Food For Thought

Investigating the effect of food prices and local resources on nutritional inequality in the US

Clara Bartusiak (cmb235), Eva Aggarwal (ea196), Abigail Eun (ase26), Bea Radtke (lbr17), Allison Yang (ayy11)

Part 1: Introduction and Research Questions

Nutritional Inequality is defined as the difference in diet quality and access to healthy food between groups of people. Our project aims to explore the underlying trends and factors behind nutritional inequality and give insights into how to best tackle the issue. We are directly applying data science and analysis skills from this class to perform data wrangling, produce visualizations, and conduct statistical analysis to explore nutritional inequality across the U.S. This data analysis is important to understand how food insecurity varies across the U.S. and in aiding policymakers and nonprofit organizations to best target areas of significant need.

We explore the following questions: (1) How do the percent changes in price of nutrient-dense foods (defined as fruits, vegetables, lean protein, whole grains, and low-fat dairy products) compare to less nutritious foods (processed, packaged, finished foods, etc.)? (2) Do certain geographic factors (such as low access to grocery stores) restrict access to healthy foods, and do areas that experience these improve over time? (3) Do certain socioeconomic factors (such as median household income or breakdown of the state population by race) contribute to disparities in food insecurity?

Part 2: Data Sources

To address our research questions regarding food pricing trends, accessibility, and nutritional disparities, we utilized four datasets from the U.S. Department of Agriculture (USDA). These datasets were selected for their comprehensive coverage of food-related economic indicators and geographic disparities in food access:

USDA Economic Research Service – Food Price Outlook

Datasets: *Annual percent changes in selected Consumer Price Indexes (CPI), 1974–2023; Annual percent changes in selected Producer Price Indexes (PPI), 1974–2023*

Content & Relevance: The CPI dataset tracks historical and projected changes in the retail prices paid by consumers, while the PPI dataset reflects the profits made by producers. These metrics are categorized by food type (e.g., beef, fruits, cereals) and level of processing (e.g., farm-level, wholesale, processed). This allowed us to analyze price changes across different food groups and assess how price changes for both consumers and producers varies for what we categorized as healthy vs. processed foods.

Data Preparation:

- Merged CPI and PPI datasets on “Year” and unified product categories for comparative analysis
- Cleaned missing values and converted “Year” and “Percent Change” columns to numeric types using dropna() and pd.to_numeric()
- Resolved naming inconsistencies between CPI and PPI categories (e.g., "Beef and Veal" vs. "Wholesale Beef") using a manual mapping dictionary and custom functions to standardize labels
- Grouped food products by processing level and “healthy” (Unprocessed, Farm-level) vs. “unhealthy” (Finished, Processed)

USDA Economic Research Service – Food Access Research Atlas

Datasets: *State and County Data; Variable List*

Content & Relevance: These datasets provide demographic, socioeconomic, and spatial data on food accessibility across U.S. counties. Variables include proximity to supermarkets, poverty status, income level, and food insecurity metrics. This data supports our investigation into how geographic and socioeconomic factors influence access to nutritious food, especially in food deserts.

Data Preparation:

- Linked “State and County” data to variable definitions using the “Variable List” to decode column codes (e.g., LACCESS_POP15 for population with low store access).

- Filtered and grouped variables by state to compute aggregated metrics like state-level averages for low access populations and food insecurity (CH_FOODINSEC_14_17).
- Merged filtered dataframes on state to create unified datasets
- Conducted EDA to assess relationships like supermarket proximity and household food insecurity
- Created a binary is_food_desert column by labeling counties as 1 if they met federal criteria for limited food access and high poverty (> 30% of the population has low grocery access), and 0 otherwise, based on USDA definitions

Part 3: Modules Used

(3) Visualization: Visualization was used during data investigation, analysis, and the final report stages, which helped uncover trends about CPI and PPI changes, regional patterns in food disparities, and relations between socioeconomic factors and household insecurity. We used line plots and scatter plots (e.g., CPI vs. PPI trends), histograms/density plots/heat maps, choropleth maps, and faceted plots by foot category and price changes.

