Location Analysis for Restaurant in Singapore

Tan Yong Sin

1st May 2020

1. Introduction

1.1 Background

Among all the factors that make or break a restaurant, location is definitely one of the most important. Prime location is deemed such importance as it guarantees the entrepreneur one less factor to worry about: Customer Traffic. Nevertheless, in reality there are only so much prime location in an area, which therefore came the term "prime". Although there are many aspects that constitutes a good location, one cannot neglect the fact that an unsuitable location will cause unnecessary loss, and in worst case scenario causing the restaurant unsustainable.

This project aims to provide some analysis of a good location through the usage of existing restaurant distribution and existing population distribution. In this project, I choose Singapore as current target study location and I'll be providing some analysis based on available data.

1.2 Problem

A suitable location will consist of 2 factors. The first factor will of course being having adequate space for the desired restaurant business model. The second factor of a suitable location will be the one that have a lot of customer traffic flow around the location. This study will revolves around the second factor.

One of the indicators that helps identify the second factor of a location is the current distribution of restaurant around the area. Restaurant distribution not only shows the competition around the area, but it can also assume that current customer traffic around the area is able to sustain these restaurants.

The second indicator that will help will be the population distribution around the area. The probability of customer traffic flow is higher as the population density around the area increases. Albeit there are also cases like CBD where there will be high traffic flow due to offices but population is actually non-existent.

1.3 Objective

Other than using basic mapping and graphs for visualization, I'll also be using correlation to confirm the relationship of these indicators with the amount of restaurant in its respective area. The assumption made in this study will be that the amount of restaurant in an area will correlate the suitability of a location for a new restaurant, which means it will be a good location.

Therefore, this study will determine the relationship between the amount of restaurant in an area with other indicators. This study will also determine a suitable location for the next restaurant using the correlation result.

2. Data

2.1 Data Sources

The first prerequisite required to conduct this study will be to identify the areas in Singapore. The Master Plan 2014 Singapore Planning Area and Boundaries shows areas in Singapore and divide them into 55 areas.

The distribution of restaurant over Singapore are collected through Foursquare. Due to the limitation of 50 results per call, the data are compiled by searching restaurant within 6km radius of every MRT & LRT station in Singapore. The coordinates for MRT and LRT station from Singapore is obtained from Kaggle Source.

Next, the data of the resident population by planning area and ethnic group is used to identify the correlation with different ethnicity. The distribution of ethnicity at each area are also tabulated as shown in the Resident Population by Planning Area/Subzone, Ethnic Group and Sex, 2015.

Lastly, the resident population by planning area and types of dwelling is used to identify the correlation with different income level. The income level of the population at the area is roughly estimated based on the type of household as shown in the Resident Population by Planning Area/Subzone and Type of Dwelling, 2015.

2.2 Data Cleaning

The data obtained from each source are analysed and cleaned before put into use. Data of year 2014 and 2015 are used as they are the latest data available provided by Singapore government. The accuracy of the data will be much reliable. The planning area and MRT station are confirmed to be as current location and therefore used as it is. While for the population distribution of each area according to dwelling types, the total amount does not tally with the sum of the types provided. However, sum of types provided does tally with sum of population distribution of each area according to ethnic group and sex and therefore sum of the dwellings are used instead of pre-existing total column.

For the distribution of restaurant obtained through Foursquare, there are several changes made before it is being further analyzed. The radius is set to be 6km as that is the maximum distance between 2 station, while for stations less than 6km will have duplicated restaurant names. The restaurant in Malaysia but also included in the list are filtered. The data listed in the location and categories are in the form of lists, they are further processed and concatenate back to the original data before being used.

Some of the restaurants does not have types assign to them as well. The restaurants that do not have assigned types mostly have similar names with their venues right beside one another when after confirming, the location has only one restaurant of that name. Also, some of the restaurant does not exist or ended their operation sometimes ago. Therefore, those without assigned types are omitted and the types are rearranged according to different cultures, differentiated by areas and countries.

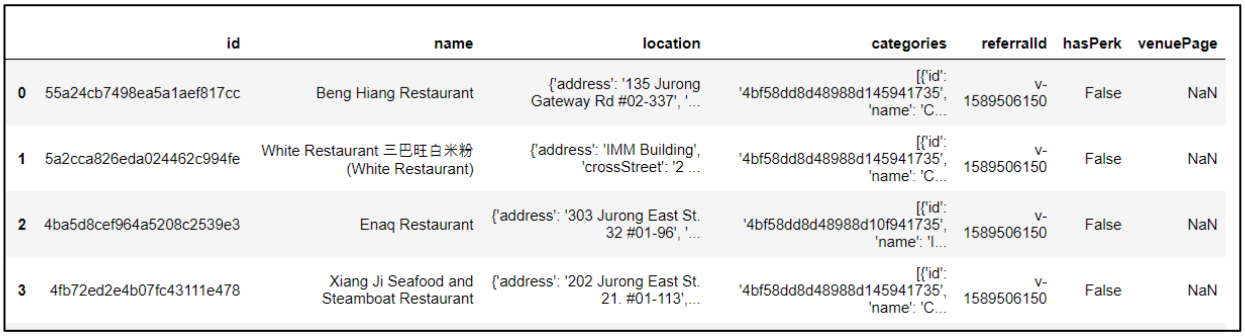
2.3 Feature Selection

The restaurants are roughly divided into 90 types on the first query. The types are rearranged according to different cultures, differentiated by areas and countries. For example, restaurants under Halal category are grouped accordingly to the type of food they offered, Malay food into the Malay category while Indian Muslim food will be under Indian food category.

