Where is a restaurant located in Hong Kong?

1 Introduction

Background

People do not only go to a restaurant for the good food and service, but also for convenience. This leads to a combination of the restaurants and other service facilities and sites for a certain activity, such as shopping, sightseeing, family activity etc. Figuring out the combination of restaurants and other service facilities and sites can help restaurant owners decide where to set up a restaurant and what kind of service they may provide to make the restaurant more popular.

Hypothesis

The hypothesis is that certain combinations of nearby facilities exist for a restaurant. The study aimed at proving if the hypothesis is true, figuring out the common combinations of restaurants and other facilities in Hong Kong and exploring the combinations for different types of restaurants.

2 Data preparation

Data structure

The analysis was based on the restaurant entities. The basic attributes include the ID, name, longitude and latitude of a restaurant, and the relationship of the restaurant to other facilities and sites was summarized as the number of venues of different categories within a distance of 200 m and 2000 m from the restaurant. The 5 most popular categories were selected representing the combination of the restaurant and others facilities and sites.

Data collection and cleaning

The data was collected from website Foursquare (https://foursquare.com/). Because of a maximum limit for a query, the map of Hong Kong city was divided progressively into different sizes to build a series of search queries. All the available venues of varies categories of the city was collected. The collected venue information includes ID, name, category, latitudes and longitudes. The higher hierarchy of the categories was added according to the hierarchical category trees of the foursquare service. The restaurants were subsetted, and the number of likes was scraped to analyze the relationship of likes and nearby venues. The distances between the restaurants and the venues were calculated using geodesic function of geopy in Python.

Clustering

Before clustering, principle components analysis (PCA) was applied to visually check the possibility of clustering and the approximate number of clusters. However, not every dataset has enough variance explained by the first 3 principle components (PC), so the approach is only applied on the datasets which have over 80% of the total variance explained by the first 3 PCs. Then, two algorithms were tested in the clustering, density based spatial clustering of application with noise (DBSCAN) and k-means clustering. The datasets with clear clusters in PCA was labeled using DBSCAN clusters and the datasets with a relatively continuous distribution of the venues was labeled with K-means clusters.

For analysis of nearby venues within 2000 m, the radius of DBSCAN was tuned according to the distribution plots of the distances and by trial and error. For analysis of nearby venues within 200 m, the structure of the nearby venues has more variation, and I aimed at finding more possibilities of clustering, the plot of radius and cluster number was applied to find a radius which gives more clusters. The number of core element needed for a cluster was set as a fixed value 60, because the restaurants are densely distributed in Hong Kong, too small clusters do not make sense as it will fall into a local area of some restaurants.

The cluster numbers of K-means were set according to 3D PCA plot of the first 3 PCs or by trial and error according to the mean attribute values of each cluster, e.g. emerging of two similar cluster will lead to reduction of the cluster number.

Comparison of the likes number

As the distribution is unknown with different clusters of unequal sizes, a Mann-Whitney U test was applied to compare the number of likes. The mean, standard deviation, minimum and maximum of likes were also summarized.

Correlation of the likes number with the nearby venues

Spearman rank correlations were applied to the number of likes of each restaurant and their basic categories. For a more detailed category, the categories with a frequency higher than 50% in more than half clusters were selected first, then the rank correlation was applied to the selected categories.

