Quantitative Analysis of Obesity Factors in the United States: Food Deserts, Food Swamps, and Food Heavens

Citadel Datathon Spring 2024 Team 28 Report

Aarav Dogra, Alicia Zhang, Saikhanbileg Tsogtgerel, Yule Fu

July 2024

Contents

1 Introduction

1.1 Motivation

Picture a common scene across America: fast-food chains line the streets, while grocery stores with fresh produce are few and far between. This has been the lived reality for many residents in what are called "food deserts" across the United States. Numerous studies have documented the health impacts of these food deserts: people living in areas with limited access to healthy food options are significantly more likely to suffer from obesity and related health issues.

This problem is more multifaceted than just a lack of access to fresh food. There is strong evidence for disparities in food access by income and race in America [1]. Additionally, a growing body of research for what are known as "food swamps", areas with higher access to fast food options, have linked the prevalence of these regions to higher rates of obesity and related issues. Food swamps are inextricably linked to food deserts, a fact that is a product of the footprint of processed food in the United States. A 2023 study found that counties with high food swamp scores were 77% more likely to have high obesity-related cancer mortality rates [2].

The obesity epidemic in the U.S. is driven by a complex interplay of socioeconomic and cultural factors. Food deserts and food swamps exacerbate the issue by limiting access to nutritious foods and perpetuating the consumption of processed food, contributing to higher obesity rates among affected populations. Thus, addressing food access is a crucial step toward mitigating the obesity epidemic. In this report, we wish to explore the link between food access and obesity by examining food access indicators, such as grocery store accessibility and fruit and vegetable consumption, within the context of socioeconomic factors.

1.2 Topic Questions

To investigate this issue further, we have two focused questions that we wish to answer:

1. How are food access indicators related to obesity rates?
2. How can we employ machine learning techniques to predict the obesity rate of an area based on certain indicators?

We performed data analysis to explore the first question, while we built a model to answer the second. We focus primarily on food desert indicators, but we also explore food swamp indicators as well.

1.3 Nontechnical Summary

We explored the link between food desert indicators and obesity rates in different parts of the U.S. We examined these factors nationwide on a state-wise basis, after which we narrowed our scope to the state of California on a county-wise basis.

We found that significant and logical predictors of obesity both nationwide and within California were **low access to fresh food** (specifically, people classified as low income or without a vehicle more than 1/2 a mile from a grocery store) and **food consumption habits** (including fruit, vegetables, fast food, and soda).

For the purpose of interpretability, we also created our own blanket metrics for food access: **food desert**, **food swamp**, and **food heaven**, which represent low access to fresh food, high access to fast food, and high access to fresh food, respectively. Nationwide, there were logical and significant relationships between each metric and obesity. The results were mostly inconclusive within California, except for a moderate correlation between food swamps and obesity.

We also used machine learning to create a **predictive model for mean obesity percentage as proof of concept of our best predictors**. The model employed a Gradient Boosting Regressor, and used key predictors such as fruit and vegetable consumption, proximity to grocery stores, and access to transportation. The model performed well in predictin obesity, underscoring **the importance of both dietary habits and food accessibility in managing obesity rates**. The model highlights the need for **public health initiatives** to improve access to healthy foods and promote better dietary habits, particularly in low-income and vehicle-less populations.

2 Methods

2.1 Scope and Variables

We defined our geographical scope as **all 50 U.S. states and the District of Columbia**. In addition, we conducted a more granular analysis on the state of **California**, which we observed as having both a high food swamp and food heaven score, indicating more diversity with respect to food access across counties.

All analysis was conducted on **only 2019 data**, with the exception of the number of grocery stores, convenience stores, fast food restaurants, and farmers' markets, of which more recent data was not available. We made the choice to include this 2016 and 2018 data due to the fact that likely not much change occurred in 1-3 years.

In addition, we made the choice to examine only 2019 data because of the lack of data across multiple years (sources were limited to only 2 years or only years before 2020), which would result in an inaccurate time series analysis. In addition, 2019 would likely most accurately reflect the current state of food access and obesity in the United States, as the COVID-19 epidemic had consequences that we might not be able to account for. Thus, our choice of 2019 also serves as a limitation of the confounding variable of COVID.

Below are all of the variables we examined within each region.

- Urban percentage
- Total population
- Median family income
- Obesity rate

Below are the variables calculated per 10,000 people.

- Convenience stores, 2016
- Farmers' markets, 2018
- Fast-food restaurants, 2016
- Grocery stores, 2016

Below are the variables calculated in terms of percentage of total population. The variable codes are shown in parentheses to improve readability of figures later in this paper. Low income is defined as less than 185% of the Federal Poverty Level (FPL). The FPL is a poverty threshold that the government uses to determine if a family is in poverty based on their total income before taxes.

- Population more than 1/2 mile from a grocery store (*Population_Half_Percent*)
- Low income population more than 1/2 mile from a grocery store (*Low_Income_Half_Percent*)
- Population that does not own a vehicle and is more than 1/2 mile from a grocery store (*Vehicle_Half_Percent*)
- Population more than 1 mile from a grocery store (*Population_1_Percent*)
- Low income population more than 1 mile from a grocery store (*Low_Income_1_Percent*)
- Population that does not own a vehicle and is more than 1 mile from a grocery store (*Vehicle_1_Percent*)
- Adults that consume fruit less than once daily (*Low_Fruit_Consump_Pct*)
- Adults that consume vegetables less than once daily (*Low_Veggie_Consump_Pct*)

2.2 Data Sourcing and Cleaning

Our data is taken primarily from the 2023 USDA Food Access Research Atlas [7], the 2020 USDA Food Environment Research Atlas [6], the 2023 CDC Behavioral Risk Factor Surveillance System (BRFSS) [4], and the California Department of Public Health Community Obesity Profiles [3].

We dropped rows with missing values, dropped unnecessary columns/fields, standardized variable names, standardized units, and sampled based on scope (state vs. county).

More precisely, for the latter two: we standardized units so that all population variables were percentages, and all store variables were per 10,000 people. Fields were averaged or summed across smaller regions to produce data points for larger regions. For example, the Food Access dataset provided data by census tract, so fields had to be averaged or summed across each tract in each county or state. Urban percentage in particular was averaged across the binary assignment for each tract (1 for urban and 0 for rural).

After processing the individual datasets, we merged them into a single comprehensive dataset.

2.3 Feature Engineering

We engineered three new features by averaging existing data points. Food deserts, swamps, and heavens are defined as follows:

- deserts: low access to grocery stores and/or fresh food
  - average of Vehicle_Half_Percent, Population_Half_Percent, Low_Income_Half_Percent, Population_1_Percent, Low_Income_1_Percent, and Vehicle_1_Percent
- swamps: high access to fast food
  - average of Convenience stores, 2016, and Fast-food restaurants, 2016
- heavens: high access to fresh food
  - average of Grocery stores, 2016, Farmers' markets, 2018, Low_Fruit_Consump_Pct, and Low_Veggie_Consump_Pct.

We tested adding different weights to the variables that compose the food scores. In particular, we tried using the results from our LASSO regression as weights. This resulted in very high correlation between food scores and obesity. However, there were also signs of overfitting, so we chose to determine the food scores by evenly weighting the variables.

It is also important to note that high consumption of a food by an individual does not necessarily mean that individual is in a high-access area for that food, and vice versa. For example, a low-income individual might not eat many fruits and vegetables due to their cost and low caloric value, regardless of if they live in a food heaven. An individual might enjoy fast food regardless of if there is an abundance of restaurants in their immediate vicinity. However, in order to extract actionable insights given the available data, we somewhat simplify the problem.

