Desert or Dessert? An Investigation into the Causes of Obesity in the United States

Brayden Zhang[1], Archita Khaire[2], Ryan Wu[3], and Claire Xu[4]

[1]University of Toronto

[2]MIT

[3]Johns Hopkins University

[4]Harvard University

Contents

1. Introduction

1.1 Motivation

As of 2024, nearly 2 in 5 U.S. adults are obese (WOO, 2024), compared to just 1 in 8 adults who are considered obese around the world (WHO, 2024). Obesity is a serious issue, one that can lead to chronic diseases such as diabetes and heart disease. But why is obesity so prevalent in the U.S.? Is it due to our lifestyle? Our love for processed and fast foods? Or is it something else entirely?

A key aspect of this report is the exploration of **"food deserts"**—areas with limited access to affordable and nutritious food. Food deserts are predominantly found in low-income neighborhoods and are characterized by a scarcity of supermarkets and a proliferation of convenience stores and fast-food outlets. In fact, nearly 1 in 6 Americans, live in USDA-classified food desert zones (Rhone, 2017). **We have developed a method to better predict a region's "food desert" level, which we will use to identify areas with higher obesity risks and processed food consumption.**

1.2 Research Questions

Many inter-playing factors have contributed to America's heavy reliance on processed food and skyrocketing obesity rates. A food desert is generally understood to be a geographic area with limited access to affordable and nutritious food. The US Department of Agriculture's Economic Research Service defines regions of low-income and low-access (low-access to grocery stores within a mile radius) as an indicator of a county obtaining food desert status (Dutko, 2021).

We hypothesize that a strong correlation exists between the presence of food deserts and obesity rates. Our goal is to construct a more comprehensive definition of food deserts to be used as a predictor of obesity rates and enable an understanding of the impact of processed foods in America. Particularly, we expand utilize data sources for prediction beyond income and access, including demographics, education, vehicle access, and more.

To help guide our report, and to thoroughly investigate the factors impacting obesity, we aim to answer the following questions:

1. **How do socioeconomic factors and dietary behaviors influence obesity rates across different regions?**

2. **What role does access to healthy food (e.g., food deserts) play in rising obesity rates? Can we predict the probability of a region becoming a food desert to take preventative measures earlier?**

   These questions were chosen based on observed trends in our Exploratory Data Analysis, which indicated that obesity is influenced heavily by location and nutritional access. The motivations behind these questions include understanding who and how is most affected by the rise in obesity across the country.

1.3 Non-Technical Executive Summary

As we analyzed the data on obesity amongst Americans, it was fascinating to find a steady growth in obesity rates over time in the US. Despite the spread of knowledge about the detriments of processed foods, there appears to be limited changes in American lifestyle in this regard. Furthermore, the US has a significantly higher obesity rate compared to the global obesity rate (World Obesity Foundation, 2024). Why is this the case? What factors are contributing most to obesity? How can we use these factors as a predictive measure for obesity?

1.3.1 Key Findings

We found that factors such as **dietary habits, location, education, and income correlated strongly with obesity.** In contrast, other factors such as exercise and meat production/consumption had much less significant of a relation to obesity rates. Diet was the most correlated factor, as the lack of fruit and vegetable consumption was the most highly correlated to obesity in children and adults (see Figure 10).

Due to these findings, we shifted focus toward notable factors such as diet, education, location, and income in an attempt to create a model to predict areas with a higher risk of obesity. We were inspired by the definition of a "food desert" created by the ERS and hypothesized that labeling counties as food deserts would be an effective method to consider the relationship between these underlying factors and obesity rates.

To verify that this research direction had potential, we found that counties labeled as a food desert by the ERS had a greater percent of obese individuals, as shown in Figure 1. Therefore, in this research, **we created our own food desert index** that takes into account additional factors such as location, unemployment, college education, vehicle access, and demographic data, expanding beyond the ERS definition of food deserts. Using this data, we trained a Spatio-Temporal Graph Convolutional Network (STGCN), a deep learning model **to more accurately predict food desert status at the county-level** across the US.

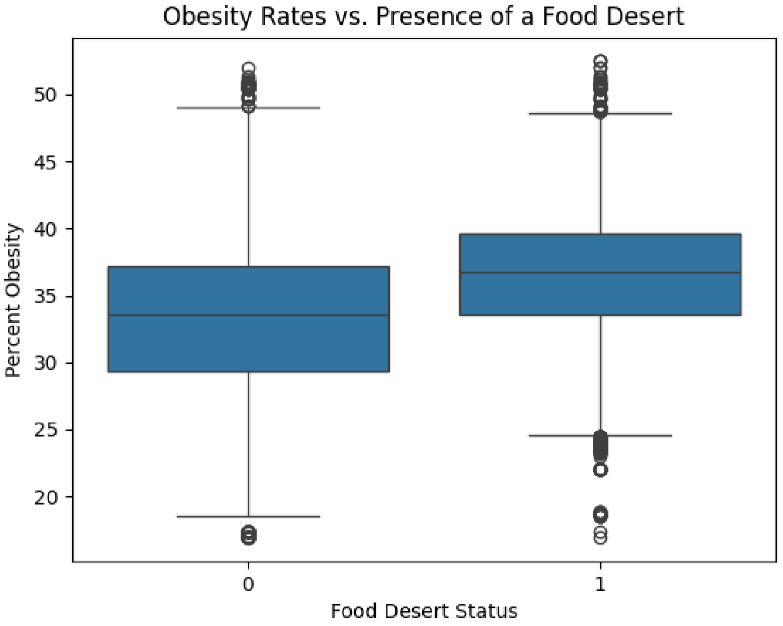


Figure 1: **Obesity Rates vs Food Desert Status (2021 ERS data).**

1.4 Overview of Methods

1.4.1 Datasets Used

1. **Nutrition Physical Activity and Obesity Data**. Most of our initial analysis used the Nutrition Physical Activity and Obesity Data given, especially to determine the behavioral and socioeconomic factors that would impact obesity.

2. **PLACES Local Data for Better Health County Data 2023**. This is a dataset from the PLACES Project completed by the CDC. Due to the large amounts of variation within states, we decided to look further for more specific data on obesity rates within different counties in the US. We used this outside data source that contained information on specific counties to investigate the links between income and obesity as well as other factors (CDC, 2023).

3. **Food Access Research Atlas Data 2010, 2015, 2021**. These datasets were utilized to create labels of food desert status for our model, as they contain information on income and grocery store access at the county level. We also incorporated an additional feature of vehicle access from this data in our model (Dutko, 2021).

4. **USDA Economic Research Service Education and Unemployment Data**. Combining the results of the American Community Service (ACS) estimates (provided to us) and Department of Commerce Population Censuses, the ERS compiled unemployment and education data at the county-level (Dutko, 2021).

5. **National Cancer Institute: Surveillance, Epidemiology, and End Results Program Data**. As our research approach had a large emphasis on county-level

differences, we needed yearly demographic data for every county in the US. This dataset provided race, sex, age, and population information from 1969-2022 (NationalCancerInstitute, 2024).

6. **US Census Bureau County Adjacency**. It was necessary to have information on spatial dependencies for our model to effectively train on county-level data. This dataset enabled creation of an adjacency matrix with information of adjacent counties in the US (USCensusBureau, 2023).

7. **Consumer Data Research Centre UK – E-food Desert Index**. This dataset provided another measure of a multidimensional index for a neighborhood's food desert level. This was useful to serve as an international comparison to our results obtained in U.S. counties (Newing, 2024).

8. **NiceRX Fast Food Restaurants per 100,000 people by state and restaurant chain.** We used this data to explore the relationship between the number of fast food restaurants and the obesity rate in each state (NiceRx, 2023). See Figure 14.

