Prioritized List of Candidate Places for Starting up a Restaurant

Wang Xuan

October 1st 2021

1. Introduction

1.1 Background

Shanghai has a population of 30M, offering infinite things to tourists and its citizens, where restaurants play a very important role which gives Shanghai a unique taste. According to answers.com, there are more than 100,000 restaurants listed by the Shanghai Statistics Bureau.

1.2 Problem

Due to its nature of multiple ethnics, it’s a problem for those who wants to start up a restaurant: which borough, neighborhoods and even venues are the best choice? What should be considered if we want to make a prioritized list of available places by doing some data analysis?

When exploring this problem, the first idea coming up is, we want to find a place with most population but least restaurant density. The other choice is the ratio between the GDP or average income per person versus the number of restaurants.

And we want to find out such kind of lists for each restaurant types: Chinese, Jewish, Italian, etc.

1.3 Interest

Target Audience of this research will be those who want to start up a business (especially a restaurant) in a big city like New York City.

2. Data acquisition and cleaning

2.1 Data sources

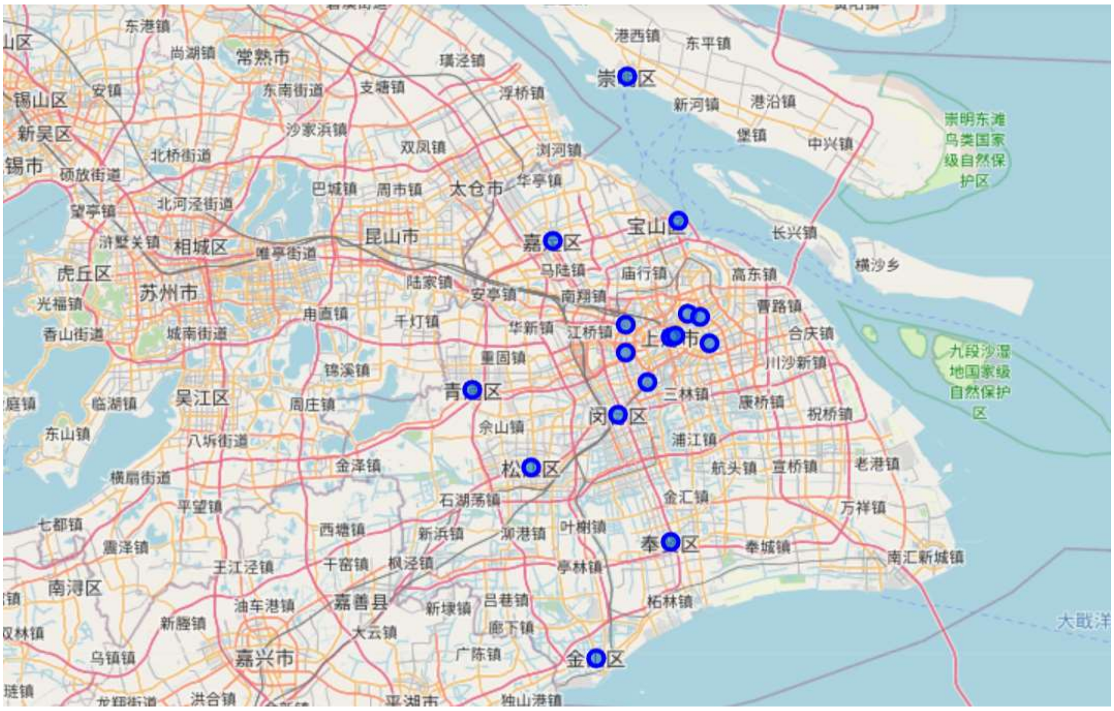
- Shanghai neighborhood data, which include Neighborhood Area (km2), Total population 2017, Seat, Postal code. https://en.wikipedia.org/wiki/Shanghai. I used BeautifuSoup library to get the data and parse the table of neighborhoods and other data inside it.

