altREU: Computational Modeling Serving Your Community
Exploring Food Desert Impact on Health in Chicago
Paulina Grzybowicz
August 6th , 2020

Abstract

Food deserts are defined as, “an impoverished area where residents lack access to healthy foods”. This lack of access can be due to a combination of socioeconomic, geographic, and food-related variables, and has been proven to impact the health of residents in the area. In this project, several statistical and machine learning techniques are used to model the impact of food desserts on health in both the City of Chicago and the state of Oregon. The models are then used to evaluate the influence of specific food desert factors on health, in order to help evaluate potential solutions.

Purpose

The purpose of this project is to use computational modeling to showcase the proposed correlation between a wide-variety of food desert variables and diabetes rates. Such a correlation would indicate that several factors contribute to the creation of a food desert, as opposed to just physical food access. It would also prove food desert impact on health outcomes.

Understanding the significance of the predictor variables would then allow for a better evaluation of potential solutions to lowering food-desert and diet-related health outcome rates.

Introduction

According to ___, food deserts are “impoverished areas where residents lack access to healthy foods. Food deserts may exist in rural or urban areas and are associated with complex geographic and socioeconomic factors, as well as with poor diet and health disorders such as obesity”. Food deserts have been shown to affect the health of residents, specifically impacting the prevalence of diet-related health outcomes such as diabetes and cardiovascular disease.

Access to food can be misunderstood as simply an individual's distance to a grocery store. This is a common misconception when discussing food deserts. In fact, food access consists of a variety of socioeconomic and geographic factors, including income level, employment, access to transportation, access to child care, available time, nutritional literacy, and even culture. For example, the amount of time available to shop and cook is significantly lowered for an individual who is working long hours, lives far from a grocery store, is reliant on public transportation, or doesn’t have access to childcare. Additionally, one’s ability to purchase and cook healthier options requires an understanding of nutrition and certain recipes, knowledge that may not be accessible to all.

Food deserts are very prominent in Chicago, and correlate significantly with the socioeconomic status of a community. Additionally, in Chicago, communities of color are disproportionately impacted by food

deserts. The city has a very diverse population, with White, Black, and Hispanic/Latino populations making up 32.3, 26.7, and 30.9 percent of the total population respectively. This makes for an almost even split between the three groups, making it one of the most diverse cities in the United States.

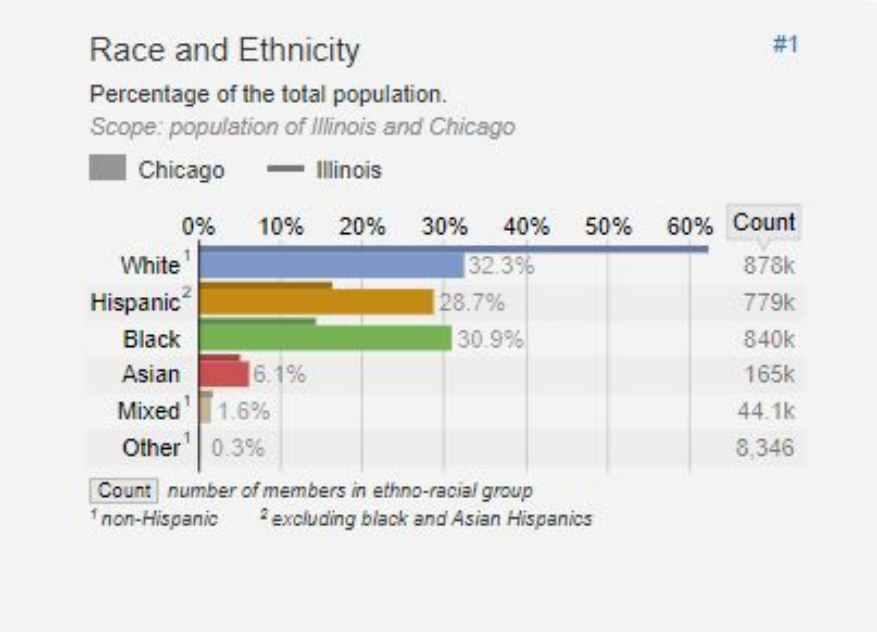
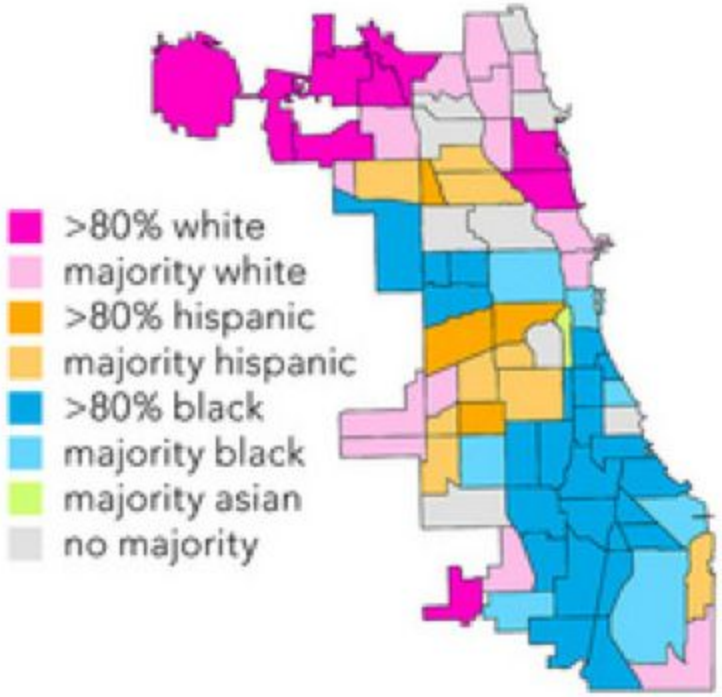


Figure 1: Race and Ethnicity in Chicago

The city is also extremely segregated, due to a long history of racist housing and investment policies. Decades of unfair investments in housing, education, medical resources and food, has left many populations in Chicago at a disadvantage to this day. This divide is very unsurprisingly correlated to race.

Figure 2: Racial Distribution by Community Area



There are significant inequalities among Chicago community areas, and food systems are impacted. This is why we see communities of color being disproportionately impacted by food deserts, as shown in Figures 3 and 4.

