My Capstone project to learn course "Python for Data Science, AI & Development"

Cost effective analysis to open a shop close to metro stations in Shanghai, China

By Ziyue Sheng/ 2024 June

1. Introduction

1.1 Introduction of city Shanghai

Shanghai is the economics capital of China. In terms of GDP volume , consumer's capacity as an individual city, it has long been the top one among all cities in China; quoted from Wikipedia " with around 24.87 million inhabitants in 2023, while the urban area is the most populous in China, with 29.87 million residents, *it is the second most populous city proper in the world. Shanghai is a global center for finance, innovation and transportation.* "

The city embraces rich lifestyles, the business facilities to support such variety is substantial. So how to explore the business opportunities in such a giant city is interesting and exciting, especially if we look at the chance brought by the convenient metro transportation system in Shanghai;

1.2 Introduction of Shanghai's metro system

The Shanghai metro system is the world's second largest metro system by route length, reaching total 837 kilometres (520 mi) in Year 2024, it is also the second largest by the number of metro stations, meanwhile it ranks second in the world by annual ridership;

SHANGHAI METRO NETWORK MAP

1.3 Business chance exploration and problems

Such gigantic metro transportation not only provides much convenience to the daily life for people living in this city, but also renders big chance to setup commercial facilities as business chances, when the metro lines cover most of the city's major residence/commodity areas, more than half of the people movements within the city is by metros, people gets food, drink, shopping during their movement along the metro stations. On another hand, metro stations provide fast reachability so people are willing to go there when shopping is considered; Hence usually the café, coffee shop, drink selling, convenient store, shopping malls, restaurants of all flavors do have good flow of people to support their business;

When at the same time, the real estate prices are already significant in such a big city, utilizing the location advantage also brings solid pressure of the housing rental cost for which, when you think of starting a new business around the metro station you also need to check what you have to pay for that commercial real estate, these have to be well balanced to allow your investment gets profitable return;

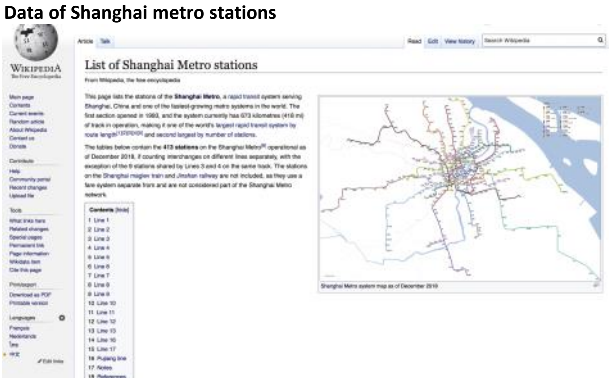
1.4 target people who are interested in the project

In this project, we will go through all the metro stations in Shanghai, by their location geo data of latitude/longitude, we explore each of their nearby venues, finding their current business categories, classify these categories into different patterns, then with the information of the real estate price level and the opportunities analyzed from the clustering pattern, giving out the recommendation what and where to start your business closing to Shanghai metro stations;

So target people of this project are those interested to explore business opportunities in nearby area of Shanghai metro stations, considering business categories, right stations as well as budget cost for real estate;

2. Data

2.1 Data of Shanghai metro stations



List of Shanghai Metro stations

From Wikipedia, the free encyclopedia

This page lists the stations of the **Shanghai Metro**, a rapid transit system serving Shanghai, China and one of the fastest-growing metro systems in the world. The first section opened in 1993, and the system currently has 673 kilometres (418 mi) of track in operation, making it one of the world's largest rapid transit system by route length[1][2][3][4] and second largest by number of stations.

The tables below contain the **413 stations** on the Shanghai Metro[5] operational as of December 2019, if counting interchanges on different lines separately, with the exception of the 9 stations shared by Lines 3 and 4 on the same track. The stations on the Shanghai maglev train and Jinshan railway are not included, as they use a fare system separate from and are not considered part of the Shanghai Metro network.

Contents [hide]
1 Line 1
2 Line 2
3 Line 3
4 Line 4
5 Line 5
6 Line 6
7 Line 7
8 Line 8
9 Line 9
10 Line 10
11 Line 11
12 Line 12
13 Line 13
14 Line 16
15 Line 17
16 Pujiang line
17 Notes
18 References

Shanghai Metro system map as of December 2019

In this project, it is crutial to have all the accurate information of Shanghai metro stations;
I got the data from the Wikipedia website "List of Shanghai Metro stations" at the following address: https://en.wikipedia.org/wiki/List_of_Shanghai_Metro_stations
There are entry pages for all 16 metro lines, all the station listed along each line with their name, belonged District; And for each station, there is a link to each of the station page which contains the detail of the station information; Here what I am interested is the longitude and latitude of the station;
So from the above wikipedia pages, I got the station information as following:

- station name
- district the station belongs to
- station longitude/latitude

I use python to webscrape these data that includes total 422 stations related information;

2.2 Data of Shanghai District map

The data is used to be the input of python folium library to generate the map of Shanghai with each district by different color, by calling folium.Map.choropleth method;
Data obtain procedure:

- Get the Arcgis data format of "Shanghai district border" from googling website and download from
https://www.arcgis.com/home/item.html?id=105f92bd1fe54d428bea35eade65691b
- converting the Arcgis data format to the GeoJson data format by online tool:
https://ogre.adc4gis.com/
The instruction of doing that is in
https://www.statsilk.com/maps/convert-esri-shapefile-map-geojson-format

2.3 Data of average estate price of each district in Shanghai

The data is obtained from a big estate trade agency of China
https://shanghai.anjuke.com/market/
In this website, the price of average estate price of each district in Year 2020 May is got, it is stored then in the District_price.csv and later to be read into a pandas dataframe for further processing;

2.4 Venue data of shops nearby each Shanghai metro station

With all the Shanghai metro station data obtained from 2.1(longitude, latitude), by calling Foursquare API for each metro station with radius of 500 meters and limit of 100 venues, I got all the venue data with my foursquare free account;
The data I extracted out is venue name, venue latitude, vanue longitude and venue category; In this project, I am particulary interested in the venue category which should help me to analyze all the venue data and cluster their nearby metro station with different categorical characteristic, then this can be the reference together with their belonging district average estate price to have a business decision review;

So This is a project that many data science knowledges & skills are utilized, from web scraping (to Wikipedia pages), working with API provided by Foursquare, applying data cleaning & data wrangling to machine learning (K-means clustering) and map visualization (Folium).
In the next section, I will present the Methodology in which I will explain the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

