

My Capstone project:

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

*By ShengJie(Kenny) / 2020 June*



# Introduction of Shanghai

- Shanghai is the economics capital of China. In terms of GDP volume , consumer's capacity as an individual city, it has longly been the top one among all cities in China; quoted from Wikipedia “ *With a population of 24.28 million of 2019, it is the most populous urban area in China and 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;

# Introduction of Shanghai's metro system



The Shanghai metro system is the world's second largest metro system by route length, reaching total 676Kilometers (420mi), it is also the second largest by the number of metro stations, meanwhile it ranks second in the world by annual ridership;

# 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:
  - more than half of the people movements within the city is by metros,
  - metro stations provide fast reachability so people are willing to go there when shopping is considered;
  - 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, 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;

target people who are interested in the project

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;

# Data of Shanghai metro stations

WIKIPEDIA The Free Encyclopedia

Main page  
Contents  
Current events  
Random article  
About Wikipedia  
Contact us  
Donate

Contribute  
Help  
Community portal  
Recent changes  
Upload file

Tools  
What links here  
Related changes  
Special pages  
Permanent link  
Page information  
Wikidata item  
Cite this page

Print/export  
Download as PDF  
Printable version

Languages  
Français  
Nederlands  
עברית  
★ 中文

Article Talk

## List of Shanghai Metro stations

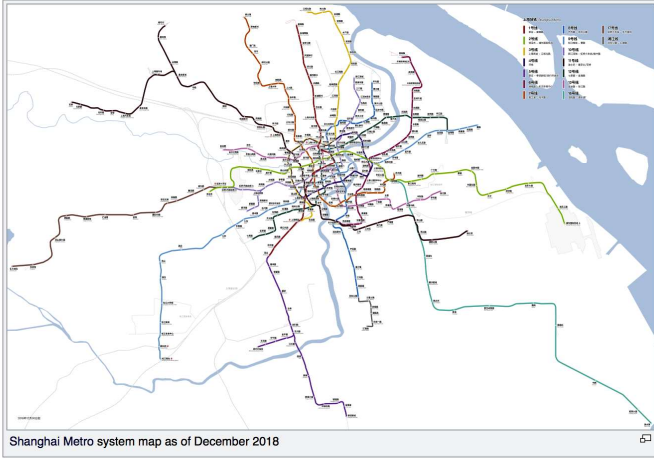
From Wikipedia, the free encyclopedia

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<sup>[1][2][3][4][5]</sup> and second largest by number of stations.

The tables below contain the **413 stations** on the Shanghai Metro<sup>[6]</sup> operational as of December 2018, 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]

- Line 1
- Line 2
- Line 3
- Line 4
- Line 5
- Line 6
- Line 7
- Line 8
- Line 9
- Line 10
- Line 11
- Line 12
- Line 13
- Line 16
- Line 17
- Pujiang line
- Notes
- References



Shanghai Metro system map as of December 2018

[https://en.wikipedia.org/wiki/List\\_of\\_Shanghai\\_Metro\\_stations](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, 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;

# Data of map & real estate price

## **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
- ✓ converting the Arcgis data format to the GeoJson data format by online tool:

## **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;



# Venue data of metro station shops

- With all the Shanghai metro station data obtained ( longitude, latitude), by calling Foursquare API for each metro station with radius of 500 meters and limit of 100 venues, all the venue data with foursquare free account are obtained;
- The data extracted out are venue name, venue latitude, venue longitude and venue category; In this project, I am particularly 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;