1.4.2 Data Cleaning

In order to handle our Nutrition, Physical Activity, and Obesity Dataset, we performed the following steps:

1. We handled missing values by removing rows with the label "Data not available because sample size is insufficient or data not reported." This because we would not have been able to get a target value, using our Random Forest model, so we wouldn't have been able to make use of those rows anyway.

2. The dataset is filtered to include only the target and selected features, and then pivoted to transform it into a wide format, where each row represents a unique combination of *YearStart* and *LocationDesc*.

3. Feature normalization was performed using *sklearn's 'StandardScaler'*, which scales the features to have zero mean and unit variance. We used this because normalization is crucial for Random Forest models, that may not be sensitive to feature scaling but ensures consistency in the preprocessing pipeline and can improve the performance of models sensitive to feature scales.

4. As for feature engineering, we only used existing features based on relevance to the target variable (obesity percentage), focusing on specific questions related to physical activity and nutrition.

The PLACES Local Data for Better Health County Data 2023 contained multiple data entries for over 2000 counties in the US. However, there were many duplicate entries for each county due to the fact that data was aggregated from specific locations within each county. Thus, each county had multiple entries because of the way the data was entered. To transform this data, we merged rows that contained the same county name and took a weighted mean (weighted by sample size) of the data values of these rows. This produced a much more accurate representation of the obesity rates within each county.

These approaches made it simple to create different models for each state and each socioeconomic category, which allowed to thoroughly investigate the relationship between our input features and target features.

The ERS unemployment dataset required transformation, converting columns contains yearly unemployment rates into two columns of year and unemployment rate. The county names were also properly cleaned and extracted. The ERS education dataset posed a few more difficulties. Due to inconsistent data collection, the data was only provided for the following range of years: 2000, 2008-2012, 2018-2022. In other words, we were forced to make the assumption that education levels were static during the remaining years. In the interest of time and because of a lack of availability of county-level data, we filled in the missing years with education values from the closest available range. This also enabled us to merge with datasets containing other features on the year value. After transformation and cleaning the county values, we were able to obtain yearly county-level estimates of the percentage of adults with college education.

The demographic data from the National Cancer Institute contained a number of categorical values: race, age, and sex, which we one-hot encoded. We also note that all numerical features were normalized. Additionally, all the county values were standardized and converted from their FIPS code (an identification means) to an actual county name using a mapping file.

The Food Access Research Atlas contained data that flagged each county as low-income, low-access, and high vehicle access. Similar to the ERS education data, the data was only collected for the years 2010, 2015, and 2021. We filled in all missing years using data from the closest available year and cleaned the county values.

During the process of merging the demographic, unemployment, education, and food access data, we noticed a discrepancy in the number of data points. Most significantly, since the National Cancer Institute collected demographic data from multiple census tracts for each county, there were many values per county-year combination. To merge properly and limit the data amount, we dropped all counties that did not have data for all the features. The demographic data still dominated in terms of number of records, therefore we randomly sampled 1,000 values per year and then merged. After the complete process of data wrangling, we had compiled a dataset with features and ERS-defined labels for training of our neural network.

1.4.3 Summary of Models

1. **Random Forest Regression Model** to determine the most important predictor variables for obesity prediction.
2. **Spatio-Temporal Graph Convolutional Network** to classify counties according to their food desert status.

2. Exploratory Data Analysis

2.1 Summary

Visualizations: We used scatter plots, bar charts, and time series graphs to explore relationships between obesity rates and demographic factors. For example, a scatter plot of obesity rates versus median household income (Figure 3) helped visualize the negative correlation between income and obesity.

Hypothesis Testing: We tested hypotheses such as whether lower income areas have higher obesity rates and whether educational attainment correlates with lower obesity rates. These hypotheses were informed by initial exploratory results and guided our subsequent analyses.

Synthesized Results: We combined findings from different analyses (e.g., food deserts & obesity rates) to form a comprehensive understanding of obesity's drivers. For instance, the increase in obesity rates among students and adults over time was linked to dietary habits, supporting our hypothesis about the importance of nutrition.

2.2 Obesity Rates

Although much has been made of the obesity "epidemic", we wanted to see whether this issue had been improving, and how it varied for students and adults over time. From Figure 2:

1. The percentage of overweight and obese students is on the rise over the last two decades
2. The percentage of obese adults is increasing at an average rate of about 0.6% per year.
3. The percentage of overweight adults is on a slight decline, but this is more than made up for by the greater increase in obese adults.

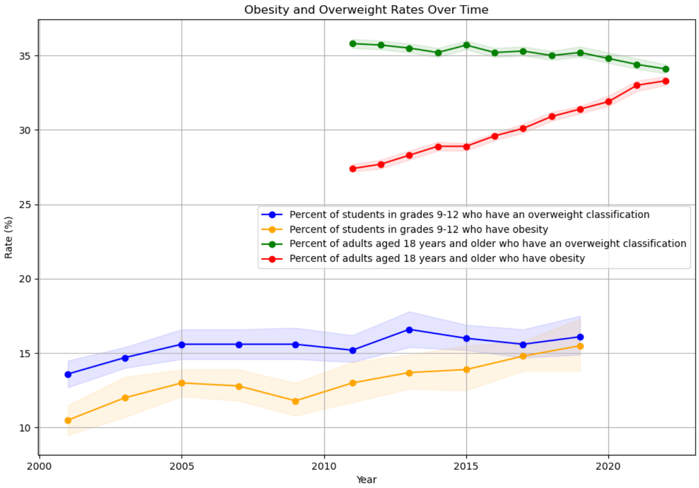


Figure 2: **Obesity and Overweight Classification Rates Over Time Nationally.** The shaded region around each line represents the area between the low-confidence and high-confidence limits.

2.2.1 Obesity Rates by Demographics

In this section, we examine the obesity rates across various demographic categories to identify general trends or patterns related to education, gender, income, and other factors. Understanding these demographic distinctions is crucial for tailoring public health interventions and policies.

1. **Obesity Rates By Median Household Income**: Income level is a significant determinant of obesity, influencing dietary choices, access to healthcare, and opportunities for physical activity. Figure 5 shows that **lower-income populations experience higher obesity rates.** This may be due to limited access to nutritious food and recreational facilities, alongside higher consumption of inexpensive, calorie-dense foods.

2. **Obesity Rates By Education Level**: Education level has been shown to correlate with obesity rates. We analyzed data to observe how obesity prevalence varies among individuals with different educational backgrounds. Generally**, higher education levels are associated with lower obesity rates,** potentially due to increased health literacy and better access to resources promoting a healthy lifestyle.

3. **Obesity Rates by Race/Ethnicity**: Figure 5 shows that different races have drastically different obesity rates over time. We surmised that this could be due to cultural differences as well as disparities in income, education, and location between races.

