The Battle of Neighbourhoods

Applied Data Science Capstone

Contents

Introduction and Business Problem

The best neighbourhoods to settle down for a relocation

A common question from a relocation service client

- A client is relocating from Central, Hong Kong to Shanghai.
- The client is asking for a recommendation of the best neighborhood to settle down in Shanghai.
- The client prefers the similarity to his/her current home location in terms of the accesses to certain nearby venue types.
- There are other considerations like population density and housing cost.

Data

Base Data Sets

List of administrative divisions of Shanghai - Wikipedia

	Unnamed: 0_level_0								County Level
	Unnamed: 0_level_1	Name	Chinese	Hanyu Pinyin	Division code[2]	Division code[2].1	Area (km²)[3]	Population (2018 census)[3]	Density (/km²)
0	NaN	Huangpu District[4](City seat)	黄浦区	Huángpǔ Qū	310101	HGP	20.46	653800	31955
1	NaN	Xuhui District	徐汇区	Xúhuì Qū	310104	XHI	54.76	1084400	19803
2	NaN	Changning District	长宁区	Chángníng Qū	310105	CNQ	38.30	694000	18120
3	NaN	Jing'an District	静安区	Jìng'ān Qū	310106	JAQ	36.88	1062800	28818
4	NaN	Putuo District	普陀区	Pǔtuó Qū	310107	PTQ	54.83	1281900	23380
5	NaN	Hongkou District	虹口区	Hóngkǒu Qū	310109	HKQ	23.48	797000	33944
6	NaN	Yangpu District	杨浦区	Yángpǔ Qū	310110	YPU	60.73	1312700	21615
7	NaN	Pudong New Area	浦东新区	Pǔdōng Xīnqū	310115	PDX	1210.41	5550200	4585
8	NaN	Minhang District	闵行区	Mǐnháng Qū	310112	MHQ	370.75	2543500	6860
9	NaN	Baoshan District	宝山区	Bǎoshān Qū	310113	BAO	270.99	2042300	7536
10	NaN	Jiading District	嘉定区	Jiādìng Qū	310114	JDG	464.20	1588900	3423
11	NaN	Jinshan District	金山区	Jīnshān Qū	310116	JSH	586.05	805000	1374
12	NaN	Songjiang District	松江区	Sōngjiāng Qū	310117	SOJ	605.64	1762200	2910
13	NaN	Qingpu District	青浦区	Qīngpǔ Qū	310118	QPU	670.14	1219100	1819
14	NaN	Fengxian District	奉贤区	Fèngxián Qū	310120	FXI	687.39	1152000	1676
15	NaN	Chongming District	崇明区	Chóngmíng Qū	310151	CMG	1185.49	688100	580

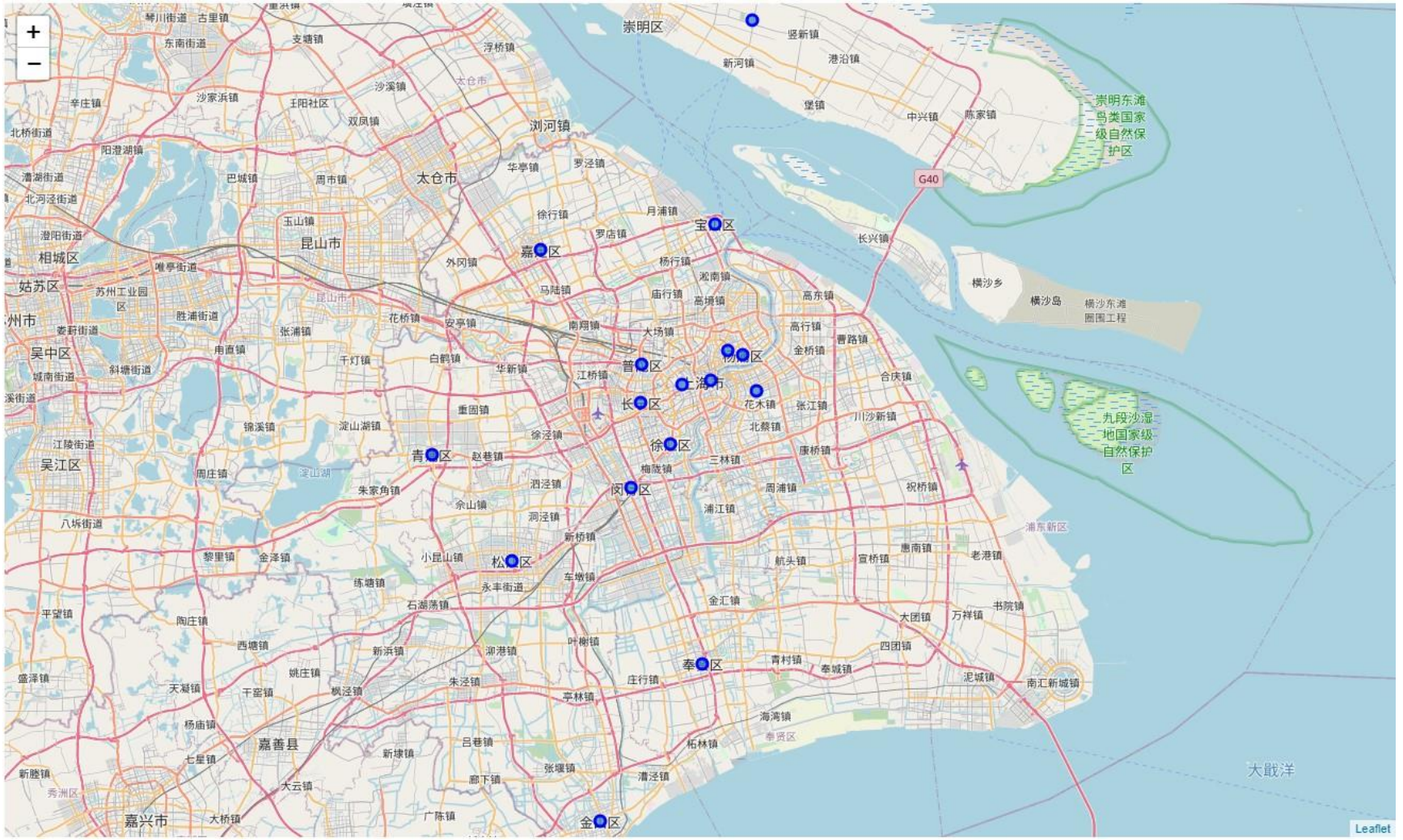
Shanghai Housing Costs (creprice.cn)