(4) Data Wrangling: We used data wrangling during the data gathering and cleaning stages of the project. This was important when preparing the USDA CPI, PPI, and Food Access Research Atlas datasets for analysis as our data contained missing values, inconsistent formats, and non-numeric analysis, and thus wrangling was important to allow us to perform our analysis. Concepts applied included handling missing values with dropna(), type conversion (ex. pd.to_numeric()), creating new variables (food categories), and standardizing column names and categories (ex. unifying food item labels across CPI and PPI).

(6) Combining Data: We used concepts from this module during the data cleaning phase while working with multiple data sources. We merged our datasets to analyze relationships between economic and accessibility factors (e.g., linking food insecurity with proximity to grocery stores) and integrated PPI and CPI datasets to compare trends between the two overtime. Specific concepts we applied included merging with different join methods, using mapping functions to align data sets, and multi-level grouping by state, food type, year, etc.

(7) Statistical Inference: We used statistical inference to assess linear and logistic relationships between food pricing, accessibility, and food insecurity. Linear regression analyzed CPI and PPI trends, while logistic regression predicted whether county-level food insecurity decreased from 2012 to 2017 based on economic, demographic, and geographic factors. We evaluated statistical significance using specific topics like p-values, confidence intervals, and model metrics like accuracy, sensitivity, and specificity.

Libraries used: scipy, seaborn, matplotlib, numpy, scikit-learn, geopandas

Part 4: Results + Methods

(*Full Implementation: https://github.com/lbr17/cs216-project*)

Our project started by comparing CPI and PPI changes overtime, after categorizing our observations into five main food categories (Unprocessed, Processed, Finished, Farm Level and Wholesale)

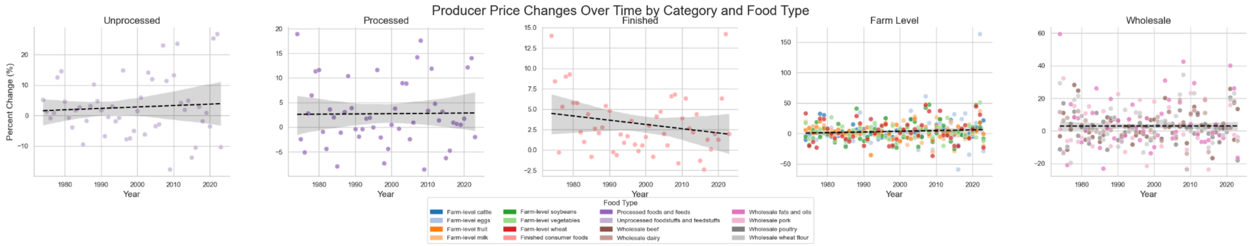


Figure 1: Trends in Producer Price Changes: Food Categories Over Time

Figure 1 does not reveal a strong linear relationship between Producer Price Changes over time when faceted by category and food type. However, it suggests that Farm Level and Wholesale Products experience larger Producer Price Index (PPI) fluctuations than Processed foods. This implies that Processed Foods may offer more stable profits for producers, potentially making them preferable. From here, we decided to categorize our observations further into healthy (Farm level, unprocessed), unhealthy (Processed, finished), and other (Wholesale) foods. From here, we visualized and fit a linear regression model predicting PPI percent change from CPI percent change for both healthy and unhealthy foods.

