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.
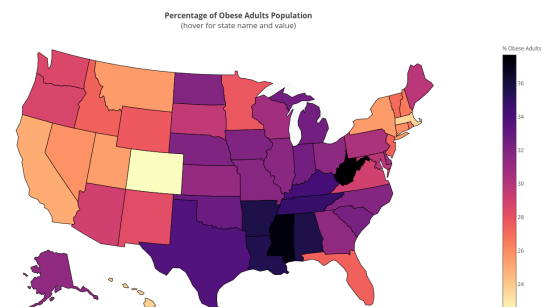

Figure 1. Obesity by State

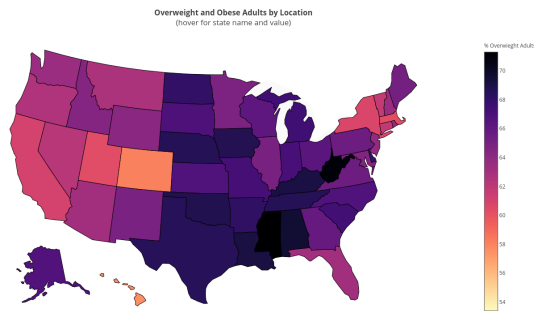*Figure 2.* Overweight Adults by Location

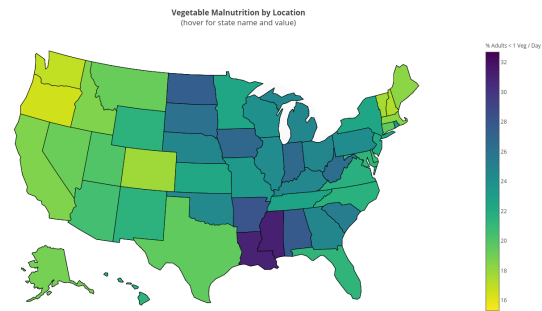*Figure 3.* Inactivity by Location

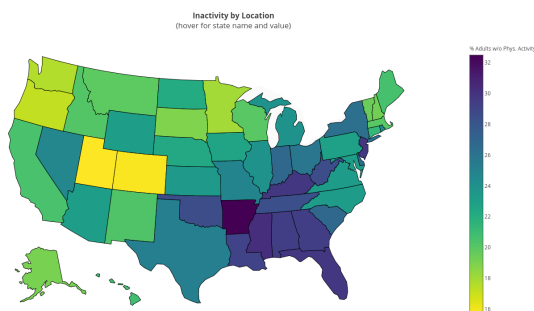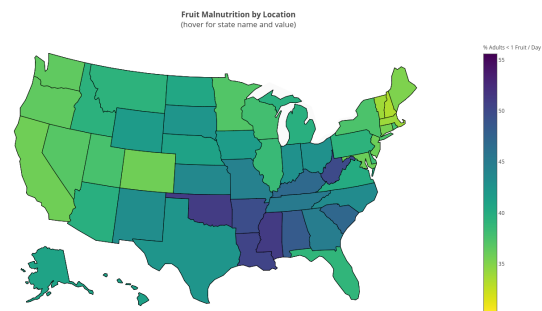*Figure 4.* Vegetable Malnutrition by Location

Figure 5. Fruit Malnutrition by Location

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 to 8.

We also wanted to create plots to visually look at the percentages of adults in these other conditions and obesity based on age and educational risk factors. We have included a few of these plots in figs. 9 to 11 below.