3 Exploratory Data Analysis

3.1 Correlations between Variables

In Figure 1, we present a correlation matrix that includes various socio-economic and health-related variables, with a focus on understanding their relationships with obesity rates.

The correlation matrix also reveals several significant insights and the factors that had the strongest correlation with obesity are:

- Median family income
- Population more than 1/2 mile from a grocery store
- Low income population more than 1/2 mile from a grocery store
- Population that does not own a vehicle and is more than 1/2 mile from a grocery store
- Low income population more than 1 mile from a grocery store

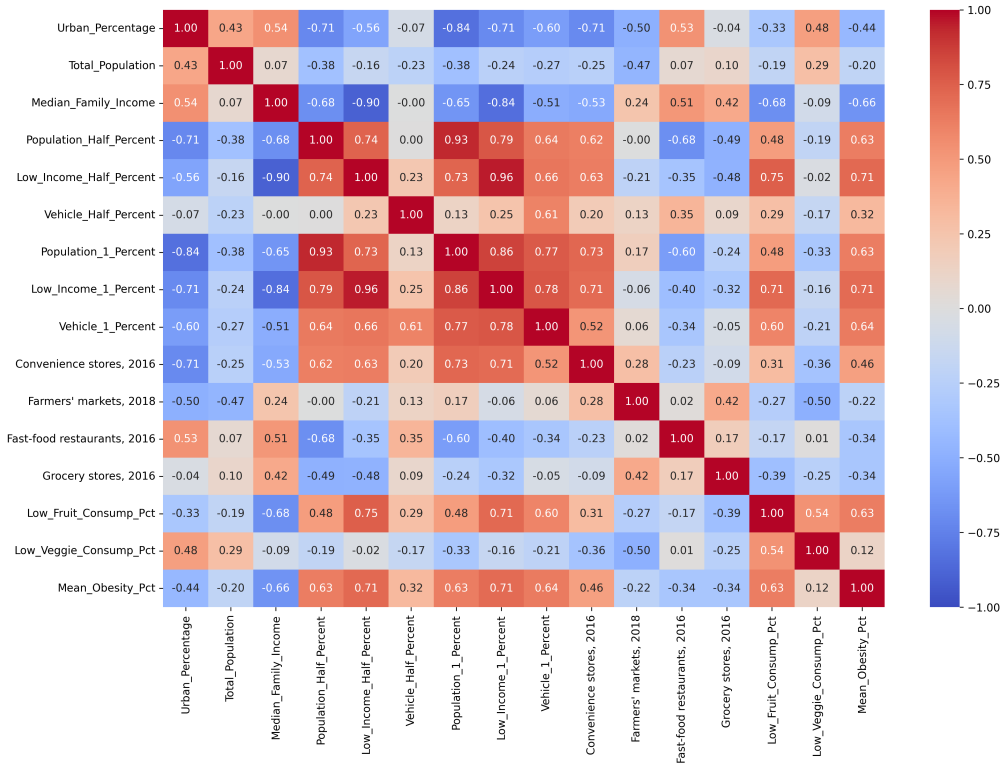

Figure 1: Correlation between all variables

This gives us a general idea of the factors that are most predictive of obesity. We plot the top five correlations in Figure 2.

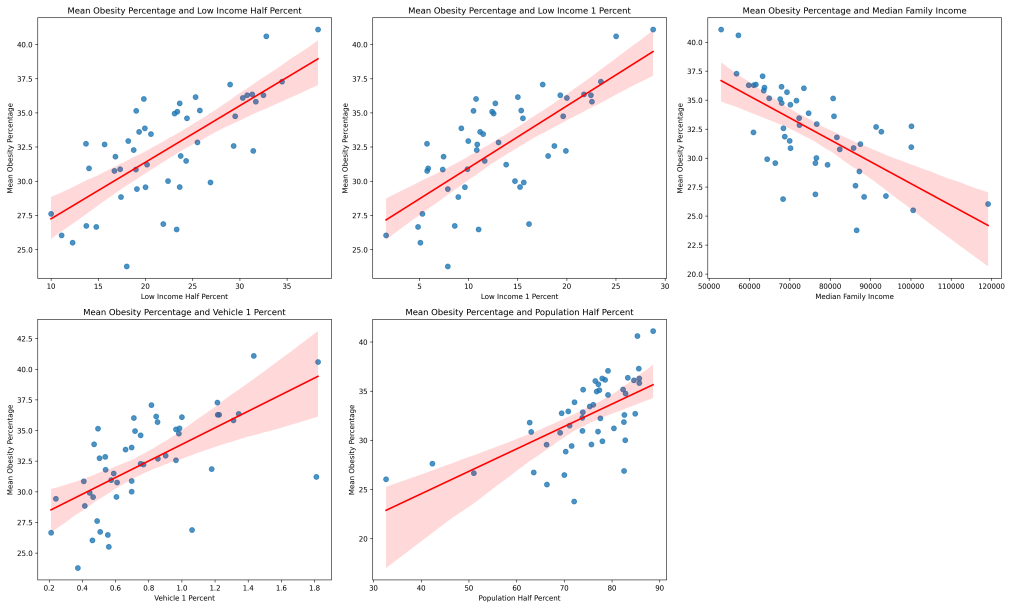

Figure 2: Top Five Correlations between Variables and Obesity

3.2 LASSO Regression

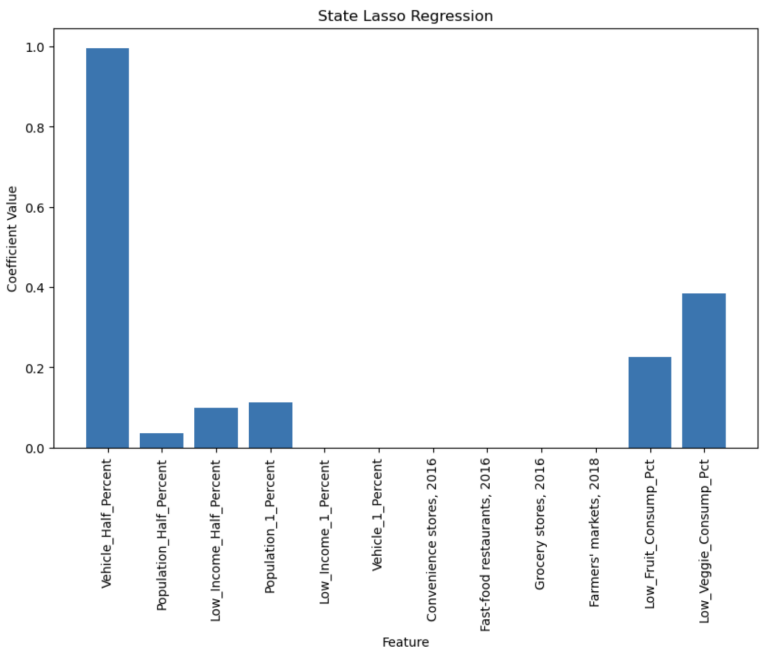

Figure 3: LASSO Regression on obesity rate

To gain even more insight into feature importance, we conducted LASSO regression using obesity as the dependent variable. LASSO regression can aid us in creating a **sparse model that highlights the most important predictors**. Depending on the regularization parameter α, LASSO regression might result in zeroed coefficients for some variables, indicating that they are less important. This is beneficial for understanding which features are driving the predictions.