For the population distribution by types of dwellings, all types of HDB flats are group into a same category for further analysis. While only the population is only separated by ethnic group but not by gender. Total male and female at each area is sum up and only separated according to ethnic group.

Figure 2.3.1 below shows the initial table obtained from Foursquare. The **id** is unique to each venue regardless of restaurant name and therefore it is used to remove any duplicate restaurant from the query.

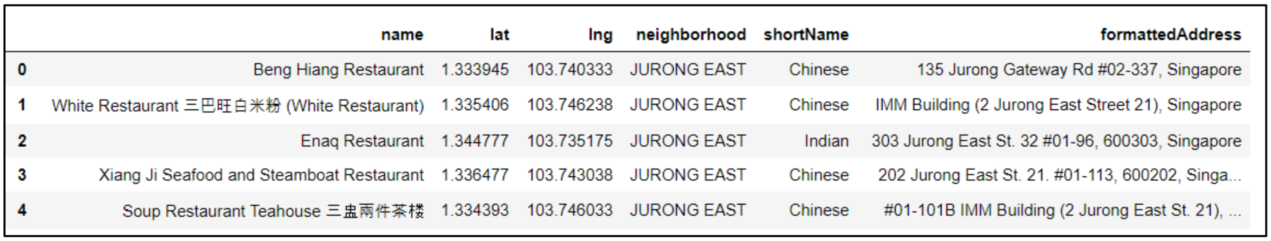
Figure 2.3.1: Table Obtained from Foursquare Query



	id	name	location	categories	referralId	hasPerk	venuePage
0	55a24cb7498ea5a1aef817cc	Beng Hiang Restaurant	{'address': '135 Jurong Gateway Rd #02-337', '...	[{'id': '4bf58dd8d48988d145941735', 'name': 'C...	v-1589506150	False	NaN
1	5a2cca826eda024462c994fe	White Restaurant 三巴旺白米粉 (White Restaurant)	{'address': 'IMM Building', 'crossStreet': '2 ...	[{'id': '4bf58dd8d48988d145941735', 'name': 'C...	v-1589506150	False	NaN
2	4ba5d8cef964a5208c2539e3	Enaq Restaurant	{'address': '303 Jurong East St. 32 #01-96', '...	[{'id': '4bf58dd8d48988d10f941735', 'name': 'I...	v-1589506150	False	NaN
3	4fb72ed2e4b07fc43111e478	Xiang Ji Seafood and Steamboat Restaurant	{'address': '202 Jurong East St. 21. #01-113',...	[{'id': '4bf58dd8d48988d145941735', 'name': 'C...	v-1589506150	False	NaN

As per Figure 2.3.1 above, information such as types and venues are all within the_columns like categories and location in the form of list. They are all extracted and transformed into data frame before concatenate back to the main table. Columns like **venuePage**, **referralId**, **cc**, **postalCode**, **pluralName** are dropped after filtering. This is because they will not be use in the further part of analysis. The column **shortName** is used as the types of restaurant assigned to them.

