HW5-kMeansClustering

Richard Barad, Dave Drennan, Jarred Randall

2023-12-04

Contents

Introduction

Complex real estate markets influence the prices of homes, and determining what features are most impactful to the value goes beyond physical characteristics. In this report, we analyze a data set containing variables for 1720 block groups in Philadelphia. The variables include:

- Median house value (MEDHVAL),
- Median household income (MEDHHINC),
- Percent of individuals with a bachelor's degree or higher (PCTBACHMOR),
- Percent of single/detached housing units (PCTSINGLES), and
- Percent of vacant housing units (PCTVACANT).

We use the K-means clustering method to uncover patterns and relationships within the data. K-means clustering can help identify distinct groups of block groups with similar characteristics based on our variables. While the method requires initially choosing a number of clusters to sort the data, K-means clustering helps us to consider the mean value per variable for each cluster, which we can then use to categorize each cluster with a more descriptive name to better understand the data.

Methods

K-means Clustering

The K-means algorithm, a form of cluster analysis, is used for large data sets with numeric variables. The method uses an iterative process that creates a division of objects into non-overlapping subsets (clusters) such that each object is in exactly one cluster. Before the six-step K-means process is initiated, the total number of clusters (K) must be specified. The six-step process is as follow:

- Randomly select K data points as cluster centers
- Calculate the distance (typically, Euclidean) between each data point and K cluster centroid.
- Assign each data point to the nearest cluster based on the calculated distances. The "nearest" means the centroid with the least distance to the data point.
- After all data points are assigned to a certain cluster, new cluster centers are recalculated by taking the average of all points assigned to each cluster.
- The distances between each data point and new cluster centers are then updated.
- Finally, if no data point was reassigned, stop. Otherwise repeat from step 3.

K-means aims to minimize the within-cluster sum of squares errors (SSE). The SSE is a measure of how close each data point in a cluster is to the cluster's centroid. SSE is calculated by computing the squared distance between each observation and the centroid of its nearest cluster and summing these squared distances. The overall K-means SSE is computed by summing these SSEs across clusters.

For more confidence in the clustering solution that the K-means algorithm chooses, running the algorithm multiple times with different initial centroid values can help select the result which yields the lowest SSE. In our analysis of block groups in Philadelphia, we run the K-means cluster algorithm 25 times and select the result which yields the lowest SSE.

Although K-means is a useful form of cluster analysis, it has its limitations. These limitations include:

- Requiring the number of clusters to be specified in advance, which is not always evident without domain knowledge.
- Not being suitable for categorical data because it relies on Euclidean distances, which are not defined for categorical attributes.
- A chance of encountering difficulties when clusters are of differing sizes, densities, or non-globular shapes.
- Being unable to handle noisy data and outliers.
- Occasionally, the final clustering solution will be incorrect, because K-means will find the local minimum of SSE rather than the global minimum.

Additional Clustering Methods

Other clustering algorithms include hierarchical clustering and density-based clustering (DBSCAN).

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters by either successively dividing or combining clusters based on distance metrics until a desired structure is achieved. Unlike K-means, it is not necessary to input the number of clusters. Any number of clusters may be chosen based on a dendrogram, or a cluster tree. Hierarchical clustering does not work well for large data sets, so it would not be an appropriate method for this use case given that the dataset is large.

DBSCAN can identify irregular cluster shapes, observations that have many neighbors nearby grouped together in a single cluster, and observations whose nearest neighbors are too far away which are outliers and aren't part of any cluster. DBSCAN would likely be appropriate for this data set. This method could potentially cluster points that reflect neighborhoods or regions of Philadelphia when they share similar characteristics - however, if DBSCAN generates too many clusters, it may hamper interpretability.

