Coursera IBM

Data Science Professional Certificate Capstone Project

Topic: Investigate the optimal location for setting up a new café business along the Ma On Shan Rail in Hong Kong

Prepared by: Lawrence Lee

Date: 22 May, 2020

1. Introduction

The Ma On Shan Rail in Hong Kong is a rapid transit line that serving the new towns of Shatin and Ma On Shan in the northeastern N.T. The ridership is 153,100 weekday average in 2014 research figure. Due to the extension of Ma On Shan Rail to Kai Tak station on 14 Feb 2020, it is predicted that the ridership in Ma On Shan Rail will be sharply increased. As a result, venues near the railway stations in Ma On Shan lines are suitable for small business to setup as more people will pass by those area.

2. Business Problem

In order to find the suitable location and type of business to start with, it is important to grab the venues data for the places nearby for further analysis. Recently, many Hongkongers like to go for café for meal or chatting with friends. The cafés are not only treated as "café" and they always offer entertainment services like karaoke, boardgames and books for reading. It is world to invest a café as a start up business as the capital involved is low. The main objective of this project is to find out the ideal location near the Ma On Shan Line for starting a new café business. The locations with more restaurants but less café will be selected as business location.

3. Data

The data for this project had been processed with different sources and ensured the method of analysis is accurate.

3.1 MTR Station

The MTR Station data of Ma On Shan line will be scraped from a Wikipedia webpage. AS there are only 9 stations in Ma On Shan line, the data will be directly saved and written to a list without using BeautifulSoup library in Python.

3.2 Geocoding

For the geometric data like the latitude and longitude of each of the MTR stations, Google Maps Geocoding API will be used to extract the data and those dataframe will be placed into the station dataframe.

3.3 Venue Data

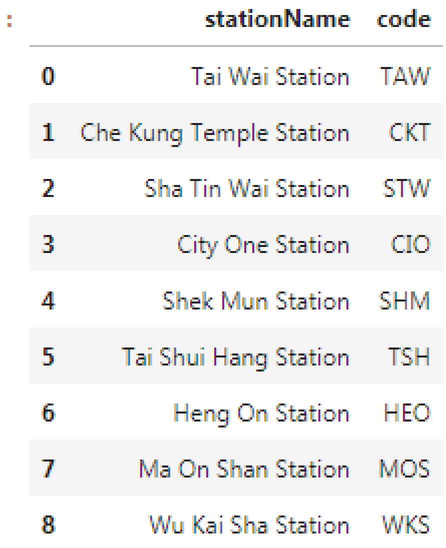
After the location data is achieved from geocoding and Wikipedia, the FourSquare API will be used to find out the nearby venues and create a new dataframe to involve all the venue data nearby the MTR station. It is assumed that maximum 30 nearby venues to be explored for each MTR Stations and the radius of exploration is 1000m.

4. Methodology

The method, rules and procedures used in this data analysis project will be explained through this section.

4.1 MTR Station Data

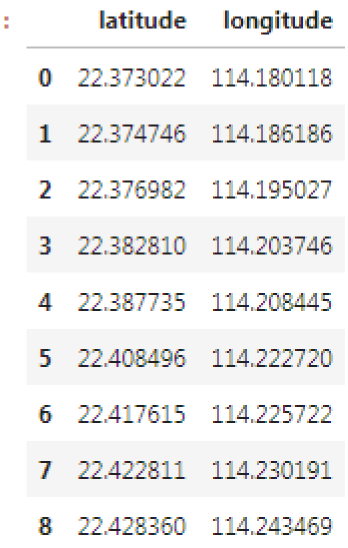
The MTR Station data of Ma On Shan line were scraped from a Wikipedia webpage and store in Panda dataframe.



