Coursera

IBM Data Science Professional Certificate

Applied Capstone Project

Finding Locations for Restaurants Business Operation in Singapore

Thomas Chua

7 Sep 2020

1 Introduction

1.1 Singapore is reputed as a global city with a vibrant Food and Beverage (F&B) industry. Arguably, it is celebrated as the culinary capital of Asia with cuisines from all parts of the world including the Mediterranean, Italian, Thai, Indian, Malaysian, Moroccan, Mexican and African, all finding their space here. The sector has crossed over $ 8.3 billion in annual receipts with over 6,500 establishments and touching one percent contribution in the country's GDP.

1.2 With eating out being a national passion, more and more Singaporeans putting in longer work hours, and ever-growing tourism sector, the F&B sector is set to grow even further.

2 Problem Statement

2.1 As much that a F&B business in Singapore can be potentially rewarding, the industry is highly competitve with a market of fast changing taste, not to mention high rental and overhead costs among the challenges.

2.2 This project aspires to provide data based insights to help addressing the "where", "what" and "who" questions for starting a F&B business:

2.2.1 "Where" – This relates to the location of the business

2.2.2 "What" – The cuisine type to offer

3 Data

3.1 For performing analysis, the following data are obtained:

3.1.1 List of Singapore Planning Areas (http://en.wikipedia.org/wiki/Planning_Areas_of_Singapore). The Planning Areas shall be referred to as Towns in this study.

3.1.2 List of MRT Stations with the Towns they are located (https://en.wikipedia.org/wiki/List_of_Singapore_MRT_stations)

3.1.3 Latitude and Longitude (lat/long) coordinates of MRT Stations (https://www.onemap.sg/)

3.2 Instead of using Towns as location references, MRT stations are used because they are nexuses of activities with linkages to shopping malls and town centres located in close proximity.

3.3 Web scraping on (i) and (ii) was done using beautifulsoup packages. Python Geocoder package was used to get the lat/long coordinates of Singapore. To get lat/long coordinates of the MRT stations, a python wrapper for client OneMapSG API was used. Documentation for this wrapper is provided in the link https://pypi.org/project/onemapsg/.

3.4 Foursquare API requests are then made to obtain restaurants within 350m of every MRT station. This is achieved with the condition set for venue category to contain "Restaurant". The

search radius of 350m is choosen because this is deemed to be an acceptable walking distance for diners who commute by train.

4 Methodology

4.1 Data acquisition and wrangling

4.1.1 A datafram *df_plan_areas* is created by scraping a tabulated data on Singapore towns and related information, such as area, population and population density, from https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore. Web scrapping is done using beautifulsoup package.

[4]:

	Name(English)	Malay	Chinese	Pinyin	Tamil	Region	Area	Population	Density(/km2)
0	Ang Mo Kio		宏茂桥	Hóng mào qiáo	ஆங் மோ கியோ	North-East	13.94	163,950	13,400
1	Bedok	*	勿洛	Wù luò	பிடோக்	East	21.69	279,380	13,000
2	Bishan		碧山	Bì shān	பீஷான்	Central	7.62	88,010	12,000
3	Boon Lay		文礼	Wén lǐ	பூன் லே	West	8.23	30	3.6
4	Bukit Batok	*	武吉巴督	Wǔjí bā dū	புக்கிட் பாத்தோக்	West	11.13	153,740	14,000
5	Bukit Merah	*	红山	Hóng shān	புக்கிட் மேரா	Central	14.34	151,980	11,000
6	Bukit Panjang	*	武吉班让	Wǔjí bān ràng	பக்கிட் பஞ்சாங்	West	8.99	139,280	15,000
7	Bukit Timah	*	武吉知马	Wǔjí zhī mǎ	புக்கித் திமா	Central	17.53	77,430	4,400
8	Central Water Catchment	Kawasan Tadahan Air Tengah	中央集水区	Zhōngyāng jí shuǐ qū	மத்திய நீர் நீர்ப்பிடிப்பு	North	37.15	*	*
9	Changi	*	樟宜	Zhāng yí	சாங்கி	East	40.61	1,830	80.62

4.1.2 The dataframe is then wrangled to drop certain columns that will not be involved in subsequent analysis, replace any zero value for “Population” with mean. This new dataframe is *df_plan_areas3.*

	Towns	Region	Town Area Coverage (km2)	Population
0	Ang Mo Kio	North-East	13.94	163950
1	Bedok	East	21.69	279380
2	Bishan	Central	7.62	88010
3	Boon Lay	West	8.23	30
4	Bukit Batok	West	11.13	153740
5	Bukit Merah	Central	14.34	151980
6	Bukit Panjang	West	8.99	139280
7	Bukit Timah	Central	17.53	77430
8	Central Water Catchment	North	37.15	73203
9	Changi	East	40.61	1830

4.1.3 A dataframe *df_mrt_stns* is created by scraping tabulated information on MRT stations in Singapore from https://en.wikipedia.org/wiki/List_of_Singapore_MRT_stations. The

dataframe is then wrangled to drop rows and columns that will not be used in subsequent analysis. Column headings are also renamed. For example, the column heading ”Locations” is renamed as “Towns” to facilitate subsequent joining with *df_plan_areas3*. Rows with NaN column values are also dropped using pandas.dropna() function.

	Alpha-numeric code(s)	Alpha-numeric code(s).1	Station name	Station name.1	Station name.2	Opening	Name(s) during planning stages	Abbreviation	Location(s)	Connection(s) to other transport
0	In operation	Future	English • Malay	Chinese	Tamil	Opening	Name(s) during planning stages	Abbreviation	Location(s)	Connection(s) to other transport
1	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)	North South Line (NSL)
2	NS1 EW24	JE5	Jurong East	裕廊东	ஜூரோங் கிழக்கு	10 March 1990	Jurong East	JUR	Jurong East	Jurong East Temporary Bus Interchange
3	NS2	NaN	Bukit Batok	武吉巴督	புக்கிட் பாத்தோக்	10 March 1990	Bukit Batok South	BBT	Bukit Batok	Bukit Batok Bus Interchange
4	NS3	NaN	Bukit Gombak	武吉甘柏	புக்கிட் கோம்பாக்	10 March 1990	Bukit Batok North	BGB	Bukit Batok	NaN

4.1.4 The eventual dataframe, *df_mrt_stns4* is obtained as partially shown below.

	Station Name	Towns	Existing Station Codes	Future Station Codes	Connecting Transport Interchange
0	Jurong East	Jurong East	NS1 EW24	JE5	Jurong East Temporary Bus Interchange
1	Bukit Batok	Bukit Batok	NS2	NaN	Bukit Batok Bus Interchange
2	Bukit Gombak	Bukit Batok	NS3	NaN	NaN
4	Choa Chu Kang	Choa Chu Kang	NS4 BP1	JS1	Choa Chu Kang Bus Interchange
5	Yew Tee	Choa Chu Kang	NS5	NaN	NaN
7	Kranji	Sungei Kadut	NS7	NaN	NaN
8	Marsiling	Woodlands	NS8	NaN	NaN
9	Woodlands	Woodlands	NS9 TE2	NaN	Woodlands Temporary Bus Interchange
10	Admiralty	Woodlands	NS10	NaN	NaN

4.1.5 Both *df_plan_areas3* and *df_mrt_stns4* are then merged on the column “Towns” to form *df_merge as partially shown below.*

	Station Name	Towns	Existing Station Codes	Future Station Codes	Connecting Transport Interchange
0	Jurong East	Jurong East	NS1 EW24	JE5	Jurong East Temporary Bus Interchange
1	Bukit Batok	Bukit Batok	NS2	NaN	Bukit Batok Bus Interchange
2	Bukit Gombak	Bukit Batok	NS3	NaN	NaN
4	Choa Chu Kang	Choa Chu Kang	NS4 BP1	JS1	Choa Chu Kang Bus Interchange
5	Yew Tee	Choa Chu Kang	NS5	NaN	NaN
7	Kranji	Sungei Kadut	NS7	NaN	NaN
8	Marsiling	Woodlands	NS8	NaN	NaN
9	Woodlands	Woodlands	NS9 TE2	NaN	Woodlands Temporary Bus Interchange
10	Admiralty	Woodlands	NS10	NaN	NaN
11	Sembawang	Sembawang	NS11	NaN	Sembawang Bus Interchange
12	Canberra	Sembawang	NS12	NaN	NaN
13	Yishun	Yishun	NS13	NaN	Yishun Bus Interchange
14	Khatib	Yishun	NS14	NaN	NaN
15	Yio Chu Kang	Ang Mo Kio	NS15	NaN	Yio Chu Kang Bus Interchange
16	Ang Mo Kio	Ang Mo Kio	NS16	CR11	Ang Mo Kio Bus Interchange
17	Bishan	Bishan	NS17 CC15	NaN	Bishan Bus Interchange
18	Braddell	Toa Payoh	NS18	NaN	NaN
19	Toa Payoh	Toa Payoh	NS19	NaN	Toa Payoh Bus Interchange
20	Novena	Novena	NS20	NaN	NaN
21	Newton	Newton	NS21 DT11	NaN	NaN

