Insight Report

SUBJECT: Education and Public Traffic Facilities Distribution Insight Report

Executive Summary

Education and public transports inequality is a crucial part of the sustainable urban planning in China. This insight report, by using clustering analysis, identified some patterns among the public transports around the public transports in Beijing. The patterns suggest, although more than 88% of the public schools has accessible public transports, yet 280 schools need improvements in term of public transport accessibility. In additions. Among these schools, most of the schools are elementary schools in the rural area. It is high time to put improve the conditions accordingly for these 280 schools in the long run, with a emphasis on the parking lots for rural elementary school in the short run.

1. Introduction

State-owned land is a characteristic of an authoritarian country like China. Therefore, proper and sustainable urban planning and land use are essential functions of the government in China. In my research design, I would like to use clustering analysis to see whether public transports are evenly located around public schools in Beijing. If the allocation of the public transport is reasonable, then it will be improving the education equality level of Beijing in the long run.

- **Relevant research that I would like to partially replicate**: In 2020, a Visual Analysis of Land Use Characteristics Around Urban Rail Transit Station in Beijing elaborated on how identifying and analyzing land use characteristics around urban rail transit stations can significantly contribute to urban rail transportation operation and management (Cai et al., 2020). My research topic could partially replicate Cai's research by focusing on the following three perspectives:

  - **Data:** In the research, the author used **Points of Interests datasets from Gaode Map API** to visualize the allocation of infrastructure along with the railway stations. I requested the datasets that they used in the research and would build my research upon the datasets they provided
  - **Technical Approaches** that I could use in my clustering analysis: In Cai's study, he calculated the distance to subway stations according to Euclidean distance. By using the Euclidean distance, we could calculate the nearest distance from schools to different public transports facilities and use them as features for K-means clustering analysis.
  - **Findings** that I would like to extend in my research: The study's key results highlighted POI of specific types of land-use characteristics usually clustered around specific transit station locations of Beijing. Based on this finding, in my research, I would focus on the education-related POI and its distance to different public transports facilites.

- **Critiques**: although this research provide us with accurate visualization of the land use characteristic of Beijing, yet the finding is a broad overview of the land use characteristics and does not use any machine learning algorithm to validate its visualization result. In additions, it does not specific whether the gathering of POIs around specific transit station is reasonable or not.

- **Innovation from my research**: Based upon the summary and discussion of the research mentioned above, my research would use the clustering analysis to identify any specific pattern to validate the result of visualization from Cai's research. In additions, by identify some patterns within the allocation of public transports and education inequality, I aim to give the Beijing government some specific recommendation in terms of planning of building new educational and public transports related infrastructures. That is , what to build and where to build.

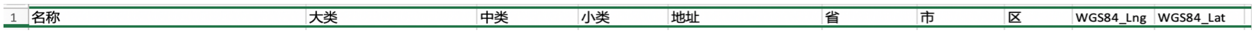
2. Data

- **Original Dataset**

The original dataset I requested from Cai is describe as below. The source of this dataset is the API of Gaode Map. The URL is https://restapi.amap.com/v3/place/text?parameters. According to the term three copyrights and limitation(许可和限制), :the data could not be used for commercial purpose. Using keywords such as school, traffic as parameters In additions, a preprocess of the data (see appendix) is carried out and the final data/features after preprocessing are discussed in detail in the next section.

In total, 14 datasets sort there by the types of POIs. The 14 types are traffic, accommodation, sports facilities, public area, company and enterprise, health, residential area, government and NGOs, lifestyles, shopping malls, business, attractions, education, and dining. There are at least 50000 observations in each dataset with the location names, categories, longitude, and latitude of the POI (Point of Interests) by location types in Beijing. the dataset that I would focus on would be the education dataset and traffic dataset.

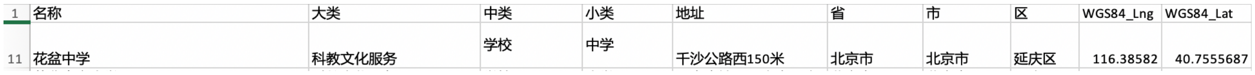
a) Ten Variables and translations: please refer to Figure 1 below (from left to right)
1) 名称(Name of the location)
2) 大类(Main Categories): This is the broadest classification
3) 中类(Medium Categories): sorted the POIs in a more detailed way
4) 小类(Minor Categories): Sorted the POIs by the most detailed categories.
5) 地址(Address)
6) 省(Province)
7) 市(City)
8) 区(District)
9) WG585_Lng (Longitude)
10) WG584_Lat (Latitude)



1	名称	大类	中类	小类	地址	省	市	区	WGS84_Lng	WGS84_Lat

Figure 1

b) Below is an example of an observation in the dataset. please refer to Figure 2 below(from left to right)



