Introduction:

The data set contains five variables and one unique ID for 1720 block groups in Philadelphia. The purpose for this assignment is to cluster 1720 census block groups into several groups with specific characteristics, based on five variables: median house value, median household income, percent of individuals with at least a bachelor's degree, percent of single/detached housing units, percent of vacant housing units in that block group. With the help of k-means clustering, we can figure out how many different groups are in Philadelphia estimated on these five fields and explore what class-label information these clusters have.

Methods

1. **How does the K-means algorithm work？**

   K-means minimizes the within-cluster sum of squared errors (SSE), which is calculated by doing the following for each cluster: compute the squared distance between each observation and the centroid of the cluster into which it falls and sum these squared distances.

   The working process for K-means is listed below:

   a. Set a prior for the number of clusters (K)
   b. Randomly select K data points as cluster centers.
   c. Calculate the distance between each data point and K cluster centers that are set before.
   d. Assign each data point to a cluster whose distance from the cluster center is minimal among all cluster centers.
   e. Recalculate new cluster centers
   f. Recalculate the distance between each data point and new cluster centers
   g. See whether the distance from each point to its allocated cluster center has the minimal distance. If all data points have the minimal distance to its allocated cluster center point, stop; otherwise repeat from step d.

2. **What are some of the limitations of the algorithm?**

First, to run K-means, you need to specify K (number of clusters) in advance. Additionally, the algorithm is only applicable for interval variables (only numeric data) – though some researchers will include binary variables. K-means may also have problems with clusters of differing sizes, densities, and non-globular shapes. The algorithm is also unable to handle noisy data and outliers. Instead of filtering outliers out, K-means will try to include outliers in a cluster, which heavily influences the clustering division. Finally, occasionally, the final clustering solution will be incorrect because the K-means algorithm will find the local minimum of SSE, rather than the global minimum SSE.

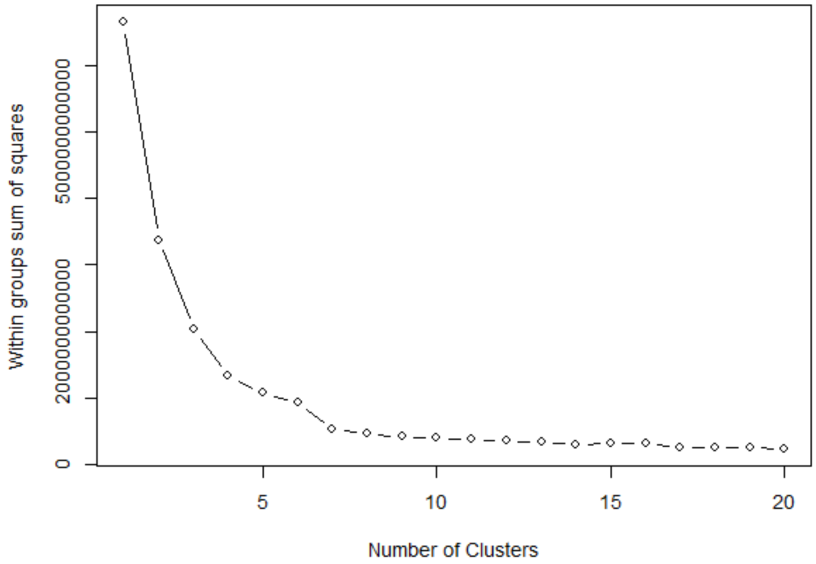
3. **What are some other clustering algorithms, and might they be more appropriate here?**

There are other clustering algorithms like "hierarchical clustering", and "Density-Based Clustering". The hierarchical clustering approach is an alternative to k-means clustering for identifying groups in a dataset. It does not require analysts to specify the number of clusters to be generated, as is the case with k-means clustering. Additionally, hierarchical clustering has the added advantage of producing a tree-based representation of the observations, called a dendrogram, which is more attractive than K-means clustering. As for the hierarchical clustering, it is not suitable for our study because the dataset is too large for it. To be more specific, the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2). To run a hierarchical clustering algorithm here, the time will be long.