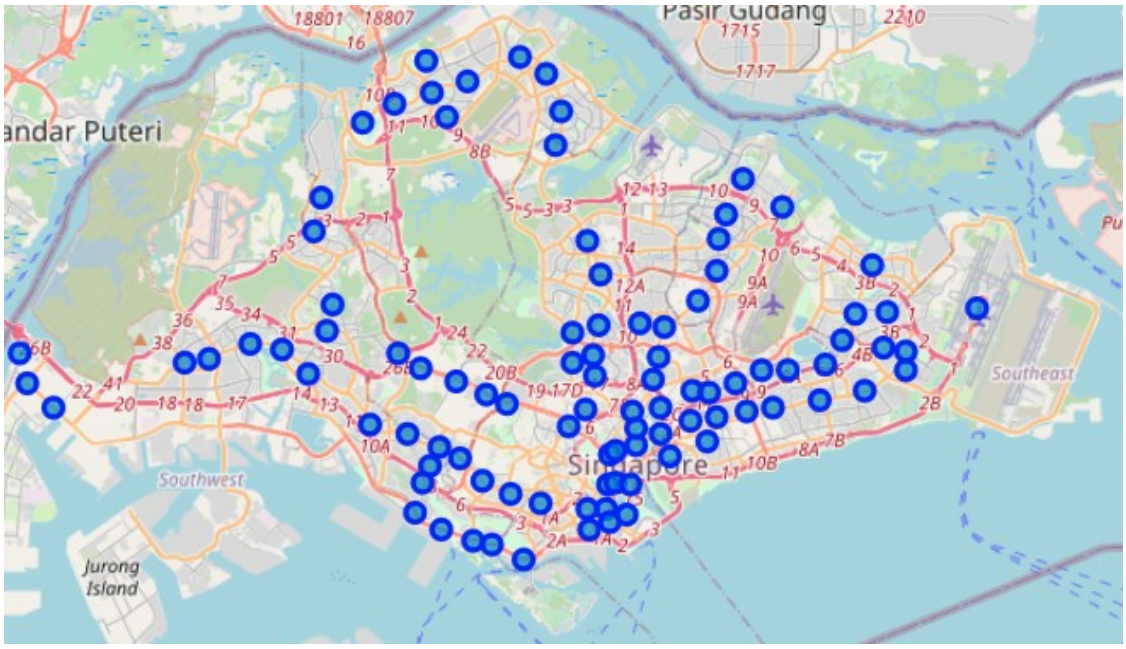
4.1.6 Columns “Stations per Town” and “Crowd Density” (“town population” divided by “Stations per Town”) are then added to *df_merge* as shown partially below.

	Station Name	Towns	Existing Station Codes	Future Station Codes	Connecting Transport Interchange	Region	Town Area Coverage (km2)	Population	Station(s) per Town	Crowd Density Per Station
0	Chinese Garden	Jurong East	EW25	NaN	NaN	West	17.83	79240	2	2222.097588
1	Jurong East	Jurong East	NS1 EW24	JE5	Jurong East Bus Interchange	West	17.83	79240	2	2222.097588
2	Bukit Batok	Bukit Batok	NS2	NaN	Bukit Batok Bus Interchange	West	11.13	153740	2	6906.558850
3	Bukit Gombak	Bukit Batok	NS3	NaN	NaN	West	11.13	153740	2	6906.558850
4	Yew Tee	Choa Chu Kang	NS5	NaN	NaN	West	6.11	190890	2	15621.112930
5	Choa Chu Kang	Choa Chu Kang	NS4 BP1	JS1	Choa Chu Kang Bus Interchange	West	6.11	190890	2	15621.112930
6	Kranji	Sungei Kadut	NS7	NaN	NaN	North	15.99	780	1	48.780488
7	Marsiling	Woodlands	NS8	NaN	NaN	North	13.59	254730	5	3748.785872
8	Admiralty	Woodlands	NS10	NaN	NaN	North	13.59	254730	5	3748.785872
9	Woodlands North	Woodlands	TE1 RTS	NaN	NaN	North	13.59	254730	5	3748.785872

4.1.7 Lat and long coordinates of every MRT station are obtained from OneMapSG and added as two seperate columns to *df_merge* and store as *df_merge3 as partially shown below.*

	Station Name	Towns	Existing Station Codes	Future Station Codes	Connecting Transport Interchange	Region	Town Area Coverage (km2)	Population	Station(s) per Town	Crowd Density Per Station	Latitude	Longitude
0	Chinese Garden	Jurong East	EW25	NaN	NaN	West	17.83	79240	2	2222.097588	1.342352821	103.7325967
1	Jurong East	Jurong East	NS1 EW24	JE5	Jurong East Bus Interchange	West	17.83	79240	2	2222.097588	1.333152816	103.7422863
2	Bukit Batok	Bukit Batok	NS2	NaN	Bukit Batok Bus Interchange	West	11.13	153740	2	6906.558850	1.349033312	103.7495665
3	Bukit Gombak	Bukit Batok	NS3	NaN	NaN	West	11.13	153740	2	6906.558850	1.3586115909999998	103.7517909
4	Yew Tee	Choa Chu Kang	NS5	NaN	NaN	West	6.11	190890	2	15621.112930	1.3975350690000001	103.7474052

With the lat and long coordinates, the locations of the MRT stations are plotted using *folium.map()*.



4.1.1 Foursquare API calls were made to acquire top five venues within 350m of every MRT station and stored as dataframe *df_FnBvenues*. Search was conditioned to return venues with

keyword “Restaurant” in the “Venue Category”. The code was tested using Chinese Garden MRT station.

	name	categories	lat	lng
0	Xing Yun Hainanese Boneless Chicken Rice	Asian Restaurant	1.344723	103.733746
1	Joo Siah Bak Koot Teh 裕城肉骨茶	Chinese Restaurant	1.344726	103.733745

4.1.2 The same search was then carried out for all the MRT stations. A total of 220 entries were returned by the search.

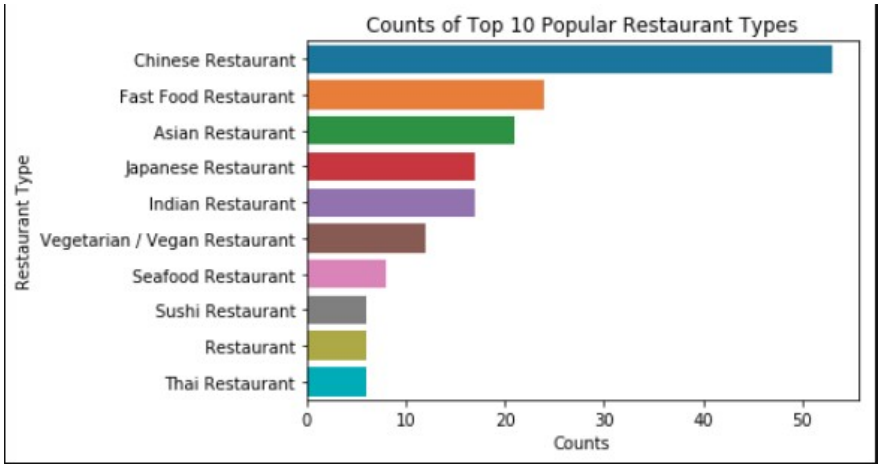
```
print(sg_FnBvenues.shape)
(220, 7)
```

4.1.3 A new dataframe *df_merge4* is then created by joining df_FnBvenues using df_merge3 using *pandas.join(df_FnBvenues.set_index(‘Station Name), on=’Station Name’)*.