1	名称	大类	中类	小类	地址	省	市	区	WGS84_Lng	WGS84_Lat
11	花盆中学	科教文化服务	学校	中学	干沙公路西150米	北京市	北京市	延庆区	116.38582	40.7555687

Figure 2

1. 1) 名称(Name of the location): 花盆中学(Hua Pen Middle School)
2. 2) 大类(Main Categories): 科教文化服务类(Education and culture service)
3. 3) 中类(Medium Categories): 学校(School)
4. 4) 小类(Minor Categories): 中学(Middle School)
5. 5) 地址(Address): 干沙公路西 150 米(150 Gansha West Road)
6. 6) 省(Province):北京(Beijing)
7. 7) 市(City):北京(Beijing)
8. 8) 区(District):延庆区(Y anqing District)
9. 9) WG585_Lng (Longitude):116.38582
10. 10) WG584_Lat (Latitude): 40.7555687

- **Final Dataset and Features**

The variables below come from the cleaned version of the Educational Point of Interests and Traffic Facilities Point of Interests' datasets from Gaode Map API, provided by the author of *a Visual Analysis of Land Use Characteristics Around Urban Rail Transit Station in Beijing.* Both original datasets include categories of the locations and geographic coordinates of the locations. After preprocessing the dataset, *there are in total 2174 observations in the dataset. The unit of analysis would be school.* This section provides summaries and graphs of relevant variables in the cleaned data, which will comprise the variables used in the clustering analysis. Finally, we should also try to find more data related to school enrollments, population, and school quality.

1. Elementary School/ Middle School

   The elementary school and Middle school together are the public schools that built to provide compulsory education in China. Both elementary school and middle school variable are binary variables. The elementary school variable indicates whether a school is an elementary school (Grade 1 to Grade 6) or not. Middle school variable is whether a school is a middle school (Grade 7 to Grade 10) or not. A "1" means that the school is a public elementary school/ public middle school in Beijing, while "0" means that the school is not a public elementary school/public middle school in Beijing.

   Below, table 1 shows that only 37.04% of the public schools are public elementary schools and 62.9% of the schools are public middle schools.

	Public Elementary School	Public Middle School
Count	1368	805
Proportion	0.3704	0.6295

Table 1 Summary Statistics for Elementary School

Figure 1, below, shows findings that are consistent with the numerical summary provided above in Table 1 that there are more public middle schools than public elementary schools.

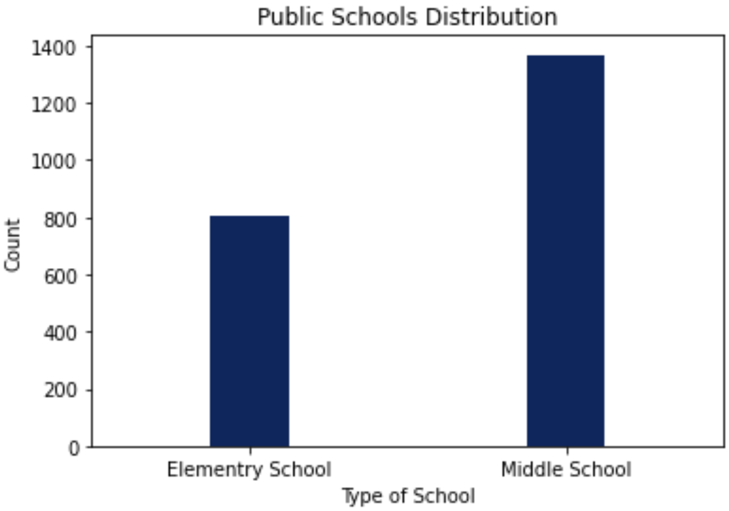


Figure 3 Public School Distribution

2. Distance to the Nearest Subway Stations

   The distance to subway stations is a continuous variable that indicates the distance from each school to the nearest subway station. It is calculated using the latitude and longitude of the subway stations as well as the longitude and latitude of the school.
   The Euclidean distance formula used to calculate the distance is Equation 1. Xi is the geographic coordinates of the school and Yi are the geographic coordinates of the subway stations. The i in the formula is the number of observations, which is 2174 in this dataset. The unit of the distance is kilometer. Using algorithm (algorithm 1) in the appendix, we could find the nearest distance calculated using Euclidean distance.

   $$\mathrm{Dist}(X_i, Y_i) = \sqrt{\sum_i^n (Xi^2 - Yi^2)} \quad \text{(Equation 1)}$$

   According to Table 3 (below), the minimum distance is 0 km and the maximum is 103km. The mean distance is 8 km. The Median is 1.01km. The median and mean are not close, so we should look further into the distribution to see if there are outliers that could potentially affect the clustering analysis.

	Distance to subway station
Mean	8.260
Std	16.322
min	0.000
25%	0.510
50%	1.010
75%	7.340
max	103.260

Table 2 Summary of Distance to Subway Station

The distribution in Figure 2 indicates that most of the observations fall within 30 km. In additions, there are a few outliers and extrema that falls outside of 40km. We should take these outliers into consideration when choosing the method of clustering analysis.

