County Clustering to Assess Counties' Health Outcomes

From: Eric LaRose
To: Professor Brodnax
Date: December 16, 2020

Executive Summary

This report replicates the analysis of Wallace, Sharfstein, and Kaminsky (2019), who use *k*-means clustering to cluster counties into eight groups based on the factors that are most important in predicting obesity, smoking, and motor vehicle crash deaths. The authors argue that clustering can be helpful to policymakers by allowing them to identify peer counties that have similar demographic characteristics but potentially much better or worse health outcomes. This report attempts to verify the choice of clustering technique, the determination of the appropriate number of clusters to use, and the groupings obtained from these techniques. We find that two other clustering techniques, hierarchical clustering and density-based clustering, perform extremely poorly, suggesting that *k*-means is the appropriate technique. We determine that five clusters should be used rather than eight, but find that either number of clusters produces reasonably similar outcomes and there is no clear "best" number of clusters. When using eight clusters, we obtain virtually the same results as the authors. Overall, we determine that the authors' findings are fundamentally replicable and open up several important avenues for future research.

Background

Government agencies, nonprofit organizations, and researchers focused on public health frequently compare the health outcomes of states, counties, or other regions within the United States (see, for instance, "State Reports," n.d., and Pflanzer, 2019). These comparison and rankings are used by health departments at all levels of government to try to improve health outcomes in their communities, particularly on metrics such as the obesity rate and the prevalence of smoking (Frieden, 2011; Beitsch et al., 2006). Mays and Smith (2009) note that there are nearly 3,000 local health departments in the United States, which tend to be financed by a mixture of state and local spending. Many of these departments are severely underfunded and rely heavily on state grants, which are often awarded based on county health outcomes (Sessions, 2010).

At the same time, a vast quantity of public health research has found individuals' sociodemographic and economic characteristics to be highly predictive of, and correlated with, their health outcomes. For instance, Nguyen et al. (2015) find that state-level demographics are highly correlated with the prevalence of chronic disease, while Levine (2011) finds a strong correlation between poverty and obesity across counties. Overall, Link and Phelan (1995) argue that "social factors such as socioeconomic status and social support are likely `fundamental causes' of disease" (p. 80).

The fact that much of the variation in local health outcomes can be explained by the variation in sociodemographic and economic characteristics implies that there is a limited role and responsibility for local health departments in determining and improving local health outcomes. Despite the best efforts of a local county health department, it is extremely unlikely that a very economically distressed county's health outcomes can be improved to match that of an extremely affluent and educated county, because "social inequities are often barriers that few local health departments have the resources to affect" (Wallace, Sharfstein, and Kaminsky, 2019). Thus, in funding situations and otherwise, health departments should be judged in reference to the population characteristics of the communities they serve, not simply in reference to all counties nationwide. Unfortunately, as established above, this is not always the case in funding, and the existence of various county and state health rankings (for example, "Explore Health Rankings," 2020) encourages public officials to think about outcomes in nationwide terms.

Wallace, Sharfstein, and Kaminsky (2019) attempt to fix this problem by creating groups, or *clusters*, of counties with similar characteristics. The authors argue that counties should be assessed on health outcomes within their own cluster, not nationwide; counties that perform poorly *relative to similar*

counties can then look to their higher-performing peers for ideas on interventions to improve health outcomes.

Particularly, Wallace et al. (2019) are interested in the three outcomes of the prevalence of obesity, prevalence of smoking, and the population-adjusted motor vehicle crash death rate. Using individual-level data on demographic characteristics and health outcomes ("Behavioral Risk Factor Surveillance System," 2020), they use a machine learning method called a *random forest* to find the characteristics that are most strongly predictive of these three health outcomes; these characteristics are "individual race and ethnicity, educational attainment, age, marital status, employment status, sex, and health insurance status." The authors then gather data on county-level versions of these variables (described in more detail below) and use a clustering technique called *k*-means clustering (also described more below) to generate eight clusters of counties with similar values on these characteristics. They then assign each county a percentile ranking on the three health outcomes of interest among all counties nationwide and among counties only within their clusters. Wallace et al. (2019) show that the percentile rankings of many counties change drastically when only comparing counties within a cluster, which underscores the need to take relatively immutable demographic factors into account when evaluating health departments in relation to counties' health outcomes.

This report takes the most important determinants identified by Wallace et al. (2019) as given and attempts to replicate the county-level clustering of their analysis. Particularly, while the authors use *k*-means clustering with $k=8$ clusters, they provide no information on why they chose this particular clustering method and almost no information on how they chose the number of clusters. This report reconstructs their dataset and obtains nearly identical results using *k*-means clustering with 8 clusters. However, when attempting to determine the optimal number of clusters, we obtain that there should be 5 clusters, rather than 8. We also show the existence of some extreme outlier counties, which can cause major problems in *k*-means clustering, and examine how the clusters change when removing potential outliers. Next, we apply two other clustering techniques to the data, agglomerative hierarchical clustering and density-based clustering. Finally, we use cluster validation methods to evaluate the performance of these clustering techniques and, potentially, offer improvements to the clustering technique used by Wallace et al. (2019). These clustering techniques and methods for cluster validation are described in much more detail later on in this report.

The findings of Wallace et al. (2019) and this report are arguably a major improvement over current methods of rankings counties' health outcomes. While the problems with simply comparing all counties nationwide have been discussed, it should be noted that the authors are not the first public health researchers to assign counties to peer groups. The Centers for Disease Control and Prevention (CDC), for instance, has assigned counties to peer groups since 2008 (Kanarek, Bialek, and Stanley, 2008). Outside of the field of public health, many other groupings of counties have already been devised for purposes such as promoting economic development and understanding voting patterns (for instance, "American Communities Project," n.d.). There are two major disadvantages of these preexisting analyses. First, Wallace et al. (2019) note that in many cases "it is unclear how the social determinants used to group counties were selected." Second, Wallace et al. (2019) and this analysis focus only on demographic factors that are most relevant to the outcome of interest. For a local public health agency trying to, say, reduce smoking prevalence, looking at actions of peer counties chosen specifically on those factors most important in predicting smoking rates will be much more useful than looking at peer counties chosen based on many potentially irrelevant characteristics.

Additionally, note that other methods besides clustering could potentially be used to identify counties that over- or under-perform on health outcomes given their demographic characteristics. For example, linear regression could be used to predict obesity rates by county based on demographic characteristics and identify counties where the predicted rates are much higher or lower than their actual rates. However, a key limitation of this analysis is that it would not give health departments peer groups to which they can compare themselves. A key advantage of clustering is that, unlike other techniques, it specifically gives local health departments a list of "peers" to which they can look for ideas, inspiration, and guidance.

Data

Data Sources and Description

Unfortunately, Wallace et al. (2019) provide neither the data nor any replication code for their analysis. Ideally, we would be able to use their exact same dataset to see if we are able to exactly replicate their results, which is not possible in this scenario. Luckily, however, all of the county-level data that the authors use is publicly available, and we are able to use the description of their data sources to approximately reconstruct their dataset.

The authors' county-level data primarily comes from two Census Bureau sources and the Robert Wood Johnson Foundation's County Health Rankings dataset ("Explore Health Rankings," 2020). Most of their data is for years during the period 2010-2014. We use the same data sources and variables as Wallace et. al but use the most recent available data when possible (generally, the period 2014-2018). Since the variables described below would be expected to change only very slowly in a county, a difference of a few years in data between our analysis and theirs should not have a substantive impact on their 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 nine that have missing data and consider 3,134 remaining counties. For comparison, Wallace et al. (2019) consider 3,139 counties. The implementation appendix provides more detail on the data sources and any transformations and cleaning steps applied.

Description of Clustering Variables

Table 1, below, provides summary statistics of the eleven variables used to cluster counties as well as the total population variable, which is not used in clustering but, as discussed below, is important in identifying potential outliers. 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! As discussed later on, outliers can pose major problems for certain types of clustering analysis, particularly *k*-means clustering (Jin and Han, 2010).

Variable Description	Data Source	Year(s)	Number of Observations	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
Total Population	ACS	2014-2018	3,134	75	10,990	25,789	103,008	67,515	10,098,052	330,294
Percent of Population Aged 25+ with at Least Some College Education	ACS	2014-2018	3,134	15.18	49.86	57.96	57.90	66.48	100.00	11.82
Percent of Population Under Age 65 without Health Insurance	ACS	2014-2018	3,134	1.95	7.52	11.09	12.19	15.43	49.45	6.01
Percent of Adult Population Currently Married	ACS	2014-2018	3,134	21.52	49.69	54.05	53.43	57.67	78.80	6.58
Median Age	ACS	2014-2018	3,134	21.70	38.00	41.20	41.30	44.40	67.00	5.40
Percent of Population Female	ACS	2014-2018)	3,134	21.00	49.41	50.38	49.92	51.11	58.61	2.38
Percent of Population Non-Hispanic White	ACS	2014-2018	3,134	0.73	64.87	84.00	76.58	92.70	100.00	20.07
Percent of Population Non-Hispanic Black	ACS	2014-2018	3,134	0.00	0.64	2.17	8.94	9.87	87.41	14.48
Percent of Population Non-Hispanic American Indian and Alaskan Native	ACS	2014-2018	3,134	0.00	0.13	0.28	1.74	0.67	82.48	7.16
Percent of Population that is Hispanic	ACS	2014-2018	3,134	0.00	2.11	4.10	9.25	9.59	99.07	13.76
Percent of Population Aged 16+ Unemployed but Seeking Work	ACS	2014-2018	3,134	0.00	2.38	3.15	3.25	3.92	16.47	1.43
Percent of Population in Rural Areas	Decennial Census	2010	3,134	0	33.22	59.48	58.58	87.79	100.00	31.48

Table 1: Summary Statistics of 11 Variables Used to Cluster Counties and Total Population

Description of Health Outcome Variables

Table 2, below, shows summary statistics for our three health outcome variables: the percent of the population that is obese, the percent of the population that smokes, and the number of motor vehicle crash deaths per 100,000 residents. We can see that the obesity and smoking outcomes appear to be roughly normally distributed, while the driving deaths outcome appears to be fairly right-tail skewed, and its maximum is over 25 standard deviations above the mean.