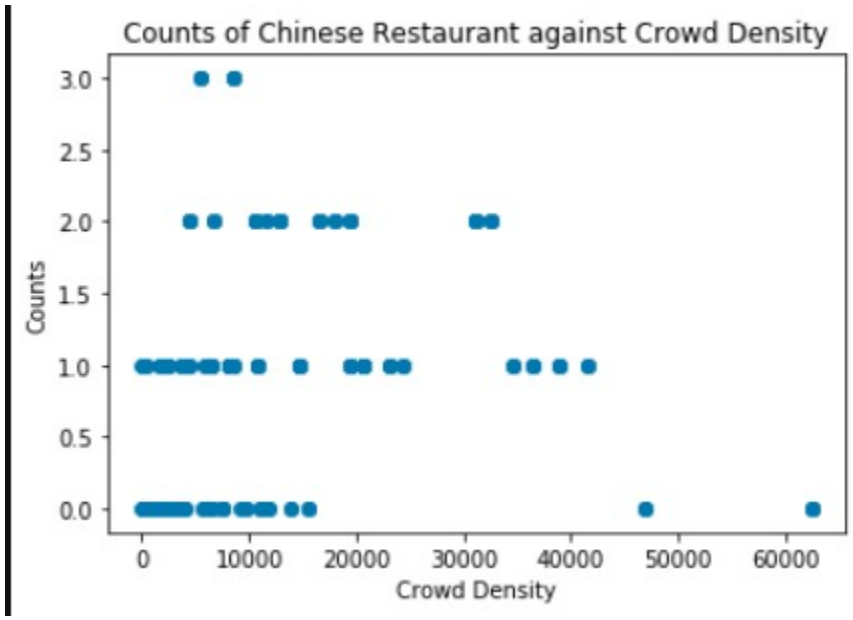
4.2 Exploratory Analysis

4.2.1 The data frame is grouped by “Venue Category”, and counts of Restaurant Type are obtained using the pandas.value_counts() method. The values are sorted in descending order. The top 10 restaurant types found within 350m of all MRT stations are shown below.

	Restaurant Type	Counts
0	Chinese Restaurant	53
1	Fast Food Restaurant	24
2	Asian Restaurant	21
3	Japanese Restaurant	17
4	Indian Restaurant	17
5	Vegetarian / Vegan Restaurant	12
6	Seafood Restaurant	8
7	Sushi Restaurant	6
8	Restaurant	6
9	Thai Restaurant	6



4.2.2 Since Chinese Restaurant is of the highest total counts, relationship between crowd density and counts of Chinese Restaurant is explored using a scatter plot.



It appears that there is no linear relationship between the two variable.

4.2.3 The relationship between crowd density and restaurant types is further examined by computing the pearson p-values for the top 10 common restaurant types.

	Restaurant Type	p_value
29	Thai Restaurant	0.360426
28	Sushi Restaurant	0.291626
6	Fast Food Restaurant	0.271617
2	Chinese Restaurant	0.181349
25	Seafood Restaurant	0.056789
1	Asian Restaurant	0.011385
24	Restaurant	0.001572
13	Indian Restaurant	-0.000178
30	Vegetarian / Vegan Restaurant	-0.078464
15	Japanese Restaurant	-0.105755

The range of p-values is $-0.5 < p < 0.5$. This indicates that there is no linear relationship between crowd density and counts of every restaurant type on the top 10 common types.

4.3 One-hot encoding

4.3.1 For further analysis, one-hot encoding is performed on "Venue Category" using pandas.get_dummies() method on the dataframe in Table 4. This converts categorical variables into dummy/indicator variables. The resulting dataframe, *sg_merge4_onehot*, contains 33 columns. This dataframe is then grouped by 'Station Name' with sum() to obtain *df_merge4grp.*

	Station Name	American Restaurant	Asian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Dim Sum Restaurant	Dumpling Restaurant	Fast Food Restaurant	Filipino Restaurant	French Restaurant	...	Portuguese Restaurant	Ramen Restaurant	Restaurant	Seafood Restaurant	Shaanxi Restaurant
0	Admiralty	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1	Aljunied	0	0	1	0	0	0	0	0	0	...	0	0	0	1	0
2	Ang Mo Kio	0	1	0	0	0	0	1	0	0	...	0	0	0	0	0
3	Bayfront	0	0	0	0	1	1	0	0	0	...	0	0	0	0	0
4	Beauty World	0	0	2	0	0	0	0	0	0	...	0	0	0	0	0

4.4 Top 5 Most Common Restaurant Type

4.4.1 The top 5 most common restaurant types within 350m of every MRT station were obtained and stored as a dataframe *df_merge4sort.*

	Station Name	1st Most Common Restaurant Type	2nd Most Common Restaurant Type	3rd Most Common Restaurant Type	4th Most Common Restaurant Type	5th Most Common Restaurant Type
0	Admiralty	Halal Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Chinese Restaurant
1	Aljunied	Vegetarian / Vegan Restaurant	Chinese Restaurant	Seafood Restaurant	Japanese Restaurant	Comfort Food Restaurant
2	Ang Mo Kio	Asian Restaurant	Fast Food Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Chinese Restaurant
3	Bayfront	Dim Sum Restaurant	Dumpling Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant
4	Beauty World	Chinese Restaurant	Korean Restaurant	Indian Restaurant	Vietnamese Restaurant	Italian Restaurant

4.5 K-Means Clustering with Machine Learning

4.5.1 K-Means clustering is a type of partition clustering that divides data into K non-overlapping subsets or clusters without any cluster internal structure or labels. It is an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.

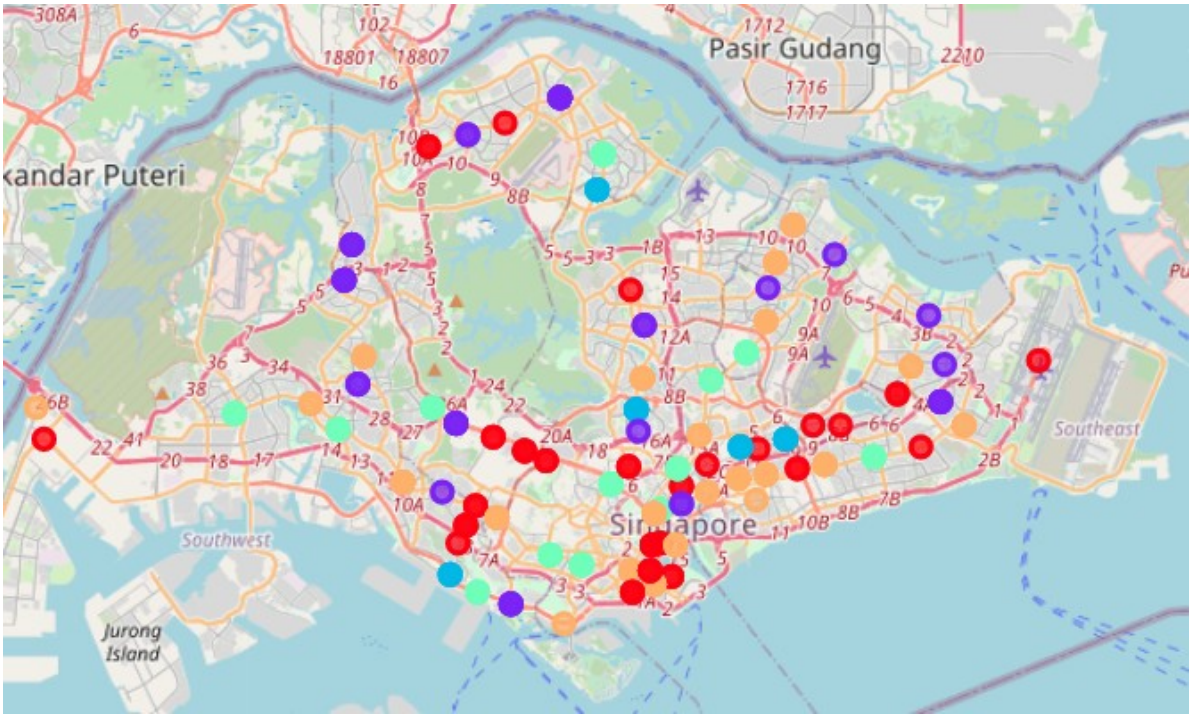
4.5.2 The standardised values were fitted using the KMeans() function from sklearn.cluster library,with number of clusters = 5. The generated cluster labels are added under a new column in *df_merge4grp*. This was then merged with *df_merge4sort* and then merged again with *df_merge4* to create *df_merge5. For* both merges, *pd.merge()* was used with *left_on = 'Station Name'* and *right_on = 'Station Name'*. Then, *df_merge5.dropna(subset = ['Station Latitude','Station Longitude'])* was implemented and resulting dataframe is stored as *df_merge6* (note: this was done after failing to plot the cluster labels onto folium map with error message pointing to NaN values for location).