Figure 2.3.2: Table after Processed



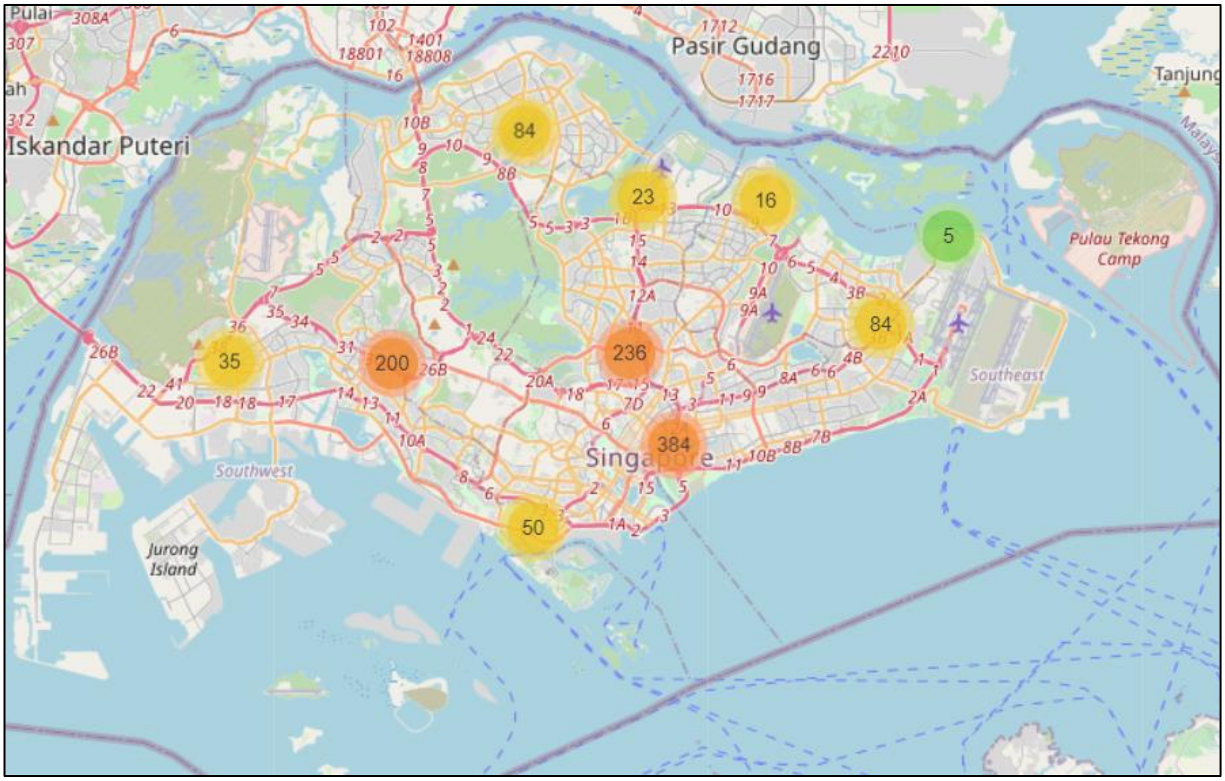
	name	lat	lng	neighborhood	shortName	formattedAddress
0	Beng Hiang Restaurant	1.333945	103.740333	JURONG EAST	Chinese	135 Jurong Gateway Rd #02-337, Singapore
1	White Restaurant 三巴旺白米粉 (White Restaurant)	1.335406	103.746238	JURONG EAST	Chinese	IMM Building (2 Jurong East Street 21), Singapore
2	Enaq Restaurant	1.344777	103.735175	JURONG EAST	Indian	303 Jurong East St. 32 #01-96, 600303, Singapore
3	Xiang Ji Seafood and Steamboat Restaurant	1.336477	103.743038	JURONG EAST	Chinese	202 Jurong East St. 21. #01-113, 600202, Singa...
4	Soup Restaurant Teahouse 三盅兩件茶楼	1.334393	103.746033	JURONG EAST	Chinese	#01-101B IMM Building (2 Jurong East St. 21), ...

Figure 2.3.2 shows the sample table being used for analysis. The neighborhood of the restaurant is calculated by checking if the coordinates of the restaurant is within the area as stated by the planning area using Shapely within() function and assign the area name accordingly.

3. Methodology

After pre-processed the restaurant data obtained from Foursquare, the restaurant is plotted onto Singapore map as shown in Figure 3.1. All the restaurants are plotted onto the Singapore and group into clusters for visualization. The restaurant is not separated according to each planning area in this figure.

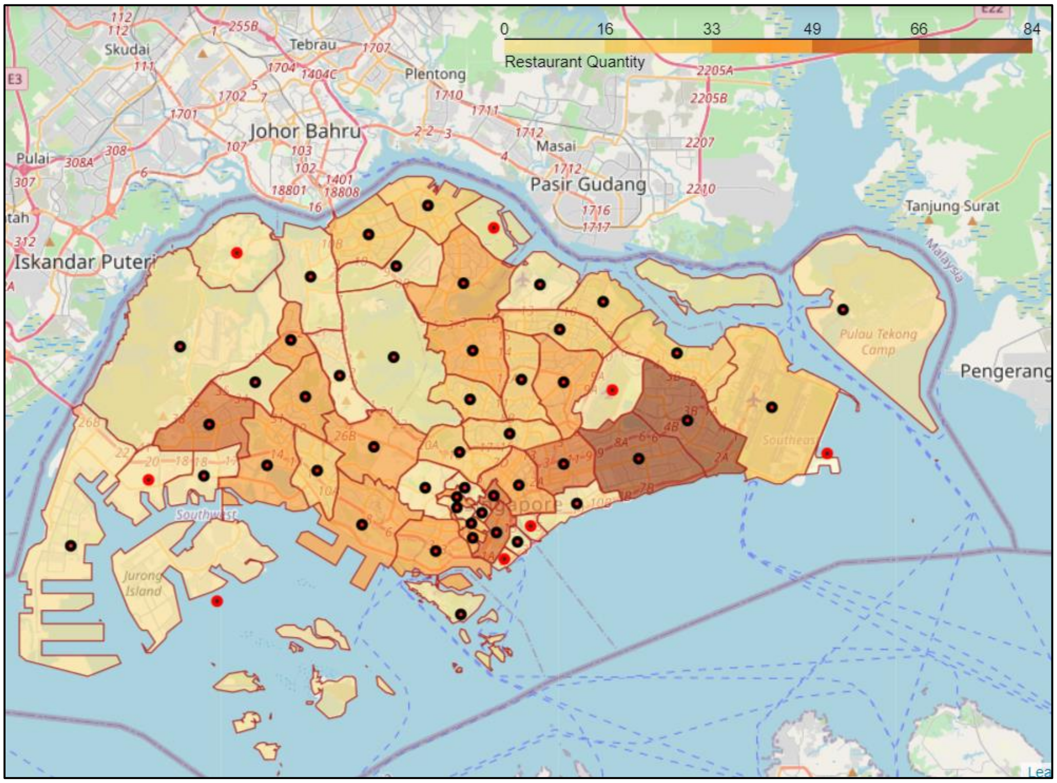
Figure 3.1: Restaurant Clusters in Singapore



According to Figure 3.1 above, it is shown that most of the restaurant is grouped near 3 areas. 1 is near Jurong East while the other 2 is located at South and South East region of Singapore.

