# Obesity and Walkability in the United States*

Cole Gaynor
Wentworth Institute of Technology
Boston, MA, USA
gaynorc2@wit.edu

## ABSTRACT

In this report, we analyze statistics surrounding obesity rates in the United States and walkability scores in the United States. We use a linear regression model to examine the potential correlation between walkability and obesity.

## KEYWORDS

Walkability, obesity, public health

## 1 Introduction

Obesity in the United States is an epidemic, with one of the highest rates in the world. Obesity is medically associated with a large number of comorbid diseases and health complications, often causing a decrease in quality of life and, on a material level, a greater strain on health services: the United States spends the most on healthcare per capita of any OECD nation. Additionally, the US is considered uniquely bad for pedestrian travel among developed nations. In this paper, we examine links within the context of the US connecting obesity rates to prevalence of pedestrian travel and walkability to examine if efforts to promote pedestrian travel and walkability have a positive effect on lowering obesity rates.

## 2 Data

### 2.1 Source of dataset

This project uses data covering walkability indices on a county level within the United States. It also contains data such as mean wage and other economic factors that can be controlled for so that variables are not identified as causal if they are both correlated to a third economic variable. This dataset was generated in 2025 and is taken from a United States .gov source, which is considered reliable.

Our other dataset covers state-level obesity rates. This dataset is from 2024 and is also from a United States .gov source.

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note


*WOODSTOCK'18, June, 2018, El Paso, Texas USA*


### 2.2 Characters of the datasets

All datasets are available for analysis in csv format. Our data includes walkability scores as well as general economic metrics at a country level in addition to obesity levels by state.

## 3 Methodology

This project uses a regression to analyze the correlation between variables related to walkability to predict obesity. We use a linear regression to make this estimation, because this is a simple model, and it seems unlikely that there would exist complex interactions that would necessitate a more complex model.

We also analyze the distribution of analyzed statistics to identify outliers.

In order to get data into comparable formats for analysis, aggregation was performed from the level of Census regions into the state level, weighting averages appropriately from all rows of the data. Per capita statistics were weighted with respect to region population, and per area statistics were weighted with respect to region area, *excluding area unavailable for development* such as environmentally protected land. This is to give a better view of actually usable land and density in livable areas of states, rather than taking into account conservation land and areas that are categorically unavailable for development and letting that skew our view of what actually inhabited areas of each state look like.

All instances with null values or unreported values, including values that are impossible, were dropped from this analysis.

Our first question is to examine outliers in specific variables. For this, a threshold of 3 standard deviations away from the mean for all states was determined to be considered an outlier. We analyze significant variables including:

- Percentage of households owning 0, 1, or 2+ cars
- Percentage of population earning less than $1,250 per month, or $15,000 per year
- Per-census area average distance of population centroid to the nearest transit stop, weighted by population of census area
- Population density on unprotected land (people per acre)
- Job density on unprotected land (jobs per acre)
- National Walkability Index score
- Obesity rate in percentage

We then also analyze regional differences. We analyze the means and standard deviations of walkability and obesity in each region of the United States (Northeast, South, Midwest, and West) to examine regional trends.

We also examine several regressions to decide their efficacy.
First, we examine a regression of economic factors to predict walkability: can population density, job density, and percentage of population earning low wages predict the walkability score of a state?
We then examine a regression that attempts to predict obesity rates. For this, we include all variables analyzed in our earlier questions as inputs to the regression and attempt to predict obesity rates for a state. For this regression, we exclude our variables representing the percentage of population owning 1 car and the percentage of population owning 2 or more cars, as these are not independent from our variable representing the percentage of population owning no cars.
All regressions will use Mean Squared Error as their loss function. Each model will use 30% of data for testing and 70% for training and will be trained using Sci-Kit-Learn. The testing training split function will have a random_state parameter of 1. All inputs were scaled using a StandardScaler from SciKitLearn prior to training.

## 4 Results

Our first result is identification of outliers of significant variables. Results are presented in the below table:

| Variable | Low Outliers | High Outliers |
|---|---|---|
| Percent owning no cars | District of Columbia, New York | None |
| Percent owning 1 car | None | West Virginia |
| Percent owning 2+ cars | District of Columbia | None |
| Low wage earners | None | Mississippi |
| Centroid distance from nearest transit | West Virginia | None |
| Population density | None | District of Columbia |
| Job density | New Hampshire | District of Columbia |
| Obesity rate | None | None |

Notably, obesity has no outliers, while the District of Colombia appears very often on lists of outliers. This is sensible, as the District of Columbia has a very different demographic makeup than any other state; it is almost entirely urban with no rural or even suburban population.

