Group 2 Update 2: United States Obesity Risk Factor Data Analysis

Brian Desnoyers BDESNOY@CCS.NEU.EDU
Alankrit Joshi ALANKRIT93@CCS.NEU.EDU
Rahul Kondakrindi RAHULKONDAKRINDI@CCS.NEU.EDU
Prasanna Vikash Peddinti VIKASH4281@CCS.NEU.EDU
Junyi Wang WANG4615@CCS.NEU.EDU

1. Introduction and Dataset

Obesity is a major challenge facing the healthcare system in the United States. Obesity rates have risen significantly over the last thirty years and if this trend continues, the United States healthcare system is projected to pay $150 billion annually (Hurt et al., 2010). In addition to genetic factors, this increase in obesity rates, inactivity, and malnutrition has been linked to the environment. Environmental changes; such as car transportation, inactive jobs, carry out food, food advertisements, and food portions; can be tied to specific regions and have a significant impact on health (National Heart, Lung, and Blood Institute, 2013; National Institute of Diabetes and Digestive and Kidney Diseases, 2012). The main goal of this project is to explore the distribution of these risk factors across the United States and visualize how these groupings correspond to obesity and health.

For this project, we will analyze the Nutrition, Physical Activity, and Obesity dataset provided by the Centers for Disease Control and Prevention (CDC). This dataset provides information on the percentage of the population suffering from adult obesity, as well as associated behaviors, such as poor nutrition and physical inactivity, for the nation, states, and selected sub-state districts. These data also include potential risk factor features, such as age, education, sex, and income (Centers for Disease Control and Prevention). In addition, this dataset provides similar population information for child obesity, infant breastfeeding, active transportation, and community policy supports.

2. Preprocessing

The main purpose of our team's work for this update was driven around exploratory analysis, which started with data pre-processing. Since the dataset from the CSV was available in CSV format, it was easy to read without file format conversion. After reading in the dataset, we focused on understanding each column to allow us to extract the columns useful for our analysis.

Unfortunately the raw dataset relies on a relatively combersome "question-based" format corresponding to different types of data values, such as percentage of adults who are overweight. To deal with this, we developed functions to filter the dataset based on specific question types, which we utilized for additional exploratory analysis.

In addition, because the overweight percentages did not include obese adults, we had to update all of the overweight values to also include them.

3. Exploratory Analysis and Results

Our initial exploratory analysis involved generating state-wise interactive plots for each condition in the dataset. These plots were shown on an interactive map, which allowed our team to begin to explore and understand the dataset. The conditions plotted were the percentage of adults who are obese, the percentage of adults who are overweight, the percentage of adults who engage in no physical activity, the percentage of adults who eat less than one fruit per day, and the percentage of adults who eat less than one vegetable per day. Screenshots of these plots are shown in figs. 1 to 5 below.

From the exploratory analysis of making these plots, our team has identified several regional patterns, which will be important to understand when working on our clustering analysis, which was specified in the project proposal. We also wanted to start visualizing relationships between specific features to understand the dataset in finer detail. One way we did this was to look at the relationship between some of the other conditions and obesity on a state-by-state basis. These plots are shown in figs. 6 and 7.

Figure 1.

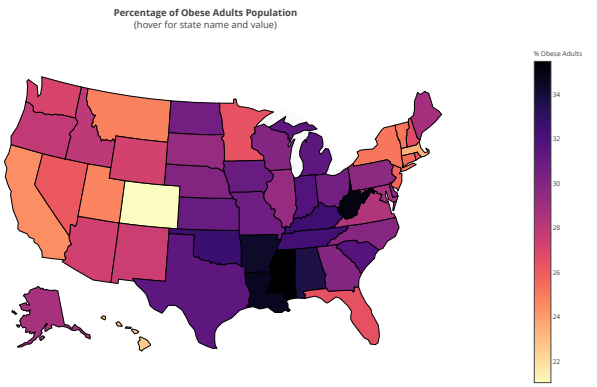


Figure 2.