	排名	行政区	平均单价（元/㎡）	环比
0	1	宝山区	46155	+2.52%
1	2	崇明区	23517	+7.42%
2	3	奉贤区	25331	+0.62%
3	4	虹口区	72213	+7.74%
4	5	黄浦区	108298	+7.70%
5	6	嘉定区	38504	+2.70%
6	7	静安区	79793	+3.59%
7	8	金山区	18071	-4.33%
8	9	闵行区	55592	+4.24%
9	10	浦东新区	65675	+4.50%
10	11	普陀区	66007	+4.18%
11	12	青浦区	37930	+1.04%
12	13	松江区	38313	+3.60%
13	14	徐汇区	83942	+6.42%
14	15	杨浦区	68258	+4.70%
15	16	长宁区	77511	+3.36%

Combined and Transformed Data Set with Geospatial Coordinates

	District	Area	Total Population	Population Density	Housing Cost	Latitude	Longitude
0	Huangpu	20.46	653800	31955	108298	31.233593	121.479864
1	Xuhui	54.76	1084400	19803	83942	31.163698	121.427994
2	Changning	38.30	694000	18120	77511	31.209276	121.389986
3	Jing'an	36.88	1062800	28818	79793	31.229776	121.443060
4	Putuo	54.83	1281900	23380	66007	31.251326	121.391229
5	Hongkou	23.48	797000	33944	72213	31.266703	121.501751
6	Yangpu	60.73	1312700	21615	68258	31.262011	121.521430
7	Pudong	1210.41	5550200	4585	65675	31.221783	121.538740
8	Minhang	370.75	2543500	6860	55592	31.114767	121.376943
9	Baoshan	270.99	2042300	7536	46155	31.406634	121.485158
10	Jiading	464.20	1588900	3423	38504	31.377756	121.260612
11	Jinshan	586.05	805000	1374	18071	30.744817	121.337257
12	Songjiang	605.64	1762200	2910	38313	31.034405	121.223208
13	Qingpu	670.14	1219100	1819	37930	31.152164	121.119552
14	Fengxian	687.39	1152000	1676	25331	30.920449	121.469383
15	Chongming	1185.49	688100	580	23517	31.631339	121.533777

The 16 Districts in Shanghai on Map



The Nearby Venues Data Sets

- All the nearby venues around these locations within 5km radius using the Foursquare Places API. We also append the same info of the client's home location in the Central, Hong Kong to the data frame.

	District	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Huangpu	31.233593	121.479864	Campanile Hotel and Restaurant	31.232123	121.479144	Hotel
1	Huangpu	31.233593	121.479864	Waldorf Astoria Shanghai on the Bund (外滩华尔道夫酒店)	31.235479	121.485378	Hotel
2	Huangpu	31.233593	121.479864	Goodfellas	31.234878	121.486730	Italian Restaurant
3	Huangpu	31.233593	121.479864	The Bund (外滩)	31.239316	121.486065	Waterfront
4	Huangpu	31.233593	121.479864	Mercato	31.236220	121.486530	Italian Restaurant
...	...	...	...	...	...	...	...
1016	Central, Hong Kong	22.279328	114.162813	PiCi	22.283248	114.152088	Italian Restaurant
1017	Central, Hong Kong	22.279328	114.162813	Pololi	22.282837	114.153309	Hawaiian Restaurant
1018	Central, Hong Kong	22.279328	114.162813	The Diplomat	22.282525	114.155017	Cocktail Bar
1019	Central, Hong Kong	22.279328	114.162813	Chachawan	22.285581	114.148066	Thai Restaurant
1020	Central, Hong Kong	22.279328	114.162813	Sushi Gin (鮨吟)	22.277499	114.181388	Sushi Restaurant