DBSCAN is a clustering method where observations that have many neighbors nearby grouped together in a single cluster, while observations whose nearest neighbors are too far away are outliers and aren't part of any cluster. The advantage of DBSCAN is that its goal is to identify dense regions so that it can identify irregular cluster shapes and cluster data more accurately. DBSCAN is more suitable here because it can identify outliers and not assign them to a cluster. In that way, clusters will be more compact and the members insider the cluster are more similar. Besides, DBSCAN is fine regardless of whether densities are the same, however, k-means will have a problem when densities aren't the same.

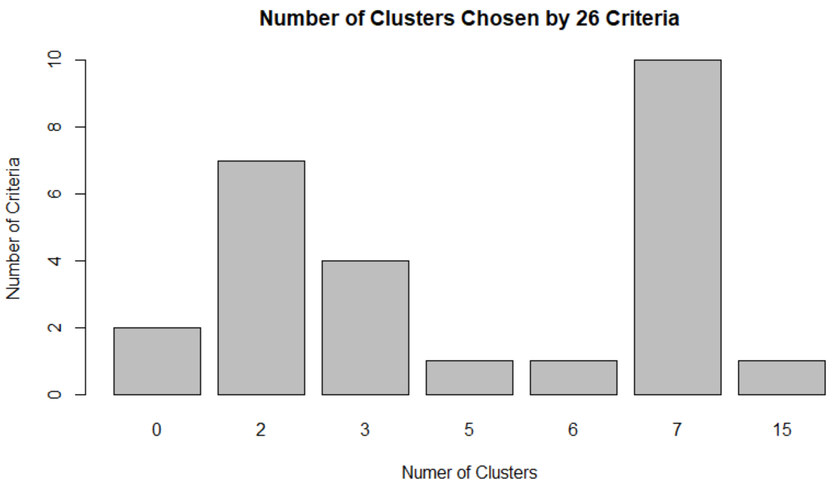
Results:

To identify the optimal number of clusters to run a K-means, we first calculated the SSE for a range of possible cluster options. The scree plot below illustrates the decrease in SSE as the number of clusters increased.



While there are significant decreases in SSE from one to five clusters and six to seven clusters, the optimal cluster number is illustrated on the scree plot at the location of the ‘elbow’, or the point at which there are no longer large decreases in SSE. In this case, this occurs at seven clusters.

To confirm this, The R package NbClust was used to identify the optimal number of clusters for our data. This bar plot illustrates the number of NbClust’s methods identified seven clusters as the optimal number of clusters.

