Group 2 Project Proposal: 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

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.

2. Dataset

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.

3. Purpose

The primary purpose is to correlate risk factors such as poor nutrition and physical inactivity with obesity and find if they affect population's health. We will start by analyzing the distribution of obesity, physical activity and nutrition conditions across the United States using state-by-state data from 2011 to 2016. This will eventually lead to generating visualizations which we shall compare with those provided by the CDC. There will also be a need to link the regions which are closely associated. We plan to start analysis first on obesity statistics and then physical activity and nutrition.

The next step is to figure out how the regions have similar obesity risk factors, including physical inactivity and nutrition. This will involve comparing the risk factor regions with obesity statistics. Different socioeconomic, racial, and ethnic groups will also be studied to learn more about how inequalities contribute to higher obesity rates in certain communities.

The final part will be to map out the prevalence and trend of obesity along with its dependence on risk factors. We aim to create multiple visualizations to better understand how such risk factors contribute to health.

All of these steps will eventually help us answer whether environmental risk factors have a definite impact on population's health and obesity.

4. Methodology

We will attempt to use the Nutrition, Physical Activity, and Obesity dataset to generate heatmaps of obesity, physical inactivity, and poor nutrition which can be compared to those provided by the CDC (Centers for Disease Control and Prevention, 2017).

To begin to understand the best grouping of regions based on their obesity statistics, we will perform clustering analysis. We will start this analysis by performing hierarchical clustering with a Euclidean distance metric so that we will be able to visualize and present the links between different regions through a dendrogram. Starting with average linkage clustering, we will also experiment with the results of complete and single-linkage clustering, explaining the reasoning for any differences. We

will compare these results with those gained from the density-based spatial clustering of applications with noise (DBSCAN) algorithm, which has the benefits of finding arbitrarily-shaped clusters and being more robust to noise. We then plan to perform a similar clustering analysis for physical inactivity and malnutrition statistics.

We will then attempt to perform a clustering analysis to begin to understand ways in which geographical regions have similar obesity risk factor conditions. Initially, we will utilize a k-means clustering analysis for this. To do this, we will first perform a k-means clustering analysis for the obesity statistics, similarly to when we performed hierarchical clustering and DBSCAN. We will perform this analysis for several values of k, plotting the objective function value for each value of k. We will select the value for k corresponding to the "kink" in the objective function and use this k value for our clustering of risk factor conditions. This will allow us to compare the risk factor condition regions with those of obesity statistics. We will then experiment with clustering of risk factor conditions for other values of k and plot the results.

To better understand how much these environmental risk factor values contribute to population obesity, we then plan to perform dimensionality reduction on these features. We will start with principal component analysis (PCA) and use the first two principal components to make a two-dimensional plot of the risk factors. We will also make three-dimensional visualizations using the first three principal components. These plots will be color-coded based on obesity prevalence groupings identified through previous analyses.

We then plan to explore similar dimensionality reduction analysis to produce the best human-readable visualization of these environmental risk factors. From this visualization, using `D3.js`, we will then create a 3D interactive visualization that allows for the visualization of obesity prevalence in the United States and its dependence on risk factors. This visualization will hopefully demonstrate that environmental risk factors have a clear effect on population health.

5. Distribution of Work

- Brian Desnoyers - Data cleaning, heatmap, hierarchical cluster analysis on obesity
- Alankrit Joshi - Hierarchical analysis on nutrition and physical activity, cluster visualizations
- Rahul Kondakrindi - Average linkage clustering, spatial clustering
- Prasanna Vikash Peddinti - k-means clustering, clustering visualizations
- Junyi Wang - Dimensionality reduction analysis, `D3.js` visualizations

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.

Centers for Disease Control and Prevention. Adult obesity prevalence maps. https://www.cdc.gov/obesity/data/prevalence-maps.html, 2017. 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.