We chose to omit median family income as a factor because it dominated the model, causing all other coefficients to go to 0. Thus, in order to gain more insight on other variables, we eliminated it. Then, LASSO regression was conducted with a regularization factor of $\alpha = 0.18$, optimized through 3 trials of 10-fold cross-validation to minimize

error. The most important features and their coefficients are shown in Figure 3.

It should be noted that the scaling of the four store statistics (convenience, fast-food, grocery, farmers' market) to per 10,000 people might affect their disappearance in LASSO regression. All of the other variables are scaled to be percentages, and are distributed around the center of 0 and 100. However, in the correlation plot, they each have a lower magnitude correlation with obesity rate than other features, so the scaling may not be a confounding factor.

Interpretation: It is logical that all populations more than a half mile away from a grocery store, **especially** the population of individuals without an available vehicle, would be important predictors of low food access and thus obesity. It is unclear why the magnitude of the coefficient on this vehicle variable is so high, especially when low income and without-vehicle populations more than a mile away do not show up at all. It is also logical that low fruit and vegetable consumption would directly predict obesity, as these are more direct metrics than e.g., the number of grocery stores. (The latter would typically lead to higher produce consumption.)

3.3 Feature Analysis

We created the aforementioned features food desert, food swamp, and food heaven score and calculated each score by state. Figure 4 shows a strong correlation between food desert score and obesity, a moderate correlation between food swamp score and obesity, and a moderate inverse correlation between food heaven score and obesity.

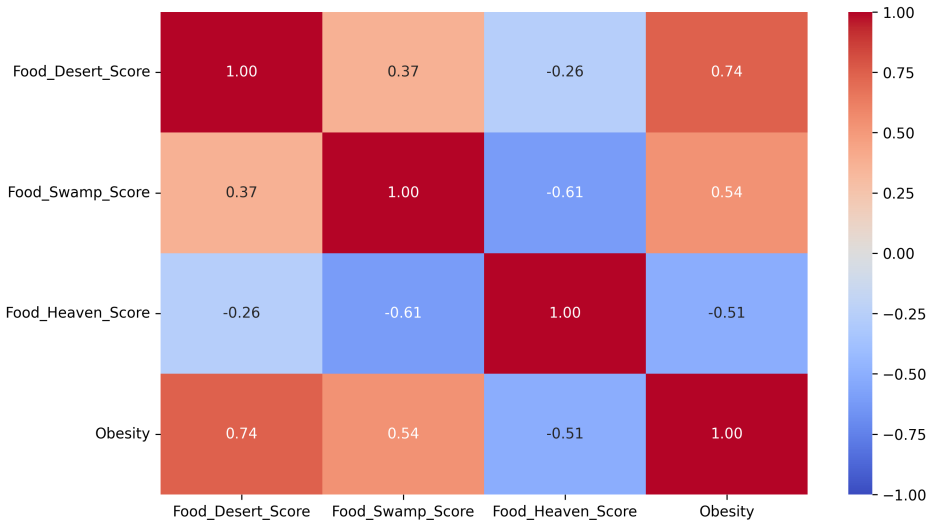

Figure 4: Correlation Matrix of Obesity and Food Scores by State

The correlation matrix in Figure 4 presents the pairwise correlations between obesity rates and three food environment scores: Food Desert Score, Food Swamp Score, and Food Heaven Score.

- **Food Desert Score**: Strong positive correlation with obesity (0.74), indicating that states with higher food desert scores tend to have higher obesity rates.
- **Food Swamp Score**: Moderate positive correlation with obesity (0.54), suggesting that states with higher food swamp scores also have higher obesity rates.
- **Food Heaven Score**: Moderate negative correlation with obesity (-0.51), indicating that states with higher food heaven scores tend to have lower obesity rates.

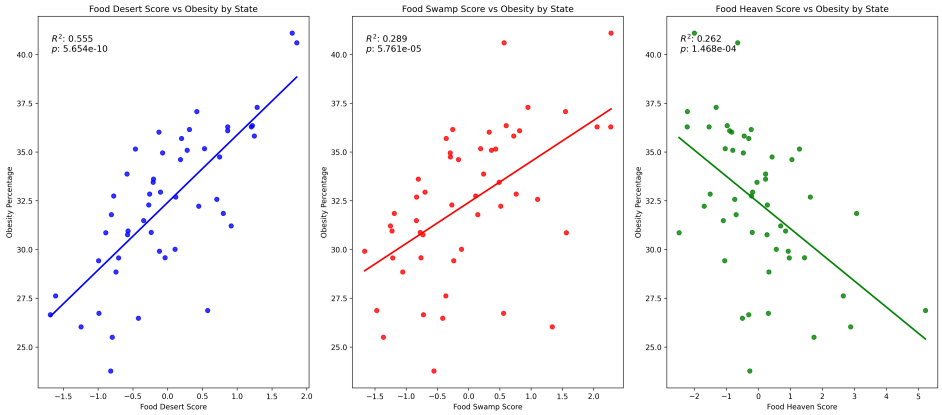

Figure 5: Scatterplot of Obesity and Food Scores by State

- **Food Desert Score vs Obesity**:
  - Strong positive relationship ($R^2 = 0.555$, $p = 5.654e-10$). States with higher food desert scores have significantly higher obesity percentages.
  - This scatterplot shows a clear upward trend, indicating a direct association between poor access to stores and higher obesity rates.
- **Food Swamp Score vs Obesity**:
  - Moderate positive relationship ($R^2 = 0.289$, $p = 5.761e-05$). States with higher food swamp scores, indicating greater availability of unhealthy food options, show higher obesity percentages.
  - The upward trend is visible, but with more variance compared to the food desert score.
- **Food Heaven Score vs Obesity**:
  - Moderate negative relationship ($R^2 = 0.262$, $p = 1.468e-04$). States with higher food heaven scores, indicating better access to healthy food options, tend to have lower obesity percentages.
  - The downward trend in the scatterplot indicates that improving access to healthy food options could be an effective strategy to reduce obesity rates.

Lastly, we mapped a geographic visualization of the food environment scores and obesity rates across different states.

- **Food Desert Score by State**:
  - The map shows higher food desert scores concentrated in the southeastern United States, indicating poorer access to healthy food options in these areas.
  - States like Mississippi, Alabama, and Kentucky have notably high food desert scores.
- **Food Swamp Score by State**:
  - The map highlights regions, particularly in the southeastern United States and parts of the Midwest, with high food swamp scores.
  - States like Louisiana, Arkansas, and Mississippi show the highest food swamp scores, reflecting a high availability of unhealthy food options.
- **Food Heaven Score by State**:
  - The map indicates that states in the northeastern United States, such as Vermont, New Hampshire, and Massachusetts, have high food heaven scores, suggesting better access to healthy food options.
  - Conversely, states in the southern and midwestern regions generally have lower food heaven scores.
- **Obesity Rate by State**:
  - The obesity rate map shows higher percentages in the southeastern United States, mirroring the patterns seen in the food desert and food swamp score maps.
  - States like Mississippi, West Virginia, and Alabama exhibit the highest obesity rates.