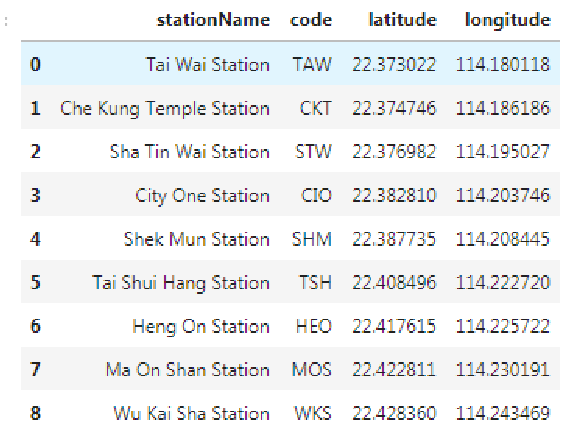
	stationName	code
0	Tai Wai Station	TAW
1	Che Kung Temple Station	CKT
2	Sha Tin Wai Station	STW
3	City One Station	CIO
4	Shek Mun Station	SHM
5	Tai Shui Hang Station	TSH
6	Heng On Station	HEO
7	Ma On Shan Station	MOS
8	Wu Kai Sha Station	WKS

4.2 Geocoding API

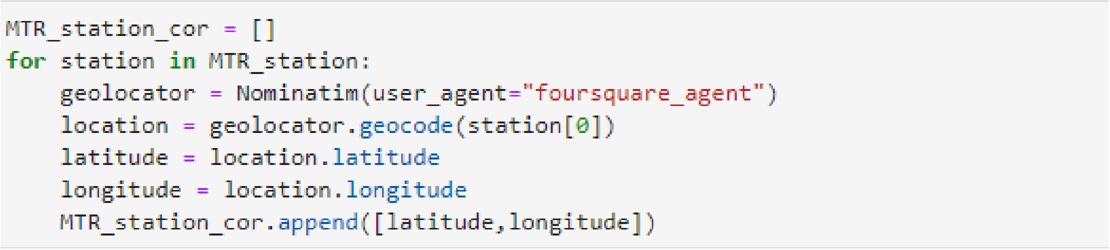
Google Map Geocoding API was used to obtain the geometric data of 9 nos. of MTR Stations in Ma On Shan Line.



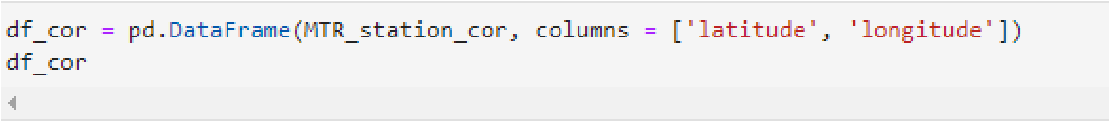
	latitude	longitude
0	22.373022	114.180118
1	22.374746	114.186186
2	22.376982	114.195027
3	22.382810	114.203746
4	22.387735	114.208445
5	22.408496	114.222720
6	22.417615	114.225722
7	22.422811	114.230191
8	22.428360	114.243469



	stationName	code	latitude	longitude
0	Tai Wai Station	TAW	22.373022	114.180118
1	Che Kung Temple Station	CKT	22.374746	114.186186
2	Sha Tin Wai Station	STW	22.376982	114.195027
3	City One Station	CIO	22.382810	114.203746
4	Shek Mun Station	SHM	22.387735	114.208445
5	Tai Shui Hang Station	TSH	22.408496	114.222720
6	Heng On Station	HEO	22.417615	114.225722
7	Ma On Shan Station	MOS	22.422811	114.230191
8	Wu Kai Sha Station	WKS	22.428360	114.243469



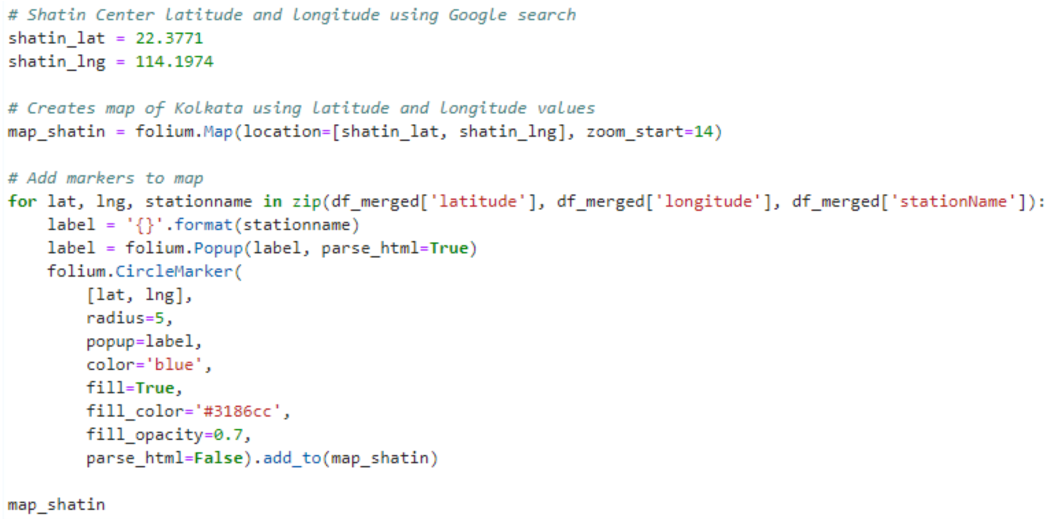
```
MTR_station_cor = []
for station in MTR_station:
    geolocator = Nominatim(user_agent="foursquare_agent")
    location = geolocator.geocode(station[0])
    latitude = location.latitude
    longitude = location.longitude
    MTR_station_cor.append([latitude,longitude])
```