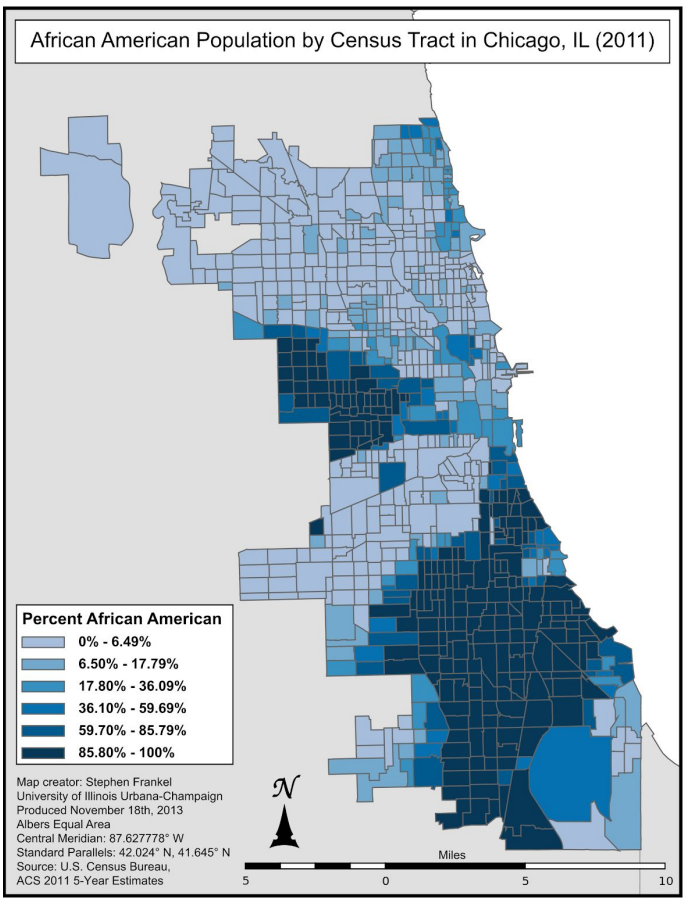


Figure 3

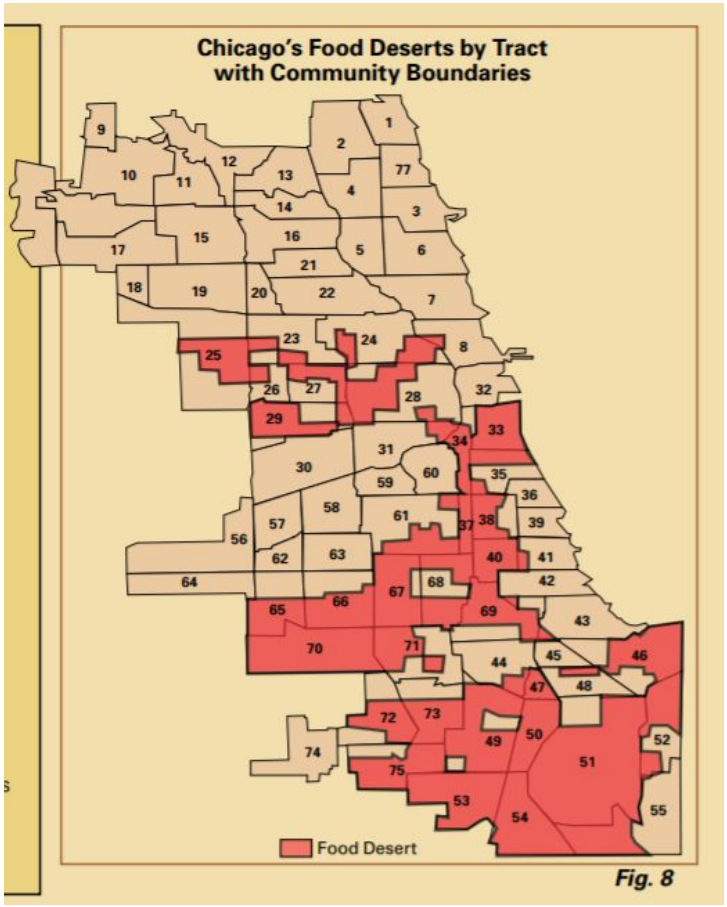


Figure 4

These patterns are accentuated by other profit-driven motives. For example, " fast-food companies are 60 percent more likely to advertise to children in predominantly black neighborhoods than in white neighborhoods", while more expensive options, like farmers markets, are more likely to be placed in predominantly white neighborhoods.

Because the formation of food deserts are so clearly correlated to racist policy and profit-driven tactics, many academics have begun to use the term "food apartheid" instead of "food desert", to bring attention to the structures (put in place by humans) have produced these inequities. ____ defines food apartheid as "a human-created system of segregations, which relegates some people to food opulence and other people to food scarcity".

When exploring the relationship between food access and health outcomes, it's essential to understand that food inequality exists because of structural elements in our society. Food access and health are incredibly complex, and there's no way to analyze them without considering larger contexts of racial and income inequalities. These underlying ideas guided every portion of the project.

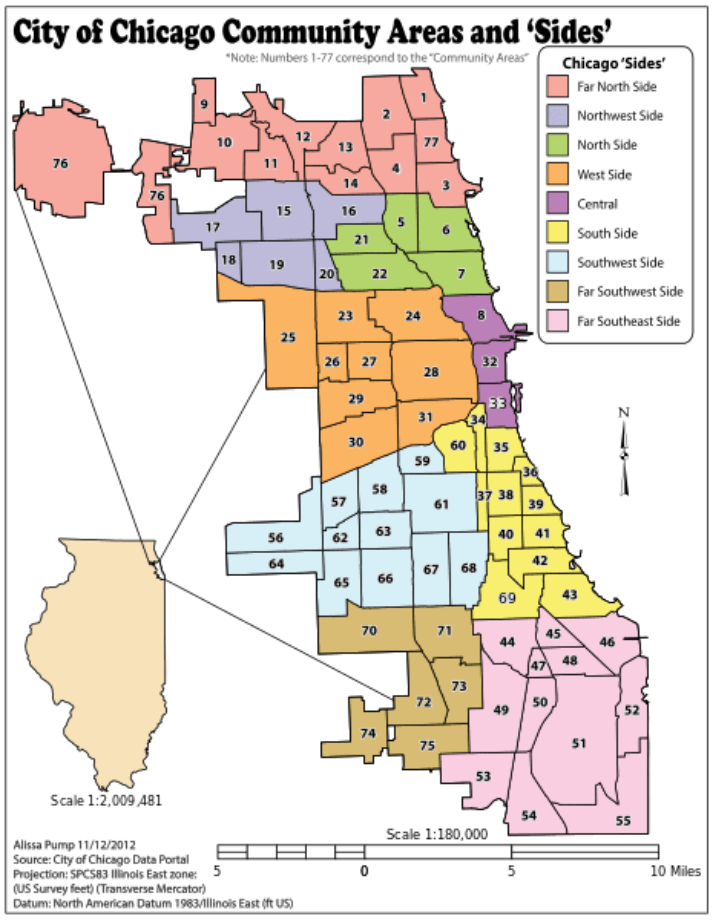
Assumptions

The models in this project work under the widely-accepted definition that a food desert is an "impoverished area where residents lack access to healthy foods".

Additionally, the data used in this project is based upon the 77 Chicago Community Areas. This is a low amount of data points, but because they each represent a significant amount of Chicago citizens, the assumption is made that these data points will yield significant results.

Datasets

The data collected for this project was all sourced from Chicago Health Atlas and Chicago Data Portal, online sources that provide datasets about Chicago citizens. This data was divided by Chicago Neighborhood, the 77 communities the city is split into based upon census tracts.



https://www.chicagodetours.com/chicago-community-areas-explainer/

The variables that were collected reflect the presumed factors that make up a food desert, and include socioeconomic status indicators, race, and food-related indicators like access to fruits and vegetables and number of grocery stores.