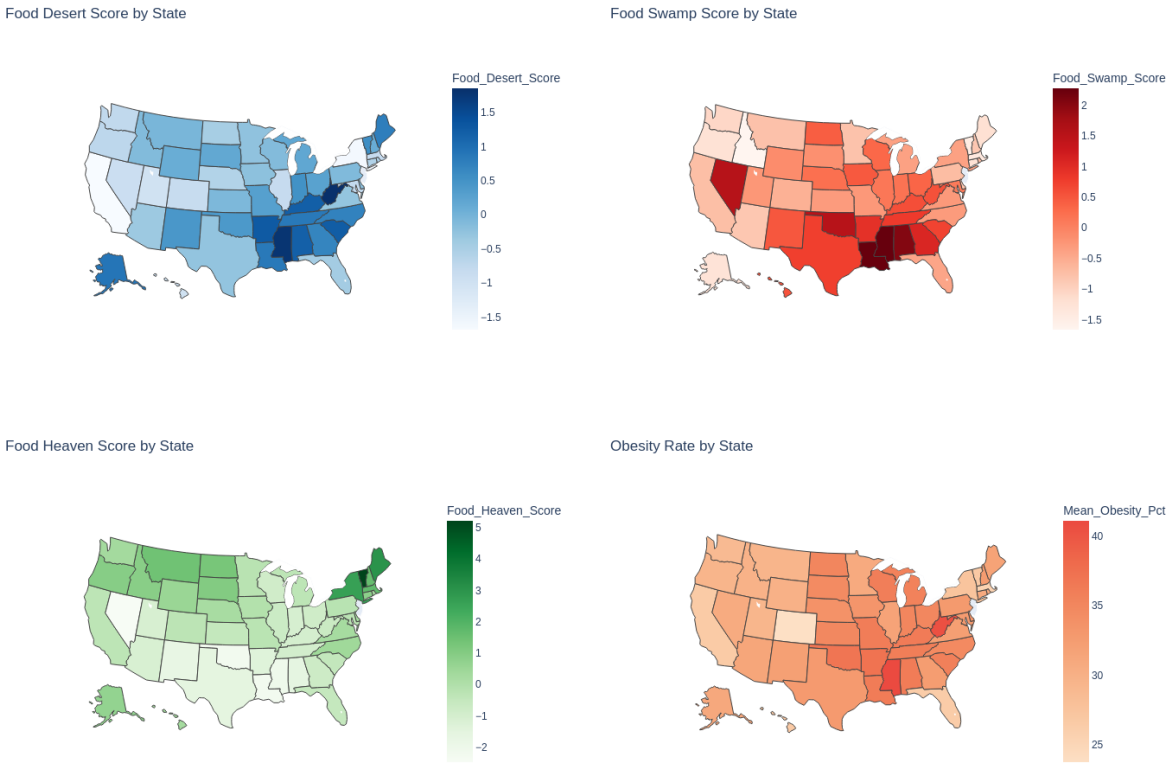

Figure 6: Maps of Obesity Rate and Food Score Rates

3.4 Key Findings

1. **Median Family Income**: Median family income is negatively correlated with obesity rates, indicating the **significant impact of socio-economic factors on obesity**. Family income can also affect other variables such as affordability of fresh food, neighborhood quality, and accessibility to fresh food, which in turn affects the family's obesity and health.
2. **Low-Income Populations**: Low-income populations living far from grocery stores (more than 1/2 mile and 1 mile) have higher obesity rates. This underscores the compounded effect of low income neighborhoods and poor access to healthy food options.
3. **Accessibility**: The LASSO regression model (Figure 3) shows that the variables related to **the distance from grocery stores and vehicle availability are the most critical predictors of obesity.**
4. **Food Environment Scores**: The analysis of food desert, food swamp, and food heaven scores (Figure 6) reveals strong correlations with obesity rates.
5. **Regional Disparities**: South-western parts of United States have especially high amounts of food deserts and food swamps, also showing higher obesity rates. States that are more urban (such as California, New York) contains more food heavens and lower obesity rates. It is interesting to note that a state can be both a food desert and a food heaven, suggesting a more granular county-level analysis is necessary (which will be conducted in the next section). However, most food swamps are not food heavens. Therefore, we hypothesize there are a limited number of stores in an area, and if there are more grocery stores and farmers' markets, then there would be less opportunities for convenience stores and fast food restaurants (and vice verse).
6. **Policy Implications**: The findings suggest that improving access to grocery stores and ensuring vehicle availability could be effective. Furthermore, policies focused on increasing the number of grocery stores in underserved areas and improving transportation infrastructure may help reduce obesity rates.

4 California County Analysis

At this moment, our focus shifted from general citizens to specific area so we could precisely determine which factors are important. We needed a reliable state with a diverse range of backgrounds to analyze as well as up-to-date data.

We needed to find data from a different source, as the nutrition and obesity dataset was not sufficiently granular to provide county-by-county data. The California county data was collected manually from the interactive obesity profiles webpage [3].

From our new source, we were able to gather more variables that were specific to each county.

- Fruit (1+ Serving of Fruit Per Day)
- Fruit Low Income
- Veggie (1+ Serving of Vegetable Per Day)
- Veggie Low Income
- Soda (7+ Serving of Soda Per Week)
- Soda Low Income
- Fast Food (1+ Serving of Fast Food Per Day)
- Fast Food Low Income

The 8 variables above were self-reported by California residents, showing the percentage of people who consume each food. When the variable ends with 'Low Income', it represents the percentage of low income people who consume each food. As before, low income is defined as less than 185% of the Federal Poverty Level (FPL). We should note that these are consumption rates that were self-reported. Thus, it implies these residents have access to fruits, veggies, soda, and fast food, but it does not tell us the amount they have access to, such as the number of fast food restaurants or farmers' markets.

4.1 Correlations between Variables

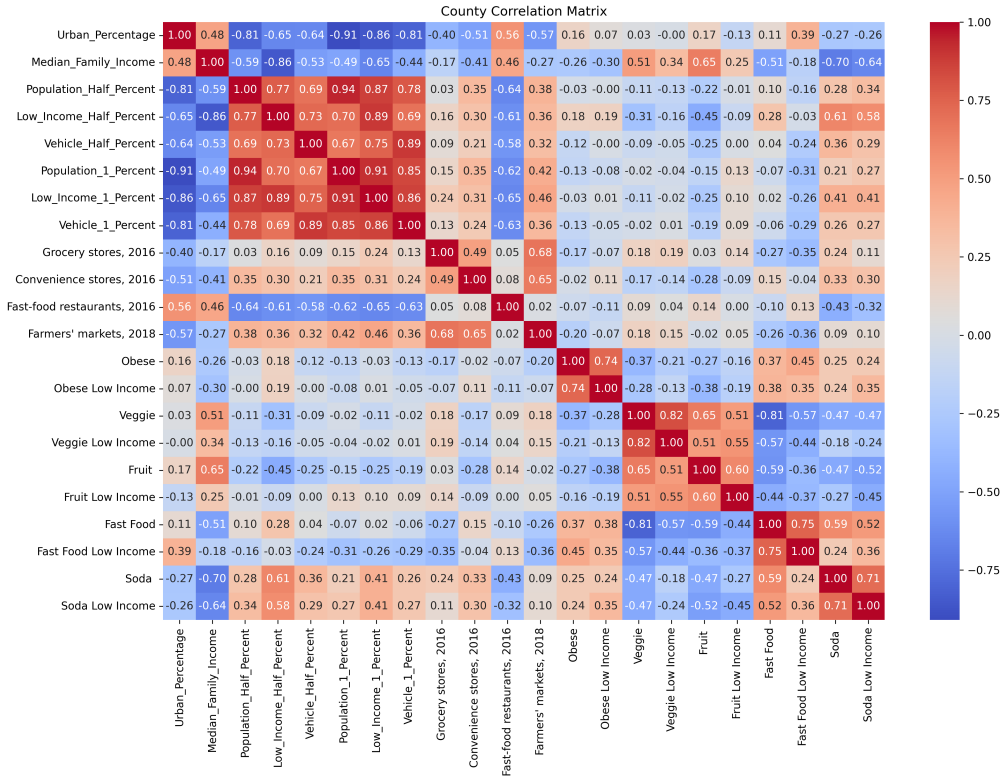

Figure 7: Correlation Matrix of All Variables by County

To start, we followed a similar process as analyzing the nationwide data. Figure 7 shows the correlation matrix of all our variables by county. It is interesting to note how **Median Family Income is the most polarized, either highly positively correlated or highly negatively correlated.** This polarization shows how **income may be a huge factor that affects other variables in our lives, including the amount of fast food we eat or whether we live close to grocery stores.**

