Group 2 Update 1: United States Obesity Risk Factor Exploration

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 Figures 1 to 11 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.

4. Next Steps and Future Work

Now that this exploratory analysis is complete, we have a fair idea of how to move forward with grouping regions based on their obesity and nutrition statistics. Our team plans to move forward with cluster analysis and visualizations.

Figure 1. Obesity by State

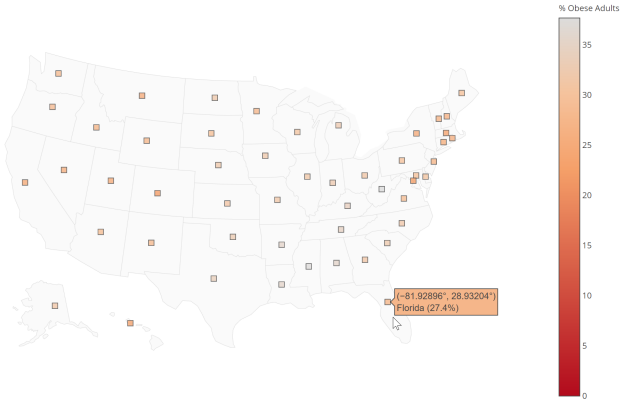


Figure 2. Overweight Adults by Location

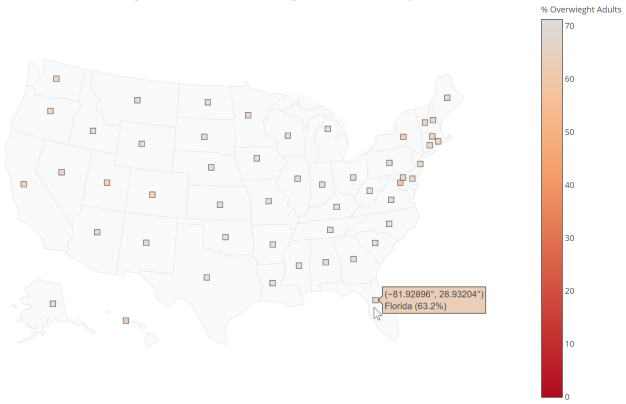


Figure 3. Inactivity by Location

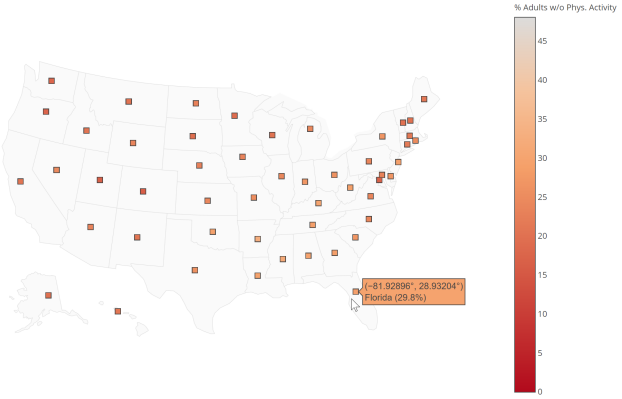
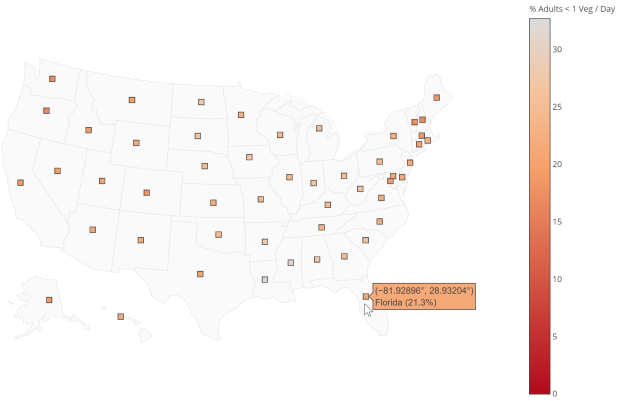


Figure 4. Vegetable Malnutrition by Location



We will start with hierarchical clustering with euclidean distance metric and visualize by presenting the links between different regions using a dendogram.

The next step, as detailed in the proposal, would be to experiment with single-linkage and complete-linkage clustering and reason out the differences.

Comparing these results with DBSCAN would give us a more precise assessment of how accurate our clustering is by identifying noise and arbitrarily shaped clusters.

The final step of the cluster analysis would involve performing k-means for several values of k, plotting objective function for each. We will select a k based on the "kink" in objective function and then proceed with plotting the results.

References

Centers for Disease Control and Prevention. Nutrition, physical activity, and obesity data. https://chronicdata.cdc.gov/health-area/nutrition- 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/weca 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/weigh 2012. Accessed: 2017-10-05.

Figure 5. Fruit Malnutrition by Location

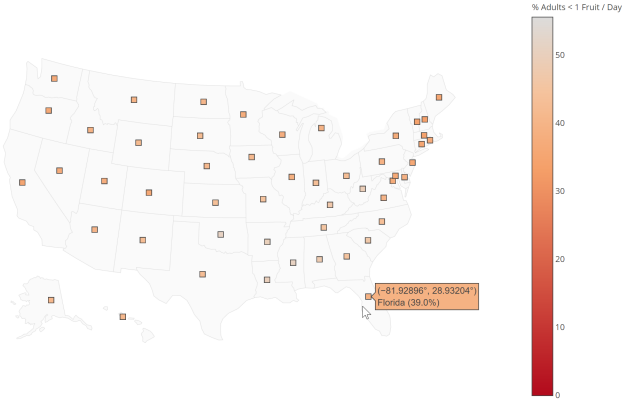


Figure 6.

