

Capstone Project Report

By Xinyu Zhang

Table of Contents

Motivations and Hypothesis	3
Data Collection	3
Small Area Health Insurance Estimates	4
Cleaning and Unifying Data	5
Small Area Health Insurance Estimates	5
Cleaning food environment atlas data	5
Cleaning Indicators of Diabetes, Obesity and Physical Inactivity data	6
Cleaning and Integrate Coverage of Insurance Datasets	7
Cleaning Rural-urban Categorization Data	8
Unifying Data	8
Exploratory Data Analysis (EDA)	8
Mapping Variables by US Counties	8
Distribution of Diabetes, obesity prevalence by Urban-rural Categorization	12
Correlation Among Important Features	13
Clustering Analysis	14
Data preprocessing	14
Principal Component Analysis (PCA)	15
Hierarchical clustering analysis of all US counties	16
Hierarchical clustering analysis of the most diabetic and obese counties	21
Regression Analysis	25
Data preprocessing	25
Cross Validation	26
Regression Analysis Results	27
Summary	35

Motivations and Hypothesis

Food environment has been considered one important determinant of population health. Food choice determines the nutrition intake and in turn affects people's health. However, there are also other factors that take effects as well, such as extent of physical activities and medical care.

Diabetes and obesity are considered to highly relevant to food choice and physical activity.¹⁻³ Research shows that physical activity reduces the risk of type II diabetes and obesity while poor food choice can increase risk of both conditions.¹⁻³ In this project, I am interested in how food environment, physical activity and medical insurance affect prevalence of diabetes and obesity in US counties.

US counties differ in multiple dimensions. Population in each county have a unique combination of sociodemographic status, and culture. It is thus reasonable to infer that counties where diabetes and obesity are highly prevalent may differ in these characteristics.

In summary, in this project I want to explore following questions,

- i. How do food environment, physical activity and availability of medical insurance differ among US counties?
- ii. What similarities and difference do US counties have? Can we categorize them into distinct groups?
- iii. Are food environment, physical activity and availability of medical insurance important determinants of prevalence of diabetes and obesity?
- iv. Do food environment, physical activity and availability of medical insurance similarly impact prevalence of diabetes and obesity?

Data Collection

The whole analysis is designed at US county level. Therefore, only data that are available for most US counties are collected. Briefly, they are,

- i. food environment of US counties such as access to grocery stores, food price, available farms and number of different types of stores
- ii. prevalence of diabetes, obesity and physical inactivity of US counties
- iii. coverage of medical insurance of US counties
- iv. urban-rural classification of US counties

The names, source (link) brief description and contents of these datasets are summarized in table 1.

Table 1. names, source (link), description and contents of the datasets

Datasets name	source	description	contents
Food environment atlas data	link: https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/	A collection of datasets that is publicly available on United States Department of Agriculture Economic Research Service website. It contains nine datasets that provide the information that defines the food environment of US counties.	All nine datasets are collected for analysis (data at US county level): <ul style="list-style-type: none"> - Access and proximity to grocery store - Availability of different types of stores - Availability and expenditures of different types of restaurants - Food assistance - State food insecurity - Food prices and taxes - Types and availability of local foods - Data regarding health and physical activity - Socioeconomic characteristics
US county data indicators of diabetes, obesity and physical inactivity	Link: https://www.cdc.gov/diabetes/data/countydata/countydataindicators.html	A collection of datasets publicly available on CDC website.	Three datasets are collected for analysis: <ul style="list-style-type: none"> - Estimated prevalence of diabetes of US counties from 2004-2013 - Prevalence of obesity of US counties from 2004-2013 - Prevalence of physical inactivity of US counties from 2004-2013
Small Area Health Insurance Estimates	Link: https://www.census.gov/data/datasets/time-series/demo/sahie/estimates-acs.html	A collection of datasets publicly available on data.gov	Time series of estimated number of uninsured population of US counties
Rural-Urban Continuum Codes	Link: https://www.census.gov/geo/reference/urban-rural.html	This dataset is publicly available on United States Census Bureau website. It provides the categorization of counties along the spectrum from urban to rural based on criteria such as level of development, residential and commercial land use and population size, etc.	Urban-rural categorization of US counties

Cleaning and Unifying Data

The goal of data preprocessing is to clean data for missing values, identify any inaccurate value input and integrate the multiple datasets into one. The data preprocessing is described as steps of data cleaning of each collection of dataset and unifying data into a single dataset for analysis.

Firstly, each collection of data as listed in table 1 are cleaned and saved as a .csv file (Listed in Table 2).

Table 2. Source of data and corresponding cleaned dataset

Original data source	Cleaned dataset
Food environment atlas data link: https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/	food_environment.csv with FIPS code as index column for data integration Github link: https://raw.githubusercontent.com/cathyxinyz/Capstone_Project_1/master/Datasets/food_environment.csv
US county data indicators of diabetes, obesity and physical inactivity link: https://www.cdc.gov/diabetes/data/countydata/countydataindicators.html	Db_ob_phy.csv with FIPS code as index column for data integration Github link: https://raw.githubusercontent.com/cathyxinyz/Capstone_Project_1/master/Datasets/Db_ob_phy.csv
Small Area Health Insurance Estimates link: https://www.census.gov/data/datasets/time-series/demo/sahie/estimates-acs.html	Uninsured.csv with FIPS code as index column for data integration Github link: https://raw.githubusercontent.com/cathyxinyz/Capstone_Project_1/master/Datasets/Uninsured.csv
Rural-Urban Continuum Codes Link: https://www.census.gov/geo/reference/urban-rural.html	Rural_urban.csv with FIPS code as index column for data integration Github link: https://raw.githubusercontent.com/cathyxinyz/Capstone_Project_1/master/Datasets/Rural_urban.csv

In the following sections, I describe the details of how each collection of datasets listed in table 1 are cleaned and integrated into the .csv files listed in table 2.

Cleaning food environment atlas data

The food environment atlas data contains nine datasets (as shown in Table 1). Preprocessing them one by one would be tedious. Therefore, I first integrated these nine datasets are integrated into one .csv file- **food_environment.csv**.

These datasets all share a common column: the Federal Information Processing Standard (FIPS) Code of US counties. It is a five-digit code with the first two digits as the state code and the last three digits as the county code. Each county has a unique FIPS code. Given its simple form and unique correspondence to each county, I set it as the index of each dataset. These nine datasets are then integrated using outer join.

There are in total 277 variables in the integrated dataset. Some of them has missing values.

I first dropped the columns with more than 5% missing values. This percentage is arbitrarily chosen based on two criteria: firstly, it is not too small to guarantee that plenty of potentially valuable variables are included for analysis; secondly, it is not too large to guarantee that imputing missing values would not introduce considerable bias.

The next step is to single out the variables that are useful for analysis. According to the variable descriptions provided by the USDA website the food environment atlas data largely have three types of variables:

- a. absolute number, such as number of people with poor access to grocery stores
- b. percentage, such as fraction of people with low income
- c. number per 1000 people
- d. change in numbers or percentage over time

Type b, c variables are most appropriate and kept for further data cleaning. In contrast, type a does not account for the population size: one person out of ten is far from grocery store is different from one person out of a thousand far from grocery store. Type d is not relevant to the analysis proposed in this project. It would be more appropriate for analysis of changing trend of food environment which could be the direction of future projects.

The next step is to check for any irrational input. For example, variables which are “percentage” cannot be negative or larger than 100, while variables which are number per 1000 people cannot have negative value or be greater than 1000. Any irrational input would be changed to nan values.

Some of the statistics are measured in different years. The year of measurement is indicated by the last two digits or last four digits of the variable name. The average of a statistic across different periods would better represent what a county has experienced to have the current prevalence of diabetes or obesity than a single measurement. Therefore, I first group the same type of statistics measured in different years together, and calculated the average of each statistics as a new column.

Through these steps of data cleaning and integration, I got the food environment dataset where there are 52 variables for analysis. It is saved as the food_enviroment.csv file (Table 2). It has a column of counties’ FIPS codes so that it is ready to be unified with other datasets.

Cleaning Indicators of Diabetes, Obesity and Physical Inactivity data

There are nine datasets in the collection of US county data indicators of diabetes, obesity and physical inactivity (Table 1). I chose three of them: prevalence of diabetes of US counties from 2004-2013, prevalence of obesity of US counties from 2004-2013, and prevalence of physical inactivity of US counties from 2004-2013.

I am interested in most recent prevalence of diabetes and obesity which best approximate their current states (data not available). Therefore, I extracted the columns of prevalence of

diabetes and obesity in US counties in 2013 as series. In addition, I extracted the columns of prevalence of physical inactivity over the most recent three years (2011-2013) and create a new column as the average of them. This new column represents the level of physical inactivity that US counties have experienced in recent years.

Then I investigated the missing values for each variable. In this collection of datasets, missing values are coded as “No Data”. I replaced “No Data” with nan. In addition, given prevalence is percentage it should be between 0 and 100. Therefore, any values that are negative or greater than 100 are also replaced with nan values.

In summary, the cleaned dataset has three variables: prevalence of diabetes of US counties, prevalence of obesity of US counties and prevalence of physical inactivity averaged over most recent three years of US counties. I saved the cleaned and integrated dataset as **Db_ob_phy.csv** file with FIPS code column that will be used for further data integration.

Cleaning and Integrate Coverage of Insurance Datasets

There are eight datasets in the collection of Small Area Health Insurance Estimates (SAHIE) data, which are the annual estimation of uninsured population of US counties from 2008 to 2015. Given that the prevalence of diabetes and obesity, which are potential dependent variables, are estimation of 2013, I only consider the SAHIE data by 2013. I choose SAHIE data at year 2011, 2012 and 2013.

Dataset in each year has a column which is the number of uninsured population of US counties. However, this number is measured for different age groups, gender, race/ethnicity groups, and people with different income levels. Which subpopulation this number is measured for are indicated by other columns of categorical variables.

I am interested in the most general measures: the number of uninsured population of US counties under 65 years old, including both genders, all races, and all income levels. Therefore, I wrote a function to only select relevant rows.

However, number of uninsured population is not a justified measure since population size is not adjusted for. Fortunately, population size of US counties in year 2011, 2012 and 2013 can be found in Food atlas environment dataset. Therefore, I need to merge the selected rows and columns in SAHIE dataset and population size columns from food atlas environment dataset. However, the demographic file in Food atlas dataset does not have FIPS code column. Instead, there is a state code column and a county code column. I need to combine them to be the FIPS code.

Then I am able to merge the dataset with number of uninsured population and dataset with population size by their common column: FIPS codes.

I calculated the fraction of uninsured population under 65 years old for US counties in year 2011, 2012, and 2013. Then I calculate their average which I save as the only column in the final cleaned dataset.

The dataset with the fraction of uninsured population of US counties (average of 2011-2013) is saved as Uninsured.csv.

Cleaning Rural-urban Categorization Data

I extracted the FIPS code column and the rural-urban category column from the Rural-Urban Continuum Codes data, and save it as Rural_urban.csv.

Unifying Data

Given that all .csv files share the FIPS code column, I read in every .csv file using pandas read_csv function and set 'FIPS' as the index column. Then I integrate all data together row wisely using inner join.

Exploratory Data Analysis (EDA)

Mapping Variables by US Counties

Maps of several variables by US counties show that prevalence of diabetes and obesity, poverty rate, physical inactivity and access to grocery stores all greatly differ among US counties.

Firstly, counties in Southeast states have high prevalence of diabetes and obesity (Figure 1, 2, 3). These states include Oklahoma, Arkansas, Tennessee, Mississippi, Alabama and Kentucky. However, population in these counties have moderate level of access and proximity to grocery stores.

In addition, counties in Colorado, states along Pacific Ocean shoreline, Atlantic Ocean shoreline, or Great Lake shoreline have low prevalence of diabetes, obesity and physical inactivity (Figure 1, 2, and 3).

There are also some counties where prevalence of diabetes is high but obesity prevalence is relatively low. For example, some counties in Texas and Florida are more vulnerable to diabetes than to obesity (Figure 1, 2, 3). The analysis show that these counties tend to have high fraction of elderly people (≥ 65 years old).

There are also some counties where prevalence of obesity is high but diabetes prevalence is relatively low. For example, some counties in Alaska are more vulnerable to obesity than to diabetes (Figure 1, 2, 3). The analysis who that these counties tend to have high fraction of young people < 18 years old).