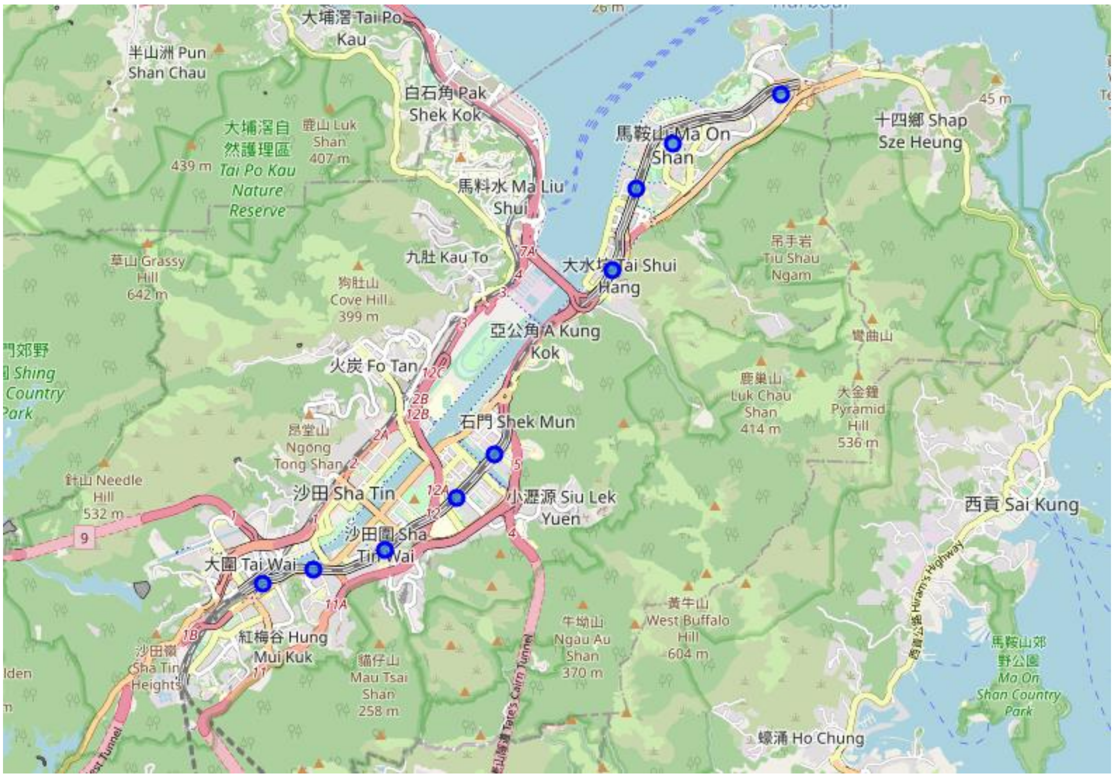
```
df_cor = pd.DataFrame(MTR_station_cor, columns = ['latitude', 'longitude'])
df_cor
```

4.3 Folium

Folium is used as the tool for visualization of the location of different MTR stations. At the beginning of plotting of Folium Map, assumed the center point to be taken as Shatin Center (22.3771,114.1974).