Variable	**Description**
Population	Number of people residing in community
Square Milage	Size of the community
Active Transportation - Percent	Percent of workers 16 years and over who walk, bike, or take public transportation to commute to work.
Percentage Non-Hispanic White (2012-2016)	Percentage of population that identifies as white with no Hispanic or Latino ancestry
Percentage Non-Hispanic African American or Black (2012-2016)	Percentage of population that identifies as African American or Black with no Hispanic or Latino ancestry
Percentage Hispanic or Latino (2012-2016)	Percentage of population that identifies as Hispanic or Latino
Median Income	Average income by community area
Poverty	Percent of households whose income is below the poverty line
Unemployment Rate	Percent of Civilian Population aged 16 years and older who were unemployed.
Food Stamps	Percent of Households receiving food assistance. .
Rate Limited Food Access	Percent of population considered low income residents with limited access to affordable, healthy foods.
Percent with Easy Access to Fresh Fruits and Vegetables	Estimated number of adults (18 years and older) who reported that it is very easy for them to get fresh fruits and vegetables divided by the estimated number of adults, expressed as a percent. This number is weighted to represent the population from which the sample was drawn, the household population of adults 18 years of age

	and older who reside in the City of Chicago.
Number of Grocery Stores	Number of grocery stores per community area
Adult Obesity Percentage	Percent of Adults with a body mass index (BMI) of 30 or greater.
Diabetes Rate	Percent of Adults Diagnosed with Diabetes (excluding gestational and pred-diabetes

Methods

Gathering Data

The above datasets were gathered and prepared to be used in future data analysis. Datasets were prepared using Microsoft Excel and manipulated using Python's pandas library.

Linear Regression

In order to explore the relationships between these variables initially, linear regression was used. Each individual variable, and several combinations, were plotted against both Diabetes Rate as well as Adult Obesity Percentage in order to observe the correlation between the variables and spot possible patterns.

Several python libraries, including pandas, numpy, matplotlib, and sklearn were used to conduct these linear regressions and visualize them. Datasets were uploaded as csv files and cleaned using pandas, linear regressions were performed using sklearn, and the regressions were visualized using matplot.

Mediation Analysis to Explore

Mediation analysis was conducted to further explore the relationships between variables. These tests were conducted under the hypothesis that Food-Related variables (such as food access, grocery store number) acted as a mediator between Socioeconomic Status and Diabetes Rate. This proposes that the relationship between SES and Diabetes Rate is influenced by a third intervening variable, Food Access. SES influences one's access to food, which in turn influences their health.

The mediation analysis was conducted using the mediation library in R. Baron and Kenny's steps for mediation were followed, and the bootstrap technique was used to test the significance of the mediation variable.

Baron and Kenny's steps for mediation involve 3 sets of linear regressions. The first checks to see if the independent variable predicts the dependent variable. The second regresses the mediator on the independent variable to check if it predicts the mediator. The third regresses the dependent variable on

both the mediator and independent variable. The significance of these regressions is measured by their p-values. Bootstrapping then confirms whether the mediation effect is significant.

ML techniques to predict

The previous two techniques allowed for a constructive exploration of the variables at hand. The findings from these techniques were then used to create several machine learning models, with the goal of predicting a community area's diabetes rate using several SES and Food-Related indicators.

Regressing on several combinations of the predictor variables allowed for a basic understanding of which were more correlated with diabetes rate. Using this knowledge, several training sets were created. The goal of creating several sets was to manually assess which set of variables were most accurate in predicting variables rates, based on previous analysis.

The most notable sets included:

- Training sets with and without **Obesity Rate**, which has a very strong correlation with obesity rate and likely influences the model very strongly.
- Training set without SES indicators
- Training set without Race
- Training set without Food-Related indicators

Several different ML algorithms were used as well. Support Vector Regression, a variation of the Support Vector Machine algorithm, tries to find a line in space that is used to predict continuous output. The SVR algorithm was used with several kernels (linear, rbg, poly, and sigmoid). Another popular regression technique, Ridge Regression, was used, which is good for datasets with multicollinearity, or correlations between predictor variables. Lasso Regression and ElasticNet Regression, both very similar to the Ridge Regression model were also used. Finally, the Random Forest method was used, which is a flexible model that constructs a number of decision trees.

These models were run using a 75-25 split for training and testing data.

After running these models, both their accuracy and variable importance was evaluated. Accuracy is measured based upon how well the model is able to predict the testing set after being trained on the training dataset. Variable importance is evaluated based upon factor coefficients in the final regression. These coefficients provide an understanding of which predictor variables were most important in the model.

Findings

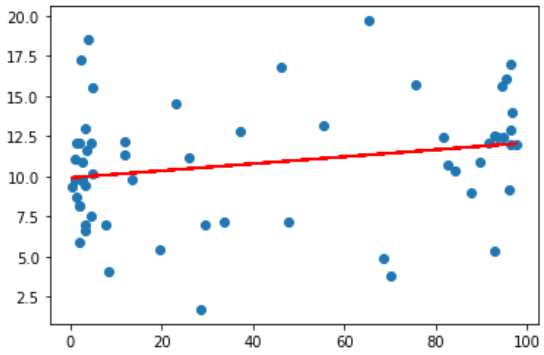
Linear Regression

All variables considered to create a Food Desert were plotted against Diabetes Rate to search for relationships between variables. Variables with a stronger correlation to Diabetes Rate include: **Limited**

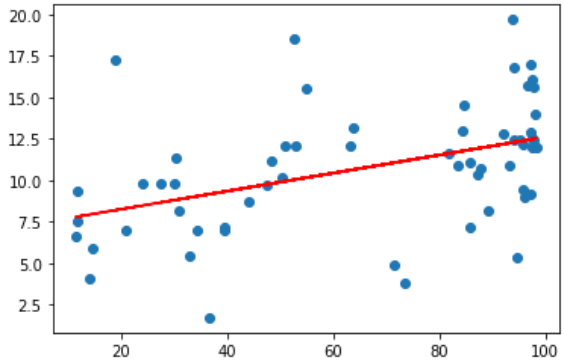
Food Access, Active Transportation, Unemployment Rate, Percent Poverty. Variables with unexpectedly low correlations include **Median Income, Percent Non-Hispanic White, Percent Non-Hispanic African American.**

The stronger correlations were evenly distributed between SES and food-related variables, meaning there wasn't an indication that one had more influence on diabetes rate than the other. This supported the idea that food-related variables may act as a mediator between SES and diabetes rates, which would be tested next.

The lack of correlation between race variables and diabetes rates was initially surprising, given the rampant segregation in Chicago. However, when a combination of Percent Non-Hispanic African American or Black and Percent Hispanic/Latino was tested, the correlation was much stronger. This showcases the overall correlation between communities of color and higher diabetes rates.