	Division code[110]	Neighborhood	Area (km2)[111]	Total population 2017[111]	Seat	Postal code
0	310000	Shanghai	6340.50	24,183,300	Huangpu	200000
1	310101	Huangpu	20.46	654,800	Waitan Subdistrict	200001
2	310104	Xuhui	54.76	1,088,300	Xujiahui Subdistrict	200030
3	310105	Changning	38.30	693,700	Jiangsu Road Subdistrict	200050
4	310106	Jing'an	36.88	1,066,200	Jiangning Road Subdistrict	200040
5	310107	Putuo	54.83	1,284,700	Zhenru Town Subdistrict	200333
6	310109	Hongkou	23.46	799,000	Jiaxing Road Subdistrict	200080
7	310110	Yangpu	60.73	1,313,400	Pingliang Road Subdistrict	200082
8	310112	Minhang	370.75	2,534,300	Xinzhuang town	201100
9	310113	Baoshan	270.99	2,030,800	Youyi Road Subdistrict	201900
10	310114	Jiading	464.20	1,581,800	Xincheng Road Subdistrict	201800
11	310115	Pudong	1210.41	5,528,400	Huamu Subdistrict	200135
12	310116	Jinshan	586.05	801,400	Shanyang town	201500
13	310117	Songjiang	605.64	1,751,300	Fangsong Subdistrict	201600
14	310118	Qingpu	670.14	1,205,300	Xiayang Subdistrict	201700

- Geographic data of Shanghai neighborhoods: I used Nominatim to get the locations and join them with the neighborhood table.

	Division code[110]	Neighborhood	Area (km2)[111]	Total population 2017[111]	Seat	Postal code	Latitude	Longitude
0	310000	Shanghai	6340.50	24,183,300	Huangpu	200000	31.232276	121.469207
1	310101	Huangpu	20.46	654,800	Waitan Subdistrict	200001	31.233593	121.479864
2	310104	Xuhui	54.76	1,088,300	Xujiahui Subdistrict	200030	31.163698	121.427994
3	310105	Changning	38.30	693,700	Jiangsu Road Subdistrict	200050	31.209276	121.389986
4	310106	Jing'an	36.88	1,066,200	Jiangning Road Subdistrict	200040	45.408600	123.132716
5	310107	Putuo	54.83	1,284,700	Zhenru Town Subdistrict	200333	31.251326	121.391229
6	310109	Hongkou	23.46	799,000	Jiaxing Road Subdistrict	200080	31.266703	121.501751
7	310110	Yangpu	60.73	1,313,400	Pingliang Road Subdistrict	200082	31.262011	121.521430
8	310112	Minhang	370.75	2,534,300	Xinzhuang town	201100	31.114767	121.376943
9	310113	Baoshan	270.99	2,030,800	Youyi Road Subdistrict	201900	31.406634	121.485158
10	310114	Jiading	464.20	1,581,800	Xincheng Road Subdistrict	201800	31.377756	121.260612
11	310115	Pudong	1210.41	5,528,400	Huamu Subdistrict	200135	31.221783	121.538740
12	310116	Jinshan	586.05	801,400	Shanyang town	201500	30.744817	121.337257
13	310117	Songjiang	605.64	1,751,300	Fangsong Subdistrict	201600	31.034405	121.223208
14	310118	Qingpu	670.14	1,205,300	Xiayang Subdistrict	201700	31.152164	121.119552

- Geographic map: I used folium to display the neighborhoods of Shanghai:



- Venue data of Shanghai neighborhood: I used Foursquare to get all the venues of the neighborhoods and join them with the neighborhood table.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Shanghai	31.232276	121.469207	JW Marriott Hotel Shanghai at Tomorrow Square ...	31.232216	121.465260	Hotel
1	Shanghai	31.232276	121.469207	Shanghai Grand Theater (上海大剧院)	31.231030	121.467263	Theater
2	Shanghai	31.232276	121.469207	Homeslice	31.225189	121.461093	Pizza Place
3	Shanghai	31.232276	121.469207	The Peninsula Shanghai	31.243049	121.484564	Hotel
4	Shanghai	31.232276	121.469207	The Bund (外滩)	31.239316	121.486065	Waterfront
5	Shanghai	31.232276	121.469207	Starbucks Reserve Roastery (星巴克臻选)	31.232402	121.457684	Coffee Shop
6	Shanghai	31.232276	121.469207	Fraser Residence Shanghai	31.225639	121.476177	Hotel
7	Shanghai	31.232276	121.469207	Goodfellas	31.234878	121.486730	Italian Restaurant
8	Shanghai	31.232276	121.469207	Din Tai Fung (鼎泰丰)	31.230923	121.458491	Dumpling Restaurant
9	Shanghai	31.232276	121.469207	Le Royal Club Lounge	31.236404	121.471364	Lounge

2.2 Feature selection

The target problem requires these features,

- the neighborhood granularity
- the population
- Venue categories.