Results

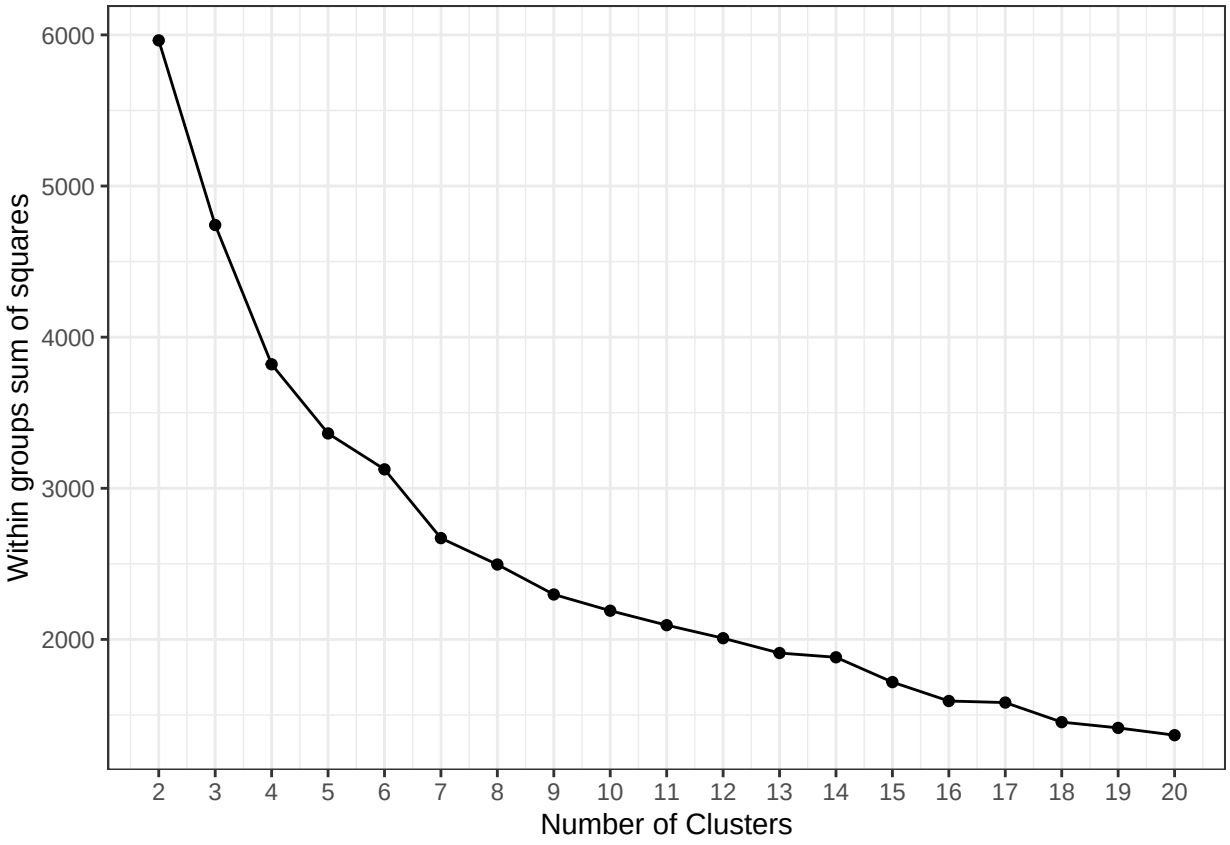
Cluster Criteria

We consider the results of two different methods to optimize the number of clusters for Philadelphia Census block groups in our analysis - interpretation of a scree plot and output of the NbClust package in R, which provides the results from a compilation of up to 30 methods for choosing the optimal number.