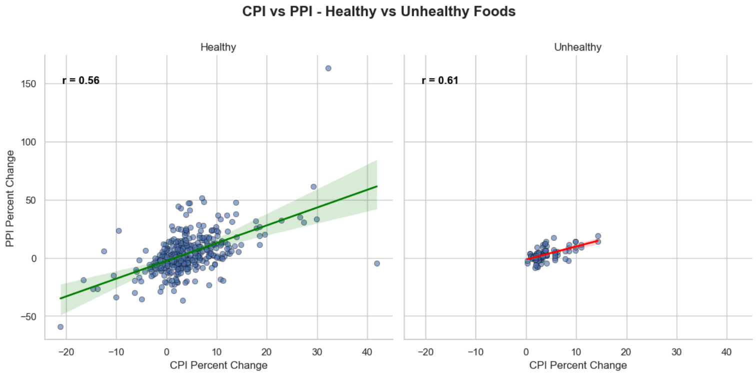
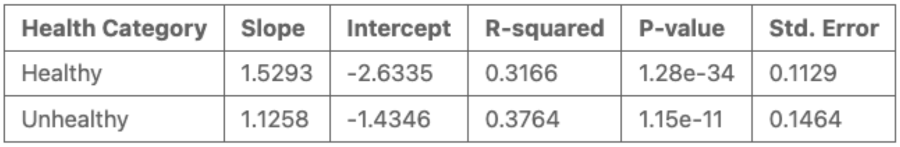


Figure 2: CPI vs PPI - Healthy vs. Unhealthy Foods



Health Category	Slope	Intercept	R-squared	P-value	Std. Error
Healthy	1.5293	-2.6335	0.3166	1.28e-34	0.1129
Unhealthy	1.1258	-1.4346	0.3764	1.15e-11	0.1464

Figure 3: Linear Regression Model for CPI vs. PPI based on Food Group

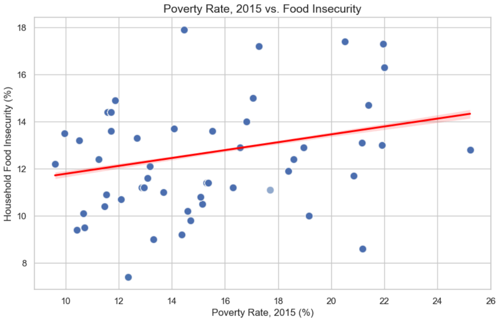
There is a moderate positive relationship between percent change in CPI and PPI for both Healthy and Unhealthy foods. This is reaffirmed by our fitted linear regression models, which suggests that for Healthy foods, for every one percent increase in CPI, we would expect, on average, for the PPI to increase by around 1.53%, while for Unhealthy foods, for every one percent increase in CPI, we would expect, on average, for the PPI to increase by 1.13%. Both models have p-values less than 0.05, suggesting that the relationship between changes in CPI and PPI is statistically significant. This suggests that producers make more profit when the price for consumers are increased for Healthy foods compared to Unhealthy foods, and thus producers may be incentivized to increase the price of Healthy foods over those of Unhealthy.

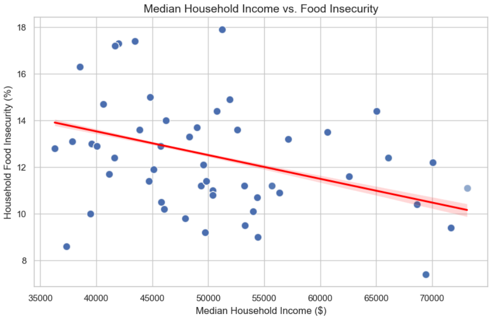
Visualizations

Based on the predictors in our data set, we conducted initial data visualization/exploration, broken into socioeconomic and geographic factors to give us an idea of predictors worth exploring.

Socioeconomic

Socioeconomic factors we decided to look at included household income, poverty rate, and how these factors interacted with food insecurity.





Figures 4/5: Median Household Income/Poverty Rate vs. Household Food Insecurity (%)

Figures 4 and 5 examine the relationship between median household income/poverty rate and household food insecurity, showing a weak positive linear relationship between poverty rate and food insecurity but a moderate negative linear relationship between median household income and food insecurity. This suggests that these are predictors worth examining further/including in a future model.