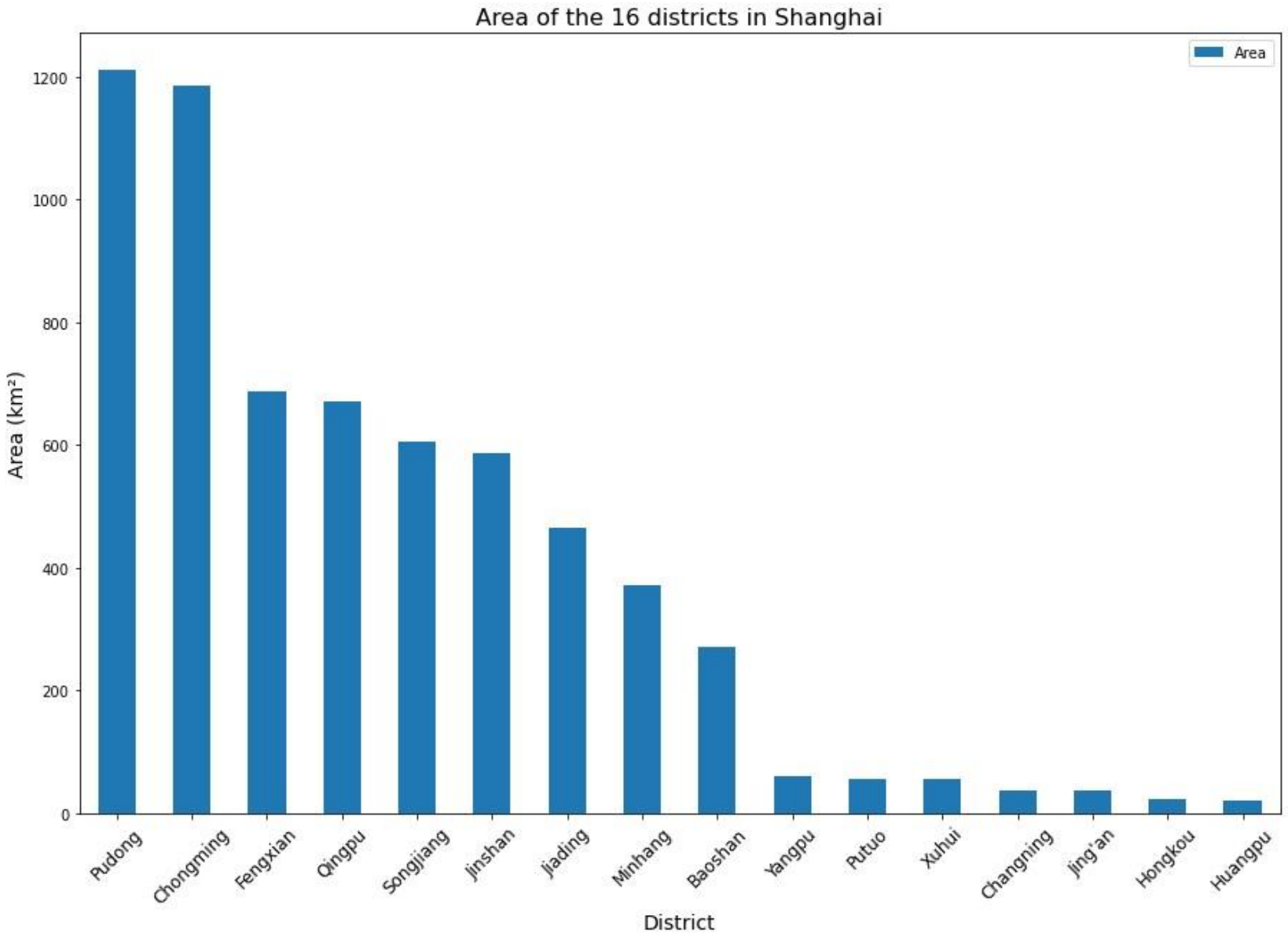
1021 rows × 7 columns

- Grouped venue types appearance frequency info by locations (first 5 rows shown)

	District	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	...	Water Park	Waterfront	Whisky Bar	Wine Bar	Wine Shop	Xinjiang Restaurant	Yoga Studio	Yunnan Restaurant	Zhejiang Restaurant	Zoo
0	Baoshan	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	...	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.0	0.00
1	Central, Hong Kong	0.0	0.0	0.01	0.01	0.01	0.01	0.01	0.01	0.0	...	0.0	0.0	0.0	0.01	0.01	0.00	0.02	0.00	0.0	0.01
2	Changning	0.0	0.0	0.01	0.00	0.00	0.01	0.00	0.00	0.0	...	0.0	0.0	0.0	0.03	0.00	0.01	0.00	0.01	0.0	0.00
3	Chongming	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	...	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.0	0.00
4	Fengxian	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.0	...	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.0	0.00