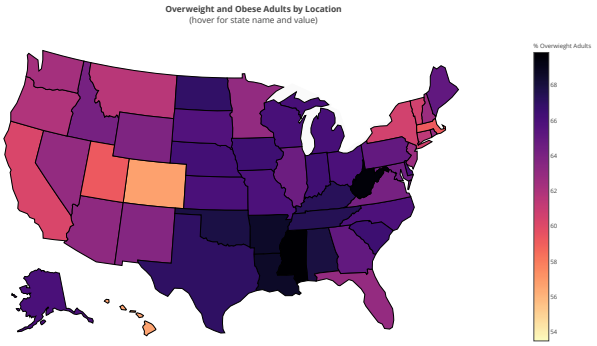


Figure 3.

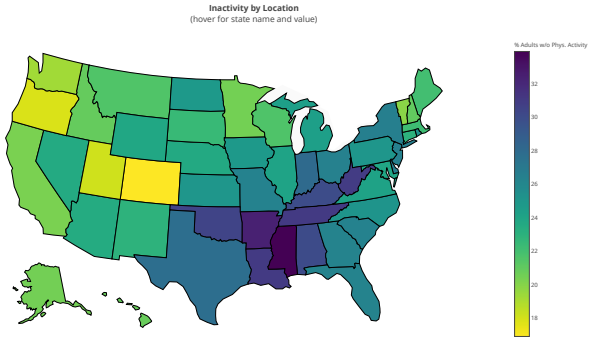
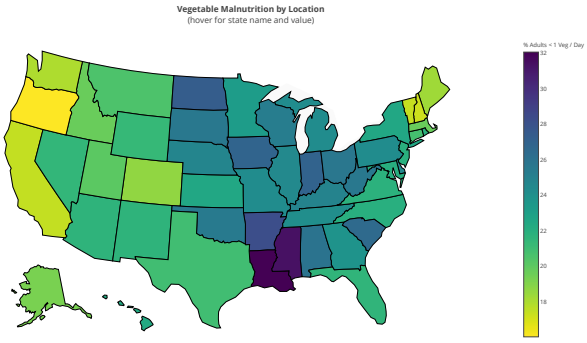


Figure 4.



4. Data Mining Analysis

In our data analysis, we were interested in clustering the CDC health behavior outcomes data we explored previously to investigate regional health outcomes and trends. We also collected data on socioeconomic risk factors (education status, race, and income) from Wikipedia (Wikipedia, a;b;c), which was clustered separately and compared to the health behavior outcomes clusterings. We were interested in looking at similarities between these two sets of clusterings.

To do this, we performed k-means, DBSCAN, and hierarchical clustering across these two datasets (CDC health behavior outcomes and the socioeconomic risk factors we put together from several articles on Wikipedia).

4.1. Socioeconomic Risk Factors

Before clustering, we first normalized the risk factor data to be scaled from 0 to 1. The specific features utilized were: percentage of residents who are high school graduates, percentage of residents with a Bachelor's degree, percentage of residents with an advanced degree, percentage of residents who are white, percentage of residents who are black or African American, percentage of residents who are American Indian or Alaska Native, percentage of residents who are Asian, percentage of residents who are Native Hawaiian or Other Pacific Islander, percentage of residents who are some other race, percentage of residents who are two or more races , per capita income in dollars, median household income in dollars, and median family income in dollars.

4.1.1. k-MEANS

To select a value of k, we performed k-means clustering for several values of k and plotted the objectives as shown in fig. 8. From this plot, we identified 5 as a value to use for clustering. We then performed k-means clustering with $k = 5$ and plotted the results on a two-dimensional scatter plot (fig. 9) and state map (fig. 10). The data was reduced to two dimensions for plotting by applying principal component analysis (PCA) and selecting the first two principal components.

Figure 5.