Furthermore, there are neighbor counties that greatly differ in population health and level of physical activity. For example, counties in northeast of Texas have lower prevalence of diabetes, obesity and physical inactivity as compared to their adjacent counties in Oklahoma, Arkansas, and Louisiana. This indicates that geographical location cannot fully explain the level of physical inactivity which may considerably affect population health (Figure 1 and 2).

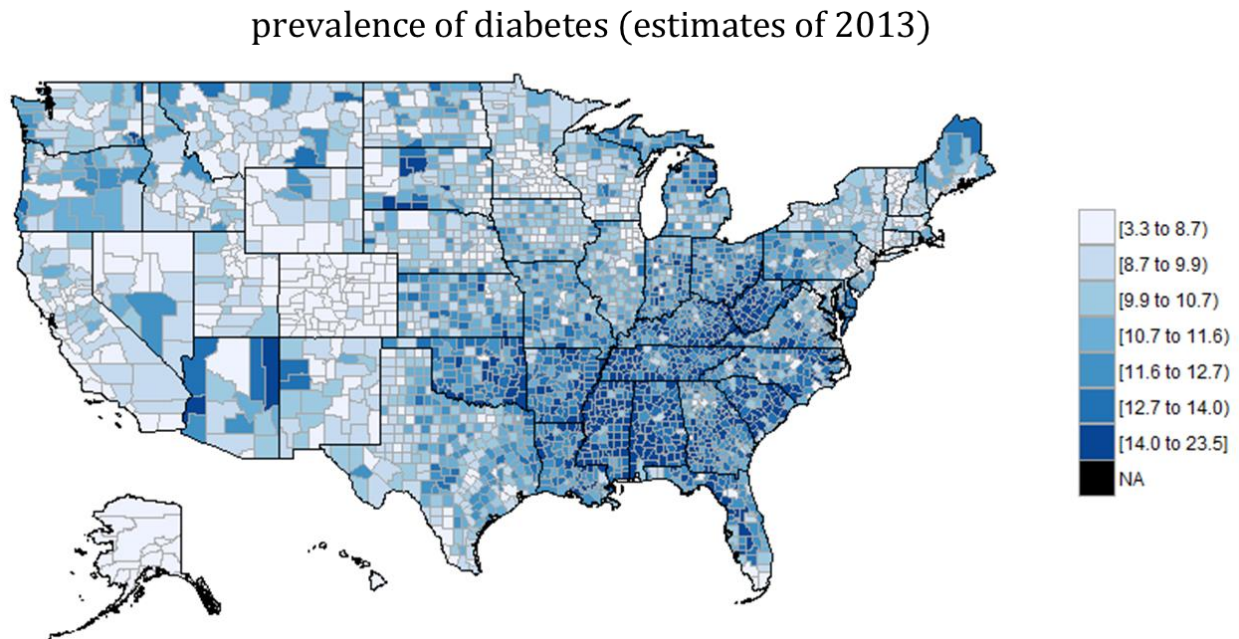


Figure 1. Map of diabetes prevalence (year 2013) of US counties

prevalence of obesity (estimates of 2013)

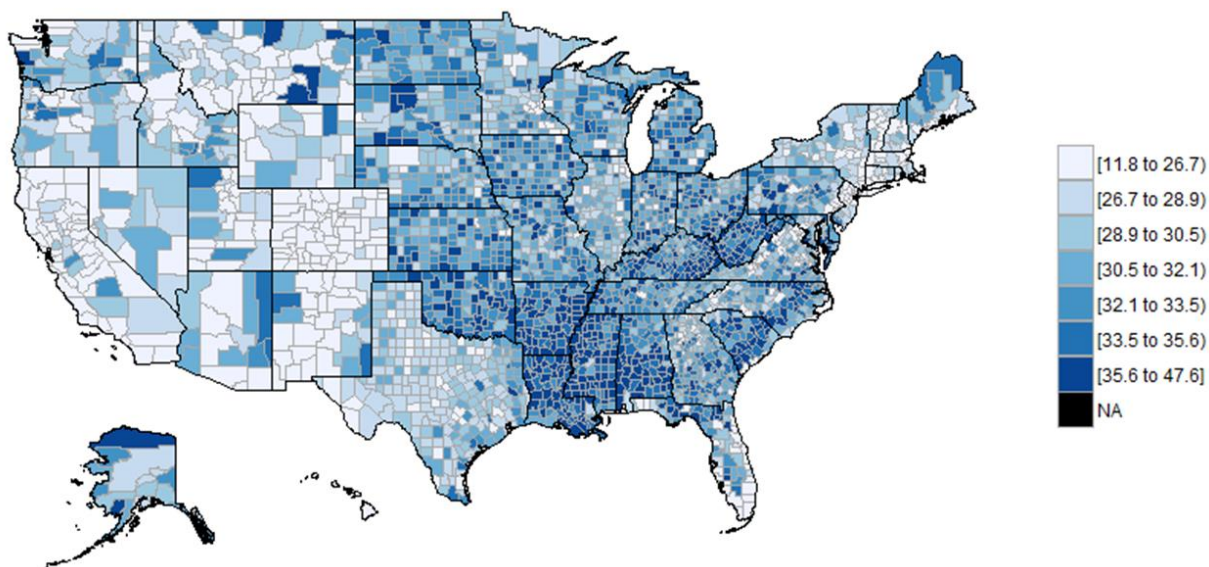


Figure 2. Map of obesity prevalence (year 2013) of US counties

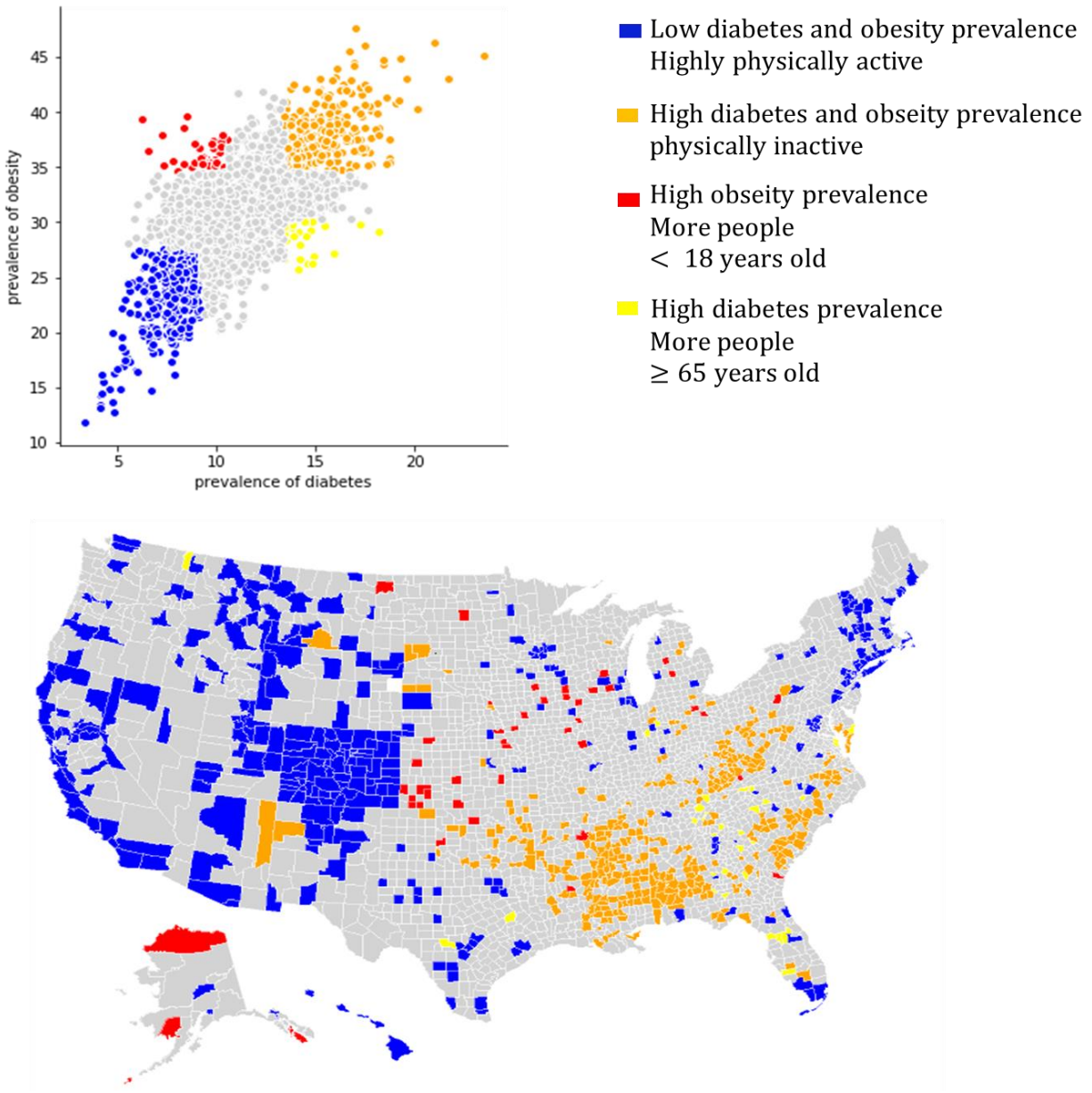


Figure 3. Upper panel: Scatter plot of prevalence of diabetes and prevalence of obesity of US counties. Colors mark counties with different combinations of prevalence of diabetes and prevalence of obesity. Blue: counties with lowest 20% of diabetes prevalence and lowest 20% of obesity prevalence; Orange: counties with highest 20% of diabetes prevalence and highest 20% of obesity prevalence; Red: counties with highest 20% of obesity prevalence and lowest 40% of diabetes prevalence; Yellow: counties with highest 20% of diabetes prevalence and lowest 40% of obesity prevalence.

Distribution of Diabetes, obesity prevalence by Urban-rural Categorization

It is shown that prevalence of diabetes, obesity and physical inactivity all greatly differ among US counties. My next question is that does this have anything to do with the degree of urbanization? Would people in urban area have lower prevalence of diabetes or obesity than people in more rural area or the opposite?

To explore this question, I make boxplots of prevalence of diabetes and obesity by the urban/rural categorization of US counties. In addition, I do permutation test to see whether distribution are statistically significantly different among different categories of counties.

Figure 4 and Figure 5 shows that there is no considerable difference in the prevalence of obesity or prevalence of obesity between urban areas and rural areas. Permutation test also confirms that there is no statistical significant difference between any two urban/rural categories of counties.

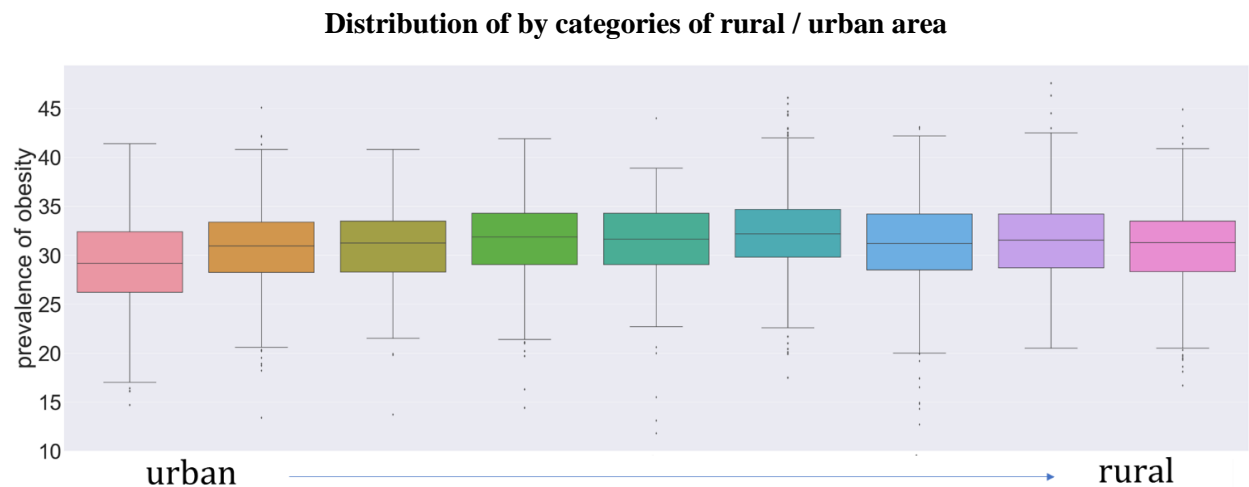


Figure 4. Prevalence of diabetes by urban-rural categorization (types of areas changes from urban to rural from left to right side)