4.6 For ease of further analysis, a subset dataframe, *df_merge6grp*, is created from *df_merge6* using only the columns shown in the screenshot below.

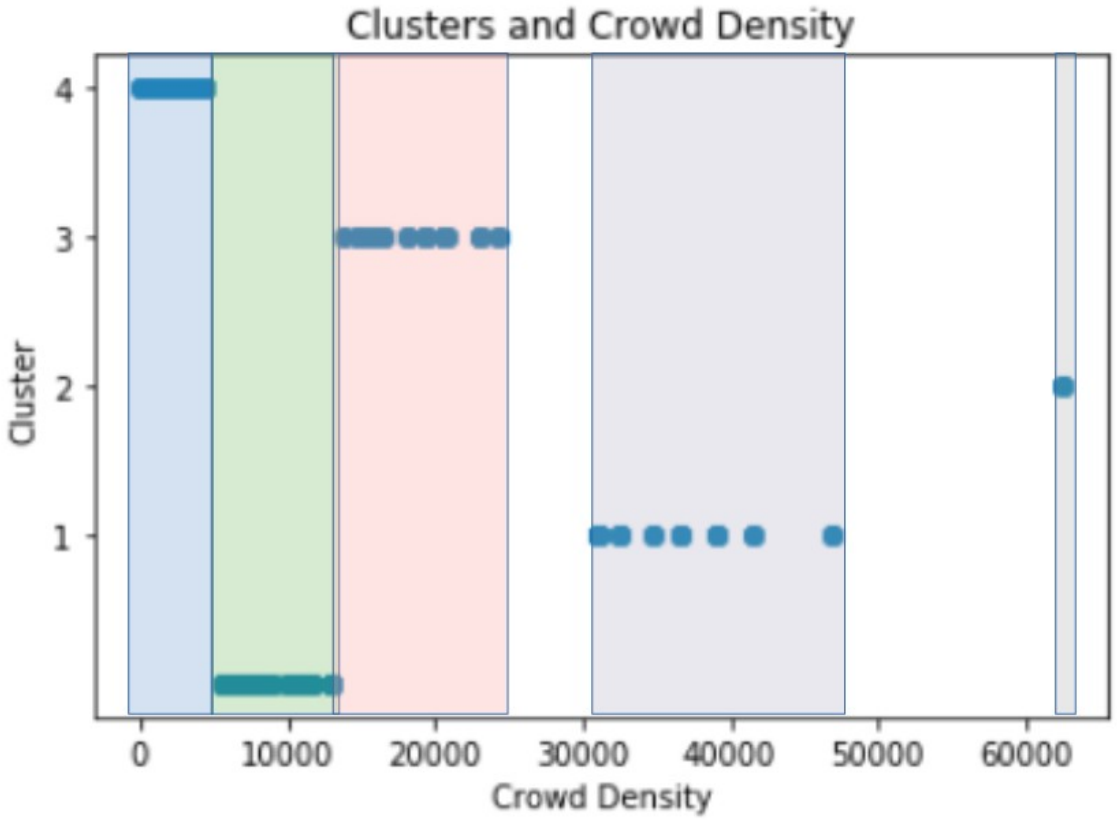
	Station Name	Station Latitude	Station Longitude	Crowd Density	Cluster Labels	1st Most Common Restaurant Type	2nd Most Common Restaurant Type	3rd Most Common Restaurant Type	4th Most Common Restaurant Type	5th Most Common Restaurant Type
0	Admiralty	1.440589	103.800990	3748.785872	5	Halal Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Asian Restaurant	Chinese Restaurant
1	Aljunied	1.316433	103.882906	1633.076467	0	Vegetarian / Vegan Restaurant	Chinese Restaurant	Seafood Restaurant	Japanese Restaurant	Comfort Food Restaurant

5 Results

5.1 All MRT stations are plotted onto Singapore map as circle markers in different colors according to cluster labels. This was achieved using folium.Map() library function. The resulting plot was as shown below.



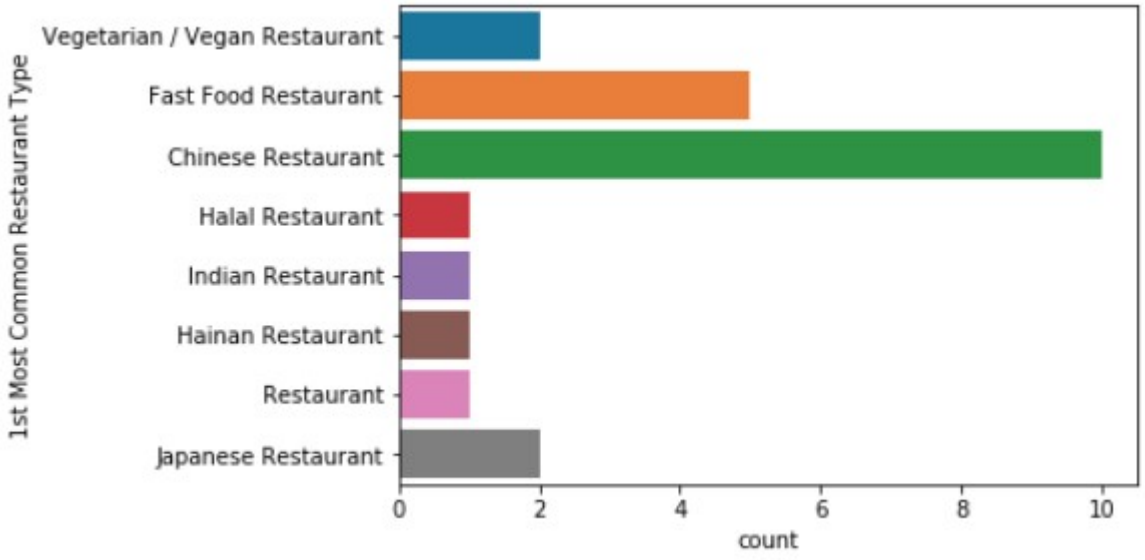
5.2 The relationship between Clusters and Crowd Density is explored. It was found that clusters are assigned according to Crowd Density.