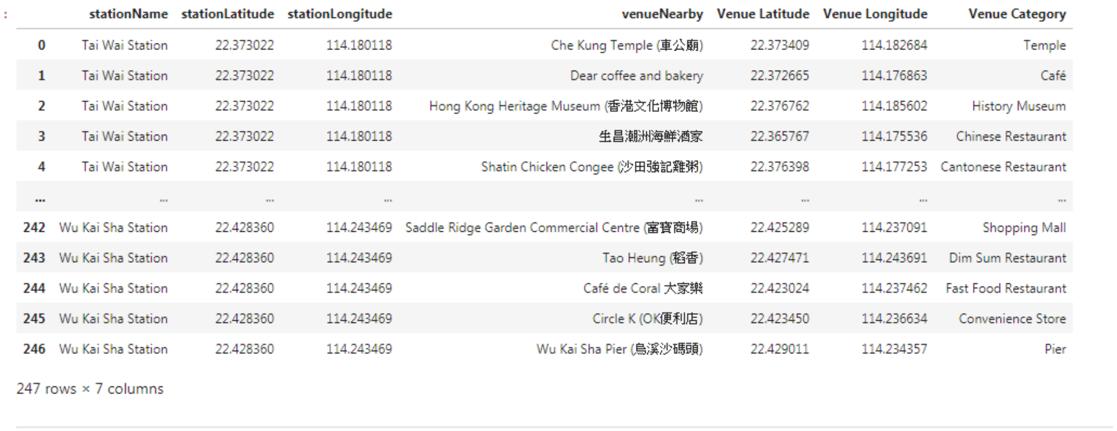


```
# Shatin Center latitude and longitude using Google search
shatin_lat = 22.3771
shatin_lng = 114.1974

# Creates map of Kolkata using latitude and longitude values
map_shatin = folium.Map(location=[shatin_lat, shatin_lng], zoom_start=14)

# Add markers to map
for lat, lng, stationname in zip(df_merged['latitude'], df_merged['longitude'], df_merged['stationName']):
    label = '{}'.format(stationname)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_shatin)

map_shatin
```



4.4 Venue Data

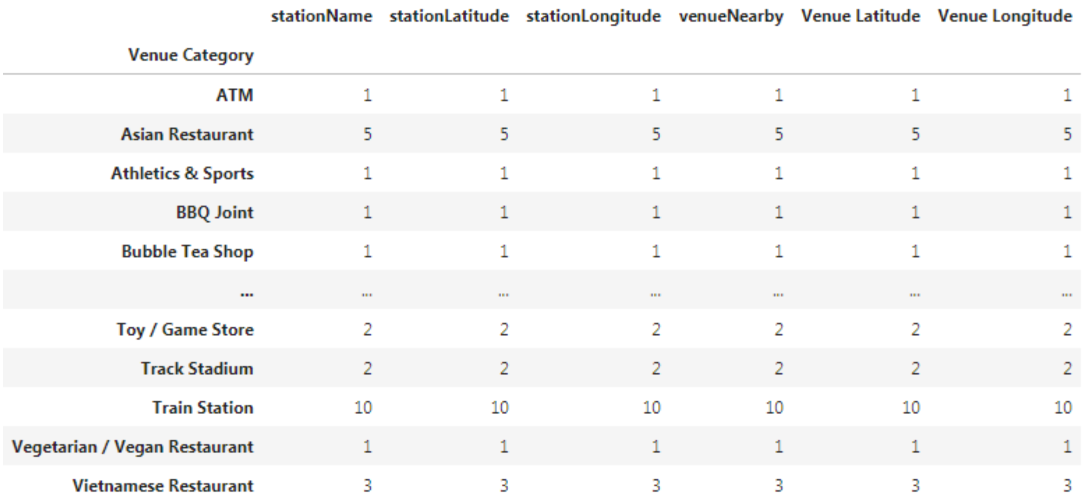
For each of the MTR stations along the Ma On Shan Line, a maximum of 30 venues were set to explore near the station with a radius of 1000m. As the distance between each of station is 2000m in average, setting 1000m as our exploration radius is reasonable. As a result, a total no. of 247 venues were found near the 9 nos. of station.