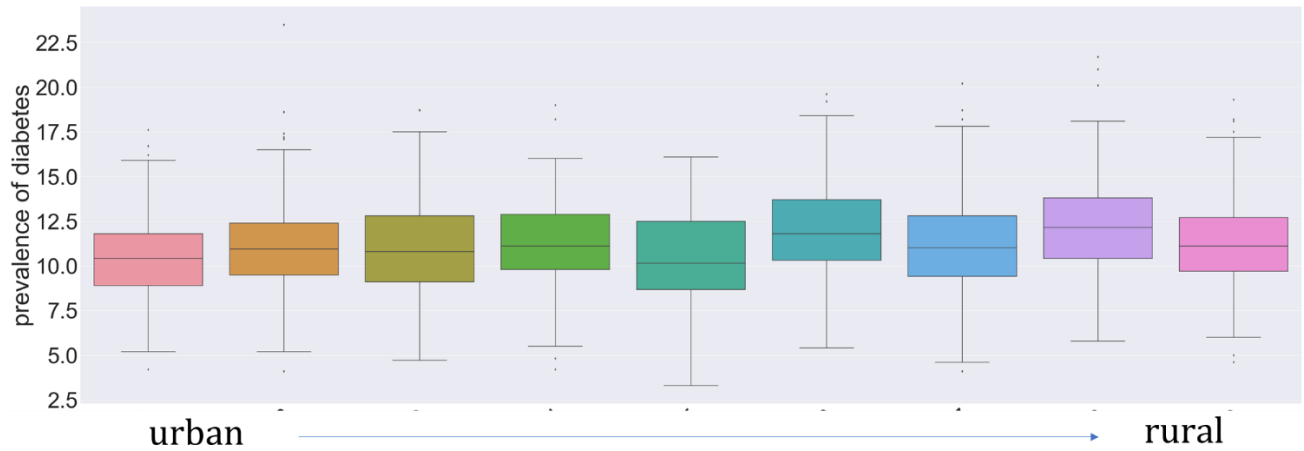


Figure 5. Prevalence of diabetes by urban-rural categorization (types of areas changes from urban to rural from left to right side)

Correlation Among Important Features

The correlation matrix among several important features show that most of these features have statistically significant correlation (Figure 6). Since this is a set of multiple tests, a Bonferroni correction is used to avoid Type I Error. That says, the maximum p value that indicates statistical significance is corrected as $0.05/n$, where n is the number of tests. Here there are 10 variables among which correlations are examined. Therefore, there are in total $10 \times (10-1)/2 = 45$ pairs for correlation test, or $n=45$. Therefore, Bonferroni corrected p value is $0.05/45 = 0.0011$.

For example, prevalence of diabetes, prevalence of obesity, prevalence of physical inactivity and fraction of uninsured population (≤ 65 years) are highly positively correlated with each other. All four are also highly correlated with poverty rate.

In addition, fraction of Black population is positively correlated with diabetes, obesity, physical inactivity and poverty rate. By contrast, fraction of White population is negatively correlated with these features.

Fraction of population above 65 years old is positively correlated with prevalence of diabetes but negatively correlated with obesity. By contrast, fraction of population under 18 years old is positively correlated with prevalence of obesity but negatively correlated with prevalence of diabetes.

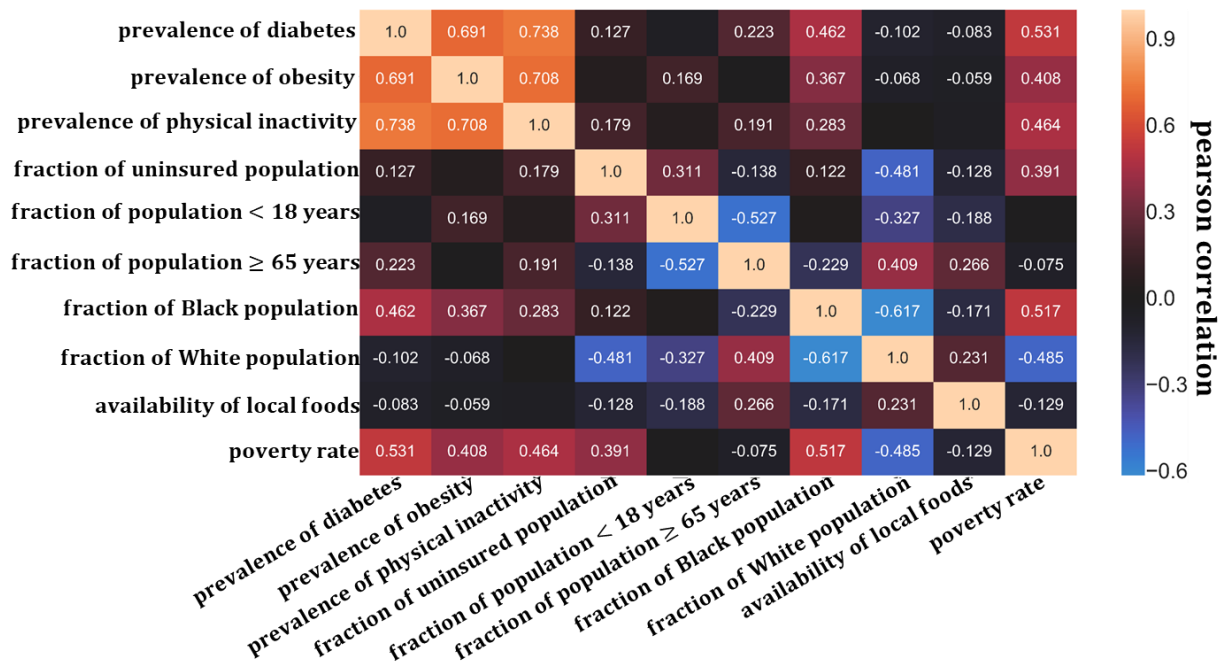


Figure 6. Correlation matrix of important features, statistically significant correlation coefficients are marked in corresponding cells, where

Clustering Analysis

The exploratory data analysis shows that US counties differ in many dimensions such as population health, food environment, physical activity and medical care. However, can one find similarities among these counties? Or can one put these over 3000 counties into several groups where counties within each group share some common features? This question is worth exploring since answering it helps simplify the US county atlas and identifying the important features to summarize and distinguish the 3000 counties.

In addition, it is also worth exploring that whether the counties where diabetes and obesity are highly prevalent are largely similar or different. For example, do they all have high poverty rate, or poor physical activity?

To answer both questions, I use clustering analysis, which can put counties into different groups based on their features (variables in the dataset). Specifically, I use hierarchical clustering method. It can give a complete picture of the similarity/difference among counties.

To do the clustering analysis, I first preprocess the data using the cleaned and unified dataset. Then I do principal component analysis to reduce the dimensionality and control for the impact of multicollinearity. Finally, I do hierarchical clustering on all US counties and counties with the highest prevalence of diabetes and obesity.

Data preprocessing

The data preprocessing is done in two steps: inspecting missing value, normalizing data.

Hierarchical clustering method cannot be used for dataset with missing values. Therefore, the first step is to inspect the columns and rows of missing values. I keep all of the numeric variables but only drop the categorical columns including state names, county names and urban/rural categorization. This reduces the number of features to 52. Then I drop the rows with missing values, so that 3114 counties out of 3143 are retained for clustering analysis.

Clustering analysis requires data to be normalized to adjust for the impact of variables with large variance. I normalize each column to have mean at zero and variance around 1. The calculation is done with `sklearn.preprocessing.scale` module.

Principal Component Analysis (PCA)

Exploratory data analysis suggest that there is multicollinearity among these features. Therefore, I do principal component analysis (PCA) to reduce the dimensionality and control for the impact of multicollinearity on clustering analysis.

I use `sklearn.decomposition.PCA` function in Python.

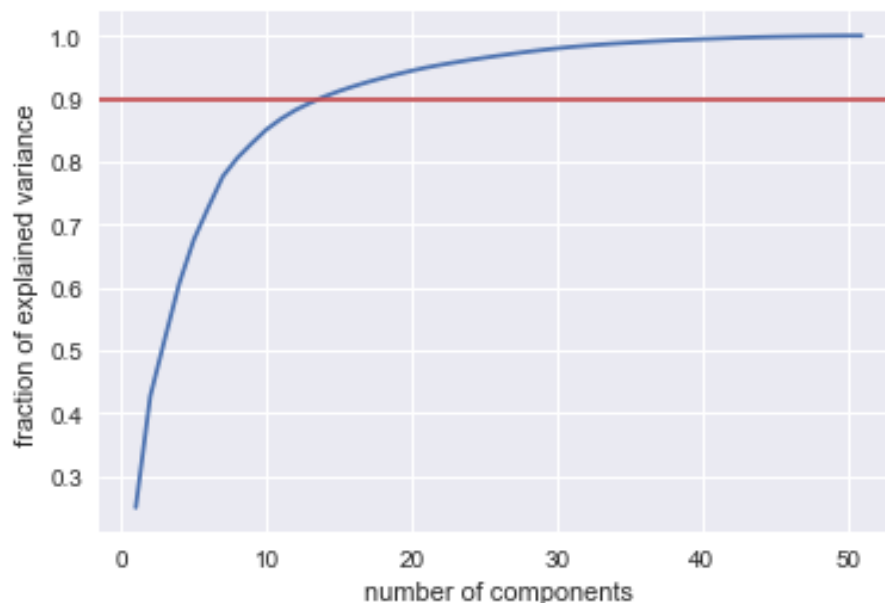


Figure 7. Fraction of explained variance by number of components

The results suggest that 14 to 15 components can explain 90% of the variance (Figure 7). It is considered a large enough fraction of variance that can be explained by a small set of features (14 ~ 15 out of 52). Therefore, I choose the 15 components and transform the dataset correspondingly.

Hierarchical clustering analysis of all US counties

Hierarchical clustering is done based on the distance among the subjects to be clustered.

Hierarchical clustering use bottom-up approach. Subjects with small distances from each other are clustered first. Then small clusters close to each other (in terms of average, median or maximum distance depending on the method used) form larger cluster, until all subjects fall into a single cluster. It produces a tree-like dendrogram, where clusters at specific cutoff distance are visualized as leaves descending from the same branch.

There are several methods to compute the distance between clusters, such as ‘average’, ‘complete’, ‘centroid’ etc. Different methods of computing distance can result in quite different dendrograms. I tested every method and pick one that gives clearly structured dendrogram where clustered subjects are tightly related and there is also clear distinction between clusters. It turns out that ‘complete’ distance is the best method in this regard.

The dendrogram is shown in Figure 8. I choose the cutoff distance to cut the dendrogram into eight clusters (dashed line in Figure 8). These eight clusters can be distinguished by at least one of the following factors: race/ethnicity composition, age distribution, poverty rate, physical activity, and fraction of uninsured population.

