CS 6140 P2: Data Collection

Grace Liu, Kurt Namini

We obtained our data from the US Department of Housing and Urban Development. Specifically, we downloaded the "2007 - 2022 Point-in-Time Estimates by CoC (XLSX)" from the HUD Exchange website: **https://www.hudexchange.info/resource/3031/pit-and-hic-data-since-2007/**. The original excel file is 7.061 megabytes, with 16 different sheets (one for each year, 2007 – 2022). Our plan is to combine each year's table, which may result in a table, or matrix, with the dimensions 577 x 6176 (roughly) depending on how preprocessing is done.

The data is stored as a point set, with each CoC representing a single point, and the various demographic counts of homeless cared for by that CoC representing each coordinate, of which there are 577 (for now). Data across years varies in two ways: the number columns, which represents the homeless count for a specific demographic, and the number of CoC's as some have merged over the years. In order to properly combine the tables across all years, both of these issues must be addressed before performing any analysis. The more recent years have 577 columns, but overall, there is an average of 252 columns across all years. What we will most likely do is ignore features/columns that do not appear across all years, and additionally only use years that have enough features to be of any use. The data we are using includes 2007 – 2009, but we have decided to use data from 2010 – 2022, as stated in the project proposal.

For combining tables there are multiple possibilities. It mainly depends on how we intend to cluster the data. For example, we may want to cluster points from each year separately, which would mean we need to determine which features appear in all years used in analysis, and not necessarily combine the datasets into a single, large set. Another possibility would be to use the selected features across all years to cluster a single point (CoC). For the next steps in this project, we will discuss what will work best with clustering and makes the most sense for the context of the data.

Clustering this data will provide insight into where certain demographics struggle with and need support for homelessness across multiple years. However, this is just the first step in understanding the problem. Once we have some definite information/results to work with we can begin researching why a certain demographic deals with higher levels of homelessness. This may require analyzing population data across the U.S.A., income data, or even housing price trends. The point is that clustering is the first step and will allow us to research further into why homelessness is an issue, not just for who/where/when it is an issue.