Focusing on Obesity, there is an unexpected correlation with the Grocery stores, Convenience stores, Fast-food restaurants, and Farmers' market. Logically, and based on past research [5], convenience stores and fast-food restaurants contain mostly unhealthy foods, while grocery stores and farmers' markets contain a higher concentration of healthy foods. Therefore, we hypothesized the convenience stores and fast-food restaurants should have a positive correlation with obesity. Yet, the correlation is -0.02 and -0.07, respectively, showing **almost no correlation between obesity and convenience stores/fast food restaurants**.

Based on the correlation matrix, we extracted the top 5 correlated variables (Figure 8): Fast Food Low Income, Veggie, Fast Food, Fruit, Median Family Income. Again, we see a negative correlation of Median Family Income and Obesity. We see a positive correlation between fast food and obesity, both generally and for low income. Finally, eating vegetables and fruits are negatively correlated with obesity. In particular, the low income population who consume fast food is the most correlated with obesity. This led to the investigation of whether there is a difference between the general and low income groups.

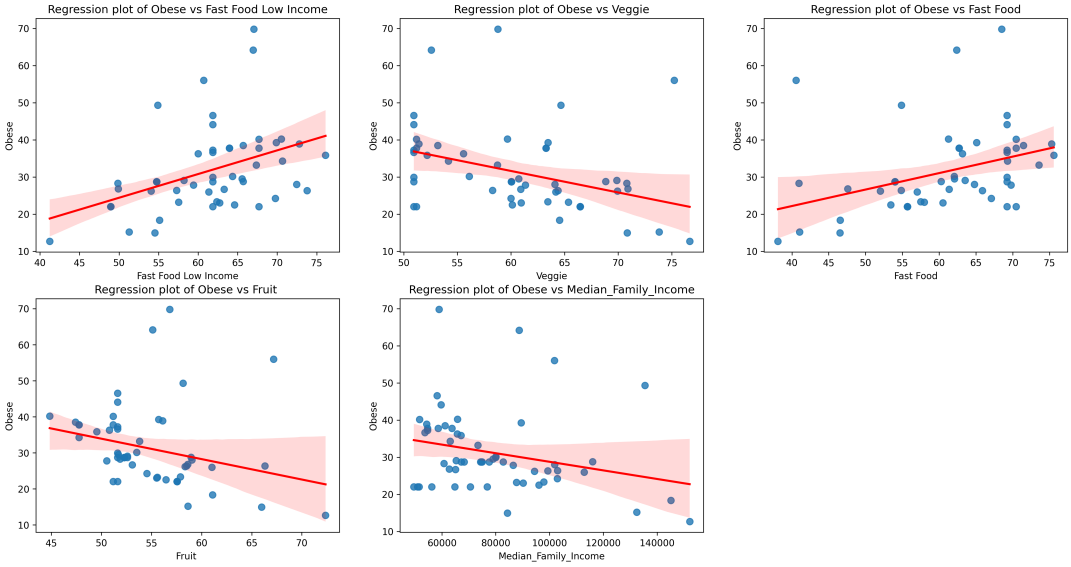

Figure 8: Top 5 Correlations between Variables and Obesity Rates by County

4.2 LASSO Regression

Similar to the state-wise data, we conducted LASSO regression on a county-wise basis with three trials of 10-fold cross validation resulting in $\alpha = 4.54$. The coefficients are shown in Figure 9.

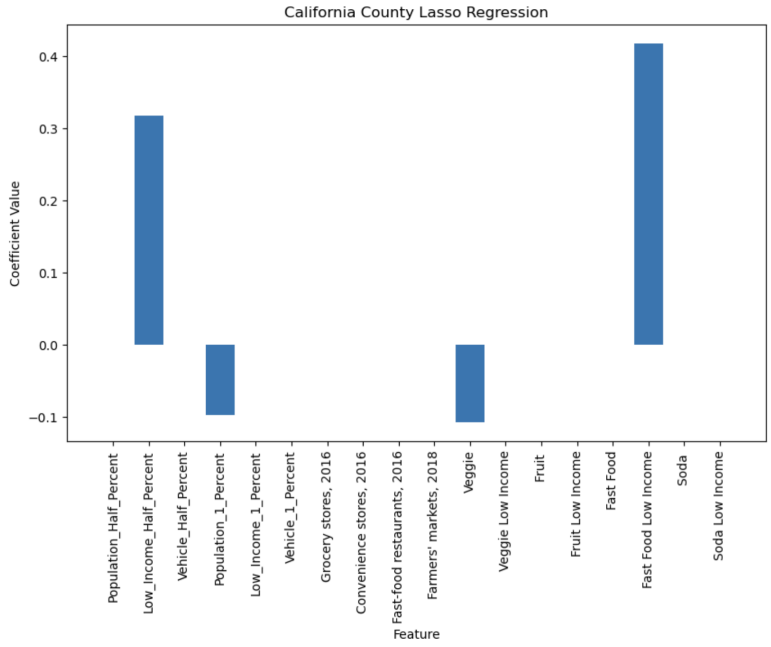

Figure 9: LASSO regression for California counties on obesity rate

Interpretation: It is logical that the low income population more than 1/2 a mile from a grocery store would have a direct relationship with obesity. It is also logical that fast food consumption (especially that of the low income population) and vegetable consumption are directly and inversely related, respectively. It is unclear why population 1 mile away from a grocery store is inversely related with obesity, but this might be due to several factors, such as the specific urban/rural landscape in California or the prevalence of vehicle ownership. Its relationship is also not as significant as other factors.

4.3 General vs Low Income Food Consumption

Since the dataset included both general and low income food consumption, we wanted to analyze the difference and its significance. We focused on finding the t-statistic and p-value of veggie, fruit, fast food, and soda consumption. The key analysis was:

Variable	t-statistic	p-value
Veggie	3.626782	4.335209e-04
Fruit	-0.582747	5.612350e-01
Fast Food	-0.202327	8.400282e-01
Soda	-5.630355	1.350428e-07

Table 1: T-statistic and p-value for food consumption variables in California

- Significant Differences:
  - Veggie Consumption: The general population consumes more veggies compared to the low-income population.
  - Soda Consumption: The low-income population consumes more soda compared to the general population.
- No Significant Differences:
  - Fruit Consumption: No significant difference between the general and low-income populations.
  - Fast Food Consumption: No significant difference between the general and low-income populations.