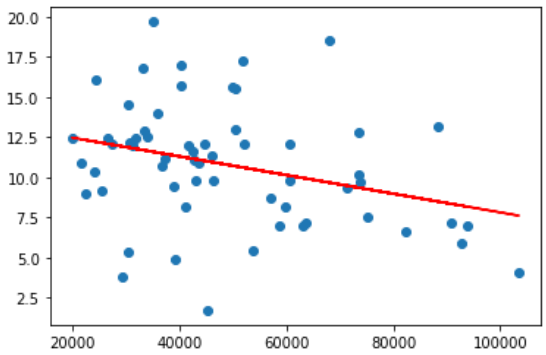


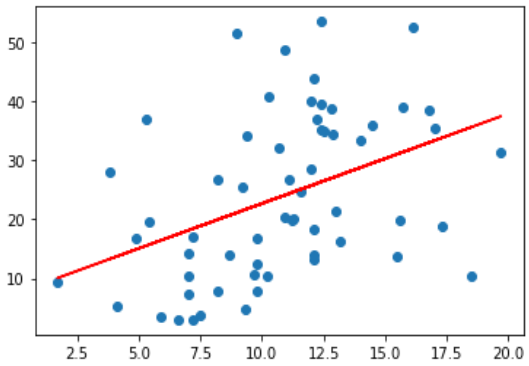
Percentage Black vs Diabetes Rate
R^2 = 0.05152733436080459



Percent Black + Hispanic vs Diabetes Rate
R^2 = 0.19599998330485546

The lack of strong correlation between median income alone and diabetes rate is possible, since SES is much more than just income. When looking at SES indicators like Unemployment, Food Stamp Usage, and public transport use, the correlation is much clearer. A similar conclusion can be made for the lack of strong correlation between **Grocery Store Number** and Diabetes Rate- grocery store amounts alone don't indicate food deserts, especially when the size of the community isn't reflected in that number.



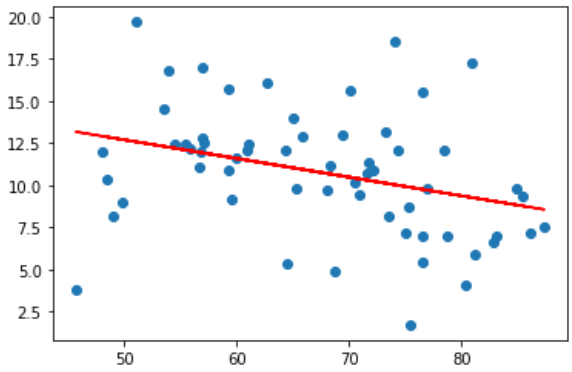


Median Income vs Diabetes Rate
R^2 = 0.09948359220222736

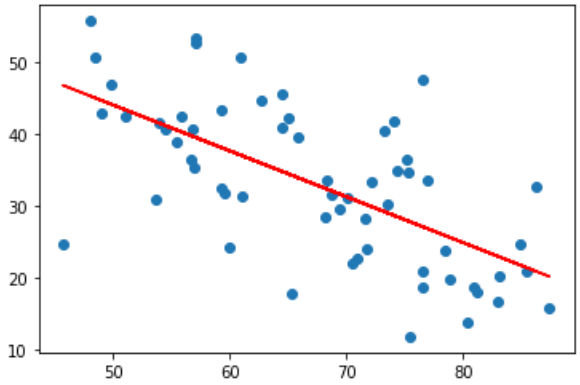
Percent Below Poverty Line vs Diabetes Rate
R^2 .0.17213094600984258

Though there were overall correlations between these variables, error was noticeably high for almost every regression considered. This is largely due to the small dataset.

All food desert variables were also plotted against **Adult Obesity Percentage,** in order to observe potential impacts on other health outcomes. All correlations were significantly stronger. This can be explained by obesity being more common than diabetes. There's a wider range of obesity rates available, and thus stronger correlations are found. Since obesity and diabetes are linked, this was also helpful information.



% Easy Access vs Diabetes Rate
R^2 = 0.10662956294919368



% Easy Access Vs Adult Obesity Rate
R^2 = 0.4151324607149999

Mediation Analysis

Mediation Analysis was then used to further investigate the relationships between the variables. The hypothesis being tested was that different food-related variables (Access to fresh fruits and vegetables, number of grocery stores, etc) act as a mediator between SES indicators and Diabetes Rates. A community area's socioeconomic status influences the area's access to healthy foods, which in turn influences health outcomes in the area.

Several combinations of SES and food-related indicators were tested against diabetes rate. (Figure)

Independent Variables	**Potential Mediator**	**Dependent Variables**	**Conclusion**
Median Income	Percent Easy Access to Fruits and Vegetables	Diabetes Rate	Not very significant
Median Income	Percent Easy Access to	Obesity Rate	Significant

	Fruits and Vegetables		
Poverty Rate	Percent Easy Access to Fruits and Vegetables	Diabetes Rate	Significant
Unemployment Rate	Percent Easy Access to Fruits and Vegetables	Diabetes Rate	Signifiant
Percent Non-Hispanic White	Percent Easy Access to Fruits and Vegetables	Diabetes Rate	Somewhat Significant

When testing whether Easy Access to Fruits and Vegetables served as a mediator between Median Income (SES) and Diabetes Rate, we saw a significant p-value when plotting Median Income against Diabetes Rate (0.0141), meaning there is a correlation. There was a very significant p-value when plotting Median Income against Easy Access to Food, showing that our independent variable did influence our Mediator. Finally, when plotting Median Income and Easy Access to food against Diabetes Rate, we saw a reduced p-value between our independent and dependent variable (indicating that our mediator's is significant). However, the p-value between Easy Access to Food and Diabetes rate was only slightly significant, indicating the mediating relationship was not very strong. This was confirmed by the Bootstrapping test of significance.

When testing whether Easy Access to Fruits and Vegetables served as a mediator between Poverty Rates and Diabetes Rate, there were much stronger results. When plotting Poverty against Diabetes Rate, the p-value was very significant, confirming the two variables were correlated. When plotting Easy Access and Poverty Rate, we also saw a very significant p-value. Finally, our third regression showcased a significantly weakened p-value between our independent and dependent variables, yet a very significant p-value between our mediator, Easy Access, and our dependent variable, Diabetes Rate. Therefore, we conclude that a mediation effect is present. This was confirmed by the Bootstrapping test of significance

Machine Learning

The next portion of the project involved creating a machine learning model that was able to use a variety of variables to predict the diabetes rate in any Chicago Community Area. Several models and training sets were used in search of the most accurate model, and variable importance was evaluated during this process.

The goal of the model was to predict a community area's rate of diabetes. Thus, several regression models were used and reviewed. This included: Support Vector Regression (with linear, rbf, poly, and sigmoid kernels), Ridge Regression, Lasso, ElasticNet, and Random Forest models.
Several training sets were also used. These training sets involved different combinations of variables in order to explore what variables were necessary to create an accurate model. The following training sets were used:

#1	**#2**	**#2.1**	**#2.2**
All Variables	***No Easy Access, No Obesity Rate***	***No Easy Access***	***No Limited Access***
- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Median Income - Unemployment Rate - Food Stamp Usage - Percentage Limited - Access to Food - Percentage Easy Access to Fruits and Vegetables - Number of Grocery Stores - Percent Adult Obesity	- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Median Income - Unemployment Rate - Food Stamp Usage - Percentage Limited - Access to Food - Number of Grocery Stores	- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Median Income - Unemployment Rate - Food Stamp Usage - Percentage Limited - Access to Food - Number of Grocery Stores - Percent Adult Obesity	- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Median Income - Unemployment Rate - Food Stamp Usage - Percentage Easy Access to Fruits and Vegetables - Number of Grocery Stores - Percent Adult Obesity