By using the planning area data from Singapore Data, the area which each restaurant belongs to is updated into the table. The distribution of restaurant by planning area is then plotted in the form of choropleth map as shown in Figure 3.2 using the data given.

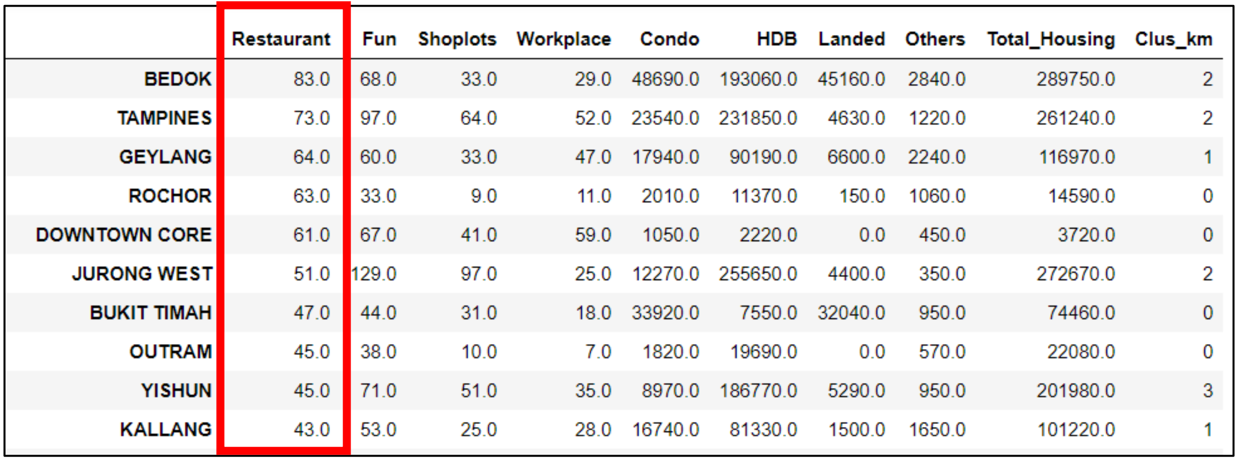
Figure 3.2: Restaurant Distribution by Planning Area



In Figure 3.2, the black dot shows the area name and restaurant quantity. Meanwhile dots in red are those planning area that have no restaurant. Color intensity of the map shows the quantity of restaurant at each respective area with darker color means high quantity of restaurant at that area.

The top 10 area in Singapore with most amount of Restaurant is tabulated as shown in Table 3.1 below. Among the top 10 areas, 7 of them are located at South – South East Region of Singapore. They account for around 35% of total restaurant in Singapore. In Table 3.1, only Jurong West, Bukit Timah and Yishun is not in that region.

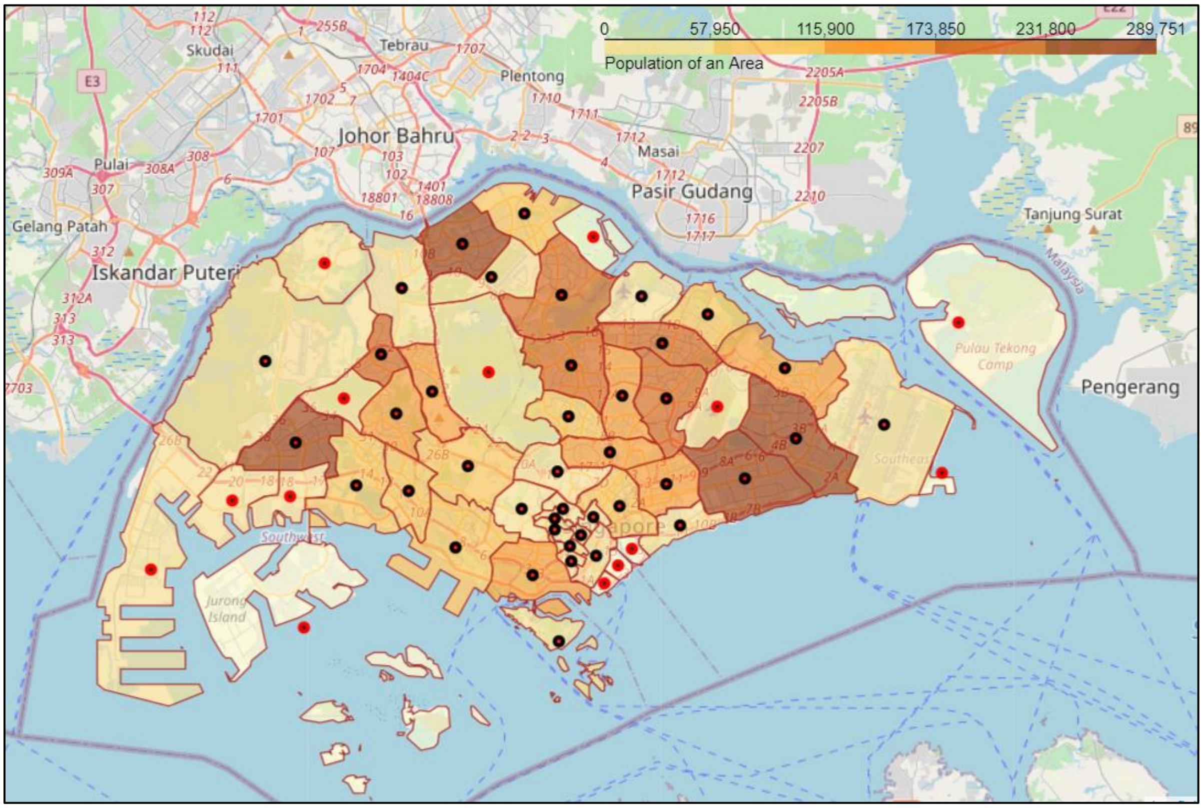
Table 3.1: List of Restaurant by Planning Area