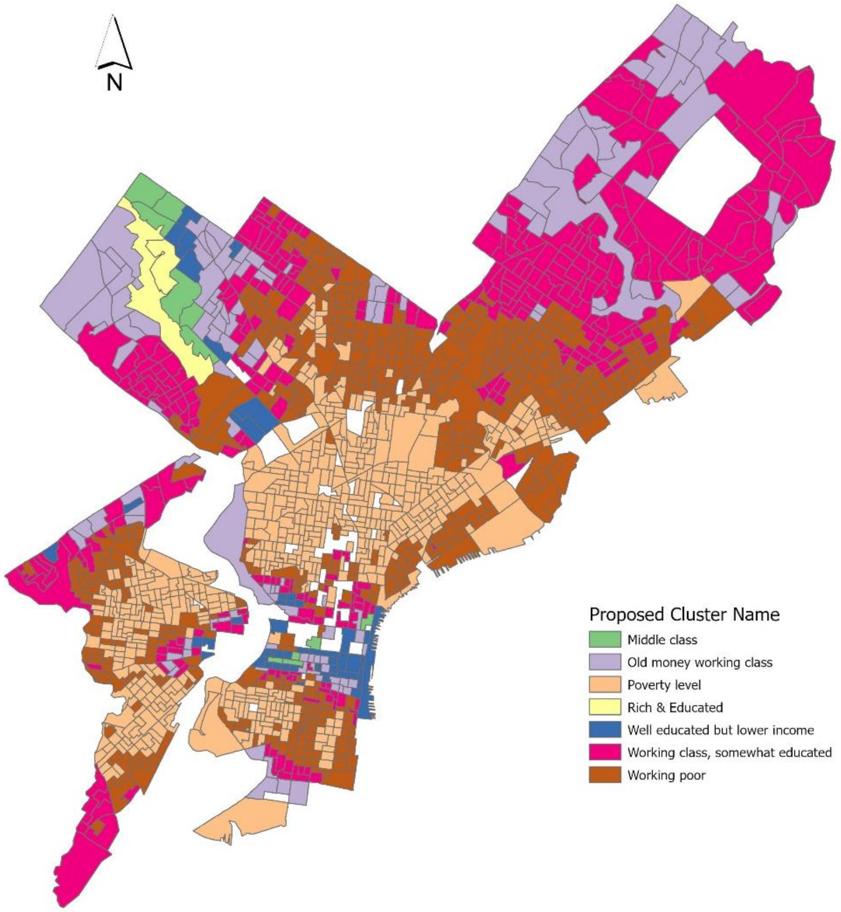


Now that the optimal number of clusters has been identified, the data were grouped into seven distinct clusters. The average values of each variable were calculated for each cluster. These values are presented in the table below, along with a proposed group name for each cluster.

CLUSTER	MEDHVAL	MEDHHINC	PCTBACHMOR	PCTSINGLES	PCTVACANT	fit.km$size	Proposed Name
1	86469.9	40886.8	24.1	9.6	5.6	346	Working class, somewhat educated
2	921900.5	200001	60.1	65.1	2.9	2	Rich & Educated
3	233078.7	54109.3	68.1	14	6.5	47	Well educated but lower income
4	436254.5	73972.3	68.9	44	6.5	11	Middle class
5	29901.7	20447.9	5.5	7.9	17.4	632	Poverty level

6	55570.6	31357.7	11.6	5.6	9.8	558	Working poor
7	133812.1	47802.4	43.1	25.4	5	124	Old money working class

While the proposed descriptive names for each cluster were very ‘back-of-the-napkin’ a priori knowledge, they were created by examining the average values for each cluster and attempting to attribute some group classification. These groupings, however, can yield important information regarding the distribution of wealth and its drivers across the City of Philadelphia. To visualize this graphically, the map below is proposed.



It is readily apparent that these categories are clustered in space, and therefore, K-means cluster membership is spatially autocorrelated. This map indicates that the K-means analysis is effective at identifying possible groups of individuals based on income, property values, educational attainment, number of vacant buildings, and number of single-family homes within regions in Philadelphia. However, it may be wise to adjust the proposed cluster names as a result of mapping, because the names could be seen as problematic.

Discussion:

The input variables, MEDHVAL, MEDHHINC, PCTBACHMOR, PCTSINGLES, and PCTVACANT, are indicators of socioeconomic status and it is evident that these indicators are spatially-autocorrelated in Philadelphia. Therefore, it is not surprising that the K-means algorithm can identify spatial clusters within Philadelphia. We observe that clustering of low incomes and low house values as illustrated by our proposed clusters of ‘Poverty Level’ and ‘Working Poor’ in parts of West-Philadelphia and North-Philadelphia, areas that have been historically segregated and disinvested in. Conversely, we observe wealthier parts of the city such as Center-City and the area around Wissahickon Park as illustrated by the clusters of ‘Rich & Educated’, ‘Middle Class’, “Well educated but lower income”. If one is familiar with the demographic distributions of Philadelphia, these results are not surprising. However, one startling observation is the number of tracts that were assigned to the ‘Poverty Level’ and ‘Working Poor’ clusters. Ultimately, these results point to a gap between the wealthy and poor people of Philadelphia.