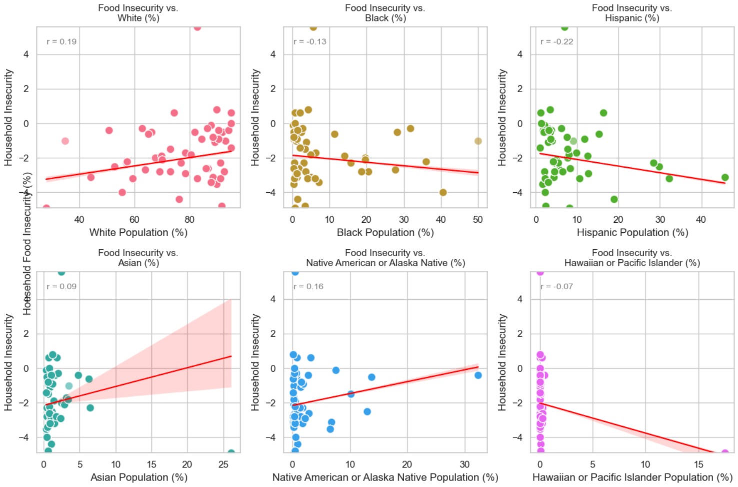


Figure 6: Racial Demographics and Household Food Insecurity Across U.S. States

Figure 6 reveals weak associations between racial composition and food insecurity. White population percentage shows a weak positive correlation (r = 0.19), while Black (r = -0.13) and Hispanic (r = -0.22) populations show weak negative correlations. Indigenous populations exhibit a slightly stronger positive correlation (r = 0.16), suggesting a mild increase in food insecurity. These weak correlations suggest racial composition alone is not a primary determinant of food insecurity but interacts with economic and policy factors.

Geographic

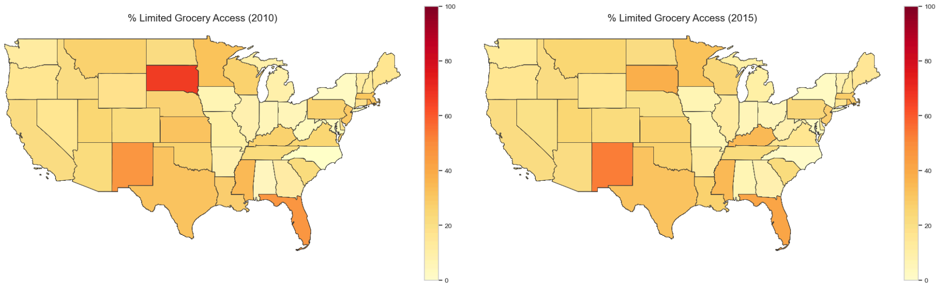

% Limited Grocery Access (2010)
% Limited Grocery Access (2015)
100
80
60
40
20
0

Figure 7: Food accessibility across U.S. states (2015), based on percentage of population with limited grocery store access

Figure 7 illustrates the percentage of each U.S. state's population with limited access to grocery stores in 2015 and 2010. States are shaded according to the proportion of residents experiencing limited food access, with darker colors (red) representing higher percentages. The visualization highlights regional disparities in food accessibility across the United States over 5 years, excluding Alaska, Hawaii, and Puerto Rico for mapping clarity. This helps us in identifying areas that have remained food deserts from 2010-15 and showed little improvement, highlighting regions in need of government support.