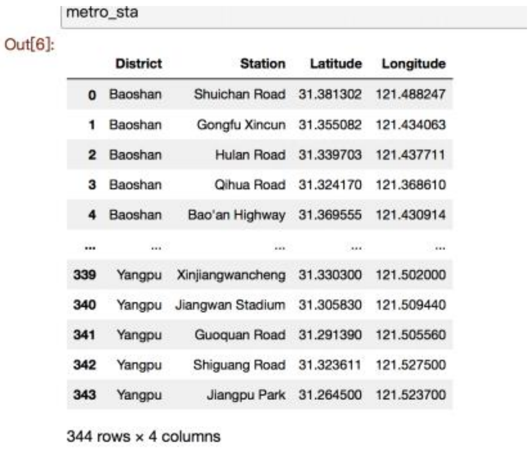
3. Methodology

3.1 webscraping Shanghai metro stations' data

I use python to webscrape these data by applying Python requests, beautifulsoup and selenium packages. From wikipedia page I got all Shanghai metro stations' data including total 422 stations related information by going through all the stations information web page from each of 16 metro lines;
These information are read into pandas dataframe of metro_sta, with columns of each station as:

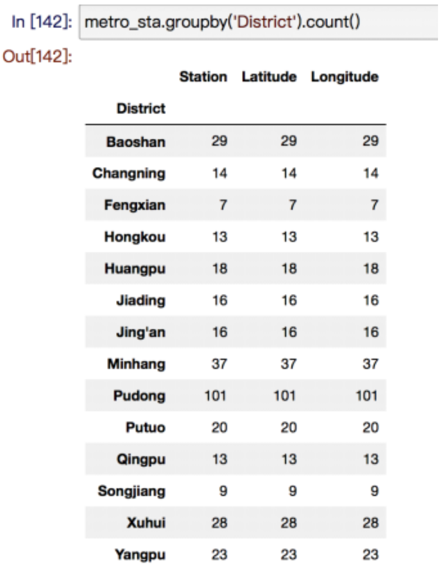
District belonged to	Station Name	Latitude	Longitude

By removing the station duplication (some metro stations are joint transfer stations so they can belong to different metro lines) I got total 344 distinct stations;



```
      metro_sta

Out[6]:
           District          Station   Latitude   Longitude
        0   Baoshan     Shuichan Road  31.381302  121.488247
        1   Baoshan    Gongfu Xincun  31.355082  121.434063
        2   Baoshan        Hulan Road  31.339703  121.437711
        3   Baoshan        Qihua Road  31.324170  121.368610
        4   Baoshan    Bao'an Highway  31.369555  121.430914
      ...       ...              ...        ...         ...
      339    Yangpu  Xinjiangwancheng  31.330300  121.502000
      340    Yangpu  Jiangwan Stadium  31.305830  121.509440
      341    Yangpu     Guoquan Road  31.291390  121.505560
      342    Yangpu    Shiguang Road  31.323611  121.527500
      343    Yangpu      Jiangpu Park  31.264500  121.523700

      344 rows × 4 columns
```

Group by district we can get the number of metro stations in each district

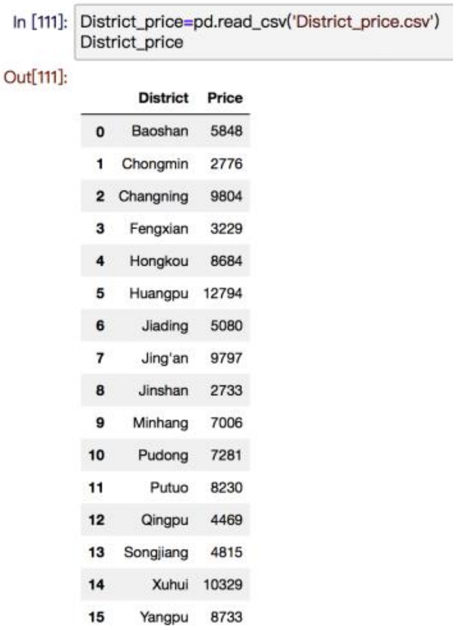


```
In [142]: metro_sta.groupby('District').count()
Out[142]:
```

District	Station	Latitude	Longitude
Baoshan	29	29	29
Changning	14	14	14
Fengxian	7	7	7
Hongkou	13	13	13
Huangpu	18	18	18
Jiading	16	16	16
Jing'an	16	16	16
Minhang	37	37	37
Pudong	101	101	101
Putuo	20	20	20
Qingpu	13	13	13
Songjiang	9	9	9
Xuhui	28	28	28
Yangpu	23	23	23

3.2 Obtaining the real estate price information from the web

The average real estate price information is obtained from the website of one big China real estate trade agency .
The data is simple and brief, so it is put in one csv file and read into one pandas dataframe
Total 16 districts of Shanghai are listed, the latest data of May/2020 when this report was written is used;
Note: the price is in the value of US$/per square-meter(m^2)

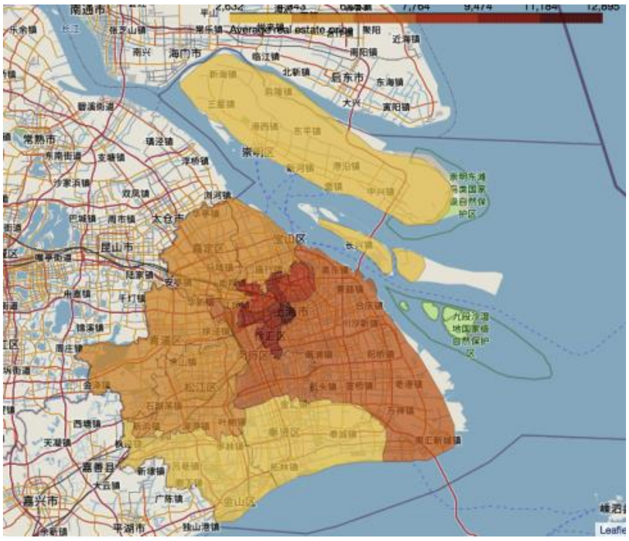


```
In [111]: District_price=pd.read_csv('District_price.csv')
          District_price
Out[111]:
```

	District	Price
0	Baoshan	5848
1	Chongmin	2776
2	Changning	9804
3	Fengxian	3229
4	Hongkou	8684
5	Huangpu	12794
6	Jiading	5080
7	Jing'an	9797
8	Jinshan	2733
9	Minhang	7006
10	Pudong	7281
11	Putuo	8230
12	Qingpu	4469
13	Songjiang	4815
14	Xuhui	10329
15	Yangpu	8733