Variable Description	Data Source	Year(s)	Number of Observations	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
Annual Number of Driving Deaths Per 100,000 Residents	County Health Rankings	2010-2014	3,134	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	3,134	10.70	28.50	31.20	30.94	33.70	46.60	4.46
Percent of Adult Population that Smokes	County Health Rankings	2014	3,134	6.90	15.70	17.80	18.38	20.70	40.00	3.75

Table 2: Summary Statistics of 3 Health Outcome Variables

Additionally, Figure 1 below shows a pairwise scatterplot of our three health outcome variables. We can see that obesity and smoking appear to be highly correlated, with a correlation coefficient of 0.61. On the other hand, driving deaths are very weakly correlated with either of the other two variables.

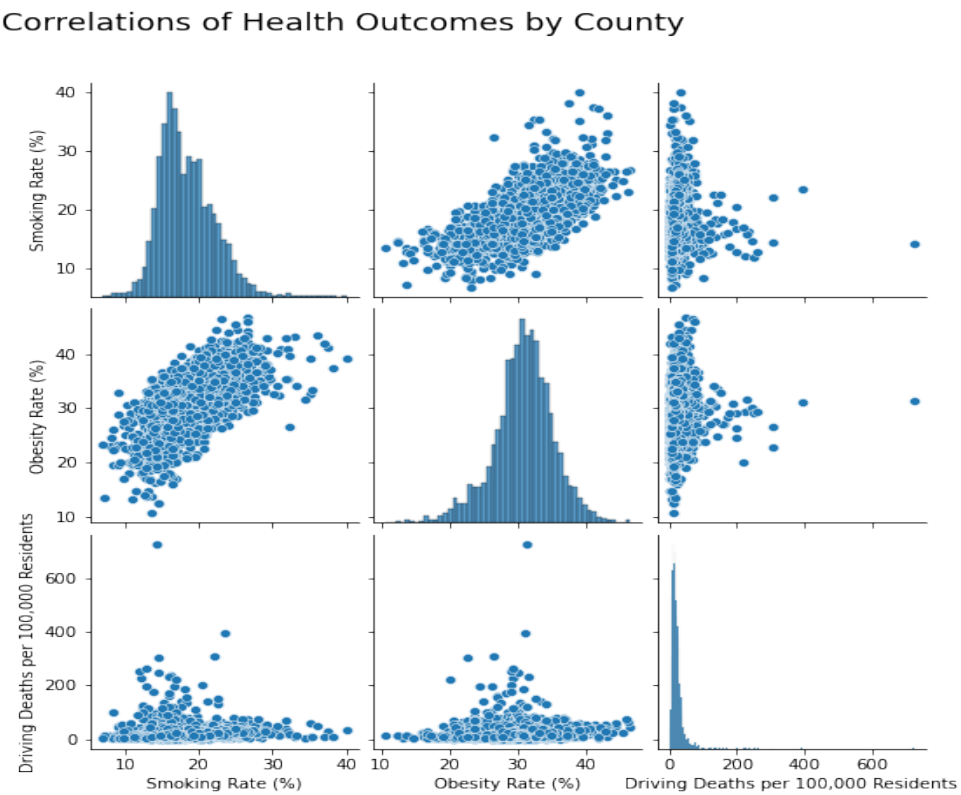
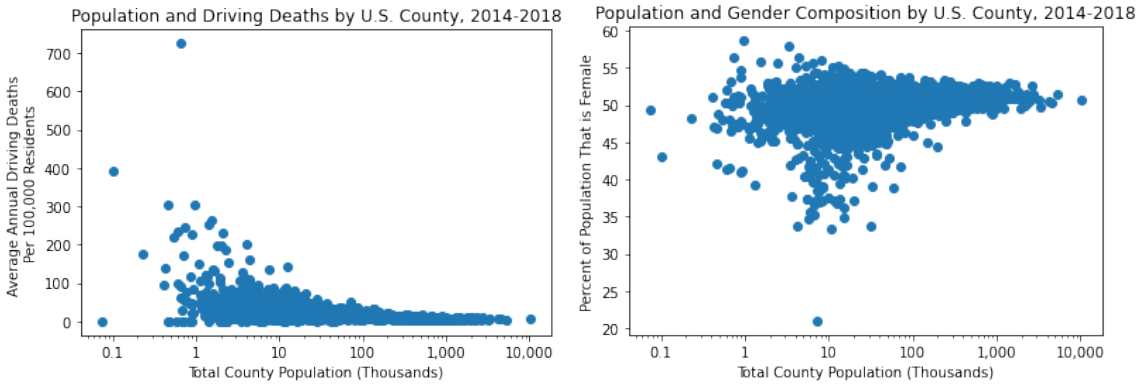


Figure 1: Correlations of County Health Outcomes

Identifying Outliers

As previously mentioned, some variables used in our analysis appear to have extreme outliers. In practice, virtually all major outliers are counties that have very small populations. This fact is illustrated in Figure 2 below. Figure 2(a) shows that all of the counties with annual driving deaths above 200 per 100,000 residents have a population of under three thousand residents, meaning even a single driving death can greatly skew the numbers. Likewise, Figure 2(b) shows that most of the counties that are under 40 percent female have fewer than 10,000 residents.



(a) (b)

Figure 2: County Population, Driving Deaths, and Gender Composition

In practice, removing counties with fewer than 10,000 residents eliminates virtually all of the most extreme outliers in the data. This threshold is somewhat arbitrary, but removing counties with a population below this threshold should make *k*-means clustering much less sensitive to very small counties with extreme values. Thus, in the analysis below we try removing these counties to see how our this affects our results from this particular clustering method.

Limitations of Data

Several limitations of the data exist. The first is the existence of outliers, as discussed above. A second major limitation is the fact that our data is not all from the same year. Particularly, ten of the 11 clustering variables are from the 2014-2018 ACS, but the rurality variable could only be obtained from 2010. Additionally, our health outcomes are only available over the period 2010 through 2014. While most of these variables likely change only very slowly over time, it is still possible that using a consistent data source for all years could meaningfully change our results.

Third, the 5-year American Community Survey presents 5-year estimates of annual surveys of approximately one percent of the population in each county. In small counties, as few as several hundred or even dozen individuals might be sampled over a 5-year period, which means that such counties tend to have very high margins of error in their estimates, "at times exceeding the estimate" itself ("Margins of error in the ACS," n.d.). This substantial measurement error could also meaningfully affect our results, and provides another justification for exploring how our analysis changes when we exclude very small counties.

Methodology

In this analysis, we use three clustering techniques: *k*-means clustering, agglomerative hierarchical clustering, and density-based clustering, and we use several validation methods to compare the results of these techniques and, hopefully, choose a "best" method. In this section, we describe in detail the intuition behind each clustering method, examples of the method being used in prior research, and the pros and cons of each method. Additionally, we describe the metrics we use to evaluate each clustering method.

First, however, we need to precisely define clustering. Clustering is an unsupervised learning technique, meaning that "we lack a response variable that can supervise our analysis" and instead "seek to understand the relationships between the variables or between the observations" (James, Witten, Hastie, and Tibshirani, 2017, p. 26-27). In other words, we are not trying to predict any particular outcome but instead, wish to find groups of observations that are relatively similar in regard to a set of variables. We may be interested in finding clusters because we believe that observations in each cluster are relatively similar in terms of other variables. In this report, for instance, we believe that counties within a given cluster will tend to be relatively similar in terms of our health outcomes of interest.

k-Means Clustering

Intuition

In *k*-means clustering, the number of clusters, *k*, is chosen before conducting clustering analysis. Given *k*, this clustering technique then seeks to assign each data point to a cluster in order to minimize the total squared Euclidean distance from each centroid to the midpoint of its cluster. Given two data points in *n*-dimensional space, we can define the Euclidean distance as the square root of the sum of squared differences of each dimension. For example, given two data points (x_1, x_2) and (y_1, y_2), the Euclidean distance is defined as $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. Then *k*-means clustering attempts to minimize the sum of squared Euclidean distances, given by $\sum_{i=1}^{k} \sum_{x \in C_i} \sum_{n=1}^{N} \left(c_{i,n} - x_n\right)^2$. Here, *k* denotes the number of clusters, *n* the number of dimensions (or variables), C_i cluster *i*, c_i the centroid of cluster *i*, $c_{i,n}$ the *n*th element of the centroid, and x_n the *n*th element of data point *x* (Tan, Steinbach, Karpatne, and Kumar, 2019). Here we define the centroid by assigning each element as the average of all the corresponding element of all data points in the cluster.