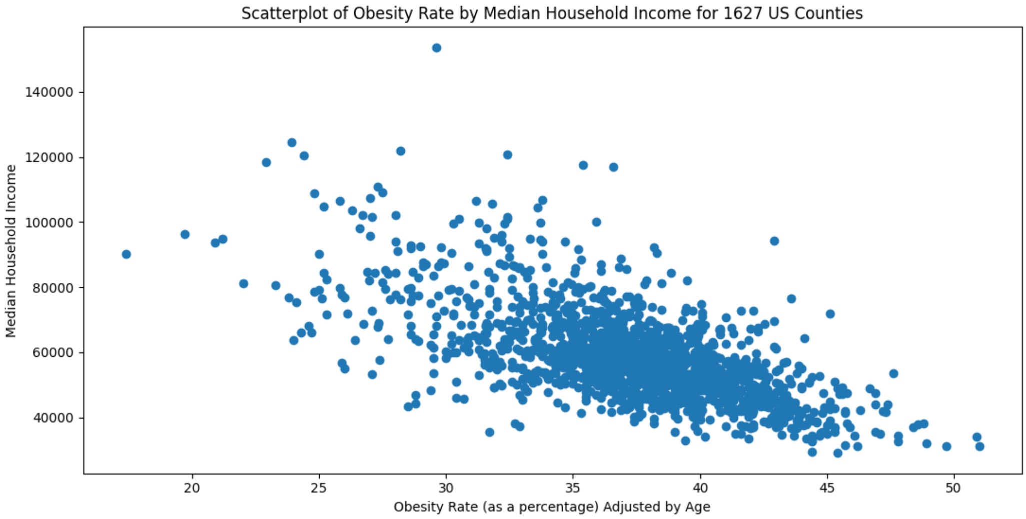


Figure 3: **Scatterplot of Obesity Rate vs. Median Household Income for 1627 U.S. Counties.** The scatter plot shown above illustrates the relationship between income and obesity. Each point plotted is one of 1627 selected US counties, with the x-coordinate containing the obesity rates of adults ages 18 and old, and the y-coordinate containing the median household income within that county. The data used was 2021 data from the PLACES project. There is a slight inverse relationship between household income and obesity, meaning the lower the household income, on average, the higher the rate of obesity. We found that the correlation coefficient is around -0.64, indicating a moderate negative correlation between income and obesity.

2.2.2 Obesity Rates and Food Deserts in 2010

Our research question was motivated through analyses of food desert classifications and obesity rates at the county-level. To do this we collected data using the USDA's ERS definition of a food desert and graphed counties that were a food desert (low income & low access to grocery stores) on a choropleth map, and also the level of obesity in counties; see Figure 6. We used a choropleth map, as it made it extremely easy to visualize patterns across the country. We then merged these two maps, in Figure 7, to show easily visualize the overlap of these two features. These results show us that the overlap is very present, especially at the county level, pushing us to look further into a quantitative relationship between food deserts and obesity rates.

A quick ANOVA analysis (Table 1) confirmed that a statistically significant relationship existed, allowing us to reject the null hypothesis of no relationship between food deserts and obesity rates.

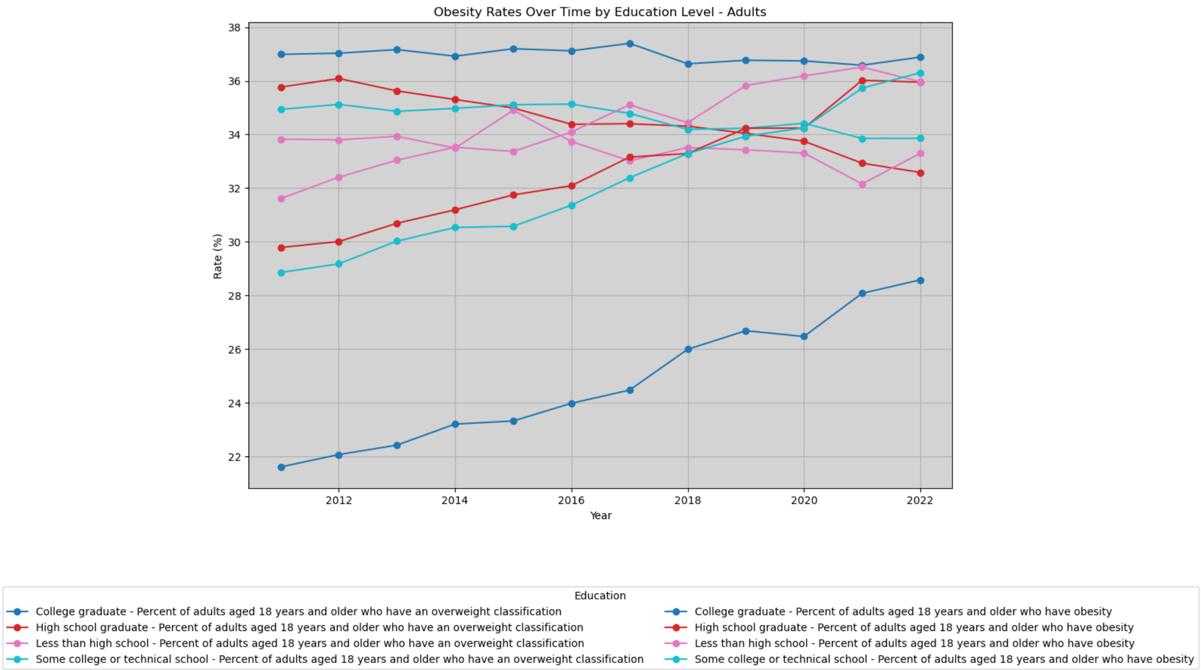


Figure 4: **Graph of Obesity Rates vs. Education Level for Adults.** College graduates are represented by dark blue lines for overweight classification and light blue lines for obesity rates. This group shows the highest overweight rates, while their obesity rates remain comparatively lower. High school graduates are indicated with dark red lines for overweight classification and light red lines for obesity rates, showing relatively stable trends in both overweight and obesity rates. Individuals with less than a high school education are depicted by dark magenta lines for overweight classification and light magenta lines for obesity rates, exhibiting fluctuating trends in both categories. Those with some college or technical school education are represented by dark cyan lines for overweight classification and light cyan lines for obesity rates, showing significant variation over the decade.

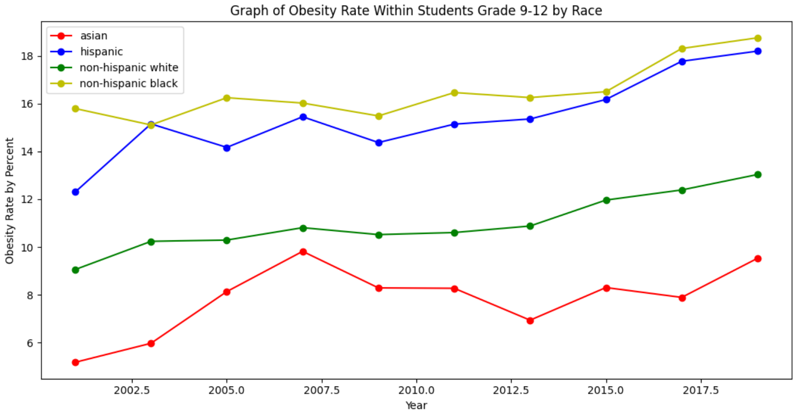
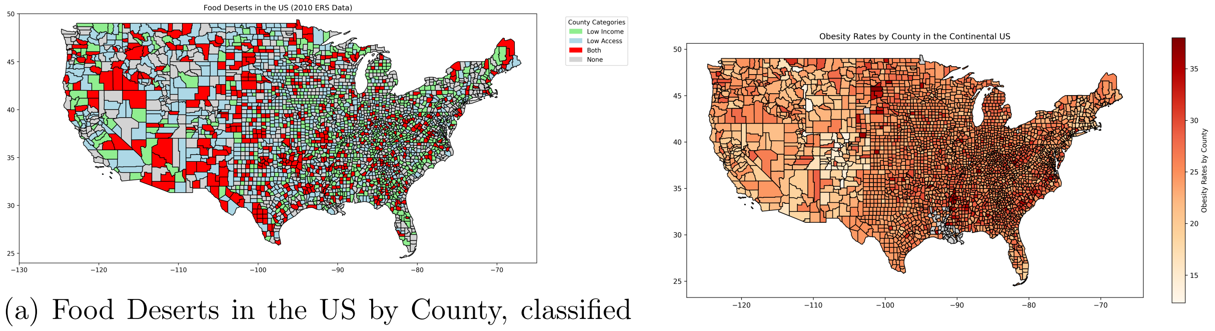


Figure 5: **Graph of Obesity Rates Over Time By Race Asian, Hispanic, Non-Hispanic White, and Non-Hispanic Black students in grades 9-12.** and their respective obesity rates over time are represented in this graph. As seen, there is a large difference between the obesity rates within the different races. This could be due to cultural eating habits, location, and other socioeconomic factors that correlate with race.



