Report of my Capstone Project of the IBM Data Science Certification Course

Abstract

This is the report of my capstone project of the IBM Data Science Certification Course. In this project I use Python data analysis , Python data visualization and K-means clustering machine learning algorithm to solve the problem that how to help the buyer make decisions when buying a house in my residence city - Nanjing, China. I completed this project on Jupyter Notebook running on IBM Watson Studio.

1. Introduction

- Background

Nanjing is the capital of Jiangsu province of China and the second largest city in the East China region. Nanjing has served as the capital of various Chinese dynasties, kingdoms and republican governments in Chinese history. Till 2017, the total population of Nanjing was 8.335 million. As a resident of this city, I decided to use Nanjing in my project.

In recent years, Nanjing has been developing its economy, commerce, industry, as well as city construction. More and more people are attracted to work and settle in Nanjing. To own one or more houses in Nanjing is the dream of most of them. Usually people have different purpose of buying a house. Some of them would like to live in a house in a good neighborhood which can bring them a convenient life. Others may consider it as an investment with an expectation of price appreciation. So, I am willing to use my project to help them make

decisions based on data analysis when buying a house with different purposes in Nanjing.

- Data Description

Firstly, I need the housing price history data of Nanjing. There are some housing sales agencies in China who place the historical housing price of each residence community(or neighborhood) in most cities in their web sit. I can use Beautifulsoup to grasp all the historical housing price of Nanjing. Fortunately, someone has already done this and I can download it directly from Internet. From this data, I can analyze the average price growth rate from the given years of each residence community, predict the average price in the next year, and recommend the top 10 communities to the buyers.

After that, based on the communities data, I can use Baidu API to acquire their coordinate data (Baidu API is the most suitable tool for acquiring the latitude and longitude of a given place in China). To simplify the problem, select top 100 communities who have the most housing deal records, and use Forsquare API to acquire the most common nearby venues data of each community. Then I use Python Pandas DataFrame to consolidate these data, use K-means clustering algorithms to seperate these 100 most popular communities to 3 clusters based on the nearby venues data. At last analyze their respective characteristics and give a understandable label, then recommend them to the different buyers with different preferences.

2. Methodology

I downloaded housing price data(csv format file) of Nanjing from the year 2012 to 2017 in Internet as mentioned in the previous section. After removing some duplicated and useless data, I stored the cleaned data on my Github project. Then I used pandas dataframe to read the data from the file URL.This loaded my master data is looked as below.

	district	Avenue	Neighborhood	dealYear	totalPrice	unitPrice	quotedPrice	layout	area
0	雨花台	能仁里	凤凰和美	2017	10.0	7640	17.0	1室0厅	13.09
1	雨花台	小行	名城世家花园	2016	10.3	7613	10.3	1室1厅	13.53
2	雨花台	小行	名城世家花园	2017	10.5	7761	11.0	1室0厅	13.53
3	雨花台	小行	名城世家花园	2016	10.5	7761	10.5	1室0厅	13.53
4	雨花台	小行	名城世家花园	2015	10.5	7761	10.5	1室1厅	13.53

Here I use unitPrice for my analysis. The unitPrice means the housing price per square meter, which is common used in China housing market.

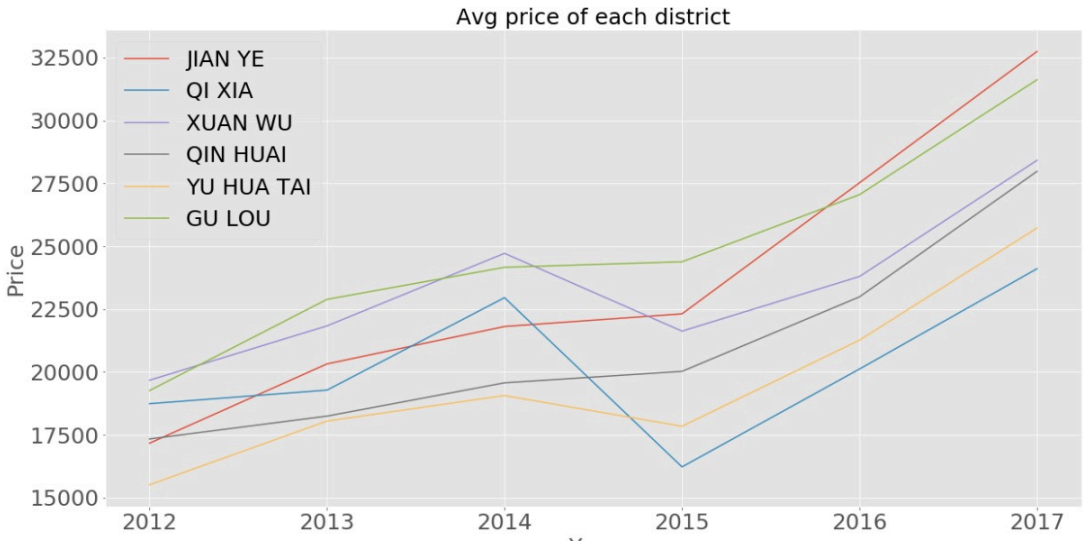
The contents in the first 3 columns are Chinese characters. For better understood by Non-Chinese readers, I converted district and Neighborhood columns to Pinyin, which is the Chinese phonetic alphabet written in English alphabet. the library pypinyin in Python can help with this.

