Analyzing Obesity Rates in Numerous Demographics in the U.S.

CS418 Final Report

Link to GitHub Repository: CS418Project

Written by: Rasleen Dhaliwal, Kyla Gonzalez, Shimra Hazmi, Fatimah Qadri, and Fariha Siyadath

Problem & Scope

Obesity is an expansive problem that affects over 1 billion adults all over the world, and it affects a staggering 42% of Americans. Obesity can lead to severe health problems such as heart disease, high blood pressure, diabetes, cancers, liver diseases, etc. These concerning statistics encouraged us to look into the problem of obesity. In particular, we researched how different demographics can affect obesity rates for people in the United States over the age of 18. For example, we looked at how a person's income, ethnicity, education, race, education, and state conditions contribute to their chance of obesity. Obesity is a problem that can be avoided, and we wanted to look at how these demographics can contribute to obesity in the United States.

Hypothesis

Our main research question is, how do specific demographic and lifestyle factors influence obesity rates in the United States? To investigate this, our group analyzed variables that we believed to be most strongly associated with obesity, such as income level, education level, race/ethnicity, age, gender, and access to recreational and food resources. Therefore, our hypothesis was that if certain demographic and lifestyle factors, such as age, gender, income level, race, and availability of recreational centers, are associated with higher obesity rates, then individuals from specific demographics will have a higher likelihood of obesity compared to others. By analyzing the most influential variables, we hope to better understand the main causes of obesity disparities and help work towards a future with data driven public health policies to help decrease obesity.

Data Sources & Preparation

To further examine our hypothesis, we used two data sets. The first data set came from the Behavioral Risk Factor Surveillance System (BRFSS). This data set was available in a CSV file, and it contained over 104,000 rows and 33 variables with detailed information. This dataset was last updated in February 2025, and its notable variables include: physical activity, weight status (where the obesity rates are derived from) age, race, sex, income, education, etc. This data set came with some null values. The second dataset came from the USDA Food Environment Atlas. This data was also in a CSV file, and it contained important information about over 280 variables divided in county level data. Notable variables include: recreational facilities per thousand, agrotourism operations, snap usage, farm accessibility, store accessibility, etc. This data set had no null values, but the county level data needed to be converted into state level data in order to merge with the first data set.

Exploratory Data Analysis

In the EDA section of our project, we strived to understand the data on a level that will assist us in gaining some preliminary insight into how we can analyze it and make meaningful solutions/findings. First, we used .info() to gain some statistical information of our dataset such as row and column count, data types, null data counts, and more. This helped us understand the dimensions of our dataset and some high-level information of what we will be working with. This also helped us get an idea of whether we would need any extra datasets to move forward with our project.

Since this is a dataset that stored information about obesity rates amongst individuals in various demographics, we wanted to see what the overall spread of obesity rates would be like throughout the data. Analyzing the obesity rates based on age seemed to be the most

straightforward method as it was the data that could be grouped and thus easier to understand. The best way we could do this was to visualize this data using a histogram plot. We plotted it relative to the Data_Value column which held the percentage of obesity rate for each group. We found that there are about 17, 500 groups in which the obesity rate falls within 30-40%.

Looking at the dataset from another lens, we wanted to explore how obesity rate is affected by factors such as income levels and poverty levels. We plotted this data using scatterplots to see the correlation between the two demographic variables. With income levels, we found that obesity rates are decreased with higher levels of income as we can see with the downward slope. This could be attributed to how individuals with higher incomes are able to purchase food that is organic and healthier which costs more than food that is more affordable like fast food. Next, we analyzed obesity rates in relation to poverty levels amongst people. We see that there is an upwards slope, meaning that as poverty levels increase, so does obesity rates. This solidifies our observation from before where we saw how healthier foods are more accessible to individuals with higher incomes and how individuals with lower incomes do not have this same access, thus increasing the obesity rates in this specific group of people.