3.3 Generate a Shanghai_map of all districts each with its real estate price info.

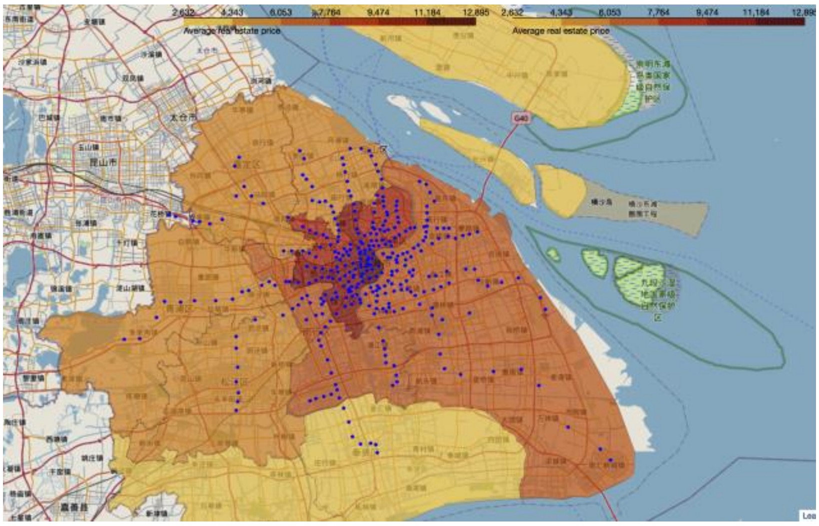
Using folium.Map.Cholopleth method, with input of Shanghai district borders geodata information(obtained with the steps presented in the above 2.2) dataframe, the real estate price dataframe of each district (obtained from 3.2) , The shanghai_map is generated:



The price level of each district of Shanghai is visualized, the darker of the color, the more expensive of the real estate;

3.4 Add all metro stations to the shanghai_map

With the above generated shanghai_map, using **folium.circleMarker** method to add all the 344 metro stations for visualization

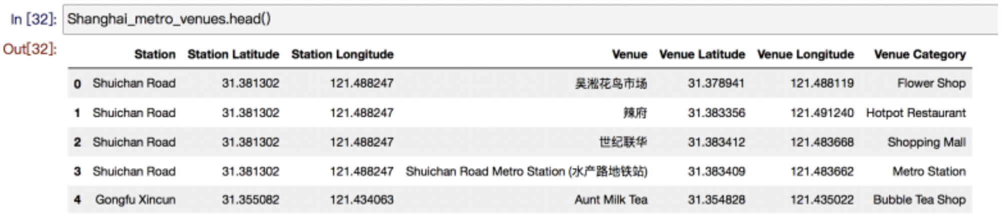


3.5 Obtaining all venue information of each metro station

With each metro station geo data (longitude, latitude), call Foursquare API to get the venue data of its nearby shops; Set the venue limit number to be 100 and the nearby radius as 500 meters;

- URL formulated
  url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, station_latitude, station_longitude, VERSION, radius, LIMIT)

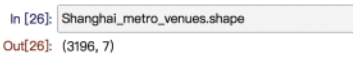
- Then using the python requests library to get the venue information
  results = requests.get(url).json()
  By iterative calling of the Foursquare API, all venue data are returned for each of the metro station,
  Foursquare will return the venue data in JSON format and the venue name, venue category, venue latitude and longitude are extracted. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. all interested venue data are input to a dataframe shanghai_metro_venues, with structure as:



```
In [32]: Shanghai_metro_venues.head()
Out[32]:
```

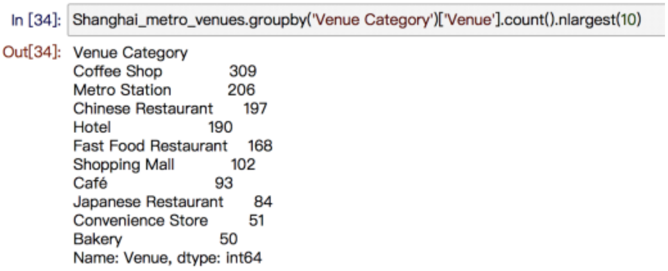
	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Shuichan Road	31.381302	121.488247	吴淞花鸟市场	31.378941	121.488119	Flower Shop
1	Shuichan Road	31.381302	121.488247	辣府	31.383356	121.491240	Hotpot Restaurant
2	Shuichan Road	31.381302	121.488247	世纪联华	31.383412	121.483668	Shopping Mall
3	Shuichan Road	31.381302	121.488247	Shuichan Road Metro Station (水产路地铁站)	31.383409	121.483662	Metro Station
4	Gongfu Xincun	31.355082	121.434063	Aunt Milk Tea	31.354828	121.435022	Bubble Tea Shop

total returned venues are 3196 and stored in dataframe shanghai_metro_venues



```
In [26]: Shanghai_metro_venues.shape
Out[26]: (3196, 7)
```

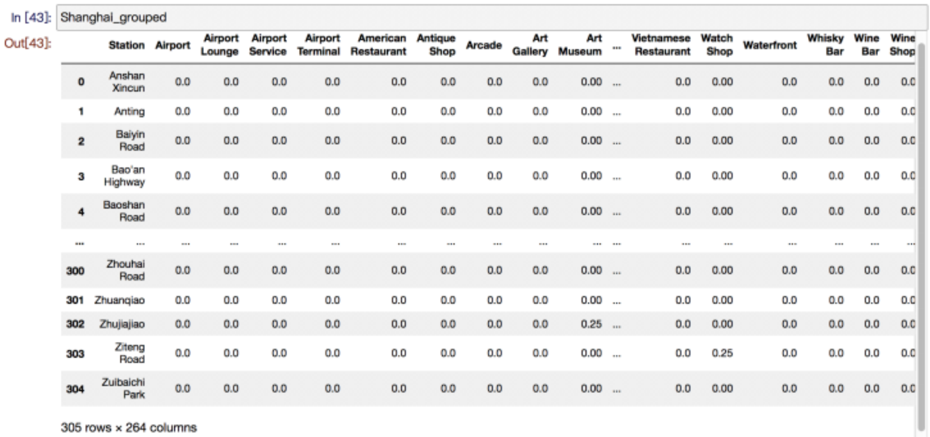
- Examing the top 10 venue categories



```
In [34]: Shanghai_metro_venues.groupby('Venue Category')['Venue'].count().nlargest(10)
Out[34]: Venue Category
         Coffee Shop            309
         Metro Station          206
         Chinese Restaurant     197
         Hotel                  190
         Fast Food Restaurant   168
         Shopping Mall          102
         Café                    93
         Japanese Restaurant     84
         Convenience Store       51
         Bakery                  50
         Name: Venue, dtype: int64
```

3.6 Analysis of each metro station venue category

- This is done by grouping the rows of metro station name and taking the mean of the frequency of occurrence of each venue category.