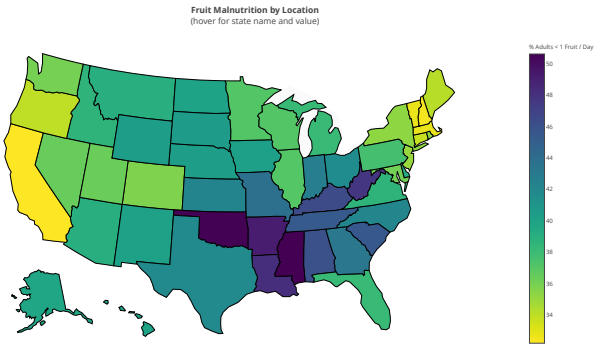


Figure 6. Correlation between Mean (over Years) Vigerous Aerobic Activity and Obesity

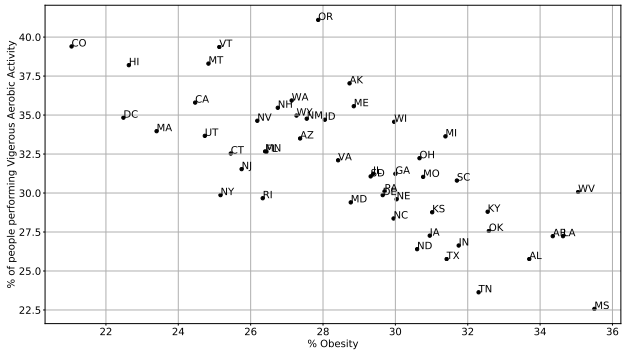


Figure 7. Correlation between Mean (over Years) Fruit Malnutrition and Obesity

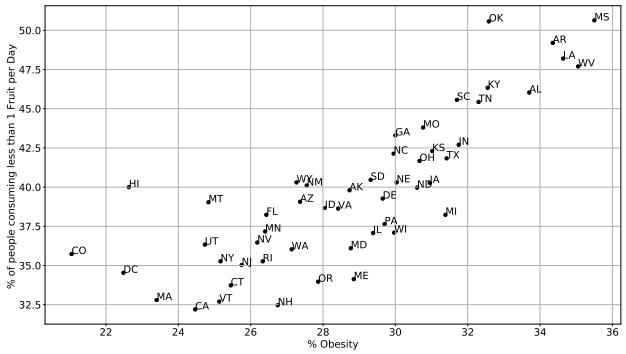
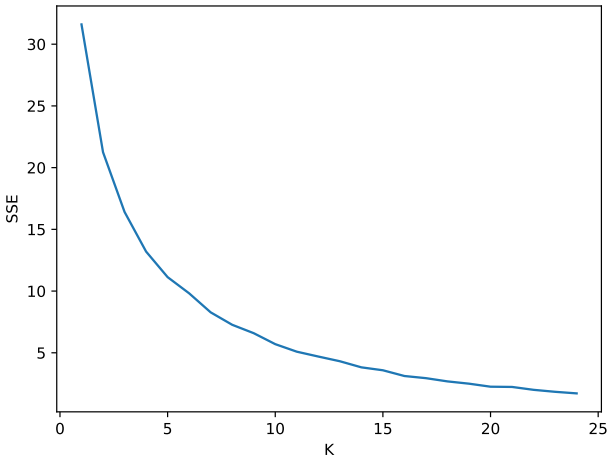


Figure 8. Objective function results as a function of k used to select a value of k for k-means clustering of socioeconomic risk factor data.



4.1.2. DBSCAN

We then performed DBSCAN with $\epsilon = 0.6$ and $minPts = 6$ because they achieved less than 5% noise. This resulted in a single cluster with four noise points as seen in figs. 11 and 12. Like performed previously, the data was reduced to two dimensions for plotting by applying principal component analysis (PCA) and selecting the first two principal components.

4.1.3. Hierarchical Clustering

We then performed agglomerative hierarchical clustering using a Ward distance metric. The Ward metric was selected after comparing results from the single, complete, and average linkage metrics. This clustering is shown in fig. 13.

4.2. CDC Health Behavior and Outcomes

Before clustering, we first normalized the CDC health behavior and outcomes data to be scaled from 0 to 1. The specific features utilized were: percentage of residents who are obese, percentage of residents who are overweight, percentage of residents who do not engage in physical activity, percentage of residents who engage in vigerous aerobic activity, percentage of residents who consume less than one vegetable per day, and percentage of residents who consume less than one fruit per day.

4.2.1. k-means

To select a value of k, we performed k-means clustering for several values of k and plotted the objectives as shown in fig. 14. From this plot, we identified 3 as a value to use for clustering. We then performed k-means clustering with $k = 3$ and plotted the results on a two-dimensional scatter plot (15) and state map (16). The data was reduced to two dimensions for plotting by applying principal component analysis (PCA) and selecting the first two principal components.