Challenges

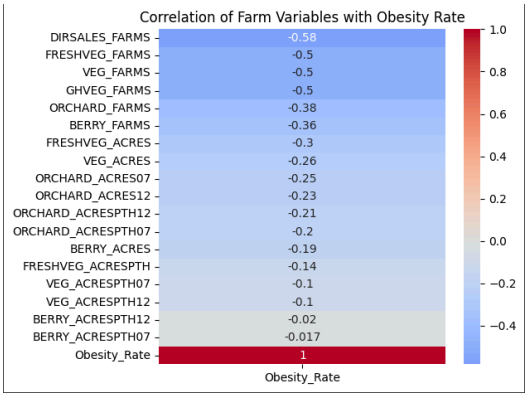
While completing this project, we faced a variety of challenges. One particular challenge that we faced is that we would sometimes encounter repeated data. For example, when we wanted to look at how different taxes correlated with obesity artes, we originally planned to use a general food tax as a baseline compared to soda tax and chips and pretzel taxes, but we found that the graphs for the general food tax and chips and pretzel tax were identical because chips and pretzels are types of food. Thus, we pivoted to just comparing the soda tax to the chips and pretzel tax. In addition to this, we ran into an issue in combining our two datasets. The BRFSS dataset contained information about a variety of demographics, and it only included data from 2017 on. At times, we found that our second data set had values we wanted to look at from 2007 or 2008; and thus, we had to pivot to use a different year. Additionally, if we wanted obesity from a specific demographic from a specific year to match the year from our second data set, we at times found that there would be only about one or two data values for the obesity rate for that demographic, and the rest would be null. We found that it would not make sense to perform a specific data imputation method on this data, as one or two data points would likely not accurately represent a variety of populations across different states. All in all, although we faced a few challenges, our data still provided us with much helpful information to continue our research.

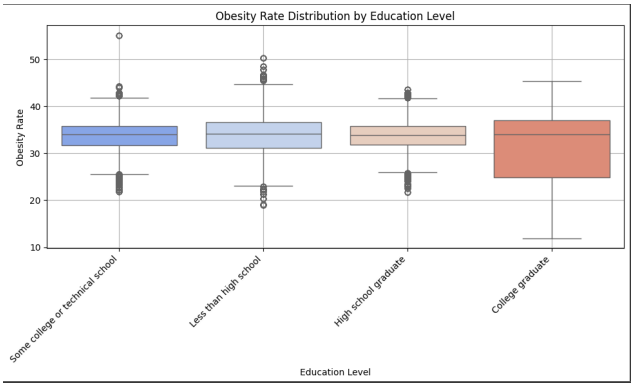
Solutions and Findings

There were multiple areas within our datasets we explored that helped us understand how obesity rates are affected across the different demographics and to what extent. One such example is how different farms can affect obesity rate. We thought that this was an area that is worth looking into, considering that food is also a large playing factor in obesity rates so it's important to analyze how and where these foods are sourced from. To visualize this data, we used a heatmap to graph the correlation between obesity rates and multiple farms like direct sales farms, fresh vegetable farms, berry farms, and more. After analyzing the heatmap, we noticed how direct sales farms have the most effect in reducing obesity rates compared to all other farms.

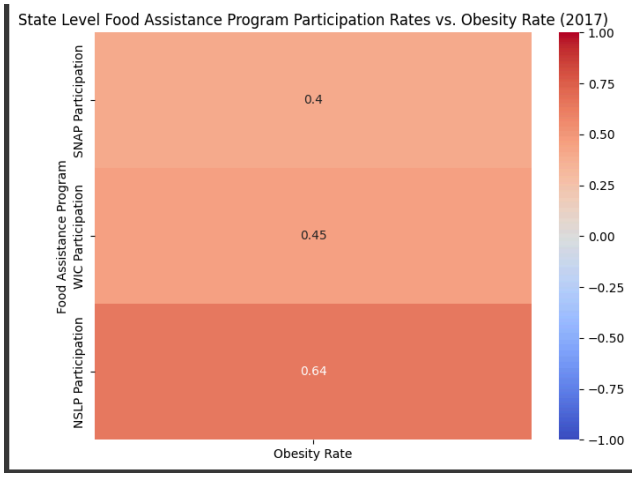
Upon closer analysis, we concluded that this is a plausible outcome since direct sales farms sell fresh produce to consumers, thus ensuring that individuals have access to healthy food. From our earlier analyses, we know that access to healthy food directly lowers obesity rates the most.

