Segmenting Hong Kong MTR Station Areas

1. Introduction

Hong Kong is a major Asian destination for tourists. On average, more than 60 million tourists visit Hong Kong every year. When tourists visit Hong Kong, the most commonly used public transportation is MTR (Mass Transit Railway). The MTR system in Hong Kong currently has 95 heavy rail stations as of March 2020, and the daily ridership is around five million.

In this project, I would like to provide some guidance to tourists who are not very familiar with Hong Kong but may plan to visit it. As tourists are likely to make use of the MTR system to move around when they are in Hong Kong, I use MTR stations to represent different areas in town. I segment MTR station areas and explore the top venue categories in each area in Hong Kong. The segmentation analysis will be interpreted together with the population density of different districts in Hong Kong. Such information is helpful for tourists to plan their trip in Hong Kong.

2. Data

2.1. List of MTR stations

A table of MTR heavy rail stations is available on Wikipedia (https://en.wikipedia.org/wiki/List_of_MTR_stations). However, this table of MTR stations has duplicates because some stations are in multiple lines. In addition, if a station used a different name in the past, the table will show the current and the old station names in the same cell. Therefore, I cleaned the table scraped from Wikipedia to make sure that it only shows the 95 current heavy rail stations of the MTR system in Hong Kong. Only the columns useful for the analysis (e.g., station name, district) were retained.

2.2. Geolocations of MTR stations

The geolocations (latitudes and longitudes) of the MTR stations were fetched with ArcGIS. ArcGIS is a geographic information system for working with maps and geographic information maintained by Esri. These geolocation data were merged into the table of MTR stations.

2.3. Venues in station areas

The Foursquare API was used to explore the station areas. At maximum 200 venues within the 500-meter radius of each station were fetched.

2.4. List of districts in Hong Kong

As I also planned to visualize the population density in Hong Kong, a list of Hong Kong districts was also be scraped from Wikipedia (https://en.wikipedia.org/wiki/Districts_of_Hong_Kong).

A dataframe was created to show the population densities of the 18 districts in Hong Kong. Only the columns useful for analysis (e.g., district, density) will be retained.

2.5. Geolocations of Hong Kong districts

As I planned to use choropleth to visualize the population densities in different Hong Kong districts, I used the GeoJSON file provided by Esri (http://opendata.esrichina.hk/datasets/eea8ff2f12b145f7b33c4eef4f045513_0). Using this GeoJSON file, I was able to create a map of Hong Kong that shows the boundaries of different districts.

3. Methodology

In this project, I primarily employed cluster analysis to segment the different MTR station areas in Hong Kong. The MTR station areas were clustered and segmented based on the similarities and differences in the venues around them.

Specifically, I first fetched the geolocations (i.e., latitudes and longitudes) of the MTR stations in Hong Kong. Then, I used the Foursquare API to explore the venues around each MTR station within a radius of 500 meters. At maximum, 200 venues were fetched for each MTR station.

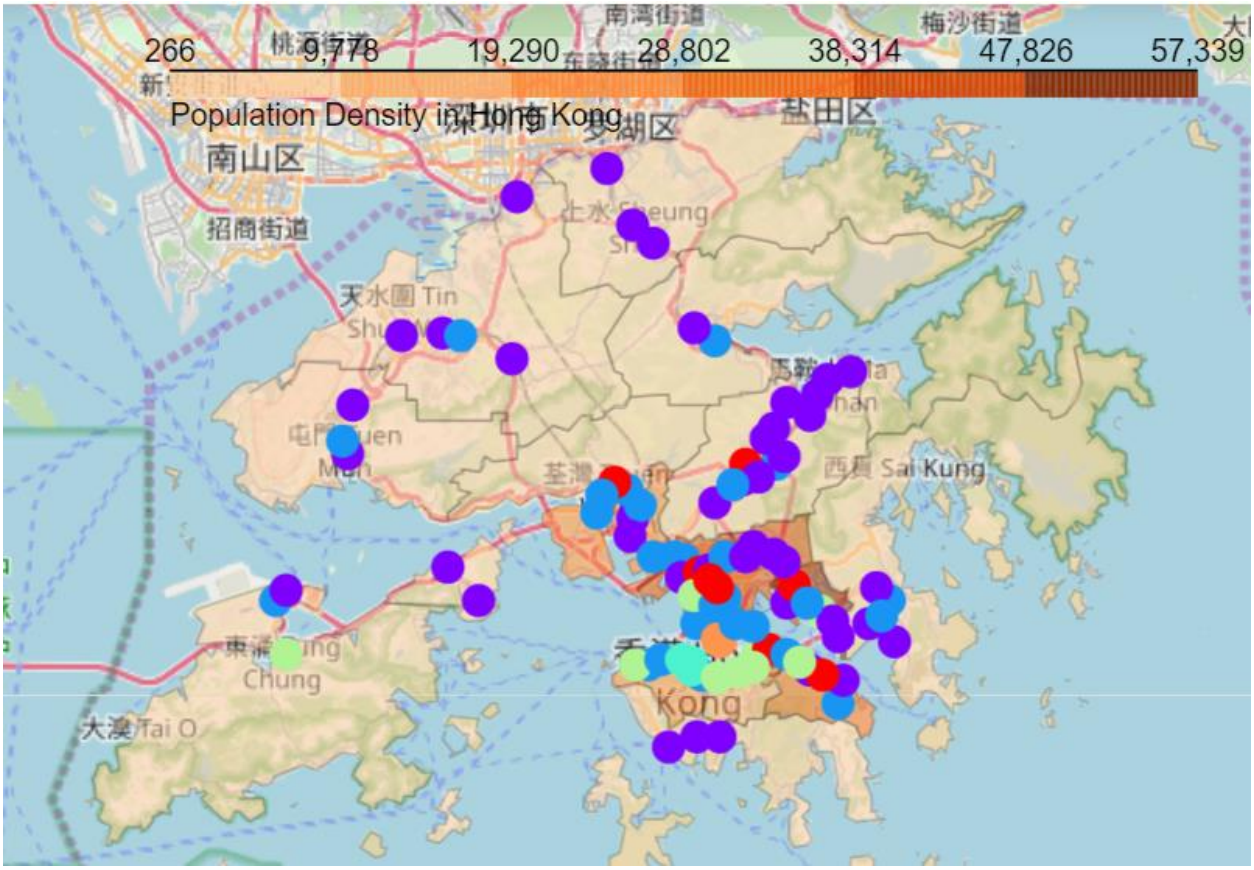
After the venues around each MTR station were fetched, the frequencies of each venue categories were calculated. Next, k-means clustering was performed to segment the MTR station areas using the frequencies of different venue categories. As an unsupervised machine learning method, k-means clustering categorizes different objects by minimizing within-cluster variances. Therefore, the more similar the venues were between two station areas, the more likely they were in the same cluster. In total, six clusters were generated, and the top five venue categories were examined, which was used as a basis to label the clusters.