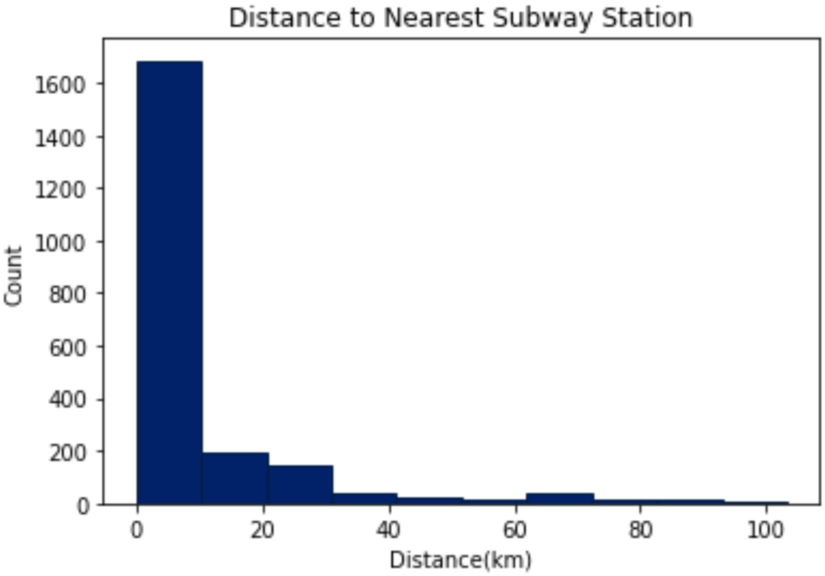


Figure 4 Distance to the Nearest Subway Station

4. Distance to the Nearest Bus Station
The distance to bus stations is a continuous variable that indicates the distance from each school to the nearest bus station. It is calculated using the latitude and longitude of the bus stations and the longitude and latitude of the school.
Same as above, it is calculated using the Euclidean distance formula referred to in Equation 1. Xi is the school's geographic coordinates, and Yi is the geographic coordinates of the bus stations. The i in the formula is the number of observations, which is 2174. Using the algorithm (1) shown in the appendix, we could find the **nearest** distance.

According to Table 3 (below), the minimum distance is 0.000052 km, and the maximum is 0.05. The mean distance is 0.002 km. The median is 0.002 km. The median and mean are equal. In addition, the standard deviation is not high, which is equal to 0.002.

	Distance to bus station
Mean	0.220
Std	0.155
min	0.000
25%	0.080
50%	0.170
75%	0.250
max	1.520

Table 3 Summary Statistics for Distance to Nearest Bus Station

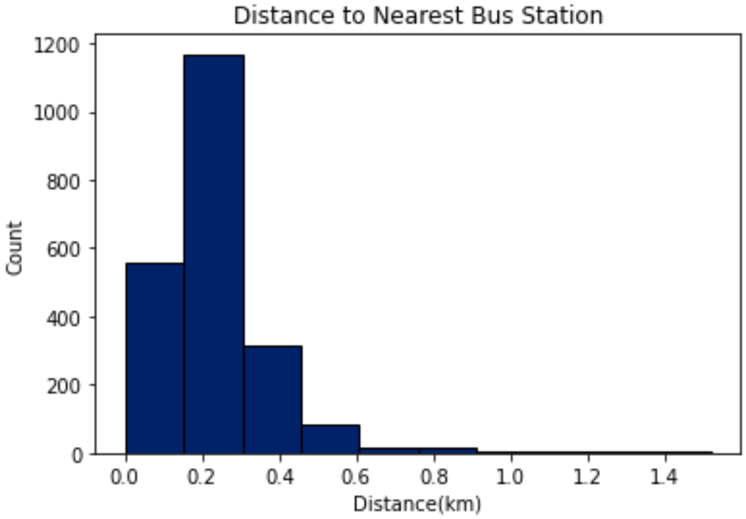


Figure 5 Distance to the Nearest Bus Station

According to the histogram Figure 4, All of the observations fall with 1.5km, with the most observations gathers around 0.2km.

5. Distance to the Nearest Parking Lot

The distance to parking is a continuous variable that indicates the distance from each school to the nearest public parking built and managed by the government. Similarly, it is calculated using geographic coordinates as well as the Euclidean formula. To find the nearest distance to the parking lot using equation one and algorithm 1, Xi is the school's geographic coordinates, and Yi is the geographic coordinates of the public parking. The unit of the distance is also kilometer.

	Distance to the parking lot
Mean	0.466
Std	0.892
min	0.000
25%	0.080
50%	0.080
75%	0.340
max	8.600

Table 4 Summary Statistics for Distance to the Parking Lot

The summary statistics shown in table 4 reveal a skewed distribution with median value of 0.08 km, which is much lower than the mean of 0.466 km. In addition, the maximum value of 1.112 is far from a minimum of 0.