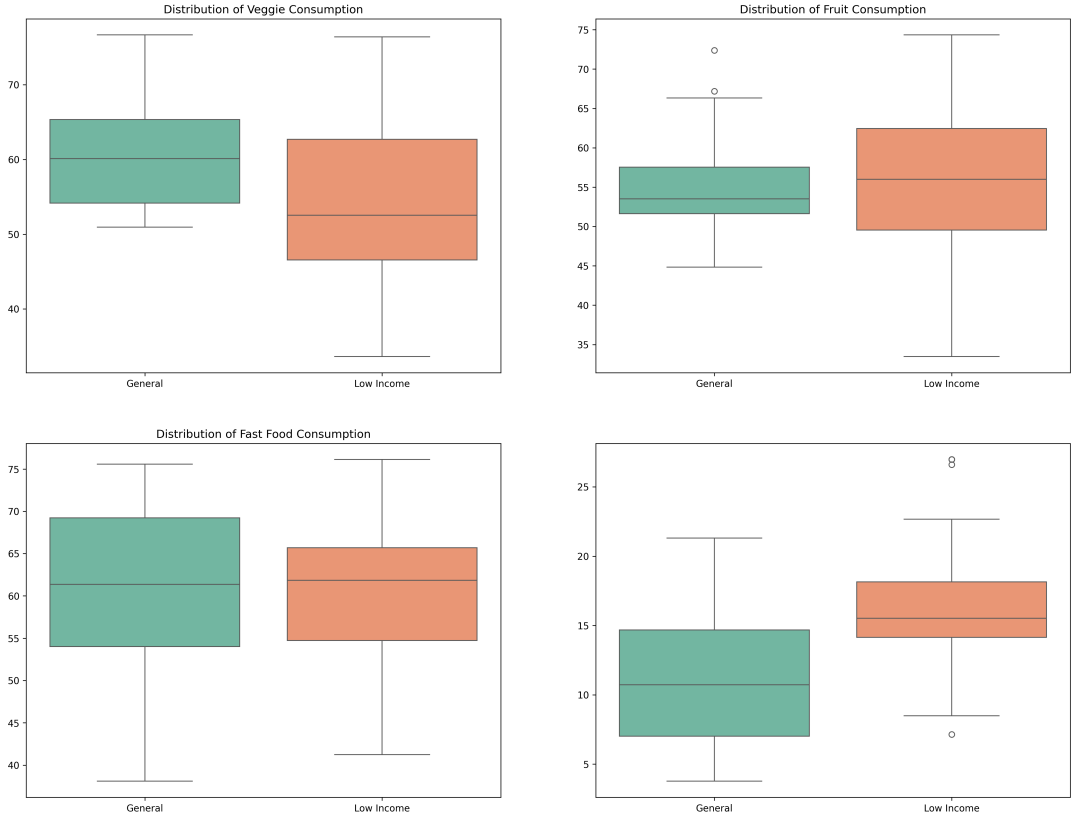

Figure 10: Differences in General and Low Income Consumption by County in California

4.4 Feature Analysis

The California counties are much smaller than states and therefore we hypothesized that our analysis will be more focused and accurate. Specifically, **we hypothesized the correlation between the food scores and obesity would be stronger.**

The correlation matrix (Figure 11) presents the correlation coefficients between various food scores (Food Desert Score, Food Swamp Score, Food Heaven Score) and obesity rates. The color scale ranges from blue (negative correlation) to red (positive correlation), with the intensity indicating the strength of the correlation. Notably, there is a moderate negative correlation between the Food Heaven Score and obesity rates, as reflected in the corresponding scatter plot in Figure 12. The scatter plots illustrate the relationship between the three scores and the obesity rate across various counties in California.

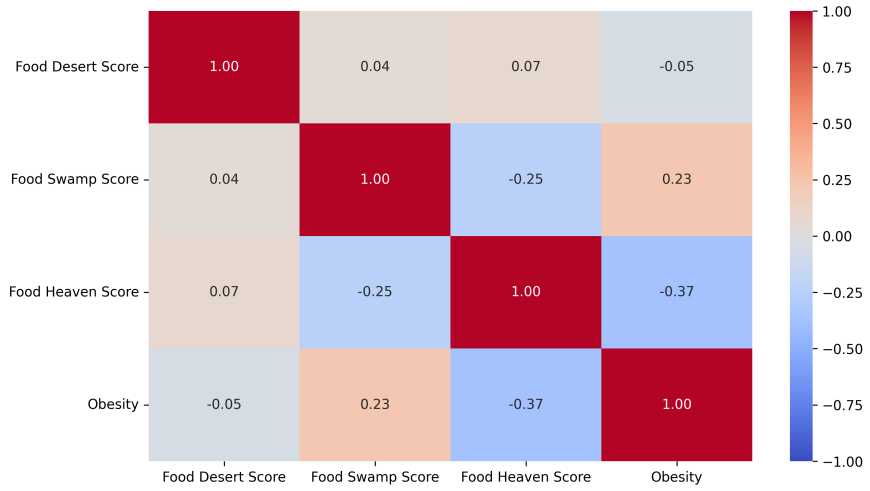

Figure 11: Correlation between Food Scores and Obesity Rates by County

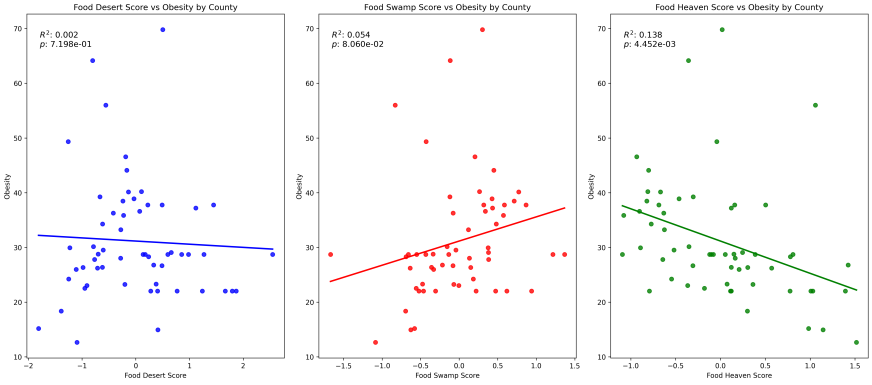

Figure 12: Scatterplot of Food Scores and Obesity Rates by County

The correlations between food scores and obesity were much weaker by county than by state, which can be shown by comparison to our previous analysis. In particular, the food desert score is -0.05, which signifies no correlation and even slight negative correlation.

- Food desert: The regression line shows a very weak negative correlation, with an R^2 value of 0.002 and a p-value of 0.719. This indicates **no significant relationship** between the food desert score and obesity rates.
- Food swamp: The regression line indicates a positive correlation, with an R^2 value of 0.054 and a p-value of 0.081. This suggests a weak positive relationship between the Food Swamp Score and obesity rates with **low significance**.
- Food heaven: There is a negative correlation, with an R^2 value of 0.138 and a p-value of 0.004. This suggests a **moderate negative relationship** and is **statistically significant**.