	Restaurant	Fun	Shoplots	Workplace	Condo	HDB	Landed	Others	Total_Housing	Clus_km
BEDOK	83.0	68.0	33.0	29.0	48690.0	193060.0	45160.0	2840.0	289750.0	2
TAMPINES	73.0	97.0	64.0	52.0	23540.0	231850.0	4630.0	1220.0	261240.0	2
GEYLANG	64.0	60.0	33.0	47.0	17940.0	90190.0	6600.0	2240.0	116970.0	1
ROCHOR	63.0	33.0	9.0	11.0	2010.0	11370.0	150.0	1060.0	14590.0	0
DOWNTOWN CORE	61.0	67.0	41.0	59.0	1050.0	2220.0	0.0	450.0	3720.0	0
JURONG WEST	51.0	129.0	97.0	25.0	12270.0	255650.0	4400.0	350.0	272670.0	2
BUKIT TIMAH	47.0	44.0	31.0	18.0	33920.0	7550.0	32040.0	950.0	74460.0	0
OUTRAM	45.0	38.0	10.0	7.0	1820.0	19690.0	0.0	570.0	22080.0	0
YISHUN	45.0	71.0	51.0	35.0	8970.0	186770.0	5290.0	950.0	201980.0	3
KALLANG	43.0	53.0	25.0	28.0	16740.0	81330.0	1500.0	1650.0	101220.0	1

The population distribution in Singapore is collected from Singapore Data by getting the population data based on types of dwelling at an area. Overall number of people living at an area is the total of the people from all types of dwellings in that area. The population is plotted using choropleth map on Singapore for simple visualization as shown in Figure 3.3 below.

Figure 3.3: Population Distribution by Planning Area

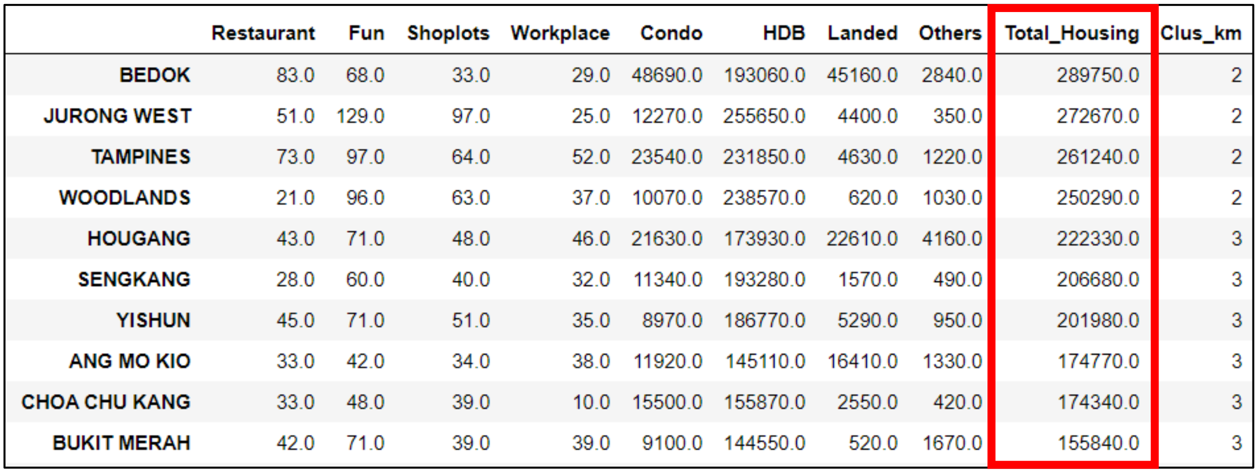


The red dot in the area signifies the area with 0 population. Do note the work permit holders that are staying in dormitories are not taken into account. As shown in Figure 3.3, one can see some difference with Figure 3.2 of Restaurant Distribution by Area. In Figure 3.3, Woodlands and Jurong East have a lot of people staying there but the amount of restaurant at these areas are much less significant.

