

Does Surrounding Venues Affects Hotel Occupancy in Hong Kong?

Introduction

Before building up a hotel, many factors must be considered, such as land price, construction cost, as well as the anticipated income. Given that the income is basically proportional to the occupancy, this project will find out whether surrounding venues have influence on the hotel's occupancy and calculate the approximate coefficient.

Data

Source

The Hong Kong Tourism Board (HKTb, <https://partnernet.hktb.com/en/home/index.html>) provides reports of [Hotel Supply Situation\(Seasonally Published\)](#) and [Hotel Room Occupancy Report\(Monthly Published\)](#) in .xls form.

By 30 Sept 2018, there're a total of 287 hotels and 1,504 guesthouses authenticated in Hong Kong. During the month, 206 hotels and 79 guesthouses responded the HKTb's monthly survey, which will be the main data source of this project.

And the project will scrape a part of these two reports and store the data in .csv file, then uploaded into DB2 for further analysis.

Parameters

The project will use the following three datasets:

1. HOTEL_CATEGORY

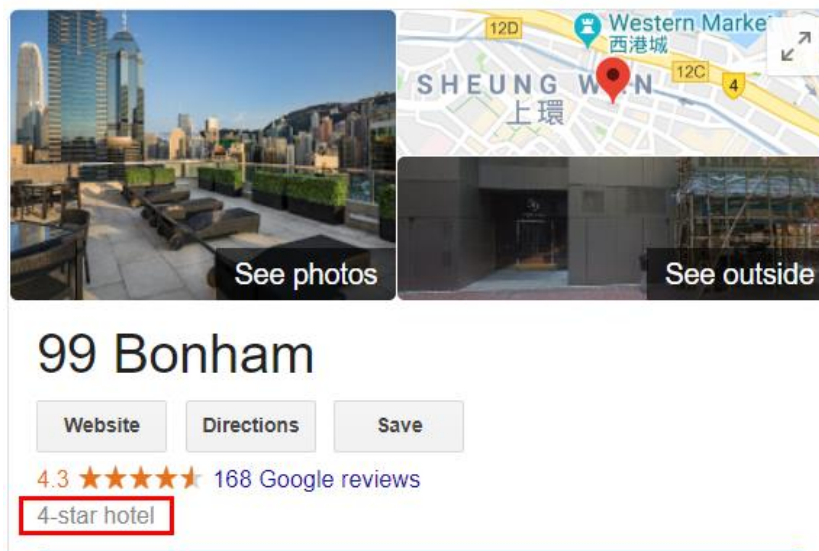
	HOTEL_NAME VARCHAR(80)	LOCATION VARCHAR(18)		TOTAL_ROOMS SMALLINT	CATEGORY VARCHAR(19)
1	338 Apartment	Central & Western		62	Unclassified Hotels
2	60 West Suites Hotel	Central & Western		60	3-star
3	99 Bonham	Central & Western		84	4-star
4	AKVO Hotel	Central & Western		30	Unclassified Hotels
5	Aveny	Central & Western		11	Unclassified Hotels
6	Best Western Hotel Harbour View	Central & Western		432	4-star
7	Best Western Plus Hong Kong	Central & Western		307	4-star
8	Bishop Lei International House	Central & Western		227	4-star
9	Butterfly On LKF	Central & Western		34	4-star
10	Butterfly On Waterfront	Central & Western		90	4-star

HOTEL_NAME --- Names of all authenticated hotels in Hong Kong. Guesthouses will not be included in the analysis.