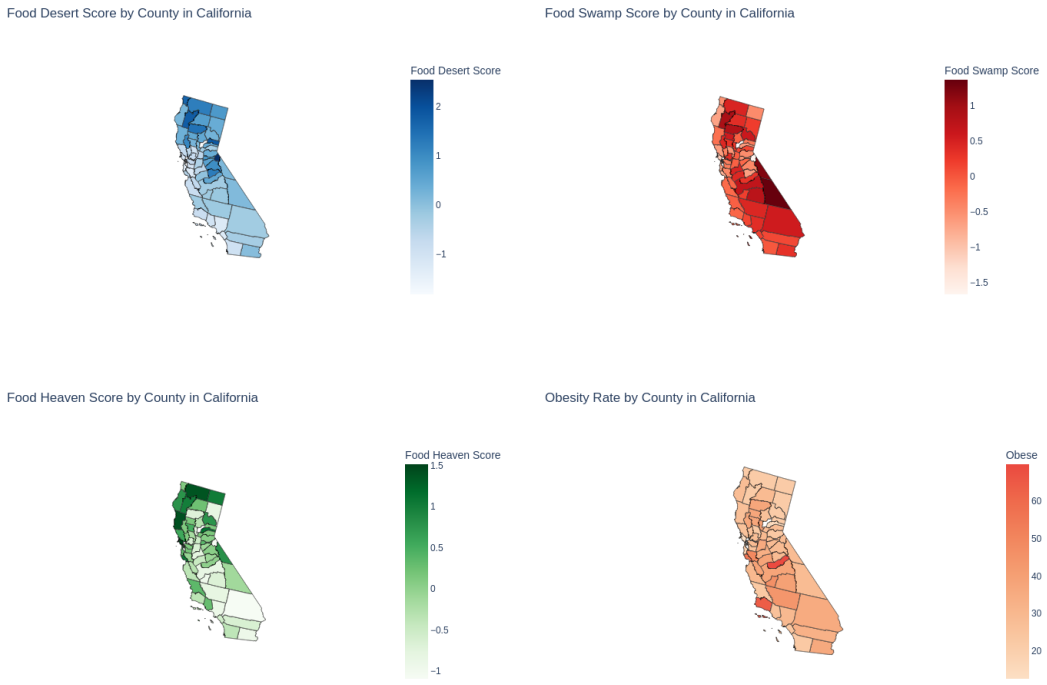

Figure 13: Maps of Obesity Rate and Food Score Rate by County in California

We also visualized these statistics on maps of California. The maps in figure 13 illustrate the Food Swamp Score, Food Heaven Score, Food Desert Score, and Obesity Rate by county in California. The color scale, ranging from light to dark, indicates the severity of each, with darker shades representing a higher score.

Notably, a clear takeaway here is that our hypothesis does not align with our findings, i.e. there is a much weaker correlation between each food score and obesity than we found with the nationwide data. In particular, the food desert score had the lowest correlation with no statistical significance. We have two hypotheses for why this might be the case:

1. We hypothesize that our model **does not sufficiently capture** the state of food deserts in **rural areas**. Low food access in rural areas is typically defined as having no grocery store within 10 miles, not 1 or 1/2 mile.
2. The food desert metric does not sufficiently take into account **income levels and other factors**. All population percentages were averaged together to determine food access, including the overall population, low income population **and** population without vehicles more than 1/2 or 1 mile away from a grocery store. In a rural area, these values will approach general population statistics, where the overall population percentage will approach 100% and dominate the metric.

As an example: In a more affluent rural area, the variables for population 1/2 and 1 mile away from a grocery store might dominate the food desert score, while the vehicle and low income variables remain low. Thus, obesity rates are still low while our calculated food desert score is high. These findings indicate that **obesity is a complex problem with many factors that we were unable to account for in our simplified model and available data**.

4.5 Key Findings

The results showed the following outcomes:

- Low Income Consumption: In relation, the general population consumes significantly more vegetables and significantly less soda than the low-income population.
- Food Heaven: The Food Heaven Score shows a moderate negative correlation to the obesity rate, meaning that higher access to fresh food is related to lower obesity rates.
- The low income population more than 1/2 a mile from a grocery store has a strong relationship with higher obesity. Therefore we would recommend policy-makers to focus on increasing access to stores in low-income neighborhoods and counties.

5 Modeling

In this section, we describe the methodology used to build a predictive model for mean obesity percentage across different states. The goal is to provide a proof of concept demonstrating that the predictors selected as the most meaningful by LASSO (shown in Figure 14.) can indeed be used to build a substantively predictive model. We note that while this model shows promising performance, it should not be used immediately in real-world settings. Instead, it confirms that the features used are predictive, and with some rigorous fine-tuning and more training data, such a model could be effectively deployed.

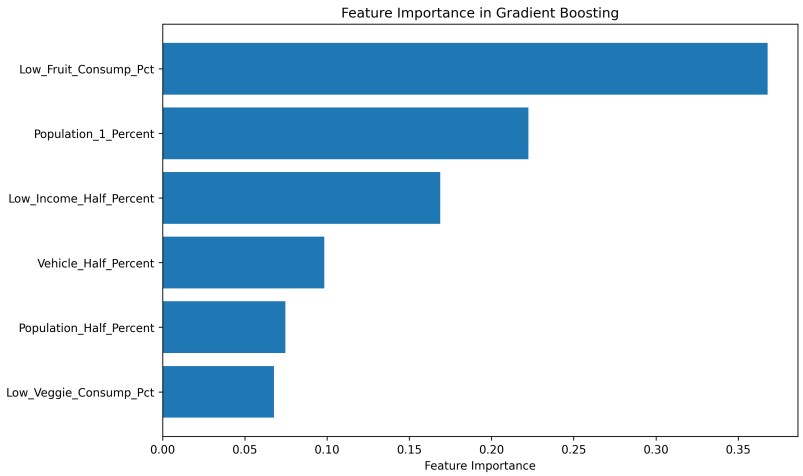

Figure 14: Ranking of feature importance

5.1 Gradient Boosting Regressor

To predict the mean obesity percentage, we used a Gradient Boosting Regressor due to its robustness in handling non-linear relationships and interactions between features. The model yielded a **Mean Squared Error of 1.44 and an R^2 score of 0.72**, indicating that the it was fairly predictive of mean obesity rates. The results again confirm that not only is fruit and vegetable consumption predictive, but access to grocery stores is an equally significant determinant. Populations that live **further away from grocery stores or lack transportation options to access them** are more likely to experience **higher obesity rates**, something particularly evident in low-income populations who might face **additional barriers to accessing healthy food**.

5.2 County Prediction Model

We aim to develop a predictive model for obesity rates in California counties based on various food environment variables. The variables are chosen based on feasibility of implementation, therefore we chose stores and consumption.

Policymaker or the general public could use this model to simulate the impact on obesity if a store was added or eating habits were changed. The variables considered in this study include:

- Convenience stores, 2016
- Fast-food restaurants, 2016
- Fast Food
- Soda
- Grocery stores, 2016
- Farmers' markets, 2018
- Fruit
- Veggie

The target variable is the obesity rate (*Obese*).

The dataset was standardized using the StandardScaler from scikit-learn to ensure all features contributed equally to the model. The data was then split into training and testing sets with an 80-20 split.

We used a similar Gradient Boosting Regressor due to its robustness and ability to handle complex relationships between variables. The model parameters were optimized using Optuna, a hyperparameter optimization framework. The optimization targeted minimizing the Mean Squared Error (MSE) on the test set. The hyperparameters used and their values can be found in the **??**.

The performance of the optimized model was evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics. Feature importance was also analyzed to identify the most significant predictors of obesity rates.