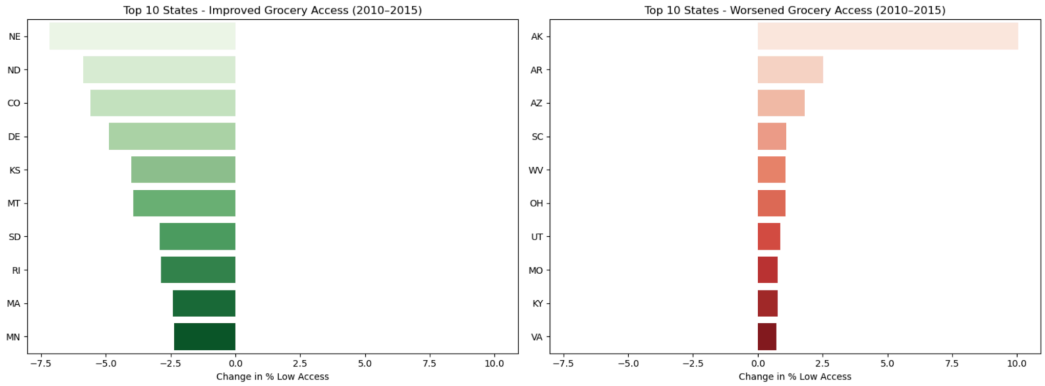


Figure 8: Top 10 U.S. States with greatest improvement (L) and worsening (R) in grocery access (2010-15)

Figure 8 illustrates how progress in food access has been highly uneven across different regions, with some states making substantial strides while others faced setbacks during the same period. On the left, states like Nebraska (NE), North Dakota (ND), and Colorado (CO) achieved the largest reductions in the percentage of residents with limited grocery access between 2010 and 2015, suggesting notable improvements in food availability. On the right, states like Alaska (AK), Arkansas (AR), and Arizona (AZ) experienced the most significant increases in limited access, indicating worsening conditions for food accessibility. This suggests that while food deserts with overall low access rates but signs of improvement may be less concerning, states showing worsening access need closer attention and intervention.

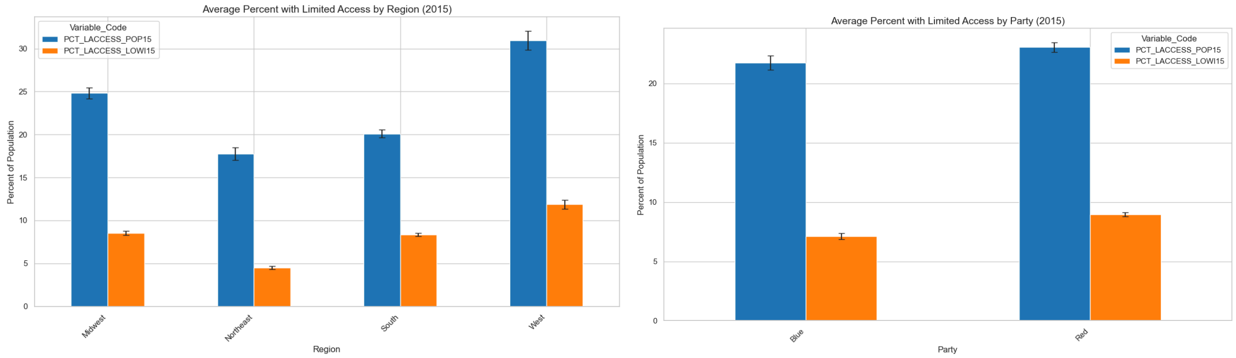


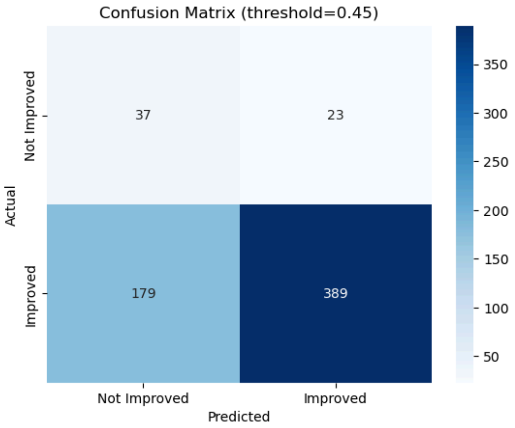
Figure 9: Low access population between U.S. states sorted by region and by political party (2015)