Given *k*, we can use an algorithm to determine the optimal assignment of each data point into clusters. First, we randomly pick *k* observations as initial centroid locations. Next, we assign each data point to the cluster whose centroid has the smallest Euclidean distance from the given point. We then re-compute centroid locations by taking the average of the data points in the cluster, and then re-assign data points. We repeat this process until none of the assignments of data points to clusters change (Géron, 2019).

Choosing the Number of Clusters

There are two common methods of choosing the optimal number of clusters, *k*, for clustering. The first is to apply the clustering for a range of values of *k* (say, for *k* = 1, 2, …, 9, 10) and then examine a plot of the total sum of squared errors (SSE, defined above) for each value of *k*. In general, this sum will always decrease as *k* increases, because with more clusters, data points will be closer to the centroids. In practice, we will see that for low values of *k* the SSE sharply decreases while for high values it very slowly decreases. We generally choose a point at which there is a distinct "elbow" in the plot, meaning that SSE goes from sharply to slowly decreasing at that point (Géron, 2019).

The second method is to plot the average silhouette score across observations. For each observation *i*, we define the silhouette coefficient as $s_i = \frac{b_i - a_i}{max(a_i, b_i)}$, where a_i is the average distance from to all other points in the cluster and b_i is the average distance to all the objects in another cluster that has the minimum average distance among all other clusters (Tan et al., 2019). The basic intuition behind the silhouette score, which ranges from -1 to 1, is that we ideally desire clusters that are cohesive, meaning points in a cluster are close to each other, and well-separated, meaning that points in a cluster tend to be far from points in other clusters. Thus, a_i acts as a measure of cohesiveness and b_i acts as a measure of separation, and higher values of the coefficient indicate more well-separated and distinct clusters. We can simply take an average value of the silhouette scores for all observations and plot this as a function of the number of clusters, *k*. In such a plot we want to choose a *k* corresponding to a "knee," meaning that the average silhouette scores increases, reaches a maximum, and then decreases.

Advantages and Limitations

k-means clustering has several advantages. First, it is a relatively simple clustering technique and the algorithm described above converges very quickly on most data sets ("k-Means Advantages and Disadvantages," 2020). Additionally, it is a partitional technique, meaning every data point is assigned to one and only one cluster (Tan et al., 2019).

However, several disadvantages exist. The first is that the number of clusters must be chosen before applying the clustering. Thus, "after-the-fact" inspection must be used to determine the optimal number of clusters, as described above. Second, the clusters obtained can change depending on the random initialization of the cluster centroids. In other words, it is not guaranteed that the sum of squared errors obtained is the global minimum of this function. Third, the clusters obtained can be sensitive to any outliers. This is because outliers will tend to have a very large Euclidean distance from cluster centroids, so the minimization of the sum of squared errors will tend to "pull" the centroids toward any outliers (Jin and Han, 2010). Fourth, *k*-means clustering generally produces globular clusters with relatively similar shapes and densities because of the Euclidean distance metric (Tan et al., 2019). If the underlying structure of the clusters does not satisfy these properties, then *k*-means clustering will produce relatively poor clusters. Overall, then, *k*-means clustering should work best on datasets with few or no outliers and globular cluster shapes of relatively equal size and density.

Example Uses

k-means clustering is an extremely common form of clustering analysis and has been used by many researchers for varied topics. For instance, Helbich, Brunauer, Hagenauer, and Leitner (2013) use this approach to construct housing "submarkets" within regions by grouping neighborhoods based on housing and geographic characteristics. In a very different vein, Ramachandran, Shah, and Moss (2018) use this technique to cluster firms in sub-Saharan Africa based on their characteristics in order to look at the relationship between firm growth and availability of electric power. Finally, it should be noted that *k*-means clustering is frequently used by major retailers to group customers based on demographic and behavioral characteristics. Advertisements and coupons can then be targeted to different groups of customers (Sagar, 2019).

Agglomerative Hierarchical Clustering (Complete Link)

Intuition

The second clustering method we consider is agglomerative hierarchical clustering (AHC). AHC is a "bottom-up" method that starts by treating every observation as a single cluster. It then computes the proximity between every possible pair of clusters and then merges the closest clusters. It repeats this process until there is only one cluster remaining (Tan et al., 2019).

Several different methods can be chosen to calculate the proximity between clusters. In this analysis, we use the complete link clustering method, which calculates the proximity between clusters as the maximum distance between the two points in the clusters that are farthest away from each other (Tan et al., 2019). We use this method because it is the most resistant to outliers and we have established the presence of extreme outliers in our data. In particular, this method is resistant to outliers because it calculates the distance between the two most distant points, rather than the two closest; this prevents two clusters from being merged because of an outlier that is far away from points in its own cluster, but close to points in another cluster.

Choosing the Number of Clusters

Note that AHC does not require the number of clusters to be pre-specified; instead, it creates a natural hierarchy of clusters. Thus, to choose the number of clusters, the user can examine the order in which the clusters are merged and choose a certain point in the process at which to cut off. This order can be visualized in a dendrogram, which is a tree-like structure showing the order in which each cluster is merged. Figure 3 below shows a set of sample data points and a sample dendrogram provided by Bock (2019). On the x-axis is each data point; the y-axis represents the distance between clusters. We read the dendrogram from the bottom up. We can see that points E and F initially had the closest proximity and were merged first, followed by points A and B. We can draw any horizontal line through the dendrogram,

and the number of vertical lines we cross equals the number of clusters. A commonly used rule is to find the largest vertical distance for which no horizontal line crosses the dendrogram, and cut the dendrogram there (Bock, 2019). For instance, in Figure 3 we can see that there is a very large vertical distance between the topmost horizontal line and the next horizontal line. Thus, we would cut the dendrogram in that region and obtain two clusters, one containing points A and B and the other containing points C, D, E, and F.

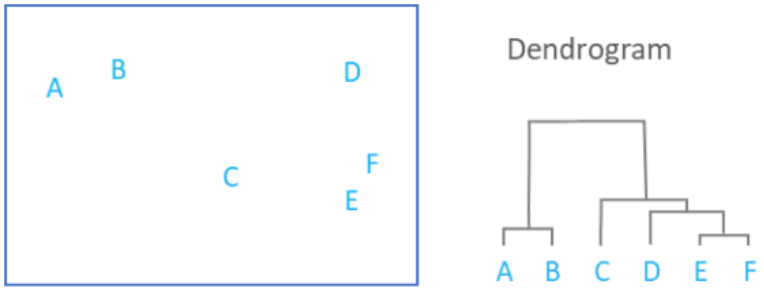


Figure 3: Sample Dendrogram

Advantages and Limitations

AHC with the complete link method has several advantages. First, it is generally robust to the inclusion of outliers, much more so than *k*-means clustering. Second, it does not require the number of clusters to be chosen up front, but rather a dendrogram can be visually inspected. Like *k*-means clustering, it also assigns every point to a cluster, but unlike *k*-means, it produces a nested hierarchy of clusters. This may be an advantage if the data has an inherently hierarchical structure (for instance, local governments may belong to counties, which belong to states).

A key disadvantage of AHC with complete link clustering is that there is no global objective function such as the SSE (Verma, 2009). Rather, clusters are merged one step at a time, and once two points are put into the same cluster, they can never be separated. Additionally, the complete link method favors compact and globular clusters and may perform poorly on data with an underlying structure that is not compact and globular (Tan et al., 2019). A final disadvantage of AHC is that the approach is quite computationally expensive, especially for larger datasets, because at every step a square similarity matrix needs to constructed with the number of rows and columns each equal to the number of observations (Patlolla, 2018).

Example Uses

Like *k*-means clustering, hierarchical clustering is often used to perform customer segmentation. Hung, Lien, and Ngoc (2019) perform customer segmentation using hierarchical clustering on credit card data so that different marketing strategies can be targeted toward different groups of customers. As another example, in a blog post Pipis (2020) performs AHC on countries in Europe, using the voting results of countries in the Eurovision 2016 Song Contest. Pipis (2020) finds that countries in each cluster tend to be "close geographically or culturally."

Density-Based Clustering

Intuition

Density-based clustering, technically called "density-based spatial clustering of applications with noise" (DBSCAN), is a clustering method that determines clusters based on some measure of density rather than proximity or distance. In particular, DBSCAN works by finding high-density regions surrounded by lower-density areas (Tan et al., 2019). Particularly, given some $\epsilon > 0$ DBSCAN counts the number of observations within a Euclidean distance of less than ϵ. Given an integer a, a point is labeled as a core point if at least a observations are within a distance of ϵ. A point that is not a core point but is within ϵ of a core point is considered a border point. Then, a point not within ϵ of a core point is

classified as a noise point and is eliminated (Géron, 2019). Importantly, noise points are not assigned to a cluster, a key difference from the two methods previously discussed. Among border and core points, any core points within ϵ of each other are part of the same cluster, while border points within ϵ of a core point are part of that core point's cluster (Tan et al., 2019). Thus, the number of clusters in DBSCAN is not explicitly specified but is rather determined organically by the algorithm.

Advantages and Limitations

A key advantage of DBSCAN is that, as just stated above, the number of clusters does not need to be explicitly stated; rather, the algorithm determines the optimal number of clusters. A disadvantage is that, unlike with the other two clustering methods previously discussed, ϵ and a need to be specified. For simplicity, we do not go into the details of determining these parameter values here, but note that commonly specified values are $\epsilon = 0.5$ and $a = 5$ ("sklearn.cluster.DBSCAN," n.d.).