3 Data Analysis

Hierarchical categories of the venues

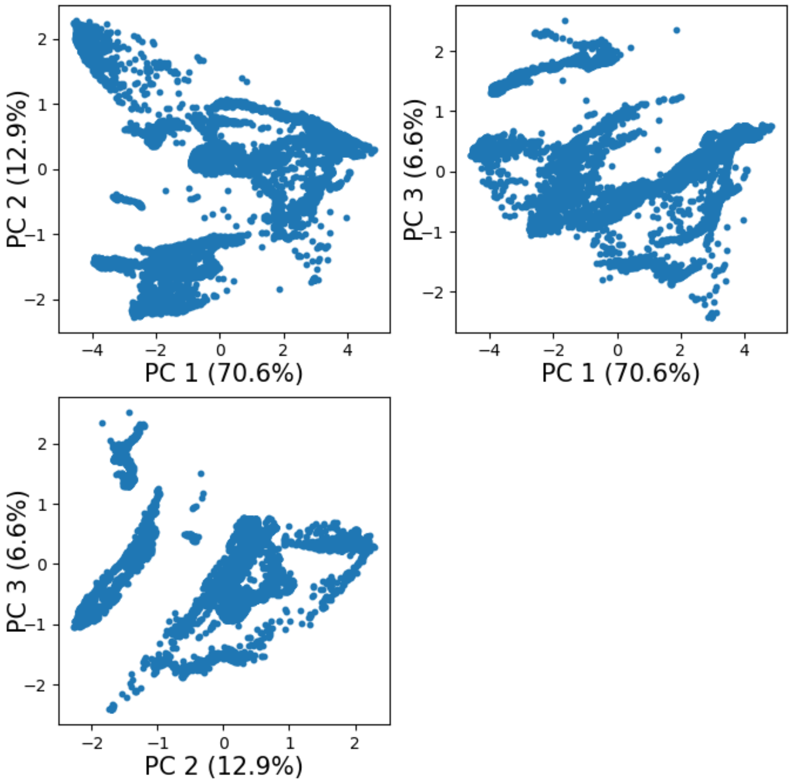
The venues are divided into 10 basic categories by foursquare, and then divided into more detailed subcategories. The basic categories and some popular subcategories in Hong Kong are listed below:

Basic Categories	Popular Subcategories
Arts & Entertainment	Art Gallery
Food	Asian Restaurant Coffee Shop Café Bakery
Nightlife Spot	Bar
Outdoors & Recreation	Athletics & Sports
Professional & Other Places	Office Medical Center
Residence	Residential Building (Apartment / Condo)
Shop & Service	Spa Food & Drink Shop Clothing Store Jewelry Store Bank Cosmetics Shop Salon / Barbershop
Travel & Transport	Hotel Bus Stop Bus Station
Event	
College & University	

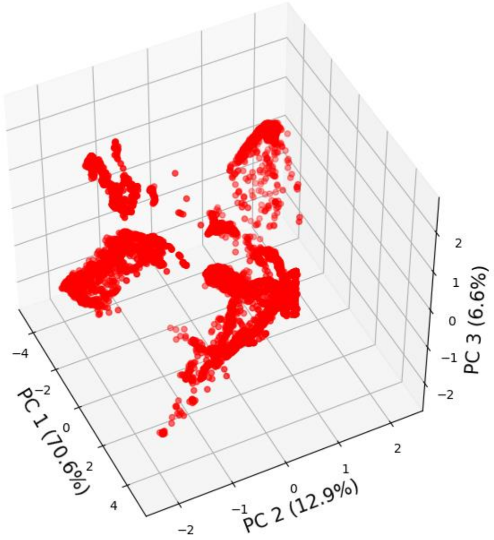
Features of venues within 2000 m of a restaurant

The counts of the venues in each basic category within 2000 m of a restaurant are good structured and clustered. The first 3 PCs of the basic categories explained 70.6%, 12.9% and 6.6% of the total variance. More than 5 clusters can be visually found easily in the PCA plot:

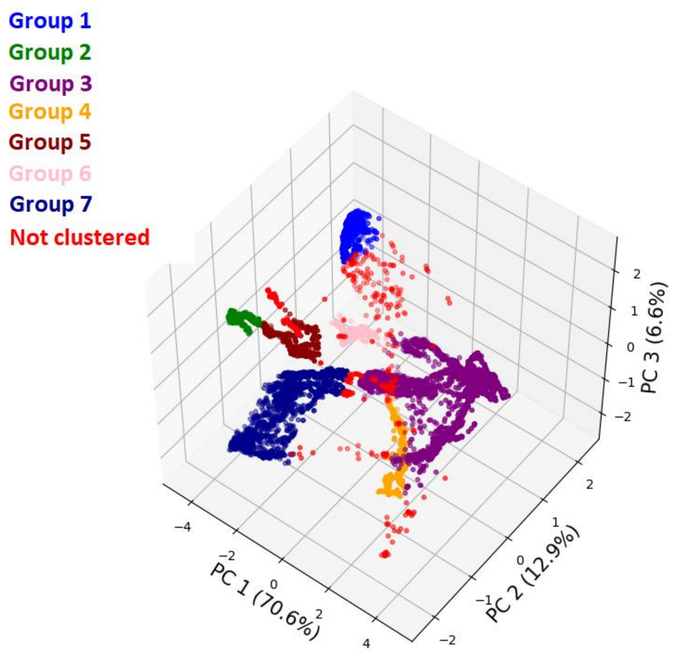
2D PCA biplot of the first 3 PCs



3D plot of the first 3 PCs



Using DBSCAN clustering, 7 clusters were found:



The statistics of the clusters are summarized below:

Cluster	Not clustered	1	2	3	4	5	6	7
Arts & Entertainment	164	316	216	72	94	166	175	139
Professional & Other Places	902	1865	1214	340	341	904	1004	787
Nightlife Spot	283	561	445	62	63	267	358	340
Food	1627	2312	2780	1163	912	1682	2444	2519
Shop & Service	1650	2892	2841	828	546	1257	2729	1839
Outdoors & Recreation	145	234	223	74	85	171	197	161
Travel & Transport	161	237	216	113	94	165	203	295
College & University	78	56	77	30	116	58	68	91
Event	0	0	2	0	0	2	0	1
Residence	184	296	211	109	195	159	192	118

The cluster of venue numbers are dominated by the density of the venues. The cluster 1 has the most venues in most categories except Food, Travel & Transport and College & Universities. The not clustered restaurants have a moderate average counts of venues in all categories, and the largest cluster, cluster 3, have the least venues in most categories.

However, difference can be found between the categories. Cluster 2 and 5 have the most venues in food, shop & service, Outdoors & recreation. The cluster 4 have the most venues in College & University but the second least venues in all other categories except residence. The cluster 7 have most venues in category Travel & Transport and many venues in category Food.

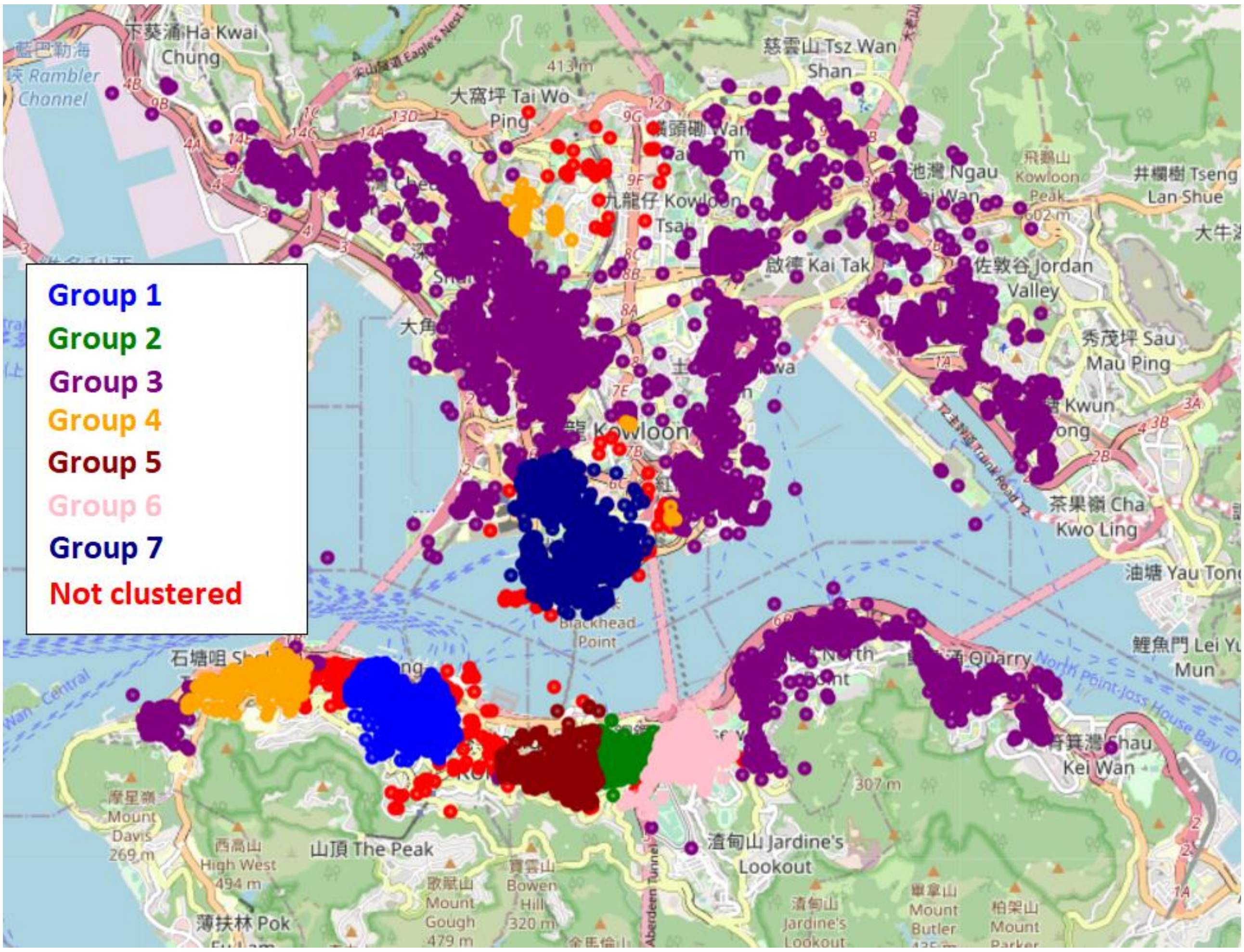
A detailed distribution of popular subcategories:

Basic Category	Subcategory	Outliers	1	2	3	4	5	6	7
Arts & Entertainment	Art Gallery	72	174	54	15	43	53	38	22
Food	Asian Restaurant	742	901	1463	648	441	817	1300	1307
	Coffee Shop	73	132	95	24	39	77	80	69
	Café	119	176	194	77	69	97	187	141
	Bakery	63	101	101	50	36	68	77	89
Nightlife Spot	Bar	215	412	343	51	49	199	277	286
Outdoors & Recreation	Athletics & Sports	90	152	153	36	37	104	136	93
Professional & Other Places	Office	314	712	425	80	92	347	321	190
	Medical Center	215	500	247	80	62	121	244	227
Residence	Residential Building (Apartment / Condo)	170	279	201	90	176	152	179	109
Shop & Service	Spa	95	182	171	22	11	56	170	102
	Food & Drink Shop	106	184	137	62	81	79	120	96
	Clothing Store	309	552	551	102	30	217	584	401
	Jewelry Store	61	120	109	15	3	32	114	81
	Bank	52	104	68	27	24	58	60	44
	Cosmetics Shop	82	115	199	32	11	62	208	117
	Salon / Barbershop	107	177	220	36	26	67	212	137
Travel & Transport	Hotel	54	52	87	40	29	53	86	158
	Bus Stop	39	56	61	33	37	51	53	36
	Bus Station	16	25	25	19	6	18	26	22

Most subcategories show the consistent results with the basic categories in venue counts. The cluster 1 has most bars, coffee shops and offices. The cluster 2 and 5 has most Asian restaurants and Cafes. The cluster 6 has most clothing stores and cosmetics shops, it also has a large number in all other shops.

However, for the basic category travel & transport, the detailed subcategories are not consistent with basic categories. The cluster 1,2,5,6 all have large number of bus stops and bus stations, which accounts for most transportation service. The cluster 7 have the largest number of hotels, giving it the highest number of venues counts in Travel & Transport category, but it only has a moderate number of bus stops and bus stations.

The location of restaurants of the clusters are visualized in maps as below:



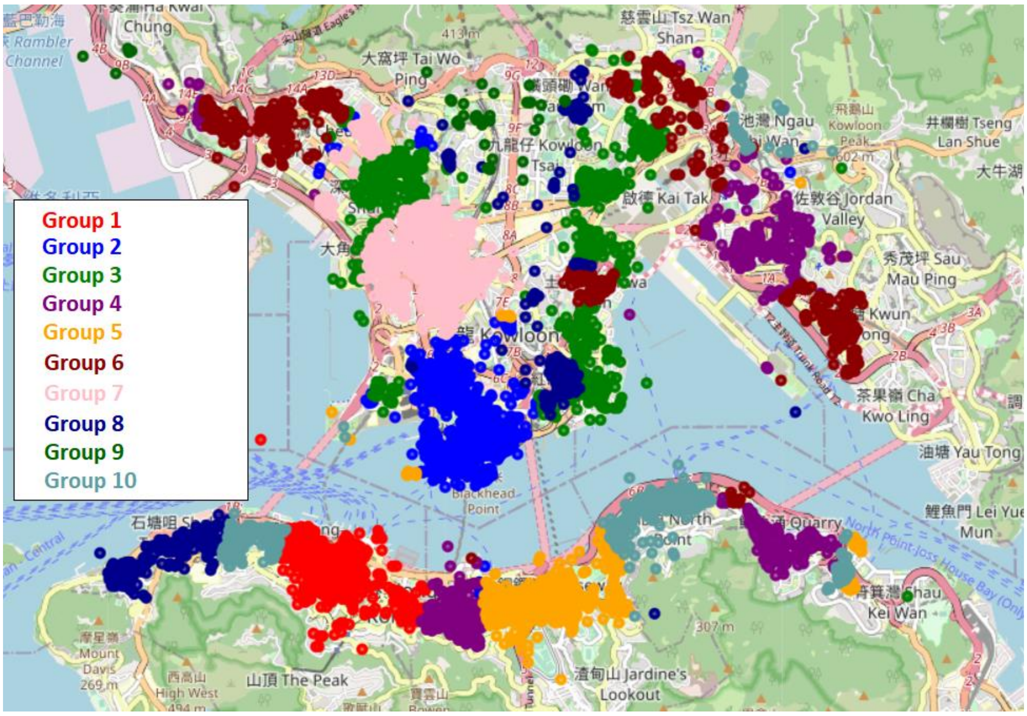
According to the map, the clusters of venue counts of basic categories within 2000 meters majorly gives the information of the district where the restaurants are located. The cluster 1 is located in the Central District, which is the political and commercial center of Hong Kong. So, it has the highest density of venues and highest number of offices, coffee shops and bars. And the cluster 4 is mostly located in Sai Wan District, which is also the location of the Hong Kong University. The cluster 6 is located at the center of Wan Chai District, which is also a commercial center of Hong Kong, it could be the reason why there is the most shops. Cluster 7 is located at Tsim Sha Tsui District, which is surrounded by the strait and at the center of Hong Kong City. This strict is by geographic and cultural reason the center of the tourism of Hong Kong, which gives it the largest number of hotels.

As the clustering of counts is more or less dominated by the density of the venues, the weight ratios of the venues of every restaurant in Hong Kong are analyzed. As the density is more even in the attribute distribution, the k-means clustering is applied. The results are showed below:

Unit: %

Cluster	1	2	3	4	5	6	7	8	9	10
Arts & Entertainment	4	2	2	3	2	4	3	3	2	3
Professional & Other Places	21	13	13	21	14	21	9	13	19	17
Nightlife Spot	6	5	2	3	5	1	2	3	1	2
Food	26	40	47	36	33	40	44	40	24	33
Shop & Service	33	29	22	26	36	22	32	21	20	28
Outdoors & Recreation	3	3	3	4	3	4	2	4	5	4
Travel & Transport	3	5	4	3	3	3	4	5	4	4
College & University	1	1	1	1	1	1	1	5	17	2
Event	0	0	0	0	0	0	0	0	0	0
Residence	3	2	5	3	3	4	3	7	7	8

Food and shops take the largest parts of the venues. The cluster 1 is featured with Professional places and nightlife spots, the cluster 3 is featured with other food venues, the cluster 5 is featured with shops and the cluster 9 is featured with college and universities. The featured clusters correspond to the same district with the counts, as the maps below:



However, different from the count clusters. The weights reveal 3 clusters featured with residential venues, cluster 8, 9, 10. They forms the areas between the clusters with more food, professional places and shops respectively.

Correlation of the nearby venues and popularity

The popularity of a restaurant is influenced by the location a lot. The influence is achieved in many ways, such as flow of population, the convenience and the combination of activities which include a dinner, etc. It is also observed that the number of likes varies a lot in the clusters found above, the statistics of the number of likes of the restaurants of the DBSCAN clusters of venue counts within 2000 meters are showed below:

Cluster	Mean	StDev	Min	Max
Not clustered	10	33	0	328
1	13	41	0	479
2	7	24	0	229
3	2	11	0	393
4	3	10	0	96
5	8	16	0	195
6	7	22	0	577
7	6	25	0	680

The comparison result using Mann Whitney U Test is summarized below:

	Not clustered	1	2	3	4	5	6	7
Not clustered								
1	-							
2	-	-						
3	<	<	<					
4	<	<	<	>				
5	-	-	>	>	>			
6	-	-	-	>	>	-		
7	<	<	-	>	>	<	<	

* The comparison is in the form the cluster of the row compared to the cluster of the column. The sign "-" indicates no significant difference between the cluster of the row and the cluster of the column, ">" indicates that the cluster of the row has more likes than the cluster of the column and "<" indicates that the cluster of the row has less likes than the cluster of the column.