Exploratory Data Analysis

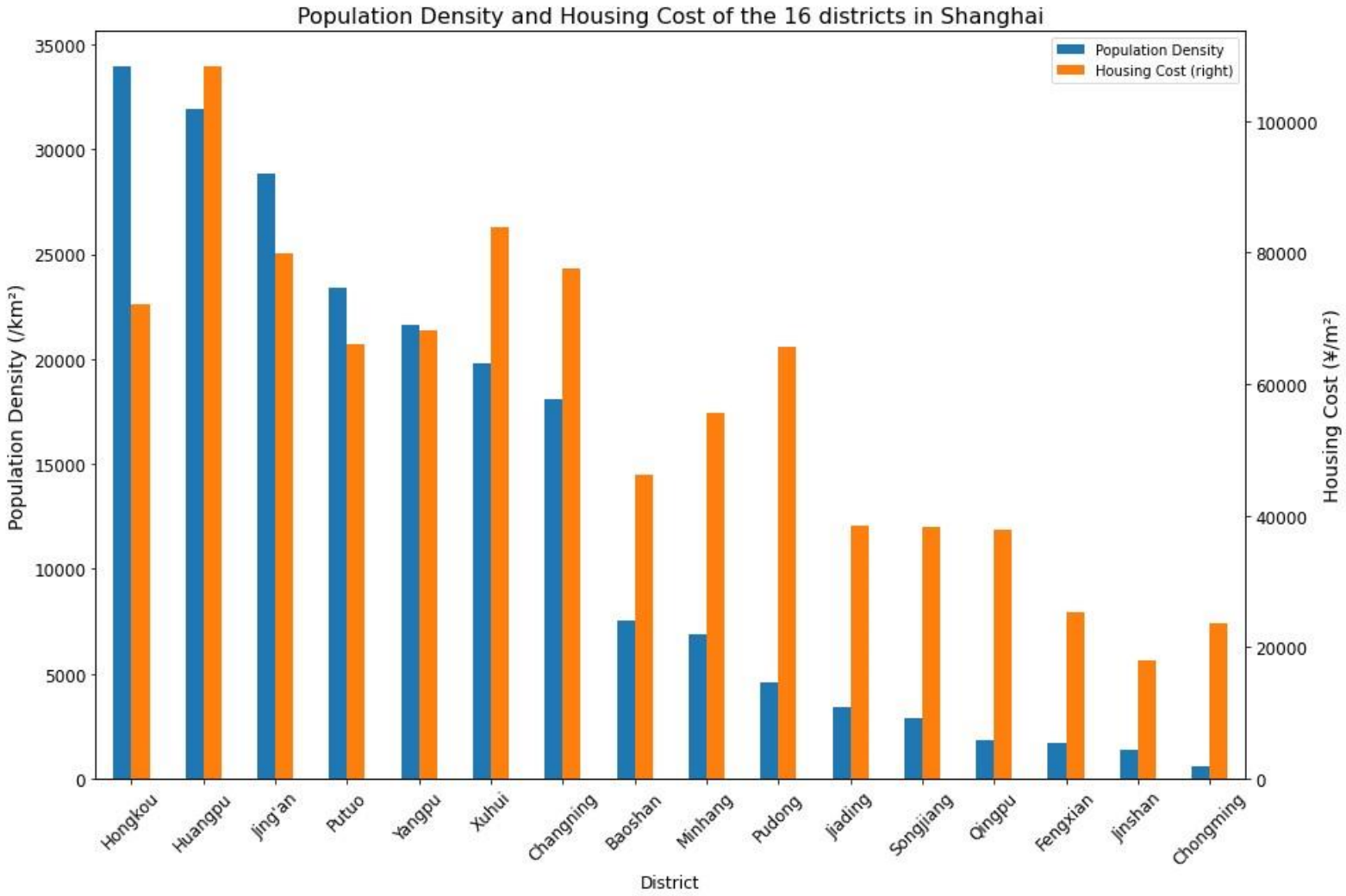
Areas of the 16 Districts in Shanghai



Observations:

Pudong and Chongming are the two largest districts in Shanghai while Hongkou and Huangpu are the two smallest districts.