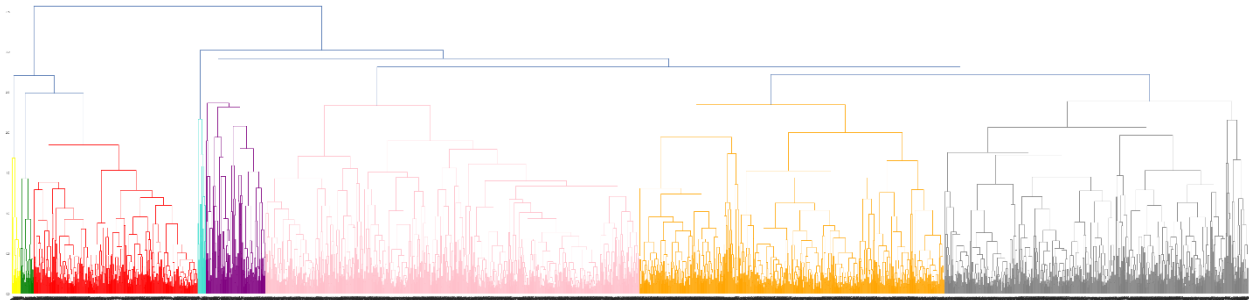


Figure 8. Dendrogram of hierarchical clusters

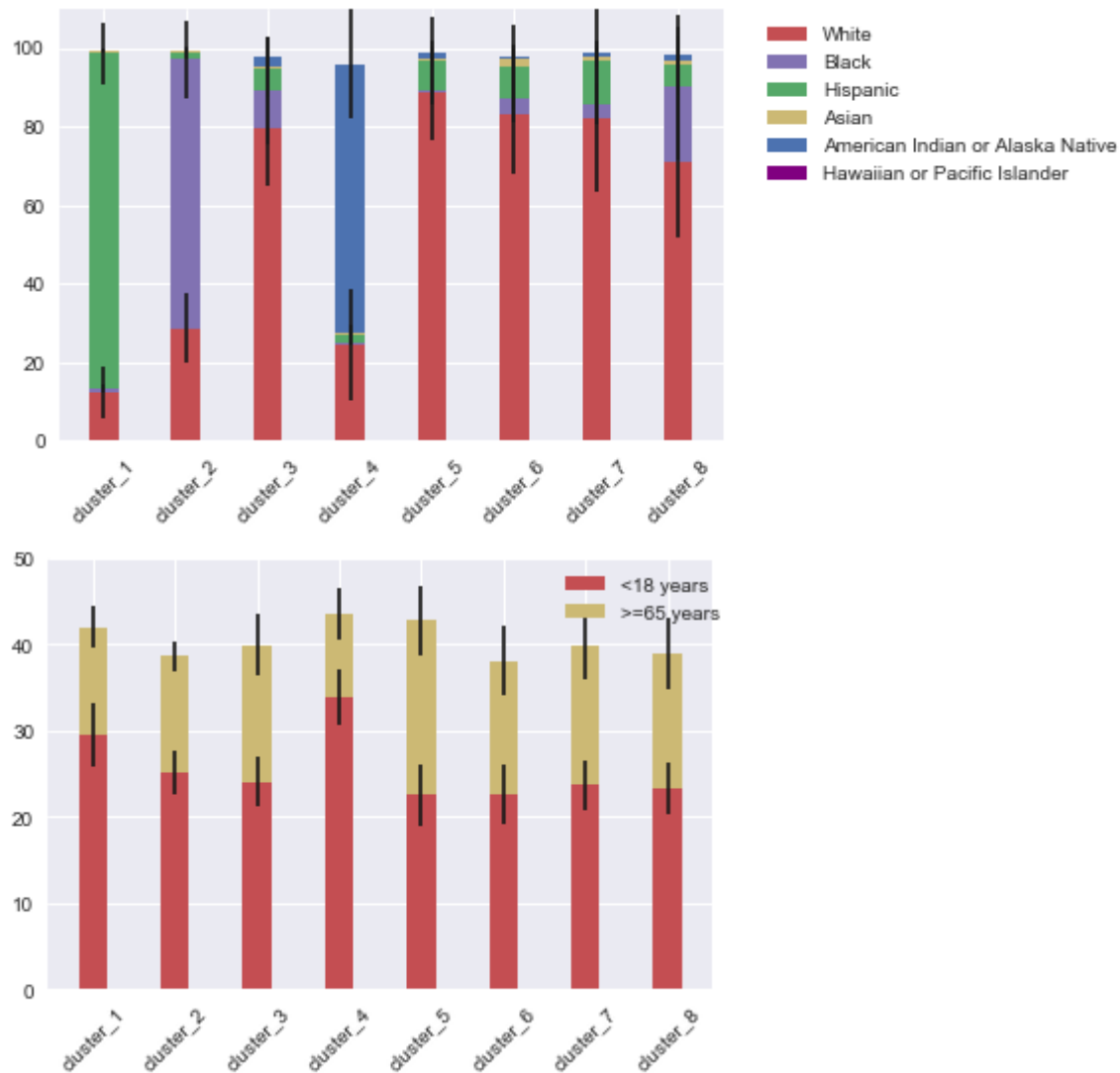


Figure 9. Race composition (upper panel) and age composition (lower panel) of clusters of US counties identified by hierarchical clustering analysis.

Major Features of Eight Groups

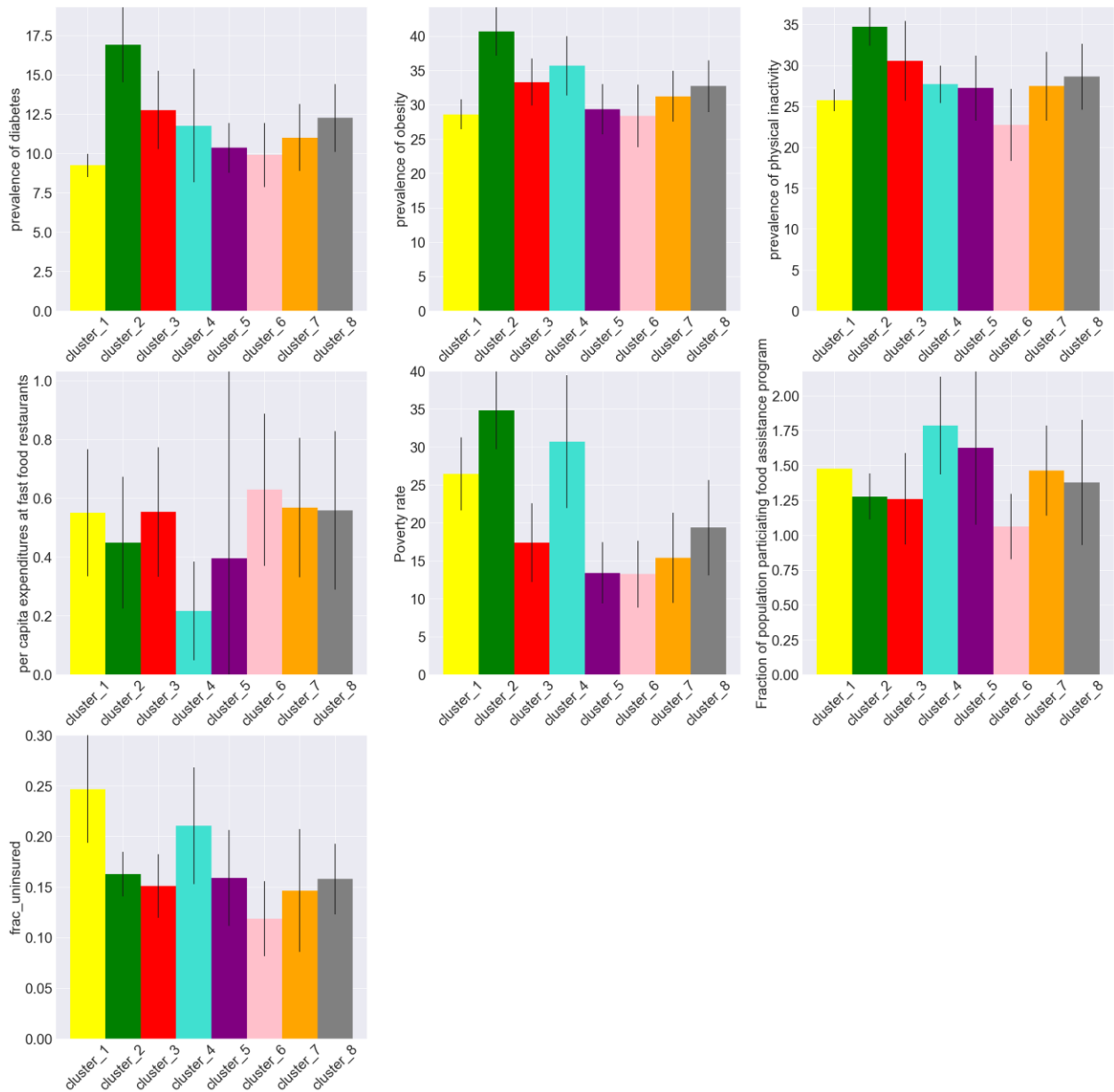


Figure 10. Prevalence of diabetes, obesity and physical inactivity, Food environment, poverty rate and fraction of uninsured of clusters identified by hierarchical clustering.

eight clusters identified by hierachical clustering analysis

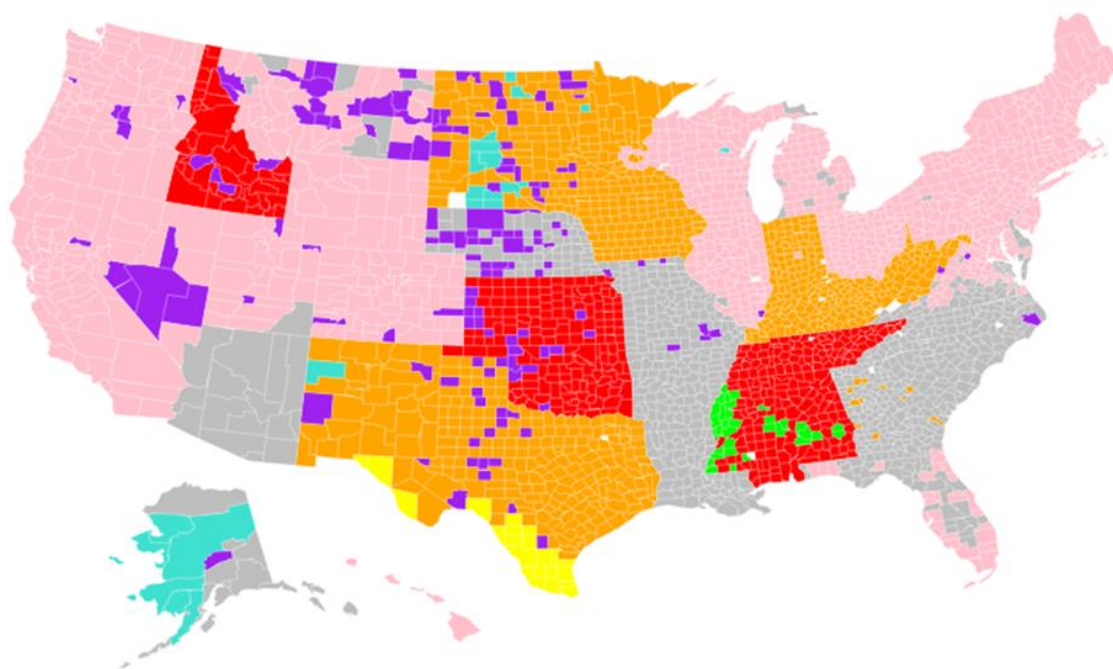










Figure 11. Map of eight clusters of US counties identified by hierarchical clustering analysis

Table 3. Characteristics and Locations of Counties in Each Cluster

Cluster id	Color on map	Characteristics	Location (states, regions)
1		Hispanic dominated, young, low prevalence of diabetes and obesity, physically active, moderate poverty rate, high fraction of uninsured people.	Texas at the Mexico-America border
2		Black dominated, high prevalence of diabetes and obesity, highly physically inactive, high poverty rate, low fraction of uninsured people.	Mississippi and Alabama.
3		White dominated, moderately old, moderately high prevalence of diabetes and obesity, physically inactive, high per capita expenditure at fast food restaurant	Idaho, Kansas, Oklahoma, Mississippi, Alabama, and Tennessee
4		American Indian and Alaska Native dominated, moderate prevalence of diabetes and high prevalence of obesity, moderate physical activity, high poverty rate and high fraction of uninsured people	Alaska, North Dakota and South Dakota
5		White dominated, old, low prevalence of diabetes and obesity, moderately physically active, low poverty rate	Widely distributed, with majority located at Western US
6		White dominated, low prevalence of diabetes and obesity, highly physically active	i. states with shoreline on Pacific ocean: California, Oregon, and Washington; ii. states near Rocky mountains: Montana,

			Utah, Wyoming, and Colorado; iii. states near Great Lakes: Wisconsin, Illinois, Michigan, Ohio, Pennsylvania and New York; iv. states with shoreline on the Atlantic ocean: Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Maryland, Virginia, and Florida
7		White dominated, moderately old, moderate prevalence of diabetes and obesity, moderately physically active	Texas, North Dakota, South Dakota and Minnesota
8		White dominated with Black as second dominant race, moderately high prevalence of diabetes and obesity, physically inactive	Arizona, Nebraska, Missouri, Arkansas, Louisiana, Georgia, North Carolina, South Carolina, South of Virginia and North of Florida.

The eight clusters greatly differ in race/ethnicity composition, age distribution, prevalence of diabetes or obesity, level of physical activity, or poverty rate. Several clusters have unique patterns that are worth noting. Their characteristics and locations are summarized in Table 3.

Firstly, cluster 1 (yellow) is quite unique. They locate in Texas at the Mexico-America border. It is group of counties where Hispanic is dominant and poverty rate is relatively high. However, counties in this cluster have the lowest average prevalence of diabetes and obesity (panel A and B in Figure 10). It is also worth noting population in these counties are highly physically active (low prevalence of physical inactivity as indicated by panel C in Figure 10). This implies that physical activity is an important determinant of population health in these counties.