(a) Food Deserts in the US by County, classified by Low Income & Low Access

(b) Obesity Rates

Figure 6: **Food Deserts in the USA by County and Obesity Rates by County Mapped.**

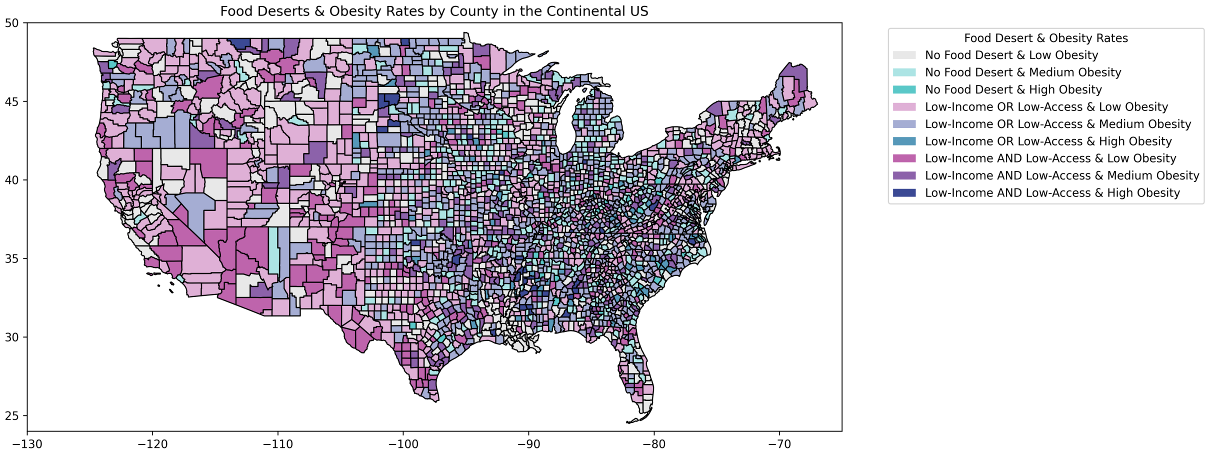

Food Deserts & Obesity Rates by County in the Continental US
Food Desert & Obesity Rates
No Food Desert & Low Obesity
No Food Desert & Medium Obesity
No Food Desert & High Obesity
Low-Income OR Low-Access & Low Obesity
Low-Income OR Low-Access & Medium Obesity
Low-Income OR Low-Access & High Obesity
Low-Income AND Low-Access & Low Obesity
Low-Income AND Low-Access & Medium Obesity
Low-Income AND Low-Access & High Obesity
50
45
40
35
30
25
−130
−120
−110
−100
−90
−80
−70

Figure 7: **Combined Map of Food Deserts & Obesity Rates by County**. The darker colors represent counties with a higher correlation.

Source	**Sum of Squares**	**Degrees of Freedom**	**F-statistic**	**p-value (PR <F)**
Food Desert Status	150,537.6	1	5199.779206	0.01
Residual	3,834,472	132,448	NaN	NaN

Table 1: **ANOVA Table: Analysis of Variance for the Impact of Food Desert Label on Obesity Rates.**

2.3 Trends in Dietary and Exercise Habits

Next, we investigate whether the responses of adults and students have changed over time, looking to see whether there would be similar trends observed of obesity and dietary/exercise habits (such as eating fruits less than once per day).

On a national scale, it can be seen in Figure 8 that the percent of students who consumed fruit less than once daily, and the percentage of students who consumed consumed vegetables less than once per day, increased from about **35% to well over 40%.**

Meanwhile, other questions such as the percentage of students that achieved daily physical education showed relatively constant numbers. This suggested to us that physical education, while important for maintaining a healthy weight for an individual, was not the main cause of the spike in obesity seen across the country.

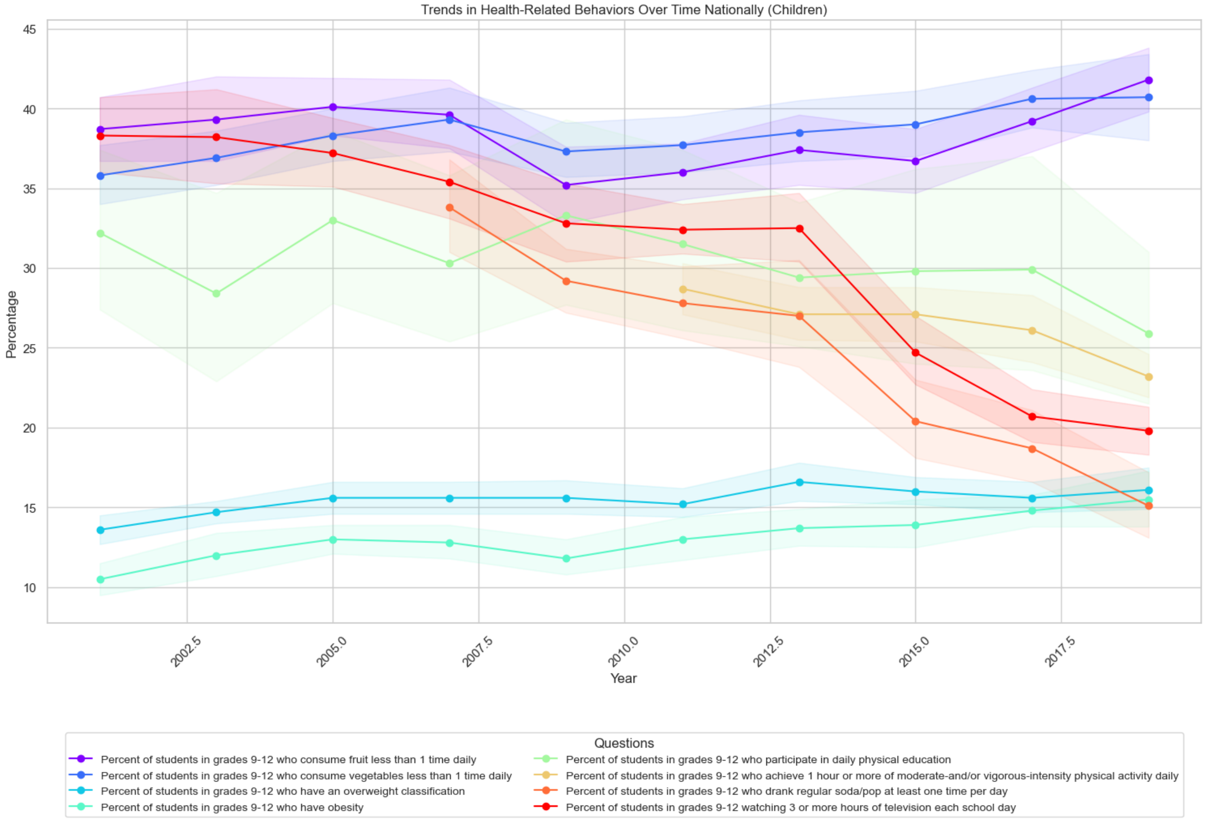
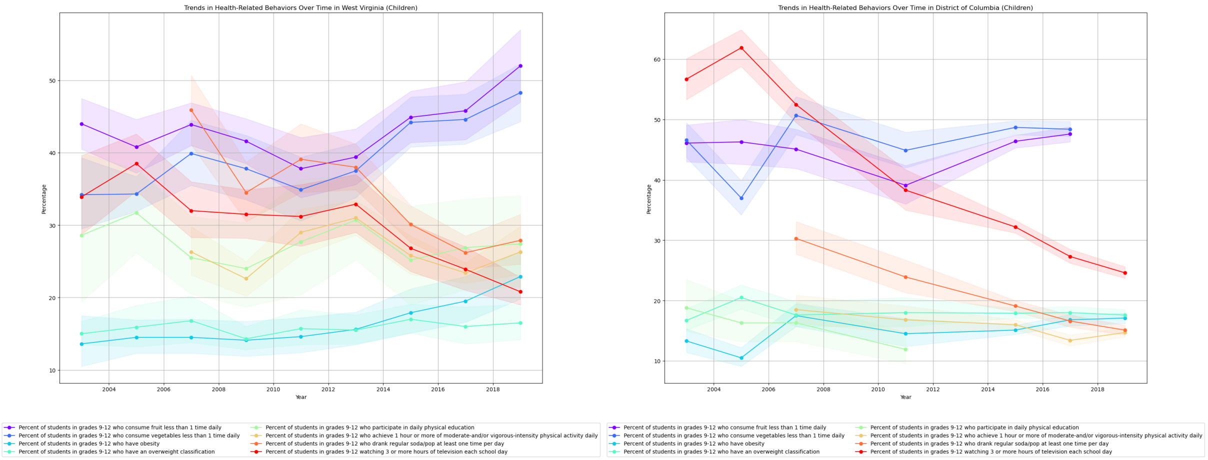