	Neighborhood	NEIGHBORHOOD_PY	district	DISTRICT_PY
0	怡景花园	YI JING HUA YUAN	鼓楼	GU LOU
1	瑞金北村	RUI JIN BEI CUN	秦淮	QIN HUAI
2	名仕嘉园	MING SHI JIA YUAN	建邺	JIAN YE
3	育才公寓	YU CAI GONG YU	鼓楼	GU LOU
4	银城花园北片	YIN CHENG HUA YUAN BEI PIAN	鼓楼	GU LOU
5	芳草园	FANG CAO YUAN	鼓楼	GU LOU
6	金舟花园	JIN ZHOU HUA YUAN	鼓楼	GU LOU
7	新河一村	XIN HE YI CUN	鼓楼	GU LOU
8	王府园小区	WANG FU YUAN XIAO QU	秦淮	QIN HUAI
9	华阳佳园华彩苑	HUA YANG JIA YUAN HUA CAI YUAN	鼓楼	GU LOU

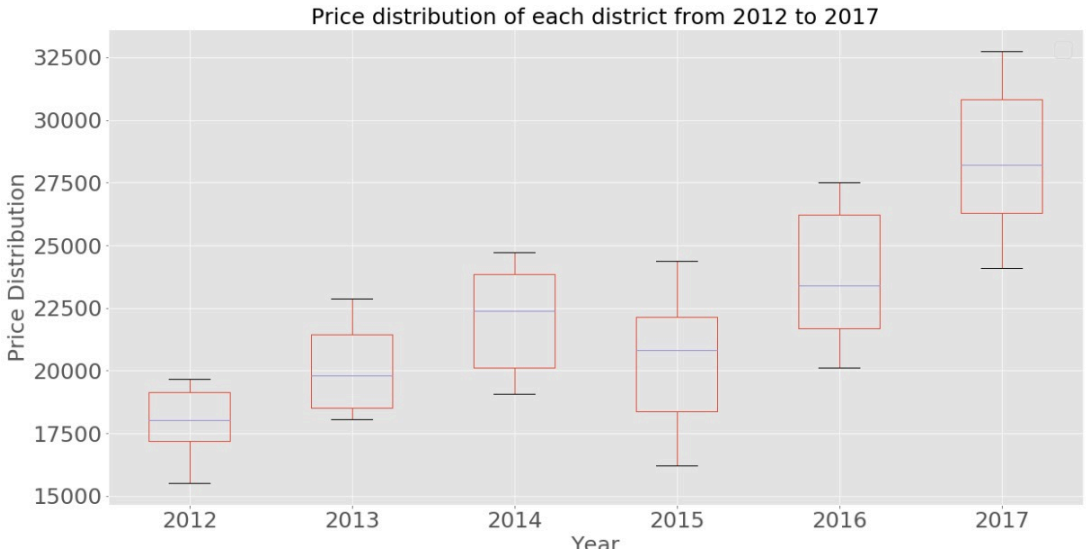
Here is the data grouped by district, shows the average house price of each district from the year 2012 to 2017. For simplicity I dropped those district that has a NaN.

	district	DISTRICT_PY	2012	2013	2014	2015	2016	2017
0	建邺	JIAN YE	17168	20321	21809	22311	27524	32745
1	栖霞	QI XIA	18739	19279	22955	16226	20118	24106
2	玄武	XUAN WU	19673	21834	24718	21625	23797	28418
3	秦淮	QIN HUAI	17336	18246	19569	20024	22988	27978
4	雨花台	YU HUA TAI	15516	18049	19063	17841	21267	25725
5	鼓楼	GU LOU	19262	22888	24163	24382	27052	31621

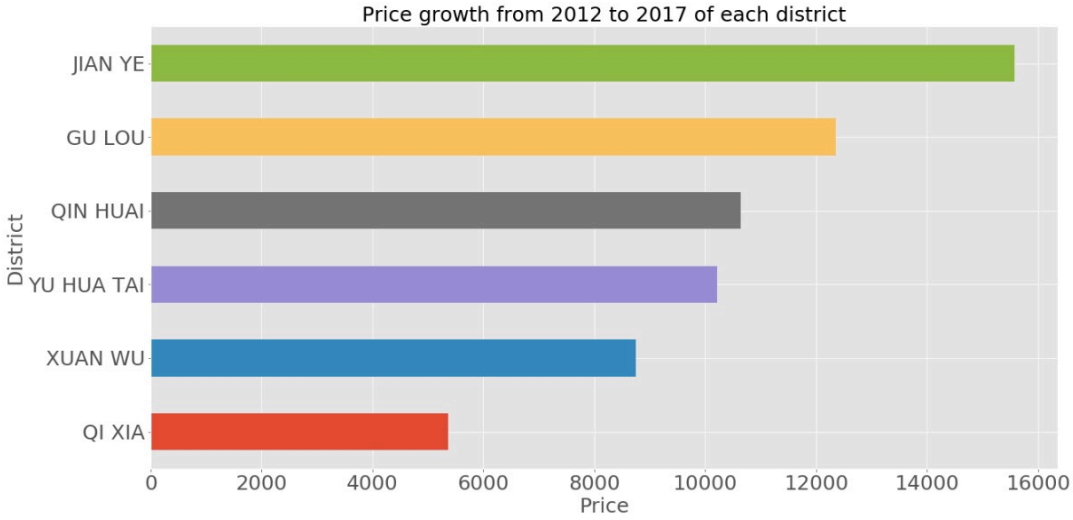
We can use a line chart to show the average price of each district in one chart and understand their trends.



We can also draw a box chart to display the average price distribution.



So, which district has the most average host price growth from the year 2012 to 2017? we can see below chart.



