Data Science Final Capstone Project

Introduction:

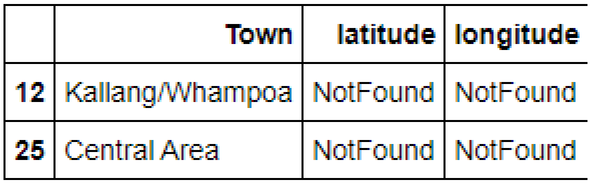
Singapore is a small and condensed country. With increasing population and land shortage, it is difficult to determine which location will be the best to begin a recreational business.

Thus, this project aims to provide business owner an overview on where Singaporean visits the most in each town of Singapore. From there, we can predict where an ideal location should be.

Recreational business owner will be interested in this project.

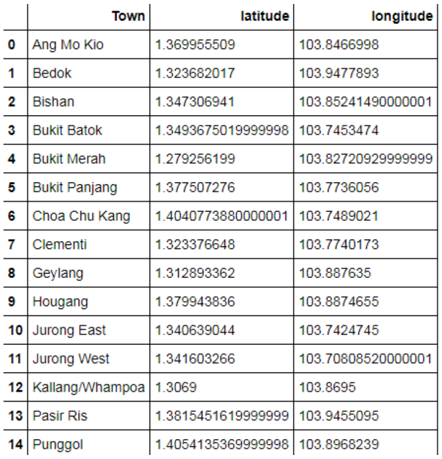
Data Extraction and Cleaning:

I have downloaded the list of Towns in Singapore from Wikipedia and made used of OneMap API to obtain longitude and latitude of each town in Singapore. There are two rows in which OneMap API does not return results.



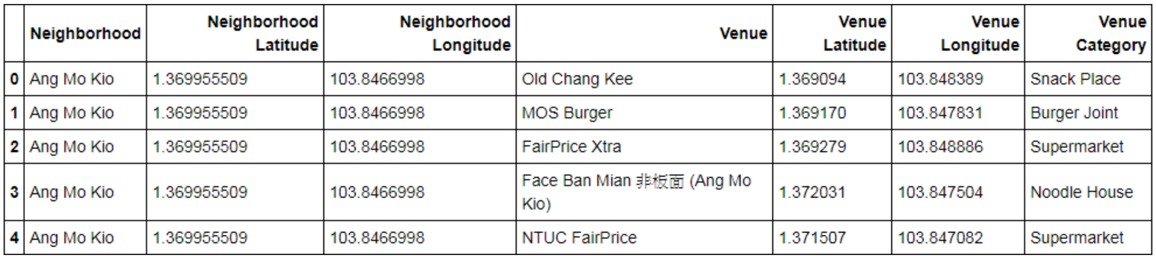
	Town	latitude	longitude
12	Kallang/Whampoa	NotFound	NotFound
25	Central Area	NotFound	NotFound

Based on local knowledge, “Kallang/Whampoa” refers to Kallang area, “Central Area” refers to City hall or nearby area. Thus, I have manually added the latitude and longitude for the mentioned above two rows. Below is a partial view of the complete table with respective coordinates.



	Town	latitude	longitude
0	Ang Mo Kio	1.369955509	103.8466998
1	Bedok	1.323682017	103.9477893
2	Bishan	1.347306941	103.85241490000001
3	Bukit Batok	1.3493675019999998	103.7453474
4	Bukit Merah	1.279256199	103.82720929999999
5	Bukit Panjang	1.377507276	103.7736056
6	Choa Chu Kang	1.4040773880000001	103.7489021
7	Clementi	1.323376648	103.7740173
8	Geylang	1.312893362	103.887635
9	Hougang	1.379943836	103.8874655
10	Jurong East	1.340639044	103.7424745
11	Jurong West	1.341603266	103.70808520000001
12	Kallang/Whampoa	1.3069	103.8695
13	Pasir Ris	1.3815451619999999	103.9455095
14	Punggol	1.4054135369999998	103.8968239

Next, I will obtain Foursquare location data to further understand the pattern/behavior of locals in the country by running Foursquare API. Below is the table which includes all results which have gotten from Foursquare.

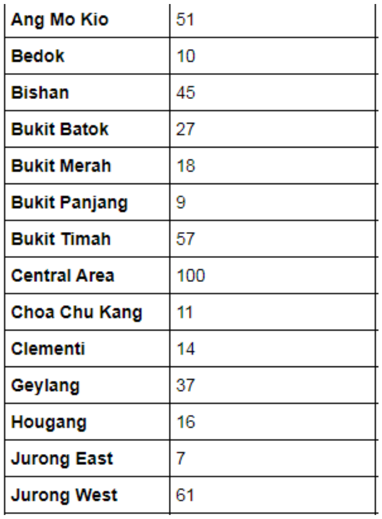


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ang Mo Kio	1.369955509	103.8466998	Old Chang Kee	1.369094	103.848389	Snack Place
1	Ang Mo Kio	1.369955509	103.8466998	MOS Burger	1.369170	103.847831	Burger Joint
2	Ang Mo Kio	1.369955509	103.8466998	FairPrice Xtra	1.369279	103.848886	Supermarket
3	Ang Mo Kio	1.369955509	103.8466998	Face Ban Mian 非板面 (Ang Mo Kio)	1.372031	103.847504	Noodle House
4	Ang Mo Kio	1.369955509	103.8466998	NTUC FairPrice	1.371507	103.847082	Supermarket