#2.3	**#2**	**#2.1**	**#2.2**
No Limited Access, No Obesity Rate	***No Race Variables***	***No Race Variables, No Obesity Rate***	***No SES variables***

- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Median Income - Percent Poverty - Unemployment Rate - Food Stamp Usage - Percentage Easy Access to Fruits and Vegetables - Number of Grocery Stores	- Square Milage - Population - Percent use of Public - Transport - Median Income - Percent Poverty - Unemployment Rate - Food Stamp Usage - Percentage Limited - Access to Food - Percentage Easy Access to Fruits and Vegetables - Number of Grocery Stores - Percent Adult Obesity	- Square Milage - Population - Percent use of Public - Transport - Median Income - Percent Poverty - Unemployment Rate - Food Stamp Usage - Percentage Limited - Access to Food - Percentage Easy Access to Fruits and Vegetables - Number of Grocery Stores	- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Percentage Easy Access to Fruits and Vegetables - Number of Grocery Stores - Percent Adult Obesity

#5	**#6**
No Obesity Rate	***No Food-Related Variables***
- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Median Income - Percent Poverty - Unemployment Rate - Food Stamp Usage - Percentage Limited - Access to Food - Percentage Easy Access to Fruits and Vegetables - Number of Grocery Stores - Percent Adult Obesity	- Square Milage - Population - Percent use of Public - Transport - Percentage White - Percentage Black - Percentage Hispanic - Median Income - Percent Poverty - Unemployment Rate - Food Stamp Usage - Percent Adult Obesity

The goal of training sets #2.1-2.4 is eliminating the possibly contradictory values of Percent Limited Access to Food and Percent Easy Access to Fruits and Vegetables. Since these are generally opposite values, the thought was they could potentially be misleading.

The goal of training sets #3.1-#3.2 was to eliminate race-related variables. Since Chicago is a very racially diverse and racially segregated city, a well-recognized hypothesis is that communities of color are disproportionately impacted by food deserts. Thus race was thought to impact diabetes rate in a community.

The goal of training set #4 was to further explore the relationship between SES and food deserts, and whether or not the model's accuracy would weaken without these variables present.

The goal of training set #5 was to test the influence of Obesity Rate on the model's accuracy. Because obesity is so strongly linked to diabetes and because the correlations to these variables to Percentage Obesity Rate was fairly strong, Percentage Obesity Rate was thought of as having significant influence on the model. However, it's incorporation in the model was controversial, as it is a health indicator that perhaps should not be included when attempting to predict Diabetes Rate from food-desert related variables.

Finally, for consistency, the goal of training set #6 ways to eliminate the most obvious food-desert variables: any variables related to food. This included food stamp use, grocery store number, and access to food/fruit/vegetables.

The following results were found:

#1	**#2.1**	**#2.2**	**#2.3**	**#2.4**
SVR → linear: 59% → poly: 27% → sigmoid: 28% → rbf: 25%	SVR → linear: 40% → poly: 31% → sigmoid: 18% → rbf: 24%	SVR → linear: 32% → poly: 23% → sigmoid: 25% → rbf: 23%	SVR → linear: 21% → poly: 23% → sigmoid: 30% → rbf: 17%	SVR → linear: 41% → poly: 38% → sigmoid: 37% → rbf: 24%
Ridge: 26%	Ridge: 48%	Ridge: 6%	Ridge: 26%	Ridge: 46%
Lasso: 38%	Lasso: 11%	Lasso: 33%	Lasso: 30%	Lasso: 47%
ElasticNet: 42%	ElasticNet: 32%	ElasticNet: 19%	ElasticNet: 26%	ElasticNet: 54%
Random Forests: 33%	Random Forests: 40%	Random Forests: 33%	Random Forests: 26%	Random Forests: 50%
Average: 34.75	Average: 30.5	Average:24.25	Average: 24.88	Average: 42.125

#3.1	**#3.2**	**#4**	**#5**	**#6**
SVR	SVR	SVR	SVR	SVR

→ linear: 40% → poly: 42% → sigmoid: 32% → rbf: 35%	→ linear: 30% → poly: 34% → sigmoid: 25% → rbf: 22%	→ linear: 52% → poly: 28% → sigmoid: 38% → rbf: 32%	→ linear: 22% → poly: 24% → sigmoid: 19% → rbf: 21%	→ linear: 13% → poly: 20% → sigmoid: 20% → rbf: 20%
Ridge: 54%	Ridge: 61%	Ridge: 26%	Ridge: 43%	Ridge: 24%
Lasso: 56%	Lasso: 62%	Lasso: 27%	Lasso: 37%	Lasso: 28%
ElasticNet: 54%	ElasticNet: 62%	ElasticNet: 27%	ElasticNet: 36%	ElasticNet: 29%
Random Forests: 48%	Random Forests: 32%	Random Forests: 21%	Random Forests: 21%	Random Forests: 21%
Average: 45.125	Average: 37.5	Average:31.75	Average:27.89	Average: 21.875

Though the accuracy of the models can vary greatly due to the small training set and testing set size, these fairly consistent accuracies indicate which groups of variables and which models were most effective in predicting diabetes rates.

The most accurate training sets were: #1, #2.4, and #3.1.

In the case of **#1,** Including all variables was very effective, meaning the algorithms are able to do fairly well in deciphering which variables are most important.

For training sets under **#2,** It seems that choosing **Easy Access to Fruits and Vegetables** gives stronger results than including **Limited Food Access.** But because the average accuracy score of training set #1 was relatively high , it seems the two variables don't contradict themselves too much.

The sets that did not include racial variables (**#3)** yielded surprisingly high results. This raises the question if race is relevant when predicting diabetes rate in Chicago. Because there is so much evidence that suggests so, this also raises the question of whether or not these accuracy scores are actually significant, due to the small number of datapoints.

Data sets that did not include SES indicators and Food-Related Variables (**#4, #6** respectively) both had significantly lower average accuracy rates, indicating the importance of these variables.

And finally, all sets including **Percent Adult Obesity** yielded higher average accuracies than those without. This showcases the large influence of this variable, and further puts to question its relevance in this dataset.

Model	**Average Accuracy**

SVR: linear	35
SVR: RBF	24.3
SVR: sigmoid	26.2
SVR: poly	29
Ridge Regression	36.5
Lasso	36.7
Elastic Net	38.1
Random Forest	32.5

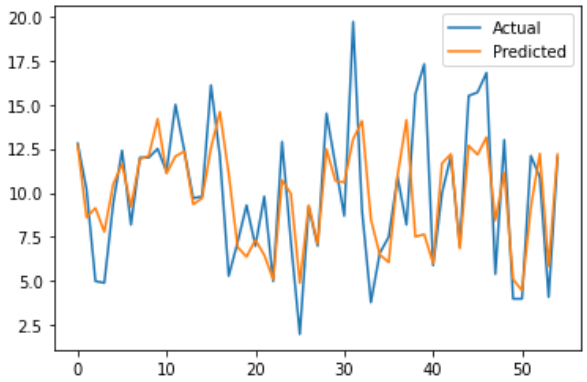
The average accuracies of the models used has a smaller range: from around 24-38. This is likely because the models are similar in nature (use regression) and also because of the small dataset used. The **Ridge, Lasso, and ElasticNet** models have the highest accuracies, as well as the SVR (linear kernel) model. The SVR models using **RBF and sigmoid** kernels had the lowest accuracies.