Next, we examine regional differences in means:

| Region | Percent no cars | Percent 1 car | Percent 2+ cars | Low wage | |
|---|---|---|---|---|---|
| Northeast | 17.04 | 38.60 | 44.19 | 23.15 | |
| South | 11.58 | 41.65 | 46.24 | 25.46 | |
| Midwest | 10.63 | 38.47 | 50.64 | 25.48 | |
| West | 7.36 | 34.25 | 58.28 | 22.51 | |

| Region | Transit Distance (m) | Walk. Index | Pop. Density | Job Density | Obesity |
|---|---|---|---|---|---|
| Northeast | 495.4 | 14.01 | 6.13 | 4.31 | 26.40 |
| South | 559.4 | 13.36 | 5.88 | 3.84 | 31.85 |
| Midwest | 496.4 | 13.50 | 5.11 | 3.24 | 31.00 |
| West | 554.5 | 13.92 | 5.33 | 2.98 | 26.43 |

We note that the Northeast is the region with the lowest car ownership and the highest walkability scores, as well as the highest density of both inhabitants and jobs. It also boasts the lowest average obesity.

For our model prediction of Walkability scores using density metrics and low wage earners, we present our error metrics below:

| Metric | Value |
|---|---|
| Mean Absolute Error | 0.534 |
| Mean Squared Error | 0.430 |
| $R^2$ | 0.114 |

Generally, these metrics suggest that the model does hold some value, if not extremely precise prediction capabilities. We then look at the coefficients for our model:

| Variable | Coefficient |
|---|---|
| Population Density | -0.021 |
| Job Density | 0.122 |
| Low Wage Percent | -1.77 |

The negative variable on population density is likely due to a high correlation between population density and job density: a high job density is more impactful on walkability because urban city centers are where jobs are concentrated, and likely to be more walkable. As such, the correlated variable of population density has a negative coefficient since it is not as strongly correlated with the target as job density is. The negative coefficient on low wage percentage suggests that poorer areas are significantly less likely to be walkable.

For our model prediction of obesity prevalence using walkability metrics, we present the following metrics:

| Metric | Value |
|---|---|
| Mean Absolute Error | 2.067 |
| Mean Squared Error | 5.748 |
| $R^2$ | 0.491 |

This suggests a much more predictive model. With a higher $R^2$ and error values that are lower in comparison to the variance in what it predicts, this shows that this model is quite good. Here we present coefficients:

| Variable | Coefficient |
|---|---|
| Population Density | 0.151 |
| Low Wage Percent | 91.76 |
| Percent Owning No Cars | -7.818 |
| Distance to Transit | 0.00315 |
| Walkability Index | -0.0762 |

These coefficients suggest that there is a very strong correlation between low wage earning and obesity, as well as a significant negative correlation between the population owning no cars and obesity. There are also small correlations between density and distance to transit, as well as a small negative correlation between walkability and obesity.

## 5 Discussion

Notably, this report does not use granular, census-block-level obesity data, and is forced to aggregate the wealth of granular data available into state data. A much more predictive study would include lower-level state obesity data. Additionally, the presence of null values in the data set did significantly cut down on the availability of data; around 50% of rows contained null values. Null values were also more likely to occur outside of the continental united states, making it infeasible to analyze US territories in this project.

## 6 Conclusion

This report has shown a connection between obesity and the percentage of populations that own cars; that is, the greater proportion of the population not owning cars, the less prevalent obesity is. We also examine that the Northeast region of the United States, as well as the District of Columbia, are particularly good in these metrics: they have, respectively, the lowest average car ownership of any US region and the lowest car ownership of any singular US state-level subdivision in our analysis. With this information, states and counties can better make decisions on how to fight the obesity epidemic in the United States at a city planning and urbanist level, by promoting walkability and walking in order to reduce car reliance and combat obesity.

## REFERENCES


[1] U.S. Environmental Protection Agency (EPA). 2021. Walkability Index (Version 3). Data.gov. Retrieved from https://catalog.data.gov/dataset/walkability-index8

[2] U.S. Department of Agriculture (USDA), Economic Research Service. 2020. National Obesity by State. Data.gov. Retrieved from https://catalog.data.gov/dataset/national-obesity-by-state-d765a