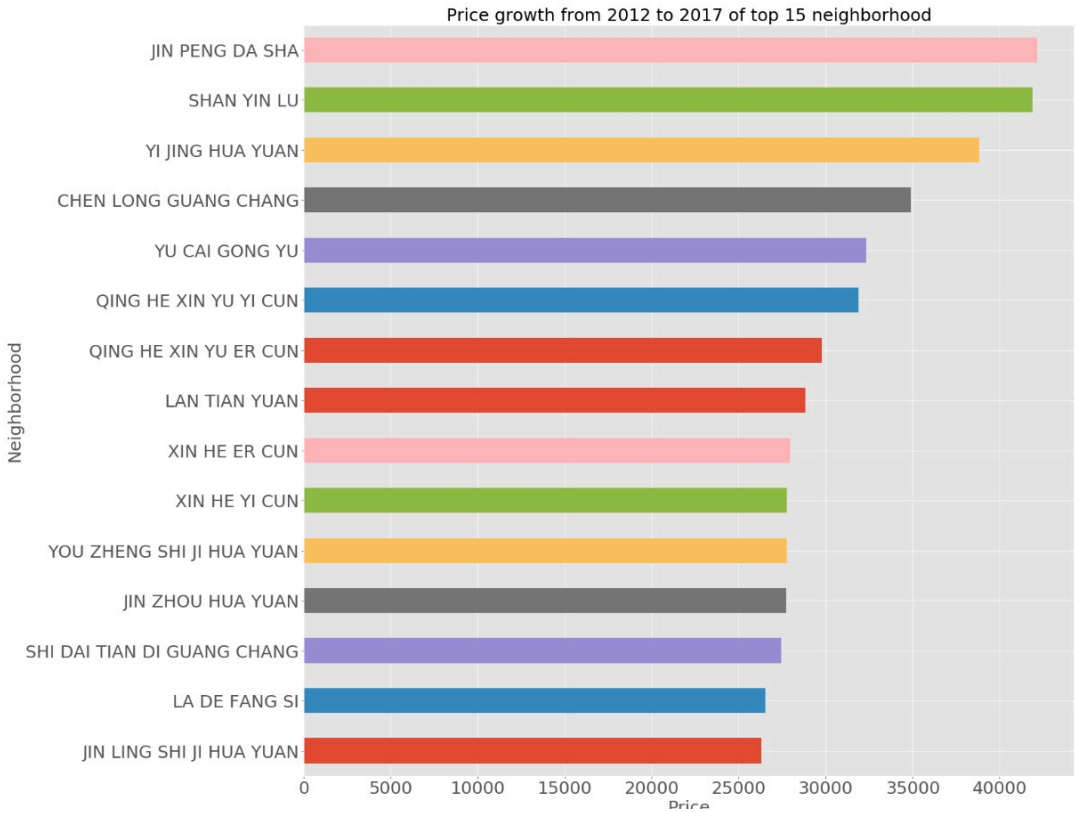
Yes, JIAN YE district. The average house price in JIAN YE district has increased about 15500CNY per square meter from the year 2012 to 2017.

But, for a house buyer, he usually want to know what exactly are the residence communities (neighborhoods) which have the most price growth. So I get the price growth data for each neighborhood and sorted it with the price growth from high to low.

NEIGHBORHOOD_PY	DISTRICT_PY	2012	2013	2014	2015	2016	2017	growth
JIN LING SHI JI HUA YUAN	GU LOU	21297	28106	27757	32159	41619	47580	26283
LA DE FANG SI	JIAN YE	18256	23520	26172	30144	39838	44775	26519
SHI DAI TIAN DI GUANG CHANG	GU LOU	21303	26100	31484	34287	54212	48759	27456
JIN ZHOU HUA YUAN	GU LOU	21426	25853	28022	28831	41930	49173	27747
YOU ZHENG SHI JI HUA YUAN	GU LOU	23432	29818	30787	32519	44721	51198	27766
XIN HE YI CUN	GU LOU	15038	23921	25600	28909	37018	42821	27783
XIN HE ER CUN	GU LOU	13315	23076	25644	27806	36687	41279	27964
LAN TIAN YUAN	GU LOU	26579	31733	34308	33898	43004	55422	28843
QING HE XIN YU ER CUN	GU LOU	19984	30746	32581	32955	45701	49786	29802
QING HE XIN YU YI CUN	GU LOU	20648	28252	31612	33421	46344	52532	31884
YU CAI GONG YU	GU LOU	21491	28752	31886	33398	45566	53812	32321
CHEN LONG GUANG CHANG	GU LOU	18934	25734	27348	31353	46095	53858	34924
YI JING HUA YUAN	GU LOU	22264	30023	34049	32071	40118	61114	38850
SHAN YIN LU	GU LOU	26796	30383	41847	42686	47399	68669	41873
JIN PENG DA SHA	GU LOU	19942	30119	27767	29188	41830	62109	42167

What interesting thing do you find from the above data? Although the highest average price growth is in JIAN YE district, but 14 of the top 15 neighborhoods with the highest house price growth are located in GU LOU district.

Here I use barh chart to show the price growth.



Next, I will get the coordinate of each neighborhood. To simplify it, I will select the top 100 neighborhoods which has the most housing deal records. It means these 100 neighborhoods are most concerned by the house buyers. I use Baidu API to get the latitude and longitude of a neighborhood name. At first you need to register an account at Baidu Map, then acquire a developer key that will be used in your request. After merging the coordinate data to the selected top 100 neighborhoods data, I got below data.