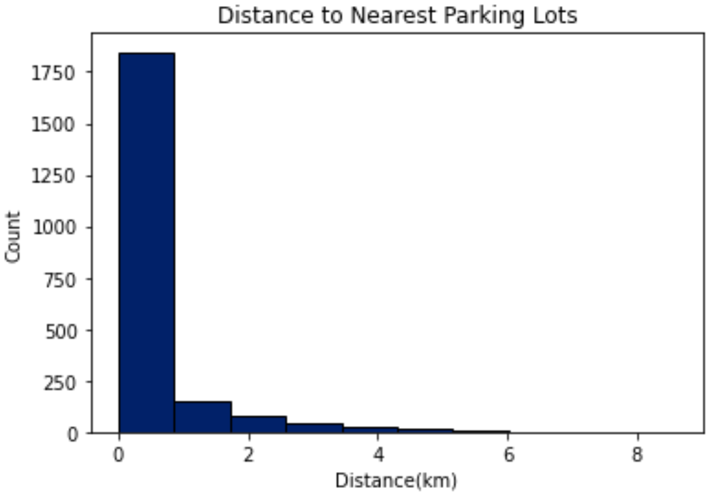


Figure 6 Distance to the Nearest Parking Lot

Consequently, we could observe that most of the observations gathered within 2km. Meanwhile, fewer observations fall outside 4km.

6. Distance to the Nearest Coach Station
The distance to parking is a continuous variable that indicates the distance from each school to the nearest coach station built and managed by the government. Similarly, we use the Euclidean formula (Equation 1) and algorithm 1, while Xi is the school's geographic coordinates, and Yi is the coach station's geographic coordinates.

From table 5, the mean value is 4.282 km, and the median is 3.040 km. The minimum value is around 0, while the maximum is 27.33km. That said, outliers may present in this distribution as well.

	Distance to Coach Station
Mean	4.282
Std	4.019
min	0.080
25%	1.520
50%	3.040
75%	5.650
max	27.330

Table 5 Summary Statistics of Distance to Coach Station

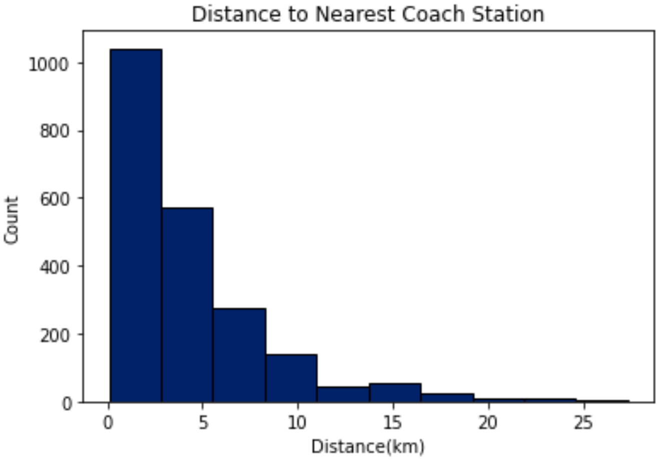


Figure 7 Distance to the Nearest Coach Station

According to figure 5, while most of the observations fall with a distance of 20km, there is a little peak, with much fewer observations presented within 15-18 km. Also, there are a few observations that are outside of 25km.

- **Limitation of the dataset**

After an overview and discussion of the dataset and based upon the research question, it would be more ideal if additional information regarding the school quality, the demographic characteristics of the districts that the schools are located in and average length of time for students to commute to work could be found.

3. Methodology

To identify some patterns among public schools within providing compulsory education[1]in Beijing and traffic facilities distribution in Beijing, China. We could apply clustering analysis to the features mentioned in the above dataset. Both k-means clustering, and hierarchical clustering are used.

- **K-means clustering**

K-means clustering aims to observations into clusters. Each observation belongs to the with the nearest cluster, serving as a prototype of the cluster. On GitHub, an analysis is trying to cluster around 1000 universities in UK based on features to predict if it is a private school or a public school. In other research, k-means are used to mine different groups of teachers and evaluate the teaching quality automatically (Sangita & Dhanamma, 2011). Therefore, we could also try to group these schools into groups and make improvement to infrastructure building based on the result.

Since we have a dataset with 2174 observations, we would intuitively choose K- mean clustering because it is friendly to a large dataset. However, with the summary presented above, the dataset needs further processing before using the K-means approach. First, since K-means clustering minimizes within-cluster variances, it might not work well with many outliers, presented in the above summary statistics of features used in the analysis. We would consider dropping the outlier when using K-mean clustering. Also, since K-means works better with equal cluster sizes and Gaussian distribution features, it might not work well on the features in our analysis, especially the distance to the nearest parking lots and the distance to the closest bus stations, with numerous observations gather in a specific range. In addition, we should also consider normalizing the datasets when using K-mean clustering, with two binary variables presented. Finally, given the context of our model, it would be harder to choose the k for K-means clustering analysis.

