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 overwieght 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 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

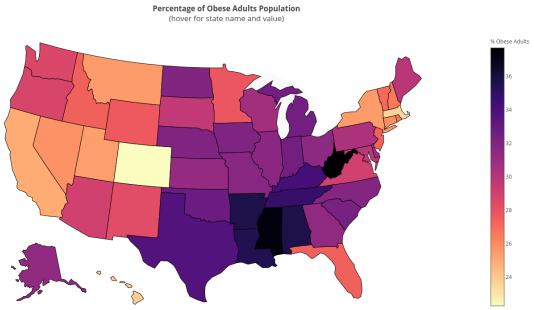


Figure 2. Overweight Adults by Location

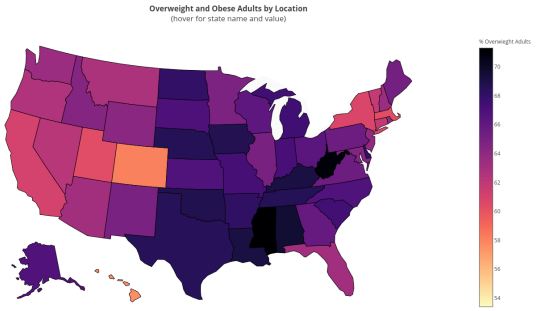
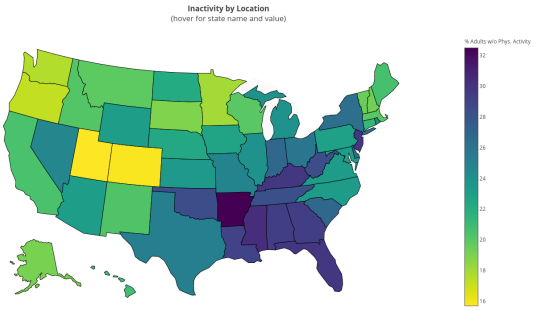


Figure 3. Inactivity by Location



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.

Figure 4. Vegetable Malnutrition by Location

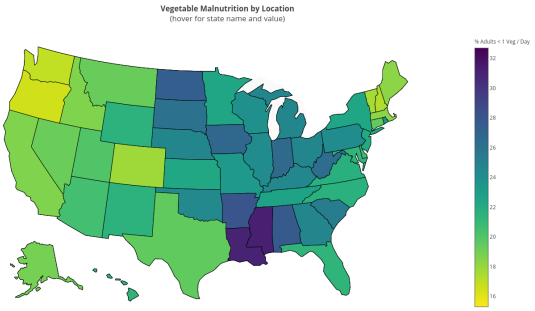
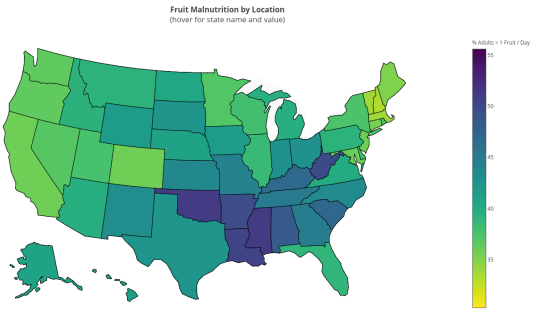


Figure 5. Fruit Malnutrition by Location



4. Data Analysis

4.1. Socioeconomic Risk Factors

4.1.1. K-MEANS

4.1.2. DBSCAN

4.1.3. HIERARCHICAL CLUSTERING

4.2. CDC Health Behaviour and Outcomes

4.2.1. K-MEANS

4.2.2. DBSCAN

4.2.3. HIERARCHICAL CLUSTERING

5. Discussion

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,

Figure 6.

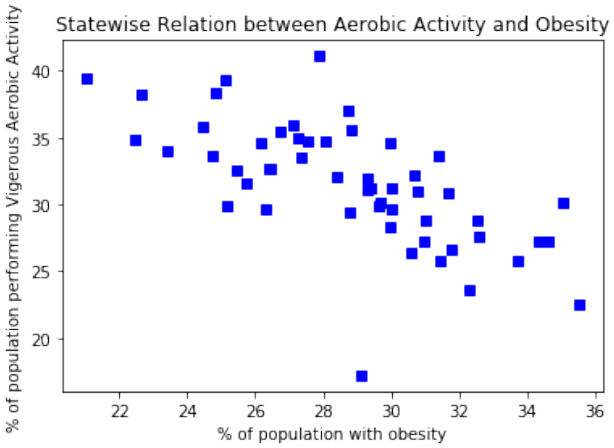


Figure 7.