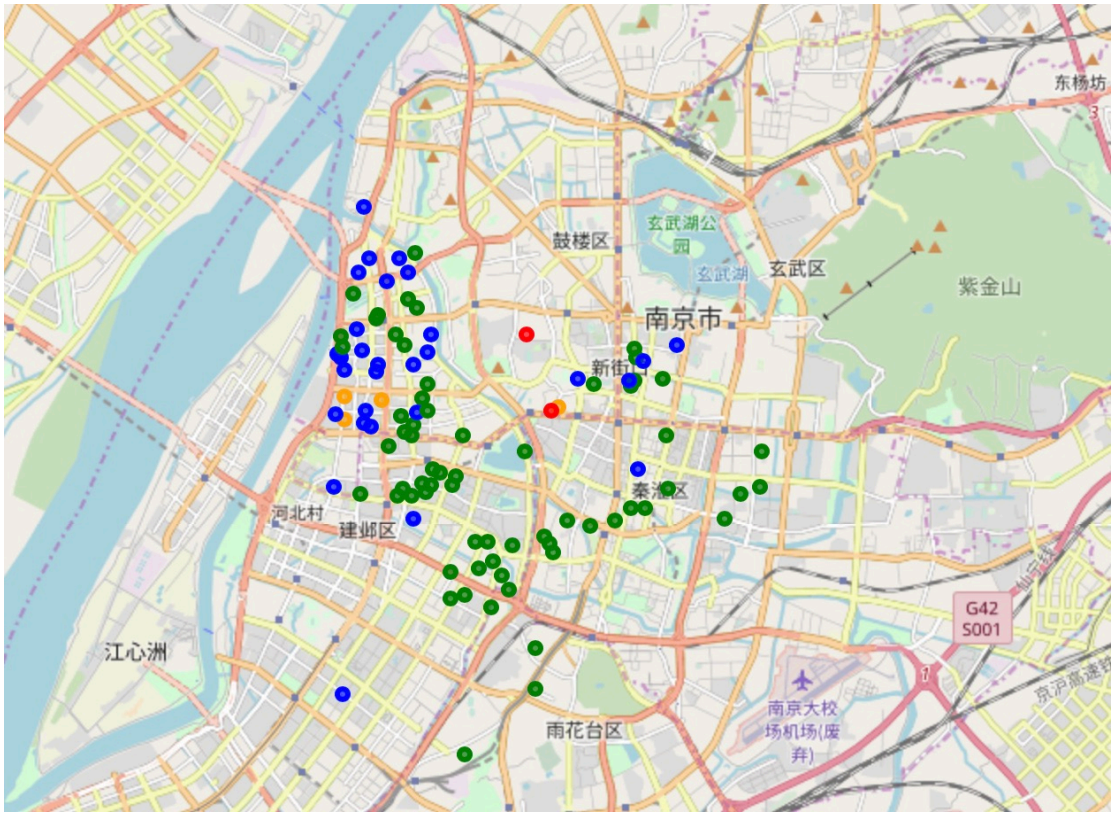
	Neighborhood	NEIGHBORHOOD_PY	district	DISTRICT_PY	Latitude	Longitude	2012	2013	2014	2015	2016	2017	growth
0	怡景花园	YI JING HUA YUAN	鼓楼	GU LOU	32.049894	118.779258	22264	30023	34049	32071	40118	61114	38850
1	瑞金北村	RUI JIN BEI CUN	秦淮	QIN HUAI	32.042576	118.818969	17860	22406	25557	25348	29892	35167	17307
2	名仕嘉园	MING SHI JIA YUAN	建邺	JIAN YE	32.043443	118.746524	16646	18898	20292	18898	24098	29535	12889
3	育才公寓	YU CAI GONG YU	鼓楼	GU LOU	32.051603	118.737733	21491	28752	31886	33398	45566	53812	32321
4	银城花园北片	YIN CHENG HUA YUAN BEI PIAN	鼓楼	GU LOU	32.057935	118.737041	17925	22199	25931	26751	36651	41731	23806
5	芳草园	FANG CAO YUAN	鼓楼	GU LOU	32.061804	118.754646	23051	30671	28753	30167	37693	48447	25396
6	金舟花园	JIN ZHOU HUA YUAN	鼓楼	GU LOU	32.056846	118.744352	21426	25853	28022	28831	41930	49173	27747
7	新河一村	XIN HE YI CUN	鼓楼	GU LOU	32.074297	118.742595	15038	23921	25600	28909	37018	42821	27783
8	王府园小区	WANG FU YUAN XIAO QU	秦淮	QIN HUAI	32.033224	118.793586	14912	16379	19117	21022	27379	33195	18283
9	华阳佳园华彩苑	HUA YANG JIA YUAN HUA CAI YUAN	鼓楼	GU LOU	32.051364	118.753064	18844	21549	24148	23253	27035	33286	14442
10	凤凰花园城静幽园	FENG HUANG HUA YUAN CHENG JING YOU YUAN	鼓楼	GU LOU	32.048406	118.748848	19189	22462	25290	26594	31939	37493	18304

While the coordinate is ready, I can show these 100 neighborhoods in

a map. I use folium map to display it with different color. Below are the classification of the used colors.

For the price growth:

- Red: >= CNY40000
- Orange: CNY30000~CNY40000
- Blue: CNY20000~CNY30000
- Green: <CNY20000



Then, I will find the nearby venues of each neighborhoods base on the coordinate by using foursqure API. I set the limit number of found venues to 100 and set the radius to 500. Through this API I got a json format response data. Parse the data and store it into a pandas dataframe. For each neighborhood, the returned nearby venues are looked like below.

	name	categories	lat	lng
0	Blue Frog (蓝蛙)	Burger Joint	32.049356	118.778968
1	Apple Nanjing IST (Apple 南京艾尚天地)	Electronics Store	32.047783	118.779065
2	Costa Coffee (Costa Coffee (咖世家))	Coffee Shop	32.052492	118.780254
3	Wagas 沃歌斯	Sandwich Place	32.049189	118.778826
4	Deji Plaza (德基广场)	Shopping Mall	32.046455	118.779676
5	Element Fresh (新元素)	New American Restaurant	32.049268	118.778835
6	大渔铁板烧 Tairyo Teppanyaki	Japanese Restaurant	32.046840	118.778955
7	IST Mall (艾尚天地)	Shopping Mall	32.048175	118.779109
8	Poets Restaurant & Lounge	Thai Restaurant	32.048617	118.779280
9	Starbucks (星巴克)	Coffee Shop	32.045406	118.779064
10	Costa Coffee (咖世家)	Coffee Shop	32.046432	118.779308
11	BHG Market Place 高级食品超市	Food & Drink Shop	32.047360	118.779332
12	McDonald's (麦当劳)	Fast Food Restaurant	32.052519	118.779525
13	中心大酒店 Central Hotel	Hotel	32.047669	118.778437
14	Golden Eagle Mall (金鹰天地)	Shopping Mall	32.052692	118.779251

The return data contains the venues name and categories. Here I will only use categories for data analysis.