Additionally, DBSCAN identifies and remove "noise" points; thus, unlike in *k*-means and AHC, not all points are assigned to a cluster under DBSCAN. This can be an advantage because it makes DBSCAN quite robust to outliers, which are dropped by the algorithm; however, it can be a disadvantage if it is important for the researcher to assign all points to a cluster.

Finally, because DBSCAN is a density-based metric, it does well at identifying clusters of different shapes and sizes as long as those clusters consist of higher-density regions separated by lower-density regions. On the other hand, this means that DBSCAN will perform poorly at identifying clusters that have different densities (Tan et al., 2019). These facts can be either an advantage or disadvantage depending on the underlying structure of the data. Finally, Tan et al. (2019) note that DBSCAN "has trouble with high-dimensional data because density is more difficult to define for such data" (p. 569). Importantly, our data has 11 dimensions, which is an indication that DBSCAN may struggle on our data.

Example Uses

While DBSCAN tends to be less commonly used than AHC or *k*-means clustering, it still has been used in several important applications, and particularly in situations where outlier detection is important. For example, when Netflix experiences server problems, the company uses DBSCAN to identify servers in its farm that are outliers in terms of their performance, helping them identify the source of the problem ("Tracking Down the Villains: Outlier Detection at Netflix," 2015). Somewhat similarly, Yang, Lian, Li, Chen, and Li (2014) use a DBSCAN algorithm to mark suspicious financial transactions, which can be used by financial regulators to identify money laundering.

Cluster Validation

Because clustering is an unsupervised technique (James et al., 2017), there is no predicted outcome variable for which we can compare our predictions to the actual values. This makes the notion of cluster validation quite difficult and ill-defined. Nevertheless, cluster validation is important because "almost every clustering algorithm will find clusters in a data set, even if that data set has no natural cluster structure" (Tan et al., 2019, p. 571).

It is important to note that cluster validation is a rather subjective process and there are several different metrics by which the results of different clustering algorithm can be compared. For simplicity, we only discuss the three validity metrics that we consider in this report. These metrics are described below.

Intuition

The simplest but most subjective method for cluster validity is to simply examine the clusters and check that they make intuitive sense. In this case, this means that we can examine the counties in each cluster and make sure that they are in fact demographically similar. One simple way to do so is to map the resulting clusters and visually examine the results. We might reasonably expect that most clusters will be geographically contiguous (e.g., one county might consist of rural southern counties, while another might consist of counties along the border with Mexico), and we should make sure that there are no major "red flags." For instance, if we obtained one cluster which consisted only of the county containing Los Angeles, one county in rural Utah, and one county in rural Alabama, we would probably deduce that this cluster is very poorly estimated. Because of its inherent subjectivity, we do not use intuition to determine

precisely how well a clustering method works, but we can use it to discard clustering techniques that clearly perform very poorly and produce nonsensical results.

Silhouette Scores

The second method for cluster validation is to use average silhouette scores, which were defined and discussed as a method for choosing *k* in *k*-means clustering. Average silhouette scores can also be used to evaluate the results from AHC and DBSCAN (Tan et al., 2019). To recap, silhouette scores are defined for each observation to give an idea of both cohesiveness and separation, and we take the average of the silhouette score across all observations. Higher values indicate clusters that are more well-separated and cohesive.

Predictive Power for Health Outcomes

As previously discussed, Wallace et al. (2019) specifically cluster counties using the set of variables that they found to be most strongly predictive of individuals' health outcomes. Thus, it is reasonable to expect that the clusters determined by our algorithm should be fairly highly predictive of these health outcomes. In other words, the variance of these outcomes within each cluster should be much lower than the variance across all counties, so that knowing a county's cluster allows us to make a reasonably accurate "guess" about its outcomes. Thus, as our third method of cluster validation we see how predictive these measures are of each health outcome.

We do this as follows. For a given health outcome (the smoking rate, obesity rate, or motor vehicle death crash rate), if we have *k* clusters from a given clustering technique, then we can run a linear regression of the health outcome on a set of *k*-1 dummy variables, where dummy variable *j* equals 1 for a county in cluster *i* and 0 otherwise. More formally, we estimate the following regression where *i* indexes each county (observation):

$$HealthOutcome_i = \alpha + \sum_{j=1}^{k-1} \beta_{j,i} + \epsilon_i.$$

Here, α is a constant that corresponds to the mean value of the outcome for the cluster that is excluded from the set of dummy variables, and ϵ_i is an error term. For any other cluster *j*, the mean value of the outcome is $\alpha + \beta_j$. To assess how well the clusters predict these health outcomes, we can look at the R-squared value of the resulting regression, which ranges from 0 to 1 and indicates the fraction of variance in the dependent variable of interest that is explained by the model (James et al., 2017). Values close to 1 indicate a better fit to the data. Thus, for each set of clusters we obtain three R-squared values, one for each health outcome.

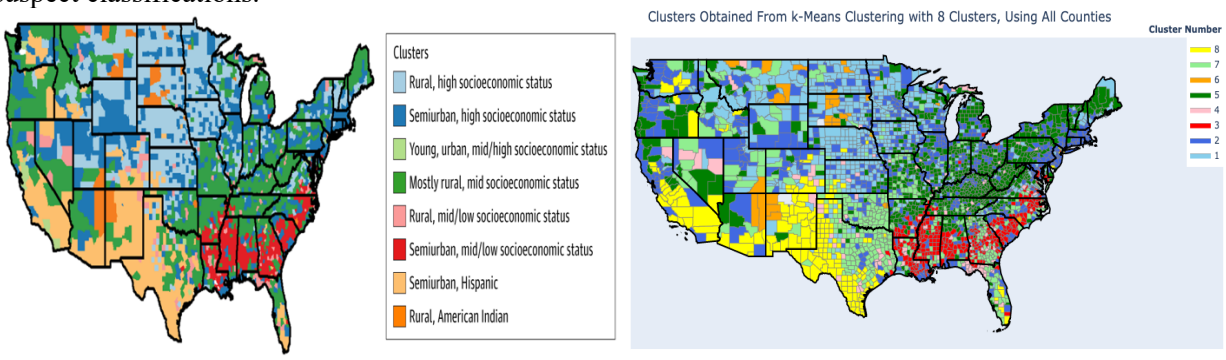
Findings

In this section, we work through the results of five clustering methods. First, we replicate exactly what Wallace et al. (2019) do by performing *k*-means clustering with eight clusters to see how our results compare with their. Second, since the authors provide very limited information on why they use $k = 8$ clusters, we use the methods previously described to determine the optimal number of clusters in *k*-means clustering. We obtain $k=5$ and discuss the results obtained from this. Third, we try removing counties with fewer than 10,000 residents and perform *k*-means clustering again, and again re-determine the optimal number of clusters. Fourth, we perform AHC with the complete link method on the complete dataset. Fifth, we perform DBSCAN on the complete dataset. Because AHC and DBSCAN are generally resistant to outliers, for simplicity we do not try removing outliers for these algorithms. Finally, we apply the cluster validation methods previously discussed to each of our five sets of clusters and compare them.

Exactly Replicating Wallace et al. (2019)

First, we simply use *k*-means clustering with eight clusters and without dropping any potential outliers, as done by Wallace et al. (2019). We use a random initialization for this technique, but the results tended to be extremely similar regardless of the seed used for initialization. Unfortunately, Wallace et al. (2019) provide neither replication code nor any data containing the cluster assignments for each county,

making it impossible to determine the exact extent of the overlap of our clusters with theirs. However, the authors provide a map of their eight clusters and a table containing the total population and number of counties in each cluster, which we reproduce for comparison. Figure 4(a) maps the eight clusters obtained by Wallace et al. (2019), while Figure 4(b) maps the eight clusters obtained in our analysis. First, we can see that both sets of clusters easily pass the "sniff test" based on the variables used for clustering. For instance, in both cases one cluster contains primarily counties in the rural South that have large African-American populations and relatively low incomes; another cluster consists of heavily Hispanic counties in Florida and states along the border with Mexico; a third cluster consists mostly of rural counties in the upper Midwest and interior West. Overall, these clusters make intuitive sense without having any clearly suspect classifications.



(a) (b)

Figure 4: Clusters Produced by Wallace et al., 2019 (a) and Our Analysis (b) in *k*-Means Clustering, *k*=8

Second, we can see that the clusters produced in our analysis and in Wallace et al. (2019) appear to be extremely similar. The vast majority of counties appear to be in the "same" cluster in both analyses. To quantify these results, Table 3 shows the distribution of counties and total population across clusters in Wallace et al. (2019) and our analysis, where we put clusters in descending order based on their number of counties. In all cases, these clusters have the same ordering; that is, the cluster that contains the most counties in Wallace et al. (2019) overlaps primarily with the cluster containing the most number of counties in our analysis, and so on.