Secondly, counties in cluster 2 (green) are mostly Black population dominant. They are in Mississippi and Alabama. These counties have the highest average prevalence of diabetes and obesity, despite the fraction of uninsured population is low in this cluster. These counties also have high poverty rate and are highly physically inactive. This implies that physical activity and poverty rate can be two important determinants of population health for this cluster.

Thirdly, counties in cluster 4 (turquoise) are mostly American Indian or Alaska Native dominated. They are in states including Alaska, New Mexico, North Dakota and South Dakota. They have high obesity prevalence and moderate physical activity. In addition, counties in this cluster have high poverty rate. It is reasonable to infer that poverty rate and poor health care can play important roles in the population health among these counties.

Fourthly, counties in cluster 6 (pink) have lowest average prevalence of diabetes and obesity and highest average level of physical activity, although they have the highest average per capita expenditures in fast food restaurants. Counties in this cluster are in regions where the land types are more suitable for recreations (details listed in Table 3).

Lastly, counties in cluster 8 (grey) are White dominant counties where Black population is the second largest race groups. They locate in Arizona, Nebraska, Missouri, Arkansas, Louisiana, Georgia, North Carolina, South Carolina, South of Virginia and North of Florida. Counties in these states have moderately high prevalence of diabetes and obesity, and physically inactive.

Hierarchical clustering analysis of the most diabetic and obese counties

The hierarchical clustering in last section shows that clusters with high prevalence and obesity can differ greatly in race, age and socioeconomic composition. To further explore this phenomena, I do hierarchical clustering analysis for counties whose prevalence of diabetes, prevalence of obesity are both in the top 20% level, which I refer as the most diabetic and obese counties.

The data preprocessing follows the same steps in section “hierarchical clustering analysis of the all counties”: columns with few missing values are selected, rows with null values are then dropped, and all numeric columns are normalized to have mean at zero and variance at 1. PCA was also done to prepare data for clustering analysis.

The method of linkage is chosen among “average”, “median”, ”complete”, and “centroid”. The method is chosen using the same criteria as last section: the one that gives tight clusters and clear distinctions among clusters. Three clusters are identified by hierarchical clustering (Figure 12).

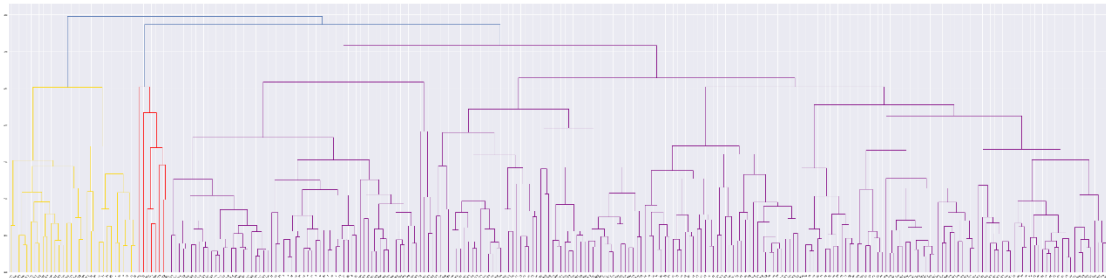


Figure 12. Dendrogram of hierarchical clustering of counties with top 20% prevalence of diabetes and top 20% prevalence of obesity.

Although these counties all have relatively high diabetes and obesity prevalence, they still differ in many ways.

Cluster 1 is Black dominant and most physically inactive (Figure 13 and Figure 14). Cluster 2 is American Indian and Alaska Native dominant, relatively young, and has the highest poverty rate and highest fraction of uninsured population (Figure 13 and Figure 14). Cluster

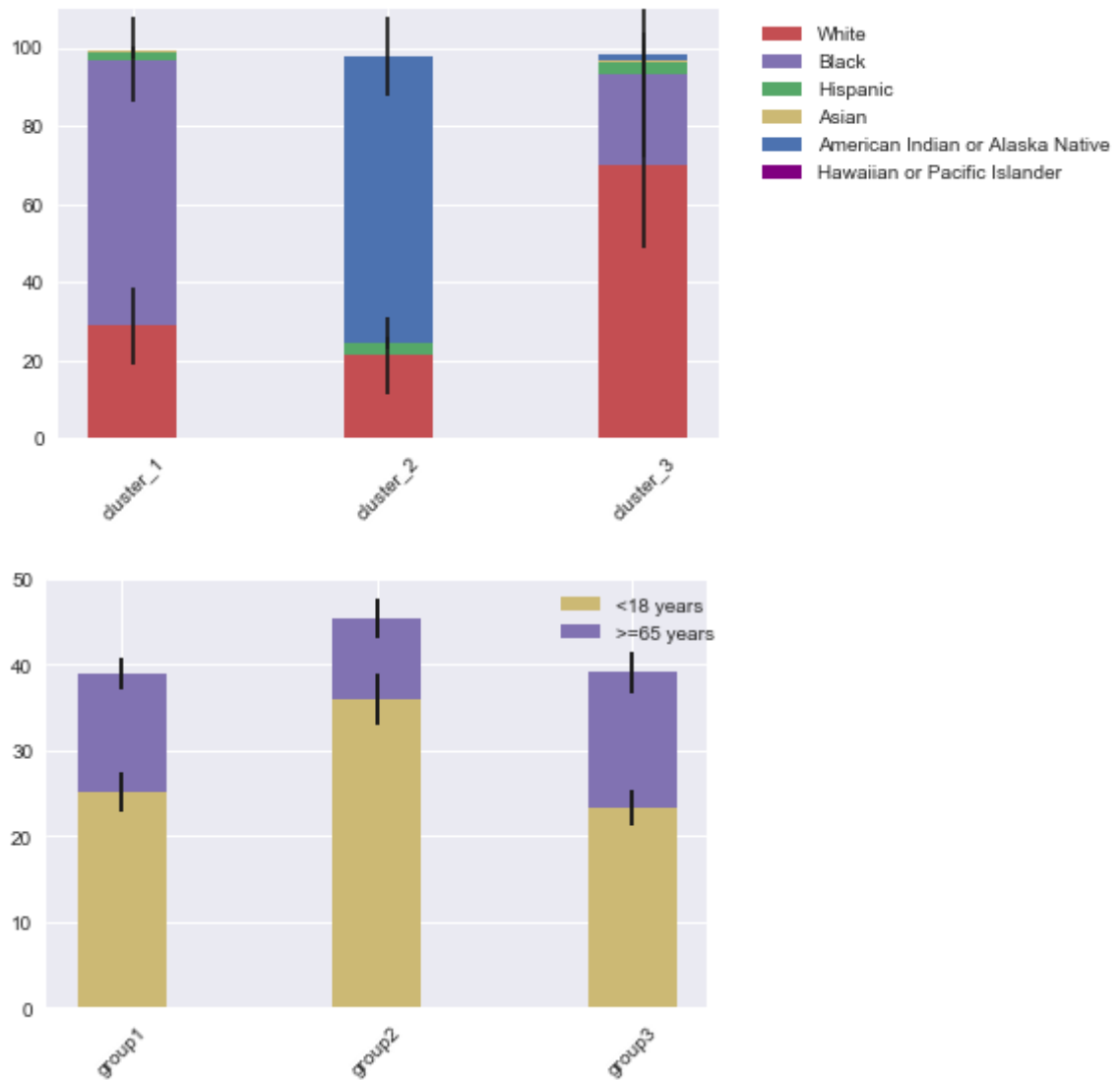


Figure 13. Race composition (upper panel) and age composition (lower panel) of the three clusters with top 20% prevalence of diabetes and top 20% prevalence of obesity.

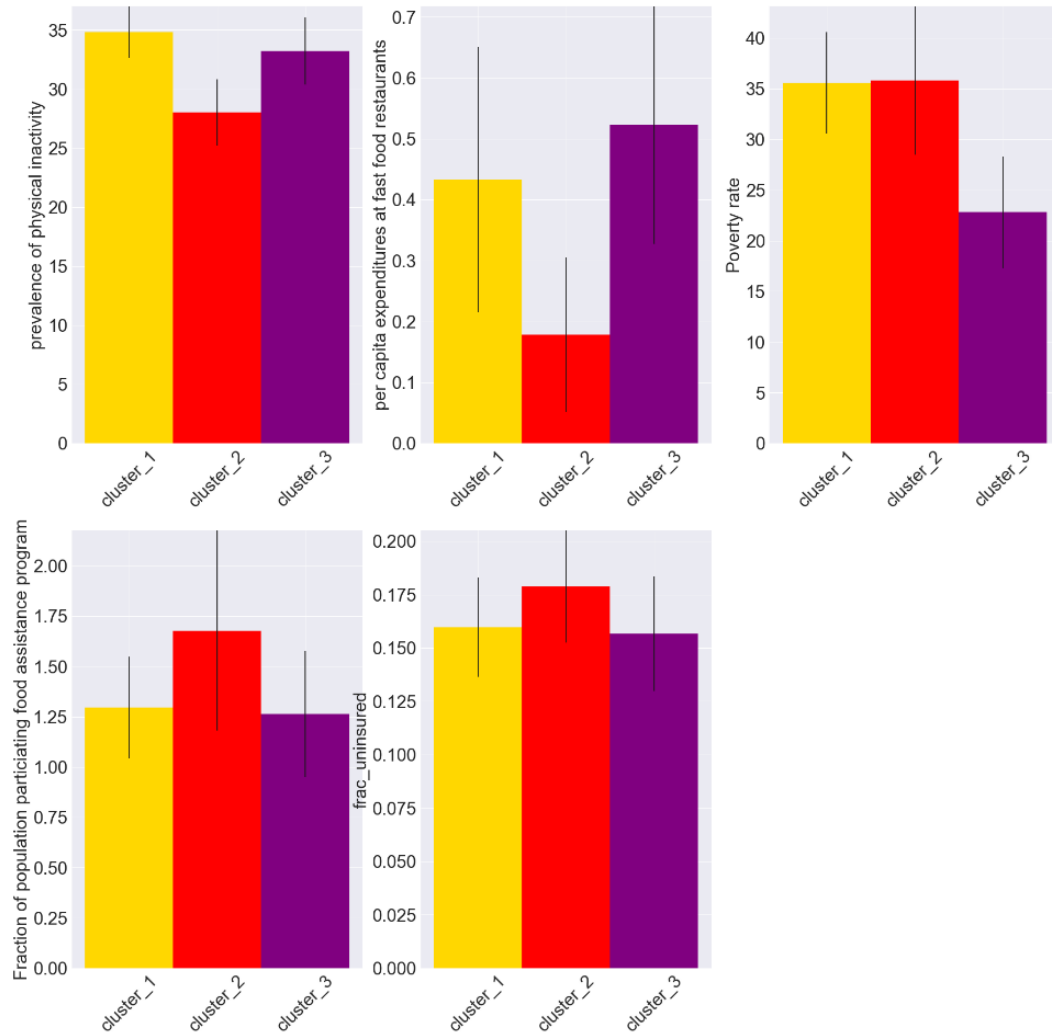


Figure 14. Features of three clusters of counties with top 20% prevalence of diabetes and top 20% prevalence of obesity.

A map of these counties clearly shows that they mostly locate at Southeast US (Figure 15).