Because **ElasticNet** combines both **Ridge** and **Lasso** regressions, this would likely be the model of choice. This model works to __simplify something idk__.

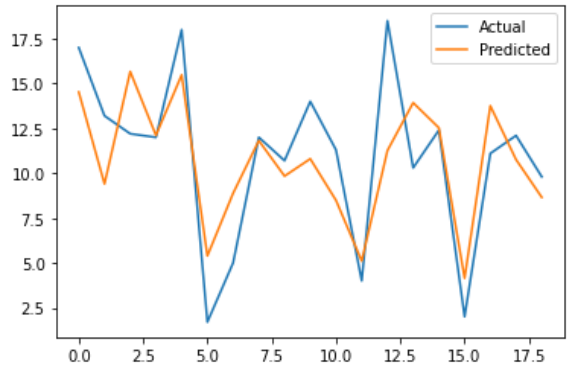
After running the machine models, the importance of the variables was evaluated for the most accurate models. For simplicity, the importance of the variables for training set #1 will be evaluated, as this model uses all variables

SVR: linear	
Feature	**Score**
Square Milage	0.18465
Population	-0.50061
Percent use of Public - Transport	0.22129
Percentage Non-Hispanic White	0.31764
Percentage Non-Hispanic Black or African American	-0.63748
Percentage Hispanic or Latino	0.72555

Median Income	-0.37569
Unemployment Rate	-0.36821
Food Stamp Usage	1.14310
Percentage Limited - Access to Food	0.78979
Percentage Easy Access to Fruits and Vegetables	-0.27690
Number of Grocery Stores	0.28974
Percent Adult Obesity	1.98738



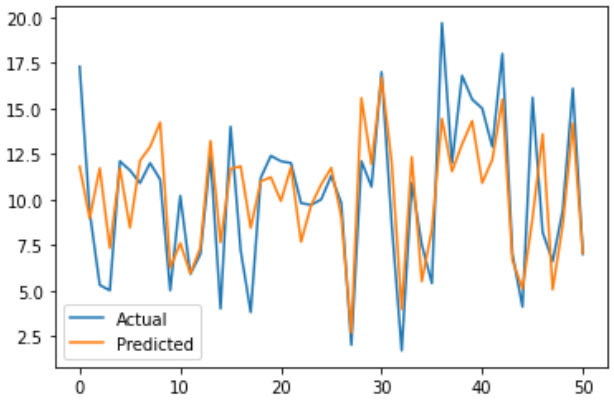
SVR: Linear - Training Set Accuracy



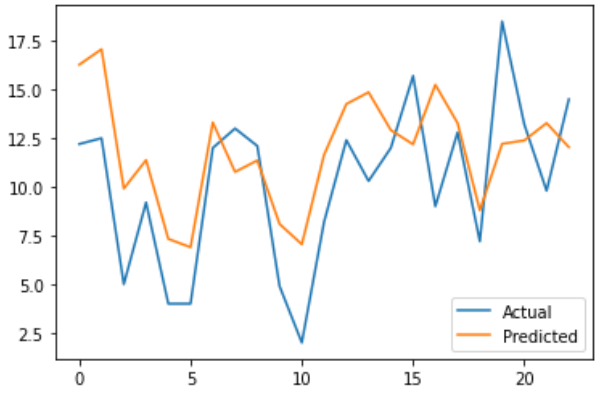
SVR: Linear - Training Set Accuracy

Ridge Regression	
Feature	**Score**
Square Milage	0.36461
Population	-10.24489
Percent use of Public - Transport	11.77182
Percentage Non-Hispanic White	-2.26138
Percentage Non-Hispanic Black or African American	-7.35754
Percentage Hispanic or Latino	-9.18160

Median Income	6.61936
Unemployment Rate	-4.92132
Food Stamp Usage	0.30923
Percentage Limited - Access to Food	-7.26857
Percentage Easy Access to Fruits and Vegetables	-7.26472
Number of Grocery Stores	7.61936
Percent Adult Obesity	2.85824



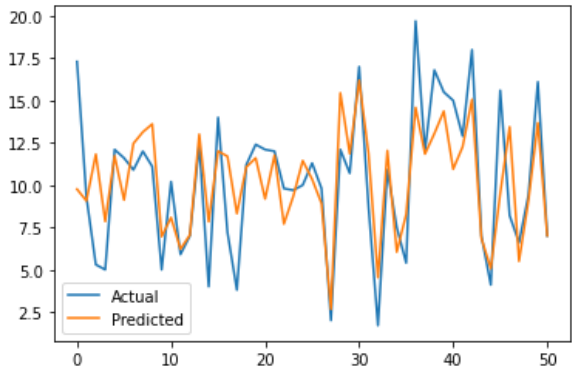
Ridge Regression - Training Set Accuracy



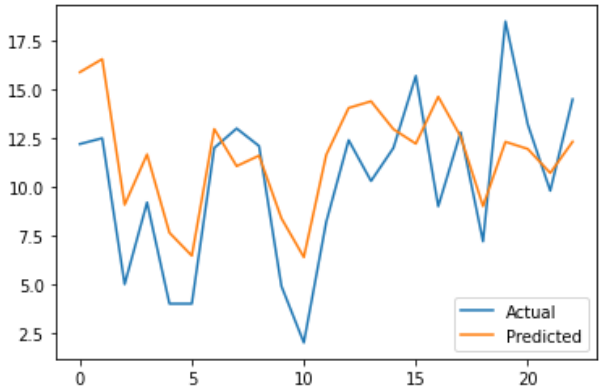
Ridge Regression - Training Set Accuracy

Lasso	**Regression**
Feature	**Score**
Square Milage	0.00000
Population	-2.04663
Percent use of Public - Transport	10.44928
Percentage Non-Hispanic White	-3.28397
Percentage Non-Hispanic Black or African American	-5.34245
Percentage Hispanic or Latino	-1.30007

Median Income	5.50140
Unemployment Rate	0.00000
Food Stamp Usage	-5.39966
Percentage Limited - Access to Food	-3.78046
Percentage Easy Access to Fruits and Vegetables	-0.00000
Number of Grocery Stores	4.73663
Percent Adult Obesity	2.80722