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
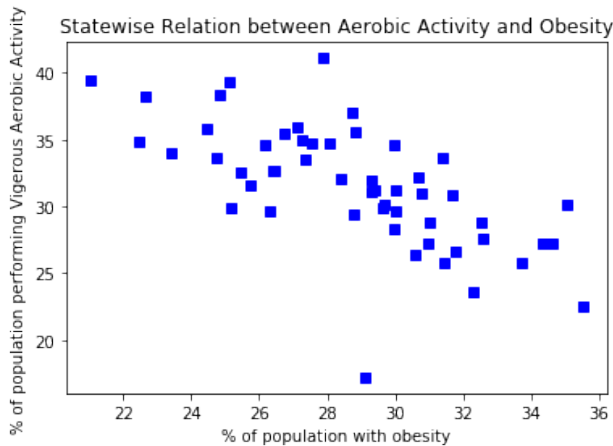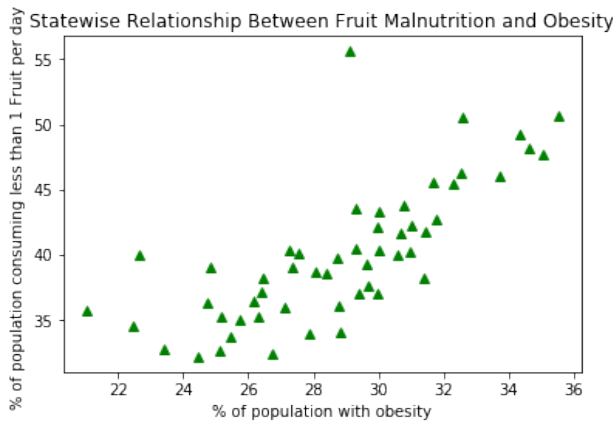
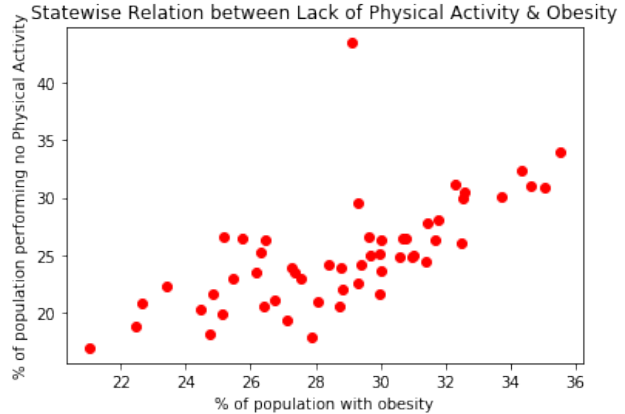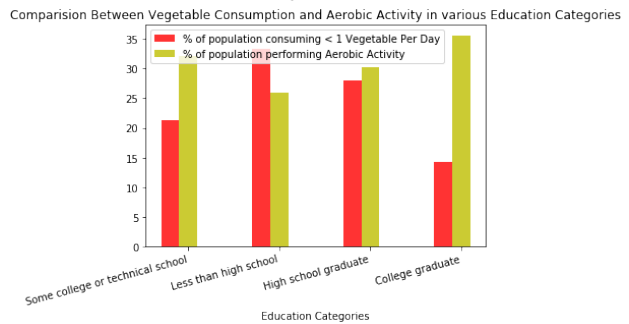Figure 6.

Figure 7.

Figure 8.

Figure 9.

??. 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 **??**.

4.2. CDC Health Behaviour 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 con-

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 **??**. 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 (**??**) and state map (**??**). The data was reduced to two dimensions for plotting by applying principal component analysis (PCA) and selecting the first two principal components.

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 **??** and
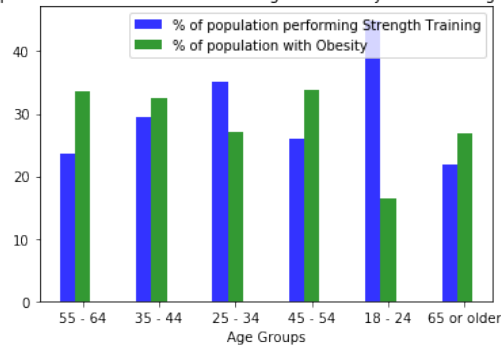
Figure 10.
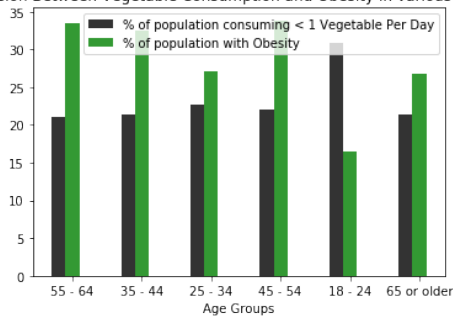
sume less than one fruit per day.

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 **??**. 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 (**??**) and state map (**??**). The data was reduced to two dimensions for plotting by applying principal component analysis (PCA) and selecting the first two principal components.

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 **??** and **??**. 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.

Figure 11.

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 **??**.

5. Discussion

When visualized in two dimensions, both sets of data do not display any clear clusters. Through these plots, clusters appear

6. Future Work

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.