- **Hierarchical clustering**

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering has been previously used in numerous education related academic research. For example, in *Application of Hierarchical Cluster Analysis in Educational Research: Distinguishing between Transmissive and Constructivist Oriented Mathematics Teachers.* The author used hierarchical clustering to find out two groups(P&Q) of teachers from 30 questionnaires. The author concluded that the main difference is in students' activity in the classroom. In classrooms taught by the cluster Q teachers there is a space for individual activity of students and strong focus on applications of mathematics (Medova&Bakusova,2019). Another research, by using hierarchical clustering have grouped universities in EU into different groups according to features related to competitiveness. Similarly, In this insight report, we could group the public schools in Beijing according to features that could evaluate the level of public transports infrastructure building （Kabók, Radišić, & Kuzmanović, 2017).

Generally, there are two types of hierarchical clustering, agglomerative and divisive clustering. Agglomerative hierarchical clustering treats observations as leaves and fuses them into clusters based on similarity. However, divisive hierarchical clustering starts with one large cluster and splits the cluster until only observations remain. Also, different from K-mean clustering, it is not necessary to be prototyped based. Based on the dataset that we need to analyze; we could also choose the appropriate linkage method using graph-based methods. Generally, there are three types of linkage methods: single linkage, complete linkage, group average. While single linkage methods are sensitive to outliers due to its nature of Computes cluster proximity based on the two closest points in different clusters, complete linkage, and group average work well with datasets with or without noise. They Compute cluster proximity based on the average pairwise proximity for all points in the different clusters.

In the context of our project, a hierarchical clustering with the complete linkage or group average stands out since it is likely to have different cluster sizes and many outliers, according to the summary statistics shown in the data section of the memo. With little information told about the potential number of clusters from the data set and background reviews, it is an advantage to K-means clustering since we do not need to choose an initial k value. However, the large sample size could be computational expensive using hierarchical clustering. Also, it might be hard to visualize using a dendrogram.

4. Analysis

By applying both K-means and hierarchical clustering to our dataset, different results are observed. In addition, evaluation and comparison of two results are performed using plots and silhouette coefficient.

- **K-means clustering**

In the clustering analysis, six features in the dataset are included. First, we have chosen a range of K for applying the k-means clustering. Since there is no indicator or any previous background information regarding potential number of clusters. I would choose k equals to 1 to 10. In order to decide the optimal number of clusters for k-means clustering, we would use the elbow method. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of *k* (say, *k* from 1 to 10), and for each value of *k* calculate the sum of squared errors (SSE). Our goal is to choose a small value of *k* that still has a low SSE, and the elbow usually represents where we start to have diminished returns by increasing *K*. Figure 6 shows that the optimal K would be around 3, where the slope of the line begin to be changing smoother. Therefore, by looking at the elbow plot, K=3 is the optimal choice, while the sum of square errors is about 6300, which is still relatively large.

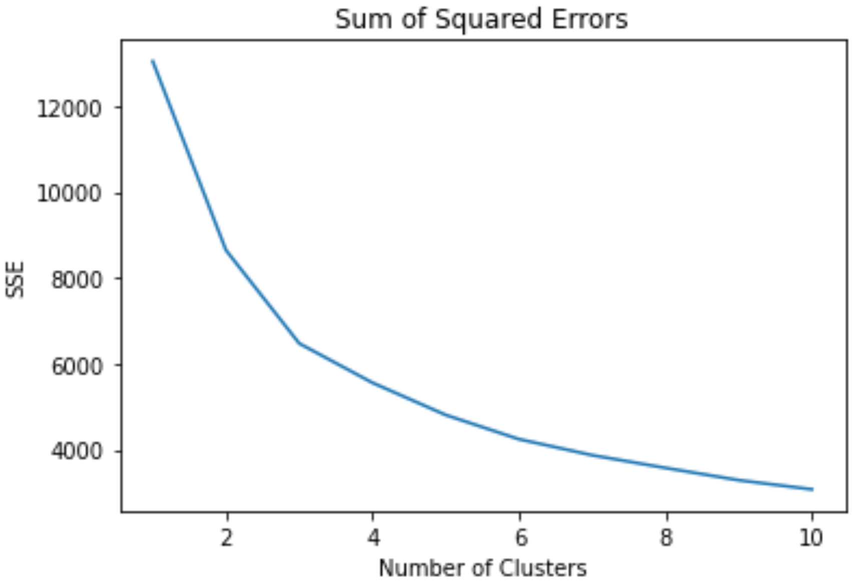


Figure 8 Elbow Plot

Further, another method could be used to validate the optimal k as well as how well it is clustering, we could plot calculate the Silhouette score for K ranging from two to ten. We could choose the K number that has the optimal silhouette score. The range of the silhouette score is -1 to 1. The silhouette score of 1 means that the clusters are very dense and nicely separated. The score of 0 means that clusters are overlapping. The score of less than 0 means that data belonging to clusters may be wrong/incorrect. Figure 7 shows that when k=6, the silhouette coefficient is the closest to 1. Meanwhile, when k equals to 3, 4 or 5. The silhouette coefficient is also closer to 1, comparing to other value of k.