# 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
----------------------	--------------	----------	-----------

# Input to metra\_sta pandas dataframe

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

# Obtaining the real estate price information from the web

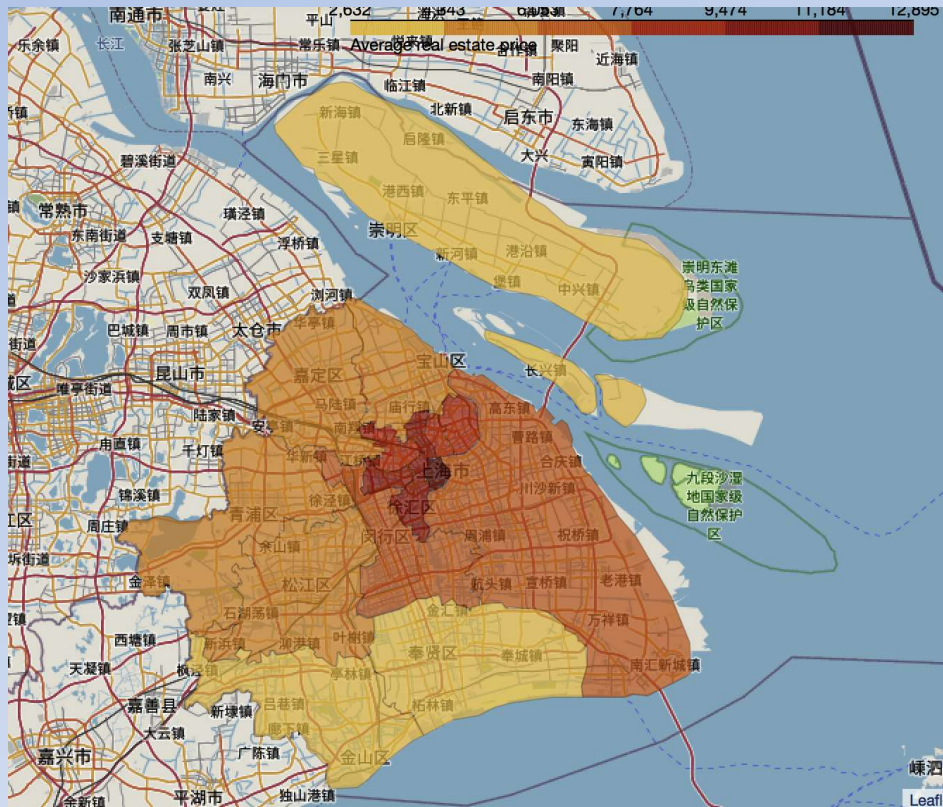
- The average real estate price information is obtained from the website of one big China real estate trade agency .
- The data is simple and brief, so it is put in one csv file and read into one pandas dataframe
- Total 16 districts of Shanghai are listed, the latest data of May/2020 when this report 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