Lasso - Training Set Accuracy

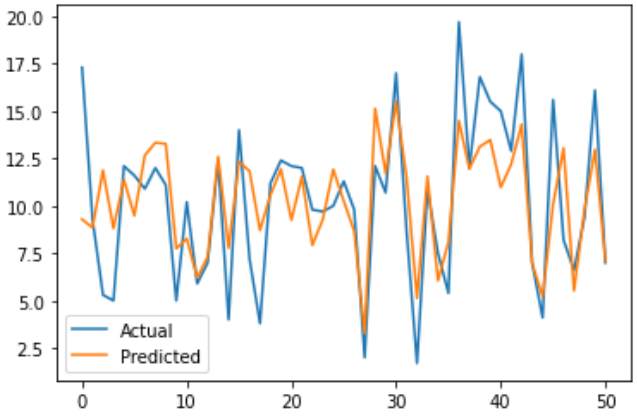


Lasso - Training Set Accuracy

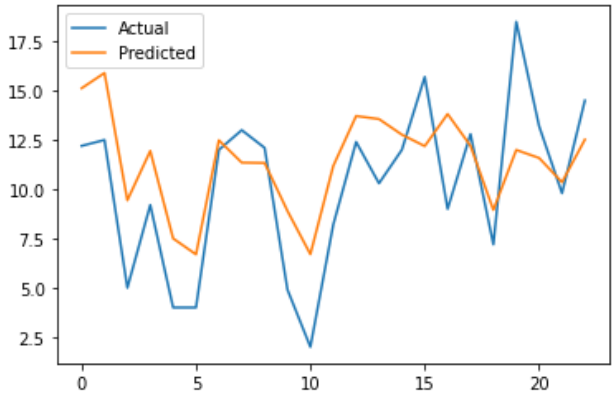
-

ElasticNet	**Regression**
Feature	**Score**
Square Milage	0.00000
Population	-0.89884
Percent use of Public - Transport	8.32261
Percentage Non-Hispanic White	-3.10952
Percentage Non-Hispanic Black or African American	-2.69750
Percentage Hispanic or Latino	-0.17894
Median Income	4.14247

Unemployment Rate	-3.62936
Food Stamp Usage	0.00000
Percentage Limited - Access to Food	-3.59694
Percentage Easy Access to Fruits and Vegetables	-0.38726
Number of Grocery Stores	2.44706
Percent Adult Obesity	2.46026



ElasticNet - Training Set Accuracy



ElasticNet - Testing Set Accuracy

In the SVR model (linear), Food Stamp Usage, Easy Access to Fruits and Vegetables, and Obesity Rate influence the prediction of diabetes rate most dramatically. In the Ridge, Lasso, and ElasticNet models, use of public transportation, median income, and number of grocery stores play the largest role in predicting diabetes rate.

Among all the models, a combination of Obesity Rate, Use of Active Transportation, Number of Grocery Stores, Use of Food Stamps generally have the largest importance among the variables in the training set. These variables are not concentrated in one grouping, rather, they represent a wide range of categories (health, infrastructure, SES, and food-related).

Conclusions

Statistics

The use of linear regression for general exploration and mediation analysis showcased a wide array of variables that influenced diabetes rate. This supports the idea that health outcomes are linked to many different factors, and especially those thought to be correlated to food deserts.

In general, this wide array of variables also indicates that food deserts are more complicated than simply the number of grocery stores nearby: access to food is complex, and is impacted by time, money, transportation, location, and culture. These factors are closely linked, and can't be separated when considering food deserts and their implications.

This linkage was further exemplified by conducting a mediation analysis. Proving that food-related variables serve as a mediator between socioeconomic status and health outcome showcases that access to food is directly influenced by socioeconomic factors like income, food stamp usage, and race, which in turn, greatly impact health outcomes.

When considering these variables together, there is a clear correlation between them and diabetes rate and obesity rate showcasing the relationship between food deserts and diet-related health outcomes.

Additionally, these processes worked to confirm the widely-understood relationship between obesity and diabetes.

Machine Learning Models

When considering machine learning models, it is clear that out of the regression models used, the Ridge, Lasso, and ElasticNet models were most accurate. This is likely due to: [SOMETHING ABOUT THE MODELS]]

The result of training set variations wasn't entirely straightforward. The average accuracies of the sets didn't vary drastically, as they all remained within the range of 20-40 percent accurate (which is fairly low). In addition, there were some confusing results, like the training sets without race variables yielding the highest average accuracies, which is not an intuitive result.

In general, it was clear that including Obesity Rate would heavily improve the accuracy of the model. Because obesity is not a food-desert variable, but rather, a diet-related outcome, it was controversial to include this variable in the first place. The fact that it so strongly influenced the models confirmed that it should likely not be used in the final machine learning model.

One of the most important conclusions of this project had to do with dataset size. The datasets that were considered were all dictated by Chicago Community Area, which there are 77 of. Though each data point represented large populations, the machine learning models were all trained on 77 different points, which is a very low amount. The accuracy and importance were somewhat inconsistent because of this: a small dataset means you have to compromise on both the size of the training and testing set. There is a lot of

room for error with smaller size. This could work to explain some of the inconsistencies in results (i.e., the training set without race variables doing so well).

Solutions

Knowing the food-desert variables that influence health outcomes like diabetes can help in the search for a solution in communities. Health is obviously an extremely complex subject, but it is evident there is a strong correlation between food deserts and diet-related health outcomes. Knowing which variables make up food-deserts can help in the evaluation of solutions.

For example, opening up grocery stores in communities that lack them is an overly simple solution when considering the other variables that contribute to food deserts. Factors that have to do with socioeconomic status, such as transportation access, time to cook and shop (due to jobs and/or childcare), food stamp eligibility at stores, etc have to be considered when discussing solutions.

Knowing that **Percent Active Transportation Use** contributes significantly to predicting diabetes rate in our model might mean that making online grocery shopping more accessible and affordable could be helpful to those in food deserts. Knowing that **Percent Food Stamp Usage** influences our model strongly could indicate that expanding the types and amounts of stores that accept food stamps could be beneficial to those living in food deserts.

These are just two small examples, but the underlying idea is that food deserts are a complex phenomenon made up of a wide array of related variables. Because these variables are so strongly correlated with several socioeconomic indicators, including race in Chicago, it's important to realize that it's impossible to reduce food deserts in Chicago without taking income and racial inequality under consideration.

Comparison with Oregon/ Portland

This project was completed alongside Sundari Arunarasu's similar exploration of Food Deserts and Health Outcomes in Oregon. Similar data, techniques, and tools were used to analyze the impact of SES, geographical, and food-related variables.

Sundari's project had several differences. First and foremostly, her analysis was done across all of Oregon's counties, and not just Portland, the largest city in Oregon. This meant that she was analyzing both urban and rural communities, while this project on Chicago only focused on urban areas. Additionally, her datasets also included health indicators like access to physicians. Exact variables also differed based upon availability.

In both cases, Ridge/Lasso regression models were most accurate in predicting diabetes rates. This could be due to the fact that these models work wells with datasets with multicollinearity, or correlations between predictor variables. The underlying idea of both of our projects was that a complex set of

variables dictate food deserts and diet-related outcomes, meaning the predictor variables would be correlated.

When comparing variable importance, Sundari found that the most influential variables were **Number of Food Services per Person, Obesity Rate, and Number of Primary Care Providers per Person.** Compared to Chicago's **Food Stamp Usage, Number of Grocery Stores, Obesity Rate, and Transportation,** we can see the similar wide variety of variables that impacts health outcomes in both cases.

Because Oregon is a predominantly white state, race plays a significantly smaller role in both food deserts and health outcomes than it does in Chicago.

