A statistical analysis on the effect of socioeconomic status on average obesity within the United States

By Siddharth Potti

Project Summary:

The goal of my program was to create a model and visualizations that accurately represent the effect of socioeconomic status on average obesity within the United States. I also believe that average obesity rates have increased in each state over the past few years. I aim to measure socioeconomic status by using the average income level in each state which would be one of my features. Another feature would be education attainment by state based on receiving a bachelors or higher, as it would be a predictor for a higher average socioeconomic status. My third feature to measure socioeconomic status would be median home value in each state. I believe that a higher median home value would signify a higher average socioeconomic status for each state. My label would be the rate of obesity for each state. I got my data from sources outside of Kaggle that reported these features for every state in the U.S. I converted all of my data into one csv table, so that I could easily use it once coding. As of today, my code correctly visualizes the relationship between different socioeconomic status features and average obesity rate, as well as contains an Linear Regression and Multiple Linear Regression Model that has a high accuracy. I then used statistical measures like residual plots and z score tests to support the use of my models and to check my hypothesis.

Project Description:

Since my project is written in Python, I did not need to use classes. My project instead is divided into parts. The first part of my program is pre-processing. I removed all null elements from my data and organized the data into a dataframe data structure using pandas. The second part is data visualization, where I visualized the relationship between each of my features and the average obesity rate. I visualized these relationships by using matplotlib. Next, I created my linear regression model between average income and average obesity. I split my data into train and test data and ran the model using sklearn. Next, I conducted a statistical analysis on my model by measuring the R^2 and coefficients. I plotted a regression plot for this linear regression model, and verified that a multiple linear regression model could be made. After the linear regression model, I created that multiple linear regression model/algorithm using all of my features. Using numpy, sklearn, and seaborn, I ran the model and printed the predicted and actual values. I found that my model had a high accuracy and a high R^2 of 63.21. Lastly, I completed a Z-test, where I tested my hypothesis using a Z Table and finding the P value.

Citation:

Here are the sources that I used to get my data:

For measuring average obesity in each state, I used this source: https://stateofchildhoodobesity.org/adult-obesity/.

For measuring average income level in each state, I used this source: https://dqydj.com/average-income-by-state-median-top-percentiles/;

For measuring educational attainment by state, I used this source: https://worldpopulationreview.com/state-rankings/educational-attainment-by-state.

For measuring median home value in each state, I used this source: https://www.experian.com/blogs/ask-experian/research/median-home-values-by-state/

Libraries used:

Numpy, seaborn, matplotlib, sklearn, and pandas