize trends and linear regression to quantify how inactivity and income relate to obesity.

IV. Dataset

The dataset was published by the Centers for Disease Control and Prevention (CDC) and is structured as a multi-year, state-level dataset containing behavioral and health indicators derived from survey responses. The main variables used in this analysis are QuestionID Q036 (percentage of adults classified as obese (BMI $\geq$ 30)) and QuestionID Q047 (percentage of adults reporting no leisure-time physical activity). Some additional fields used are YearStart,

LocationAbbr, StratificationCategory1 (category type: Total, Income, Education), Stratification1 (specific subgroup), and Data_Value (percentage). The scope of the data was filtered to include 2021-2025, 50 U.S. states, and subgroups with sample size > or = to 100 to maintain statistical reliability.

V. Data Preprocessing

First, core Python libraries are imported including pandas for data manipulation, numpy for numerical operations, and matplotlib.pyplot for visualization. Then, the dataset is loaded into a pandas DataFrame. Because the CDC CSV file contains inconsistent column naming conventions, logic is included to detect and assign the correct column names so that they can be selected and filtered. YearStart is converted to numeric and filtered to include only the past five years. The dataset is also reduced to entries for two key health indicators identified by their QuestionID: Q036 for obesity rate (% obese adults) and Q047 for physical inactivity rate (% no leisure-time activity). Only relevant population stratification categories are retained (Total, Income, and Education) to allow for demographic comparisons. The data is then cleaned column by column: the Data_Value column (measured percentage) is converted to numeric and rows with missing values are removed, confidence limit columns are also converted to numeric if possible, and the sample size column is standardized, stripped of commas, converted to numeric, and filtered so that only groups with a sample size 100 or above remain to ensure statistical reliability. Data from U.S. territories are also excluded to maintain consistency with analyses focused on the contiguous states.

Once cleaned, the dataset is pivoted so that each row represents one group defined by year, state, and demographic category, containing two side by side variables: Obesity_Pct (obesity percent) and Inactivity_Pct (inactivity percent). Structuring it like so allows for easier direct comparison and correlation analysis between inactivity and obesity within each group. The final df_pivot table is compact and analysis-ready linking obesity and inactivity rates across years, states, and demographic categories.

VI. Methods Used

a. Exploratory Data Analysis (EDA)

This method was used to visually and statistically explore how physical inactivity and income relate to obesity rates. For the inactivity-level analysis, the numeric data columns Inactivity_Pct and Obesity_Pct are selected from the reshaped dataset and a scatter plot is generated to visualize their relationship. A linear trendline is generated using numpy.polyfit() to estimate the overall direction and strength of correlation. The Pearson correlation coefficient is calculated using np.corrcoef to quantify how strongly the two variables move together, with results showing a positive correlation, indicating that higher inactivity corresponds to higher obesity rates. For the income-level analysis, the dataset is filtered to include only records categories by income level, grouped by income brackets, and the average obesity rate is then calculated within each group. This information was visualized in a bar chart to demonstrate that

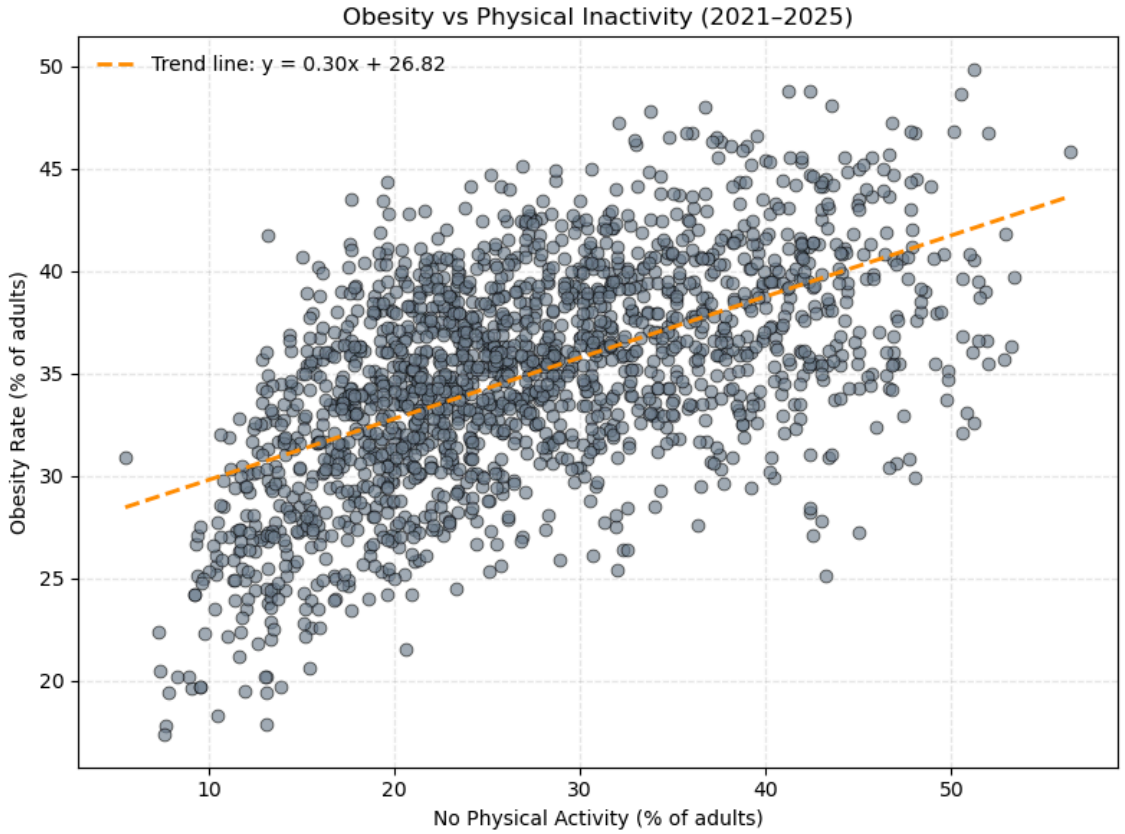
lower-income groups had higher average obesity rates–suggesting a clear socioeconomic gradient.