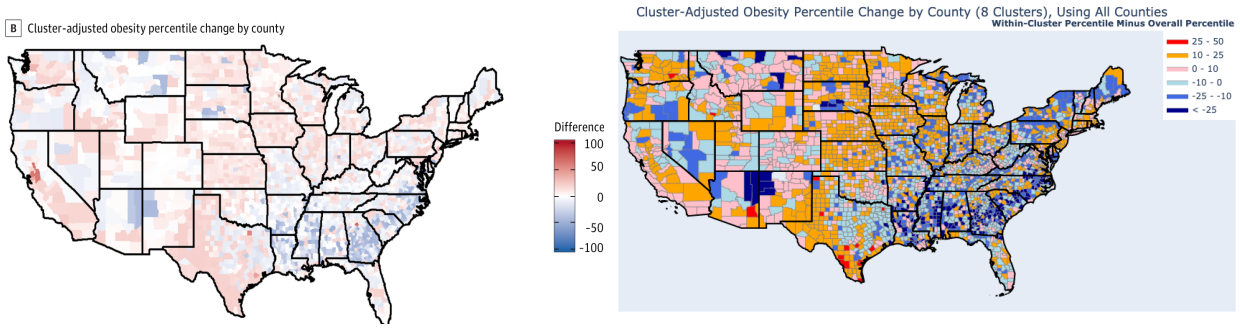
Cluster Rank in Number of Counties	Number of Counties in Wallace et al. (2019)	Number of Counties in Our Analysis	Total Population in Wallace et al. (2019), in Millions	Total Population in Our Analysis, in Millions
1	973	781	33.78	27.56
2	727	717	152.20	177.25
3	674	574	12.73	10.55
4	326	427	26.43	15.59
5	244	304	46.56	34.37
6	116	182	1.90	54.97
7	42	111	0.85	1.69
8	37	38	44.41	0.83

Table 3: Comparison of Cluster Sizes and Counties in Wallace et al. (2019) and Our Analysis

There are a few notable differences in our results. First, Wallace et al. (2019) have nearly 200 more counties in their most frequent cluster than we do. Additionally, Wallace et al. have a cluster of only 37 counties that appear to have very large average populations, which appears to be generally "missing" from our analysis. On the whole, however, these results appear very similar. As discussed previously, the observed differences in our results may result in part from the random initialization we used in *k*-means clustering and the fact that our data sources differ by several years.

As an additional robustness check, Wallace et al. (2019) assign counties percentile scores nationwide on each of the three health outcomes as well as percentile scores within their own cluster. They then map the differences in the two scores across counties, subtracting the nationwide percentile

from the within-cluster percentile. We also map these differences, and Figure 5 maps the results using the obesity outcome metric. Regardless of the health outcome used, Wallace et al. (2019) once again obtain extremely results to ours. In both sets of results, the highest obesity rates tend to occur in the rural South, but because these counties tend to be grouped within the same cluster their within-cluster percentiles are much lower. The fact that many counties have a substantial change in percentile scores underscores the need to evaluate counties on health outcomes in light of their demographic characteristics, and to compare counties not nationwide but to assess them relative to their peers.



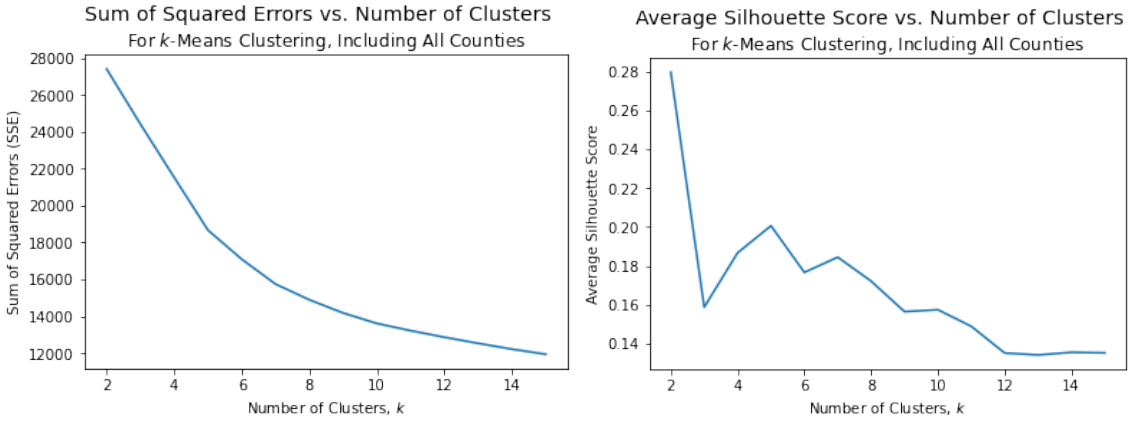
(a) (b)

Figure 5: Obesity Percentile Changed Produced by Wallace et al., 2019 (a) and Our Analysis (b) in *k*-Means Clustering, k=8

Taking $k = 8$ as given, then, we have shown that the results of Wallace et al. (2019) are fundamentally reproducible. However, the authors provide very little information on how they obtained this value of k, other than stating that "Gap statistics, Bayesian information criterion, and within-group sum of squares were used to select the optimal number of clusters" (Wallace et al., 2019). Next, we use the SSE and silhouette score metrics discussed above to find an optimal value of k and see if it compares to their results.

Determining k in k-Means Clustering with All Counties

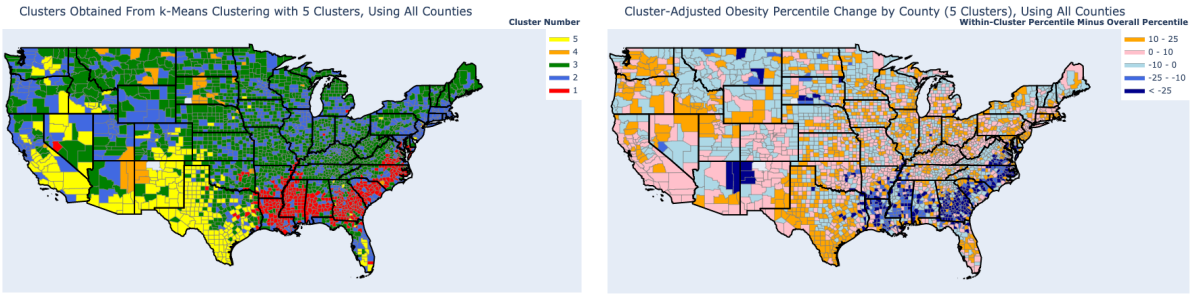
We determine k using the methodology previously. In this case, we loop over values of k from 2 to 15 and for each k calculate the sum of squared errors (SSE) and the average silhouette score across all observations. Figure 6(a) shows the SSE obtained for each value of k while Figure 6(b) shows the average silhouette score. Looking at the SSE values, we can see that there appear to be elbows at both $k = 5$ and $k = 7$. Looking at the average silhouette score, we can see that there are local peaks in the average score at both of these values. However, the peak at $k = 5$ is higher than the peak at $k = 7$. Taking both the SSE and silhouette scores into account, a value of $k = 5$ thus appears to be the best choice. Importantly, this is a smaller number of clusters than that determined by Wallace et al. (2019), so we were unable to recreate this aspect of their analysis. However, it could also be reasonably argued that $k = 7$ rather than 5 clusters exist in the data, which is obviously close to 8. However, Figure 6 shows neither an elbow in the SSE nor a local maximum in the silhouette scores for $k = 8$, so there is virtually no justification for us to choose 8 clusters.



(a) (b)

Figure 6: SSE (a) and Average Silhouette Score (b) for *k* Using All Counties

Next, we need to see whether using 5 clusters instead of 8 appears to substantially change our results. Figure 7(a) maps the five clusters produced, while Figure 7(b) maps the corresponding difference between each county's within-cluster obesity percentile and its overall percentile. Comparing Figure 7(a) to Figure 4, we can see that the overall geographic patterns look fairly similar even using 3 fewer clusters. Essentially it seems to be the case that several clusters approximately "merged" when going from 8 to 5 clusters, such as the formerly two distinct clusters prevalent in the Great Plains and Great Lakes states. Looking at Figure 7(b), we can see that the same overall patterns hold. Importantly, there are still many counties that either perform much better or much worse on obesity within their cluster than nationwide, which indicates that these clusters can be useful for policymakers by identifying counties that perform well or poorly relative to counties with similar demographics. While we were unable to replicate the determination of Wallace et al. (2019) that 8 clusters was optimal, Figure 7 makes it clear that using 5 clusters does not produce strikingly different results.