LOCATION --- There're totally 18 District Council Districts in Hong Kong: Central & Western, Eastern Hong Kong, Southern Hong Kong, Wan Chai, Kowloon City, Kwun Tong, Sham Shui Po, Wong Tai Sin, Yau Tsim Mong, Kwai Tsing, North, Sai Kung, Sha Tin, Tai Po, Tsuen Wan, Tuen Mun, Yuen Long and Islands. Among the districts, North and Tai po aren't in the list due to there's no authenticated hotel in these areas. (See https://en.wikipedia.org/wiki/Districts_of_Hong_Kong for more information)

TOTAL_ROOMS --- Total number of rooms of each hotel, rooms under repair or being refurbished are excluded

CATEGORY --- Hotel ratings collected from Google



2. OCC_BY_CATEGORY

	CATEGORY VARCHAR(20)	NO._OF_HOTELS SMALLINT	NO._OF_ROOMS SMALLINT	HOTEL_ROOM_OCCUPANCY_RATE SMALLINT	AVERAGE_ACHIEVED_HOTEL_ROOM_RATE SMALLINT
1	High Tariff A Hotels	36	18839	88	2073
2	High Tariff B Hotels	101	30800	91	1110

CATEGORY --- Instead hotel star-rating, the Hong Kong Hotel Classification System classifies hotels into 3 categories: High Tariff A Hotel (similar to 5-star hotel), High Tariff B Hotel (similar to 4-star hotel) and Medium Tariff Hotel. However, the category of each hotel is not publicly available, only individual hotels are informed of their respective category so that they can compare their own performance against their category averages in the hotel industry. Thus, this project will use the star ratings for analysis.

NO._OF_HOTELS --- Total numbers of hotels of each category

NO._OF_ROOMS --- Total numbers of rooms of each category. Rooms under repair or being refurbished are excluded.

HOTEL_ROOM_OCCUPANCY_RATE --- Based on daily rooms occupied against daily rooms available for sales of responded hotels and guesthouses to HKTB Monthly Hotel Room Occupancy Survey.

AVERAGE_ACHIEVED_HOTEL_ROOM_RATE (ARR) --- The average of ARR of

responded hotels and guesthouses to HKTB Monthly Hotel Room Occupancy Survey. ARR excluded Government tax and all non-room related components such as F&B, laundry, airport transfer, service charges, etc., which have been built into the room rate.

3.OCC_BY_DISTRICT

	DISTRICT VARCHAR(29)	NO_OF_HOTELS SMALLINT	NO_OF_ROOMS SMALLINT	HOTEL_ROOM_OCCUPANCY SMALLINT
1	Central & Western	53	9037	89
2	Wan Chai	52	11301	91
3	Eastern & Southern Hong Kong	20	6245	87
4	Tsim Sha Tsui	57	17088	93
5	Yau Ma Tei & Mong Kok	40	6810	95

DISTRICT --- Based on the 18 District Council Districts, Eastern and Southern Hong Kong are merged together, Kowloon City is merged into Other Kowloon along with Kwun Tong, Sham Shui Po and Wong Tai Sin. New Territories includes Kwai Tsing, Sai Kung, Sha Tin, Tsuen Wan, Tuen Mun and Yuen Long. And due to large number of hotels, Yau Tsim Mong is separated into Tsim Sha Tsui and Yau Ma Tei & Mong Kok.

NO._OF_HOTELS --- Total numbers of hotels of each district

NO._OF_ROOMS --- Total numbers of rooms of each district. Rooms under repair or being refurbished are excluded.

HOTEL_ROOM_OCCUPANCY_RATE --- Based on daily rooms occupied against daily rooms available for sales of responded hotels and guesthouses to HKTB Monthly Hotel Room Occupancy Survey.

Methodology

Preprocessing

In order to combine all three datasets, first we need to match the expression of 'Location' and 'Category' of the hotels.


Then, each hotel is assigned with an occupancy rate which is the average rate of its location and category. And category 'Unclassified hotel' is dropped because it is consisted of all kinds of hotels which will make the results less precise.

Next, the project will use Google Maps Places API to get the coordinates of each hotel, and require the surrounding venues with Foursquare API. The venue categories will be processed by one hot encoding.