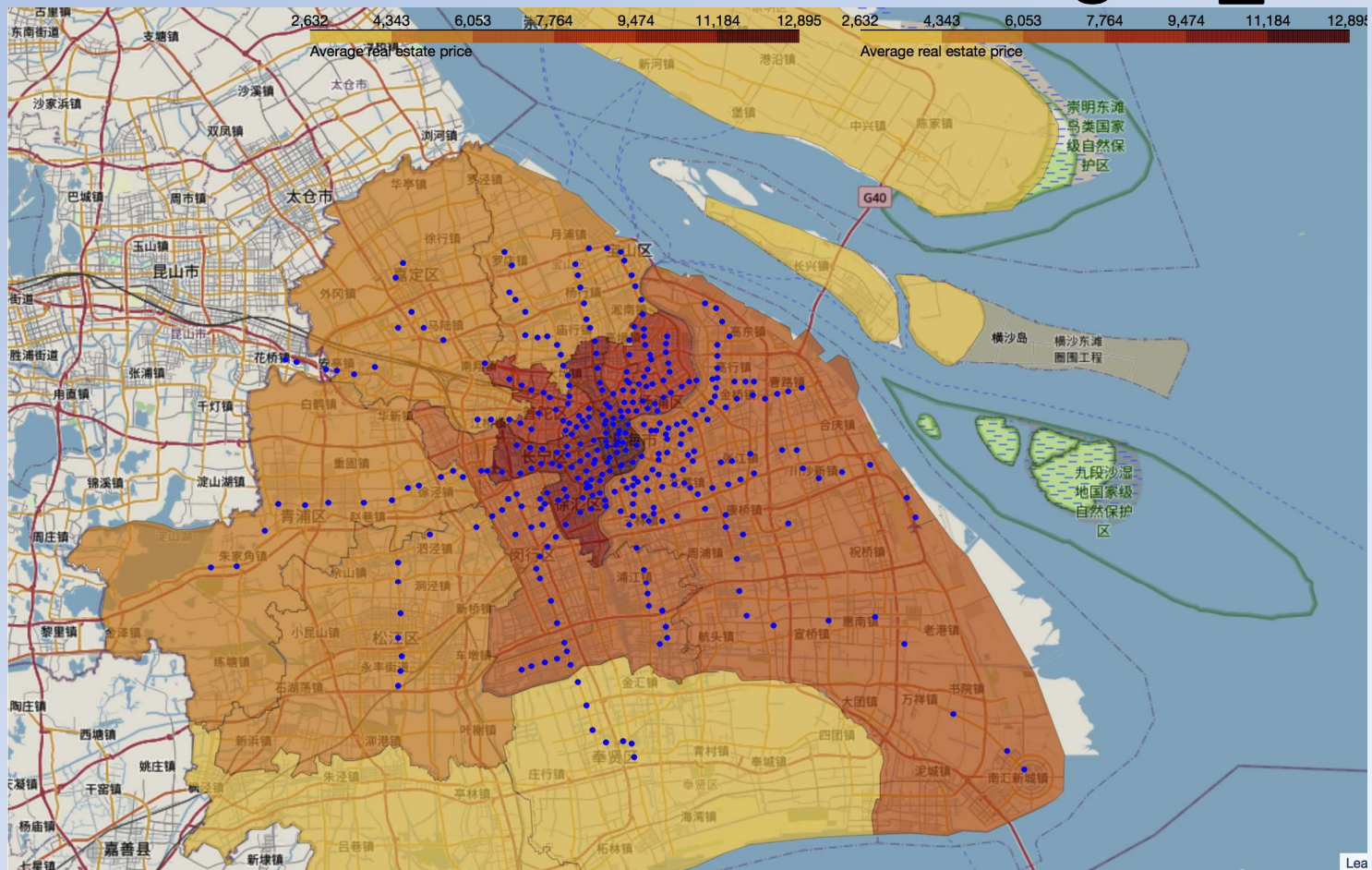
# Generate a Shanghai\_map of all districts each with its real estate price info.



- ✓ Using folium.Map.Choropleth method, with
  - input of Shanghai district borders geodata information dataframe;
  - the real estate price dataframe of each district
- ✓ The price level of each district of Shanghai is visualized, the darker of the color, the more expensive of the real estate;