Figure 8: **Trends in Health Behaviors over Time among Students Nationally.**

Zooming in further on West Virginia, the state with the highest obesity rate, we can see that this trend is only magnified, where the number of students that did not eat vegetables or fruits increased at a much higher rate, reaching about 50%.



(a) Trends in Health Behaviors over Time among Students in West Virginia

(b) Trends in Health Behaviors over Time among Students in the District of Columbia

Figure 9: **Trends in the Most Obese State (West Virginia) and the Least Obese U.S. Territory (the District of Columbia).**

Comparing these results to the District of Columbia, the federal district with the lowest obesity rate (and one of the only ones that has not seen a significant uptick in obesity rate), we also observe a relatively constant pattern of nutritional habits.

These state-level results suggest that not only is diet a key factor when analyzing obesity, but also that **these impacts are localized**. This inspired us to look beyond the state-level data, where we found obesity and other socioeconomic data at the county-level for the United States.

3. Modeling & Analysis

3.1 What habits are most correlated to obesity?

In order to get at the root of the levels of obesity seen across in the United States, we decided to build a model that would take in as input features: all nutritional/activity habits found in "Nutrition Physical Activity and Obesity Data" and output the rate of obesity. Features included all behavioral questions in the dataset excluding those directly indicating obesity/overweight to avoid circular reasoning. Specifically, we reasoned that including features like fruit and vegetable consumption would provide a clearer picture of dietary impacts.

We then created and trained a **Random Forest Regression** model using the processed training data, using a 80% / 20% train-test split. We chose this model to control for overfitting and to leverage its ability to handle complex, non-linear relationships between features and the obesity rate.

From this model, we were then able to get the relative importance of each habit, which can be seen in Figure 10. This visualization highlights which dietary and activity-related behaviors have the most significant impact on obesity rates.

Using this model, our Root Mean Squared Error was 2.49 and $R^2 = 0.38$, which was a reasonable starting point, given that we wanted to use this model to determine which behaviors are most correlated with obesity, and not predict the exact obesity percentage itself. This indicates that while our model provides useful insights into the influence of different behaviors on obesity, there is still room for improvement in terms of predictive accuracy.

From figure 10, we can see that **diet is the most correlated factor**. This can be seen by a lack of vegetable and fruit consumption being most highly correlated to obesity for both adults and students. This suggested that perhaps a diet of processed food may be a cause, especially since most processed foods are meat-heavy, and low-fruits/low-vegetables.

Note: we completed the same analysis for adult survey statistics, with similar results, namely that fruit/vegetable consumption are the most important features for predicting obesity rate. Graphs for this can be viewed in the appendix.

Hypothesis Testing:

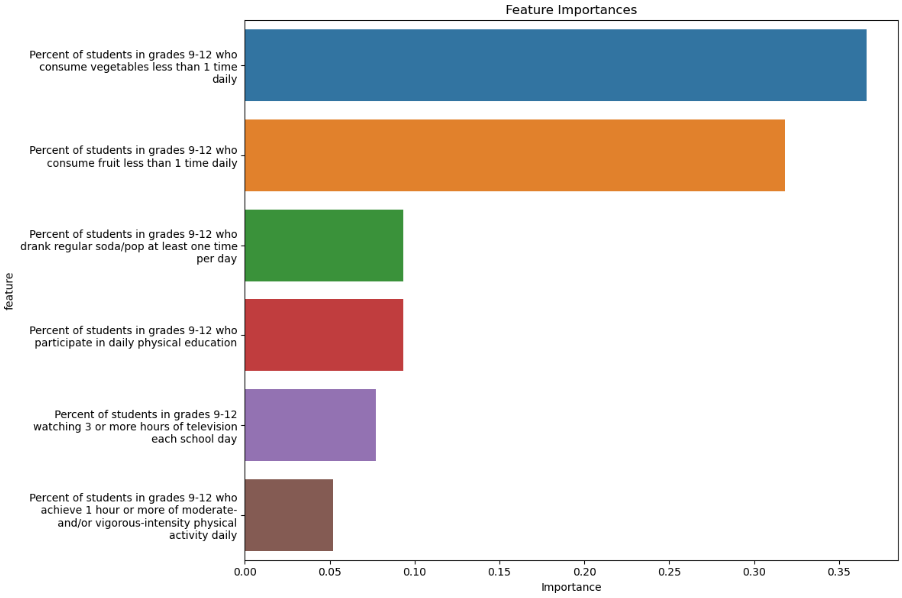
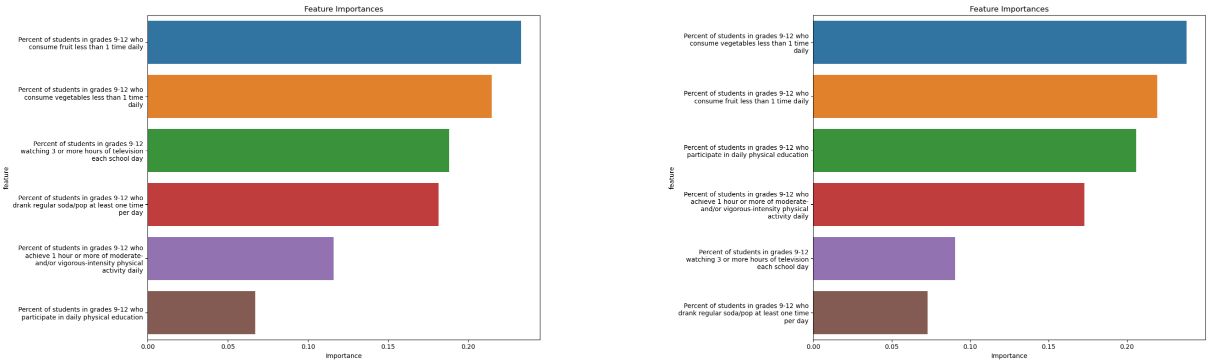


Figure 10: **Relative Importance of Various Health-Related Behaviors and Habits for Students in Grades 9-12**. The features are ranked based on their importance scores, with vegetable and fruit consumption showing the highest importance. Other factors examined include soda consumption, participation in physical education, television watching, and engagement in moderate to vigorous physical activity.



