Memo: Replicating County Clustering in Wallace, Sharfstein, and Kaminsky (2019)

Project Overview:

As detailed in my first memo, for my project I plan to replicate several of the main results of Wallace, Sharfstein, and Kaminsky (2019). Particularly, this paper uses county-level data on eleven demographic characteristics, which primarily comes from the Census Bureau's Census Population Estimates ("Population," 2020) and American Community Survey ("American Community Survey Data," 2020). The authors then apply *k*-means clustering to this county-level data in order to identify groups of counties with similar characteristics and end up identifying eight clusters. The authors then use three variables on county health characteristics from the Robert Wood Johnson Foundation ("Explore Health Rankings," 2020) to assign percentile ranks to these counties both nationwide and among counties within their cluster. Overall, I plan to reconstruct the authors' dataset and apply *k*-means clustering as well as density-based and hierarchical clustering to the data.

Data:

As previously mentioned, the authors' data primarily comes from two Census Bureau sources and the Robert Wood Johnson Foundation's County Health Rankings dataset ("Explore Health Rankings," 2020). They use demographic data to conduct the actual clustering, and then use the clusters to assign percentile ranks to counties on three metrics: population-adjusted driving deaths, the percent of the adult population that smokes, and the percent of the adult population that is obese (Wallace, Sharfstein, and Kaminsky, 2019). Most of their data is for years during the period 2010-2014. I use the same data sources and variables as Wallace, Sharfstein, and Kaminsky (2019), but generally use the most recent available data when possible (generally, the period 2014-2018). Time and space permitting, I may recreate my analysis using the years of their data to see if changes in these variables over this period substantially impacted the results.

First, most of the data on county-level demographics comes from the American Community Survey ("American Community Survey Data," 2020). Particularly, we obtain the following variables used by Wallace, Sharfstein, and Kaminsky (2019): the percent of the population with at least some college education, the percent of the population without health insurance, the percent of the population currently married, the median age, the percent of the population that is female, the percent of the population that is in the labor force and unemployed, and four variables related to the racial composition of the county. (These are all described in more detail in the table below. We also collect, and provide summary statistics for, total population but do not plan to use that variable in clustering.) The American Community Survey (ACS) is a random survey by the Census Bureau of households in each year and produces estimates of a vast array of data points at many levels of geographic aggregation. Because one-year estimates exclude counties with small populations ("When to Use 1-year, 3-year, or 5-year Estimates," 2019), we use the 2014-2018 five-year estimates, which are the most recently available such estimates and are averages of values for each county over the years 2014-2018. These five-year averages can be roughly interpreted as values for 2016, assuming a linear trend of variables over this period.

One demographic variable that is not available from the ACS is the percent of the county population that lives in a rural area, which comes from the 2010 decennial Census ("By Decade," 2018). Unlike the American Community Survey, the Census surveys the entire population and thus, in theory, has no margin of error to its estimates. For this variable, anyone who does not live in an "urban cluster" of at least 2,500 residents meeting certain population density thresholds is considered to live in a rural area ("Urban and Rural," 2020).

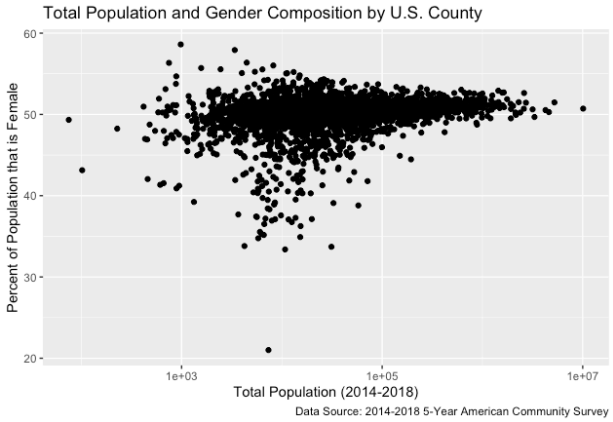
Finally, we obtain three variables related to county health outcomes from the County Health Rankings ("Explore Health Rankings," 2020) provided by the Robert Wood Johnson Foundation. These annual rankings combine data from various sources on various metrics related to healthcare and health outcomes by county. We use the 2016 rankings as this corresponds to the midpoint of the 2014-2018 5-year ACS period, and we look at three variables from these rankings. First, we consider the annual number of motor vehicle deaths per 100,000 population, which is a 5-year average over the period 2010-2014. Additionally, we consider the percent of the adult population that is obese (based on Body Mass Index) and that smokes, which come from 2012 and 2014, respectively.

