K-Means Clustering

Shengao Yi 76852392, Xuening Zhang 65401665, Yu Wang 20166947

1 Introduction

The K-means clustering analysis could recognize and locate the clusters in datasets efficiently through further dividing the datasets to be smaller groups and apply assessments for them, and it is a very algorithm for analysis. In this project, the K-means cluster analysis is used to the dataset of Homework 1 and 2 we utilized previously. Similarly, there are 5 variables included in the dataset of this project:

- **MEDHVAL**: median house value
- **MEDHHINC**: median household income
- **PCTBACHMOR**: percent of individuals with at least a bachelor's degree
- **PCTSINGLES**: percent of single or detached housing units
- **PCTVACANT**: percent of vacant housing units

With the help of K-means cluster analysis, using the R studio and ArcGIS Pro, the optimal number of clusters, numbers and characteristics of the clusters in these 5 variables could be identified. Also, using the K-means cluster analysis, questions could be answered, including How to sort the household owners? How many groups of clusters could be recognized in the dataset? Are they auto-correlated? Is there any particular types of groups clustering in a specific way? Is there any indications from the types of clusters for the public policy?

2 Methods

2.1 K-means Algorithm

Cluster Analysis is a set of data-driven partitioning techniques designed to group a collection of objects into clusters. K-means cluster method is one of the most popular algorithms. The steps are as follows:

1. Determine the number of clusters in advance before running the clustering algorithm, which can be represented as K. There are some methods for us to choose the optimal number of clusters. In R, the NbCluster package can do us a favor. We can generate a Scree plot of the sum of squared errors (SSE) for each number of clusters.
2. The second step is an iterative process:
    1. Randomly choose K data points as the centers for each cluster.
    2. Calculate the distance between each data point and K cluster centers. The distance can be calculated normally by "Euclidean distance".
    3. Distribute each data point to a cluster whose distance is the closest among all cluster centers.
    4. After giving each cluster some data points, recalculate the new cluster centers for each cluster and update the new distance between each data point and new cluster centers, then goes back to step 3 in the iterative process. If there is no data point was reassigned, stop the process.

2.2 Limitations of K-means

Despite that K-means algorithm can help us to identify the clusters in the block groups in Philadelphia, K-means function has a handful of limitations that can cause us issues when we are doing clustering analysis. Namely, these limitations are requiring number of clusters specification before doing the analysis, limited to numeric data, inaccurate results when the clusters in the data have varies sizes, densities, and non-globular shapes, heavily influenced by outliers, and doesn't guarantee global minimum.

- Sometimes when dealing with a large dataset and having limited information about the data, consequently, can be difficult to figure out what number of clusters we should use to distinguish the groups without the help from software packages like NbClust.
- K-means is designed for identifying clusters for numerical data, since the variables, MEDHVAL, MEDHHINC, OCTBACHMOR, PCTSINGLES, and PCTVACANT, in the Philadelphia median house value dataset are all numerical, we wouldn't encounter the issue in this case.
- If the clusters in our dataset have different sizes, K-means doesn't perform well. For example, if there are three clusters, and two of the clusters have the same size, and the third one has a bigger size. Then, based on the input of the number of clusters specified, the algorithm will regroup the data into the clusters that have identical size. As a result, the groups created by the algorithm will be significantly different than the original data.
- When the data have different densities, a problem similar to different group sizes will occur. The algorithm will tend to group the near data into one cluster.
- One major problem with K-means algorithm is that it assumes the clusters in our data will have spherical shapes. Thus, the shape of the clusters returned by the algorithm would be noticeably different the true pattern.
- The algorithm is sensitive to outliers in the data, and it lacks the capability to exclude outliers because all data points must be included and assigned to a cluster.
- Since K-means algorithm minimizes within-cluster variance, it will return false clusters as the algorithm emphasizes on the local minima which would split or merge clusters inappropriately.

2.3 Other Cluster Algorithms

The other clustering algorithms are density-based clustering and hierarchical clustering. And the density-based clustering might be more appropriate for the dataset in this project for the reason that there are approximately 1720 observations in the dataset we use, which is too big for the hierarchical clustering to get accurate results.

3 Results

3.1 NbClust and Scree Plot: Optimal Number of Cluster

There are a wide variety of methods to identify the best number of clusters. The NbClust package in R can help us a lot. One of the most useful methods is the Scree Plot. K-means method can minimize the within-cluster sum of squared errors (SSE). Squared errors is the squared distance between each observation and the center of the cluster into which it falls, then sum them up. The figure below shows the SSE of number of clusters from 1 to 20. The location of the elbow in the resulting plot suggests a suitable number of clusters for the K-means. We can find a significant decrease from 1 to 3, after 4, the rate of decline is starting to flatten out. The drop between 5 and 6 clusters is essentially 0, which suggests 3 or 4-clusters might be an optimal choice.