	stationName	stationLatitude	stationLongitude	venueNearby	Venue Latitude	Venue Longitude	Venue Category
0	Tai Wai Station	22.373022	114.180118	Che Kung Temple (車公廟)	22.373409	114.182684	Temple
1	Tai Wai Station	22.373022	114.180118	Dear coffee and bakery	22.372665	114.176863	Café
2	Tai Wai Station	22.373022	114.180118	Hong Kong Heritage Museum (香港文化博物館)	22.376762	114.185602	History Museum
3	Tai Wai Station	22.373022	114.180118	生昌潮洲海鮮酒家	22.365767	114.175536	Chinese Restaurant
4	Tai Wai Station	22.373022	114.180118	Shatin Chicken Congee (沙田強記雞粥)	22.376398	114.177253	Cantonese Restaurant
...	...	...	...	...	...	...	...
242	Wu Kai Sha Station	22.428360	114.243469	Saddle Ridge Garden Commercial Centre (富寶商場)	22.425289	114.237091	Shopping Mall
243	Wu Kai Sha Station	22.428360	114.243469	Tao Heung (稻香)	22.427471	114.243691	Dim Sum Restaurant
244	Wu Kai Sha Station	22.428360	114.243469	Café de Coral 大家樂	22.423024	114.237462	Fast Food Restaurant
245	Wu Kai Sha Station	22.428360	114.243469	Circle K (OK便利店)	22.423450	114.236634	Convenience Store
246	Wu Kai Sha Station	22.428360	114.243469	Wu Kai Sha Pier (烏溪沙碼頭)	22.429011	114.234357	Pier

247 rows × 7 columns

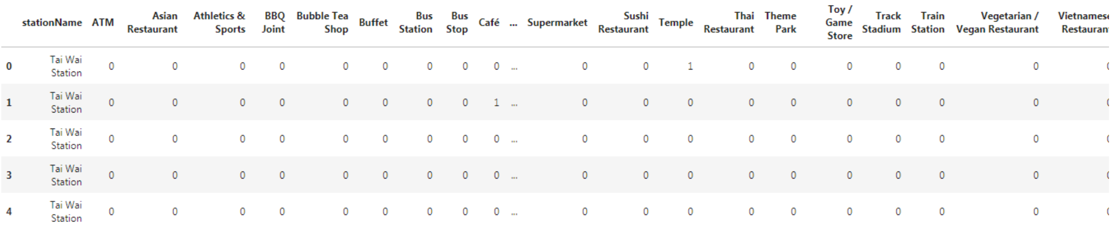
The data were also grouped by venue and counted each of the venue category.



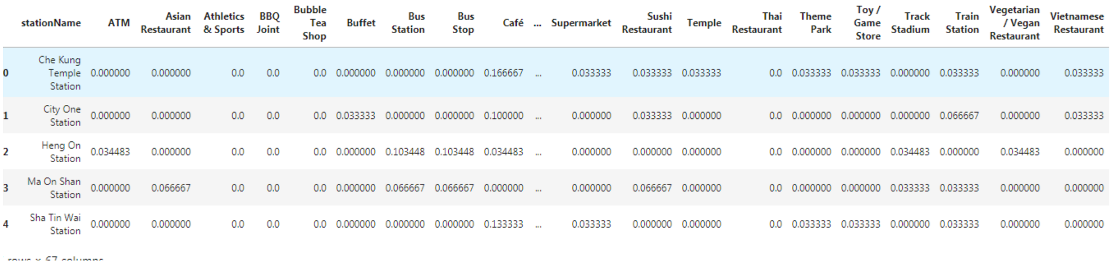
	stationName	stationLatitude	stationLongitude	venueNearby	Venue Latitude	Venue Longitude
Venue Category						
ATM	1	1	1	1	1	1
Asian Restaurant	5	5	5	5	5	5
Athletics & Sports	1	1	1	1	1	1
BBQ Joint	1	1	1	1	1	1
Bubble Tea Shop	1	1	1	1	1	1
...	...	...	...	...	...	...
Toy / Game Store	2	2	2	2	2	2
Track Stadium	2	2	2	2	2	2
Train Station	10	10	10	10	10	10
Vegetarian / Vegan Restaurant	1	1	1	1	1	1
Vietnamese Restaurant	3	3	3	3	3	3

4.5 One hot encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. In this analysis, k-means clustering algorithm will be followed after the one hot encoding.