Figure 9. Two dimensional plot of clusters from k-means clustering of socioeconomic risk factor data.

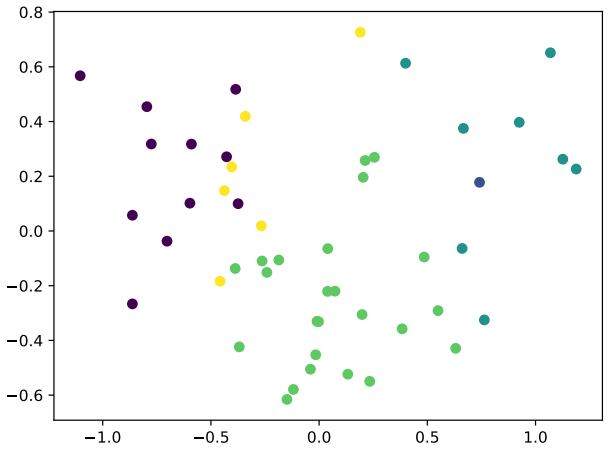
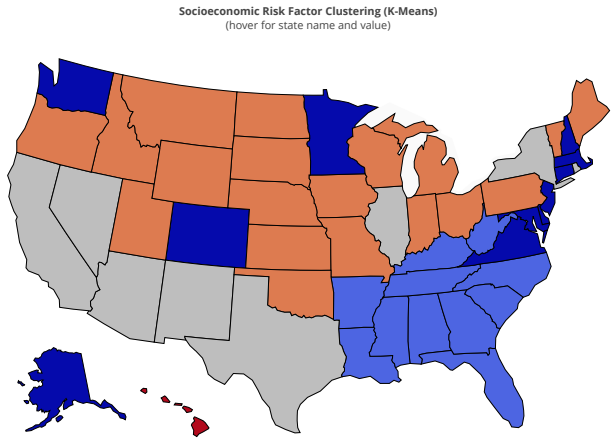


Figure 10. Geographic plot of clusters from k-means clustering of socioeconomic risk factor data.



4.2.2. DBSCAN

We then performed DBSCAN with $\epsilon = 0.2$ and $minPts = 3$ because they achieved less than 5% noise. This resulted in a single cluster with four noise points as seen in figs. 17 and 18. Like performed previously, the data was reduced to two dimensions for plotting by applying principal component analysis (PCA) and selecting the first two principal components.

4.2.3. Hierarchical Clustering

We then performed agglomerative hierarchical clustering using a Ward distance metric. The Ward metric was selected after comparing results from the single, complete, and average linkage metrics. This clustering is shown in

Figure 11. Two dimensional plot of clusters from DBSCAN clustering of socioeconomic risk factor data.

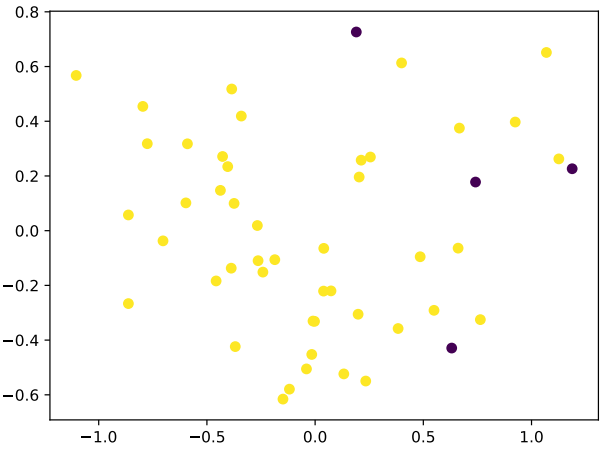
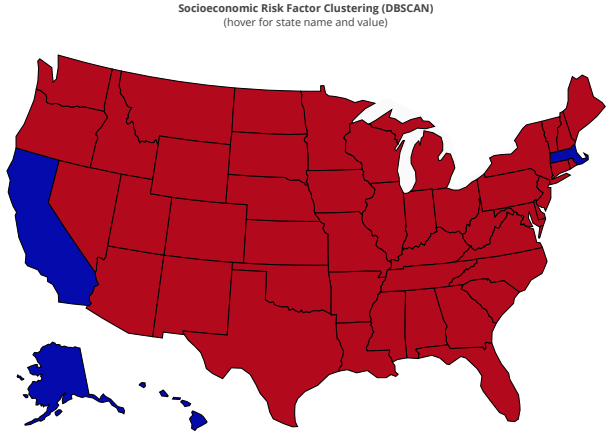