In addition, to complement the analysis based on k-means clustering, I also looked at the population density in each of the 18 districts in Hong Kong. It is possible that two station areas consist of similar venues, but they might be serving different types of customers. For example, both a station near CBD and one near downtown residential areas may have many restaurants around them, but the restaurants in the former case may be more focused on the high-end market while those in the later case may serve more affordable food. Thus, the interpretations of the k-means clustering analysis can be enriched by the additional consideration of population density.

4. Results

Six clusters of MTR station areas were generated. To facilitate the interpretation of the clustering results, the clusters were shown on a choropleth map of Hong Kong based on population density (population/km^2; Figure 1). Districts with darker colors are more populated.

Figure 1. Hong Kong Map of MTR Station Area Clusters



Next, I explored the characteristics of each cluster and labeled them accordingly.

4.1. Cluster 1

Cluster 1 has 10 station areas, indicated by the red dots in Figure 1. They are North Point Station, Shau Kei Wan Station, Kowloon Bay Station, Sha Tin Station, Sham Shui Po Station, Tsuen Wan Station, Austin Station, Mong Kok Station, Prince Edward Station, and Mong Kong East Station. The five venue categories with highest mean frequencies in this cluster are shown in Table 1.

Table 1. The Five Most Common Venue Categories in Cluster 1

Venue Category	Mean Frequency
Noodle house	3.70
Chinese restaurant	3.40

Dessert soup	3.10
Cha chaan teng (i.e., local diners)	2.40
Shopping mall	2.10

Almost all the station areas in cluster 1 are in the more densely populated districts of Hong Kong. The most frequent venue categories are mainly restaurants that primarily target local residents, such as noodle houses, Chinese restaurants, dessert soup houses and Cha Chaan Teng (i.e., local diners). The shopping malls in these densely populated MTR station areas are also more focused on serving the nearby residents. Therefore, I label this cluster as *Local Dining Spots*.

4.2. Cluster 2

Cluster 2 is the largest cluster and contains 43 station areas all over different districts in Hong Kong. Representative station areas include Heng Fa Chuen Station, Sai Wan Ho Station, Tiu Keng Leng Station, Ma On Shan Station, Racecourse Station, Sha Tin Wai Station, and Lei Tung Station. Cluster 2 are indicated by the purple dots in Figure 1. The five venue categories with the highest mean frequencies in Cluster 2 are shown in Table 2.

Table 2. The Five Most Common Venue Categories in Cluster 2

Venue Category	Mean Frequency
Fast food restaurant	1.26
Chinese restaurant	0.88
Shopping mall	0.60
Coffee shop	0.56
Bus station	0.47

Cluster 2 mainly consists of station areas in medium to low population density districts. The venues in this cluster also seem to mainly target customers who are on the go, such as fast food

restaurants. Interestingly, this cluster also includes several transportation hubs, such as Lo Wu Station and Lok Ma Chau Station that are on the border with Shenzhen. Coffee shops and bus stations are common in this cluster to serve both travelers and local residents. Therefore, I label this cluster as *Transit Spots*.