Next Steps

Moving forward, objectives are as follows:

1. **Consider Ensemble Regression.** These models use multiple learning algorithms to increase prediction accuracy and are oftentimes more flexible.
2. **Consider ways to incorporate more data by potentially involving cities with similar characteristics.** This could involve searching for cities with similar racial and income distributions and histories. More data would help better identify the most important variables when considering food deserts and their diet-related impacts. However, it would also make the results less helpful to Chicago specifically.
3. **Customize existing models to be more accurate.** Expand past the default settings offered by Python packages.
4. **Use models and their variable importance to evaluate potential solutions**

Work Cited

Info about food deserts

Kolak, M., Bradley, M., Block, D. R., Pool, L., Garg, G., Toman, C. K., . . . Wolf, M. (2018). Urban foodscape trends: Disparities in healthy food access in Chicago, 2007–2014. Health & Place, 52, 231-239. doi:10.1016/j.healthplace.2018.06.003

Jansen, T., Aguayo, L., Whitacre, J., Bobitt, J., Payne, L., & Schwingel, A. (2019). Diabetes Disparities In Illinois. *Preventing Chronic Disease, 16*. doi:10.5888/pcd16.180154

Paykin, S., », S., Howes, J., & Bose, M. (2019, April 16). Chicago Food: More supermarkets do not mean healthy food for all. Retrieved July 18, 2020, from https://chicagopolicyreview.org/2019/04/17/chicago-food-more-supermarkets-do-not-mean-healthy-food-for-all/

Pilgrim, G., Yang, C., & Nwoku, K. (n.d.). Mind the Gap: What if we treated zip codes before treating disease? Retrieved July 18, 2020, from https://www.whartonhealthcare.org/mind_the_gap_what_if_we_treated_zip_codes_before_treating_disease

Castro, M., Thernstrom, A., Achtenberg, R., Gaziano, T. F., Heriot, G., Kirsanow, P. N., . . . Yaki, M. (2011). *Food Deserts in Chicago: A Report of the Illinois Advisory Committee to the United States Comission on Civil Rights* (Rep.). Chicago, IL: U.S Commission on Civil Rights.

Datasets

Milberger, S. (2002). *Evaluation of violence against women with physical disabilities in Michigan, 2000-2001* (ICPSR version) [data file and codebook]. doi:10.3886/ICPSR03414

City of Chicago (2018). *Map of Urban Farms.* Chicago Data Portal.

City of Chicago (2013). *Map of Food Carts.* Chicago Data Portal.

Chicago Health Atlas (2016-2018). *Diabetes: Adults Diagnosed with Diabetes (excluding gestational and pred-diabetes).* *https://www.chicagohealthatlas.org/indicators/diabetes*

> *Illinois Department of Public Health, Behavioral Risk Factor Surveillance System (2000-2009); Chicago Department of Public Health, Healthy Chicago Survey (2014-2016)*

Chicago Health Atlas (2017). *Diet-related deaths: people who died due to nutrition and obesity related cases.* https://www.chicagohealthatlas.org/indicators/diet-related-deaths

> Source: *Illinois Department of Public Health, Division of Vital Records, Death Certificate Data File*

Illinois Department of Public Health, Division of Vital Records, Death Certificate Data Files; US Census Bureau, 2010 and 2000 Census; Intercensal years (2001-2009) estimated through linear interpolation by the Office of Epidemiology, Chicago Department of Public Health; For 2011-2014, 2010 Census population was used.

Chicago Health Atlas (2013-2017). *Diabetes deaths: people who died due to diabetes.* . https://www.chicagohealthatlas.org/indicators/diabetes-deaths

Illinois Department of Public Health, Division of Vital Records, Death Certificate Data Files

Illinois Department of Public Health, Division of Vital Records, Death Certificate Data Files; US Census Bureau, 2010 and 2000 Census; Intercensal years (2001-2009) estimated through linear interpolation by the Office of Epidemiology, Chicago Department of Public Health; For 2011-2014, 2010 Census population was used.

Chicago Health Atlas (2013-2017). *Active transportation: people who walk, bike, or take public transportation to commute to work* . *https://www.chicagohealthatlas.org/indicators/active-transportation*

US Census Bureau: American Community Survey 2010 5-year estimates (census and community area), 2015 5-year estimates (census and community area), 2010-2015 1-year estimates for Chicago

Source: US Census Bureau: American Community Survey 2010 5-year estimates (census and community area), 2015 5-year estimates (census and community area), 2010-2015 1-year estimates for Chicago

Chicago Health Atlas (2013-2017). *Easy Access to fruits and vegetables: Adults who find it very easy to get fruits and vegetables.* *https://www.chicagohealthatlas.org/indicators/easy-access-to-fruits-and-vegetables*

Chicago Health Atlas (2016-2018). *Adult Obesity: Adults with a body mass index (BMI) of 30 or greater.* *https://www.chicagohealthatlas.org/indicators/adult-obesity*

Illinois Department of Public Health, Behavioral Risk Factor Surveillance System (2000-2009); Chicago Department of Public Health, Healthy Chicago Survey (2014-2016)

Source: Illinois Department of Public Health, Behavioral Risk Factor Surveillance System (2000-2009); Chicago Department of Public Health, Healthy Chicago Survey (2014-2016)

Chicago Health Atlas (2012-2016). *Unemployment: Civilian Population aged 16 years and older who were unemployed.* *https://www.chicagohealthatlas.org/indicators/unemployment*.

Source: *Illinois Department of Public Health, Vital Statistics*

Source: Illinois Department of Public Health, Vital Statistics; US Census Bureau 2000 and 2010 population and populations for 2001-2009 were interpolated

Chicago Health Atlas (2012-2016). *Food Stamps/SNAP: Households receiving food assistance.* . *https://www.chicagohealthatlas.org/indicators/food-stamps-snap*

US Census Bureau: American Community Survey 2010 5-year estimates (census and community area), 2015 5-year estimates (census and community area), 2010-2015 1-year estimates for Chicago and race groups

Source: US Census Bureau: American Community Survey 2010 5-year estimates (census and community area), 2015 5-year estimates (census and community area), 2010-2015 1-year estimates for Chicago and race groups

Chicago Health Atlas (2015-2016). *Household food insecurity: households without access to affordable, nutritious food in the past year.* *https://www.chicagohealthatlas.org/indicators/household-food-insecurity*.

Sinai Community Health Survey 2.0, 2015-2016 (www.sinaisurvey.org); 2016 American Community Survey 5-year estimates (2012-2016)

Chicago Health Atlas (2015-2016). *Household emergency food use: households that accessed emergency food in the past year.* *https://www.chicagohealthatlas.org/indicators/household-emergency-food-use*

Sinai Community Health Survey 2.0, 2015-2016 (www.sinaisurvey.org); 2016 American Community Survey 5-year estimates (2012-2016)

Chicago Health Atlas (2015). *Limited food access: low income residents with limited access to affordable, healthy foods.* . *https://www.chicagohealthatlas.org/indicators/limited-food-access*

Source: *United States Department of Agriculture (USDA) Food Access Research Atlas*

Source: *United States Department of Agriculture (USDA) Food Access Research Atlas*

Chicago Health Atlas (2015). *Grocery Stores per Community Area.* *https://www.chicagohealthatlas.org/indicators/grocery-stores-per-community-area*

Source: *Chicago Open Data Portal*

Chicago Health Atlas (2012-2016). *Household Poverty: household whose income is below the poverty level.* https://www.chicagohealthatlas.org/indicators/household-poverty

Source: *US Census Bureau: American Community Survey 2010 5-year estimates (census and community area), 2015 5-year estimates (census and community area), 2010-2015 1-year estimates for Chicago and race groups*