Figure 12. Geographic plot of clusters from DBSCAN clustering of socioeconomic risk factor data.



19.

5. Discussion

Based on the first two principal components from PCA performed on the CDC health behavior outcomes data (Table 1), combined with the fact that there were only a total of 6 features, redundant features were not selected to eliminate. The same is true for the first two principal components from the socioeconomic risk factors data (Table 2). However, for the socioeconomic risk factors data, features that poorly corresponded to the first principal component, such as "% White" were incorporated more into the second principal component. Once again, the 14 features did not appear to justify more rigorous feature selection.

Figure 13.

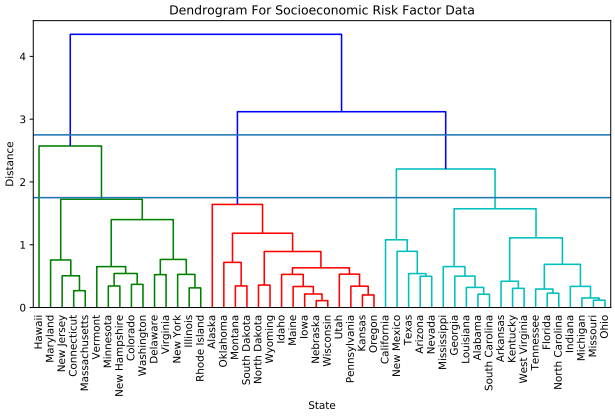


Figure 14. Objective function results as a function of k used to select a value of k for k-means clustering of health behavior and outcome data.

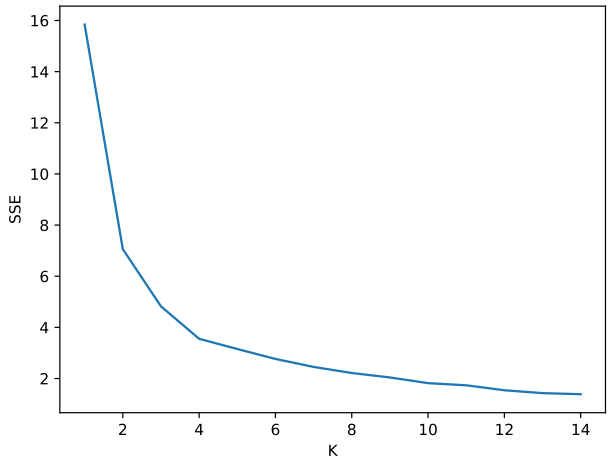


Figure 15. Two dimensional plot of clusters from k-means clustering of health behavior and outcome data.

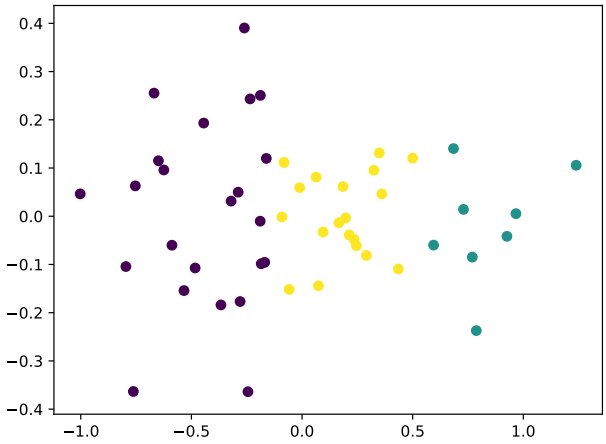
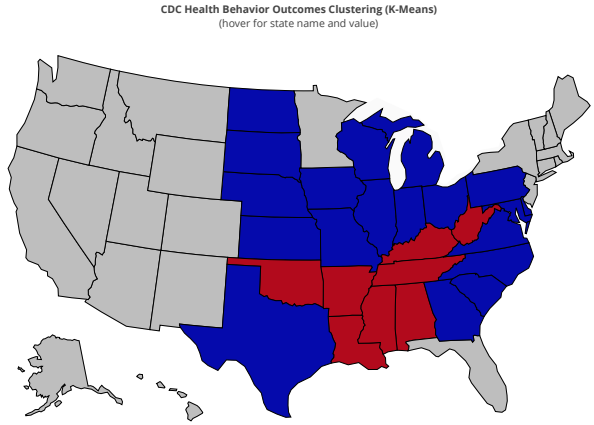