Analysis

Since 'surrounding venues' includes multiple variables, to calculate its relationship with occupancy rate, we need to build a linear regression model and fit all venues. And the square-root of the R score of the regression will be the coefficient of the total correlation.

$$R = \frac{\sum (y - \bar{y})(\hat{y} - \bar{y})}{\sqrt{\sum (y - \bar{y})^2 \sum (\hat{y} - \bar{y})^2}} \quad R^2 = \frac{[\sum (y - \bar{y})(\hat{y} - \bar{y})]^2}{\sum (y - \bar{y})^2 \sum (\hat{y} - \bar{y})^2}$$



$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

To check whether the surrounding venues are determining factors of occupancy rate, we need to get the coefficients of each venue category and them as a standard matrix. By multiplying the coefficients with one hot code of the venues, we can create a scoring system of each hotel. Hotels with more positive venues and less negative venues will have higher score. After k-Means clustering, we will find the hotels (neighborhood) which have the score that is closet to our standard and check if they have the highest occupancy rate.

Results

The analysis found out that the multiple correlation coefficient of surrounding venues and hotel occupancy rate is 0.96594, which shows that there is strong linear association between the two variables.

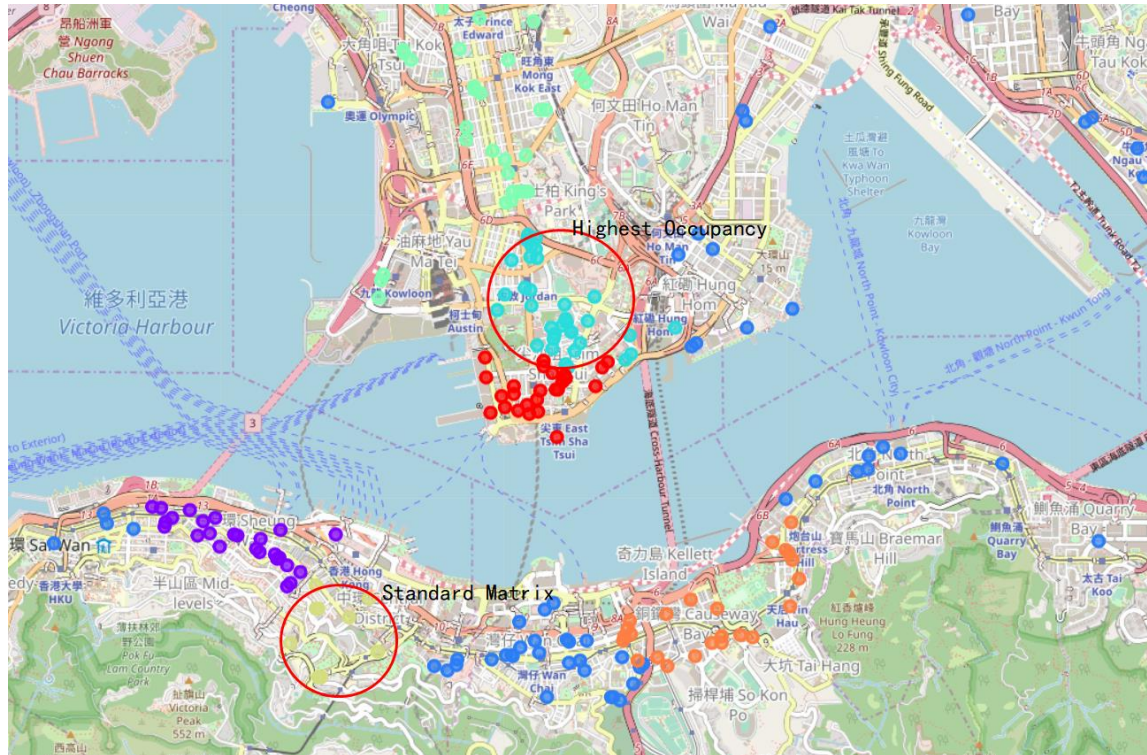
By calculating the coefficient of each venue, below categories have the most positive linear correlation with occupancy. *Only categories with p value<0.001 are considered.*

	x	spearman_coef	p_value
0	Food & Drink Shop	0.529191	8.342865e-20
1	Miscellaneous Shop	0.508328	3.678894e-18
2	Cosmetics Shop	0.507589	4.187250e-18
3	Tailor Shop	0.471320	1.658501e-15
4	Arts & Crafts Store	0.448367	5.148361e-14
6	Cricket Ground	0.436434	2.781226e-13
8	Hotel	0.433536	4.147933e-13
11	Record Shop	0.415734	4.453870e-12
12	Shoe Store	0.408053	1.189019e-11
13	History Museum	0.405692	1.599761e-11