4.3. Cluster 3

Cluster 3 is the second largest cluster, consisting of 28 MTR station areas, such as Quarry Bay Station, Kowloon Tong Station, and Tsuen Wan West Station. They are indicated by the blue dots in Figure 1. The five venue categories with the highest mean frequencies in Cluster 3 are shown in Table 3.

Table 3. The Five Most Common Venue Categories in Cluster 3

Venue Category	Mean Frequency
Chinese restaurant	2.18
Coffee shop	1.96
Café	1.54
Fast food restaurant	1.36
Cantonese restaurant	1.14

Cluster 3 is similar to Cluster 2 in that both of them contain quite a lot of Chinese restaurants and fast food restaurants. In addition, Cluster 3 are mainly in residential areas. The main distinction between these two clusters is that Cluster 3 has more coffee shops and cafés in them. Therefore, I label this cluster as *Local Café Spots*.

4.4. Cluster 4

Cluster 4 has three stations in it, namely, Central Station, Hong Kong Station, and Sheung Wan Station. These three stations are close to each other. They are indicated by the turquoise dots in Figure 1. The five venue categories with the highest mean frequencies in Cluster 4 are shown in Table 4.

Table 4. The Five Most Common Venue Categories in Cluster 4

Venue Category	Mean Frequency
Japanese restaurant	4.67
Coffee shop	4.67
Cocktail bar	4.67
Café	4.00
Bar	4.00

As seen from Table 4, Cluster 4 has quite a few places for people to mingle and socialize, such as coffee shops and bars. This is likely due to the fact that these three station areas are located at the business and financial center of the city. Therefore, I label this cluster as *CBD Social Spots*.

4.5. Cluster 5

Cluster 5 consists of nine MTR station areas, most of which are located on the Hong Kong Island. They are indicated by the green dots in Figure 1. The five venue categories with the highest mean frequencies in Cluster 5 are shown in Table 5.

Table 5. The Five Most Common Venue Categories in Cluster 5

Venue Category	Frequency
Coffee shop	4.78
Japanese restaurant	3.78
Café	3.44
Chinese restaurant	3.44
Hotel	2.56

Cluster 5 is similar to Cluster 4 in that both contain quite a few Japanese restaurants and coffee shops. The major difference is that Cluster 5 also have many hotels. Indeed, many travelers choose to stay in these areas because of their convenient locations. Therefore, I label this cluster as *Convenient Accommodation Spots*.

4.6. Cluster 6

Cluster 6 only contains two stations that are very close to each other, that is, Tsim Sha Tsui Station and East Tsim Sha Tsui Station. They are indicated by the orange dots in Figure 1. The five venue categories with the highest mean frequencies in Cluster 6 are shown in Table 6.

Table 6. The Five Most Common Venue Categories in Cluster 6

Venue Category	Frequency
Hotel	9.00
Japanese restaurant	8.00
Hotel bar	5.00
Steakhouse	4.00
Shopping mall	3.50

Cluster 6 is the Tsim Sha Tsui area in Hong Kong and on the Victoria Harbor. It is popular among tourists, which may the reason why it has many hotels in it, and also serve food that may cater to the taste buds of foreign tourists. Therefore, I label this cluster as *Central Touristy Spots*.

5. Discussion

Using k-means clustering analysis together with the visualization of population density in different districts in Hong Kong, I identified six clusters of MTR station areas in Hong Kong, that is, *Local Dining Spots*, *Transit Spots*, *Local Café Spots*, *CBD Social Spots*, *Convenient Accommodation Spots*, and *Central Touristy Spots*. *Local Dining Spots* mainly consist of MTR station areas that have many restaurants or eateries that serve local residents, so visitors who are

interested in authentic local food at affordable prices should visit these station areas. *Transit Spots* consist of MTR station areas that have restaurants and cafés that target customers on the go. Visitors will find it easy to grab a bite quickly if they happen to visit these areas on the way to other places. MTR station areas in the cluster of *Local Café Spots* have many cafés that mainly serve local customers, so visitors to these places can enjoy coffee at reasonable prices. In comparison, there are more high-end restaurants and cafés in *CBD Social Spots*, so visitors can visit these places if they prefer something fancier. If visitors wish to stay at a convenient location, they may consider the station areas in the clusters of *Convenient Accommodation Spots* and *Central Touristy Spots*, and the latter is more frequently visited by the former because there are more tourist attractions in the latter.

6. Conclusion

Visitors to Hong Kong often need to use the MTR system to move around in the city. In this report, six clusters of MTR station areas were identified through k-means clustering analysis and their characteristics were discussed. These results will help visitors to Hong Kong to better plan their stay in Hong Kong using the MTR system.

In the future, the clustering analysis can be further improved. The current analysis mainly used the similarities and differences in venue categories as the basis for clustering. Population density was taken as an additional consideration when the results of clustering were interpreted. Future analysis may try more focused clustering analysis. For example, the different MTR station areas can be clustered based on the density of sports facilities or the density of parks in order to provide more tailored recommendations for travelers who are interested in visiting areas with particular characteristics.