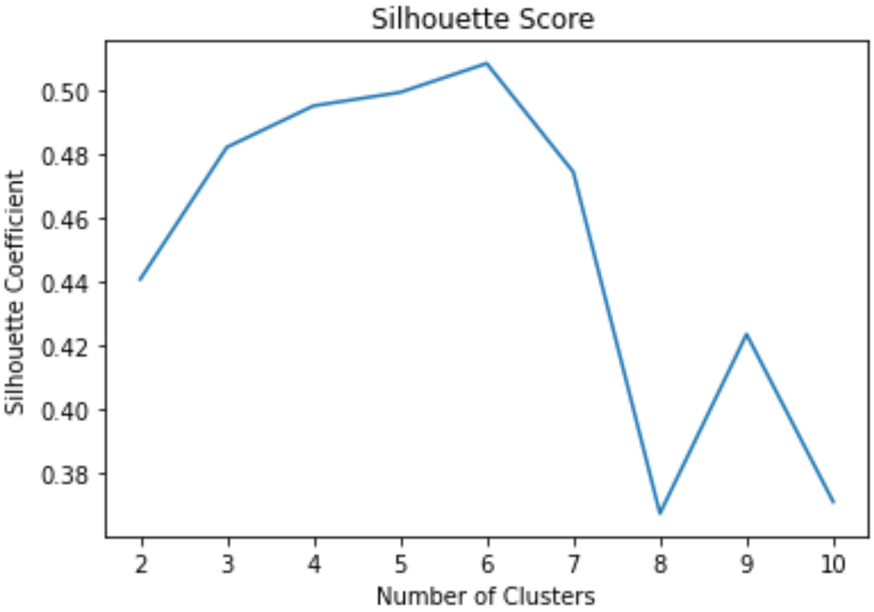


Figure 9 Silhouette Score Plot

Deciding upon the results of both elbow plot and the silhouette coefficient, I will choose K=3 to perform the K-means clustering. Although according to the silhouette coefficient, when k equals to 6, the silhouette coefficient reaches the highest, yet the elbow plot suggests that if we choose k equals to 6, it might run into risks of overfitting. Meanwhile, when k equals to 3, the silhouetee score is around 0.5, which is only slightly lower than that of k equals to *6.*

- **Hierarchical Clustering**

As mentioned in the method section of the report, the linkage method that would be most appropriate would be the complete linkage method because of the outliers. To validate the choice of the k, we could use the dendrogram to identify the potential k values. From the dendrogram, we could choose the K value to be 4,6 or 10. However, since the dataset is relatively large regarding the hierarchical clustering analysis, we could tell little information from the dendrogram, especially the hierarchical structures of the clustering.