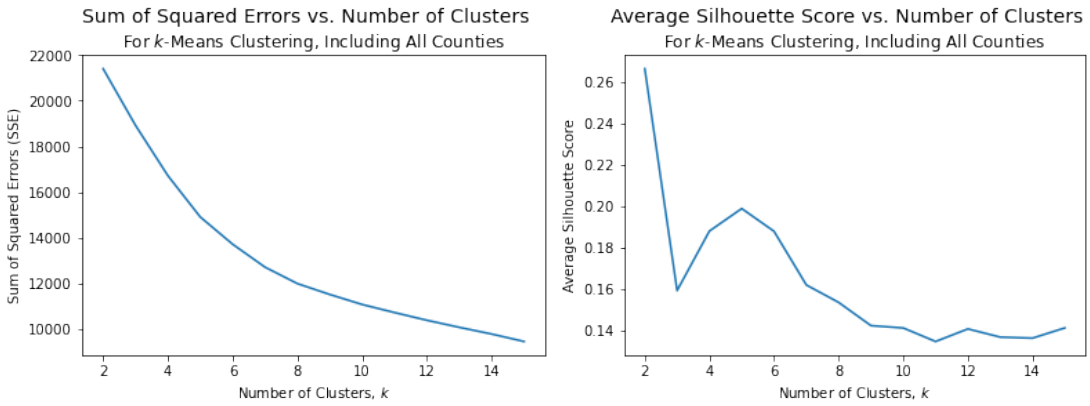
(a) (b)

Figure 7: Clusters Produced in *k*-Means, *k*=5 (a) and Corresponding Changes in Obesity Percentiles (b), Using All Counties

Determining k in k-Means Clustering while Removing Small Counties

In our exploration of the data, we noticed the existence of several extreme outliers in the data and noticed that these outliers were overwhelmingly counties with very small populations. This poses potentially severe problems for the authors' results given that the presence of outliers can drastically affect the clusters obtained from *k*-means. Here we remove counties with fewer 10,000 residents to see whether the removal of outliers substantially affects the results. This causes us to remove 703 out of 3,134 counties or approximately 22% of all observations. We first determine the optimal value of *k* using the same method that we previously used. Figure 8(a) shows the resulting SSEs while Figure 8(b) shows the

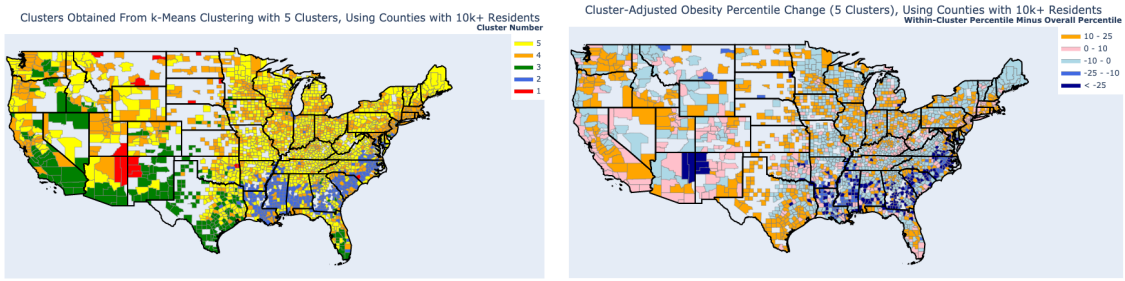
average silhouette scores. Again, the SSEs reveal possible elbows at 5 and 7; the silhouette score no longer has a local peak at 7 but still has a prominent peak at 5. Thus, we select 5 clusters.



(a) (b)

Figure 8: SSE (a) and Average Silhouette Score (b) for *k* Removing Small Counties

Figure 9(a) maps the five clusters produced when we consider only counties with at least 10,000 residents, while Figure 9(b) maps the corresponding percentile differences on obesity. First, we can see that the counties dropped are overwhelmingly located west of the Mississippi River. Comparing Figure 9 to Figure 8, we can see that the results are strikingly similar. The geographic distribution of each cluster appears to be almost the same and conditional on a county having at least 10,000 residents, it appears to almost always be placed in the "same" cluster regardless of whether small counties are excluded. Likewise, the changes in percentiles are extremely similar, and indicate that for purposes of comparing counties within peer groups, meaningful results are produced.



(a) (b)

Figure 9: Clusters Produced in *k*-Means, *k*=5 (a) and Corresponding Changes in Obesity Percentiles (b), Using Counties with at Least 10,000 Residents

Overall, then, while we documented the existence of outliers in our data, we have seen that removing these outliers has no impact on the optimal number of clusters obtained and very little impact on the patterns of such clusters. This provides support for the decision of Wallace et al. (2019) to include all counties which have non-missing data in their cluster analysis.

Results of Hierarchical Complete Link Clustering

Next, we turn to agglomerative hierarchical clustering (AHC) using the complete link method. To determine the optimal number of clusters, we first use a dendrogram, shown in Figure 10 below. Note that in both AHC and DBSCAN we are again using all 3,134 counties, because both of these methods are relatively robust to outliers. Because there are 3,134 observations, at the first level we have 3,134 clusters, which is impossible to clearly visualize; thus, we have cut Figure 10 so that it only show the topmost 15

merged clusters; the numbers in parentheses indicate the number of counties in each of these 15 merged clusters. We can clearly see that the largest vertical distance through which no horizontal line crosses corresponds to the region with three clusters; thus, we cut the dendrogram here and produce three clusters. The leftmost cluster has 19+8+16 = 43 counties, while correspondingly the middle cluster has just 12 counties and the remaining 3,079 counties are in the third cluster.

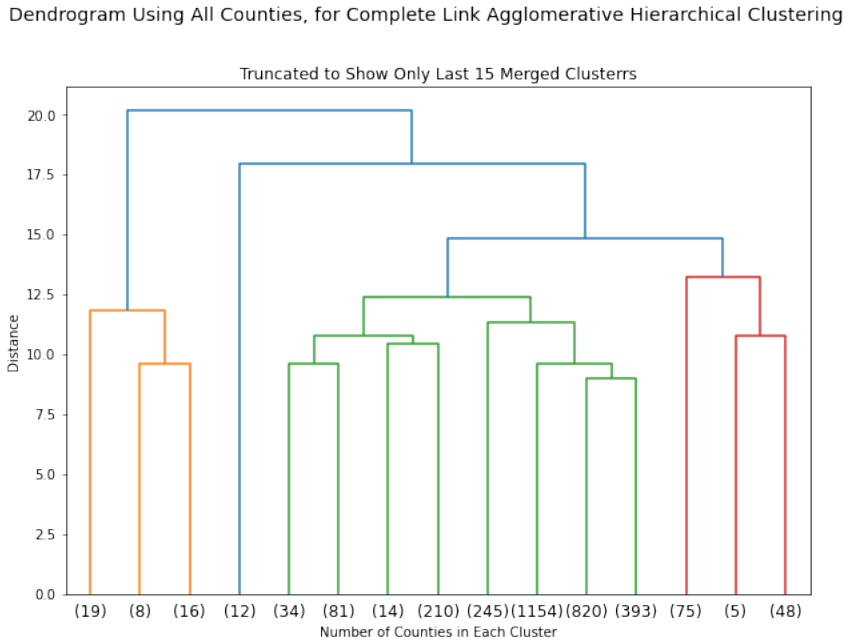


Figure 10: Dendrogram with AHC Complete Link, Truncated to Show Topmost 15 Clusters

These clusters have extremely different sizes, which likely poses problems for our analysis. Below, Figure 11 maps these clusters. Because the overwhelming majority of counties are in a single cluster, this immediately fails the "sniff test"; New York City is in the same cluster as Salt Lake City, most of rural Minnesota, and most of rural Alabama, even though the counties in these regions are generally very dissimilar. Furthermore, because almost all counties are in the same peer group, the nationwide percentiles and the within-cluster percentiles will be almost exactly the same for virtually every county. Unfortunately, the within-cluster percentiles will thus no longer be useful for assessing counties' performance on health outcomes conditional on their demographic characteristics. This clustering method has clearly failed to produce useful results.

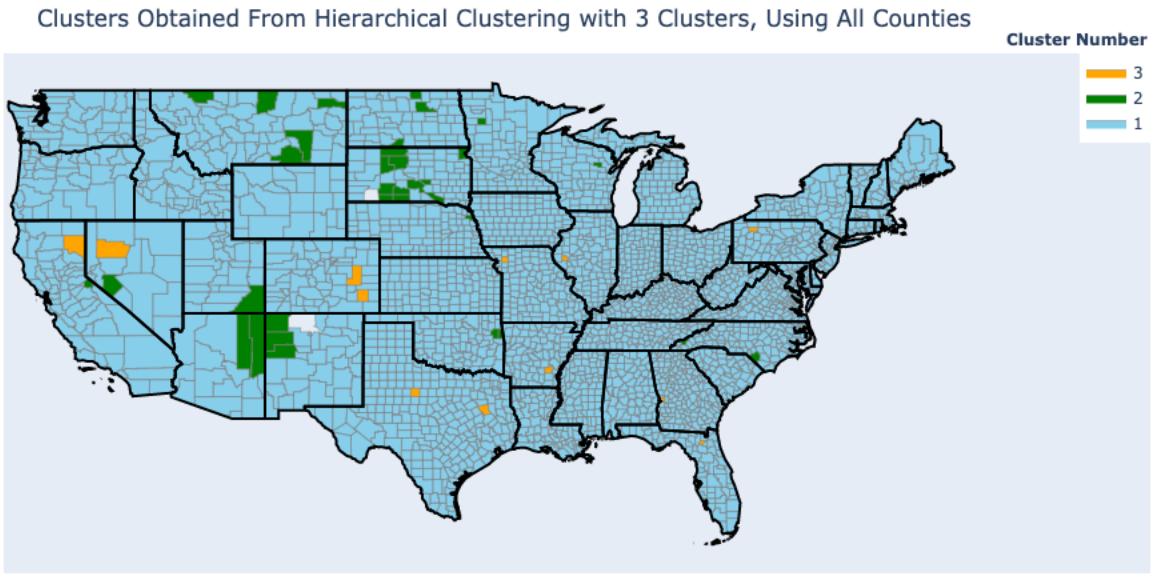


Figure 11: Map of Clusters Produced by AHC Complete Link, Using 3 Clusters

Results of DBSCAN

Fifth, we turn to the DBSCAN clustering method. As previously discussed, we do not need to determine the number of clusters but rather, this is done automatically by the algorithm and any potential outliers are removed as noise. Figure 12 maps the clusters produced by density-based scanning. Out of 3,134 counties, 2,135 or roughly 68% were removed as noise; we can see that several states had every county removed as noise. Of the 999 counties identified as core or border points, two clusters were identified and 981 were put in one cluster while the second cluster contains only 18. These results are robust to a wide range of values of ϵ and a, and the results have the same issues as with AHC; they clearly fail the "sniff test" and are not useful for purposes of identifying and comparing similar counties. Additionally, over two-thirds of counties were dropped by our algorithm; if we want to assign counties to peer groups to help local health agencies, it is clearly not helpful to tell most counties that we were unable to identify any peer counties. Like AHC, DBSCAN has failed to produce meaningful clusters for purposes of our analysis.