An additional area that we explored was an analysis of the correlation of obesity rates and education levels. The hypothesis was that the higher the education level, the lower the rate of obesity because of the widespread knowledge of health in higher education vs. lower education. First, the data was cleaned and then plotted to make a visualization of obesity levels in education levels in a box-and-whiskers plot. In this visualization, the hypothesis was proven to be incomplete because although there was a higher rate in lower education levels, there was a more widespread median in the highest education levels, which led to an inconclusive ideation in the end.

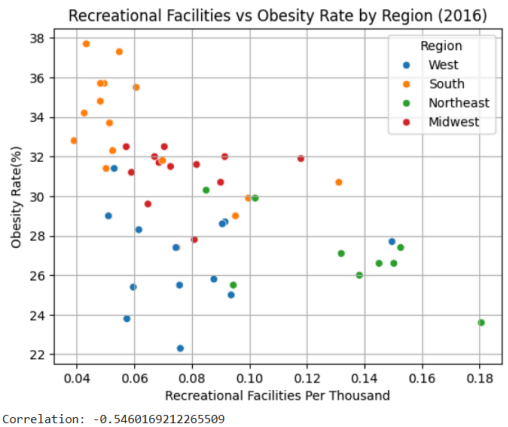




Other areas that we explored modeled the correlation coefficients between obesity rates and various factors. Kyla's visualization shows that increased participation in food stamp programs correlates to increased obesity rates, with participants in the NSLP program generally having the highest obesity rates. In addition to this, we also looked at how different grocery store types correlate to obesity rates, finding that superstores and convenience stores have the positive correlation coefficients with obesity rates. This indicates that as the amount of convenience stores and superstores increase, the obesity rate also tends to increase. However, with specialized stores and grocery stores, we found that as the number of these types of stores increased, the obesity rate tended to decrease. Because specialized stores and grocery stores likely provide less selection to unhealthy foods compared to super stores like Walmart and convenience stores like 7/11, consumers may turn to healthier options.



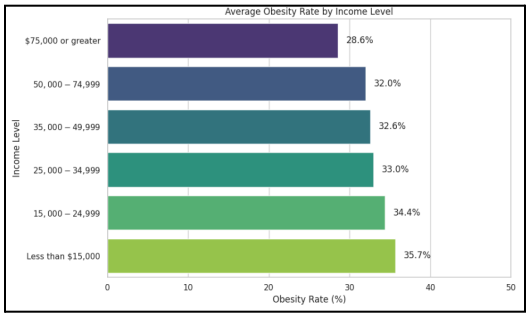
The other two visualizations were shown as scatterplots. Each state corresponded to a single data point, and the two main variables that were compared were the amount of recreational facilities per thousand and the number of agrotourism operations per state. We found that both of these had moderate negative correlations with obesity rates, showcasing that as the number of agrotourism operations and recreational facilities tended to increase, the obesity rate tended to go down. However, neither of these showed particularly