	stationName	ATM	Asian Restaurant	Athletics & Sports	BBQ Joint	Bubble Tea Shop	Buffet	Bus Station	Bus Stop	Café	...	Supermarket	Sushi Restaurant	Temple	Thai Restaurant	Theme Park	Toy / Game Store	Track Stadium	Train Station	Vegetarian / Vegan Restaurant	Vietnames Restauran
0	Tai Wai Station	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	(
1	Tai Wai Station	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	(
2	Tai Wai Station	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	(
3	Tai Wai Station	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	(
4	Tai Wai Station	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	(

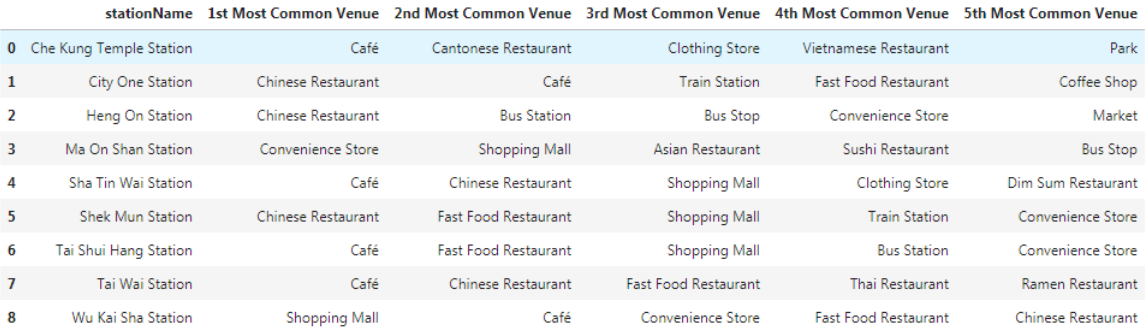


	stationName	ATM	Asian Restaurant	Athletics & Sports	BBQ Joint	Bubble Tea Shop	Buffet	Bus Station	Bus Stop	Café	...	Supermarket	Sushi Restaurant	Temple	Thai Restaurant	Theme Park	Toy / Game Store	Track Stadium	Train Station	Vegetarian / Vegan Restaurant	Vietnamese Restaurant
0	Che Kung Temple Station	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.166667	...	0.033333	0.033333	0.033333	0.0	0.033333	0.033333	0.000000	0.033333	0.000000	0.033333
1	City One Station	0.000000	0.000000	0.0	0.0	0.0	0.033333	0.000000	0.000000	0.100000	...	0.000000	0.033333	0.000000	0.0	0.000000	0.000000	0.000000	0.066667	0.000000	0.033333
2	Heng On Station	0.034483	0.000000	0.0	0.0	0.0	0.000000	0.103448	0.103448	0.034483	...	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.034483	0.000000	0.034483	0.000000
3	Ma On Shan Station	0.000000	0.066667	0.0	0.0	0.0	0.000000	0.066667	0.066667	0.000000	...	0.000000	0.066667	0.000000	0.0	0.000000	0.000000	0.033333	0.033333	0.000000	0.000000
4	Sha Tin Wai Station	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.133333	...	0.033333	0.000000	0.000000	0.0	0.033333	0.033333	0.000000	0.033333	0.000000	0.000000