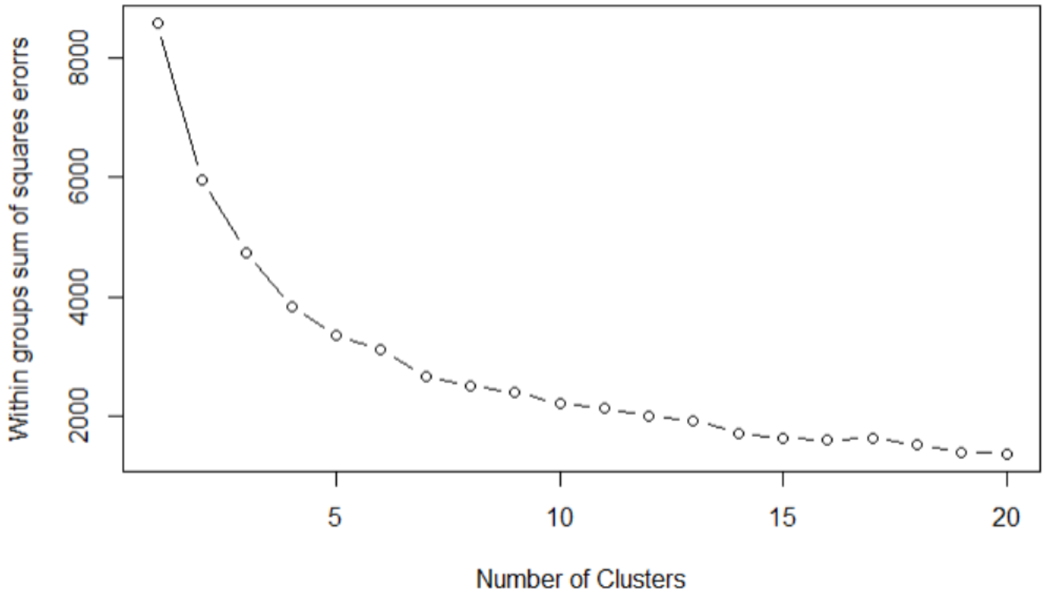
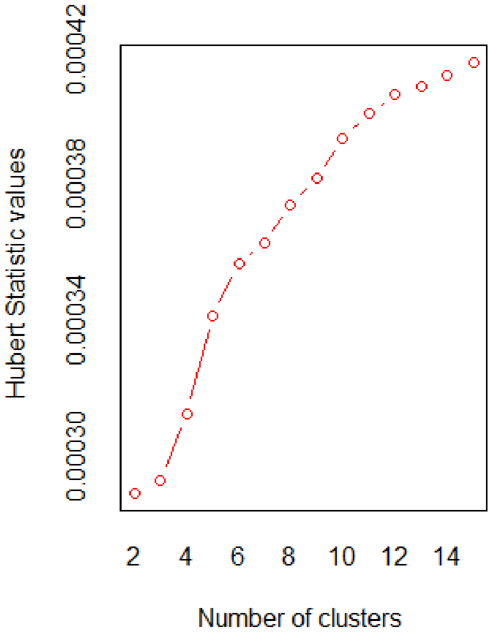
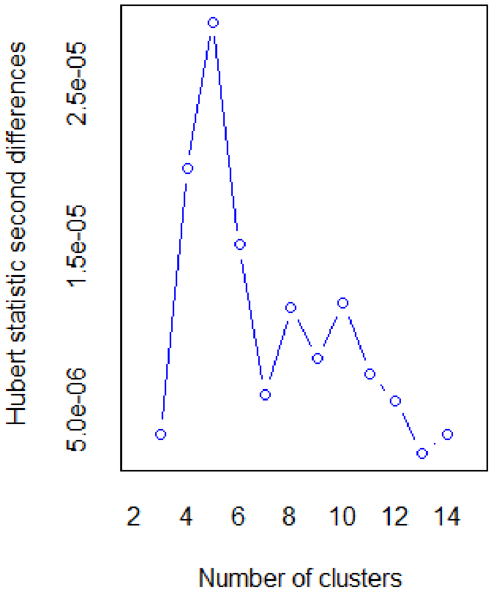
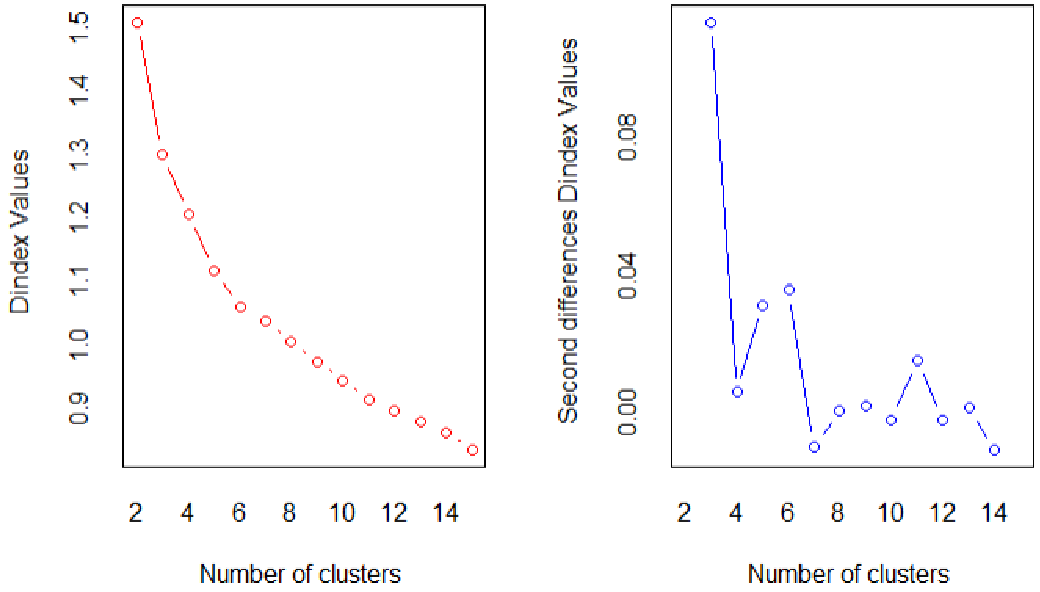