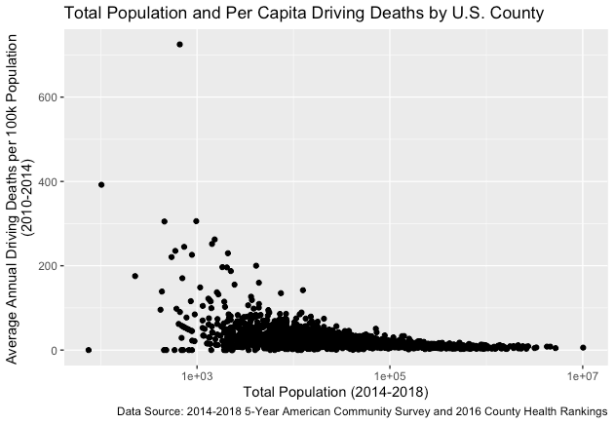
Overall, out of 3,143 total counties in the U.S. ("Reference Files," 2019), we drop eight that have primarily missing data and consider 3,135 remaining counties. For comparison, Wallace, Sharfstein, and Kaminsky (2019) consider 3,139 counties. Summary statistics of these variables across counties are provided in the table below:

Variable Description	Data Source	Year(s)	Number of Observations	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
Total Population	American Community Survey	2014-2018 (5-year Annual Average)	3135	75	11002	25796	102987	67443	10098052	330243.8
Percent of Population Aged 25+ with at Least Some College Education	American Community Survey	2014-2018 (5-year Annual Average)	3135	15.18	49.86	57.96	57.90	66.48	100.00	11.82
Percent of Population Under Age 65 without Health Insurance	American Community Survey	2014-2018 (5-year Annual Average)	3135	1.948	7.524	11.089	12.192	15.483	49.455	6.01
Percent of Adult Population Currently Married	American Community Survey	2014-2018 (5-year Annual Average)	3135	21.52	49.68	54.05	53.42	57.67	78.80	6.58
Median Age of Population	American Community Survey	2014-2018 (5-year Annual Average)	3135	21.70	38.00	41.20	41.29	44.40	67.00	5.40
Percent of Population that is Female	American Community Survey	2014-2018 (5-year Annual Average)	3135	21.00	49.41	50.38	49.92	51.11	58.61	2.38
Percent of Population that is Non-Hispanic White	American Community Survey	2014-2018 (5-year Annual Average)	3135	0.7278	64.87	83.99	76.57	92.70	100.00	20.10
Percent of Population that is Non-Hispanic Black	American Community Survey	2014-2018 (5-year Annual Average)	3135	0.00	0.64	2.17	8.94	9.87	87.41	14.48
Percent of Population that is Non-Hispanic American Indian and Alaskan Native	American Community Survey	2014-2018 (5-year Annual Average)	3135	0.00	0.13	0.28	1.74	0.67	82.48	7.16

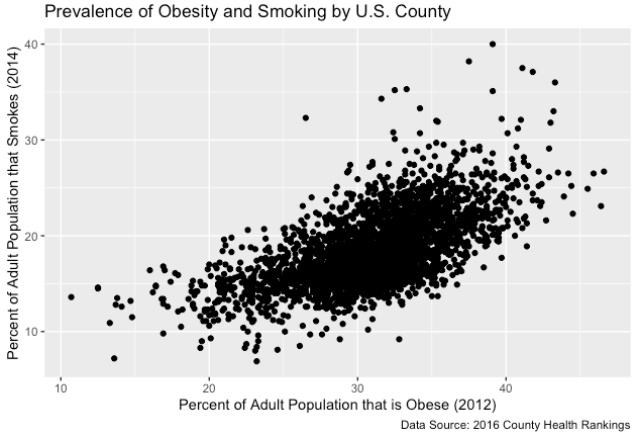
Percent of Population that is Hispanic	American Community Survey	2014-2018 (5-year Annual Average)	3135	0.00	2.112	4.101	9.272	9.563	99.069	13.80
Percent of Population Aged 16+ that is Unemployed but Seeking Work	American Community Survey	2014-2018 (5-year Annual Average)	3134	0.00	2.381	3.147	3.249	3.925	16.470	1.43
Percent of Population Living in Rural Areas	Decennial Census	2010	3135	0	33.23	59.47	58.58	87.78	100.00	31.48
Annual Number of Driving Deaths Per 100,000 Residents	County Health Rankings	2010-2014 (5-year Annual Average)	3135	0.00	11.28	17.49	22.60	26.40	725.08	25.99
Percent of Adult Population that is Obese	County Health Rankings	2012	3135	10.70	28.50	31.20	30.94	33.70	46.60	4.46
Percent of Adult Population that Smokes	County Health Rankings	2014	3135	6.90	15.70	17.80	18.38	20.70	40.00	3.75