rows × 67 columns

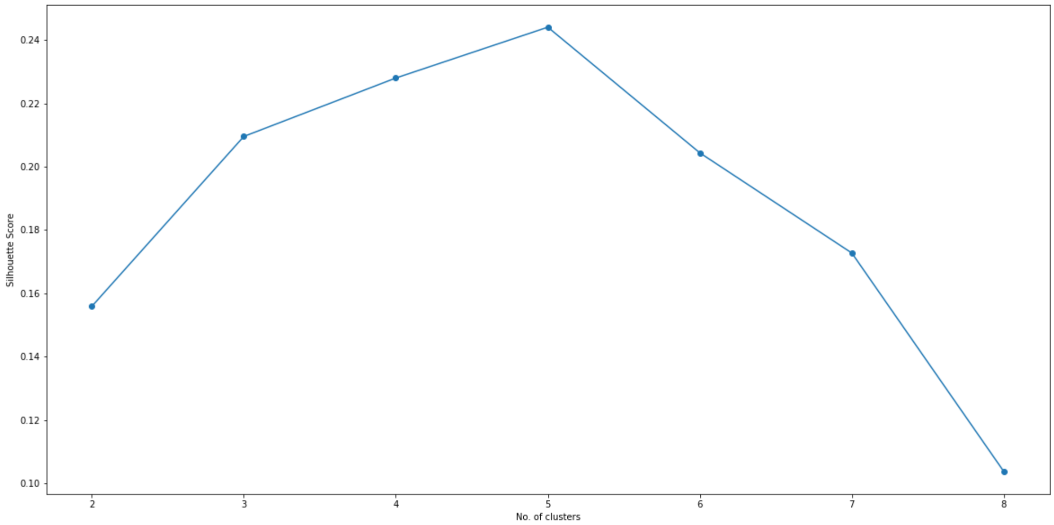
4.6 Top 5 most common venues

For easier analysis, only top 5 common venues near each of the stations were selected and used for k-means clustering algorithm.

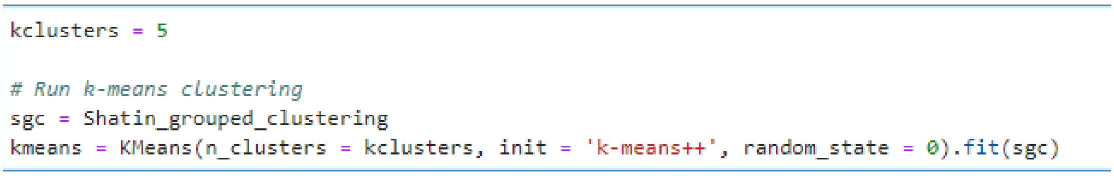


	stationName	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Che Kung Temple Station	Café	Cantonese Restaurant	Clothing Store	Vietnamese Restaurant	Park
1	City One Station	Chinese Restaurant	Café	Train Station	Fast Food Restaurant	Coffee Shop
2	Heng On Station	Chinese Restaurant	Bus Station	Bus Stop	Convenience Store	Market
3	Ma On Shan Station	Convenience Store	Shopping Mall	Asian Restaurant	Sushi Restaurant	Bus Stop
4	Sha Tin Wai Station	Café	Chinese Restaurant	Shopping Mall	Clothing Store	Dim Sum Restaurant
5	Shek Mun Station	Chinese Restaurant	Fast Food Restaurant	Shopping Mall	Train Station	Convenience Store
6	Tai Shui Hang Station	Café	Fast Food Restaurant	Shopping Mall	Bus Station	Convenience Store
7	Tai Wai Station	Café	Chinese Restaurant	Fast Food Restaurant	Thai Restaurant	Ramen Restaurant
8	Wu Kai Sha Station	Shopping Mall	Café	Convenience Store	Fast Food Restaurant	Chinese Restaurant

4.7 K-means Clustering



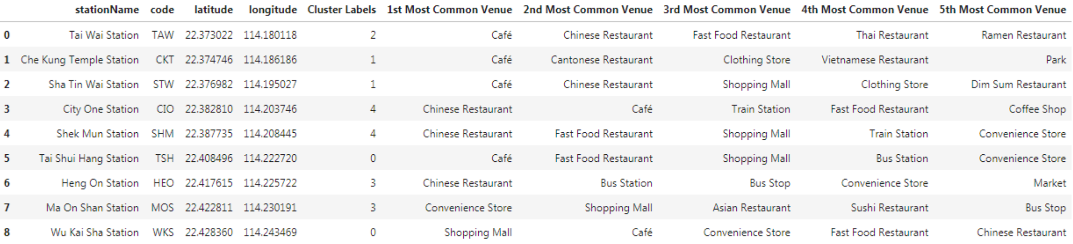
According to the above graph, k = 5 will be used as the optimal number in k-means clustering algorithm. After the k-means clustering was performed, the cluster label will be plotted in the map through Folium.