# Add all metro stations to the shanghai\_map



# Obtaining all venue information of each metro station

```
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)
```

## 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
```



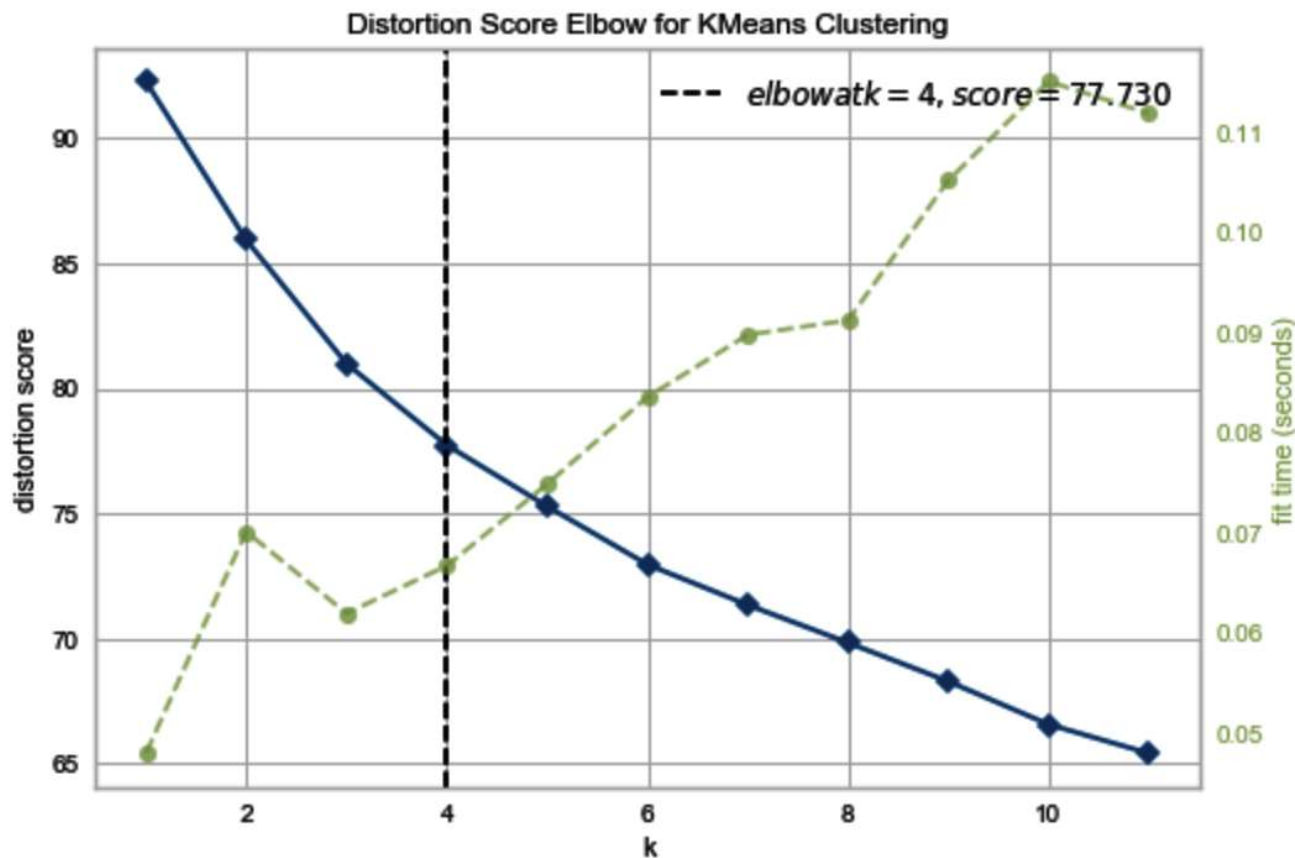
# Get the 10th most common venue of each metro station

```
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

# Perform clustering on the data by using k-means clustering



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

# Perform clustering on the data by using k-means clustering

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
igfu 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
ulan 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
in 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
...	...	...	...	...	...	...	...	...	...	...	...	...	...
gwancheng	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 Joint	Fruit & Vegetable Store	Food	2