In [43]:
```
Shanghai_grouped
```

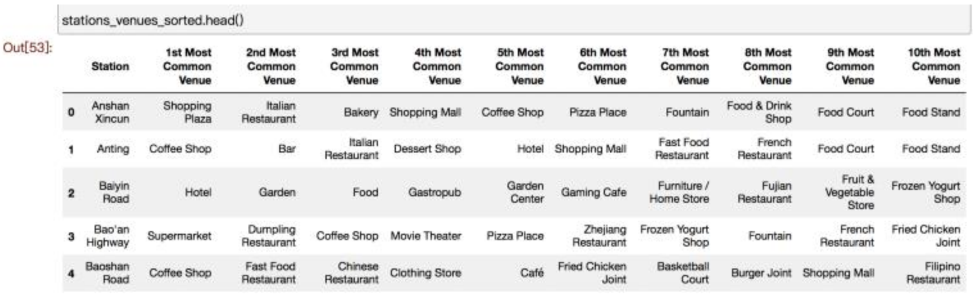
Out[43]:

	Station	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum	...	Vietnamese Restaurant	Watch Shop	Waterfront	Whisky Bar	Wine Bar	Wine Shop
0	Anshan Xincun	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0
1	Anting	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0
2	Baiyin Road	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0
3	Bao'an Highway	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0
4	Baoshan Road	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
300	Zhouhai Road	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0
301	Zhuanqiao	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0
302	Zhujiajiao	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	...	0.0	0.00	0.0	0.0	0.0	0.0
303	Ziteng Road	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.25	0.0	0.0	0.0	0.0
304	Zuibaichi Park	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	...	0.0	0.00	0.0	0.0	0.0	0.0

305 rows × 264 columns

note: some metro station has no venue data

- Then we can get the 10th most common venue of each metro station



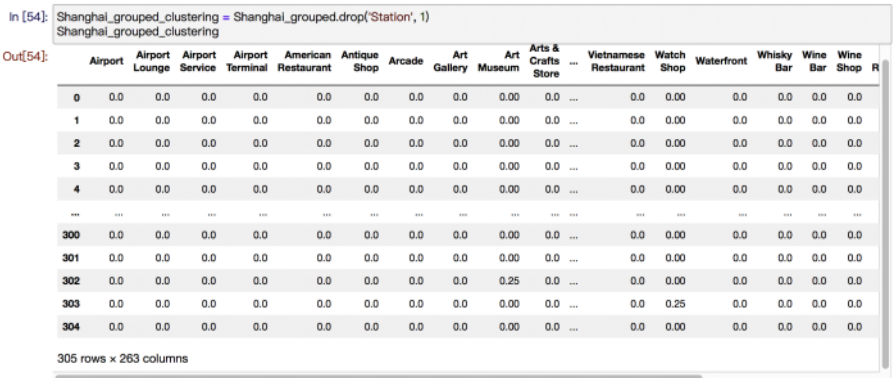
```
stations_venues_sorted.head()
```

Out[53]:

	Station	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Anshan Xincun	Shopping Plaza	Italian Restaurant	Bakery	Shopping Mall	Coffee Shop	Pizza Place	Fountain	Food & Drink Shop	Food Court	Food Stand
1	Anting	Coffee Shop	Bar	Italian Restaurant	Dessert Shop	Hotel	Shopping Mall	Fast Food Restaurant	French Restaurant	Food Court	Food Stand
2	Baiyin Road	Hotel	Garden	Food	Gastropub	Garden Center	Gaming Cafe	Furniture / Home Store	Fujian Restaurant	Fruit & Vegetable Store	Frozen Yogurt Shop
3	Bao'an Highway	Supermarket	Dumpling Restaurant	Coffee Shop	Movie Theater	Pizza Place	Zhejiang Restaurant	Frozen Yogurt Shop	Fountain	French Restaurant	Fried Chicken Joint
4	Baoshan Road	Coffee Shop	Fast Food Restaurant	Chinese Restaurant	Clothing Store	Café	Fried Chicken Joint	Basketball Court	Burger Joint	Shopping Mall	Filipino Restaurant

3.7 Clustering the metro stations with their most common venue nearby

- The 'station' venue type is dropped because we want to remove the station itself and include only the venue/shop category nearby



In [54]:
```
Shanghai_grouped_clustering = Shanghai_grouped.drop('Station', 1)
Shanghai_grouped_clustering
```

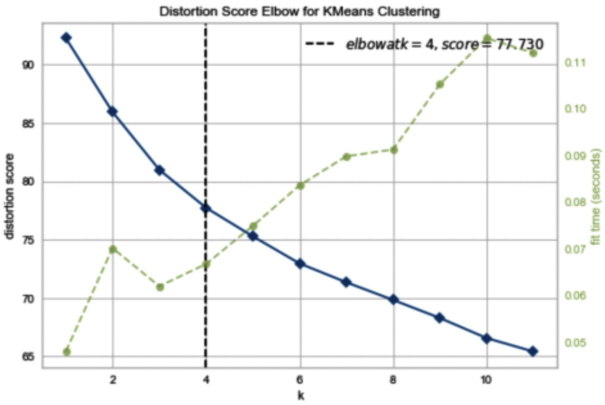
Out[54]:

	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	...	Vietnamese Restaurant	Watch Shop	Waterfront	Whisky Bar	Wine Bar	Wine Shop	R
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
300	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
301	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
302	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	
303	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.25	0.0	0.0	0.0	0.0	
304	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	...	0.0	0.00	0.0	0.0	0.0	0.0	

305 rows × 263 columns

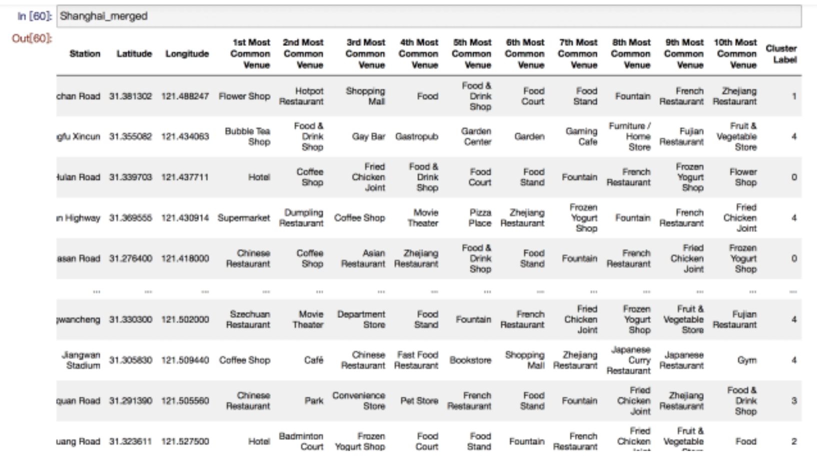
- Perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

  - First, Using cluster.KElbowVisualizer from the yellowbrick library to find the most optimal value of K