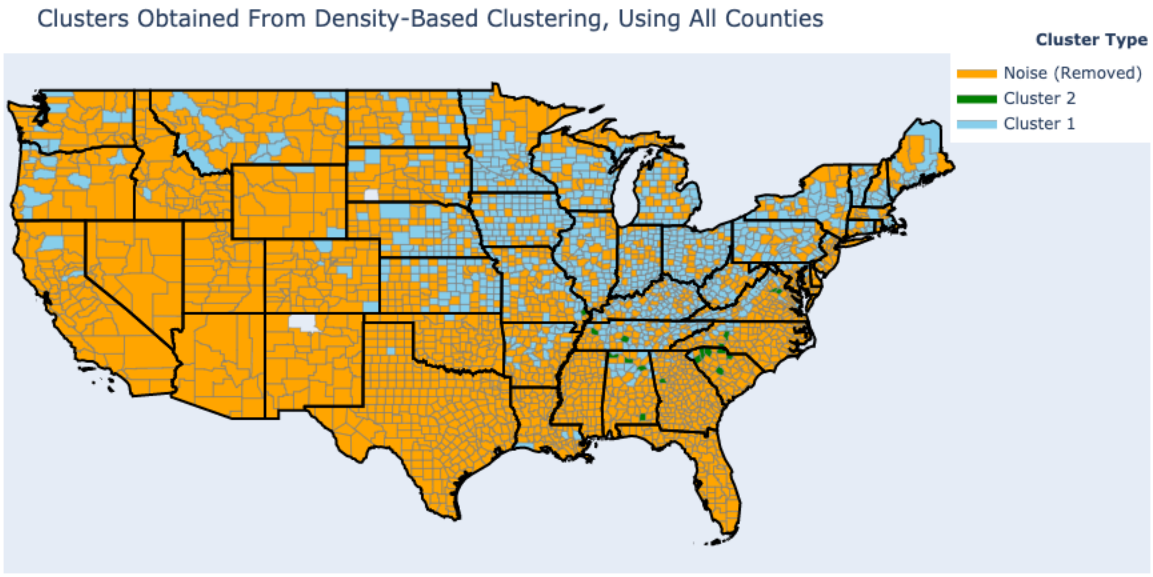


Figure 12: Map of Clusters Produced by DBSCAN

Cluster Validation

Finally, we validate the results from our five models using intuition, silhouette scores, and the predictive power of each set of clusters on our health outcomes. Intuition has already been extensively discussed. All three of our *k*-means clustering approaches appeared to pass the "sniff test," while AHC and DBSCAN clearly failed. Next, we turn to average silhouette scores and predictive power for each of our five sets of clusters. Table 4 shows for each set the number of observations placed into clusters, the average silhouette score, and three R^2 values, each corresponding to the value from a linear regression of the given health outcome on a set of dummy variables corresponding to the clusters, plus a constant.

Clustering Technique	Number of Observations Placed into Clusters	Average Silhouette Score	R^2 with Obesity Prevalence as Dependent Variable	R^2 with Smoking Prevalence as Dependent Variable	R^2 with Motor Vehicle Deaths as Dependent Variable
k-means, *k*=8, all counties	3,134	0.172	0.241	0.331	0.081
k-means, *k*=5, all counties	3,134	0.201	0.203	0.209	0.078
k-means, *k*=5, removing small counties	2,431	0.199	0.223	0.196	0.210
AHC with 3 clusters	3,134	0.478	0.007	0.09	0.003
DBSCAN with 2 clusters (excluding noise)	999	0.097	0.000	0.002	0.003

Table 4: Validation Statistics for Clustering Techniques

We can see that DBSCAN performs the worst on all metrics, with an average silhouette score of just 0.097 and virtually zero predictive power for our health outcome variables. Of the remaining four sets of cluster, the AHC clusters have by far the lowest R^2 values but also by far the highest average silhouette score. However, because AHC placed almost all counties into one cluster and clearly failed the "sniff test," we should disregard this model. Among the three sets of *k*-means clusters, no single set of clusters dominates on our performance metrics. Since removing outliers does not seem to drastically improve our performance, we should disregard that *k*-means clustering; for the purpose of giving local health departments and policymakers a group of peer counties, it is not helpful to say that nearly one-quarter of counties can't be assigned to a group of peers when their inclusion wouldn't substantially change the

validity of the peer groupings. This leaves us with *k*-means clustering with all counties and using k=8 or k=5. Using 5 clusters improves the average silhouette score by roughly fifteen percent but using 8 clusters results in higher predictive power for all three health outcomes; the predictive power is over 50 percent higher for the smoking variable when using 8 clusters. Basic intuition does not have much to say to break this tie; overall, we can say that these measures both appear to perform roughly equally well, and far better than other alternatives. It is difficult to confidently say that one number of clusters is "better" to use than the other, but we can say that *k*-means clustering is the appropriate clustering technique to use in this scenario and outliers should not be removed.

Conclusion

Key Insights and Applications

In this report, we attempted to replicate the clustering results of Wallace et al. (2019). This entailed investigating their choice of *k*-means as a clustering technique, their choice of the number of clusters *k*, the sensitivity of their results to any potential outliers or data anomalies, and the reproducibility of their results when using k=8.

Several key findings and insights emerged from this replication. First, we confirmed that two other clustering techniques, agglomerative hierarchical clustering and density-based scanning, perform very poorly, meaning that *k*-means clustering should be used to cluster counties based on the demographic characteristics most predictive of smoking, obesity, and the motor vehicle death rate. From the previous discussion, we know that *k*-means clustering is well-suited to clusters of relatively globular, equal-sized shapes, while DBSCAN is well-suited for finding regions with differing densities (Wallace et al., 2019). The relative performance of these two methods suggests that the underlying clusters in the data we use may have relatively equal densities and sizes and globular shapes.

Second, we found that while outliers exist, they do not seem to substantially affect the clustering results when using *k*-means clustering. Removing outliers is generally desirable if outliers substantially affect the clusters produced by a clustering algorithm, but if not, then in certain situations it may be undesirable to drop outliers. The discussion by Wallace et al. (2019) illustrates that this is a situation where we want to be able to include as many counties in our cluster as possible, in order to be able to provide helpful insights to local health departments.

Third, we differed from the authors in determining the optimal number of clusters to use in this analysis. However, the cluster validity methods we employed could not definitely conclude that either 5 or 8 clusters was a better choice. Using 5 clusters resulted in a higher average silhouette score, indicating that these clusters were more cohesive and well-separated; on the other hand, using 8 clusters meant that the resulting clusters could better predict our three health outcomes of interest.

For the most part, our replication thus validates the approach of Wallace et al. (2019). While our methods of determining *k* produced different results, both 5 and 8 clusters resulted in comparable validation metrics. In practice, it seems there is no clear "best" number of clusters to use; more likely, a range of values of *k* produces clusters that are intuitive, generally pass the "sniff test", and result in similar silhouette coefficients and R^2 values. This finding underscores the fact that choosing *k* is a quite subjective and somewhat arbitrary process, as much of an art as a science (Pham, Dimov, and Nguyen, 2005).

As Wallace et al. (2019), these clustering outcomes can be applied extensively in the domain of public health. Particularly, local health agencies can identify counties in the same cluster and look at counties that have the highest within-cluster outcomes. These agencies can then consult the health policies of high-performing counties to see which policies might most improve health outcomes. Additionally, these rankings can provide a new framework for all levels of government to think about regional disparities health outcomes. A wider recognition that health agencies and public officials are largely constrained by the demographic characteristics of their populations will hopefully cause desirable changes in the allocation of resources and funding to governments (Wallace et al., 2019).

Limitations and Future Work

We have previously discussed limitations of the data, but here we consider limitations of our analysis and avenues for future work. A major limitation of this replication is that it only considered the clustering aspect of the paper by Wallace et al. (2019) and takes as given the set of most important variables that they find in determining the three health outcomes we use. While we generally were able to replicate the results of the clustering in Wallace et al. (2019), we are not able to state that the findings of important determinants in Wallace et al. (2019) are replicable. Future work would attempt to reproduce this part of their analysis as well. Since the authors state that they use an algorithm called a random forest to rank variable importances but do not provide a justification for using this algorithm, this future work could explore other algorithms such as linear regression to see whether the determination of important variables is highly dependent on the choice of algorithm.