Use Forsqure API to find venues for each of the top 100 neighborhoods, then count the venue category number of each neighborhood, got below data.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
一枝园小区	2	2	2	2	2	2
万达西地二期	5	5	5	5	5	5
世茂滨江新城	1	1	1	1	1	1
丹凤新寓	2	2	2	2	2	2
五台花园	10	10	10	10	10	10
亚东国际公寓	4	4	4	4	4	4
仁园	4	4	4	4	4	4
仓顶	20	20	20	20	20	20
估衣廊	5	5	5	5	5	5
佳盛花园	2	2	2	2	2	2
健园	9	9	9	9	9	9
兆园	6	6	6	6	6	6
先锋青年公寓	4	4	4	4	4	4
凤凰花园城清溪园	4	4	4	4	4	4

Use one hot encoding to convert the venue category to binary data, then group it by neighborhood, got below data.

	Neighborhood	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bar	Beer Garden	Bookstore	...	Taiwanese Restaurant	Tapas Restaurant	Thai Restaurant	Theme Park	S
0	一枝园小区	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
1	万达西地二期	0.0	0.0	0.200000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
2	世茂滨江新城	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
3	丹凤新寓	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
4	五台花园	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
5	亚东国际公寓	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
6	仁园	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
7	仓顶	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.100000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
8	估衣廊	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
9	佳盛花园	0.0	0.0	0.000000	0.0	0.500000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
10	健园	0.0	0.0	0.000000	0.0	0.111111	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
11	兆园	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
12	先锋青年公寓	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
13	凤凰花园城清溪园	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
14	凤凰花园城金陵园	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	...	0.0	0.000000	0.000000	0.00	
15	凤凰花园城静	0.0	0.0	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000		0.0	0.000000	0.000000	0.00	

This data is ready for applying K-means algorithm. I set the number of clusters to 3, call KMeans method to get the cluster labels of each neighborhood.