Figure 16. Geographic plot of clusters from k-means clustering of health behavior and outcome data.



When visualized in two dimensions, cluster density appears to be relatively consistent which may explain the relatively poor performance of DBSCAN in finding separate clusters without classifying a large number of data points as noise. Cluster shape and size do not seem to vary too much to suggest problems with k-means clustering. However, in understanding these connections, the hierarchical clustering proved most useful as a dendrogram allowed for a clear visualization of the number of clusters and their respective distances. Looking at these clusters obtained through k-means and hierarchical clustering, these clusterings appear to follow geographical boundaries. For example, in the clustering of socioeconomic risk factor data: southern and midwest states (such as Louisiana, Mississippi, Alabama, Arkansas, Oklahoma) and coastal states (such as Connecticut, California, Massachusetts, Alaska, Hawaii) can be seen. When compared via the dendrograms obtained through hierarchical clustering, states that are most similar based on socioeconomic risk factors are often also similar in terms of health outcomes (for example, New York and Rhode Island).

The cophenetic correlation distances for the socioeconomic risk factor and CDC health behavior outcomes clustering were 0.549033284833 and 0.597008648578, respectively. While not ideal, these values are relatively close to one and indicate at least some correlation.

To attempt to quantify the similarity between the clusterings of the two datasets, we first used the dendrograms generated through hierarchical clustering to set a

Figure 17. Two dimensional plot of clusters from DBSCAN clustering of health behavior and outcome data.

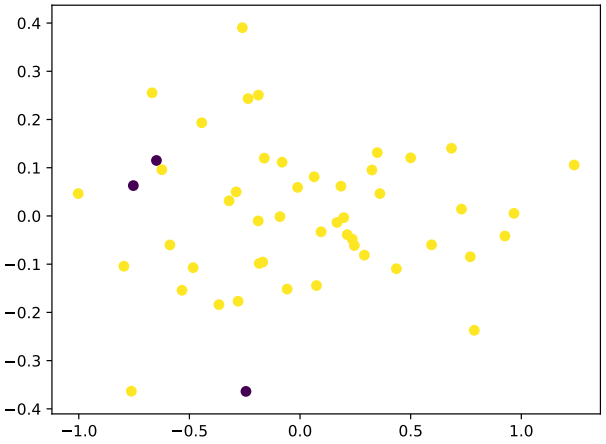


Figure 18. Geographic plot of clusters from DBSCAN clustering of health behavior and outcome data.

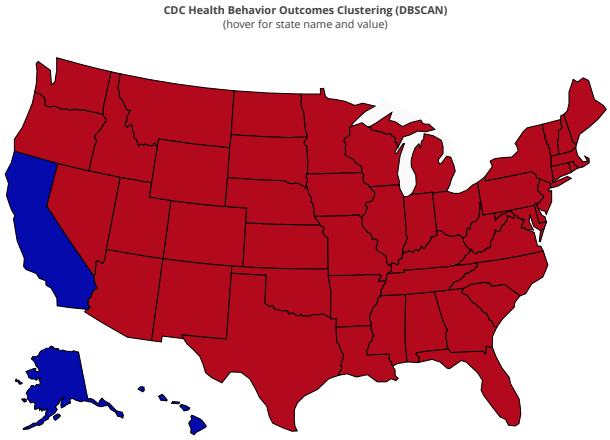
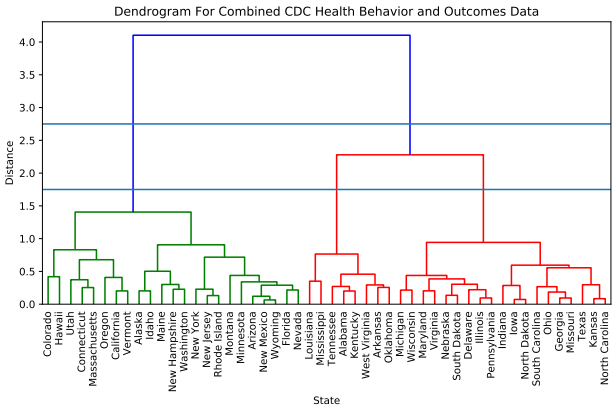