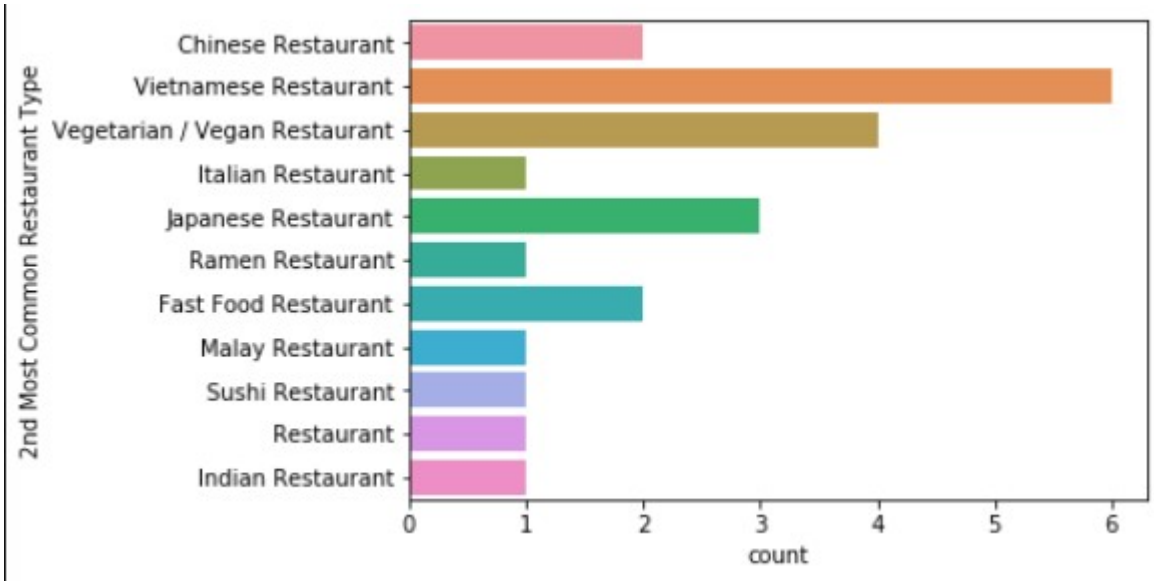


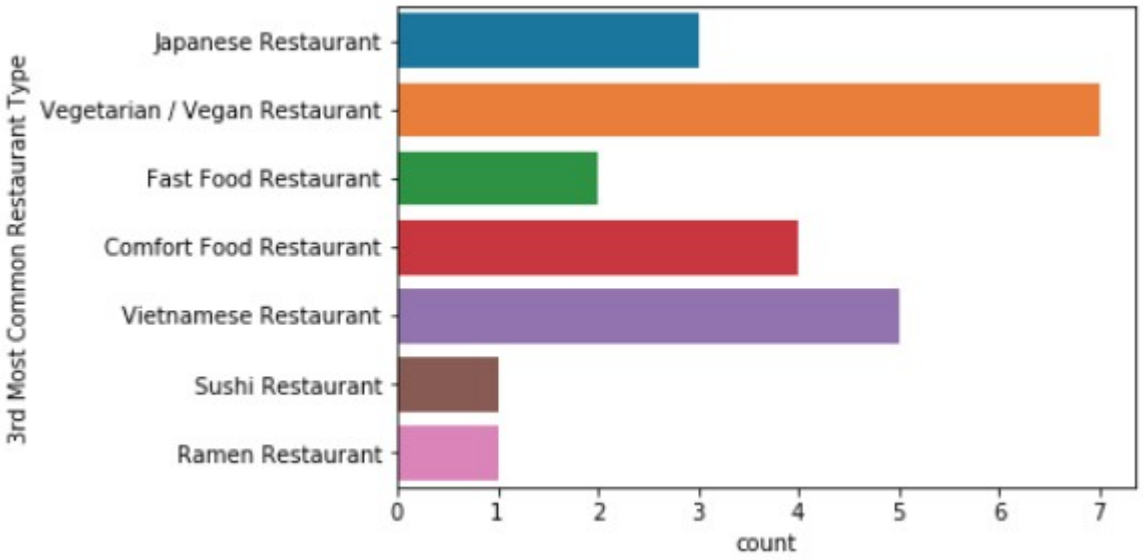
5.3 A dataframe *df_cl1* for Cluster 1 (cluster label = 0) is extracted from *df_merge6grp*. There was a total of 23 MRT Stations. Dataframe *df_cl1* was sorted by crowd density values in descending order. Finding for the top 5 highest crowd density is shown below.

	Station Name	Station Latitude	Station Longitude	Crowd Density	Cluster Labels	1st Most Common Restaurant Type	2nd Most Common Restaurant Type	3rd Most Common Restaurant Type	4th Most Common Restaurant Type	5th Most Common Restaurant Type
0	Bedok	1.323980	103.929985	12880.590134	0	Chinese Restaurant	Vegetarian / Vegan Restaurant	Fast Food Restaurant	French Restaurant	Hotpot Restaurant
1	Ang Mo Kio	1.369933	103.849558	11761.119082	0	Fast Food Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Dim Sum Restaurant
2	Buangkok	1.382870	103.893123	11548.630784	0	Fast Food Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Dim Sum Restaurant
3	Serangoon	1.350634	103.872772	11515.841584	0	Chinese Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Dim Sum Restaurant	Dumpling Restaurant
4	Novena	1.320441	103.843826	10959.910913	0	Restaurant	Ramen Restaurant	Vietnamese Restaurant	Japanese Restaurant	Comfort Food Restaurant

To find the counts of the restaurant types across Cluster 1, seaborn.countplot() is used. This is done for 1st, 2nd and 3rd Most Common Restaurant Types with the results obtained as below.



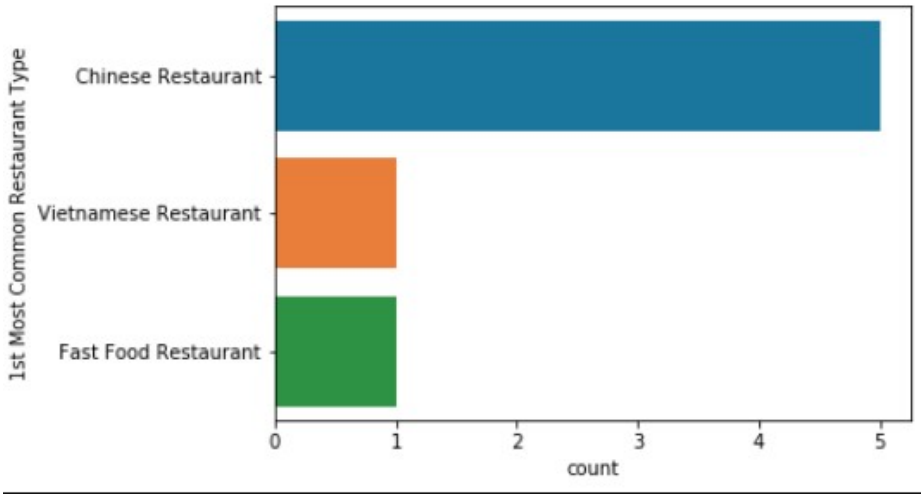


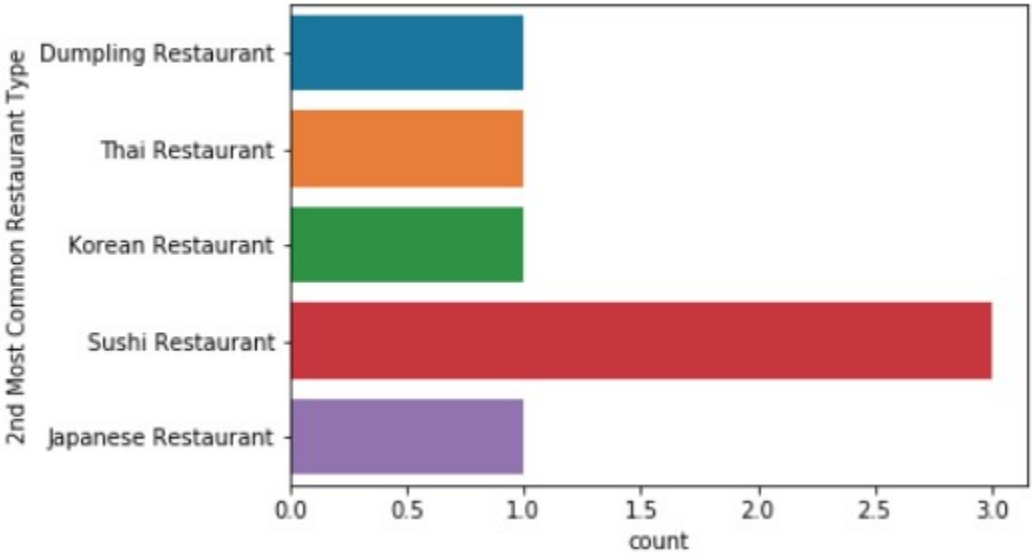


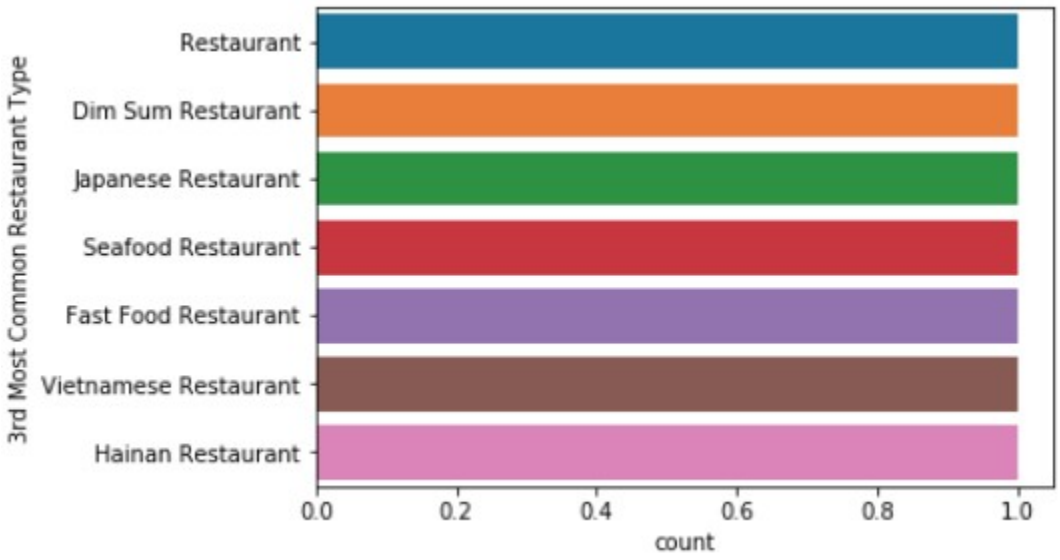
5.4 A dataframe *df_cl2* for Cluster 2 (cluster label = 1) is extracted from *df_merge6grp*. There was a total of 7 MRT Stations. Dataframe *df_cl2* was sorted by crowd density values in descending order. Finding for the top 5 highest crowd density is shown below.