- *Mean Squared Error (MSE)*: 34.19
- R^2 *Score*: 0.21

Feature importance analysis revealed the following contributions:

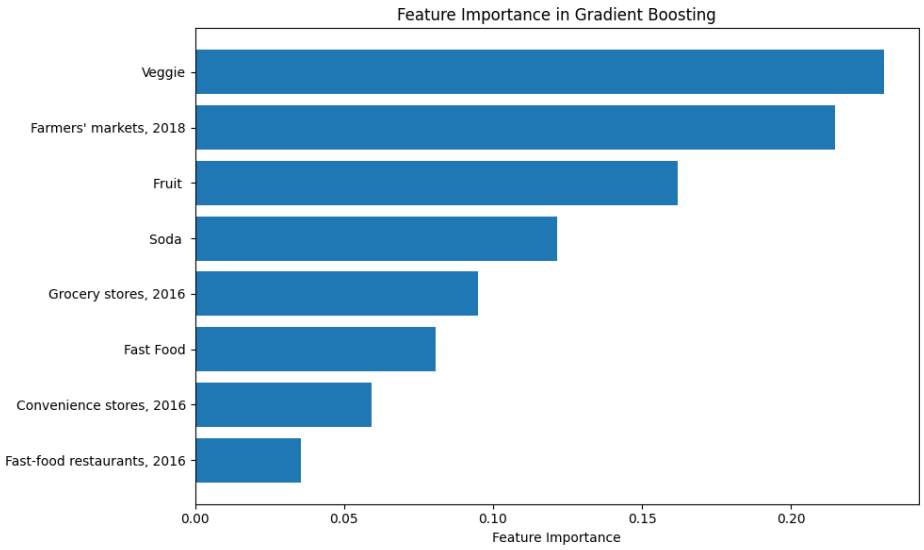

Figure 15: Feature importance for County Predictive Model

The model resulted in an R^2 score of 0.21, indicating that the model did not fit the data well. This could be due to overfitting, but we believe it could be insufficient data variability. We hypothesize the inconsistency in California county data (as stated above in section 4.4) caused the model perform worse. To improve the model's performance, we suggest following steps:

- Additional Data: Incorporate nationwide county enhance model training and capture a wider range of variability.
- Feature Engineering: Create new features or interaction terms that may better capture the underlying relationships in the data.
- Cross-Validation: Use cross-validation more extensively to ensure robust model evaluation.

We will provide the GitHub link here such that our readers can interact and try out the current model.

Following the example usage, you first select a county to observe. Then, you could increase or decrease any of the following variables (Convenience stores, 2016; Fast-food restaurants, 2016; Grocery stores, 2016; Farmers' markets, 2018; Fast Food; Soda; Fruit; Veggie) to see its effect on the obesity rate.

Despite the model's suboptimal performance, the feature importance analysis offers useful insights into the factors influencing obesity. Future work should explore additional variables, alternative modeling techniques, and more extensive data to improve prediction accuracy and provide actionable insights for public health interventions.

6 Conclusion

6.1 Quantitative Conclusions

In this study, we aimed to identify and analyze the primary factors contributing to obesity rates across the United States, with a particular focus on food access indicators and socioeconomic variables. Our analysis highlights significant relationships between these factors and obesity rates at both state and county levels.

Using both exploratory data analysis and machine learning techniques, we have established strong correlations between obesity rates and several key variables, including low access to fresh food, low fruit and vegetable consumption, and high fast food consumption. Specifically, our Gradient Boosting Regressor model demonstrated the importance of these predictors, achieving a Mean Squared Error of 1.44 and an R^2 score of 0.72, indicating substantial predictive power.

Key findings include:

- **Food Desert Score**: States with higher food desert scores exhibit significantly higher obesity rates, highlighting the critical impact of limited access to grocery stores.
- **Food Swamp Score**: There is a moderate positive relationship between food swamp scores and obesity, indicating that areas with high availability of unhealthy food options are more prone to higher obesity rates.
- **Food Heaven Score**: Conversely, higher food heaven scores are associated with lower obesity rates, underscoring the benefits of better access to healthy food options.

At the county level, our analysis of California revealed some discrepancies. While statewide trends were clear, county-level data showed weaker correlations, suggesting additional local factors at play. This indicates the need for more granular and targeted interventions.

6.2 Recommendations

Based on our findings, we recommend a multifaceted approach to address obesity rates effectively:

1. **Enhance Access to Healthy Foods**: Implement subsidies or tax incentives to encourage grocery stores to establish in food deserts. This would directly address the lack of access to fresh and nutritious food.
2. **Promote Healthy Eating Habits**: Public health campaigns should focus on increasing fruit and vegetable consumption, particularly in low-income and rural areas where these habits are less prevalent.
3. **Combat Fast Food Proliferation**: Regulations to limit the density of fast food outlets in vulnerable areas could help reduce the easy availability of unhealthy food options.

4. **Invest in Local Food Systems**: Supporting local farmers' markets and community gardens can increase the availability of fresh produce in underserved areas.

5. **Education and Awareness**: Educational programs aimed at improving dietary habits and understanding the importance of nutrition can have a significant impact, particularly in areas identified as food deserts or food swamps.

Additionally, our study underscores the importance of addressing the socioeconomic determinants of health. Efforts to improve education and economic opportunities can indirectly contribute to better dietary habits and health outcomes.

6.3 Policy Implications

The implications of our findings are clear: targeted, data-driven interventions are essential to combat the obesity epidemic. By focusing on the specific needs and vulnerabilities of different regions and demographics, policymakers can design more effective and sustainable health initiatives.

Federal, state, and local governments, alongside non-governmental organizations, have a crucial role to play in this effort. Tailored strategies that consider the unique challenges faced by different communities will be essential in reducing obesity rates and improving public health.

6.4 Future Work

While our study provides valuable insights, there is room for further research. Future studies could explore additional socioeconomic factors, such as the price of fresh produce versus processed food, which may have profound impacts on food choices. Moreover, developing more comprehensive metrics for food deserts, swamps, and heavens can enhance our understanding of food access disparities.

In conclusion, our analysis highlights the complex interplay of factors contributing to obesity and provides a foundation for targeted interventions. By addressing the root causes of poor food access and promoting healthier eating habits, we can make significant strides in reducing obesity rates and improving overall public health.

References

[1] Beaulac, J., Kristjansson, E., and Cummins, S. (2009). A systematic review of food deserts, 1966-2007. Preventing chronic disease, 6(3), A105.

[2] Bevel, M. S., Tsai, M. H., Parham, A., Andrzejak, S. E., Jones, S., and Moore, J. X. (2023). Association of food deserts and food swamps with obesity-related cancer mortality in the US. *JAMA Oncology*, 9(7), 909–916. https://doi.org/10.1001/jamaoncol.2023.0634

[3] California Department of Public Health, Nutrition and Physical Activity Branch. (2023). California community obesity profiles. Retrieved from https://www.cdph.ca.gov/Programs/CCDPHP/DCDIC/NEOPB/Pages/SNAPEdCountyProfileDashboard.aspx

[4] Centers for Disease Control and Prevention (CDC). (2023). Behavioral Risk Factor Surveillance System (BRFSS), Nutrition Physical Activity and Obesity Data.

[5] Gloria, C. T., and Steinhardt, M. A. (2010). Texas nutrition environment assessment of retail food stores (TxNEA-S): development and evaluation. *Public health nutrition*, 13(11), 1764–1772. https://doi.org/10.1017/S1368980010001588

[6] United States Department of Agriculture (USDA). (2020). Food Environment Research Atlas.

[7] United States Department of Agriculture (USDA). (2023). Food Access Research Atlas.