(a) Feature Importances in West Virginia

(b) Feature Importances in Colorado

Figure 11: **Feature Importances for the Most Obese State (West Virginia) and least Obese State (Colorado).**

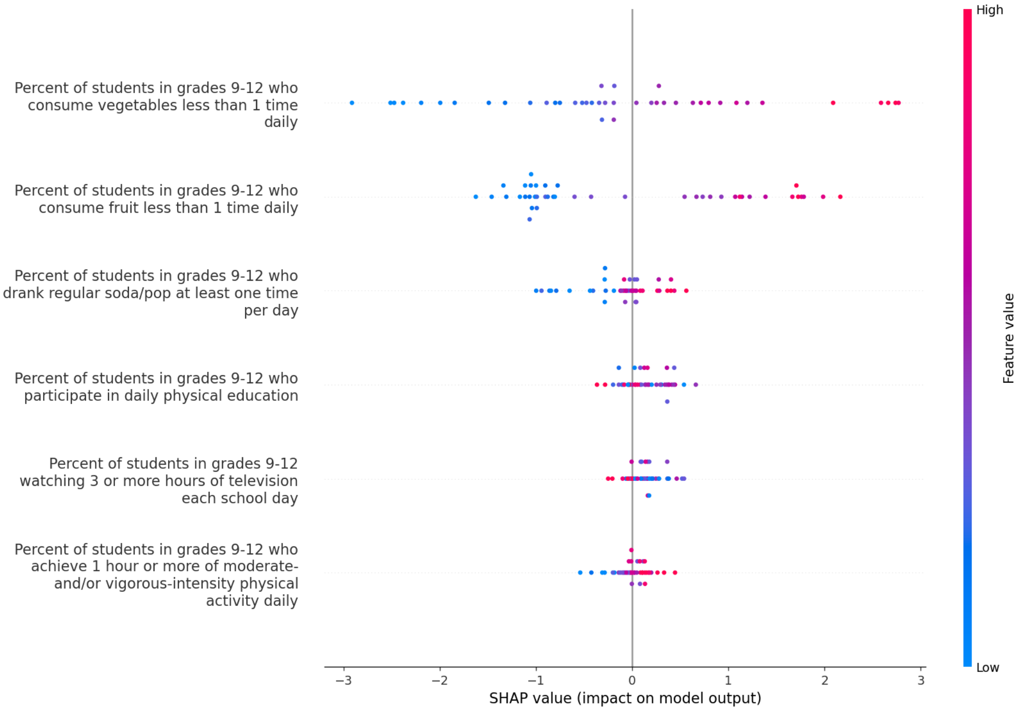


Figure 12: **SHAP Analysis of Lifestyle Factors Influencing Obesity Predictions for High School Students**. This SHAP plot demonstrates the relative importance and directional impact of six lifestyle factors on the model's output for students in grades 9-12. Vegetable and fruit consumption show the strongest effects, with lower consumption generally associated with higher model output; most data points show that lower consumption (blue) is associated with higher SHAP values. Regular soda consumption and participation in physical education display mixed impacts. Time spent watching television and achieving recommended physical activity levels show smaller, more clustered effects on the model predictions.

Using the same training data, we also built a Linear Regression Model to predict obesity percentage. We then obtained coefficients and p-values allows for statistical hypothesis testing. This involves testing whether each feature's coefficient is significantly different from zero, providing rigorous evidence of each feature's impact on obesity percentage.

feature	p value
Percent of students consuming fruit less than 1 time daily	1.01e-08
Percent of students consuming vegetables less than 1 time daily	9.4e-03
Percent of students who drank soda at least 1 time daily	1.45e-02

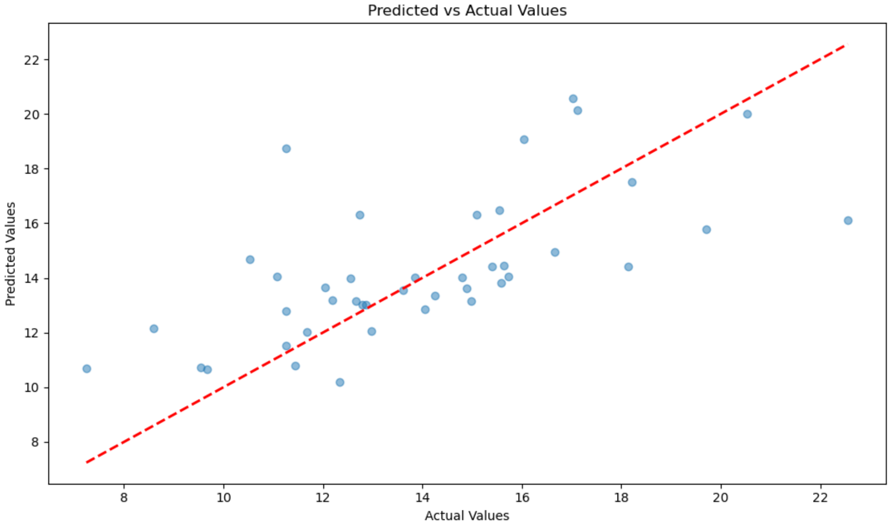


Figure 13: **Results of Linear Regression Model on a Test-Set for Obesity Percentage Prediction.**

3.1.1 Is the number of fast food restaurants a problem?

Interestingly, there was a low correlation between the obesity percentage and the number of fast food establishments in a region, as can be seen in Figure 14. This suggests that it was not so much the presence of fast food but rather the **lack of access (geographical and affordability) to healthy, nutritious foods** from places like grocery stores that was the main drive of obesity.

3.2 Which socioeconomic factors are most correlated to obesity?

Similarly, we conducted an analysis on the relative importance of location and socioeconomic factors on obesity percentage.

1. Location is the most important indicator for obesity. This suggests that spatial factors such as the **availability of healthy food options** plays a significant role

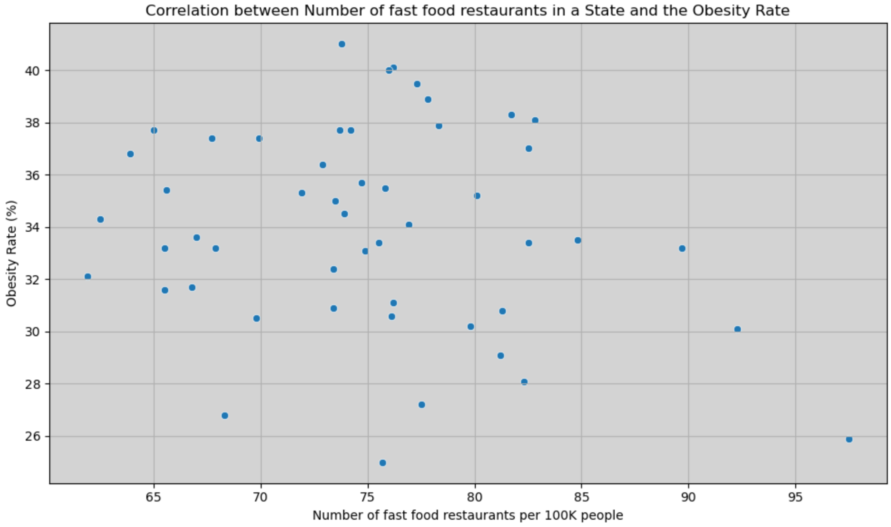


Figure 14: **The Number of Fast Food Restaurants in a State vs. the State's Overall Obesity Rate.** The Pearson's correlation coefficient $= -0.19$, showing a slight negative correlation. This led us to believe that obesity rate was not purely a function of the number of fast food restaurants, and instead look at other factors, such as access to healthy nutritional options.