b. Linear Regression

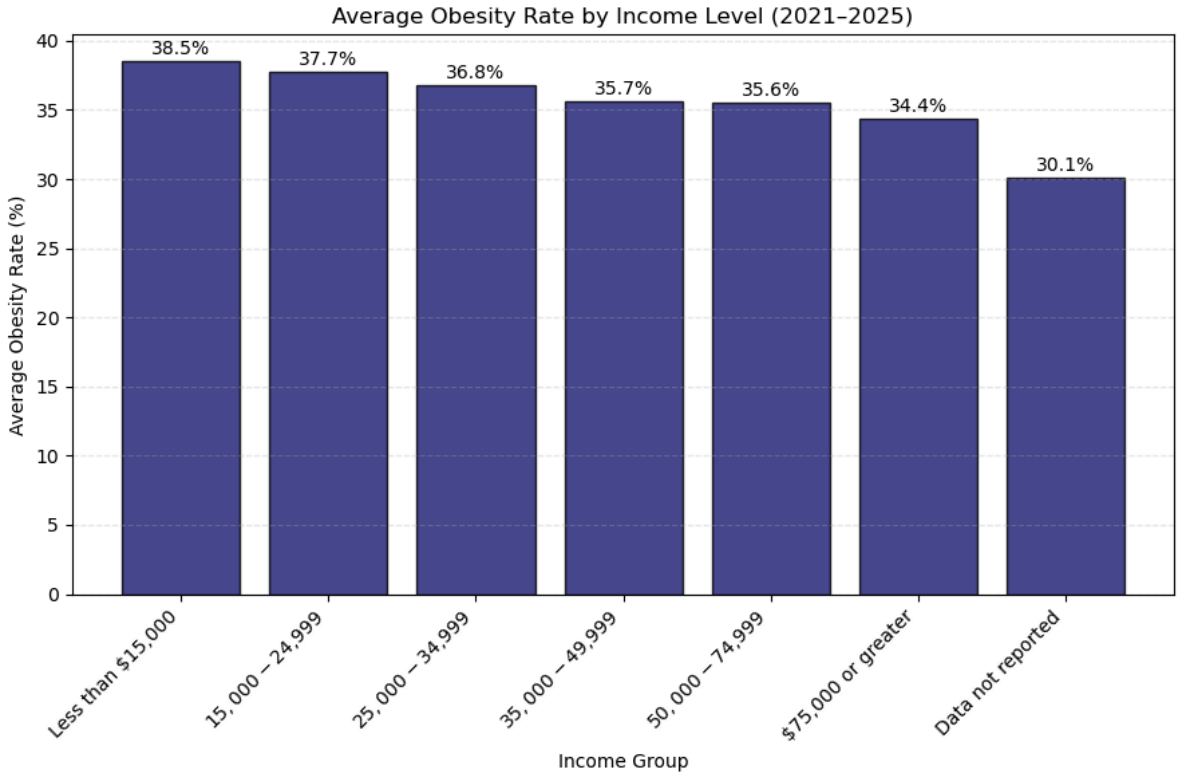
This method was used to formally model and compare how well inactivity and income level predict obesity prevalence. Numeric fields were first converted and missing values removed to ensure a clean input and income categories were encoded as numerical codes for regression compatibility. Then, two separate models were built using scikit-learn's LinearRegression() function, with model 1 relating Obesity_Pct and Inactivity_Pct and model 2 Obesity_Pct and Income_Level. Each dataset was split into training (80%) and testing (20%) subsets using train_test_split() to evaluate model performance. Both models were trained then evaluated using R^2 score, which measures how much variation in obesity is explained by each predictor. Since the inactivity model had a higher R^2 it indicates that there is a stronger relationship and inactivity is a more direct predictor of obesity. Plots for predicted vs actual obesity rates were generated for both models to further visualize fit quality. The inactivity model's points are clustered closer to the fit line which confirms the tighter predictive relationship shown by comparing the R^2 scores.