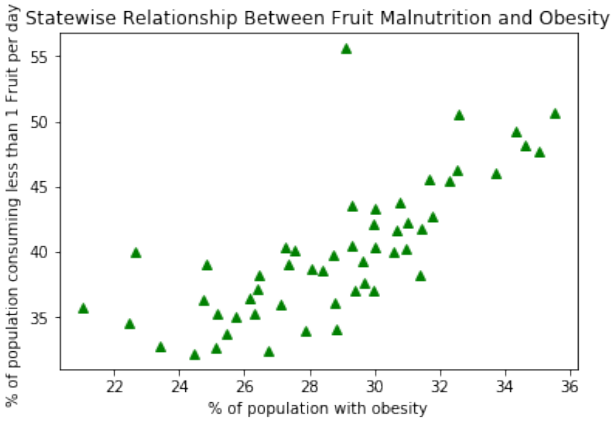


Figure 8.

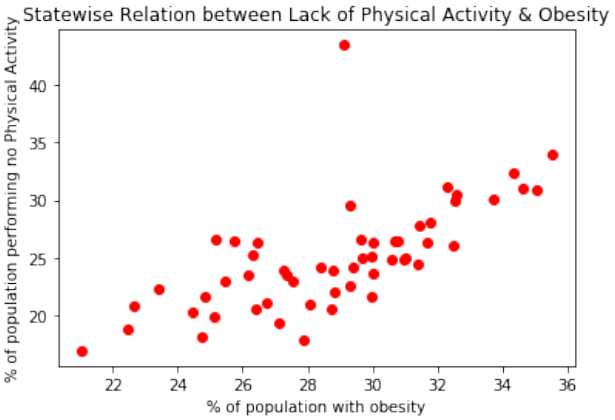


Figure 9.

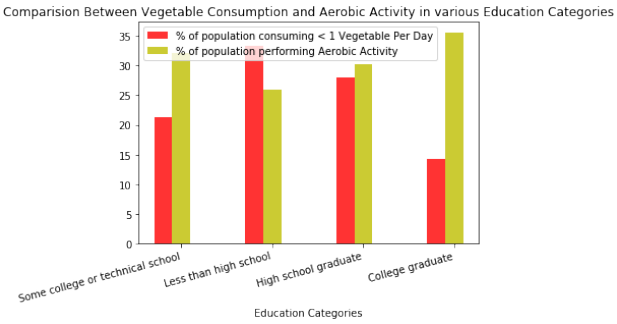
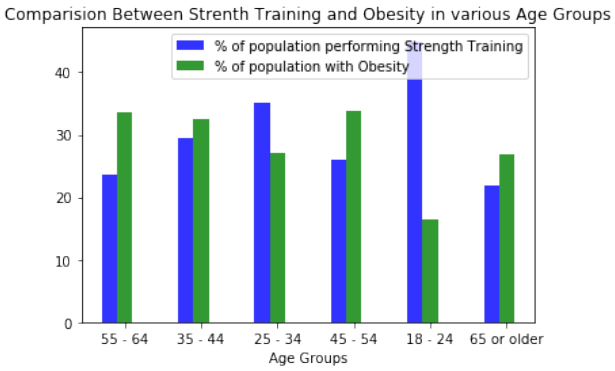


Figure 10.



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.

Figure 11.