	Station Name	Station Latitude	Station Longitude	Crowd Density	Cluster Labels	1st Most Common Restaurant Type	2nd Most Common Restaurant Type	3rd Most Common Restaurant Type	4th Most Common Restaurant Type	5th Most Common Restaurant Type
0	Yew Tee	1.397535	103.747405	46863.338789	1	Fast Food Restaurant	Japanese Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant
1	Chinatown	1.284224	103.845144	41518.248175	1	Chinese Restaurant	Dumpling Restaurant	Restaurant	Japanese Restaurant	Comfort Food Restaurant
2	Clementi	1.315116	103.765192	38954.689146	1	Chinese Restaurant	Thai Restaurant	Dim Sum Restaurant	Italian Restaurant	Japanese Restaurant
3	Punggol	1.404547	103.902073	36522.483940	1	Vietnamese Restaurant	Sushi Restaurant	Seafood Restaurant	Chinese Restaurant	Spanish Restaurant
4	Sengkang	1.391695	103.895485	34645.892351	1	Chinese Restaurant	Sushi Restaurant	Fast Food Restaurant	Japanese Restaurant	Comfort Food Restaurant

To find the counts of the restaurant types across Cluster 2, seaborn.countplot() is used. This is done for 1st, 2nd and 3rd Most Common Restaurant Types with the results obtained as below.



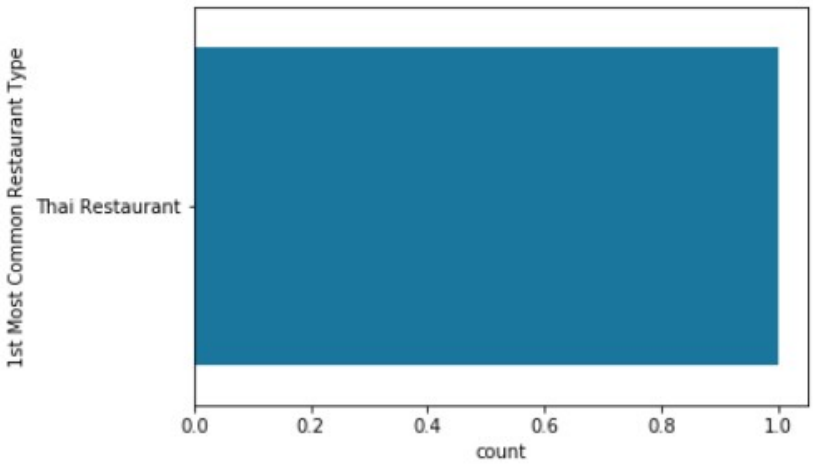


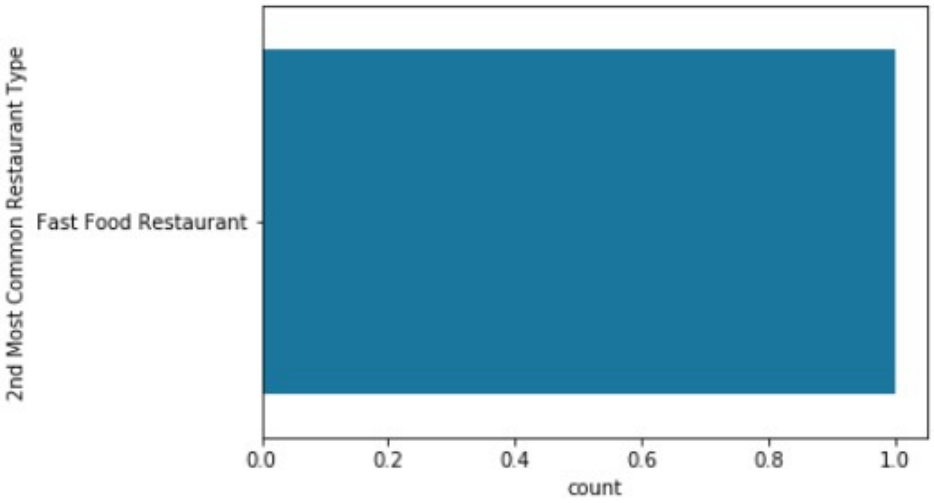


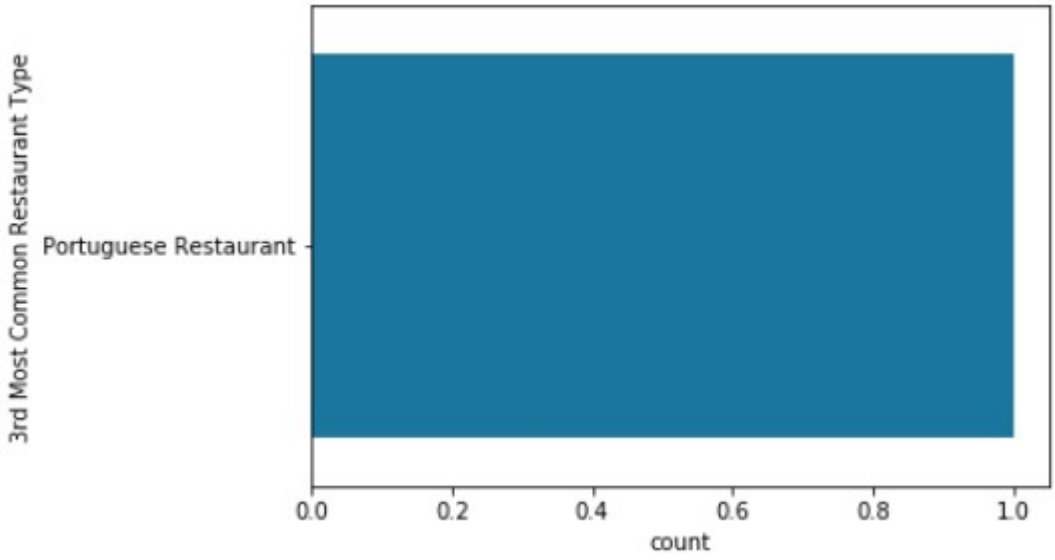
5.5 A dataframe *df_cl3* for Cluster 3 (cluster label = 2) is extracted from *df_merge6grp*. There was a total of 1 MRT Stations. Dataframe *df_cl3* was sorted by crowd density values in descending order. Finding for the top 5 highest crowd density is shown below.

	Station Name	Station Latitude	Station Longitude	Crowd Density	Cluster Labels	1st Most Common Restaurant Type	2nd Most Common Restaurant Type	3rd Most Common Restaurant Type	4th Most Common Restaurant Type	5th Most Common Restaurant Type
0	Choa Chu Kang	1.385363	103.744371	62484.451718	2	Thai Restaurant	Fast Food Restaurant	Portuguese Restaurant	Vietnamese Restaurant	Japanese Restaurant

To find the counts of the restaurant types across Cluster 3, seaborn.countplot() is used. This is done for 1st, 2nd and 3rd Most Common Restaurant Types with the results obtained as below.