It can be found that cluster 1, 5 and 6 have most likes in general, followed by cluster 2, while the cluster 3 have the least likes in general followed by cluster 7. However, great variation exists within the clusters, as the popularity of a restaurant is also and more likely decided by the quality and service of the restaurant itself.

A Spearman rank correlation between the number of likes and the nearby counts of the basic categories within 200 m is applied. The results are listed below:

Category	**Rank r**
Outdoors & Recreation	0.21
Nightlife Spot	0.20
Shop & Service	0.20
Professional & Other Places	0.18
Arts & Entertainment	0.18
Travel & Transport	0.17
Food	0.16
Residence	0.15
College & University	0.13
Event	0.03
Total Number of Nearby Venues	**0.20**

It can be found that the number of likes is generally positively correlated with the number of the nearby venues. The correlation coefficient of the number of likes and the number of nearby venues is 0.2, similar as the highest r of the single basic category.

Most of the correlations between the likes of a restaurant and the counts of the nearby venues of the subcategories are positive. A more intensive positive correlation with the venues of certain category than the correlation with the total number of venues are not found, which means the major effect is the density of the city. However, the result indicate that the bus stop, bus station and other Asian restaurants may not play an important role in the number of likes of a restaurant.

Category	**Rank r**	**Test p**	**Mean Number of Venues**	**Std of the Means**	**Std/Mean**
Coffee Shop	0.21	0.00	6	6	1.0
Athletics & Sports	0.20	0.00	9	9	1.0
Clothing Store	0.20	0.00	37	59	1.6
Bar	0.19	0.00	23	36	1.6
Office	0.18	0.00	26	34	1.3

Café	0.18	0.00	15	15	1.1
Salon / Barbershop	0.16	0.00	15	24	1.6
Food & Drink Shop	0.16	0.00	12	11	0.9
Medical Center	0.15	0.00	24	40	1.6
Residential Building (Apartment / Condo)	0.15	0.00	12	11	0.9
Bakery	0.15	0.00	9	7	0.8
Hotel	0.15	0.00	7	9	1.3
Asian Restaurant	0.12	0.00	119	92	0.8
Bus Stop	0.08	0.00	4	2	0.7
Bus Station	0.03	0.00	2	2	1.2

Considering the weight of the categories that is not influenced by the city density, all correlations of the likes with the venue counts within a 200 m range are not strong. But considering the range of 2000 m, there is a positive correlation between number of likes and the number of residential buildings and barbershops. The correlation with barbershops is resulted from that the barbershops often comes up together with residential buildings. If there is a activity influence between eating and haircuts is uncertain.

Venue counts within 200 meters:

Category	**Rank r**	**Test p**
Office	0.14	0.00
Coffee Shop	0.10	0.00
Residential Building (Apartment / Condo)	0.09	0.00
Athletics & Sports	0.07	0.00
Hotel	0.07	0.00
Bakery	0.07	0.00
Bus Station	0.06	0.00
Asian Restaurant	0.03	0.00
Salon / Barbershop	0.00	0.84
Bus Stop	0.00	0.94
Medical Center	0.00	0.60
Food & Drink Shop	-0.01	0.17
Clothing Store	-0.03	0.00
Café	-0.03	0.00
Bar	-0.06	0.00

Venue counts within 2000 meters:

Category	**Rank r**	**Test p**
Salon / Barbershop	0.25	0.00
Residential Building (Apartment / Condo)	0.25	0.00

Clothing Store	0.17	0.00
Medical Center	0.14	0.00
Hotel	0.12	0.00
Athletics & Sports	0.10	0.00
Office	0.06	0.00
Bus Station	0.02	0.04
Food & Drink Shop	0.01	0.24
Café	-0.02	0.06
Coffee Shop	-0.07	0.00
Bakery	-0.08	0.00
Bar	-0.10	0.00
Bus Stop	-0.13	0.00
Asian Restaurant	-0.17	0.00

4 Summary

In conclusion, Hong Kong city is a very dense city with a dense distribution of various types of restaurants. The high density makes the clustering of the restaurant nearby venues less obvious. But a differentiation of restaurants in different districts still exists, the restaurants were clustered with colleges, commercial center, tourism center and political & commercial center. Besides, the location with similar surroundings of the centers also has an opportunity for restaurants. The circumstances are not a major factor for the popularity of a restaurant, but the popularity does increase with the density of the venues, i.e. the city density. It also proves that an expensive location does worth its price in Hong Kong.