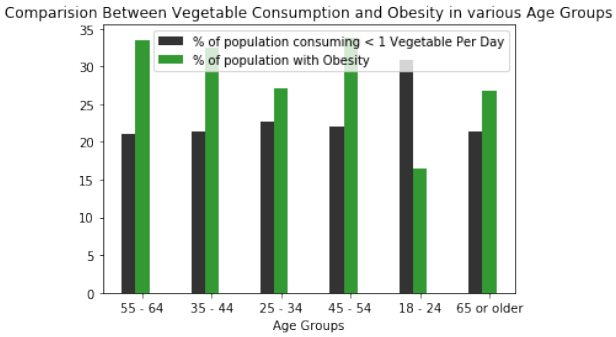


Figure 12. SSE vs k

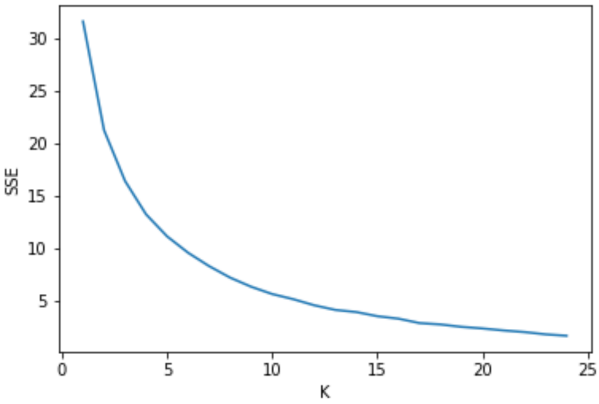


Figure 13. PCA

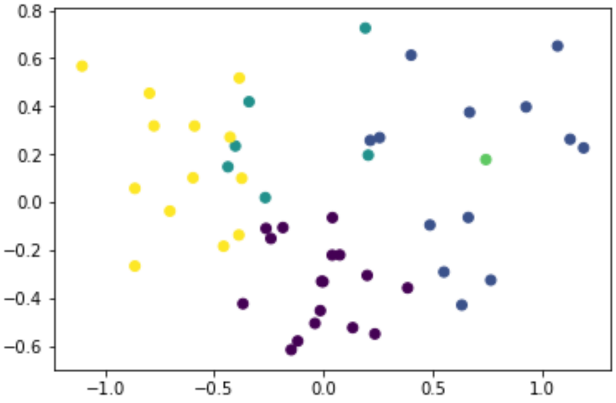


Figure 14. k-means

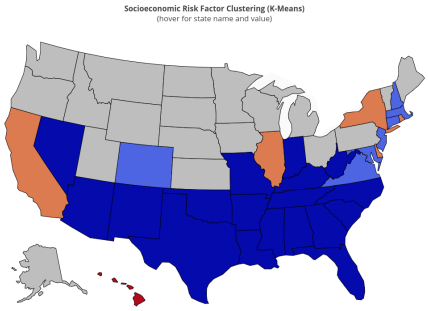


Figure 15. PCA

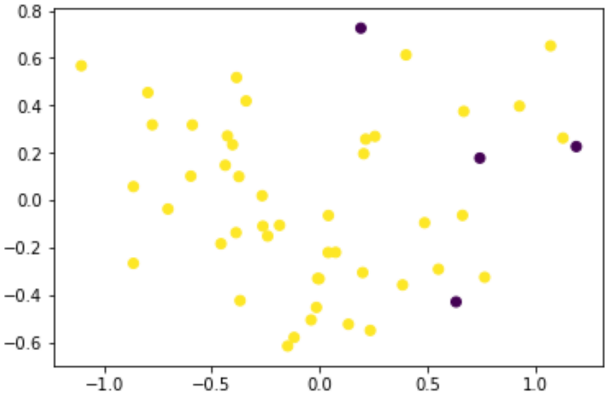


Figure 16. DBSCAN

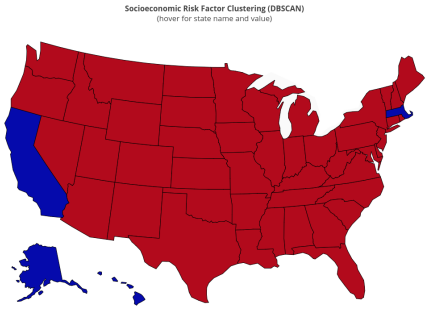


Figure 17. Agglomerative