The top 10 areas with most people staying are also tabulated as shown in Table 3.2 below. Through Table 3.2, it is found out that South East region, namely Bedok and Tampines accounted around 15% of total population. The next big cluster is Ang Mo Kio, Sengkang, Hougang and Yishun area where they accounted around 20% of total population.

Shops of other category are also collected to identify their possible correlation with amount of restaurant at the area. Offices and Business Center are collected and named as Workplace as workplaces can also generate a lot of foot traffic. Other than that, entertainment and nightlife are combined into the Fun category to see the effect of pubs, cinemas, stadium on restaurant quantity. Lastly the Shoplots category includes all sorts of boutiques or shopping malls.

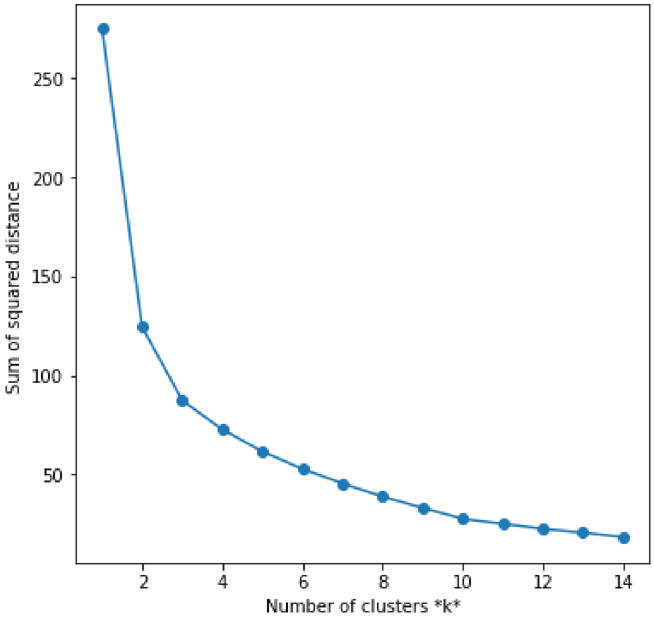
Table 3.2: List of Population by Planning Area



	Restaurant	Fun	Shoplots	Workplace	Condo	HDB	Landed	Others	Total_Housing	Clus_km
BEDOK	83.0	68.0	33.0	29.0	48690.0	193060.0	45160.0	2840.0	289750.0	2
JURONG WEST	51.0	129.0	97.0	25.0	12270.0	255650.0	4400.0	350.0	272670.0	2
TAMPINES	73.0	97.0	64.0	52.0	23540.0	231850.0	4630.0	1220.0	261240.0	2
WOODLANDS	21.0	96.0	63.0	37.0	10070.0	238570.0	620.0	1030.0	250290.0	2
HOUGANG	43.0	71.0	48.0	46.0	21630.0	173930.0	22610.0	4160.0	222330.0	3
SENGKANG	28.0	60.0	40.0	32.0	11340.0	193280.0	1570.0	490.0	206680.0	3
YISHUN	45.0	71.0	51.0	35.0	8970.0	186770.0	5290.0	950.0	201980.0	3
ANG MO KIO	33.0	42.0	34.0	38.0	11920.0	145110.0	16410.0	1330.0	174770.0	3
CHOA CHU KANG	33.0	48.0	39.0	10.0	15500.0	155870.0	2550.0	420.0	174340.0	3
BUKIT MERAH	42.0	71.0	39.0	39.0	9100.0	144550.0	520.0	1670.0	155840.0	3

K means clustering is used to identify the similarity and difference between areas. As shown in Graph 3.1 below, the changes in inertia reduces as the number of clusters increases. The sum of squared distance, also known as inertia, is able to provide a rough estimate on the suitability of the clusters. This is because inertia calculates the squared distance between each point within a cluster. Lower inertia is favorable as it means the difference between points in a cluster is less.

Graph 3.1: Inertia against Number of Clusters K & Gradient at each Inertia

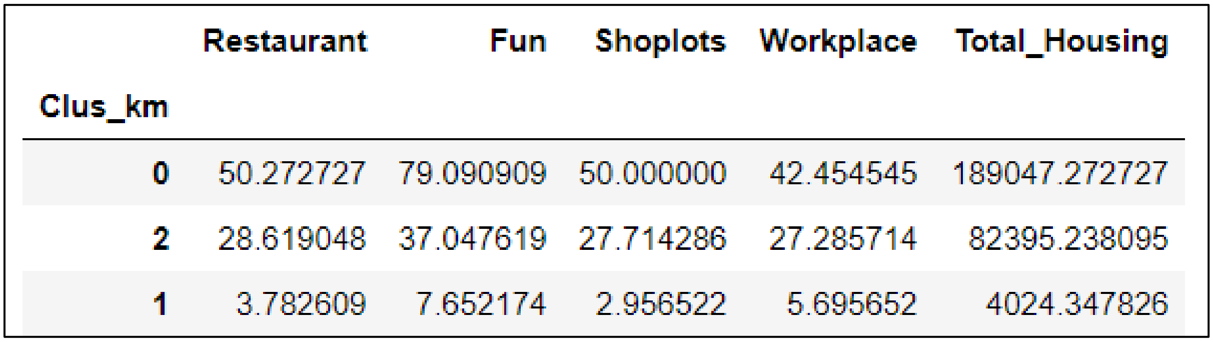


As refer to the Graph 3.1 left, the graph bends and slowly plateau. There are 2 elbows in this graph. The first obvious elbow is when k=2, while the second elbow is when k=3. In this study k=3 is used for more detail grouping of areas according to their aspects. To ensure the data is not affected by different parameters between populations and quantity of shops, the data undergoes fit transform before being used in calculation for K Means Clustering for a more accurate result.