Methodology (Data Analysis):

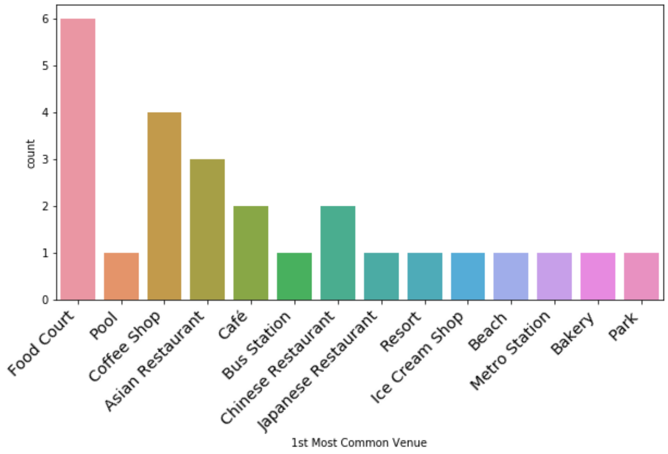
Exploratory Data Analysis:

Let’s take a look at the spread of Data which we have obtained from Foursquare. Below table shows the number of results for each town.

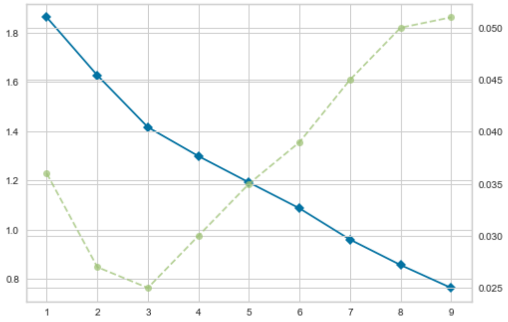


Ang Mo Kio	51
Bedok	10
Bishan	45
Bukit Batok	27
Bukit Merah	18
Bukit Panjang	9
Bukit Timah	57
Central Area	100
Choa Chu Kang	11
Clementi	14
Geylang	37
Hougang	16
Jurong East	7
Jurong West	61

I have transformed the table so that it captures both coordinates and occurrence of each venue category. Then, I have passed the table to a function so that it shows most common places in each town. Food court and Coffee shop have significantly high occurrence overall in Singapore. This aligns with the fact that Singaporean loves to visit Food Court and coffee shop. In 80% of the town, first common places are food places and eateries. Food competition in Singapore is indeed high. Recreational business seems to be less popular in Singapore and let’s dive further to observe how the distribution of each cluster.



I will use K means clustering which is an unsupervised machine learning algorithm to group each town with similar features. Let’s make use of Elbow method to find the optimal K value (number of cluster) based on this data set. The distortion value for each K represents by blue line in the graph below.



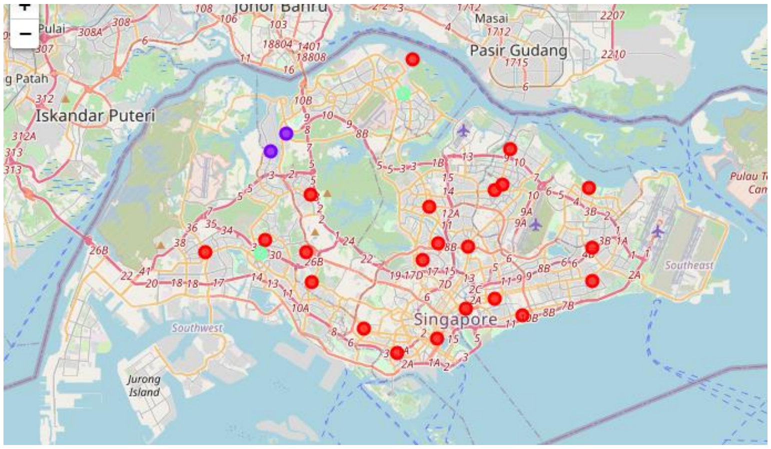
It seems that there is no elbow point. After further look into data for each town, I have decided that K to be 3 as this is the most appropriate value due to similar features for this town.

1. Cluster 1 -- Leisure and recreational activities

2. Cluster 2 -- Shopping and food such as coffee shop, Dessert shop etc.

3. Cluster 3 -- Transportation and convenience store

I have generated a map to provide us with an overview of cluster distribution. Purple – cluster 1, Red-cluster 2, Green-cluster 3.



Discussion

All 3 clusters reflect a phenomenon with food court and coffee shop being topped on most common places. Thus, the clusters are distinct from one another and the most contributing factors are highly determined by 2nd and 3rd most common places instead. This feature could be one of the reasons that why we cannot find an optimal K from the elbow method.

I observe that based on the map above, cluster 2 is mostly residential area and central area and these are where most shopping and eateries are located. These areas are most populated. Cluster 1 is located near nature reserves such as Bukit Timah nature reserve and zoos.

We can see that recreational facilities and parks are concentrated in north-west area of Singapore.

Conclusion

I have analysed most common places in each town by using K means clustering. I have identified that there are three clusters across Singapore. Food court and coffee shops are most common places. Competition is high in the north-west area of Singapore. Based on data collected, all cluster first common places are food court and coffee shop. Food Court usually comes together with shopping centers. Recreational owner may want to explore business in small scale and incorporate the idea in shopping malls since most people or traffic is located in cluster 2.