Emily Huang & Reese Burns
Predictive Analytics
28 September 2024

Paper 1: The Relationship Between Fast Food Density and Obesity Rates Across the U.S.

Obesity is one of the most significant issues in the United States with various factors, including dietary habits and environmental influences, contributing to its prevalence. Therefore, this study aims to explore the relationship between the density of fast food restaurants in each state and its corresponding obesity rate. We use two distinct datasets, “Fast Food Restaurants” and “Obesity by State”, to perform an Exploratory Data Analysis (EDA) that determines whether there is a correlation between the availability of fast food and obesity rates in different states. This is important because understanding this relationship can help policymakers and public health officials identify areas where change is most needed. With these datasets we wanted to answer if there is a correlation between the number of fast food restaurants and the obesity rate of each state.

The first dataset, “Fast Food Restaurants”, sourced from Datafiniti's Business Database, contains information about 10,000 fast food locations across the U.S. with relevant attributes such as restaurant name, city, province, and geographic coordinates. It provides a broad sample of fast food availability by state. The second dataset, “Obesity by State”, compiled by the Centers for Disease Control and Prevention (CDC) BRFSS survey, includes adult obesity rates for each state, along with state name and size by area and length. Obesity rates are calculated by taking the percent of the state population that is considered obese from the BRFSS Survey. Together, these datasets offer a comprehensive view of the relationship between fast food distribution and obesity rates across the country.

We started preprocessing by inspecting each dataset to understand its structure and contents, followed by cleaning, standardizing, and organizing the data to prepare it for analysis. In the “Fast Food” dataframe, we standardized values to ensure uniformity, such as converting date entries from object types to DateTime format and renaming restaurant names to match their canonical forms. This was achieved by calculating a Jaro-Winkler similarity score and if this similarity score was higher than a certain threshold then the name would be changed to match the first appearance of the restaurant.

For the “Obesity by State” dataframe, we renamed the state column to align with the “Fast Food” dataframe, facilitating a seamless merge. Missing data was handled by filling null values with “0” while columns deemed unnecessary for analysis in both the “Obesity by State” dataframe and the “Fast Food” dataframe–were removed as they were irrelevant to the analysis. Finally, we performed an inner merge of the two datasets based on the state, resulting in a combined dataframe with three key attributes: State, FastFood_Count, and Obesity.

For data analysis and visualization, we began by producing a scatter plot to see if any correlation between the two variables was present based on visualization alone (Figure 1).

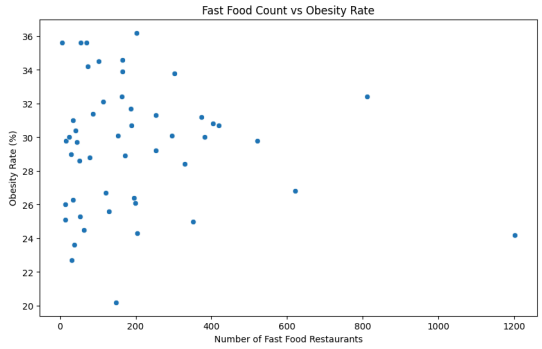

Figure 1: Scatter plot of Fast Food Count vs Obesity Rate

With no obvious correlation, we calculated the correlation coefficient to be -0.06, indicating a weak negative relationship between the number of fast food restaurants and the obesity rates by state, suggesting that as the amount of fast food restaurants increases, obesity rates slightly decrease. However, to better understand the data structure, we performed PCA to reduce dimensionality and enhance the interpretability of the dataset, preparing it for K-Means clustering (Figure 2).

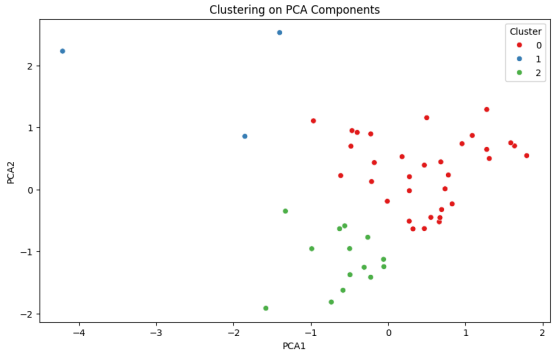

Figure 2: Clustering of PCA Components with 3 Distinct Cluster

The K-Means analysis resulted in three distinct clusters, highlighting variations in fast food density and obesity rates across states. Outlier states formed their own smaller cluster, indicating atypical characteristics in terms of fast food density and obesity rates. Before finalizing the analysis, we addressed outliers by either removing or appropriately treating them to prevent distortion of the results, ensuring more accurate insights into the data's pattern.

In the final stage of the analysis, we developed a model to predict obesity rates by state based on the number of fast food restaurants. After training on a linear regression model and the data, we evaluated the Mean Squared Error (MSE), R^2, and accuracy. A negative R^2 value of -0.13 suggests that the model failed to explain any variance in the data, demonstrating no correlation between the variables. Additionally, the model's low accuracy of 0.3 in classifying states into any obesity categories further confirmed the lack of significant linear relationship. These results indicate that the number of fast food restaurants does not significantly influence state obesity rates, suggesting that other factors play a more substantial role in determining obesity levels across the United States. For this project, we worked together on all components. Each contributing equally to both the coding and written portions.