And the following venues have the most negative linear correlation with occupancy rate.

	x	spearman_coef	p_value
5	Art Gallery	-0.442211	1.239222e-13
7	Bar	-0.434042	3.869152e-13
9	Yoga Studio	-0.427450	9.486411e-13
10	Gym	-0.419110	2.869828e-12
16	Burger Joint	-0.385734	1.794684e-10
20	Beer Store	-0.337675	3.220676e-08
21	Salon / Barbershop	-0.331866	5.697035e-08
23	Wine Shop	-0.329416	7.221350e-08
24	Furniture / Home Store	-0.326374	9.665188e-08
27	Cupcake Shop	-0.309051	4.786265e-07

The clustering result showed that although the standard matrix was located in Central & Western area, the average occupancy rate of this cluster was the lowest among all clusters while hotels in Yau Tsim Mong had the highest occupancy rate.



Discussion

According to our common sense, the occupancy of hotels should be reduced if there're more competitors around. However, the analysis showed that both hotel and hostel are positively correlated to the occupancy. It may imply that in this city of limited territory and extremely high land price, the accommodation is still in short supply.

Bar	-0.434042	3.869152e-13	0.434042
Hotel	0.433536	4.147933e-13	0.433536
Yoga Studio	-0.427450	9.486411e-13	0.427450
Gym	-0.419110	2.869828e-12	0.419110
Record Shop	0.415734	4.453870e-12	0.415734
Shoe Store	0.408053	1.189019e-11	0.408053
History Museum	0.405692	1.599761e-11	0.405692
Shopping Mall	0.402575	2.358861e-11	0.402575
Hostel	0.401118	2.824402e-11	0.401118
Burger Joint	-0.385734	1.794684e-10	0.385734

Hong Kong is well known as 'The Heaven of Shopping', which has always been one of the main purposes when visitors came here. The shops/stores on the top of the coefficient list may not only have the most positive correlation with occupancy, but also could be the most popular shops/stores among the visitors.

	spearman_coef	p_value
x		
Food & Drink Shop	0.529191	8.342865e-20
Miscellaneous Shop	0.508328	3.678894e-18
Cosmetics Shop	0.507589	4.187250e-18
Tailor Shop	0.471320	1.658501e-15
Arts & Crafts Store	0.448367	5.148361e-14
Art Gallery	-0.442211	1.239222e-13
Cricket Ground	0.436434	2.781226e-13
Bar	-0.434042	3.869152e-13
Hotel	0.433536	4.147933e-13
Yoga Studio	-0.427450	9.486411e-13
Gym	-0.419110	2.869828e-12
Record Shop	0.415734	4.453870e-12
Shoe Store	0.408053	1.189019e-11
History Museum	0.405692	1.599761e-11
Shopping Mall	0.402575	2.358861e-11
Hostel	0.401118	2.824402e-11

Despite the strong association between surrounding venues and occupancy rate, the clustering showed very different results. Not Surprisingly, hotels with the highest occupancy locate at the most popular area among tourists. Meanwhile, Central & Western area where the standard matrix fell into, is more like a commercial center, instead of tourism attraction. It may suggest that when tourists choosing their accommodation, they would take the location into account but not the specific venues. The high R value of total surroundings could be a proof for this hypothesis.

Conclusion

This project used multiple variable regression, simple correlation analysis and k-Means clustering, showing that surroundings have very strong association with hotel occupancy rate. But it also showed that occupancy would be affected by many other factors.

For further analysis, the relationship between occupancy and factors like land price, construction cost, seasons may be determined. Combining all the factors, we can build a precise recommendation system for site selection of hotels.