From above distortion score, we find k=5 is the most preferred value

  - With k=5, we call Kmeans method to cluster all the metro stations into 5, and get the merged dataframe of 10 most common venu es of each metro station, and their cluster number;

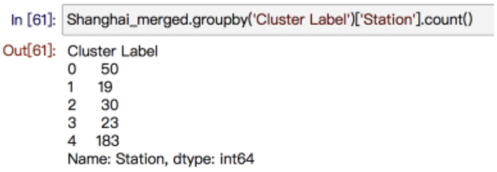


```
In [60]: Shanghai_merged
```

Out[60]:

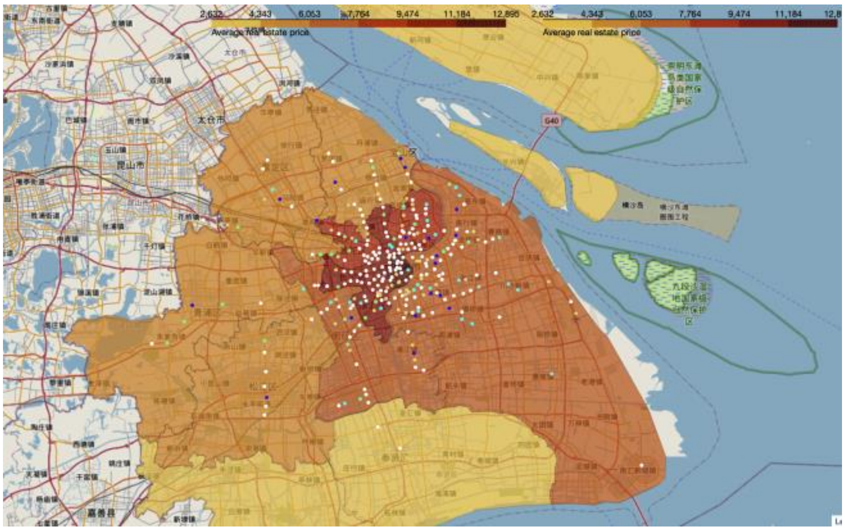
Station	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster Label
chan Road	31.381302	121.488247	Flower Shop	Hotpot Restaurant	Shopping Mall	Food	Food & Drink Shop	Food Court	Food Stand	Fountain	French Restaurant	Zhejiang Restaurant	1
gfu Xincun	31.355082	121.434063	Bubble Tea Shop	Food & Drink Shop	Gay Bar	Gastropub	Garden Center	Garden	Gaming Cafe	Furniture / Home Store	Fujian Restaurant	Fruit & Vegetable Store	4
ɪulan Road	31.339703	121.437711	Hotel	Coffee Shop	Fried Chicken Joint	Food & Drink Shop	Food Court	Food Stand	Fountain	French Restaurant	Frozen Yogurt Shop	Flower Shop	0
n Highway	31.369555	121.430914	Supermarket	Dumpling Restaurant	Coffee Shop	Movie Theater	Pizza Place	Zhejiang Restaurant	Frozen Yogurt Shop	Fountain	French Restaurant	Fried Chicken Joint	4
asan Road	31.276400	121.418000	Chinese Restaurant	Coffee Shop	Asian Restaurant	Zhejiang Restaurant	Food & Drink Shop	Food Stand	Fountain	French Restaurant	Fried Chicken Joint	Frozen Yogurt Shop	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
ȷwancheng	31.330300	121.502000	Szechuan Restaurant	Movie Theater	Department Store	Food Stand	Fountain	French Restaurant	Fried Chicken Joint	Frozen Yogurt Shop	Fruit & Vegetable Store	Fujian Restaurant	4
Jiangwan Stadium	31.305830	121.509440	Coffee Shop	Café	Chinese Restaurant	Fast Food Restaurant	Bookstore	Shopping Mall	Zhejiang Restaurant	Japanese Curry Restaurant	Japanese Restaurant	Gym	4
quan Road	31.291390	121.505560	Chinese Restaurant	Park	Convenience Store	Pet Store	French Restaurant	Food Stand	Fountain	Fried Chicken Joint	Zhejiang Restaurant	Food & Drink Shop	3
uang Road	31.323611	121.527500	Hotel	Badminton Court	Frozen Yogurt Shop	Food Court	Food Stand	Fountain	French Restaurant	Fried Chicken	Fruit & Vegetable	Food	2

- And we can get the number of metro stations in each cluster



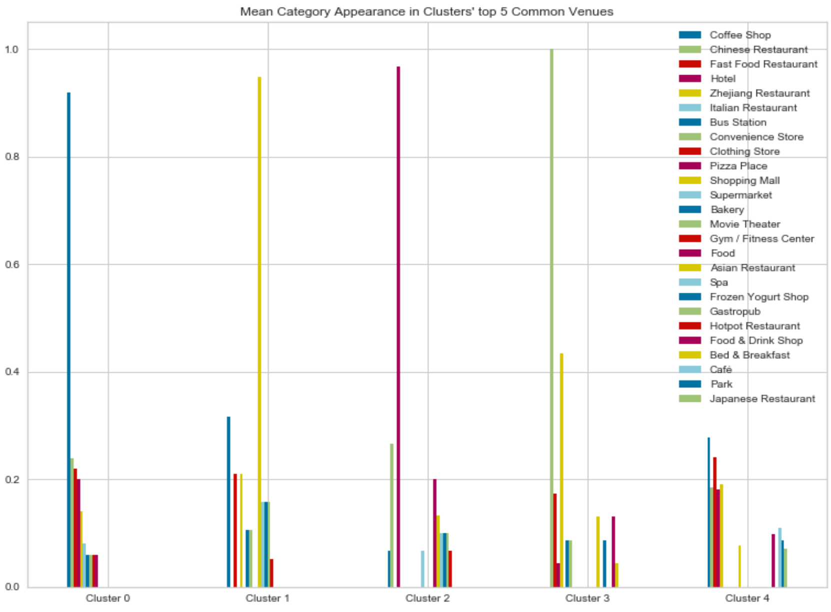
```
In [61]: Shanghai_merged.groupby('Cluster Label')['Station'].count()

Out[61]: Cluster Label
         0     50
         1     19
         2     30
         3     23
         4    183
         Name: Station, dtype: int64
```

- Then we can visualize all the metro stations with different color based on its belonging cluster on the shanghai_map



4. Result

By clustering all the metro stations we get the most common venues/shops of each cluster;
The following is the bar chart of the 5 clusters for their common venues/shops:



Cluster0: Coffees / fast food area

Metro stations with most of the coffee shops, then the Chinese Restaurants, Fast food restaurants as well as hotels and Zhejiang restaurant as top 5 categories;
Then the next top 5 shop category as Italian Restaurant, bus station , convenience store and clothing store as well as Pizza place;
Summarized as coffee/fast food area, with 2nd business tag as hotels/convenience store/clothing
There are 50 stations belonging to this cluster;

cluster0:
[('Hulan Road', 'Baoshan'), ('Dahuasan Road', 'Baoshan'), ('Shanghai University', 'Baoshan'), ('South Changjiang Road', 'Baoshan'), ('Songhong Road', 'Changning'), ('Fengxian Xincheng', 'Fengxian'), ('Huanchengdong Road', 'Fengxian'), ('Dabaishu', 'Hongkou'), ('Jiangwan Town', 'Hongkou'), ('North Sichuan Road', 'Hongkou'), ('Luban Road', 'Huangpu'), ('World Expo Museum', 'Huangpu'), ('Madang Road', 'Huangpu'), ('Shanghai Circuit', 'Jiading'), ('Anting', 'Jiading'), ('Zhaofeng Road', 'Jiading'), ('Yindu Road', 'Minhang'), ('Qixin Road', 'Minhang'), ('Ziteng Road', 'Minhang'), ('Lianhang Road', 'Minhang'), ('Pujiang Town', 'Minhang'), ('Oriental Sports Center', 'Pudong'), ('South Lingyan Road', 'Pudong'), ('Yuanshen Stadium', 'Pudong'), ('China Art Museum', 'Pudong'), ("Shanghai Children's Medical Center", 'Pudong'), ('Jinji Road', 'Pudong'), ('Shangcheng Road', 'Pudong'), ('Fanghua Road', 'Pudong'), ('Yuntai Road', 'Pudong'), ('Yaohua Road', 'Pudong'), ('Huapeng Road', 'Pudong'), ('Pudong International Airport', 'Pudong'), ('Yuandong Avenue', 'Pudong'), ('Jinjing Road', 'Pudong'), ('Dongchang Road', 'Pudong'), ('Jinke Road', 'Pudong'), ('South Qilianshan Road', 'Putuo'), ('Taopu Xincun', 'Putuo'), ('Liziyuan', 'Putuo'), ('Shanghai Swimming Center', 'Xuhui'), ('Longhua', 'Xuhui'), ('Damuqiao Road', 'Xuhui'), ('Shanghai Indoor Stadium', 'Xuhui'), ('Caohejing Hi-Tech Park', 'Xuhui'), ('Shanghai South Railway Station', 'Xuhui'), ("Dong'an Road", 'Xuhui'), ('Jiangpu Road', 'Yangpu'), ('East Yingao Road', 'Yangpu'), ('Jiangpu Park', 'Yangpu')]

Cluster1: Shopping center area

Metro stations with most of the Shopping malls & super markets, the second commercial store characteristic are coffee shops, fast food restaurants and ZheJiang restaurants, summarized most as shopping centers, with 2nd business tag as restaurants/bakery/entertainment
There are 19 stations belonging to this cluster;

cluster1:
[('Shuichan Road', 'Baoshan'), ('Liuhang', 'Baoshan'), ('Baoyang Road', 'Baoshan'), ('Youdian Xincun', 'Hongkou'), ('Jiading Xincheng', 'Jiading'), ('Nanxiang', 'Jiading'), ('Jiangyue Road', 'Minhang'), ('Gudai Road', 'Minhang'), ('Wulian Road', 'Pudong'), ('Beiyangjing Road', 'Pudong'), ('Middle Yanggao Road', 'Pudong'), ('Zhouhai Road', 'Pudong'), ('Yuqiao', 'Pudong'), ('Sanlin', 'Pudong'), ('Xianan Road', 'Pudong'), ('Chenchun Road', 'Pudong'), ('Caoying Road', 'Qingpu'), ('Songjiang Sports Center', 'Songjiang'), ('Ningguo Road', 'Yangpu')]

Cluster2: Hotels area

Metro stations with most of the hotels, followed by Chinese restaurants, pizza place, shopping mall/super markets as well as movie theater and Gym fitness
summarized as hotel area, with 2nd business tag as Food/shopping center/entertainment
There are 30 stations belonging to this cluster;

cluster2:
[('West Yingao Road', 'Baoshan'), ('Xiaonanmen', 'Huangpu'), ('Baiyin Road', 'Jiading'), ('North Zhongshan Road', "Jing'an"), ('Zhongxing Road', "Jing'an"), ('Wenjing Road', 'Minhang'), ('Hangzhong Road', 'Minhang'), ('Hechuan Road', 'Minhang'), ('Longbai Xincun', 'Minhang'), ('West Huaxia Road', 'Pudong'), ('North Waigaoqiao Free Trade Zone', 'Pudong'), ('Hangjin Road', 'Pudong'), ('Linyi Xincun', 'Pudong'), ('Jinxiu Road', 'Pudong'), ('West Gaoke Road', 'Pudong'), ('Chuansha', 'Pudong'), ('East Huaxia Road', 'Pudong'), ('Minlei Road', 'Pudong'), ('Zhangjiang Road', 'Pudong'), ('Kangxin Highway', 'Pudong'), ('Xiuyan Road', 'Pudong'), ('Zhangjiang Hi-Tech Park', 'Pudong'), ('Huinan', 'Pudong'), ('West Shanghai Railway Station', 'Putuo'), ('East Xujing', 'Qingpu'), ('Hongcao Road', 'Xuhui'), ('Shilong Road', 'Xuhui'), ('Huangxing Park', 'Yangpu'), ('Longchang Road', 'Yangpu'), ('Shiguang Road', 'Yangpu')]

Cluster3: Chinese Restaurants

Metro stations with most of the Chinese restaurants(including Zhejiang Restaurant), followed by fastfood restaurants, shopping malls and other food (including Asian food)
summarized as Chinese Restaurants area, with 2nd business tag as fastfood/shopping center/other food
There are 23 stations belonging to this cluster;

cluster3:
[('Youyi Road', 'Baoshan'), ('West Youyi Road', 'Baoshan'), ('Shanghai Zoo', 'Changning'), ('Fengzhuang', 'Jiading'), ('Shanghai Automobile City', 'Jiading'), ('West Jinshajiang Road', 'Jiading'), ('Pengpu Xincun', "Jing'an"), ('North Xizang Road', "Jing'an"), ('Chunshen Road', 'Minhang'), ('Jinping Road', 'Minhang'), ('Hongxin Road', 'Minhang'), ('Jinqiao', 'Pudong'), ('Deping Road', 'Pudong'), ('Gangcheng Road', 'Pudong'), ('Qilianshan Road', 'Putuo'), ('Zhenbei Road', 'Putuo'), ('Dianshanhu Avenue', 'Qingpu'), ('Zhujiajiao', 'Qingpu'), ('Huijin Road', 'Qingpu'), ('Sheshan', 'Songjiang'), ('Caobao Road', 'Xuhui'), ('Jinjiang Park', 'Xuhui'), ('Guoquan Road', 'Yangpu')]