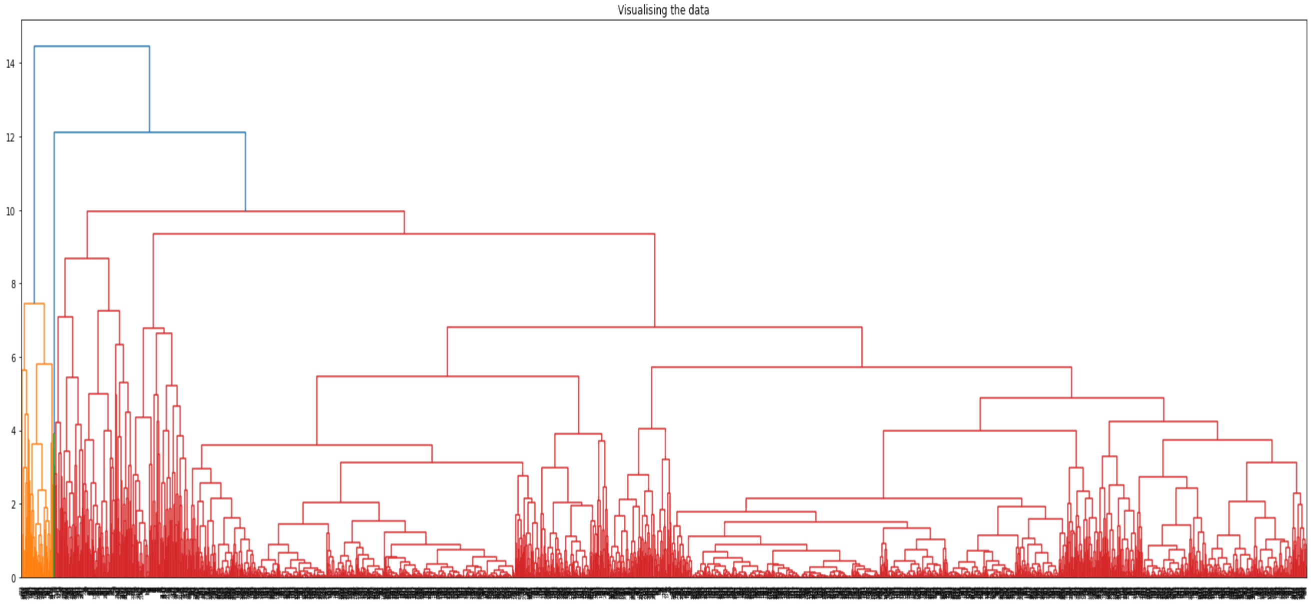


Figure 10 Dendrogram

To further validate the optimal value of k as well as evaluate the performance of the hierarchical clustering analysis, we could also use the silhouette score. From the plot, we could observe that when K=10, the silhouette coefficient is closest to 1.

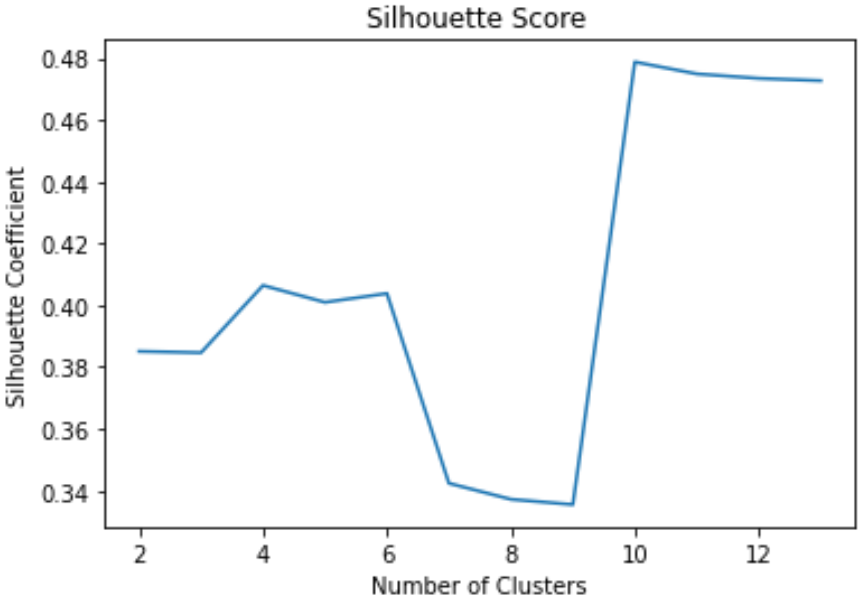


Figure 11 Silhouette Score

Finally, when comparing K-means clustering to hierarchical clustering using the silhouette score, the k-means with K=3 has a slightly higher score than K=10 using the hierarchical clustering. Also, since there are only 6 indicators, the cluster number of 3 would be more meaningful than a k value of 10 under this setting. K=10 would be hardly meaningful in this setting. Meanwhile, when K=3, the result is interpretable and meaningful under domain expertise. A discussion of the reasons in detail will be include in the next section regarding how K=3 is more meaningful.

5. Results and Insights

- **Results**

	Primary School	**Middle School**	**Distance to Subway Station**	**Distance to Bus Station**	**Distance to Parking Lots**	**Distance to Coach Station**
cluster 0	0.86	0.14	35.17	0.302	2.078	9.191
cluster 1	0	1	5.4	0.200	0.233	3.668
cluster 2	1	0	3.5	0.212	0.223	3.480

Table 6 Result for K-means clustering

Table 6 is the results of the k-means clustering. Three clusters are generated. According to the result, cluster 0 is mostly consist of elementary schools. The distance to the subway station, parking lots as well as coach stations are all relatively higher than the other two clusters. Cluster 1 and cluster 2 has similar level regarding the distance to the public transport facilities, while cluster 1 is middle school and cluster 2 is all elementary school.

	Primary School	**Middle School**	**Distance to Subway Station**	**Distance to Bus Station**	**Distance to Parking Lots**	**Distance to Coach Station**
Cluster 0	0.530	0.469	21.883	0.251	1.497	15.930
Cluster 1	0.923	0.077	4.713	0.661	1.170	6.060
Cluster 2	1	0	61.86	0.357	2.238	3.857
Cluster 3	1	0	4.797	0.203	0.295	3.774
Cluster 4	0	1	3.370	0.198	0.195	3.244
Cluster 5	1	0	29.553	0.223	7.223	21.203
Cluster 6	0.867	0.133	18.036	1.104	0.877	6.473
Cluster 7	0.556	0.444	72.778	0.150	0.748	3.946
Cluster 8	0.571	0.429	76.412	0.221	5.103	6.520
Cluster 9	0.559	0.441	14.239	0.202	3.717	5.450

Table 7 Result of Hierarchical Clustering

However, the result of hierarchical clustering shows little valuable patterns regarding schools and public transports.