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

## Perform clustering on the data by using k-means clustering

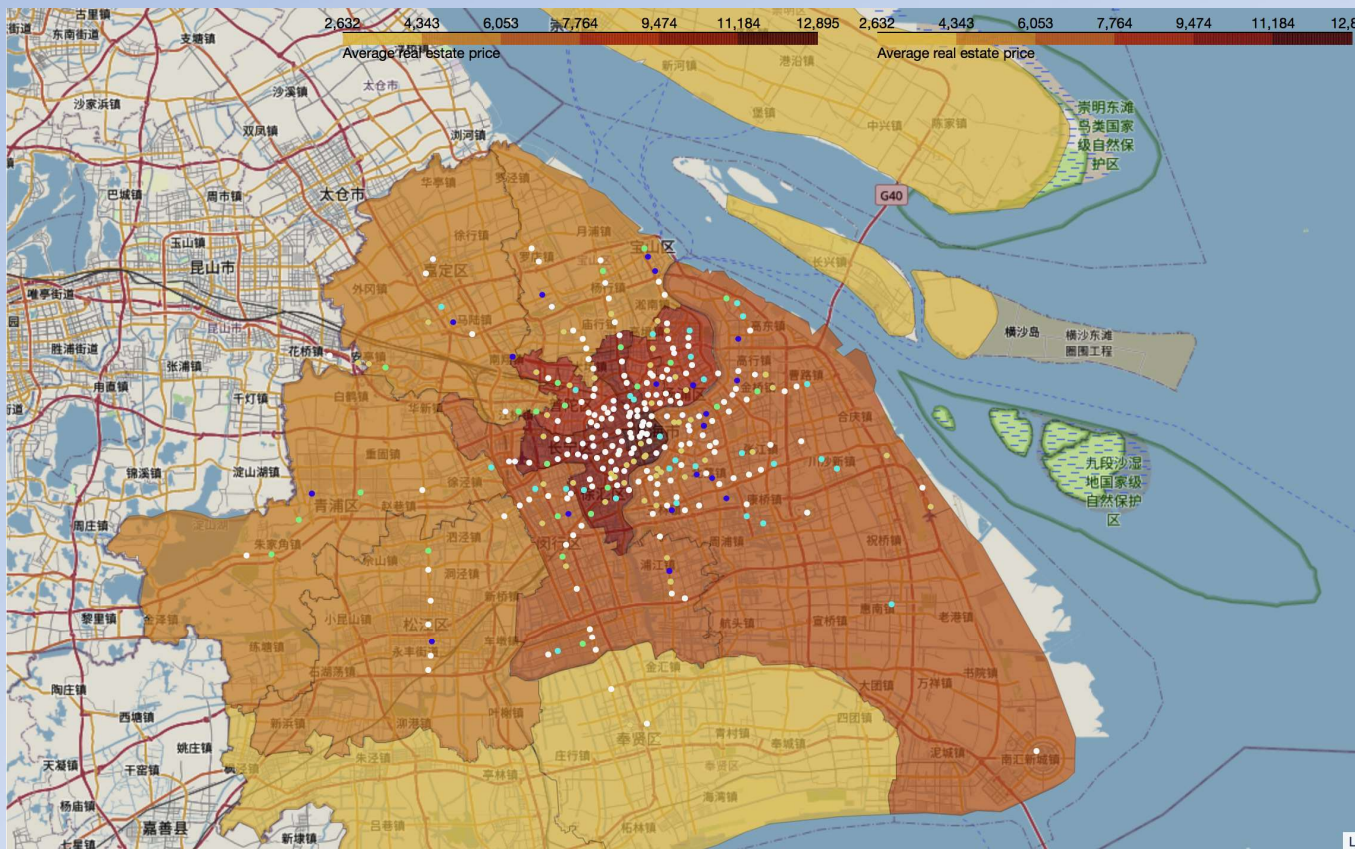
```
In [61]: Shanghai_merged.groupby('Cluster Label')['Station'].count()
```

