Group 2 Project Proposal: United States Obesity Risk Factor Exploration

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

1. Introduction

Obesity is a major challenge facing the healthcare system in the United States. With obesity rates increasing significantly over the last thirty years, if this trend continues, the United States healthcare system is projected to pay $150 billion annually (**?**). In addition to genetic factors, this increase in obesity rates, inactivity, and malnutrition has been linked to the enviroment. Enviromental 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 (**??**). 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 the incidence of obesity.

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) provides information on the percentage of the population suffering from adult obesitiy, 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 featues, such as age, education, sex, and income (**?**). In addition, this dataset provides similar population information for child obesity, infant breastfeeding, active transportation, and community policy supports.

3. Purpose

The main overarching question this project will investigate is: which regions of the United States suffer most from these negative enviromental trends and...

This project will seek to answer several...

4. Methodology

To begin, our group 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 (**?**). To begin to understand the best grouping of regions based on their obesity statistics, we will perform a clustering analysis. We will start this analysis by performing heirarchical 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 arbitrarialy-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 analyses 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 dependance on risk factors. This visualization will hopefully demonstrate that enviromental risk factors have a clear effect on population health.

5. Distribution of Work

Distribution goes here.