An Application of Machine Learning to Health and Social Vulnerability Data

Dr. Holly Rosson and Dr. David Abernathy
Student Researchers: Gracen Alleman, Neylan Visnius, Mason Hollar

Abstract

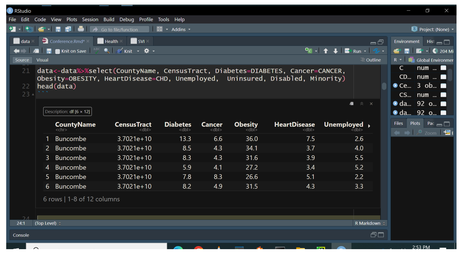
Clustering is a data analysis technique that divides a dataset into distinct groups, with each group containing similar data points. We apply clustering to CDC social vulnerability and health conditions data to identify and map geographic regions in Western North Carolina that are most suitable for targeted outreach efforts by **Bounty and Soul**: a local food bank which provides access to fresh produce and wellness education for individuals in Buncombe, Henderson, and McDowell counties.

The Data

CDC PLACES data contains percent estimates of multiple health conditions for all census tracts across the United States. **Cancer**, **diabetes**, **obesity**, and **heart disease** were chosen to be part of our clustering algorithm because individuals with these diseases would likely benefit the most from dietary interventions from Bounty and Soul.

Similarly, **CDC's Social Vulnerability Index** contains estimates of multiple social vulnerability variables for all census tracts across the United States. **Median family income** and percent **minority, unemployed, uninsured, and disabled** were also included in our clustering algorithm.

Process



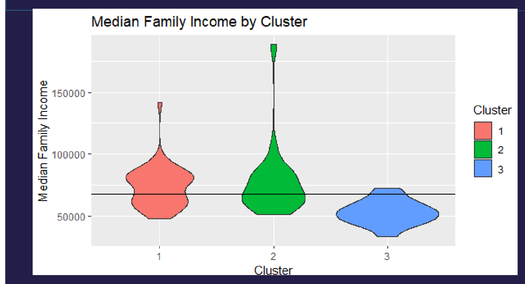
- **R Studio** was used to clean, reorganize, and merge the datasets
- 3 clusters were created using the **k-means clustering algorithm**
- **Principal component analysis**, as well as basic descriptive statistics provided a characterization of the clusters
- **QGIS** was used to map the clusters

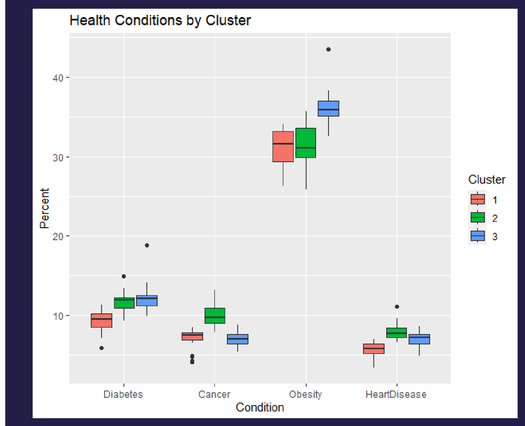
References

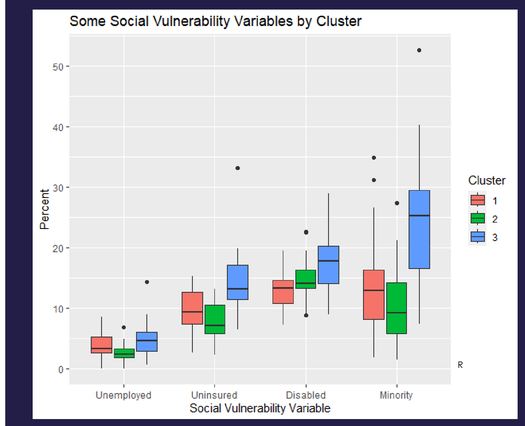
- **CDC Social Vulnerability Index** 2010, accessed November 1, 2022: https://www.atsdr.cdc.gov/placeandhealth/svi
- **CDC PLACES** 2020 for North Carolina, accessed November 1, 2022: https://www.cdc.gov/places/

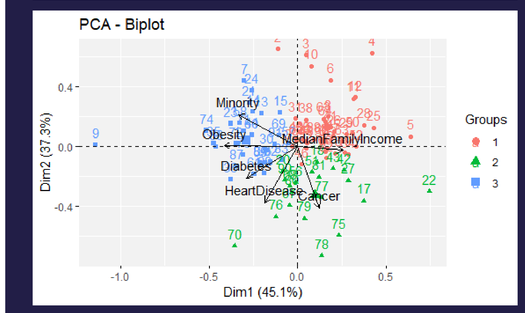
- This research was supported by a SEEDS grant from the **South Big Data Innovation Hub.**

Some Visualizations

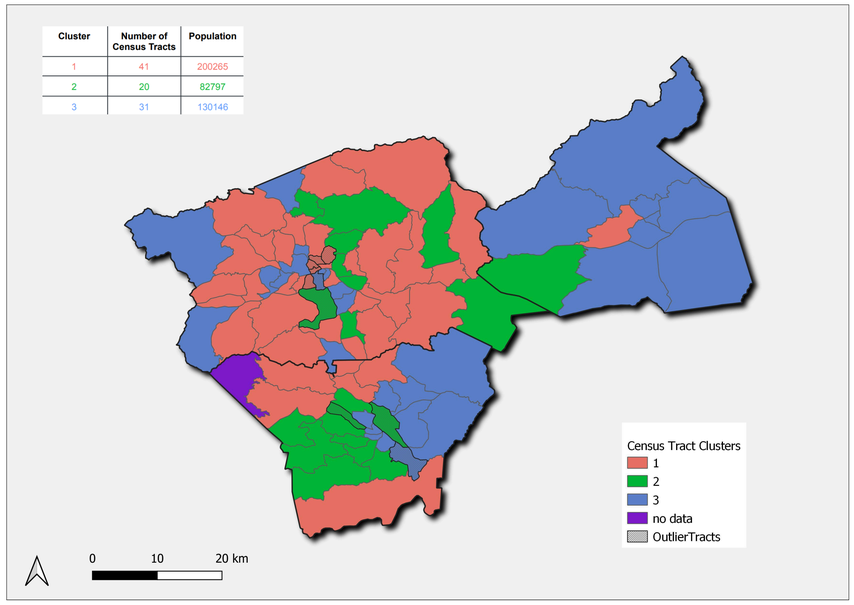








The Map



Conclusions

- **Cluster 1** consists of census tracts characterized by predominantly above-average income levels, as well as the lowest rates of diabetes and heart disease.

  There are a few outliers in this cluster. First, a group with high values of PC2—these census tracts have more minorities and less cancer (River Arts district and Montford), and a second with a high value on PC1--this tract has a much higher median family income (North Asheville).

- **Cluster 2** has a wider range of median family income than cluster 1, but many of the census tracts have above average income for the area. It also has the highest rates of cancer, as well as the lowest social vulnerability variables.

  Most extreme values in cluster 2 (in Henderson county) are quite low on PC2, so have much higher rates of heart disease and cancer, as well as lower minority residents. One census tract has a much higher median family income (Biltmore Forest in Asheville).

- **Cluster 3** has the lowest income, highest social vulnerability variables, as well as the highest rates of diabetes and obesity.

  The outlier in cluster 3 (Carrier Park area in Asheville) has a very low value of PC1. This tract has an extremely low median family income.