In [142]:
```
# set number of clusters
kclusters = 3

nbh100_grouped_clustering = nbh100_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(nbh100_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

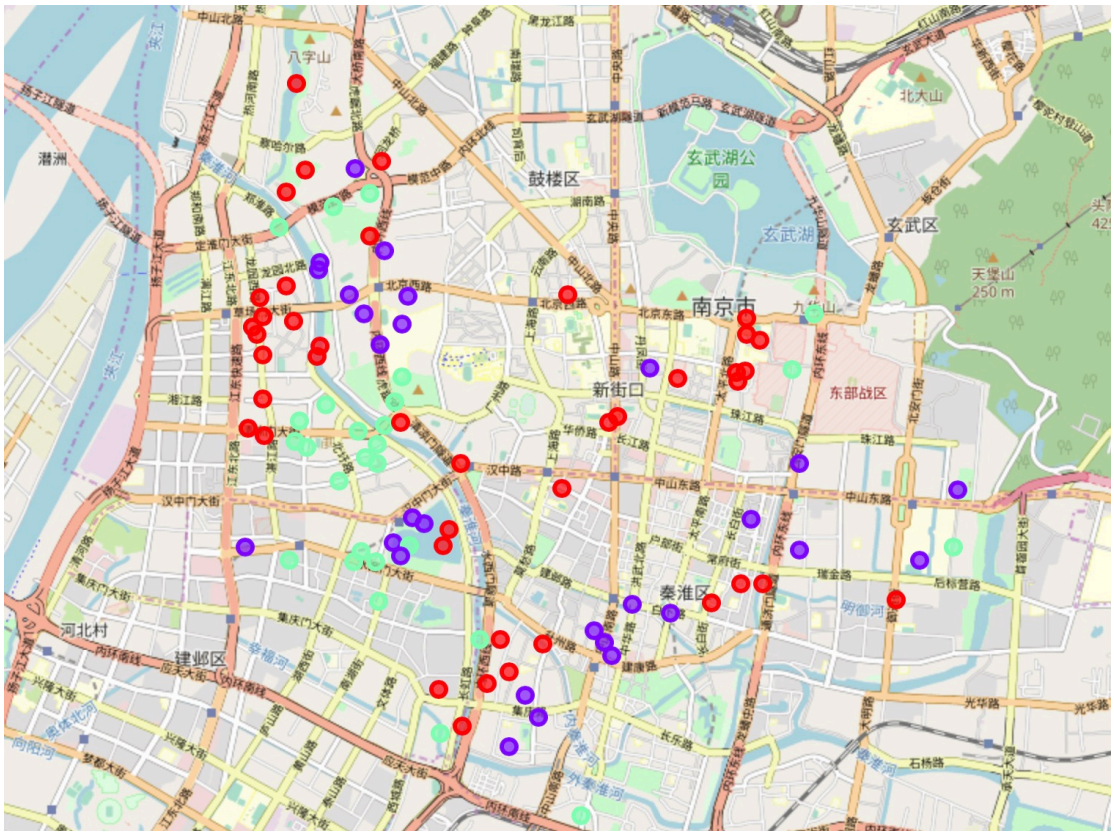
Out[142]: array([2, 2, 0, 0, 1, 2, 1, 1, 0, 0], dtype=int32)

Also, I want to know what venues categories do a neighborhood with a kind of label has. So, re-arrange the neighborhood data with the venue categories, and get the most 10 frequent venue categories for each neighborhood. Finally, merge the cluster label, got the final data, like below.

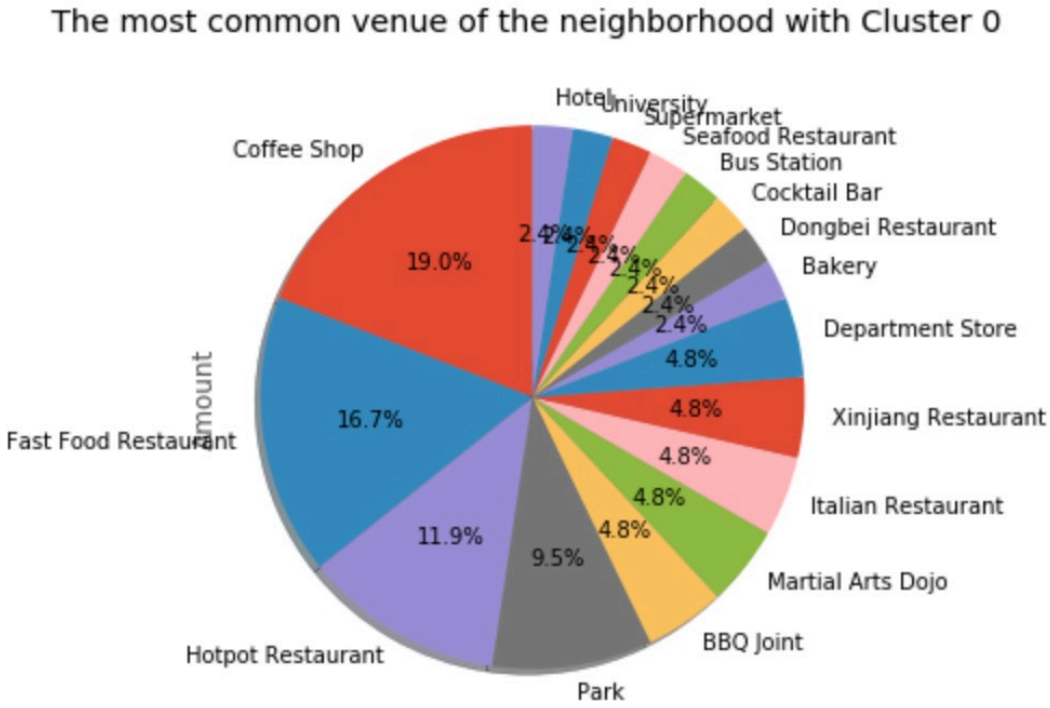
	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	2	一枝园小区	Chinese Restaurant	Xinjiang Restaurant	Fast Food Restaurant	Dessert Shop	Dim Sum Restaurant	Diner	Dongbei Restaurant	Dumpling Restaurant	Electronics Store	Farm
1	2	万达西地二期	Memorial Site	Metro Station	Asian Restaurant	Chinese Restaurant	Museum	Dim Sum Restaurant	Diner	Dongbei Restaurant	Dumpling Restaurant	Electronics Store
2	0	世茂滨江新城	Fast Food Restaurant	Xinjiang Restaurant	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dongbei Restaurant	Dumpling Restaurant	Electronics Store	Farm
3	0	丹凤新寓	Hotpot Restaurant	University	Fast Food Restaurant	Department Store	Dessert Shop	Dim Sum Restaurant	Diner	Dongbei Restaurant	Dumpling Restaurant	Electronics Store
4	1	五台花园	Hotel	Fast Food Restaurant	Metro Station	Taco Place	Café	Shopping Mall	Department Store	Coffee Shop	Food & Drink Shop	Flea Market