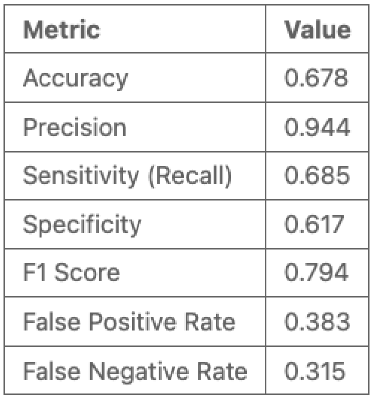
Figure 9 shows how food access varied between US regions and state political parties. The West clearly has the highest percent population low access followed by the Midwest, South, and Northeast. This ranking is followed, with statistical significance as indicated by non-overlapping 95% confidence intervals, for low income population percentages. Though difficult to see with the naked eye, the error bars on the political party map do not overlap for blue and red states (22.3 max versus 22.6 min) for low access. Red states also had a slightly higher percentage low income population.

Model

From our visualizations, we built an overall logistic regression model, predicting whether food insecurity rates would improve based on (2010) percent of the regional population with low grocery access/low income, percent regional ethnic demographic breakdowns (White, Hispanic, etc.); (2015) whether a county is metropolitan, regional poverty rate, food desert status, median household income. Our response variable was the percentage change in SNAP (The Supplemental Nutrition Assistance Program) participants from 2012 to 2017. We created a binary categorical variable, indicating whether a county improved (% participants decreased), or did not (% participants increased).

Due to the disproportionate number of observations we had for counties that improved compared to those that did not, we had to set “class_weight = balance”, to give more weight to the observations that did not improve as we had fewer of those observations. We played around with the threshold level, and found that the best threshold was 0.45, yielding said results:





Metric	Value
Accuracy	0.678
Precision	0.944
Sensitivity (Recall)	0.685
Specificity	0.617
F1 Score	0.794
False Positive Rate	0.383
False Negative Rate	0.315

Figure 10: Racial Demographics and Household Food Insecurity Across U.S. States

Our logistic regression model had 68.5%, 61.7%, 67.8%, and 94.4% sensitivity, specificity, accuracy, and precision rates respectively. The FPR and FNR were 38.3% and 31.5% respectively. These metrics suggest that our model has moderate predictive power in terms of whether a county’s food insecurity rate would decrease overtime or not, depending on certain socioeconomic and geographic factors. This model could aid policy makers in gauging which counties may require the most aid/how to most adequately aid food insecure communities.

Part 5: Limitations and Future Work

In conducting our data analysis, we faced three main limitations: (1) The column we initially chose as our response variable that indicates the percent of the population experiencing food insecurity was only recorded at the state level, rather than the county level, (2) The columns in the dataset recorded values for different years, and (3) Our dataset was heavily imbalanced with regard to the number of counties that experienced an increase in food access versus the number that did not.

Limitation 1: The column we initially chose as our response variable that indicates the percent of the population experiencing food insecurity was only recorded at the state level, rather than the county level. This gave us 50 data points to work with, which wasn’t sufficient to conduct thorough statistical analysis. Thus, we picked another column in the dataset that indicated the change in the percentage of county population that participated in SNAP (2012 to 2017). SNAP is a program that supplements a family's grocery budget and allows them to afford more nutritious food. Thus, we assumed that the population percentage on SNAP was correlated to the percentage that experienced food insecurity. However, we recognize there are some misalignments with this substitution. For one, the percentage of people who participate in the SNAP program is not necessarily comparable to the actual number of people experiencing food insecurity. There may be food-insecure individuals who aren’t eligible to receive SNAP benefits. To add, the percentage of population receiving SNAP benefits could be influenced by outside factors like political administration, as a decrease in the number of people participating in the SNAP program could just as well be caused by administrative budget cuts.