Cluster4: Less density commercial area

Metro stations with less density of commercial stores compared with other clusters, and much focused on the coffee/hotel/food and with very low existance of other business types including entertainment, shopping centers , etc;
There are 183 stations belonging to this cluster

cluster4:
[('Gongfu Xincun', 'Baoshan'), ("Bao'an Highway", 'Baoshan'), ('Luonan Xincun', 'Baoshan'), ('Meilan Lake', 'Baoshan'), ('Nanchen Road', 'Baoshan'), ('Shangda Road', 'Baoshan'), ('Changzhong Road', 'Baoshan'), ('Tonghe Xincun', 'Baoshan'), ('Xingzhi Road', 'Baoshan'), ('Fujin Road', 'Baoshan'), ('Zhanghuabang', 'Baoshan'), ('Songbin Road', 'Baoshan'), ('Gucun Park', 'Baoshan'), ('Dachang Town', 'Baoshan'), ('Hongqiao Airport Terminal 1', 'Changning'), ('Beixinjing', 'Changning'), ('Jiangsu Road', 'Changning'), ('Zhongshan Park', 'Changning'), ('Weining Road', 'Changning'), ('Loushanguan Road', 'Changning'), ('Hongqiao Road', 'Changning'), ("West Yan'an Road", 'Changning'), ('Shuicheng Road', 'Changning'), ('Yili Road', 'Changning'), ('Songyuan Road', 'Changning'), ('Longxi Road', 'Changning'), ('Xiaotang', 'Fengxian'), ('Wangyuan Road', 'Fengxian'), ('Hongkou Football Stadium', 'Hongkou'), ('Tiantong Road', 'Hongkou'), ('International Cruise Terminal', 'Hongkou'), ('Tilanqiao', 'Hongkou'), ('Hailun Road', 'Hongkou'), ('Quyang Road', 'Hongkou'), ('Chifeng Road', 'Hongkou'), ('Linping Road', 'Hongkou'), ('Dongbaoxing Road', 'Hongkou'), ("People's Square", 'Huangpu'), ('Xinzha Road', 'Huangpu'), ('South Shaanxi Road', 'Huangpu'), ('Lujiabang Road', 'Huangpu'), ('Dashijie', 'Huangpu'), ('South Huangpi Road', 'Huangpu'), ('South Xizang Road', 'Huangpu'), ('East Nanjing Road', 'Huangpu'), ('Xintiandi', 'Huangpu'), ('Dapuqiao', 'Huangpu'), ('Nanpu Bridge', 'Huangpu'), ('Laoximen', 'Huangpu'), ('Yuyuan Garden', 'Huangpu'), ('Middle Huaihai Road', 'Huangpu'), ('Malu', 'Jiading'), ('West Jiading', 'Jiading'), ('North Jiading', 'Jiading'), ('Jinyun Road', 'Jiading'), ('Guangming Road', 'Jiading'), ('Shanghai Circus World', "Jing'an"), ('West Nanjing Road', "Jing'an"), ('Shanghai Railway Station', "Jing'an"), ('Qufu Road', "Jing'an"), ('Hanzhong Road', "Jing'an"), ('Yanchang Road', "Jing'an"), ('Wenshui Road', "Jing'an"), ('Changping Road', "Jing'an"), ('Gongkang Road', "Jing'an"), ('Baoshan Road', "Jing'an"), ('Shanghai Natural History Museum', "Jing'an"), ("Jing'an Temple", "Jing'an"), ('Xinzhuang', 'Minhang'), ('Zhuanqiao', 'Minhang'), ('Hongqiao Airport Terminal 2', 'Minhang'), ('Dongchuan Road', 'Minhang'), ('Minhang Development Zone', 'Minhang'), ('Donglan Road', 'Minhang'), ('Waihuanlu', 'Minhang'), ('Lianhua Road', 'Minhang'), ('Hongqiao Railway Station', 'Minhang'), ('Jiangchuan Road', 'Minhang'), ('Jianchuan Road', 'Minhang'), ('Sanlu Highway', 'Minhang'), ('Minrui Road', 'Minhang'), ('Xingzhong Road', 'Minhang'), ('Qibao', 'Minhang'), ('Shendu Highway', 'Minhang'), ('Luheng Road', 'Minhang'), ('Zhongchun Road', 'Minhang'), ('Lingzhao Xincun', 'Pudong'), ('Jufeng Road', 'Pudong'), ('Lancun Road', 'Pudong'), ('Pudian Road', 'Pudong'), ('Shangnan Road', 'Pudong'), ('Yangsi', 'Pudong'), ('Chengshan Road', 'Pudong'), ('Boxing Road', 'Pudong'), ('Jinqiao Road', 'Pudong'), ('Yunshan Road', 'Pudong'), ('Minsheng Road', 'Pudong'), ('Taierzhuang Road', 'Pudong'), ('Gaoqing Road', 'Pudong'), ('Dongming Road', 'Pudong'), ('South Waigaoqiao Free Trade Zone', 'Pudong'), ('Lantian Road', 'Pudong'), ('Fangdian Road', 'Pudong'), ('Century Avenue', 'Pudong'), ('Pusan Road', 'Pudong'), ('Huamu Road', 'Pudong'), ('Longyang Road', 'Pudong'), ('South Yanggao Road', 'Pudong'), ('Changqing Road', 'Pudong'), ('Houtan', 'Pudong'), ('Zhongke Road', 'Pudong'), ('Xuelin Road', 'Pudong'), ('Jinhai Road', 'Pudong'), ('Gutang Road', 'Pudong'), ('Tangqiao', 'Pudong'), ('Disney Resort', 'Pudong'), ('North Yanggao Road', 'Pudong'), ('East Sanlin', 'Pudong'), ('Haitiansan Road', 'Pudong'), ('Luoshan Road', 'Pudong'), ('Pudong Avenue', 'Pudong'), ('Middle Huaxia Road', 'Pudong'), ('Lujiazui', 'Pudong'), ('Shanghai Science and Technology Museum', 'Pudong'), ('Century Park', 'Pudong'), ('Guanglan Road [f]', 'Pudong'), ('Beicai', 'Pudong'), ('Tangzhen', 'Pudong'), ('Shibo Avenue', 'Pudong'), ('Dishui Lake', 'Pudong'), ('Lianxi Road', 'Pudong'), ('Fengqiao Road', 'Putuo'), ('Caoyang Road', 'Putuo'), ('Longde Road', 'Putuo'), ('Jinshajiang Road', 'Putuo'), ('Wuning Road', 'Putuo'), ('Jiangning Road', 'Putuo'), ('Zhongtan Road', 'Putuo'), ('Zhenru', 'Putuo'), ('Changshou Road', 'Putuo'), ('Wuwei Road', 'Putuo'), ('Zhenping Road', 'Putuo'), ('Daduhe Road', 'Putuo'), ('Langao Road', 'Putuo'), ('Xincun Road', 'Putuo'), ('Oriental Land', 'Qingpu'), ('Middle Jiasong Road', 'Qingpu'), ('Jiuting', 'Songjiang'), ('Songjiang University Town', 'Songjiang'), ('Dongjing', 'Songjiang'), ('Songjiang South Railway Station', 'Songjiang'), ('Zuibaichi Park', 'Songjiang'), ('Songjiang Xincheng', 'Songjiang'), ('Guilin Park', 'Xuhui'), ('Hongmei Road', 'Xuhui'), ('Caoxi Road', 'Xuhui'), ('Yishan Road', 'Xuhui'), ('Xujiahui', 'Xuhui'), ('Yunjin Road', 'Xuhui'), ('Longcao Road', 'Xuhui'), ('Jiaotong University', 'Xuhui'), ('Middle Longhua Road', 'Xuhui'), ('Jiashan Road', 'Xuhui'), ('Zhaojiabang Road', 'Xuhui'), ('Guilin Road', 'Xuhui'), ('Shanghai Stadium', 'Xuhui'), ('Shanghai Library', 'Xuhui'), ('Changshu Road', 'Xuhui'), ('Hengshan Road', 'Xuhui'), ('Huangxing Road', 'Yangpu'), ('Fuxing Island', 'Yangpu'), ('Siping Road', 'Yangpu'), ('Dalian Road', 'Yangpu'), ('Yangshupu Road', 'Yangpu'),

```
('Anshan Xincun', 'Yangpu'), ('Middle Yanji Road', 'Yangpu'), ('Wujiaochang', 'Yangpu'), ('Nenjiang Road', 'Yangpu'), ('Xiangyin Road', 'Yangpu'), ('Tongji University', 'Yangpu'), ('Sanmen
Road', 'Yangpu'), ('Xinjiangwancheng', 'Yangpu'), ('Jiangwan Stadium', 'Yangpu')]
```