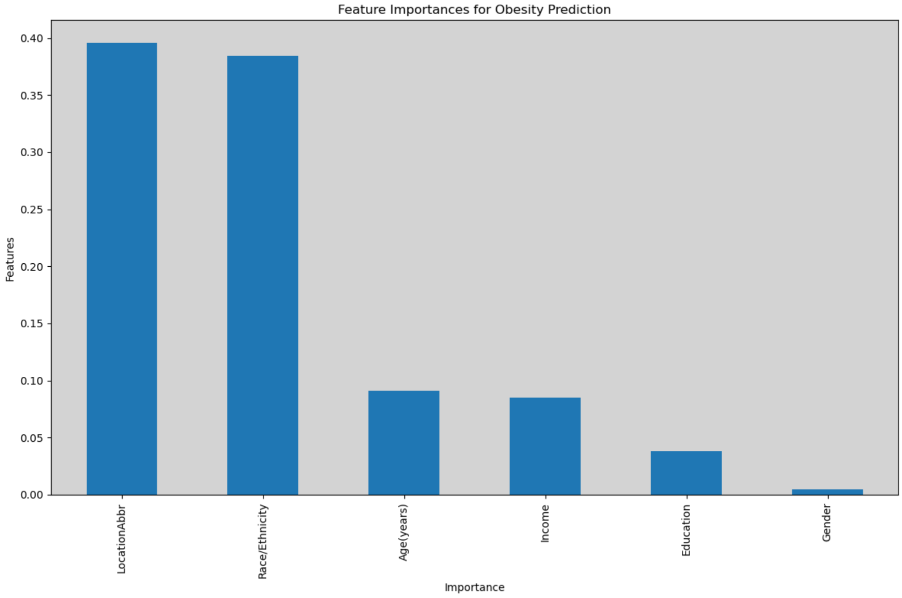


Figure 15: **Most important socioeconomic features for obesity prediction for adults nationally.**

in influencing obesity. This also signalled to us that using more detailed geographic data, including county-level information, would be better understand regional variations and to develop targeted strategies for obesity intervention.

2. Race/Ethnicity is the most important socioeconomic factor for Obesity Prediction. This indicated to us that we needed to look at a county-level, especially due to the fact that Race/Ethnicity percentage can vary wildly due to past geographical segregation and the presence of ethnic enclaves.

3.3 Predicting the Level of "Food Desertness"

3.3.1 Spatio-Temporal Graph Convolutional Network

We trained an spatio-temporal graph convolutional network (STGCN) to classify counties according to their food desert status. As discussed previously, the following 8 features were using for prediction: year, race, sex, age, population, vehicle access flag, unemployment rate, and percentage of adults completing some college. These features were collected for each county and each year from 2010 to 2021, such that we had approximately 12,000 data points.

Label construction was nontrivial, as there is no widely defined standard for what consists of a food desert. Given that our research question was motivation by the ERS definition of low-income and low-access, we removed this data from our set of features and used them to construct the label. Counties that were positive for both the low-income flag and low-access flag were classified as "high-priority food deserts" and counties that were positive for one flag were classified as "moderate food deserts". Counties that had neither flag were given the label of "not a food desert". We found that three classification buckets performed better than two, potentially because the model could classify counties with values in the intermediate range to the "moderate" flag rather than forcing them into either extreme.

The primary benefit of training an STGCN is that it learns from spatial dependencies in the data. In other words, if counties that were located closeby had similar feature values, we could feed this information into our model to improve predictions. This was particularly relevant for our dataset, as it relied regions with high obesity are likely to be clustered. We generated an adjacency matrix, using data from the US Census Bureau of which counties are adjacent. This was an undirected graph, meaning we assumed mutual adjacency. Using an 80%/20% train-test split, we generated tensors for the data and adjacency matrix.

Through testing of multiple structures, we found that a simple STGCN structure performed the best (see Figure 16). The data is first passed through two graph convolutional layers, obtained from PyTorch's geometric library. A ReLU activation function is then applied. We used the Adam Optimizer, Cross Entropy Loss function, and ran the training for 100 epochs.

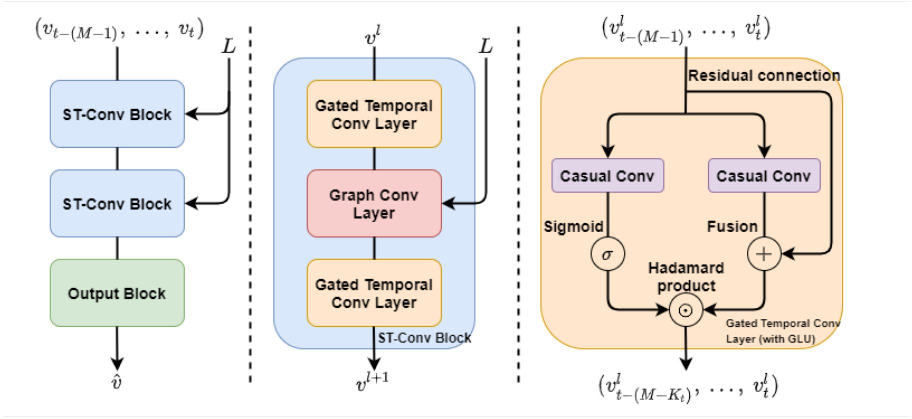


Figure 16: **Architecture of our Spatial-Temporal Graph Convolutional Network.**

Priority was given to the utilization of our chosen features in our model because of the results of our feature importance analysis. The most important factors for obesity prediction were determined to be location and other demographic factors, followed by income and education. Evaluation results can be found Table 2. Although a first glance at the accuracy does not seem promising, we note that it was most likely impacted by the unavailability of yearly data at the county level.

Accuracy	F1-Score	Mean Squared Error (MSE)
0.5517	0.4735	0.5369

Table 2: **STGCN Model Metrics.**

We note that our STGCN provides a better tool for identifying food deserts in comparison to the ERS definition, which only relies on two factors. We capture more intricate patterns and relationships between income, education, and demographics that lead to high obesity rates.

Most importantly, this model can uniquely classify food desert status at the county level, providing clearer insights into variations of obesity at a micro-scale. A lack of such approaches currently exist in literature.

3.3.2 Linear Regression Model

Building a Linear Regression Model between our food desert index and obesity rates, gave a Pearson's coefficient of $r = 0.53$, with $p = 3 \times 10^{-5} < 0.05$, showing that our result was statistically significant.

Specifically, we model the linear regression by the relationship of the form

$$Obesity = \beta_0 + \beta_1 \cdot NormalizedFoodDesertIndex + \epsilon$$

where β_0 is the intercept, β_1 is the *estimate*, and ϵ is the error term.

The choice of linear regression was motivated by its simplicity and interpretability. Linear regression assumes a linear relationship between the predictor and target variables,

Target	Predictor	Estimate	s.e.	t statistic
Obesity	Food deserts	6.6	0.92	7.17

Table 3: Regression results for Obesity in U.S. counties (2013) with a NormalizedFoodDesertIndex (from 0.0 to 1.0) as predictor.

which allows for straightforward interpretation of the coefficients. Additionally, it provides a clear framework for hypothesis testing, enabling us to assess the statistical significance of the relationship.

To validate our findings, we verified that key assumptions of linear regression (linearity, independence, and normality of residuals) were reasonably met. We also performed diagnostic checks, including residual analysis and multicollinearity assessment, to ensure the robustness of our model.