3. Exploratory Data Analysis

To get all the venues for each neighborhood, I used the latitude/longitude data as the input parameter to the Foursqaure API and get the venues. I chose the radius of 10km, which can cover all the areas of neighborhoods. The number of venues is 181, which is far less than the official data. This means the Foursquare data might be insufficient. However, my problem is to research the density of restaurants to population, so the data set is still valid.

3.1 Calculation of target variable

I use *contains* function to filter out the restaurant data. By exploring the data, I listed all the venue categories which are related to restaurant:

Restaurant|Coffee|Noodle|Bar|Breakfast|Bakery|Café|Steak|Pizza|Sandwich|Tea|Food|Seafood|Buffet|Dumpling.

As the result, I created a data-frame of restaurants venues only.

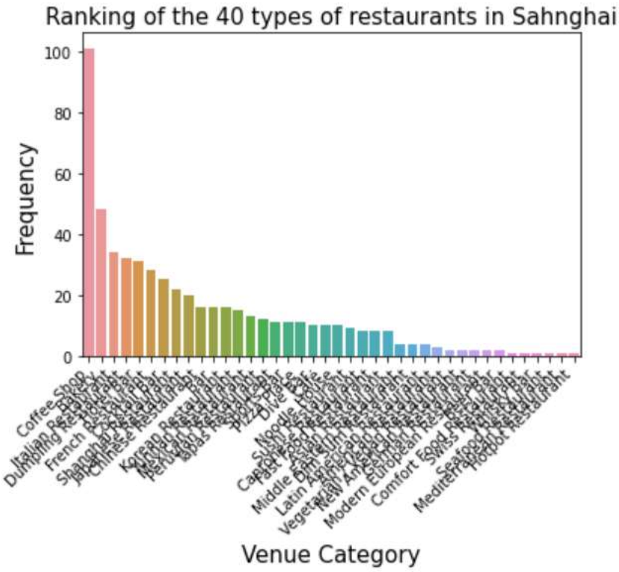
3.2 By venue category at city level

I then immediately calculated the ranking of venue categories at city level by *value_counts* function:

	Frequency
Coffee Shop	101
Bakery	48
Italian Restaurant	34
Dumpling Restaurant	32
Hotel Bar	31
French Restaurant	28
Cocktail Bar	25
Shanghai Restaurant	22
Japanese Restaurant	20
Chinese Restaurant	16
Bar	16
Korean Restaurant	16
Yunnan Restaurant	15
Mexican Restaurant	13
Peruvian Restaurant	12
Tapas Restaurant	11
Sports Bar	11

3.3 Visualization

I used seaborn library to display the relationship between venue number and venue categories:



3.4 Analysis of each Neighborhood

Now I have by neighborhood and by venue category table, so I can make an on-hot table out of these two dimensions:

	Neighborhood	American Restaurant	Asian Restaurant	Bakery	Bar	Beer Bar	Café	Cantonese Restaurant	Chinese Restaurant	Cocktail Bar	Coffee Shop	Comfort Food Restaurant	Dim Sum Restaurant	Dive Bar	Dumpling Restaurant
1	Shanghai	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Shanghai	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3	Shanghai	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Shanghai	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5	Shanghai	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6	Shanghai	0	0	1	0	0	0	0	0	0	0	0	0	0	0
7	Shanghai	0	0	0	0	0	0	1	0	0	0	0	0	0	0
8	Shanghai	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Shanghai	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Next, I group rows by neighborhood and by taking the ratio of the frequency to population of occurrence of each category:

```
shanghai_grouped_density= shanghai_grouped.loc[:,"American Restaurant":"Yunnan Restaurant"]
    .div(shanghai_grouped['Population']
    .str.replace(',', '')
    .astype(float), axis=0)*1000000
```

Because the populations are at million level, the ratio between venue number and population would be too small so I used a factor 100000. Then we have the result to reflect the number-to-population ratio between restaurant category and neighborhoods:

	Neighborhood	American Restaurant	Asian Restaurant	Bakery	Bar	Beer Bar	Café	Cantonese Restaurant	Chinese Restaurant	Cocktail Bar	Coffee Shop	Comfort Food Restaurant	Dim Sum Restaurant
0	Baoshan	0.000000	0.000000	0.082702	0.041351	0.000000	0.082702	0.041351	0.000000	0.082702	0.206754	0.000000	0.041351
1	Changning	0.000000	1.527184	7.635919	1.527184	0.000000	0.000000	0.000000	1.527184	3.054368	7.635919	0.000000	1.527184
2	Chongming	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2.756593	0.000000	0.000000
3	Fengxian	0.000000	0.000000	1.441545	0.000000	0.000000	1.441545	0.000000	2.883091	0.000000	18.740089	0.000000	0.000000
4	Hongkou	0.000000	0.000000	3.751641	0.937910	0.000000	0.937910	0.937910	0.000000	1.875821	5.627462	0.000000	0.000000
5	Huangpu	0.778392	0.000000	2.335176	0.778392	0.000000	0.778392	0.778392	0.000000	2.335176	4.670351	0.000000	0.000000
6	Jiading	0.000000	1.251564	3.754693	2.503129	0.000000	0.000000	1.251564	1.251564	1.251564	6.257822	0.000000	1.251564
7	Jing'an	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.761383	0.000000	0.000000	0.000000	0.000000
8	Jinshan	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.578345	0.000000	0.000000
9	Minhang	0.000000	0.492417	1.969667	0.492417	0.000000	0.000000	0.000000	0.492417	0.984834	2.462084	0.000000	0.000000
10	Pudong	0.632191	0.000000	1.896574	0.632191	0.000000	0.632191	0.632191	0.000000	1.896574	3.793147	0.000000	0.000000
11	Putuo	0.000000	0.180884	0.904421	0.180884	0.000000	0.000000	0.180884	0.180884	0.180884	0.904421	0.000000	0.180884
12	Qingpu	0.000000	1.247816	3.743449	2.495633	2.495633	1.247816	0.000000	4.991265	0.000000	18.717245	1.247816	0.000000
13	Shanghai	0.571004	0.000000	2.284018	0.571004	0.000000	0.571004	0.571004	0.000000	1.142009	2.855022	0.000000	0.000000
14	Songjiang	0.000000	1.659338	3.318676	1.659338	0.000000	0.829669	0.000000	3.318676	2.489007	4.978014	0.000000	0.000000
15	Xuhui	0.000000	0.865576	3.462304	0.865576	0.000000	0.000000	0.000000	0.865576	1.731152	4.327880	0.000000	0.000000

These table provides good result but it is not very user friendly if someone want to check the data of individual neighborhoods, so I give out a by-neighborhood ranking algorithm and generated the results (take Hongkou as the example),

```
26          Dim Sum Restaurant    0.00
27     Comfort Food Restaurant    0.00
28          Hotpot Restaurant     0.00
29               Cocktail Bar     0.00
30                   Beer Bar     0.00
31       Cantonese Restaurant     0.00
32                     Bakery     1.44
33         Italian Restaurant     1.44
34         Mexican Restaurant     1.44
35                       Café     1.44
36         Chinese Restaurant     2.88
37          Korean Restaurant     4.32
38       Fast Food Restaurant     7.21
39                Coffee Shop    18.74


----Hongkou----
```

4. Conclusions

For huge cities like Shanghai, the opportunity of starting up a business is numerous, and to startup a restaurant is a good choice for many people. To ensure good cash-flow, enough number of customers is essential. However, when the population is big, the number of restaurants might be even more. So it make sense to research the restaurant number to population ratio.

The online data sources are: the information of neighborhoods, population, location. The latitudes and longitudes of venues, the information list of each venue.

Using pandas dataframe and the related algorithms I can join all the necessary information in a table, which reflects the number of restaurants with two dimensions: neighborhoods and restaurant category. The most meaningful table then can be derived by using one-hot algorithm, and calculated the ratio between number of restaurants and population.

To check the ratios of each neighborhoods, simple make a ordered list by neighborhoods.