Use folium to display the neighborhoods with different label in different color.

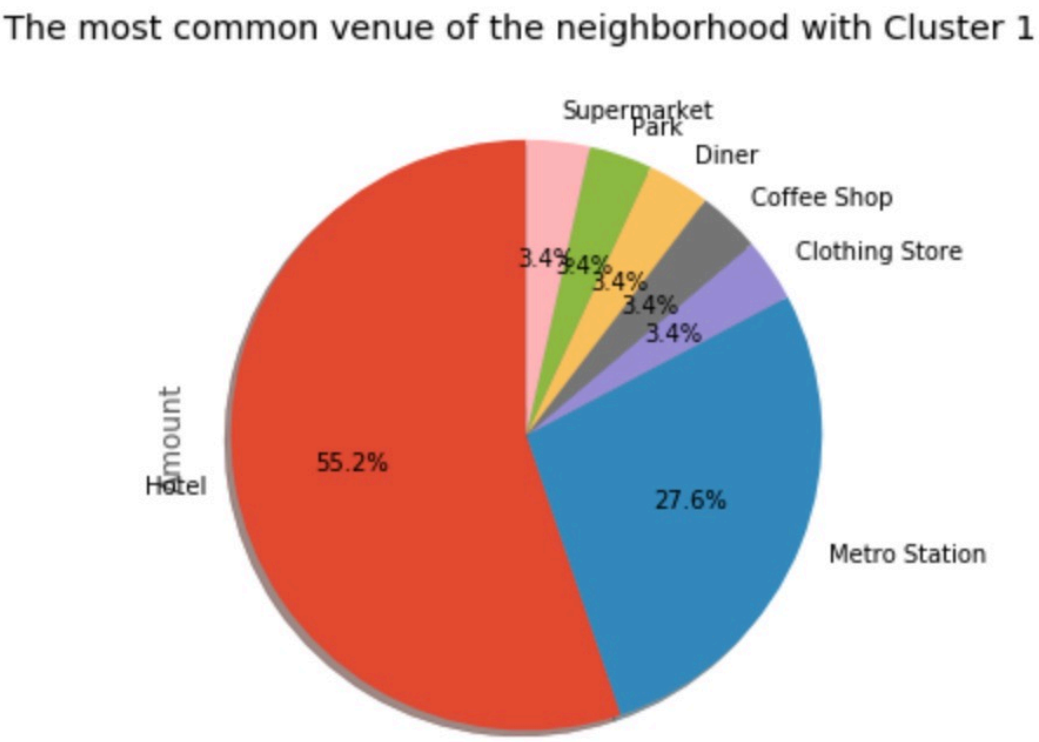
- Red: Cluster 0
- Purple: Cluster 1
- Green: Cluster 2



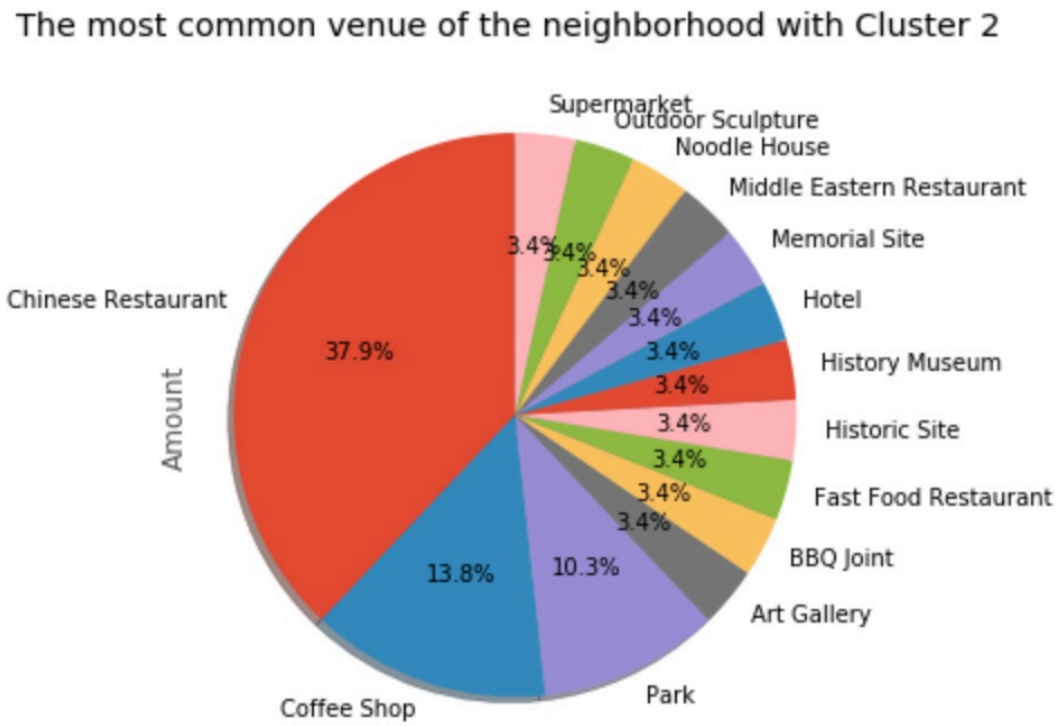
As we have separated the neighborhoods to three clusters, then what are these clusters exactly mean? Let's see the most common venues of each cluster, so we can understand it more clearly.



Cluster 0: most common venues are Coffee Shop, then Fast Food, Hotpot, Park.



Cluster 1: most common venues are Hotel, exceed 50%, then Metro Station.

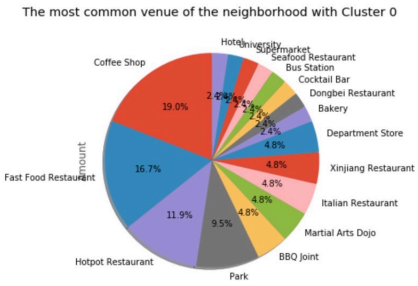


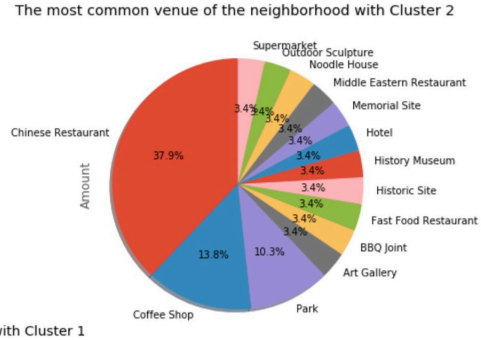
Cluster 1: most common venues are Chinese Restrautrant, then Coffee Shop and Park.

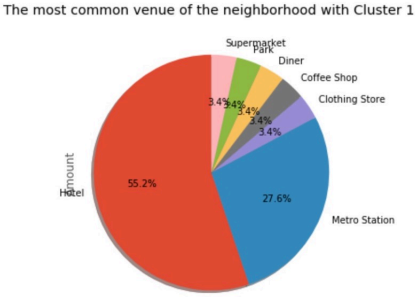
3. Result

After doing so many data analysis, now come back to the problem that I want to resolve in this project. That is how to help the buyer make decisions when buying a house in Nanjing? At the first section I mentioned there are 2 major purpose for a house buyer. One is to live in a good neighborhood which can bring them a convenient life, second is for investment.

For the first one, show the 3 clusters of the neighborhood venues pie charts.



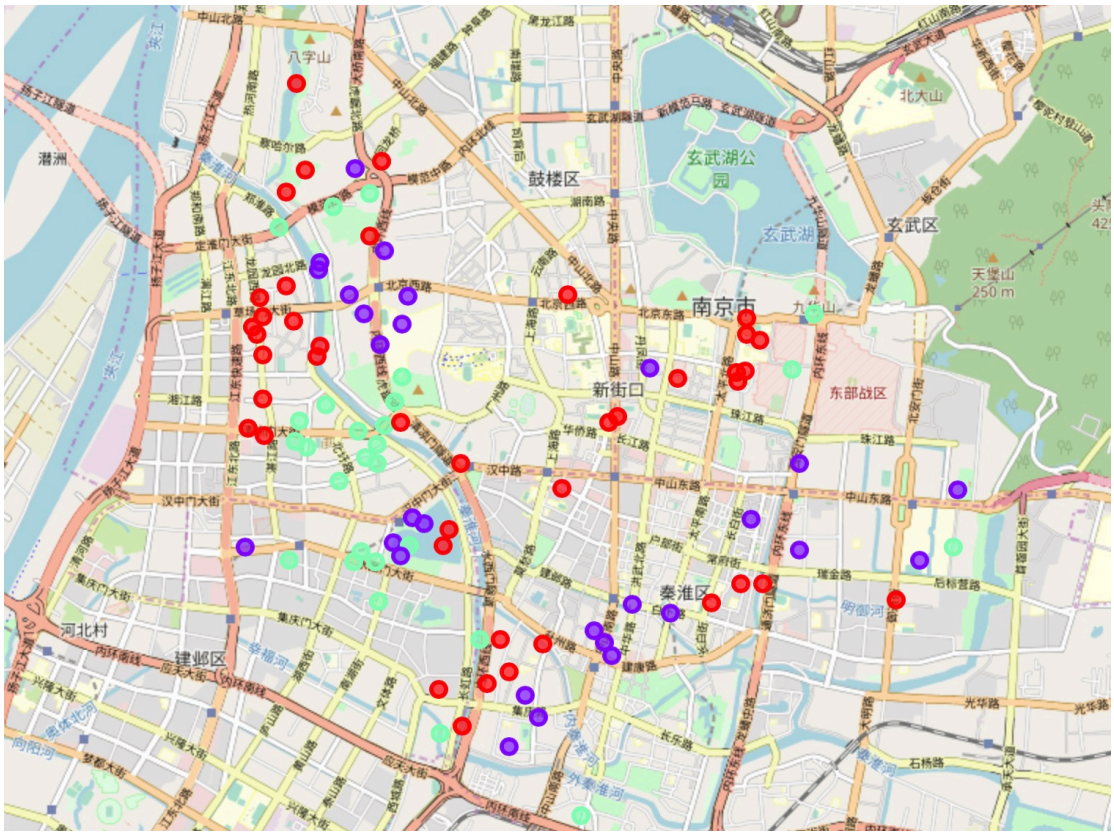




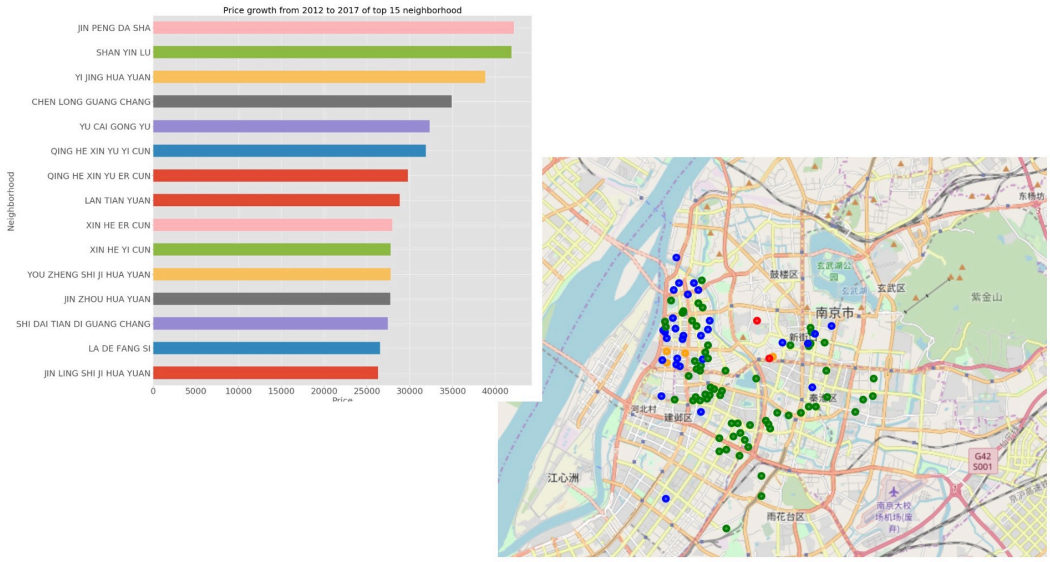
About these three cluster labels, I can name it according to it's venues information.

- Cluster 0: Recreational Type
- Cluster 1: Accommodation & Commute Type
- Cluster 2: Food Lover Type

House buyer can select what type is his preference. Also he can reference this map to select the suitable position.



For the second one, if a house buyer is for investment, then show the price growth data to him.



The buyer will know what neighborhoods have the most price growth, and he can also find in what area the house price growth higher.

4. Discussion

In this project, in order to make things simple, many complex situations are ignored.

At first, the house price history data I got from internet contains many NaN. I dropped the rows that contains NaN cells. But if this is in a real scenario, need to find a proper method to fill those NaN cells, otherwise the growth data will not be calculated correctly.

Secondly, the price growth isn't the only factor for a investor. In real word you may need to create a more complex model and require more data to give a precise prediction.

The third thing is, better to find nearby venues through Baidu API in Nanjing. Not enough information got through Forsquare API.

5. Conclusion

Data science can help people understand the world transparently and help people to make correct decision. It can be widely used in different regions.

Sincerely,

Zhu Lan

Notebook Links:

https://github.com/markzhu1974/Coursera_Capstone/blob/master/DataScience%20Capstone%20Project-Final.ipynb