Overall, **the statistically significant positive relationship between the normalized food desert index and obesity rates suggests that living in a food desert is associated with higher obesity rates**. This finding aligns with existing literature that points to limited access to healthy food options as a contributing factor to higher obesity prevalence.

3.3.3 Discussion

The results show that a STGCN is able to learn to predict the food desert level of a particular county, laying the foundation for the development of more complicated predictive models, that can be able to learn a variety of other public health indicators.

As for future improvements, we would conduct a more extensive hyperparameter search, and experiment with adding/changing the neural net layers, to see if model performance could be improved.

Limitation: Since the most recent dataset available was from 2019 (and the census data used in those were from 2010), one limitation of our model is that it missed out on the impacts of COVID-19, where even fewer people could have open access to grocery stores. Additionally, the rise of food delivery services across the U.S. in the years since then may have allowed more people in food deserts to order their groceries online – but there are also inequalities when it comes to accessing these services.

3.3.4 Does this Relationship Hold Internationally?

To explore whether the relationship between food deserts and obesity holds across different diets, we examined England's 2021 E-food Desert Index (EFDI) and compared it with reported child obesity rates. The EFDI takes into account various factors such as public transport accessibility and the availability of online grocery retailers. The addition of online grocery retailers into this index is particularly noteworthy, especially as the use of these services skyrocketed during the COVID-19 Pandemic. This surge in online grocery shopping might explain why the rate of increase of obesity decreased during this time (and overall obesity actually went down in some places).

Despite these complexities, our simple correlation analysis revealed a strong relationship between food deserts and obesity rates, with an R^2 value of 0.76. This finding suggests that, regardless of differences between English and American habits/socioeconomics (diets, physical activity, healthcare), **access to healthy food from grocery stores remains crucial for reducing obesity rates.**

It is also worth noting that the data for the EFDI was collected on the LSOA (Lower Layer Super Output Area) scale, which is much smaller than the county-level data we used in the USA. This finer granularity may contribute to the stronger correlation observed in our analysis.

4. Conclusions

4.1 Limitations and Future Directions

Data Timeliness: Both of our models rely on data before 2020, which does not account for the impacts of the COVID-19 pandemic or subsequent changes in behavior and access to resources. The pandemic has likely altered dietary and exercise habits, potentially affecting obesity trends. Future research would incorporate more recent data to provide a current understanding of these dynamics.

Granularity of Data: Our analysis primarily uses county-level data. While this provides a useful regional perspective, more granular data at smaller geographic units (e.g., ZIP code level) could offer deeper insights into localized variations in obesity rates. We could also do deeper analysis on specific counties that are situated in the largest food deserts, to see if any other patterns could emerge.

Future Directions: Besides experimenting with improving our current model, we could also investigate he economical impact and healthcare impact of having a much higher risk of obesity concentrated in certain regions. For example, we could investigate sales of grocery/food stores in food desert regions, to see what kinds of foods people are eating, and what kind of nutritional benefit/damage that is doing.

4.2 Recommendations

Overall, our analysis has shown that there is a strong predictive correlation between food deserts (among other factors) and obesity percentage in a region. Corporations and government agencies can use this information to:

1. Identify At-Risk Areas: Our food desert prediction model can help identify regions with limited access to healthy food options. By pinpointing these areas, policymakers and public health officials can prioritize resource allocation to the most affected communities.

2. Monitoring and Evaluation: Using our food desert prediction model, we can continuously monitor the effectiveness of implemented interventions. By tracking changes in obesity rates and other health metrics, we can evaluate the success of our strategies and make adjustments as needed.

Through predictive modeling of our food-desert model, we can identify the communities at highest risk for obesity, making meaningful progress in reducing obesity rates and improving public health outcomes in communities across the United States.

Acknowledgments

We thank Citadel and Correlation One for hosting this datathon, providing the datasets, and inviting us to participate.

References

CDC. (2023). https://www.cdc.gov/places/index.html

Dutko. (2021). https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data/

NationalCancerInstitute. (2024). https://seer.cancer.gov/popdata/

Newing, A. (2024). https://data.cdrc.ac.uk/dataset/e-food-desert-index

NiceRx. (2023). https://www.nicerx.com/fast-food-capitals/

Rhone, A. (2017). Low-income and low-supermarket-access census tracts, 2010–2015. *U.S. Department of Agriculture, Economic Research Service.*

USCensusBureau. (2023). https://www.census.gov/geographies/reference-files/time-series/geo/county-adjacency.html

WHO. (2024). Obesity and overweight. *World Health Organization News Room.*

WOO. (2024). Ranking (percent obesity by country). *Global Obesity Observatory.*

Appendix

A. Extra Figures

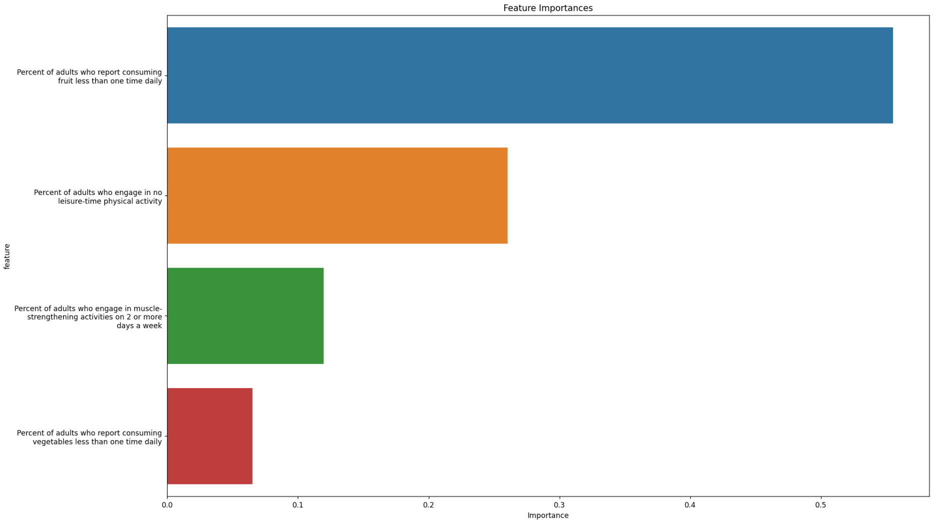


Figure A1: **Most Important Behaviors for Adults that are Predictive of Obesity.**

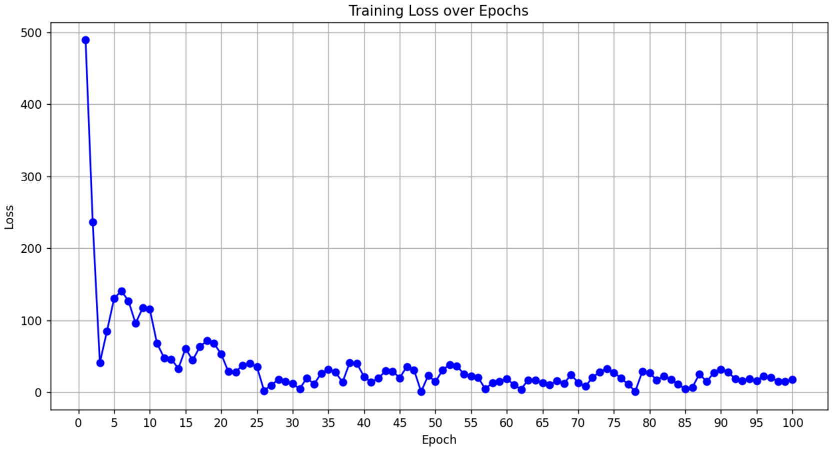


Figure A2: **Training Loss Over Epochs for our STCGN Model.**