Figure 1. SSE of each number of clusters







We have also plotted the number of clusters suggested using 26 criteria, the number of 3 and 4 is same. Thus, we have used both of them to conduct cluster analysis. The results showed that one category of 4-clusters result is separated from the 3-clusters result, which all of the variables are higher than any other category. We think it's more reasonable. Finally, we choose 4 as the number of clusters.

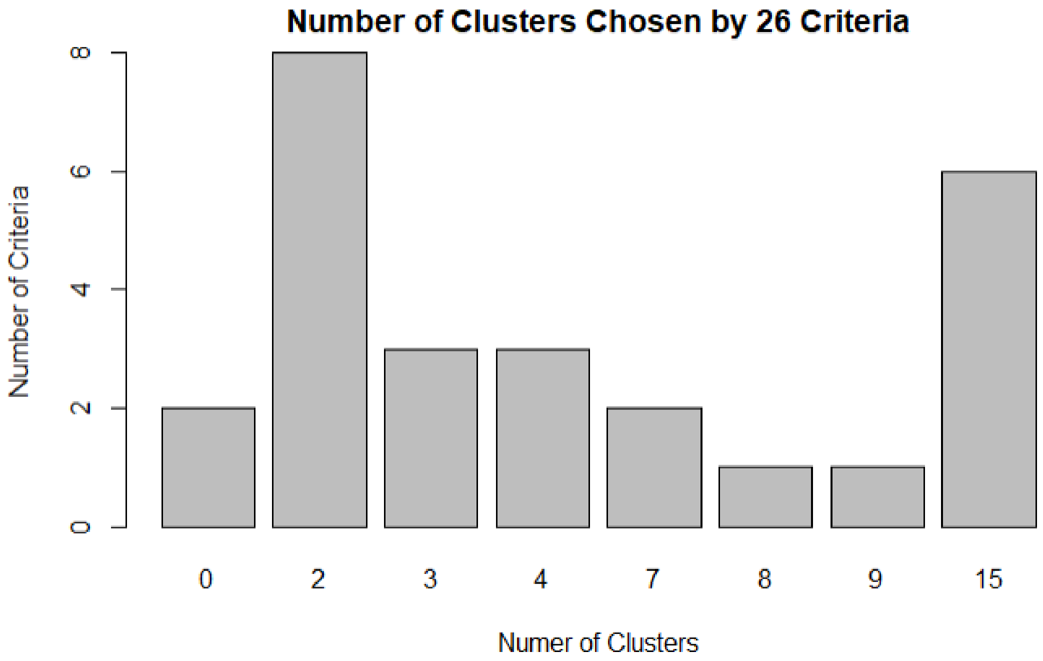
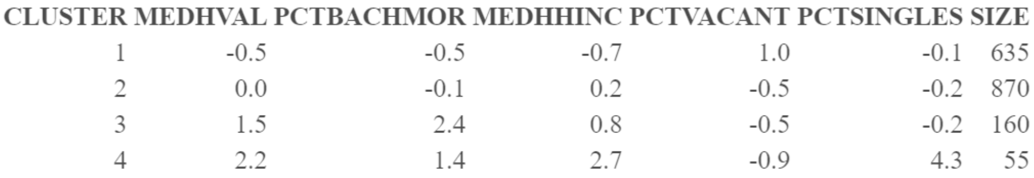