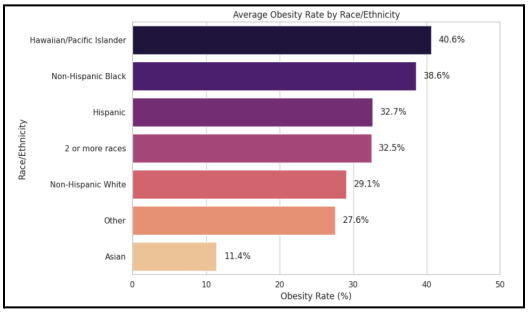


Correlation: -0.5460169212265509

strong correlations, indicating that there are likely a variety of other factors that contribute to obesity rates. The two additional visualizations we did focused on the percent of farmers markets that accept SNAP per state in 2018 and the obesity rate per state in 2018. We ended up seeing a correlation coefficient of -0.3391071844694196, meaning that as the percentage of farmers markets that accepted snap per state tended to increase, the obesity rate tended to decrease. For our other additional visualization, we looked at obesity rate and different taxes. We noticed that there was not a strong correlation between each of the soda taxes and the general food taxes and obesity rates.

Another important area our group decided to explore is how both income and race are linked to obesity and what relationship they have. Our hypothesis was that certain income ranges and certain racial or ethnic groups may have higher obesity rates because of systemic disparities and unequal access to resources. In our income graph we can see that obesity rates tend to go up as income goes down. People making under $15,000 a year have the highest average obesity rate at 35.7%, while those earning $75,000 or more have the lowest at 28.6%. Similarly, in our Race/Ethnicity graph, we see clear differences across racial and ethnic groups. Hawaiian/Pacific Islanders and non-Hispanic Black populations have the highest obesity rates (over 38%) while Asian individuals have the lowest at just 11.4%. These results support our hypothesis that some groups face greater challenges with obesity based on factors they can't always control, like income level and racial background. We also include a few extra visualizations in the colab report.





ML Analyses

For the Machine Learning Analysis, the goal was to predict whether an individual is obese or non-obese based on the feature of sex. The steps that were taken was that first, the data was cleaned. The dataset was filtered to contain only records where the "Sex" column had valid values ("Male" or "Female") and the class column we chose to look at was "Obesity / Weight Status". In our process, we used a threshold of 30 as a means to identify which individuals were obese or non-obese. Through the machine learning model of a decision tree classifier, we were able to train this classifier to classify individuals as obese or non-obese solely based on the "Sex" feature.

The results of the decision tree came to a conclusion that "Sex" is most likely a key determinant for classifying individuals as Obese or Non-Obese. Right now, the tree splits based on the "Sex" feature, and results in different classifications for males and females. Based on our results, we can see that "Sex" alone does serve as an indicator for obesity classification, but our model could have better results if it took into account other factors such as weight, or height.

In the Random Forest model, there was an alternate trend. We observed that the recall for obese individuals was very low, this indicates that the model struggled to identify them accurately. The precision for class 1 was also low, this suggested that the model predicted obesity even when it wasn't accurate. But for class 0 (non-obese), this model showed an improvement in recall, it reached 0.46 which was better than the Decision Tree's performance for non-obese

individuals.

In conclusion, both of our models showed a similar accuracy, around 40%, but in terms of recall and precision, they performed differently. The Decision Tree was better in identifying obese individuals but had trouble with non-obese classifications whereas the Random Forest model had better improvements for non-obese classifications but struggled with obese individuals. All in all, both models could use additional features such as weight, lifestyle, or height to get a better read on obesity rates and make better predictions in the future. We also have our three other ML analyses in the colab report.

Conclusion & Main Takeaways

Our study of this dataset confirms that various demographics do have an effect on obesity rates across individuals with the most significant demographics we noted being income, race, and age. Furthermore, we found that environmental factors such as having access to recreational facilities, fresh food, and agriculture-based programs can mitigate the risks that are associated with obesity. Another interesting point we found while analyzing our data was that while food assistant programs are extremely important in providing food stability to people, it is correlated with higher obesity which highlights that this area needs to be studied on a deeper level to understand the quality and accessibility of food provided by these programs.

Ultimately, we must address both the structural inequalities and the local environment conditions that disproportionately affect people of various communities and backgrounds. Some potential steps we can take in the future to lower obesity rates is to increase access to fresh food in low-income areas, public recreational facilities, and supporting urban farming initiatives.