Limitation 2: The columns in the dataset recorded values for different years. Unfortunately, not all the columns in the dataset were recorded for the same year, and even the columns that recorded values for percent change across a range of years did not record these for the same range of years. For example, each of the demographic breakdowns for the percent of the population for each race was recorded for 2010, however median household income was recorded for 2015 and the percentage of SNAP participants in each county was recorded from 2012 and 2017. To conduct our analysis, we made assumptions that these variances in years (such as 2010 compared to 2012) were not significant enough to prevent us from comparing values across these years. Future work could include finding a more consistent dataset or other datasets to supplement missing years.

Limitation 3: Our dataset was heavily imbalanced with regard to the number of counties that experienced an increase in food access versus the number that did not. After conducting statistical analysis, our dataset revealed that 90% of counties experienced an increase in food access, compared to 10% that did not. The lack of a balanced dataset in this regard made it challenging to create an accurate logistic regression model. Initially, the model predicted that every county would experience improved access to grocery stores simply because the vast majority of counties in the dataset experienced access to grocery stores. However, we were ultimately able to overcome this predictive bias by tuning model parameters and lowering the prediction threshold to 0.45, which achieved a 67.8% prediction accuracy.

Future work: In future work, we would aim to obtain data that is more consistent and granular, spatially and geographically. Data that is more consistent on the county level across specific time-periods would help counter some of the limitations we faced in this project. It would also be helpful to find more data on additional factors affecting food insecurity to fit a better model, such as political standing, local food initiatives and transportation. Incorporating these broader structural and policy-related variables would provide a more complete understanding of the drivers behind nutritional inequality across the US.

Part 6: Conclusion

Thinking back to our research questions and results, we concluded our findings as follows:

We evaluated CPI and PPI changes from 1974 to 2023 and found substantial evidence that producers earn more profit for increasing prices of healthy foods compared to unhealthy. Unhealthy food PPIs were more consistent, suggesting that unhealthy foods may be more preferable for producers. We also examined geographic factors such as low access to grocery stores and percentage of low income faceted by region and state political party, finding that Western states had the highest percent low-income and low-food-access populations, followed by the Midwest, South, and then the North. Red states also had slightly more low-income and low-access populations than Blue states. From our visualizations of socioeconomic factors (poverty rate, median household income in 2015), there is substantial evidence that increased poverty rates and lowered median household income is correlated with increased household food insecurity at least on the state level. We also found there was a weak correlation between racial demographics and food insecurity. Our overall logistic model that incorporated both geographic and socioeconomic factors had 68.5% and 61.7% sensitivity and specificity rates respectively, indicating that while these predictors have correlations with household food insecurity, their relationships are not substantial enough on their own.

Part 7: Collaboration Reflection & Sharing Plans

Our initial collaboration plan detailed the following:

Weekly meetings (potentially at 3pm on Saturdays either in person or on Zoom): We ended up meeting a little more sporadically over the course of the semester, scheduling calls and meet-ups instance-by-instance instead of having a fixed time since our schedules ended up being more variable than expected.

1-2 hours per person per week up to 3-4 by the end of the semester: This ended up being about right if we take the mean work done over the entire semester. Project workload increased in late March and April. *Text message group chat for communications outside of meetings:* This was our main form of communication throughout the semester. *Keep track of outstanding and completed tasks on a Google document:* We more or less kept up with using this document and also used it to brainstorm at times. *Use a shared google folder and a GitHub Repository to keep all of our code together:* All of our milestone documents are in our folder and the repository is how we kept up with code collaboration. *Willingness to publicly share our project as part of a portfolio of work as long as it is clear that this project was created by the five of us:* This is still true and some of our members will have the project shared on their Github/lab with clear attributions.

We pretty much followed our initial plan with some minor deviations. Assigning roles earlier in the semester could have been helpful for balancing workload. Additionally, scheduling meetings was pretty hard and could've used a little more thought at the beginning of the semester. More consistent commits, pulls/pushes could also have been helpful for merge conflicts as well as tracing our contributions and flow.