Scree Plot

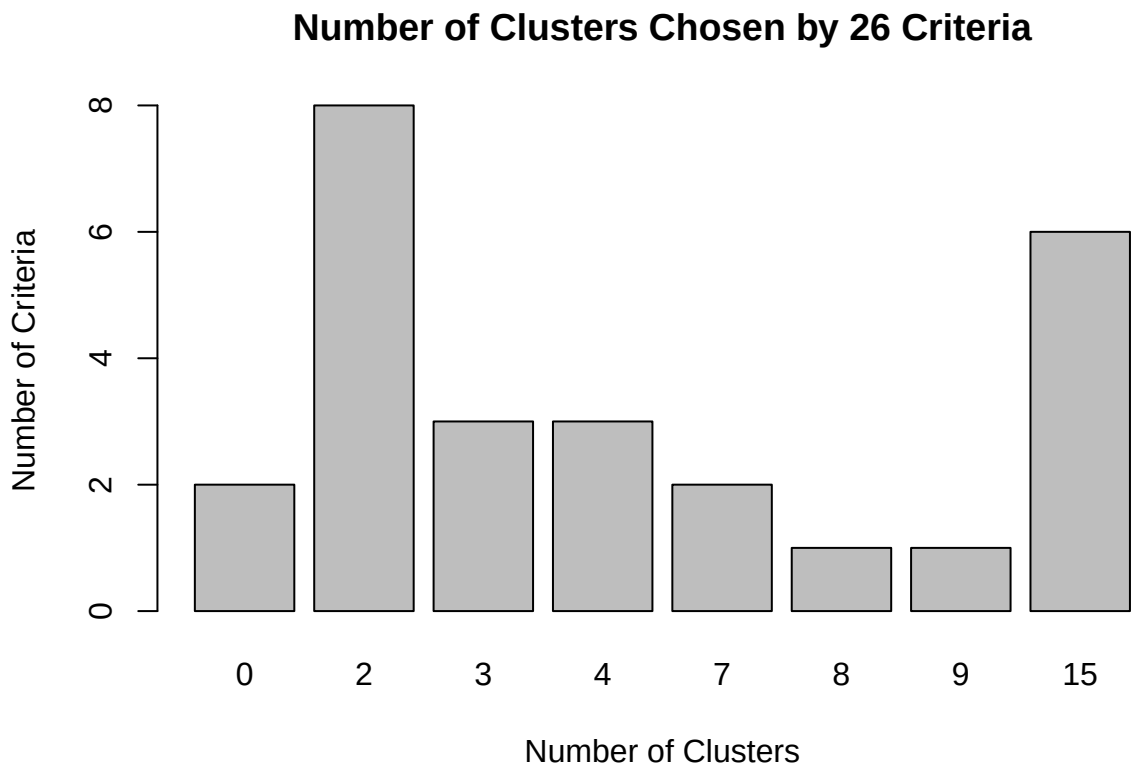
We start with the scree plot, which is interpreted by finding the "elbow" of the plot - the number of clusters in which the drop off in the within group Sum of Squares Error to the next number of clusters is minimal compared to the previous number of clusters.

According to the scree plot generated by our data, the optimal number of clusters appears to be 7. Decreases in the within group Sum of Squares Error become much smaller after this point compared to the preceding points. However, we also assess the results from the NbClust package to validate this choice, given that the package provides a more robust assessment of optimal clusters.



NbClust

NbClust identifies 2 as the optimal numbers of clusters based on criteria from 26 of the methods considered most relevant by the package. By plotting the outcomes, we see that 8 of the 26 methods chose 2 as the optimal number of clusters. The next highest chosen number is 15 clusters, which was picked by 6 of the 26 methods.



Choice for Optimal Number of Clusters

We choose 2 clusters as the optimal number to use in our analysis. While the scree plot appears to show 7 as the optimal number, the more robust nature of the NbClust assessment indicates that 8 of the 26 methods propose 2 clusters as the best number. Additionally, lower numbers of clusters can help simplify the categorization and interpretability of results.

Cluster Analysis

Cluster Size

We first examine the size of our two clusters and see that Cluster 1 is much larger, with 1446 block groups compared to Cluster 2's 274 block groups.

Cluster	Size
Cluster 1	1446
Cluster 2	274

Cluster Means

We then examine the means of each of our variables for our two clusters. The results show stark differences in our clusters:

- The average Median House Value is over three times higher in Cluster 2 block groups compared to Cluster 1, at over $150,000 compared to just under $50,000

- Cluster 2 block groups, on average, have nearly 50% of individuals with at least a bachelor's degree compared to the average for Cluster 1 of 10% of individuals per block group
- Average Median Household Income for Cluster 2 is nearly double Cluster 1, as about $52,000 compared to about $28,000
- Cluster 2 block groups average less than half the percent of vacant housing units compared to Cluster 1, at 4.8% versus 12.5%
- The average percent of single/detached housing units in Cluster 2 block groups is over three times higher than Cluster 1, at 22.3% compared to 6.8%

The variable means for our two clusters indicate that the block groups identified for Cluster 2 are wealthier, more highly educated, and have higher demand to live in them compared to Cluster 1. Given the disparities, the cluster solution appears to make sense with the two clusters. Based on this output, we choose to identify the two clusters with the following titles:

- Cluster 1: Limited Gentrification Block Groups
- Cluster 2: High Gentrification Block Groups