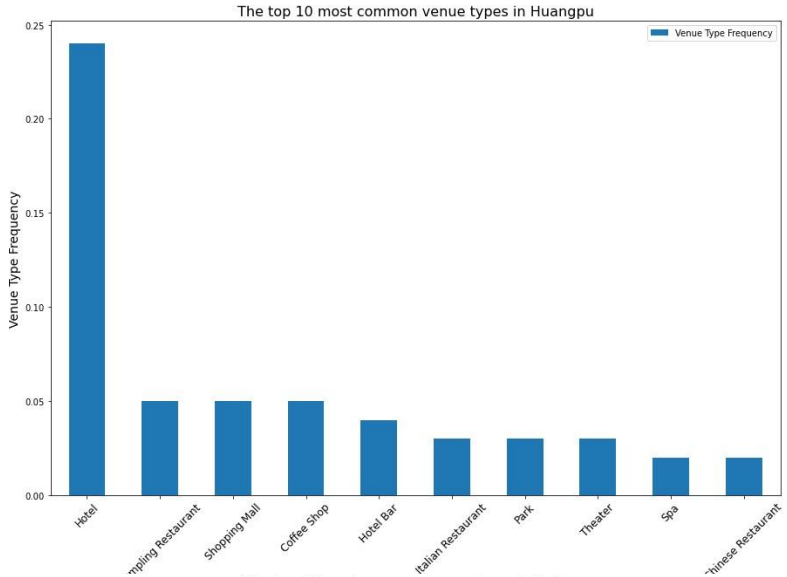
Population Density and Housing Costs

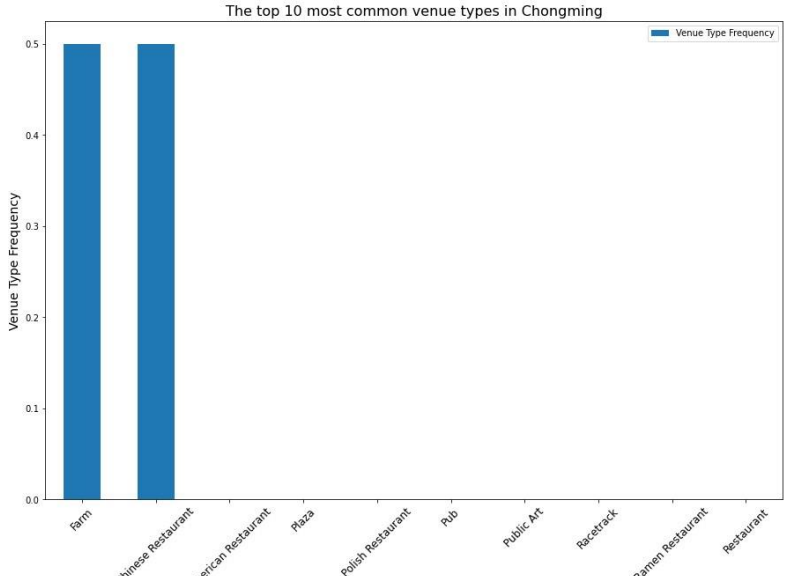


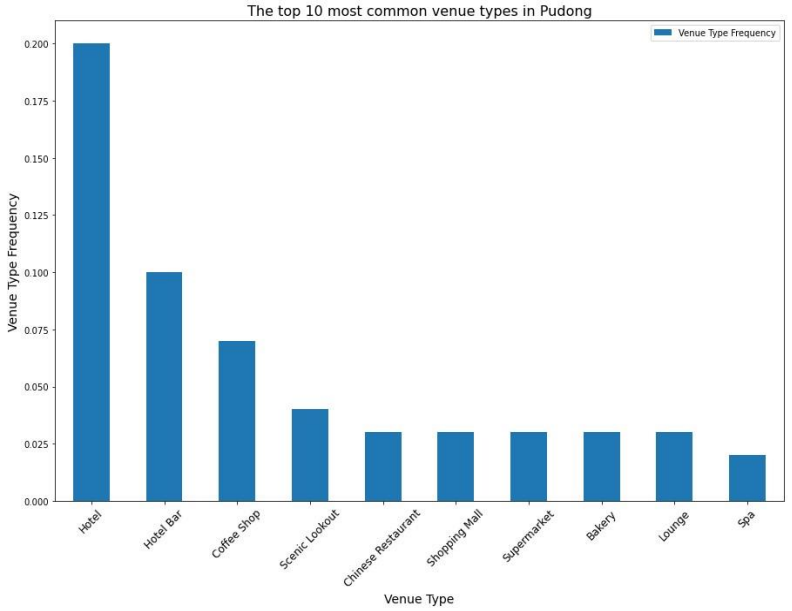
Observations:

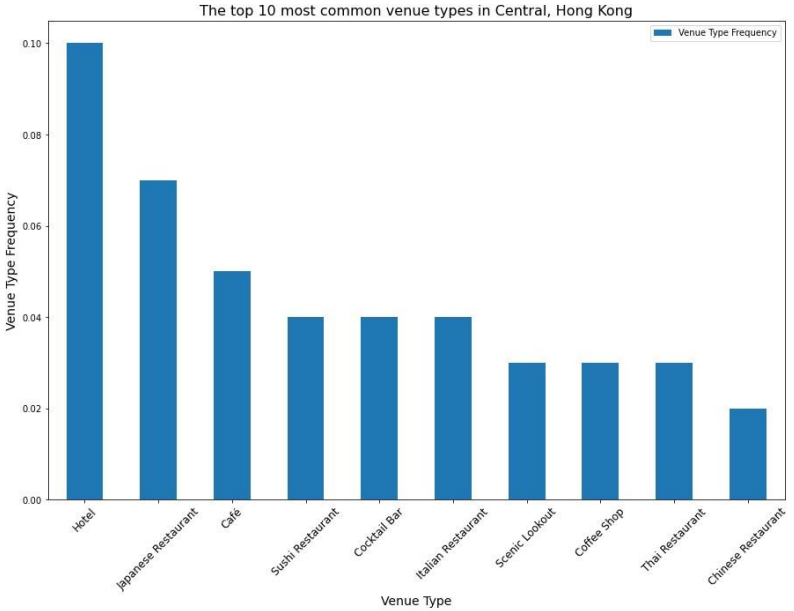
Hongkou and Huangpu are the two most crowded districts in Shanghai due to the small sizes, while Huangpu is also the most expensive neighbourhood for housing. Jinshan and Chongming are the two relative remote areas with lowest population densities and housing costs. One interesting thing to note that, while Pudong is the largest district in Shanghai which also has a very low population density, it is actually not very cheap for housing. The main reason is that Pudong is a newer district developed in Shanghai with some new industrial centers settled down in the region. Pudong is also Shanghai's international logistics hub where the main international airport is located.

Top Common Nearby Venue Types









Observations:

We selected some main locations to explore the most common nearby venue types

- Huangpu: As the most crowded and expensive and busiest district in Shanghai, the most common venue type is hotel, occupying over 25% of the venues nearby. There are also a lot of restaurants, shopping malls, parks and theatres.
- Chongming: The only remote island in Shanghai. We only got two types of venues in this area, farm and Chinese restaurant. It might not be a good choice for any new comers to this city.
- Pudong: Pudong is a newly developed district in Shanghai. There are also a lot of hotels and restaurants around. We can also notice that "scenic lookout" is among the top 5 most common venue types in Pudong.
- Central, Hong Kong: The Client's home location. The international financial hub, which has a lot of hotels, restaurants and bars. Notice that the "Scenic Lookout" is also among the top 10 common types in the Central, Hong Kong, which is similar to the Pudong district in Shanghai.

Methodology

Client Preference Scenarios

Now it's the time to analyze our venues data for our settlement location recommendation task. Different clients may have different preferences of the access to the nearby venues and other considerations when choosing a place to live. Next, let's assume two different client preference scenarios and we will then analyze our data according to these scenarios and build up a model suitable to the scenario.

- Scenario 1: The client doesn't have particular preferred venue type. The only considerations are similarity to his/her current location and if it's not very crowded.

- Scenario 2: The client only cares about the common daily life style venues and wants a place where they can afford buying an apartment

Clustering and Recommendation Methodology

Scenario 1: Overall similarity with low population density

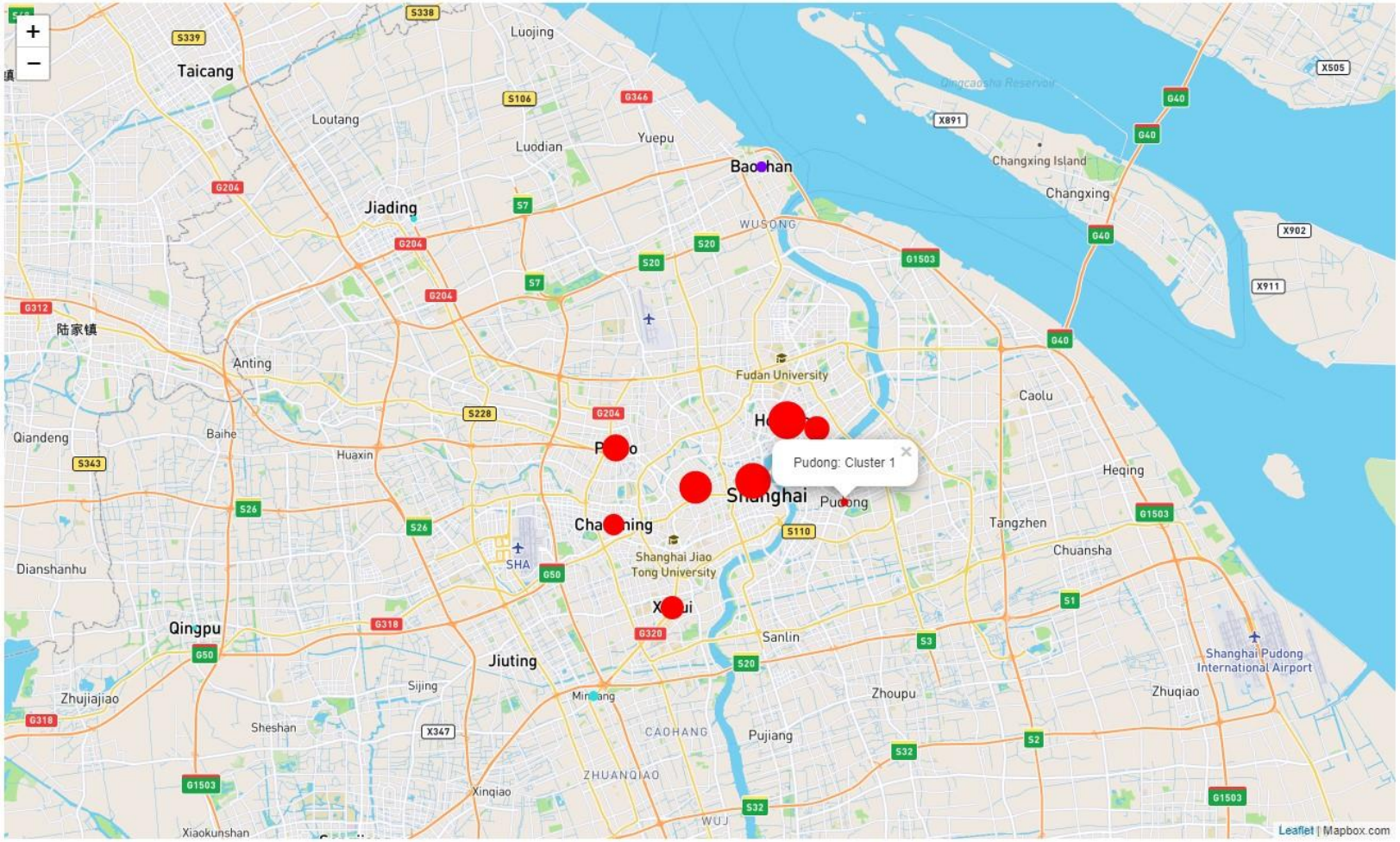
So, for this scenario, the client only wants to choose a neighborhood which will be similar to their current home, but they don't want to move to a very crowded area. So, what we can do is to cluster the venues data of Shanghai together with the venues info of the Central, Hong Kong (which we have already combined together with our venues dataframe). We can build a K-Means model to do so. And after that pick up a similar neighborhood with the lowest population density.