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

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

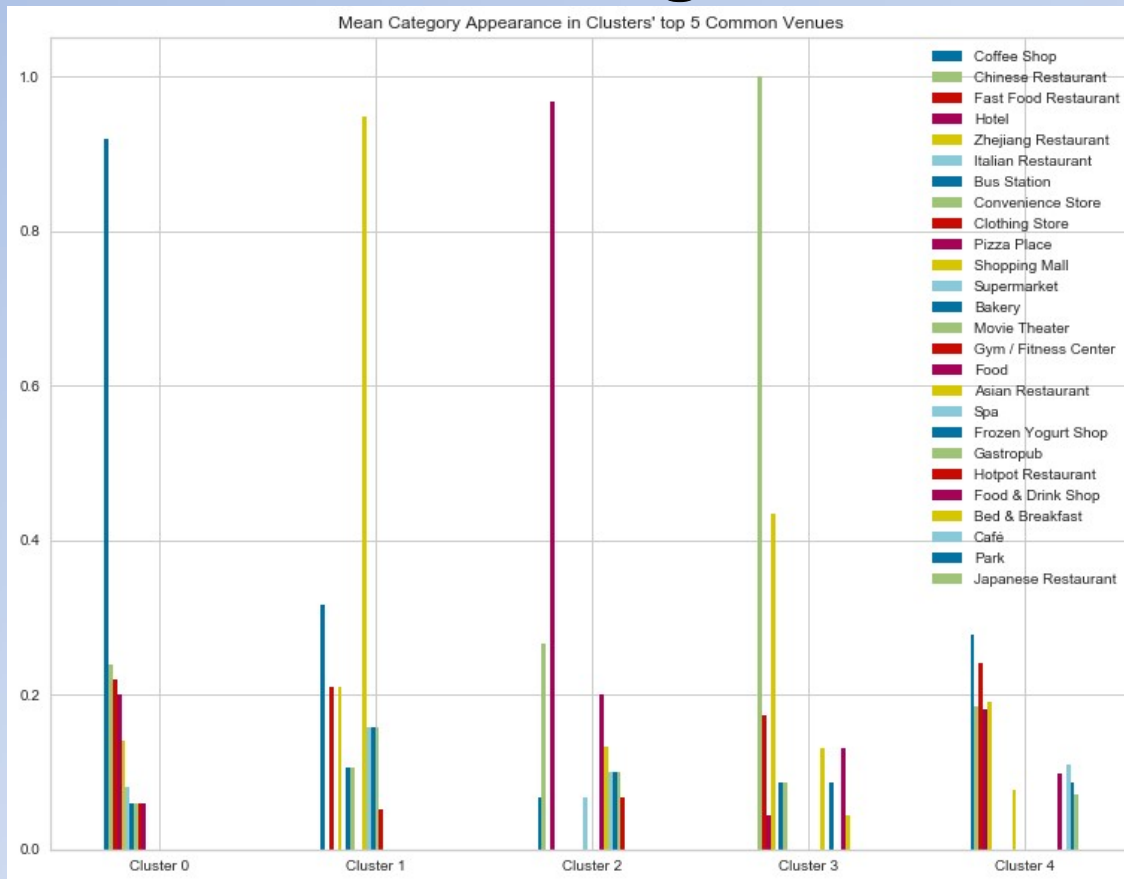


# Perform clustering on the data by using k-means clustering



Then we can visualize all the metro stations with different color based on its belonging cluster on the shanghai\_map

# Perform clustering on the data by using k-means clustering



most common top  
10 venues/shops of  
each cluster

# Perform clustering on the data by using k-means clustering

## **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;



# Perform clustering on the data by using k-means clustering

## **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;

## **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;

# Perform clustering on the data by using k-means clustering

## **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

## **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 existence of other business types including entertainment, shopping centers , etc;
- There are 183 stations belonging to this cluster

# Discussions

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;

**Differentiated competition**

**Benefits from business cluster effect**

**Explore the insufficient business offering**

**Integration business planning with real estate cost information**

# 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 geograph longitude/latitude with Foursquare API

# Conclusion

## Limitations

- The Foursquare API can provide much limited venue information for China and Shanghai, this should much limit our data sufficiency to work for our target;
- The real estate price is handled with each district in Shanghai, however, in reality even in the same district, the price can be various much for the commercial property;
- For the individual metro station, its location in specific business centers or at 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;
- Other factors are not considered and only the simplified model is built with all learned knowledge/skills , to be improved;