4. Results

As refer Table 4.1 below, the clusters are grouped to calculate the mean number of restaurant and listed accordingly. It is shown that Cluster 0 has most number of restaurant on average and also have more shops, offices, and residents compared to other areas.

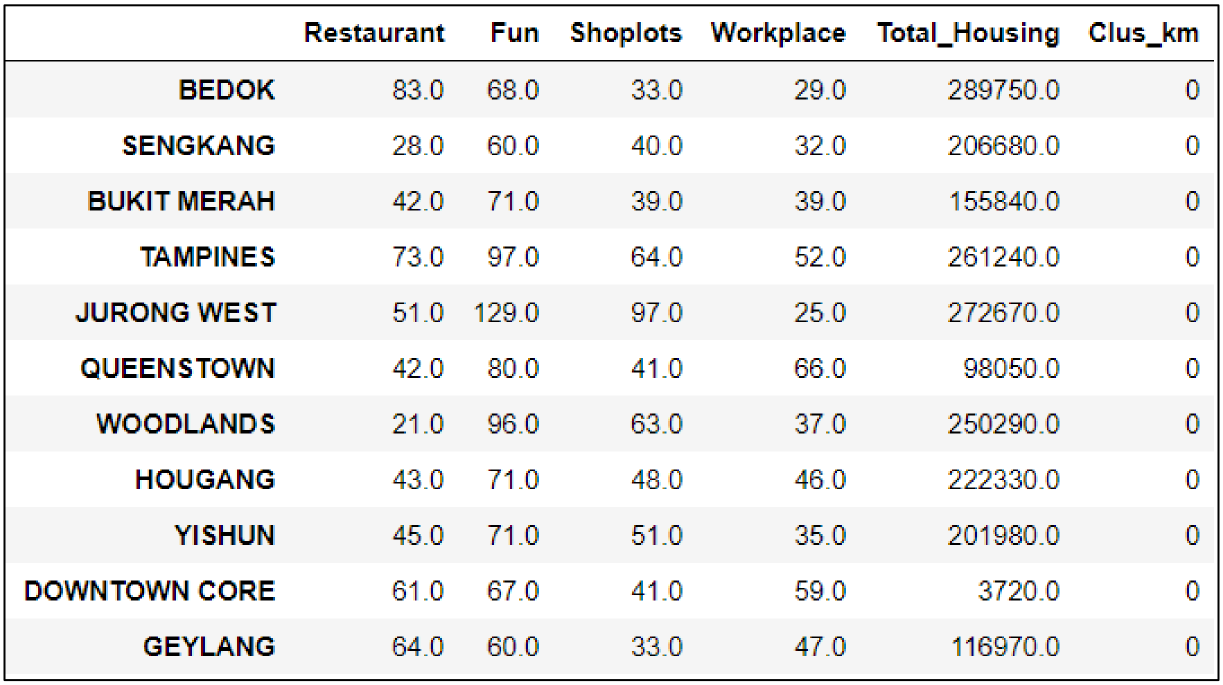
Table 4.1: List of Clusters by Number of Restaurant



	Restaurant	Fun	Shoplots	Workplace	Total_Housing
Clus_km					
0	50.272727	79.090909	50.000000	42.454545	189047.272727
2	28.619048	37.047619	27.714286	27.285714	82395.238095
1	3.782609	7.652174	2.956522	5.695652	4024.347826

After listing out the details in cluster 0 as shown in Table 4.2 below, it is shown that 6 of the top 10 areas with most restaurant is in Table 4.2 below. The remaining 4 areas are in cluster 2 as they have lower quantity of other type of shops. When compared with Table 3.2, it is found out that 7 of the top 10 areas with most population is within cluster 0. It is also shown that among the other 3 categories, the minimum quantity of shops under the Fun, Shoplots and Workplace category in cluster 0 is higher than the average quantity in cluster 2.