5. Discussion

With the above clustering of Shanghai metro stations, and together with the real estate price of the district that the station belongs to, we can have a much useful guide for our business commercial plan reference when we want to open some new business stores close to the Shanghai metro stations;

We can study these data analysis output with machine-learning assisted result to help our decision in multiple perspective, here is some of the thoughts & applications as examples;

5.1 Differentiated competition

When we want to open a new business store, one strategy is to avoid the direct competition with those spots where the same kind of business stores are already fully deployed, with the convenient metro networks in Shanghai, people are easy to go from one place to another quickly, so we can much utilize this and choose a different metro station that we can avoid much of the competition for those same kind of consumption demands;

For example, if we want to open one shopping center, we should find the chance in the metro stations other than cluster1 above; Shopping centers can accommodate many kind of shops so usually too many shopping centers in one location , or near one metro station will decrease the business efficiency due to the load balance the consumers and dilutes the profit & gains;

For another example, if we want to open one coffee shop, we should avoid to open it in cluster0 where most common shops in these metro stations are already selling coffee; We can plan the coffee shop in other clusters, or opening another drink shop in cluster0, such as juice bar, Tea shop, etc;

5.2 Benefits from business cluster effect

On the contrary , we can take advantage the business cluster effect some time to get benefits; When one metro station has already certain kinds of business category in place and is attracting target consumers, the more gathering of the same category shops can bring additional opportunities when the metro location becomes well known for the specific business shop type;

For example, if we want to open one Asian restaurant, we may seek the chance in metro station cluster 3 where more Asian restaurants are available than other clusters; As in this cluster more Asian restaurants are opened nearby the metro station, it can be favorable for people to search this specific food flavors in those metro stations' areas;

5.3 Explore the insufficient business offering

In total we can see more spread restaurants and coffee shops in all the 5 metro station clusters, but less entertainment facilities are observed, especially in cluster 0,3,4; In these clusters we can utilize the already existing coffee shops, restaurants, hotels to attract people to entertain themselves there, we can think of opening movie theatres, Gym/fitness in these metro stations;

As another example, the convenient store is not sufficient in cluster 2,3,4, where we can think of the chances to open there to facilitate the fast-buying of small things for daily life;

5.4 Integration business planning with real estate cost information

In the above Shanghai metro station clustering map, the real estate price level information is visualized together, so we can find both the cluster the metro station belongs to, and the real estate cost where it is; This gives the business planner a clear indicator of the properties' cost which is a substantial cost part of opening the shops;

So when it is planed to open one shop category nearby the metro stations, we can check its popularity based on the metro station's clustering information, depending the characteristic of the category, we may adopt the strategy of either differentiated competition in 5.1, or business cluster effect benefit in 5.2 to choose the right cluster, then we can double check the real estate price level in the same clusters to find those most optimal stations based on best ratio of profit return to total costs;

Also, we can study those less expensive district and find the metro station clusters there with insufficient deployment of specific shop categories, and planning the relevant new offering for these shop categories so as to get the most optimal return with lower costs;

6. Conclusion

This is the project that I utilized almost all the knowledge and skills I learned from the "Applied data science" courses, even beyond; Among which the following key subjects are covered:

- Data source exploration and fetch/extract (by webscraping skills with python library)
- Data collection, cleaning and processing with dataframe from python pandas library
- Data visualization with python matplotlib library
- Machine learning application (k-means clustering) with python sklearn library
- Location data visualization on map with Python folium library
- Venue data obtain based on geography longitude/latitude with Foursquare API

I selected the target of the project which mandatory requires the Foursquare API to utilize the location information, the metro station of Shanghai is a good subject for me because Shanghai is my home city and I witnessed the city's growth, the fast development of its commercial environment and even faster deployment of the public transportation system, I understand deeply the convenient transportation system over the whole city brings more chances for business development, it is a good topic to be explored in this project;

However still there are quite several limitation in my project implementation that impact the accuracy of the result;

First, the Foursquare API can provide much limited venue information for China and Shanghai, when total metro station in Shanghai reaches to 344, all the venue information with radius of 500 meters nearby each station returns total 3196 locations only, which is much less than the actual venues in the places; This should much limit our data sufficiency to work for our target;

Second, the real estate price is handled with each district in Shanghai, however, it is a rough representation of the real estate price close to each metro station, in reality even in the same district, the price can be various much for the commercial property;

Third, for the individual metro station, its location in specific business centers or not are not taken into account, and some stations are the metro line's junction , these factors will increase its popularity as well as its business cost, it is not considered so as to reduce the complexity;

So in summary, the project used a simplified model with all the knowledges/skills learned from the "Applied Data science" course , but anyhow the approach I used make my course studied looks much valuable even the result can be not accurate enough;