Clusters of Most Diabetic and Obese Counties Identified by Hierarchical Clustering

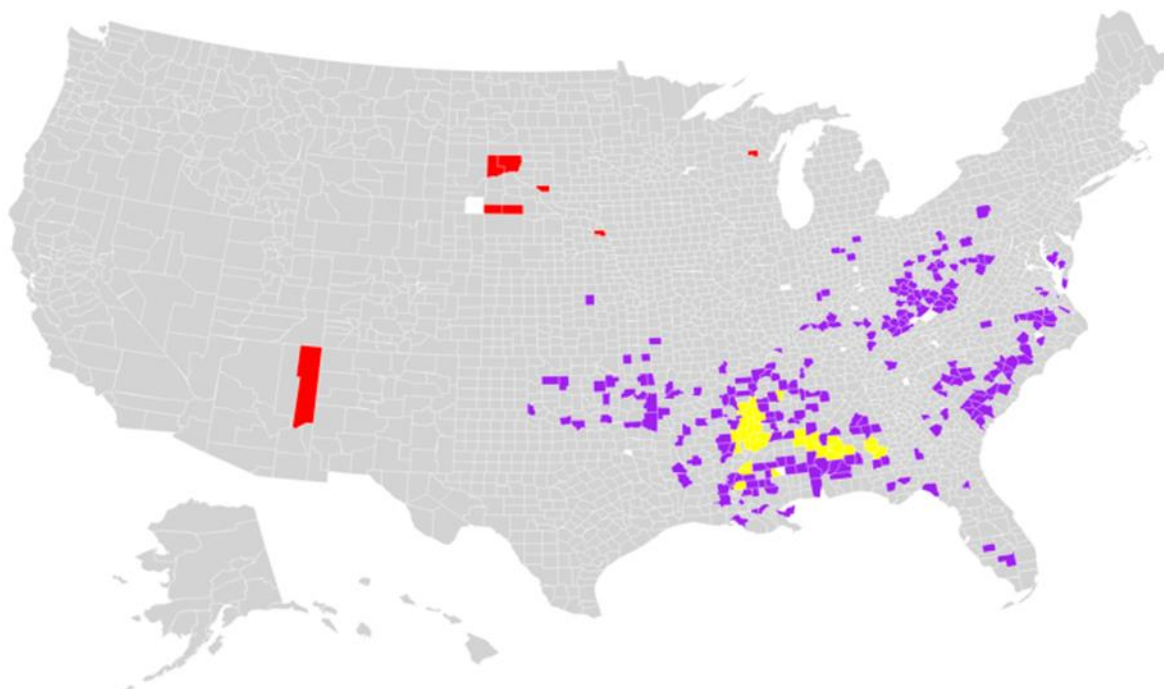





Figure 15. Map of most diabetic and obese counties (prevalence of diabetes and obesity are both in the top 20% of all US counties).

Table 4. Characteristics and Locations of Clusters of most Diabetic and Obese Counties

Cluster id	Color on map	Characteristics	Locations
1		Black dominated, most physically inactive, high poverty rate. Found in states including Mississippi, Alabama, and Georgia.	Mississippi, Alabama and Georgia
2		American Indian and Alaska Native dominated, moderate physical activity, high poverty rate and high fraction of uninsured people.	Alaska, North Dakota and South Dakota
3		White dominant with Black as second dominant race, physically inactive	Southeast US

The three clusters largely match the cluster 2, 4 and 8 identified in last section.

Regression Analysis

Regression analysis is done for two purposes. Firstly, regression coefficients reflect the impact of model features on the outcome variables. For example, how much prevalence of diabetes and obesity change if the fraction of physically inactive population drops by one percent. Secondly, it predicts the likely prevalence of diabetes and obesity of a specific county given the county's features.

Before regression analysis, data needs to be preprocessed in two steps: missing values imputation and data normalization. Then data is split into testing set and training set, and cross validation set to tune the model parameters. Finally, the coefficients of tuned model

Data preprocessing

I use the sklearn package to impute the missing values with mean imputation. Data are normalized so that each column has mean as zero and variance as one.

Predictors are singled out as numpy array while outcome variable (prevalence of diabetes or prevalence of obesity) is also singled out as numpy array.

Then 75% of the whole preprocessed data is used for training while the remaining 25% is retained as test data (Figure 16). Training data are used to tune model parameters and estimate coefficients. The test data will be then used to assess the accuracy of prediction of the tuned model with the estimated coefficients.

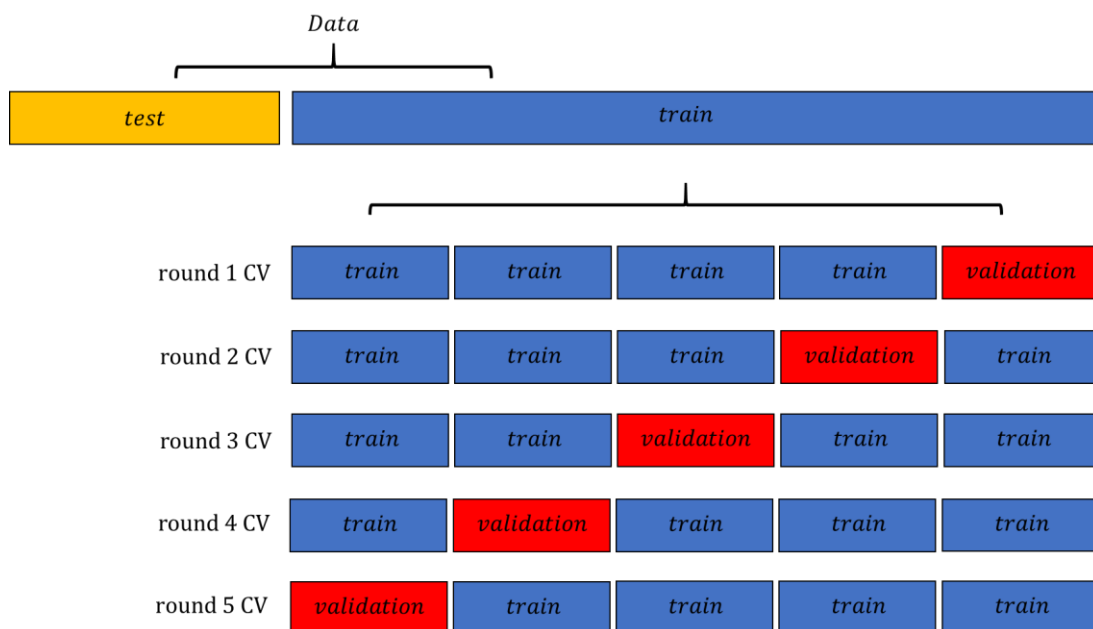


Figure 16. Schematic of data split for training, cross validation and testing.

Cross Validation

Cross validation (CV) helps improve the generalizability of the model. Its purpose is to tune the model parameters, i.e. hyperparameters, and coefficients, to find the setting that best balance bias and variance of model prediction.

Bias is usually result of underfitting, which means features do not provide enough information to explain the outcome variable. It is suggested by high training errors and high validation errors. High variance is result of overfitting, which means that model tries to use many features to fit the training data but fail to generalize to new data (validation data in this case). It is suggested by low training errors and high validation errors. (citations) The training error is calculated as the sum of squared difference between the observed values and predicted values of dependent variable in the training set. Validation error is calculated in the same way for validation data.

Validation is done in following steps: split the data (excluding test data) into training data and validation data; given a specific set of hyperparameters, regression model is fit to training data to estimate coefficients; Then fit the model with the estimated coefficients to the validation data. This is done under different hyperparameter settings until model with low training errors and low validation errors is found.

In this section, I do five-fold cross validation to tune the model parameters and estimate predictors' coefficients. Specifically, 75% data (data excluding the testing set) is randomly split into five equally sized groups, i.e. 15% data for each group (Figure 19). One group is used as validation data and the other four groups as train data. CV is done for five rounds, where each group is used as validation data once (Figure 19). CV error is calculated as the average of CV errors in each round.

To control for overfitting, I use linear regression with regularization. Regularization include the values of estimated coefficients in the cost function. Increasing coefficients would increase cost function and reduce the model fitting. It helps reduce variance of prediction: while fitting process try to minimize cost function, coefficients of less impactful features tend to be tuned to zero. Therefore, it also helps select the subset of predictors and reduce the variance of prediction.

There are two types of regularization method: l1 regularization and l2 regularization. L1 regularization uses the sum of absolute values of predictors' coefficients as the regularization term while l2 regularization uses sum of squared predictors' coefficients as the regularization term. To increase the flexibility of model, I use elastic net regression model, which includes both regularization terms. Model is tuned by adjusting the coefficients of both regularization terms.

Therefore, there are two hyperparameters, to tune model during the cross validation: the sum of coefficients of two regularization terms, α , and the weight of l1 regularization term, l1_ratio.

For cross validation, I use the ElasticNetCV function from sklearn package in python. It tuned α and l1_ratio simultaneously across a wide range of values of α and l1_ratio from zero to one.

Regression Analysis Results

When prevalence of diabetes is the outcome variable, the tuned model parameters are that α equals 3.795 e-05, and l1_ratio is 1.

In addition, R^2 is 77% when prevalence of diabetes is the outcome variable. This means that 77% of the variance of the prevalence of diabetes can be explained by the selected features. This is a relatively high R^2 , which also suggests that tuned model works relatively well for the test data.

A plot of the predictors' coefficients shows that the top three impactful factors that increases prevalence of diabetes of a county (with top three highest coefficients) are: fraction of WIC participants, prevalence of physical inactivity, fraction of population over 65 years old (Figure 17).

The top three impactful factors that reduces prevalence of diabetes of a county (with top three lowest coefficients) are: fraction of uninsured population under 65 years old, recreation and fitness facilities, and number of available grocery stores (Figure 17).

Most of them are understandable except the observation that fraction of uninsured population under 65 years old reduces prevalence of diabetes. This could be result that some areas where population are highly physically active have high fraction of uninsured population under 65 years old, such as counties in California.

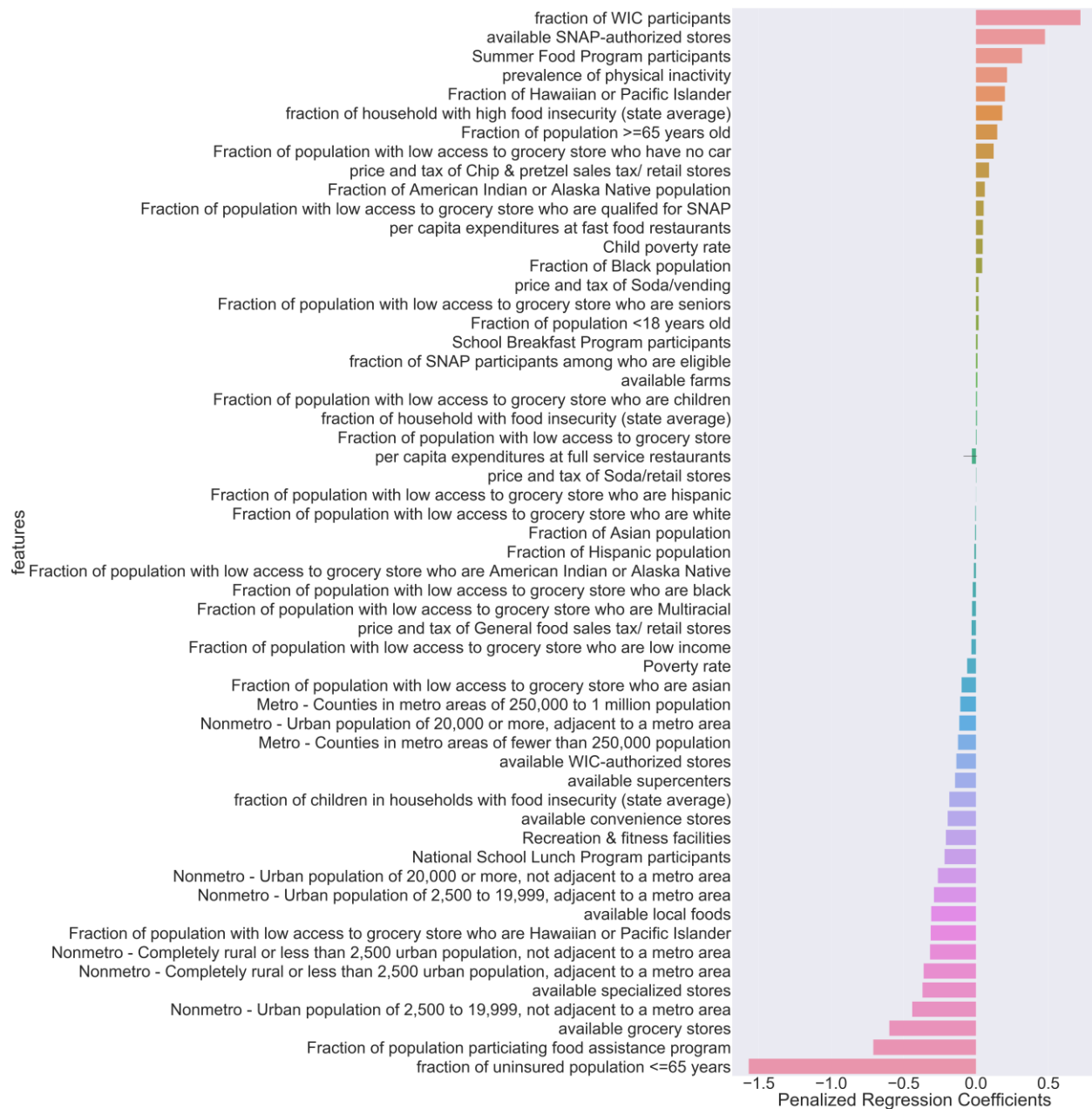


Figure 17. Coefficients of predictors of prevalence of diabetes

When prevalence of obesity is the outcome variable, R^2 is 69% when prevalence of obesity is the outcome variable.