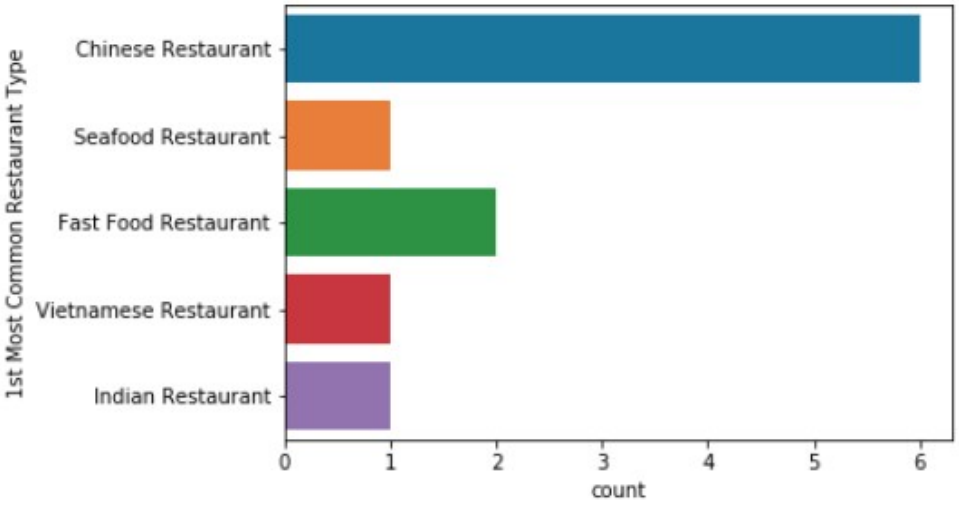

2nd Most Common Restaurant Type
Fast Food Restaurant
0.0
0.2
0.4
0.6
0.8
1.0
count

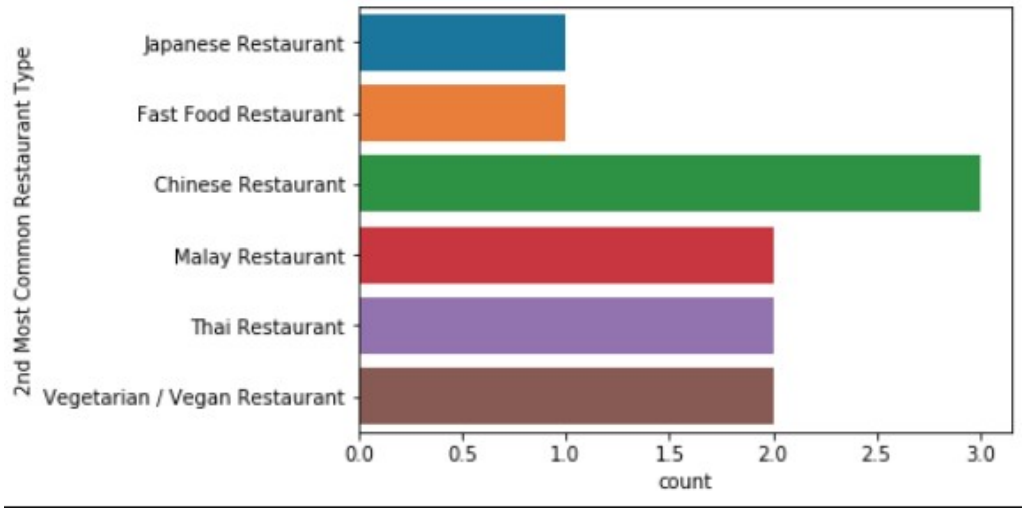

3rd Most Common Restaurant Type
Portuguese Restaurant
0.0
0.2
0.4
0.6
0.8
1.0
count

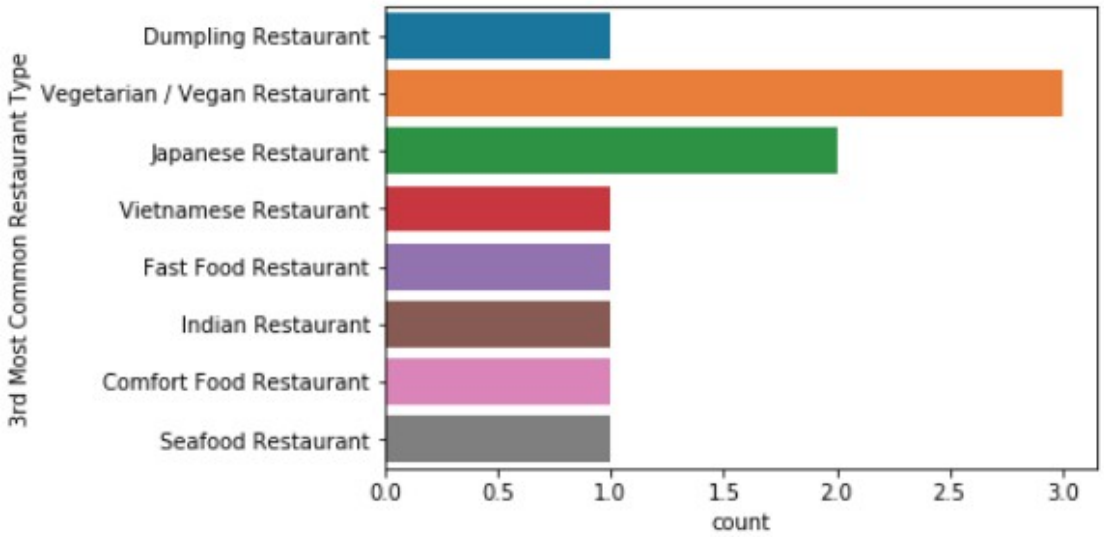
5.6 A dataframe *df_cl4* for Cluster 4 (cluster label = 3) is extracted from *df_merge6grp*. There was a total of 11 MRT Stations. Dataframe *df_cl4* was sorted by crowd density values in descending order. Finding for the top 5 highest crowd density is shown below.

	Station Name	Station Latitude	Station Longitude	Crowd Density	Cluster Labels	1st Most Common Restaurant Type	2nd Most Common Restaurant Type	3rd Most Common Restaurant Type	4th Most Common Restaurant Type	5th Most Common Restaurant Type
0	Hougang	1.371292	103.892381	24361.809045	3	Chinese Restaurant	Thai Restaurant	Fast Food Restaurant	Japanese Restaurant	Comfort Food Restaurant
1	Bishan	1.351316	103.849140	23099.737533	3	Chinese Restaurant	Japanese Restaurant	Dumpling Restaurant	Shaanxi Restaurant	Comfort Food Restaurant
2	Bukit Gombak	1.358612	103.751791	20719.676550	3	Chinese Restaurant	Malay Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Comfort Food Restaurant
3	Rochor	1.303852	103.852769	20586.419753	3	Indian Restaurant	Chinese Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Dim Sum Restaurant
4	Sembawang	1.449051	103.820046	19432.739060	3	Fast Food Restaurant	Chinese Restaurant	Japanese Restaurant	Sushi Restaurant	Comfort Food Restaurant

To find the counts of the restaurant types across Cluster 4, seaborn.countplot() is used. This is done for 1st, 2nd and 3rd Most Common Restaurant Types with the results obtained as below.