Second, our analysis used data that was relatively high-dimensional, with 11 variables used to create county clusters. As more variables are added, the Euclidean distance measures necessarily increase, and methods such as DBSCAN can work relatively poorly with many variables (Tan et al., 2019). This may partly explain why we found *k*-means clustering to perform so much better than either DBSCAN or AHC. Future work could apply dimension reduction techniques such as principal component analysis (PCA) to this data. PCA attempts to create a few composite variables (called principal components) that capture most of the variation in the data, and typically just two or three principal components account for a great deal of variability (Brems, 2017). Thus, future work could attempt to apply PCA to this data so that only 2 or 3 variables are used in clustering; we could then assess whether this dimension reduction substantially improves the performance of AHC and DBSCAN.

Third, because our analysis only identified the demographic variables most relevant to our three health outcomes of interest, these clustering results are not likely to generalize well for researchers interested in other outcomes, either health-related or otherwise. Future work could thus try to reproduce both the selection of important variables and the resulting clusters for a wide variety of health outcomes, to see how much both the clustering results and most important variables tend to change for a variety of disparate health outcomes.

The data collection process and analysis pose no major ethical issues since they all use publicly available data collected by reputed government agencies and nonprofit organizations. However, ethical concerns may arise if policy or funding decisions are made based on these clustering results. Wallace et al. (2019) emphasize that this analysis represents just one of many approaches to grouping similar areas based on their characteristics, and these groupings should not be treated as definitive.

Works Cited

American Communities Project. (n.d.). Retrieved December 16, 2020, from https://www.americancommunities.org/

Beitsch, L. M., Brooks, R. G., Menachemi, N., & Libbey, P. M. (2006). Public Health At Center Stage: New Roles, Old Props. *Health Affairs, 25*(4), 911-922. doi:10.1377/hlthaff.25.4.911

Bock, T. (2019). What is a Dendrogram? How to use Dendrograms. Retrieved December 16, 2020, from https://www.displayr.com/what-is-dendrogram/

Brems, M. (2017, April 17). A One-Stop Shop for Principal Component Analysis. Retrieved December 16, 2020, from https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

Developers. (n.d.). Retrieved December 16, 2020, from https://www.census.gov/data/developers.html

Frieden, T. R. (2004). Asleep at the Switch: Local Public Health and Chronic Disease. *American Journal of Public Health, 94*(12), 2059-2061. doi:10.2105/ajph.94.12.2059

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (Vol. 2). Beijing: O'Reilly.

Helbich, M., Brunauer, W., Hagenauer, J., & Leitner, M. (2013). Data-Driven Regionalization of Housing Markets. *Annals of the Association of American Geographers, 103*(4), 871-889. doi:10.1080/00045608.2012.707587

Hung, P. D., Lien, N. T., & Ngoc, N. D. (2019). Customer Segmentation Using Hierarchical Agglomerative Clustering. *Proceedings of the 2019 2nd International Conference on Information Science and Systems*. doi:10.1145/3322645.3322677

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R*. New York: Springer.

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

K-Means Advantages and Disadvantages | Clustering in Machine Learning. (2020, February 10). Retrieved December 16, 2020, from https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages

Kanarek, N., Bialek, R., & Stanley, J. (2008). Use of peer groupings to assess county public health status. *Preventing Chronic Disease*.

Levine, J. A. (2011). Poverty and Obesity in the U.S.: FIG. 1. *Diabetes, 60*(11), 2667-2668. doi:10.2337/db11-1118

Link, B. G., & Phelan, J. (1995). Social Conditions As Fundamental Causes of Disease. *Journal of Health and Social Behavior, 35*, 80-94. doi:10.2307/2626958

Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., Gupta, S., . . . Allender, S. (2015). Is Demography Destiny? Application of Machine Learning Techniques to Accurately Predict Population Health Outcomes from a Minimal Demographic Dataset. *Plos One, 10*(5). doi:10.1371/journal.pone.0125602

Manson, S., Schroeder, Riper, Kugler, and Ruggles. IPUMS National Historical Geographic Information System: Version 15.0 [dataset]. Minneapolis, MN: IPUMS. 2020. http://doi.org/10.18128/D050.V15.0

Margins of error in the ACS. (n.d.). Retrieved February 17, 2020, from https://walkerke.github.io/tidycensus/articles/margins-of-error.html

Mays, G. P., & Smith, S. A. (2009). Geographic Variation in Public Health Spending: Correlates and Consequences. *Health Services Research, 44*(5p2), 1796-1817. doi:10.1111/j.1475-6773.2009.01014.x

Measures of Distance between Samples: Euclidean. (n.d.). Retrieved December 16, 2020, from http://www.econ.upf.edu/~michael/stanford/maeb4.pdf

Patlolla, C. (2020, May 29). Understanding the concept of Hierarchical clustering Technique. Retrieved December 10, 2018, from https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

Pflanzer, L. R. (2019, January 06). The 50 US states ranked from most to least healthy. Retrieved December 16, 2020, from https://www.businessinsider.com/the-healthiest-and-unhealthiest-states-in-america-ranked-2018-12

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219*(1), 103-119. doi:10.1243/095440605x8298

Pipis, G. (2020, August 21). Hierarchical Clustering of Countries based on Eurovision Votes. Retrieved December 16, 2020, from https://predictivehacks.com/hierarchical-clustering-of-countries-based-on-eurovision-votes/

Ramachandran, V., Shah, M. K., & Moss, T. J. (2019). How Do African Firms Respond to Unreliable Power? Exploring Firm Heterogeneity Using K-Means Clustering. *SSRN Electronic Journal*. doi:10.2139/ssrn.3310490

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

Sagar, A. (2019, August 24). Customer Segmentation Using K Means Clustering. Retrieved December 16, 2020, from https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3

Sessions, S. Y. (2012). *For the public's health investing in a healthier future*. Washington, DC: National Acad. Press.

Sklearn.cluster.DBSCAN¶. (n.d.). Retrieved December 16, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

State Reports. (n.d.). Retrieved December 16, 2020, from https://www.countyhealthrankings.org/reports/state-reports

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

Tracking down the Villains. (2015, July 14). Retrieved December 16, 2020, from https://netflixtechblog.com/tracking-down-the-villains-outlier-detection-at-netflix-40360b31732

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

Verma, R. (2009). Key Issues in Hierarchical Clustering. Retrieved December 16, 2020, from http://www.hypertextbookshop.com/dataminingbook/public_version/contents/chapters/chapter004/section003/blue/page003.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

Yang, Y., Lian, B., Li, L., Chen, C., & Li, P. (2014). DBSCAN Clustering Algorithm Applied to Identify Suspicious Financial Transactions. *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*. doi:10.1109/cyberc.2014.89

Implementation Appendix

In this appendix, we discuss the data collection and processing steps in more detail. The data on health outcomes is downloaded directly from the Robert Wood Johnson Foundation's website ("Explore Health Rankings," 2020). The data on the percent of the population that is rural in each county is downloaded from the National Historical Geographic Information System (NHGIS) provided by Manson et al. (2020). The data from the 5-year 2014-2018 American Community Survey is imported directly into Python through the *CensusData* package ("CensusData 1.10," 2020), which pulls this data from the U.S. Census Bureau's application programming interface ("Updated," n.d.).

The following variables are downloaded directly "as is" from the relevant data sources with no transformations needed: the obesity rate, the smoking rate, total population, the median age, and the percent of the population that is rural.

To calculate our four variables related to racial composition, we obtain data from the ACS on the number of residents identifying as members of each racial group. Using the total population we then calculate the percent of each county's population that is Hispanic, non-Hispanic white, non-Hispanic African-American, and non-Hispanic American Indian and Alaskan Native. Note that these four variables do not add up to 100 percent because they exclude individuals of other racial categories.

Likewise, to calculate the percent female of each county we simply obtain the male population and female population from the ACS, and then divide accordingly.

To calculate the percent of the percent of the population aged 25+ with at least some college education, we use ACS data on the total population in each possible gender-age-educational compositional grouping.

To calculate the percent of the population that is in the labor force and unemployed, we use the ACS data on the total population in the labor force and the total population that is unemployed, noting that all individuals classified in the ACS as unemployed are also classified as being in the labor force.

For health insurance rates, the ACS provides the total population of each possible grouping of age and insurance status. We then total the population that is under 65 and does not have any form of health insurance, and divide by total population.

To calculate the percent of the adult population that is married, we use ACS numbers on the total adult population, the total male population currently married, and the total female population currently married.

The County Health Rankings data ("Explore Health Rankings," 2020) includes the total number of motor vehicle crash deaths in each county over a five-year period. We divide this number by five to get an annual average, and then merge this number with ACS data on total population to calculate the number per 100,000 residents.

There are 3,143 counties in the 50 U.S. states and the District of Columbia ("Reference Files," 2019) , but we exclude nine counties that have at least one missing value of either our 11 demographic variables or our three health outcome variables. These 9 counties have extremely small populations, which may explain the lack of publicly available data. Note that the ACS tracks counties or county-equivalents in U.S. territories such as Puerto Rico, but the County Health Rankings data does not and thus we only consider counties in the 50 states and the District of Columbia.

Finally, before conducting any clustering analysis we standardize all 11 of our demographic variables. In other words, for each variable we subtract the mean from each observation and divide by its standard deviation, which rescales all variables to have a mean of zero and a standard deviation of one. Standardizing variables is extremely important in clustering so that all variables are measured on the same scale and that our measures of distance calculated in clustering are not affected by arbitrary differences in scaling of variables ("Measures of Distance," n.d.).