The top three impactful factors that increases prevalence of obesity of a county (with top three highest coefficients) are: number of available supercenters, Hawaii and Pacific Islander, and urban types, prevalence of physical inactivity, fraction of population over 65 years old (Figure 21).

The top three impactful factors that reduces prevalence of obesity of a county (with top three lowest coefficients) are: fraction of uninsured population under 65 years old, recreation and fitness facilities, and number of available grocery stores (Figure 18).

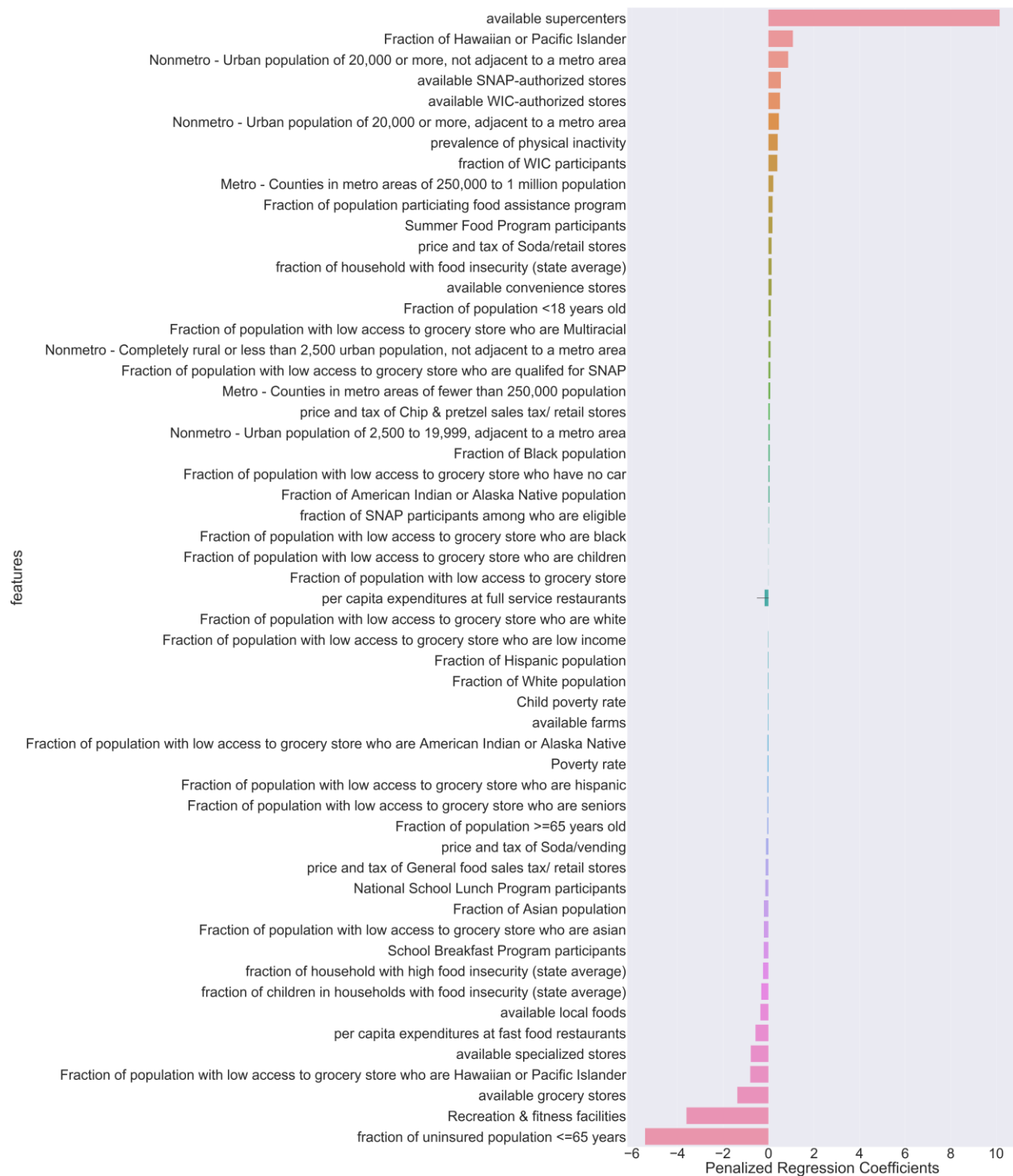


Figure 18. Coefficients of predictors of prevalence of obesity

Comparison in coefficients between diabetes and obesity

A comparison between the coefficients of several selected predictors for diabetes and obesity shows some interesting pattern (Figure 19).

Firstly, physical activity reduces the risk of both diabetes and obesity. Counties with Hawaiian or Pacific Islander tends to have greater risk of diabetes and obesity. This is something that is not revealed by the clustering analysis.

Prevalence of diabetes increases when there are more people older than 65 years old, which is not the case for obesity. By contrast, prevalence of obesity increases when there are more people younger than 18 years old.

Although the availability of recreation and fitness facilities reduces prevalence of diabetes and obesity, it affects obesity considerably more than it does for diabetes.

Lastly, counties with more supercenters tend to have higher prevalence of obesity but slightly lower prevalence of diabetes than others.

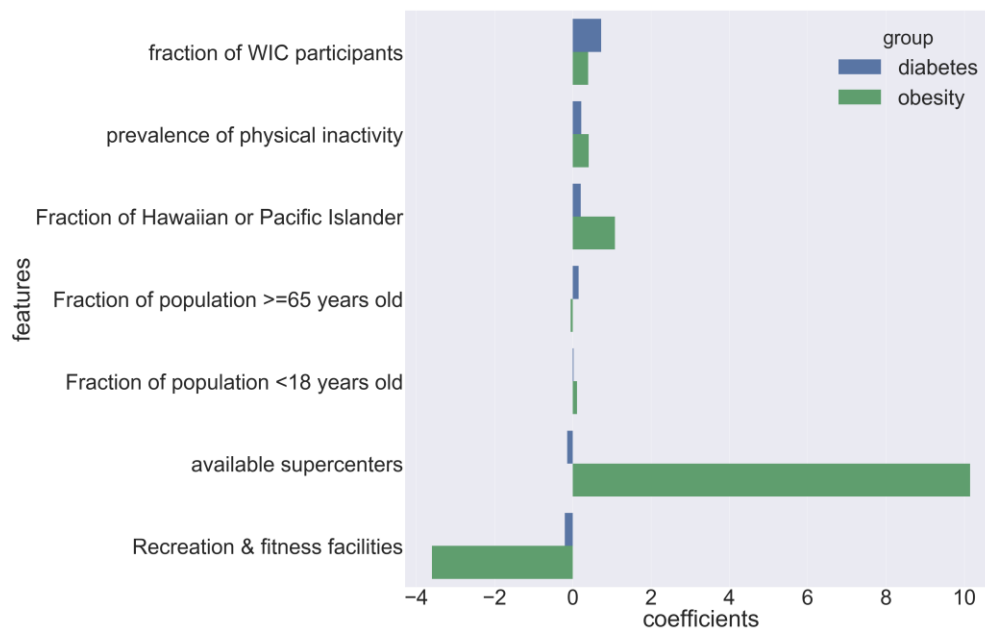


Figure 19. Comparison of coefficients of some predictors when prevalence of diabetes is the outcome variable and when prevalence of obesity is the outcome variable.

Predicting Prevalence of Diabetes and Obesity

Using the tuned regression model, I was able to predict the prevalence of diabetes if one of the variables changes. For example, if prevalence of physical inactivity decreases by 5%, one can see an obvious drop in diabetes prevalence in Southeast US (Figure 20). Also, increase in available grocery stores by 5% can also considerably reduce the diabetes prevalence in these areas (Figure 21).

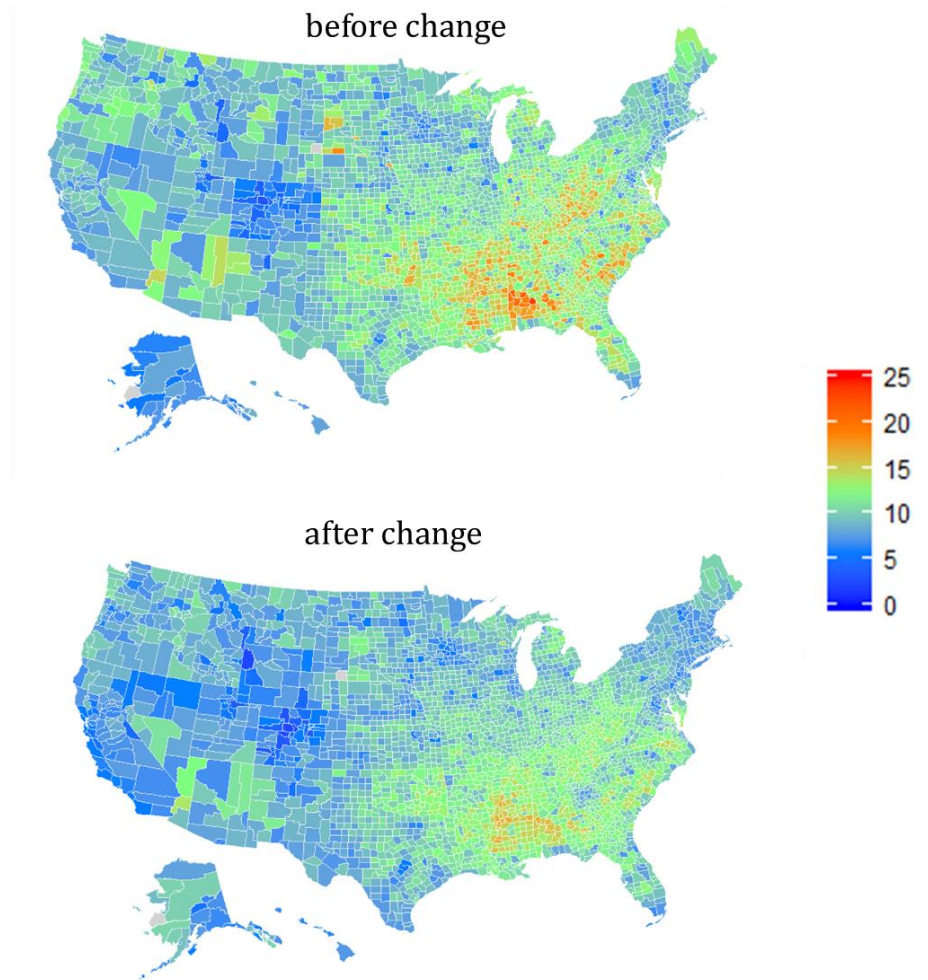


Figure 20. Upper panel: estimated prevalence of diabetes in 2013; lower panel: Predicted prevalence of diabetes if prevalence of physical inactivity decreases by 5%.