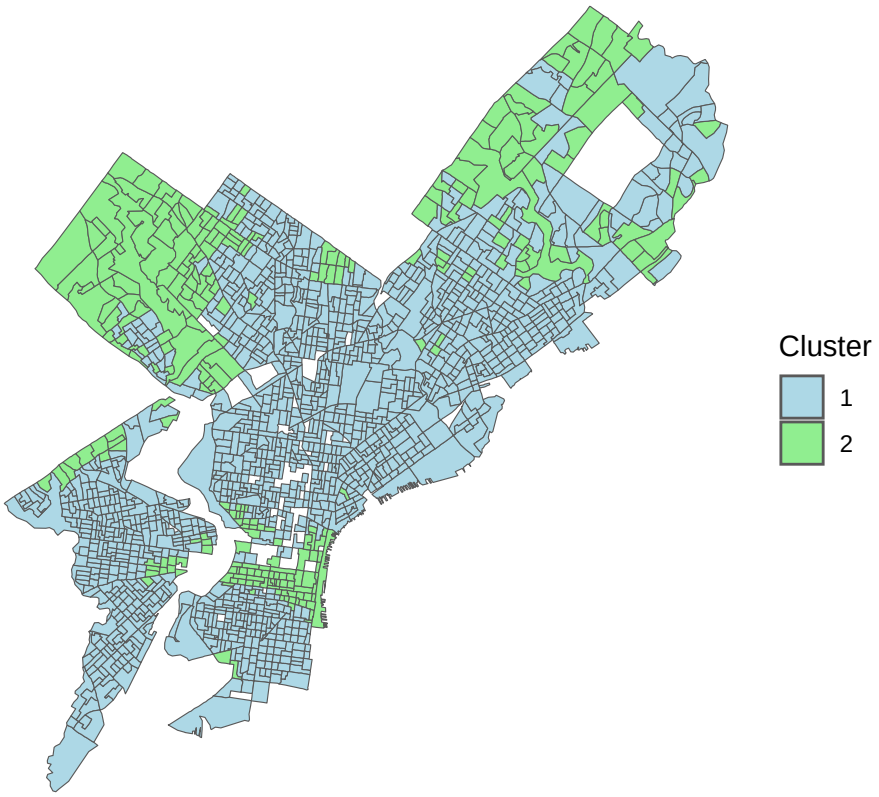
Cluster	Median House Value	% Bachelor's Degree or higher	Median Household Income	% Vacant HU	% Single/Detached Housing Units
1	49952.4	10.2	27668.5	12.5	6.8
2	152495.3	46.9	51982.6	4.8	22.3

HU: Housing Units

Cluster Map Results

By mapping our observations based on the two clusters, we observed that our K-means clusters appear to also cluster in space, indicating likely spatial autocorrelation. Generally speaking, Cluster 2 comprises areas around Center City, Northwest Philly, and Northeast Philly, with smaller areas in University City, the northern edge of North Philly and the western edge of West Philly. Cluster 1 covers most of the rest of the city.

These clusters reinforce the title choices to describe Cluster 1 and Cluster 2. Generally speaking, areas in Cluster 2 such as Center City, University City, and Northwest Philly are more gentrified compared to Cluster 1 areas, with more expensive real estate and wealthier households that hold higher education degrees. Northeastern and northwestern neighborhoods in Philly also have a more suburban feel, with higher proportions of single family detached homes.



Discussion

The map of our cluster analysis and the average values for the variables in each cluster show the patterns of wealth and gentrification in Philadelphia. While only two clusters is a relatively simplified map of the city, our High Gentrification cluster reflects the reality that wealth in Philly can be spatially segregated.

It is unsurprising to see where the K-means clustering shows divides throughout the city. The cluster map is largely similar to the historical redlining map of the city, where the Home Owners' Loan Corporation (HOLC) identified areas that it deemed high-risk for lending through racist discrimination against Black residents and other communities of color.[1] Most of the Cluster 1 areas are in sections of the city that HOLC deemed "definitely declining" or "hazardous" for mortgage lending. Our k-means clustering highlights this history of redlining, segregation and racism in Philadelphia.

[1] Jake Blumgart, How redlining segregated Philadelphia, WHYY, 10 December 2017, https://whyy.org/segments/redlining-segregated-philadelphia/