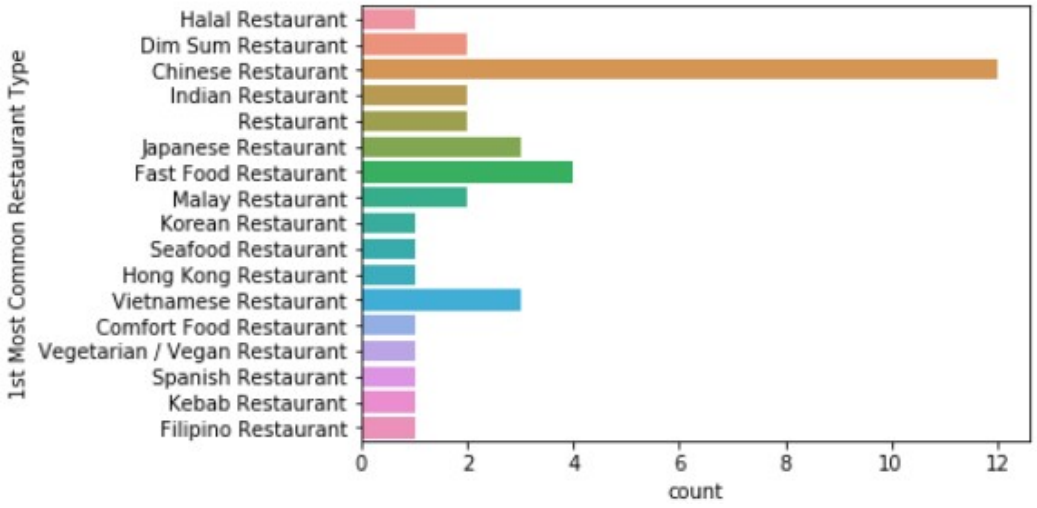


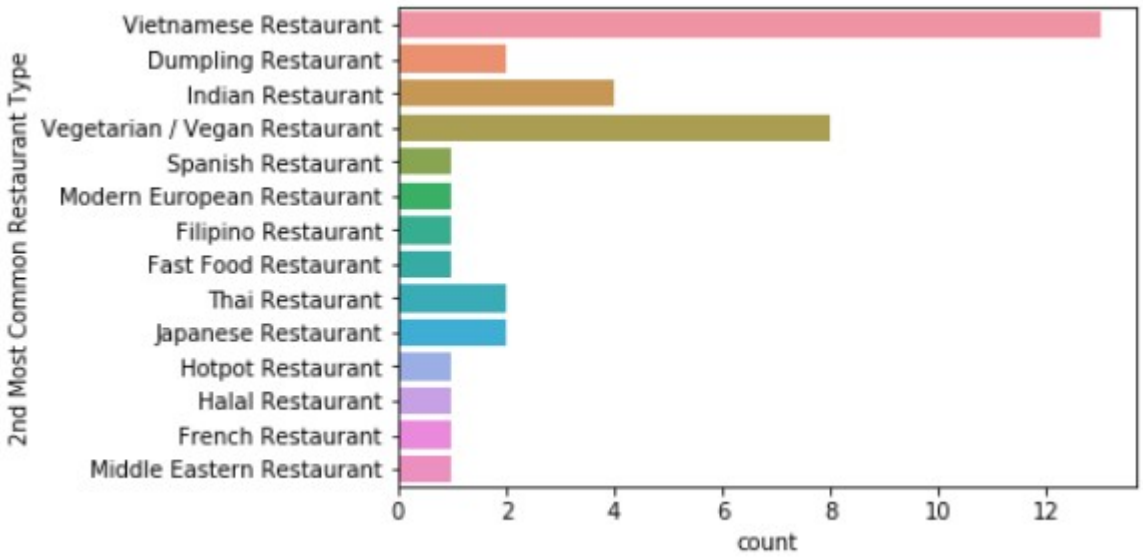


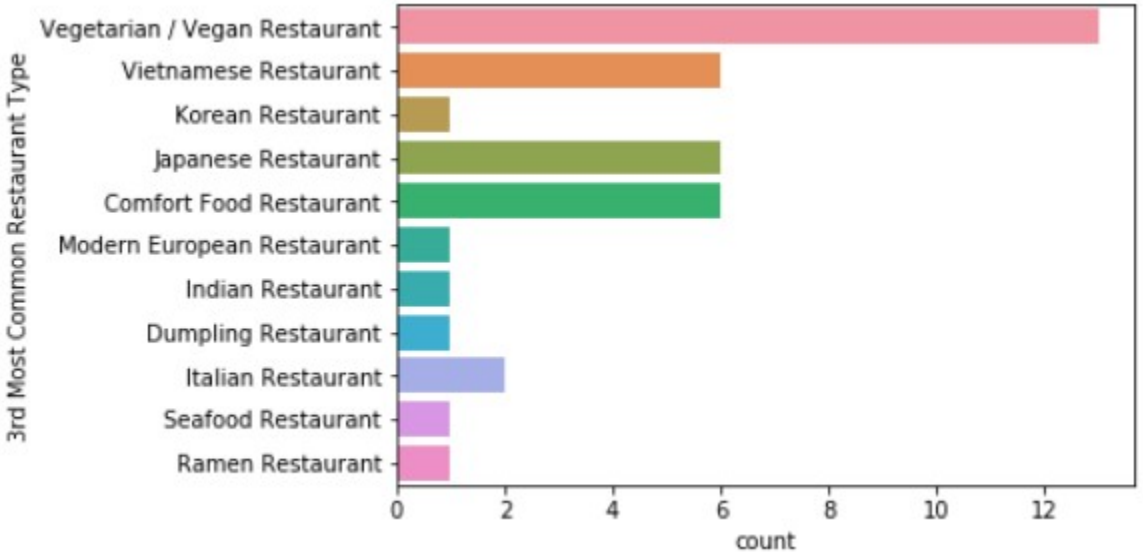
5.7 A dataframe *df_cl5* for Cluster 5 (cluster label = 4) is extracted from *df_merge6grp*. There was a total of 39 MRT Stations. Dataframe *df_cl5* was sorted by crowd density values in descending order. Finding for the top 5 highest crowd density is shown below.

	Station Name	Station Latitude	Station Longitude	Crowd Density	Cluster Labels	1st Most Common Restaurant Type	2nd Most Common Restaurant Type	3rd Most Common Restaurant Type	4th Most Common Restaurant Type	5th Most Common Restaurant Type
0	Chinese Garden	1.342353	103.732597	4444.195177	4	Chinese Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Dim Sum Restaurant	Dumpling Restaurant
1	Beauty World	1.341223	103.775794	4416.999430	4	Chinese Restaurant	Indian Restaurant	Korean Restaurant	Japanese Restaurant	Comfort Food Restaurant
2	Tampines West	1.345515	103.938437	4096.537418	4	Vegetarian / Vegan Restaurant	Halal Restaurant	Vietnamese Restaurant	Comfort Food Restaurant	Dim Sum Restaurant
3	Expo	1.334550	103.961548	4096.537418	4	Chinese Restaurant	Filipino Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant	Dim Sum Restaurant
4	Tampines	1.355150	103.943010	4096.537418	4	Chinese Restaurant	Thai Restaurant	Japanese Restaurant	Comfort Food Restaurant	Dim Sum Restaurant

To find the counts of the restaurant types across Cluster 5, seaborn.countplot() is used. This is done for 1st, 2nd and 3rd Most Common Restaurant Types with the results obtained as below.







6 Discussion

6.1 When acquiring data from Wikipedia, it was noted that certain MRT stations have very low crowd density values. This can be explained that that the population demographics provided in the source refers to residential population, and may not accurately reflect the actual daily crowd size that experienced at the MRT stations. For example, crowd density for Jurong East MRT Station could be much higher than the 2222 pax/km^2. This is because there are commercial and industrial offices located in Jurong East town. As such, we can expect the crowd size to be more than just the residents of the town. Another glaring example is City Hall MRT Station, which has a crowd density of 89 pax/km^2. It is a tourist attraction area and a commercial hub of the country. So naturally, it is not expected to have

many residential dwellings. One should also be cognizant that the reported figures in the Wikipedia source do not include approximately 1.6 million non-permanent residents of Singapore.

6.2 Basing on the overall counts of restaurant types, Chinese Restaurant has the most number. This is not surprising as majority of Singapore's population are ethnic Chinese (about 76.2% https://en.wikipedia.org/wiki/Demographics_of_Singapore#:~:text=Singapore%20is%20a%20multiracial%20and,the%20majority%20of%20the%20population) . Hence, opening a Chinese restaurant could be a straightforward choice. This is supported by the observation that Chinese Restaurant is the 1st most commont restaurant type for all clusters except Cluster 3. In these clusters, the competition may be less keen for other restaurant type. Likewise, one may consider avoiding Cluster 2 areas for opening a fast food restaurant, which has the second highest count.

6.3 It is interesting to note that Vietnamese Restaurant is not on the top 10 list for overall counts, but it appears as 2nd most common restaurant type in Cluster 1 and 5. This could indicate an emerging market for Vietnamese cuisine . In this case, one could consider areas in Cluster 1 and 5 for opening Vietnamese restaurant to tap on the rising demand there.

6.4 Some restaurant types, obtained as "Venue Categories" from Foursquare, could be misleading or unclear For example, Dim Sum Restaurant and Hong Kong Cuisine Restaurant are variants of Chinese Restaurant. Also, it is unclear what cuisine "Restaurant" and "Asian Restaurant" actually refer to. For the latter, we could probably take it to mean asian fusion.

7 Conclusion

7.1 In this Capstone project, we have extracted information from the internet, get coordinates for every MRT stations, and used Foursquare API to get venues surrounding every MRT station. Data is wrangled, correctly formatted, and normalised before further data analysis was done. Exploratory analysis and visualisations are done to gain a better understanding of the data. Finally, machine learning algorithms are used to cluster data. Clustering appeared to be in accordance to Crowd Density. The clustering results could serve as a reference to help restauranteurs decide what type of restaurants to open where.