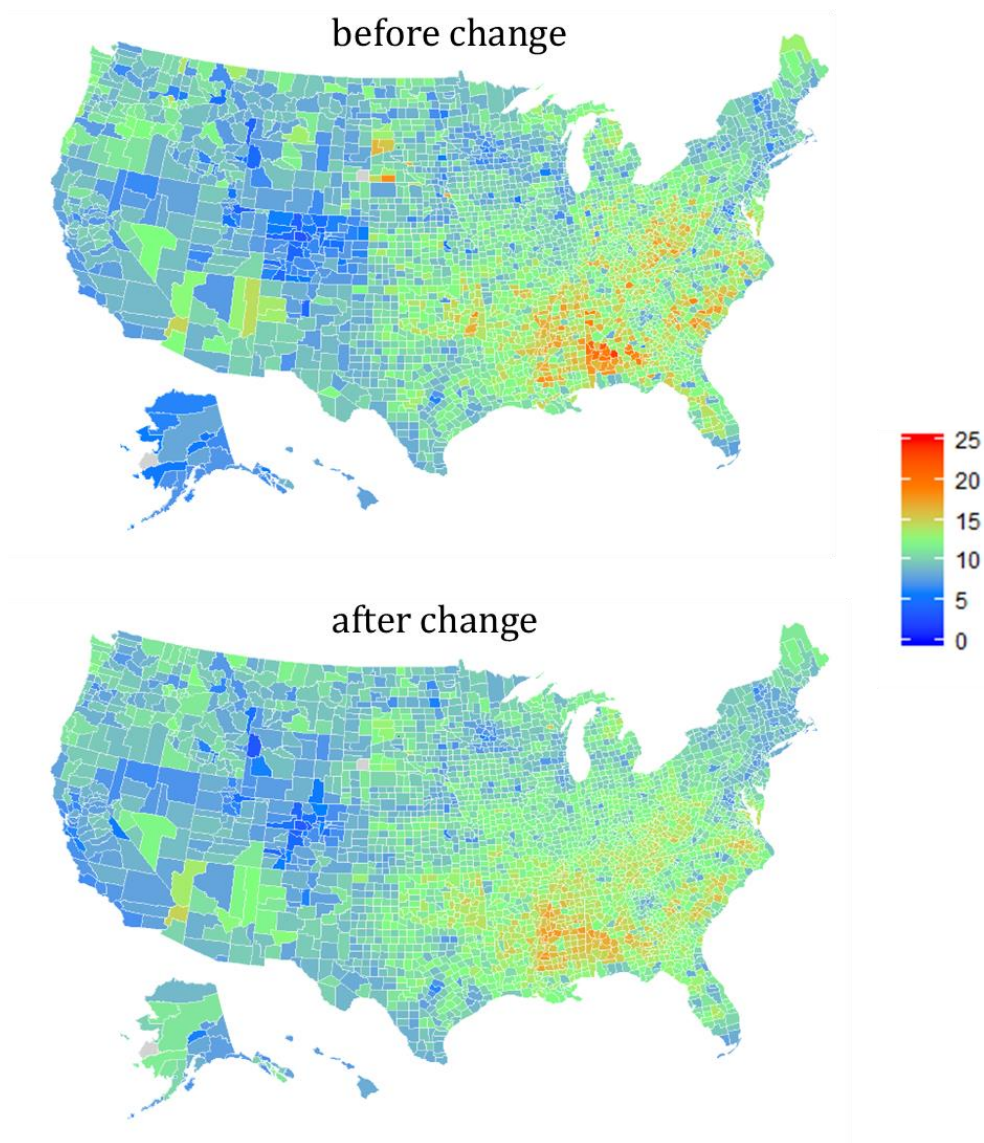


Figure 21. Upper panel: estimated prevalence of diabetes in 2013; lower panel: Predicted prevalence of diabetes if availability of grocery stores increases by 5%.

If fraction of recreation facilities increases by 5%, one can see an obvious drop in obesity prevalence in Southeast US and Midwest US (Figure 22). Also, increase in available local foods by 5% can also considerably reduce the obesity prevalence in these areas (Figure 23).

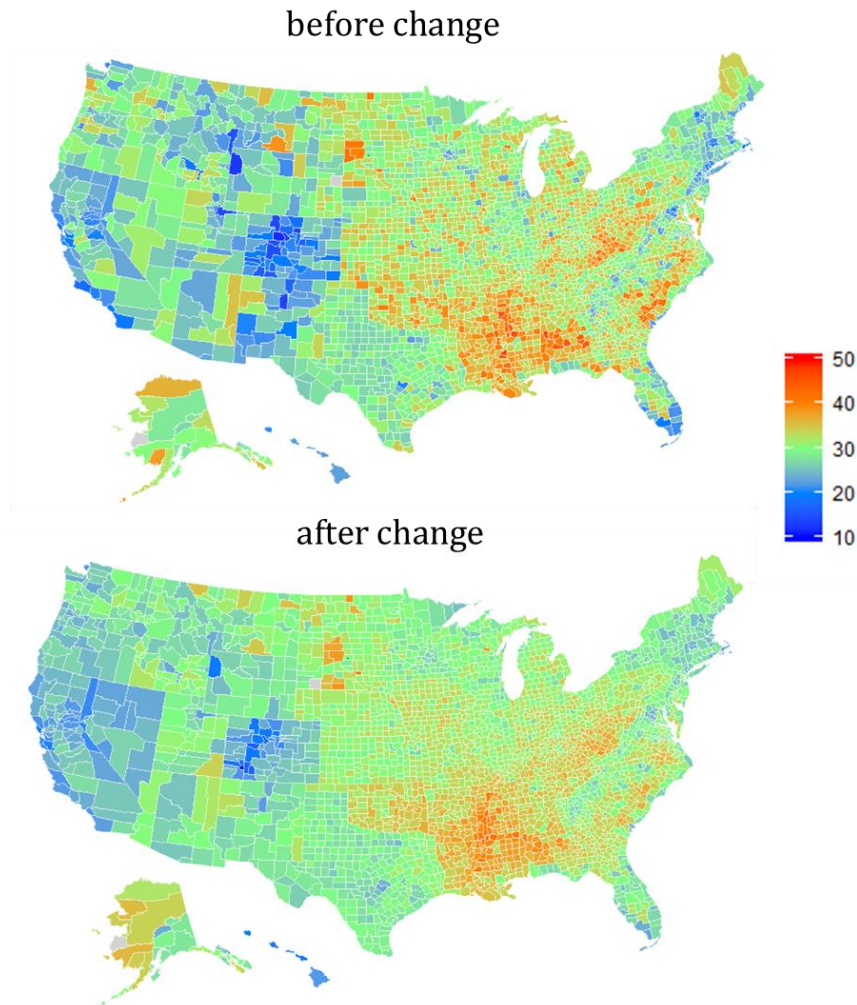


Figure 22. Upper panel: estimated prevalence of obesity in 2013; lower panel: Predicted prevalence of obesity if available recreation facilities increases by 5%.

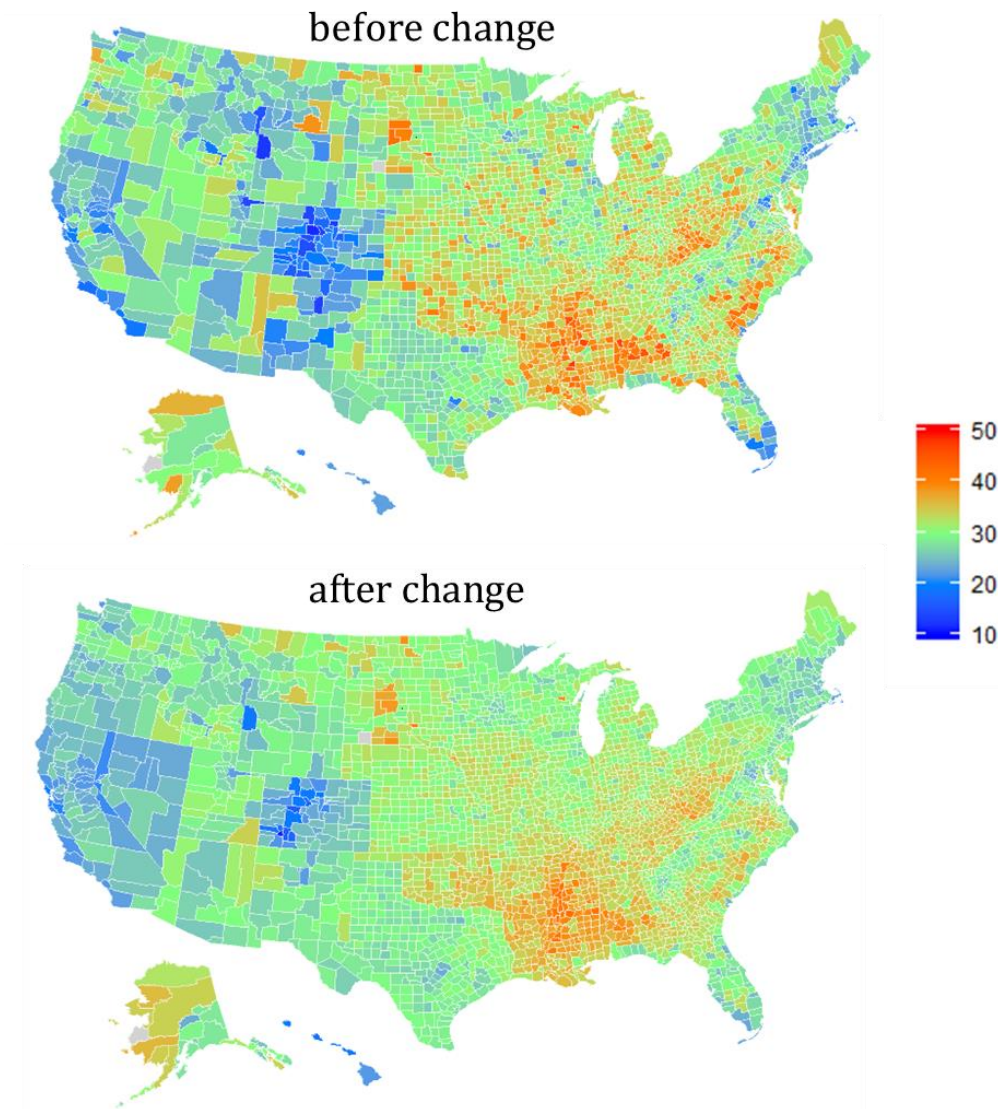


Figure 23. Upper panel: estimated prevalence of obesity in 2013; lower panel: Predicted prevalence of obesity if available local foods increases by 5%.

Summary

In this project I use exploratory data analysis, hierarchical clustering analysis and regression analysis with regularization to explore the pattern of physical activity, food environment, coverage of medical insurance, demographics, socioeconomic status and prevalence of diabetes and obesity among US counties.

Using exploratory analysis and regression analysis, I found that high prevalence of diabetes and obesity share similar predictors such as physical inactivity, but also have quite different predictors, such as age profile the population.

Using clustering analysis, I identified three major groups where prevalence of diabetes and obesity are both high. The three clusters greatly differ in age composition, race composition, poverty rate and geographical locations.

The major findings are as follows. Firstly, Counties have high prevalence of diabetes and high prevalence of obesity are mostly in Southeast US. Improving physical activity and food choice likely reduce prevalence of diabetes and obesity in these areas. Secondly, improving recreation facilities and food choice can considerably reduce prevalence of obesity, especially counties in Southeast US, Midwest US and Northern Alaska. Thirdly, High level of physical activity is likely the reason of low prevalence of diabetes and obesity in counties along shoreline of ocean and lakes, and along Rocky Mountains. Food environment likely plays minor role in affecting population health in these areas.

There are some interesting observations worth noting. Firstly, the clustering analysis result suggest that physical activity and socioeconomic characteristics are the most important variables to distinguish the US counties with different prevalence of diabetes and obesity. Although results of regression analysis confirm the effects of these variables, they are not the most impactful predictors. Rather, variables such as available supercenters, participants in WIC programs have large coefficients.

Secondly, regression analysis suggests that counties with high fraction of Pacific islander and Hawaiian tends to have high diabetes and obesity prevalence. This is not found by clustering analysis either. This could be because that there are too few counties dominated by Pacific islander and Hawaiian for them to be identified as a single cluster. Both suggest that regression analysis reveals the important phenomena that may not be identified by clustering analysis.

Thirdly, as suggested by regression analysis, fraction of uninsured population is surprisingly a protective factor for both diabetes and obesity. This is counterintuitive. One possible reason is that areas with low fraction of uninsured population happen to be those with high prevalence of diabetes and obesity. This is likely the result of Obama care, where health insurance coverage particularly increases in Southeast US.⁴

In the future, more effort can be put into collecting more information and data to improve the understanding of population health and food environment of US counties. Also, more data can be collected to further explore the similarity and difference between obese counties and diabetic counties. Time series analysis can also be done to explore the trend of change in

diabetes and obesity prevalence and how they correlate with change in physical activity and food environment.

Reference

1. Salois MJ. Obesity and diabetes, the built environment, and the ‘local’ food economy in the United States, 2007. *Econ Hum Biol.* 2012;10(1):35-42. doi:10.1016/j.ehb.2011.04.001
2. Seligman HK, Bindman AB, Vittinghoff E, Kanaya AM, Kushel MB. Food Insecurity is Associated with Diabetes Mellitus: Results from the National Health Examination and Nutrition Examination Survey (NHANES) 1999–2002. *J Gen Intern Med.* 2007;22(7):1018-1023. doi:10.1007/s11606-007-0192-6
3. Ansari RM. Effect of Physical Activity and Obesity on Type 2 Diabetes in a Middle-Aged Population. *J Environ Public Health.* 2009;2009. doi:10.1155/2009/195285
4. Sanger-Katz M. The Impact of Obamacare, in Four Maps. *The New York Times*.
<https://www.nytimes.com/interactive/2016/10/31/upshot/up-uninsured-2016.html>,
<https://www.nytimes.com/interactive/2016/10/31/upshot/up-uninsured-2016.html>. Published October 31, 2016. Accessed February 28, 2018.