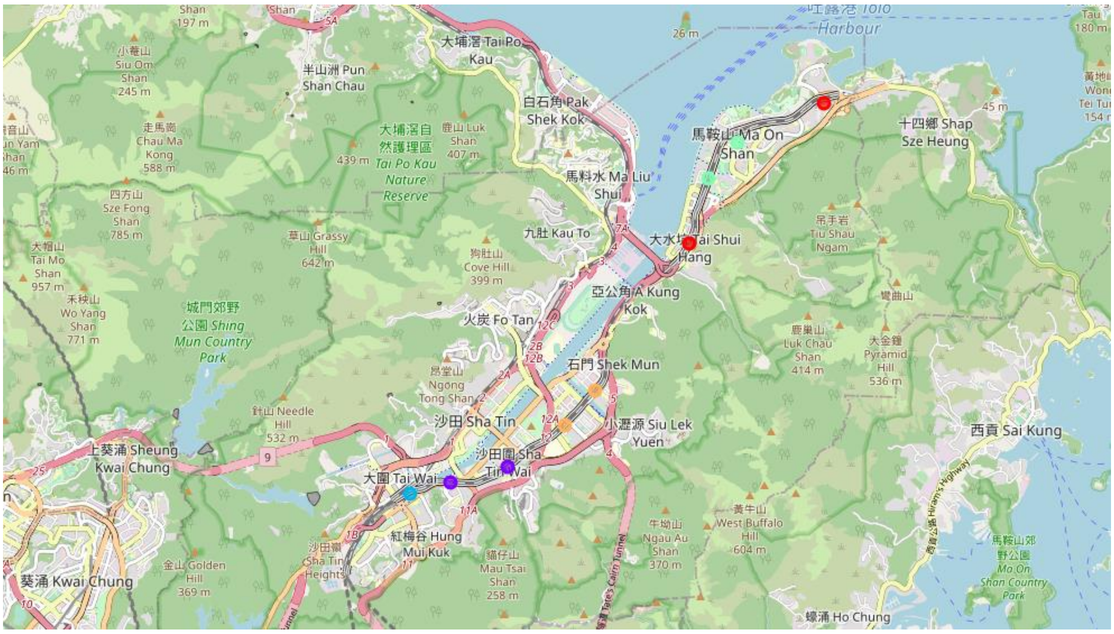
```
kclusters = 5

# Run k-means clustering
sgc = Shatin_grouped_clustering
kmeans = KMeans(n_clusters = kclusters, init = 'k-means++', random_state = 0).fit(sgc)
```

5. Result

The MTR stations were divided into 5 clusters where the value was found as optimal in section 4.7. The clustered MTR stations are visualized using different colors for easier identification in map.



	stationName	code	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Tai Wai Station	TAW	22.373022	114.180118	2	Café	Chinese Restaurant	Fast Food Restaurant	Thai Restaurant	Ramen Restaurant
1	Che Kung Temple Station	CKT	22.374746	114.186186	1	Café	Cantonese Restaurant	Clothing Store	Vietnamese Restaurant	Park
2	Sha Tin Wai Station	STW	22.376982	114.195027	1	Café	Chinese Restaurant	Shopping Mall	Clothing Store	Dim Sum Restaurant
3	City One Station	CIO	22.382810	114.203746	4	Chinese Restaurant	Café	Train Station	Fast Food Restaurant	Coffee Shop
4	Shek Mun Station	SHM	22.387735	114.208445	4	Chinese Restaurant	Fast Food Restaurant	Shopping Mall	Train Station	Convenience Store
5	Tai Shui Hang Station	TSH	22.408496	114.222720	0	Café	Fast Food Restaurant	Shopping Mall	Bus Station	Convenience Store
6	Heng On Station	HEO	22.417615	114.225722	3	Chinese Restaurant	Bus Station	Bus Stop	Convenience Store	Market
7	Ma On Shan Station	MOS	22.422811	114.230191	3	Convenience Store	Shopping Mall	Asian Restaurant	Sushi Restaurant	Bus Stop
8	Wu Kai Sha Station	WKS	22.428360	114.243469	0	Shopping Mall	Café	Convenience Store	Fast Food Restaurant	Chinese Restaurant



6. Discussion

After analyzing the 5 clusters obtained, it was found that cluster 3 is the most suitable one for solving the problem. It is because in Heng On and Ma On Shan Station, there are many shopping mall and bus stations but no café shop. The frequently occurrence of shopping mall and bus stops proved that the people will always pass by these areas and if a café is set up nearby will attract the people to come and stay for rest for a while. Moreover, there are nearly no café near the abovementioned location. Therefore, start up a café at these locations is less competitive as no competitors are nearby.