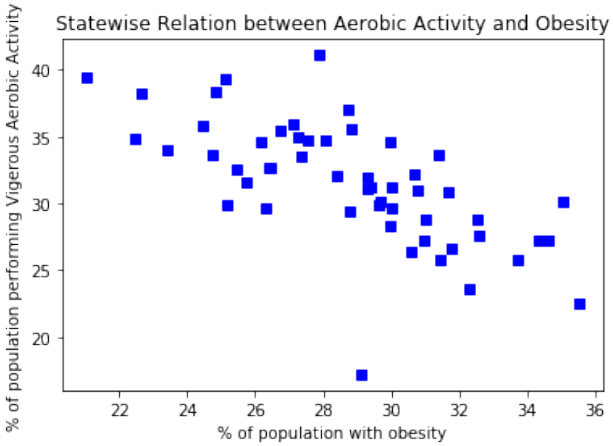


Figure 7.

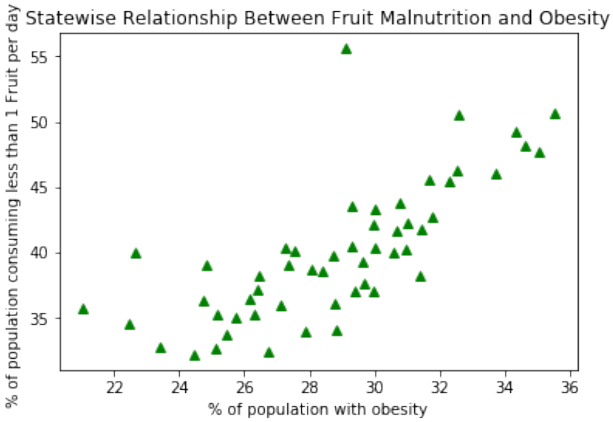


Figure 8.

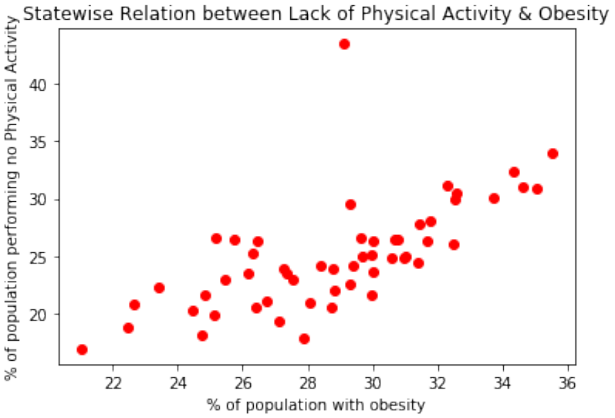


Figure 9.

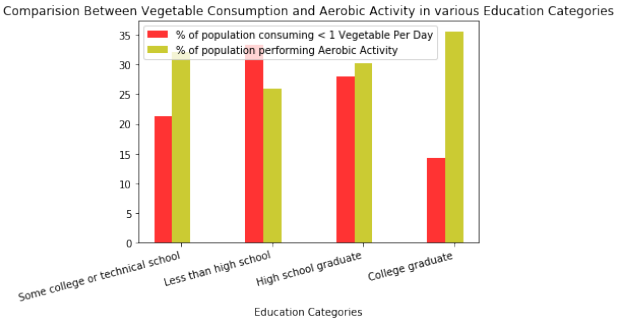


Figure 10.

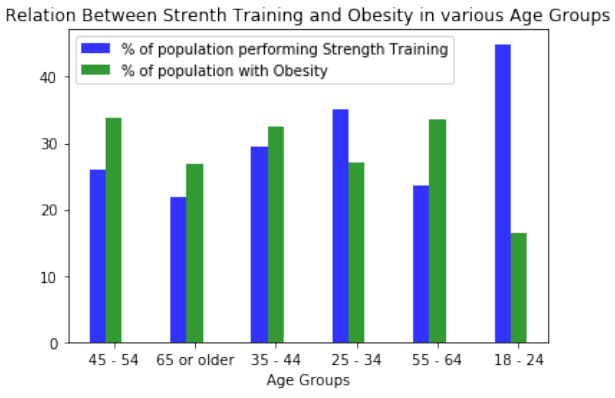
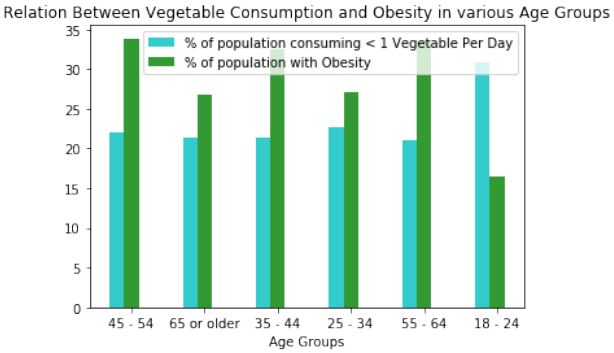


Figure 11.


Relation Between Vegetable Consumption and Obesity in various Age Groups
% of population consuming < 1 Vegetable Per Day
% of population with Obesity
35
30
25
20
15
10
5
0
45 - 54
65 or older
35 - 44
25 - 34
55 - 64
18 - 24
Age Groups