We can see that, for most of these variables (with the major exception of total population), the mean is relatively close to the median, providing evidence that these variables may be approximately normally distributed. However, the minimum and maximum values provide evidence of extreme outliers in the data. For instance, there appears to be a county that is only 21 percent female, roughly 12 standard deviations below the median! This is likely very problematic because we know that some clustering methods, including *k*-means, can be very sensitive to outliers (Jin and Han, 2010). Thus, we may want to remove values with outliers before conducting our clustering analysis. Visual inspection of the data reveals that almost all outliers result from counties with extremely small populations (under a few thousand residents). Examples are provided in the scatterplots below, where we can see that all counties that are less than 40 percent female have under 60,000 residents, with most having well under 10,000. Likewise, all counties with over 200 annual driving deaths per 100,000 residents have a population below 5,000. (Note that total population is shown on a log scale.)





Finally, the scatterplot below provides shows the prevalence of smoking and obesity by county, two of our three county health outcomes. We can see that these two variables are highly positively correlated, with a correlation of 0.61. However, both of these variables are very weakly correlated with our other health outcome, population-adjusted driving deaths.



Analytical Technique:

As mentioned above, the authors use *k*-means clustering to identify eight clusters of counties. Clustering models are unsupervised learning techniques that attempt to find similar groups of observations using information on features (Tan, Steinbach, Karpatne, and Kumar, 2019), with no dependent or outcome variables. *K*-means clustering is a centroid-based technique that calculates the Euclidean distance of each data point from a centroid, with the number of clusters (*k*) pre-specified by the user (Tan et al., 2019). I plan to replicate the analysis along several dimensions. First, I plan to determine the optimal number of clusters using the sum of squared errors for different values of *k*, as discussed in class, and compare it to the *k*=8 in the paper I am replicating. (The authors provide no information as to why they chose this number of clusters.) Second, I plan to evaluate the similarity of my clusters to the clusters obtained in Wallace, Sharfstein, and Kaminsky (2019). Third, the authors report using 3,135 counties, with no mention of outliers; this is particularly troublesome as we know outliers can substantially distort the result of *k*-means clustering . Thus, I plan to conduct this clustering both including and excluding counties with very small populations (perhaps below 5,000), to see how much the authors' results are affected by these counties.

Additionally, I plan to use two other clustering techniques, hierarchical clustering and density-based clustering. Hierarchical clustering produces a tree-based structure of the data and does not require a pre-specified number of clusters, while density-based clustering finds regions of high-density points and also does not require a pre-specified number of clusters (Tan et al., 2019). A key advantage of both methods over *k*-means clustering is that they can be much less susceptible to outliers, which makes it less likely that counties with very small populations will skew our clusters. Thus, I plan to apply both forms of clustering to this data, and compare the results to *k*-means clustering.

Once I have applied these three clustering techniques, I plan to see how similar the results are across techniques and then determine which clustering technique performs best on the data. One way to do so is to use a "smell test" – for instance, a technique that puts Manhattan in the same cluster as most counties in Wyoming is likely to be highly suspect. Additionally, the Hopkins statistic can be used to compare the clusters to those produced from randomly generated data (Tan et al., 2019). The overall similarity of the clustering techniques will provide a measure of robustness or validity of the results obtained in my analysis and in Wallace, Sharfstein, and Kaminsky (2019).

Works Cited

American Communities Project. American Communities Project. Retrieved October 7, 2020, from https://www.americancommunities.org/

Jin X., Han J. (2011) *K*-Medoids Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_426

Robert Wood Johnson Foundation. (2020). Explore Health Rankings. Retrieved October 7, 2020, from https://www.countyhealthrankings.org/explore-health-rankings

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining*. New York: Pearson Education.

US Census Bureau. (2020, March 20). American Community Survey Data. Retrieved October 7, 2020, from https://www.census.gov/programs-surveys/acs/data.html

US Census Bureau. (2018, April 30). By Decade. Retrieved November 09, 2020, from https://www.census.gov/programs-surveys/decennial-census/decade.html

US Census Bureau. (2020, October 16). Population. Retrieved November 09, 2020, from https://www.census.gov/topics/population.html

US Census Bureau. (2019, July 01). Reference Files. Retrieved November 09, 2020, from https://www.census.gov/geographies/reference-files.html

US Census Bureau. (2020, February 24). Urban and Rural. Retrieved November 09, 2020, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

US Census Bureau. (2019, September 17). When to Use 1-year, 3-year, or 5-year Estimates. Retrieved February 17, 2020, from https://www.census.gov/programs- surveys/acs/guidance/estimates.html

Wallace, M., Sharfstein, J., & Kaminsky, J. (2019). Comparison of US County-Level Public Health Performance Rankings With County Cluster and National Rankings. *JAMA Network Open*, *2(1)*. doi: 10.1001/jamanetworkopen.2018.6816