Figure 19.



cutoff distance to generate specific clusterings. These cutoff distances were 2.75 for the socioeconomic data clustering and 1.75 for the CDC health behavior outcomes data clustering. We then measured the percentage of state pairings clustered identically (either in same or different clusters) between both the socioeconomic risk factor clustering and CDC health behavior and outcome clustering. This similarity measure was 56.9%, which seems to indicate some degree of similarity between these two clusterings. More detailed methods for comparing hierarchical clusterings can be utilized in future works, which can help identify states with similar socioeconomic risk factors but very different health behavior outcomes.

Table 1. First two principal components for CDC health behavior outcomes data

PC1		PC2	
% <1 Fruit / Day	0.459181	% Vigorous Aerobic	-0.627665
% Obesity	0.418273	% Overweight	-0.410559
% Inactive	0.411431	% Obesity	-0.409127
% Overweight	0.407122	% Inactive	0.329247
% Vigorous Aerobic	-0.379775	% <1 Fruit / Day	-0.29556
% <1 Veg/ Day	0.367353	% <1 Veg/ Day	0.272643

Table 2. First two principal components for socioeconomic data

PC1		PC2	
$ Median family income	0.461288	% High school graduate	-0.575076
$ Median household income	0.452155	% Black or African American	0.526907
$ Per capita income	0.389594	% White	-0.370121
% Bachelor's degree	0.382095	% Some other race	0.322323
% Advanced degree	0.370484	% Advanced degree	0.254567
% High school graduate	0.318128	% American Indian and Alaska Native	-0.218517
% Black or African American	-0.143937	% Asian	0.127889
% Asian	0.109871	% Bachelor's degree	0.094785
% Some other race	0.072519	$ Median household income	0.077591
% Two or more races	0.070956	$ Per capita income	0.072180
% Native Hawaiian and Other Pacific Islander	0.052622	$ Median family income	0.033356
% White	-0.025696	% Native Hawaiian and Other Pacific Islander	0.018870
% American Indian and Alaska Native	-0.003568	% Two or more races	0.013513

References

Centers for Disease Control and Prevention. Nutrition, physical activity, and obesity data. https://chronicdata.cdc.gov/health-area/nutrition-physicalactivity-obesity. Accessed: 2017-10-05.

Hurt, Ryan T, Kulisek, Christopher, Buchanan, Laura A, and McClave, Stephen A. The obesity epidemic: challenges, health initiatives, and implications for gastroenterologists. *Gastroenterology & hepatology*, 6(12):780, 2010.

National Heart, Lung, and Blood Institute. Why obesity is a health problem. https://www.nhlbi.nih.gov/health/educational/wecan/healthy-weight-basics/obesity.htm, 2013. Accessed: 2017-10-05.

National Institute of Diabetes and Digestive and Kidney Diseases. Understanding adult overweight and obesity. https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity, 2012. Accessed: 2017-10-05.

Wikipedia. Demography of the united states. https://en.wikipedia.org/wiki/Demography_of_the_United_States, a. Accessed: 2017-11-18.

Wikipedia. List of u.s. states by educational attainment. https://en.wikipedia.org/wiki/List_of_U.S._states_by_educational_attainment, b. Accessed: 2017-11-18.

Wikipedia. List of u.s. states by income. https://en.wikipedia.org/wiki/List_of_U.S._states_by_income, c. Accessed: 2017-11-18.