- **Insights**

	School count	**proportion**
cluster 0	280	0.123
cluster 1	766	0.353
cluster 2	1127	0.519

Table 8 Result of K-means: School Count & Proportion

Furthermore, table 8 indicates that most of the schools has a commutable distance to public transports. The problematic schools in cluster 0 only takes up 12.3% of the total public schools in Beijing. We could look into these 280 schools.

Lastly, when we looked into the districts that the schools belong to, 279 out of 280 of them located in rural districts of Beijing: Huairou, Yanqing, Miyun, Yanqing, Pinggu, Shunyi, Changping, Fangshan, Mentougou and Daxing. Only one of the schools falls in the Haidian district, that is within the urban area of Beijing. This result is consistent with situation in Beijing, where there is a gap between the urban and rural part of Beijing regarding infrastructure Building.

6. Conclusion

To conclude, most of the public transports around public school are reasonable allocated, with a few schools, especially, elementary schools that needs improvement in terms of public transports accessibility in terms of subway stations, parking lots and coach stations. The following recommendation are proposed to the government of Beijing in the long term sustainable urban planning in terms of promoting education equality.

1. Focusing on building public transports around public schools in the rural districts of Beijing, especially subway stations and parking lots. The government should consider focusing its infrastructure building around the 280 schools in the cluster 0, where there is a relatively poor access to subway station, parking lots and coach stations. Poor access to subway stations could be problematic for students' daily travel. That is, how to commute from home to school. In addition, parking lots are essential for public transports around elementary schools, where the students are mostly uncapable of commuting to school on their own. Finally, the coach station, could be beneficial to connect to museums, stadiums as well as going on field trips, which are all meaningful parts of promoting education equality.
2. In addition, building parking lots around elementary schools among the 280 schools should be prioritize in the short run. As the result in table 7 indicates, among the 280 schools, there are more elementary schools than middle school. which is more problematic in terms of inaccessible public transports for pupils.

7. Limitation

Technical support on calculating the walking distance is needed for more accurate results. In this research, we have used the Euclidean formula to calculate the distance between two geographical points. However, in the reality, we should take into consideration the distance of the walking route one should undertake to travel between two points. In many situations, even though two places are close together on the map, one could not travel directly, but need to take a longer route instead.

Bibliography

ChrisWoodard43. (n.d.). ChrisWoodard43/KMeans-Universities. Retrieved December 15, 2020, from https://github.com/ChrisWoodard43/KMeans-Universities

Kabók, J., Radišić, S., & Kuzmanović, B. (2017). Cluster analysis of higher-education competitiveness in selected European countries. *Economic Research-Ekonomska Istraživanja,30*(1), 845-857. doi:10.1080/1331677x.2017.1305783

Medova,J., Bakusova,Jana(2019), Application of Hierarchical Cluster Analysis in Educational Research: Distinguishing between Transmissive and Constructivist Oriented Mathematics Teachers. Statistika-Statistics and Economy Journal, WOSUID: WOS:000472577100003

Sangita, O., & Dhanamma, J. (2011). An Improved K-Means Clustering Approach for Teaching Evaluation. *Communications in Computer and Information Science Advances in Computing, Communication and Control,*108-115. doi:10.1007/978-3-642-18440-6_13

Implementation Appendix

- **Preprocessing the data**

a. Step1: please refer to project_clean.ipynb. (ouput_data.xlsx are generated from the original data)

1. In order to calculate the Euclidean distance, we need to combine the latitude and longitude into geographic coordinates. Then, we need to define a function that could calculate the distance using the Euclidean formula. Finally, we could apply it to the education and all four types of public transports.
2. In order to get the nearest distance, please refer to #calculate the nearest distance in the last section of project_clean.ipynb. By using this algorithm, we could calculate the nearest distance to the school from any given public transport's.
3. To use the k-means clustering, we need to generate binary numerical variable for public school type. For these part, please refer to #generate binary variable

b. Step 2: Generate book5.csv from output.xlsx

4. Drop features that is not useful in our analysis. That is, 'name', 'district', 'kind 1', 'kind 2', 'kind 3' and 'Jaccard_sim'.
5. Drop observations that are neither public elementary school nor publi middle school by using sort function of excel.
6. The name of final data set is book5.csv

- **For visualizing the data, please refer to project_memo2.ipynb**
- **For analysis, please refer to project_jw1803.ipynb**