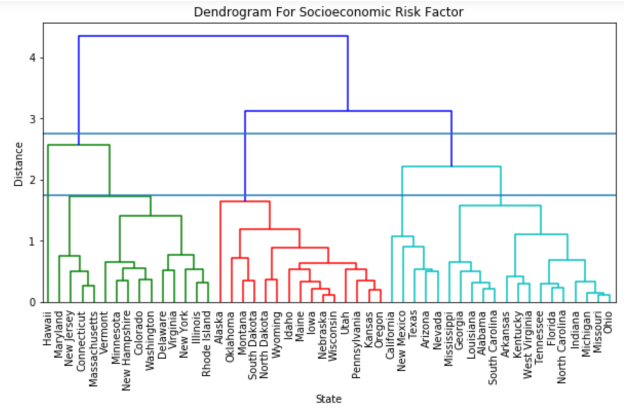


Figure 18. SSE vs k

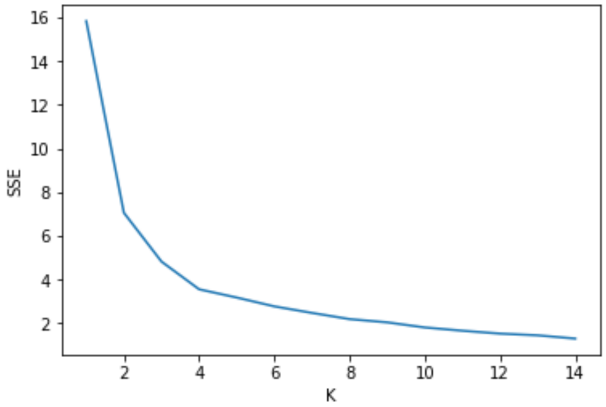


Figure 19. PCA

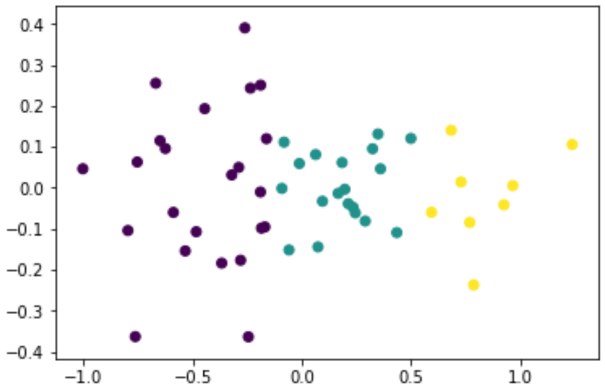


Figure 20. k-means

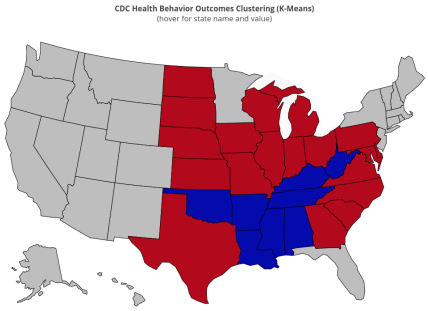


Figure 21. PCA

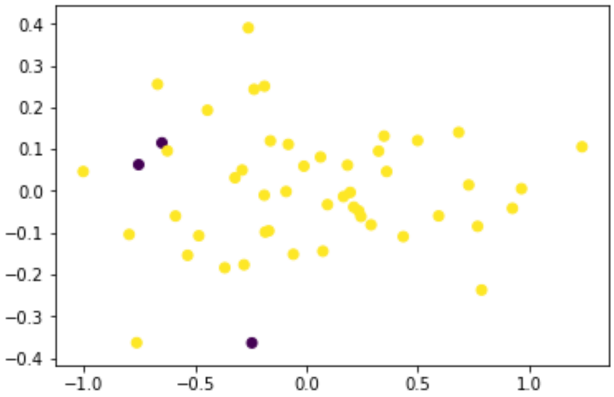


Figure 22. DBSCAN

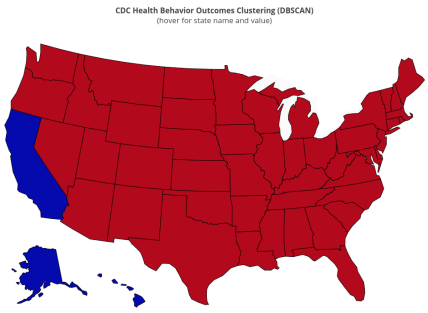


Figure 23. Agglomerative