Figure 2. Result from 26 Criteria

3.2 Four Clusters



CLUSTER	MEDHVAL	PCTBACHMOR	MEDHHINC	PCTVACANT	PCTSINGLES	SIZE
1	-0.5	-0.5	-0.7	1.0	-0.1	635
2	0.0	-0.1	0.2	-0.5	-0.2	870
3	1.5	2.4	0.8	-0.5	-0.2	160
4	2.2	1.4	2.7	-0.9	4.3	55

Table 1. Mean variable in each cluster

Table 1 shows the mean value of each variable in every cluster. For Cluster 1, we can find that it has the lowest house value, percentage of bachelor’s degree, median household income, and percentage of single family houses as well as the highest percentage of vacant house. We define it as low-income and poor-educated group. All the values in Cluster 2 are around 0, it’s average income and education group. For Cluster 3, it’s above average income and education group. In terms of Cluster 4, all the values are very positive and high, we named it as high-income and well-educated group. Then we have exported the cluster details and merged them with the census tracts of Philadelphia.

3.3 Spatial Distribution of Clusters

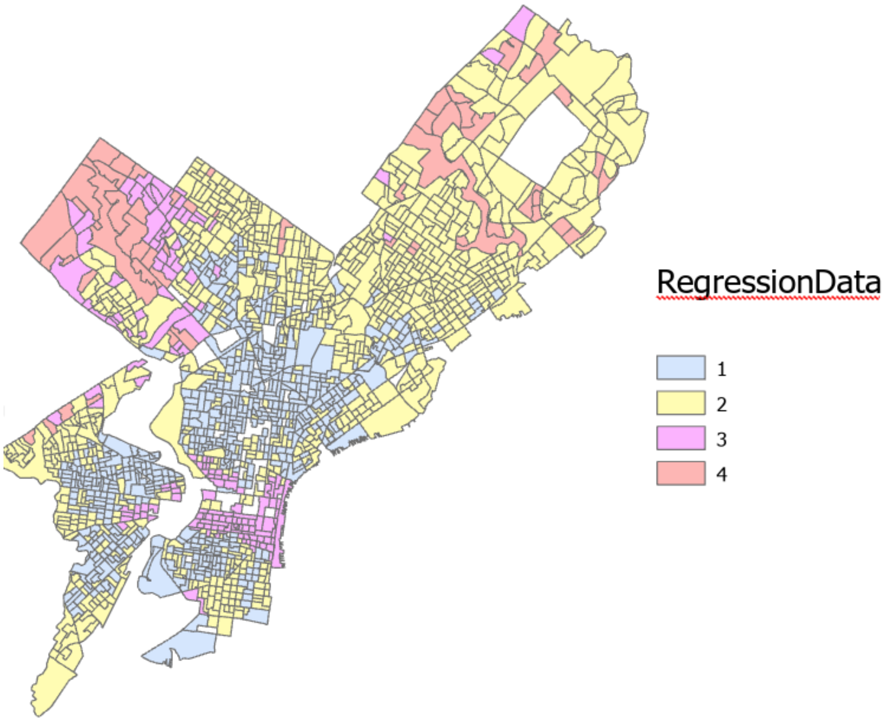


Figure 3. Spatial distribution of each cluster

Since the figure for SSE against cluster solutions shows that a 4-cluster solution is a good fit for the *Regressiondata* dataset. Based on the map of cluster distributions, we observed that group 1 (low-income and poor-educated tracts) is spatial clustered in North, Central, and West Philadelphia. Group 2 (average income and educated tracts) appeared to be located all over Philadelphia. Finally, group 3 (above average income and educated tracts) and 4 (high-income and well-educated tracts) appear to be distributed near to each other, and we think this phenomenon is due to similar traits shared among these two groups. Group 3 and 4 are spatially clustered in Northwestern Philadelphia and a few clustered in Northeast Philadelphia.

When we were naming these four groups, an important factor that influenced our decision was the median house value, income, percentage of bachelor’s degree. The reason that the names for these 4 cluster groups were not chosen based on the geographic locations is that the clusters don’t appear to be located uniformly. For example, Group 3 located in various areas across Philadelphia, these tracts would be significantly different than the lower-income groups, group 1 and 2, if we assigned the groups based on geographical locations.

4 Discussion

In this project, the K-Means Cluster analysis is applied to the 5 variables of the datasets, and greatly identified the clustering. And with the help of NbClust package and the derived scree plot, there are 4 groups of clustering generated based on the difference in median house value, median household income and the percentage of the residences with a bachelor’s degree instead of their geographical distributions. The 4-cluster solution fits the dataset *regressiondata* very well. According to our results, the tracts of Group 2 with residence of average income and education level occupies across most areas in Philadelphia, which surprises us very much.