VII. Results

a. EDA



The correlation between inactivity and obesity was calculated as 0.559. A positive correlation indicates that higher inactivity generally corresponds to higher obesity rates. This is also demonstrated in the scatterplot as the cluster of points are generally densest along the fit line.



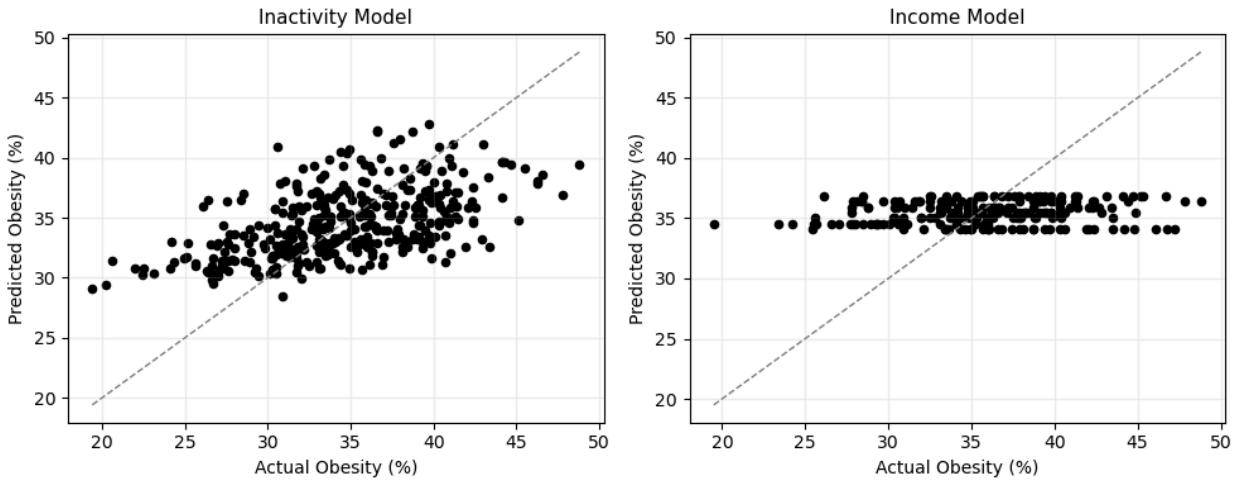
Lower-income groups typically show higher average obesity rates, suggesting a socioeconomic gradient in obesity prevalence as shown in the bar graph. The average obesity rate for individuals making less than $15,000 is 38.5% which gradually decreases by income stratification to being 34.4% for individuals making $75,000 or greater.

b. Linear Regression

Inactivity vs Obesity
Equation: Obesity = 26.77 + 0.30 × Inactivity
R^2 score: 0.291

Income vs Obesity
Equation: Obesity = 36.82 + -0.46 × Income_Level
R^2 score: 0.018

Model Comparison: Inactivity is a better predictor of obesity ($R^2 = 0.291$) than income ($R^2 = 0.018$). The higher the R^2, the stronger the relationship.



Both models show that obesity is related to inactivity and income, but the shape of the inactivity graph, with points falling closer along the fit line, shows a stronger relationship compared to the income graph which is more spread out.

VIII. Conclusion

From these simple regression models, one can see that both physical inactivity and income level have a relationship with obesity rates. The model using inactivity explains a bit more of the variation in obesity because it has a higher R^2 and the points on its plot fall much closer to the fit line, which means the model's predictions are more accurate. This suggests that how active people are may have a more direct link to obesity than income alone.
However, income still seems to matter as areas or groups with lower income often have higher obesity rates on average. This could be because income affects access to healthy foods, exercise opportunities, and overall lifestyle. In short, both factors are connected to obesity, but physical inactivity appears to be the stronger single predictor in this dataset. It's likely that the two variables also interact in real life: people with lower income might also face more barriers to being physically active.

The main limitations with the approach taken are that the data is based on state-level averages and self-reported surveys which can be biased and hide individual variation, and that the models used may not be the most accurate methods to verify the conclusions made. Linear regression models assume simple, direct relationships and don't necessarily account for confounding factors like age, education, or regional differences. Especially since inactivity and income were analyzed separately it's hard to tell how they interact with one another or which is the stronger predictor when considered together. A stronger next step would be to build a combined model using both variables and potential confounders. A future analysis could use a Random Forest model to capture nonlinear relationships and interactions between income, inactivity, and obesity that linear regression might miss.