Table 4.2: List of Planning Area by Clusters

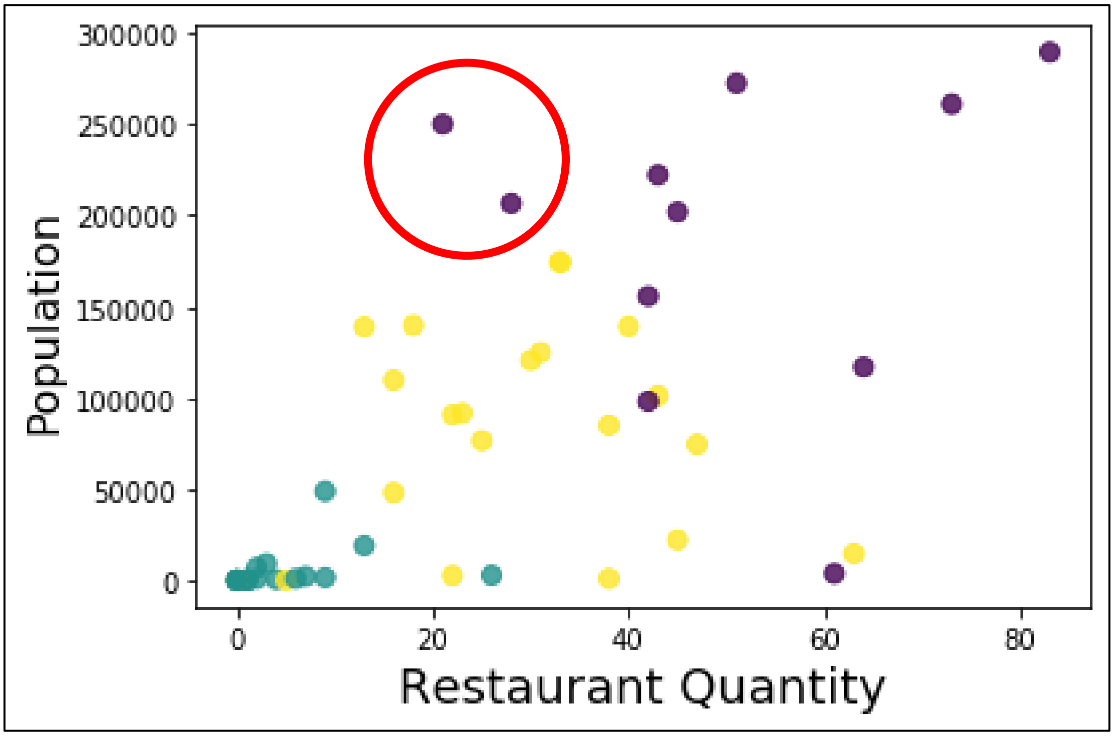


	Restaurant	Fun	Shoplots	Workplace	Total_Housing	Clus_km
BEDOK	83.0	68.0	33.0	29.0	289750.0	0
SENGKANG	28.0	60.0	40.0	32.0	206680.0	0
BUKIT MERAH	42.0	71.0	39.0	39.0	155840.0	0
TAMPINES	73.0	97.0	64.0	52.0	261240.0	0
JURONG WEST	51.0	129.0	97.0	25.0	272670.0	0
QUEENSTOWN	42.0	80.0	41.0	66.0	98050.0	0
WOODLANDS	21.0	96.0	63.0	37.0	250290.0	0
HOUGANG	43.0	71.0	48.0	46.0	222330.0	0
YISHUN	45.0	71.0	51.0	35.0	201980.0	0
DOWNTOWN CORE	61.0	67.0	41.0	59.0	3720.0	0
GEYLANG	64.0	60.0	33.0	47.0	116970.0	0

After calculation, it is found out that cluster 0 encompasses 45% of all restaurant, 48% of all shops in the Fun category, 46% of all shoplots, 40% of all workplaces and 53% of all Singapore population. Cluster 0 is only 11 areas out of total 55 areas in Singapore.

Graph 4.1 below shows the scatter plot for the quantity of restaurant against population distribution at the planning area. The three colors indicate the 3 different clusters.

Graph 4.1: Scatter Plot for Restaurant Quantity against Total Population



From Graph 4.1, one can see that Restaurant Quantity is scatter across areas of varying population. They do not shows a linear relationship whereby the restaurant quantity increases when the population increases. Nevertheless, it does shows that places having more population have higher tendencies to have more restaurant. Purple color indicates the planning area in cluster 0.

As refer to both Graph 4.1 and Table 4.2, it is found out that despite having similar quantity of offices, shoplots, places for entertainment and high amount of residents staying in the same area compared to other areas within the same cluster, the restaurant amount for Woodlands and Sengkang (as marked in Graph 4.1 in red circle) has only 50% of the next area with lowest quantity of restaurant.   Therefore, these 2 areas can be considered to be a good location for the restaurant.

5. Discussion

There are several limitations in this study as previously mention during the data section. Factors such as limitation in calls due to premium usage and the inherent nature of Foursquare which collects data through crowdsourcing efforts hinders the progress of study. Also, during this study, it is found out that there are many discrepancies in the locations provided. This can easily cause the unwanted issue of “garbage in, garbage out” which all data analyst has been trying to avoid. A lot of time would have been saved if the accuracy of the data provided could be further increased. For example the categories of the shoplots can actually be limited by Foursquare, and to promote user to identify non-existent location. I believe data of quality is also an essential base structure to provide a solid data analysis.

An attempt has been done to include type of housing available in the area in the K means clustering but it eventually skewed the clustering towards population staying in the area. Also another attempt on including quantity of ethnic races in the area has been run but no significant changes is shown to correlates with the quantity of restaurant in the region.

6. Conclusion

Through this study it is found out that Woodlands and Sengkang shows its potential to be a good location for a restaurant. These 2 areas have the quantity of shop lots required to generate the foot traffic in its area and yet having low quantity of restaurant which greatly reduces competition. The large number of entertainment shop lots, shopping centers and offices in the area will greatly promote the foot traffic in the region and therefore provide a good foundation for the restaurant to come.