Scenario 2: Most comparable access to daily life venues with low housing cost

So, for this scenario, the client only cares about those most common daily life style venue types, including restaurants, bars, supermarkets, fitness facilities and theatres. On the other hand, the housing cost is another consideration. Thus, for this preference scenario, we can just include the relevant venue types when doing the clustering and then rank by the housing costs.

Results and Discussions

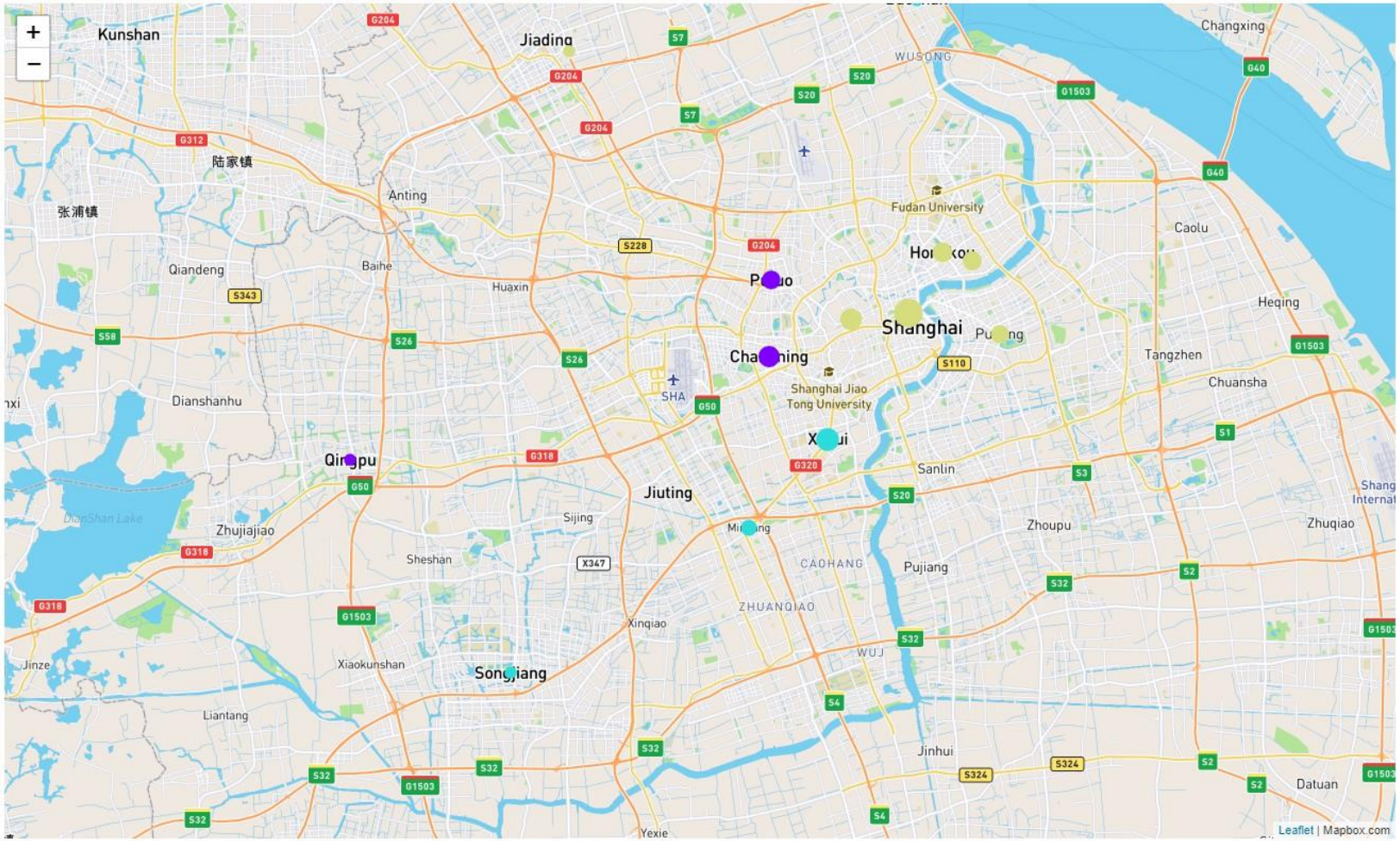
Scenario 1



Discussions:

As mentioned earlier, in this scenario, the client cares about the overall similarity to his/her current home and also prefer the lower crowdedmess. Thus we cluster the entire venues data with a K-Means model with 4 clusters. In this scenario, the client's home "Central, Hong Kong" got a cluster label 1. As we can see from the clustering map left, those districts with the red color are most similar to the client's home location in Hong Kong (cluster label 1) in terms of all nearby venue types, which are all closer to Shanghai's city center compared to other clusters. The radius of the markers represent the population density of that district. Larger marker means highly occupied region with high population density while smaller marker represents a region which is not crowded. Among the Cluster 1, Pudong stands out as an exceptional case which is similar enough to our client's home but with a very low population density. So it's a very good fit to our client's preference. Thus we will recommend **Pudong** to our client in this Scenario.

Scenario 2



Discussions:

For this scenario, given the client only cares about the certain daily life style venues. We will first filter our venues data with certain key words. We will then build a K-Means model to fit the filtered venue types only. The client also puts more weighting on the housing cost. Note in this scenario, the client's home "Central, Hong Kong" got a cluster label 0. As we can see from the clustering results left, those districts with the purple color are most similar to the client's home location in terms of all nearby restaurants, bars, supermarkets, fitness facilities and theatres. We got 3 districts in this cluster, Qingpu, Putuo and Changning. As we can see, among the 3 purple districts, Qingpu has the lowest housing cost. This could the best choice for the client. However, if the client cares about the distance to the city center, it's relatively a bit too far away. For the remaining two similar choices in this cluster, Putuo and Changning have comparable housing costs while Putuo has slightly lower margin. Thus we will recommend **Qingpu** if the client doesn't have any preference on the distance to the city center, otherwise we will recommend **Putuo**.

Conclusions

Conclusions

In this Capstone project, we performed a recommendation task for choosing the best settlement neighbourhood for our client who's relocating from Central, Hong Kong to Shanghai.

We first obtained some basic data from the web about the 16 districts in Shanghai, including the area, the population, population density and housing cost. We then retrieve the geospatial coordinates of the 16 districts and the Central, Hong Kong, the client's home location, which are then used with Foursquare Places API to explore the common popular nearby venues. And finally, we grouped the venue types based on the locations and calculated the corresponding appearance frequencies of each venue type for each location, which then becomes the base data set for building our machine learning models for clustering.

Next, we assumed two different client preferences of the access to the nearby venues and other considerations and performed our recommendation tasks accordingly:

- Scenario 1: The client doesn't have particular preferred venue type. The only considerations are similarity to his/her current location and if it's not very crowded.

So, for this scenario, the client only wants to choose a neighbourhood which will be similar to their current home in terms of all types of nearby venues, but they don't want to move to a very crowded area. So, we built up a K-Means model to cluster all the venues data of Shanghai together with the venues info of the Central, Hong Kong. From the clustering results, we observed that the districts sharing the same cluster label of the Central, Hong Kong are all very close to Shanghai city center. Among those, Pudong has the lowest population density. Thus we recommend **Pudong** to our client in this scenario

- Scenario 2: The client only cares about the common daily life style venues and wants a place where they can afford buying an apartment

So, for this scenario, given the client only cares about the certain daily life style venues. We first filtered our venues data for those daily life venues. We then built a K-Means model to fit the filtered venue types only. From the clustering results, we observed 3 districts which are similar to the client's home in terms of these daily life venues. Among the 3 districts, **Qingpu** has the lowest housing cost, which is then our top recommendation to the client in this scenario. However, it's a bit far away to the Shanghai city center. If our client has concerns about the distance to the city center, then we recommend **Putuo**, which is closer to the city center but is still a cheap place for housing if compared with all other districts around.