The Battle of the Neighborhoods - IBM Data Science Capstone Report

Joseph Radomski

August 7, 2019

1 Introduction

1.1 Background

The island city nation of Singapore is like no other. It is a concoction of many different people of many different ancestries. The country alone has four commonly used languages, namely English, Mandarin, Tamil and Malay. For such a small island filled with such variety, a majority working long hours with little time to cook, eating out is necessary. The range of cuisine is even more necessary as most Singaporeans tend to be adventurous and regularly look for new cuisine. Due to the immense amounts of trade and business within the city, there is also a large number of temporary residents. Along with the many tourists, these are people always looking for a quick and different food choice. Therefore, it is of high interest to determine where you should go in Singapore when looking for a specific cuisine.

1.2 The Question

From locational data, as well as ratings data, we want to determine an answer to the question "When looking for a specific cuisine choice, which area of Singapore should you go to?".

1.3 Interest

As discussed within the Background, there are numerous types of individuals who would be interested in the answer to this question.

- The residents of Singapore are often adventurous, however often lack time due to lengthy working hours, being able to answer this question will be of use to many of them.
- Singapore is famed for its variety, and many tourists and temporary residents coming there know this, however they may not often know where to go if they look for a specific cuisine type. This is a market of individuals this question can be aimed at.

- Aside from individuals, businesses often look to hold company events, and often a meal at a restaurant, they are another target audience for this question.
- Restaurant chains or individuals looking to open a specific style of restaurant will also be interested in the answering of this question.

Along with the answering of this specific question, we can alternatively answer the reverse from this,”We find ourselves within a specific area of Singapore, restaurants of what cuisine type should we seek?”.

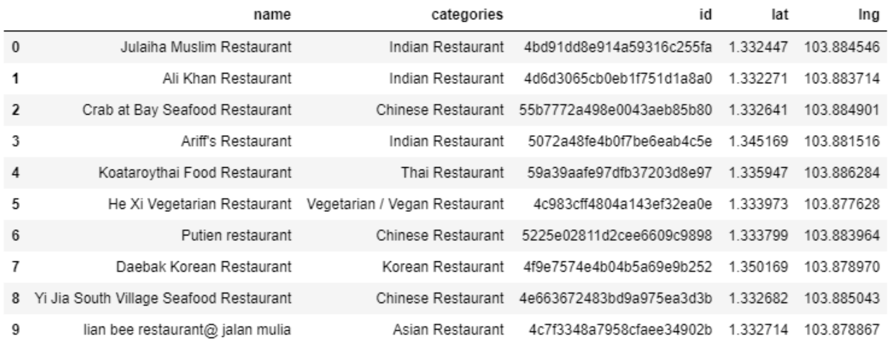
2 Data Section

2.1 Data Sources

I currently believe that the question can be answered using simply only the Foursquare API. The data will provide an insight into the geographic location, cuisine categorisation and ratings of all the restaurants within Singapore. There are ways to ensure the ratings of the restaurants are credible, these include filtering based on the number of likes a rating has and filtering based on the profile of the rating user.

2.2 Data Cleaning/Preprocessing

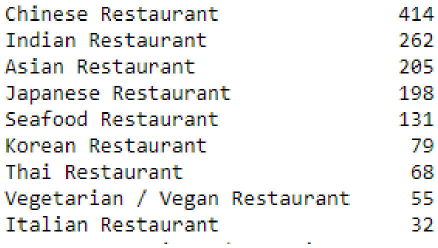
With being limited to only searches for each query to Foursquare API, I needed to retrieve the data I needed differently. To do this, I iteratively changed the geographical location i was centering my search from and then conducted a query from there. I covered the entire island this way. I built each search into a dataframe, then appended these dataframes together into a larger dataframe. We required some data cleaning after this. Removing duplicate entries to the larger data from (this done by dropping rows with duplicate id column entries). I also found there to be entries into the database from Malaysia, with it being picked up in the radius. I was able to drop all rows that didn’t contain ’SG’ within the country code column. There also existed rows where the category of restaurant (eg. cuisine type) was unknown. This was a very small percentage of cases, and hence these entries were removed.
We found there also to be many redundant columns within the dataframe, in fact the only columns kept are name, categories, id, latitude and longitude. For a few reasons, we eradicate all entries where there exists less than a total of thirty entries of a specific cuisines type. The reasons include the strength of the results we would obtain, supported by difficulties to render the original amount of data. There are also restaurants where the cuisines type is entered as ’Restaurant’, these entries are also removed. We also remove ’Coffee Shop’ and ’Food Court’ entries, food courts generally have a wide range of cuisines anyway and like coffee shops, wouldn’t be considered as restaurants.



	name	categories	id	lat	lng
0	Julaiha Muslim Restaurant	Indian Restaurant	4bd91dd8e914a59316c255fa	1.332447	103.884546
1	Ali Khan Restaurant	Indian Restaurant	4d6d3065cb0eb1f751d1a8a0	1.332271	103.883714
2	Crab at Bay Seafood Restaurant	Chinese Restaurant	55b7772a498e0043aeb85b80	1.332641	103.884901
3	Ariff's Restaurant	Indian Restaurant	5072a48fe4b0f7be6eab4c5e	1.345169	103.881516
4	Koataroythai Food Restaurant	Thai Restaurant	59a39aafe97dfb37203d8e97	1.335947	103.886284
5	He Xi Vegetarian Restaurant	Vegetarian / Vegan Restaurant	4c983cff4804a143ef32ea0e	1.333973	103.877628
6	Putien restaurant	Chinese Restaurant	5225e02811d2cee6609c9898	1.333799	103.883964
7	Daebak Korean Restaurant	Korean Restaurant	4f9e7574e4b04b5a69e9b252	1.350169	103.878970
8	Yi Jia South Village Seafood Restaurant	Chinese Restaurant	4e663672483bd9a975ea3d3b	1.332682	103.885043
9	lian bee restaurant@ jalan mulia	Asian Restaurant	4c7f3348a7958cfaee34902b	1.332714	103.878867

Figure 1: Head of Dataframe

Here we find the head of the dataframe. In total, there are 1444 Restaurants entered.



```
Chinese Restaurant                 414
Indian Restaurant                  262
Asian Restaurant                   205
Japanese Restaurant                198
Seafood Restaurant                 131
Korean Restaurant                   79
Thai Restaurant                     68
Vegetarian / Vegan Restaurant       55
Italian Restaurant                  32
```

Figure 2: Number of entries of each cuisine type

Using a value count, we then find the number of restaurants of each cuisine type. We note there being no cuisines listed with less than 30 entries, as required.

2.3 Folium Visualisation

Using the Folium library, we can then visualise the data points using the geographical location of each entry. Initially, we plot every data point to ensure data is gathered from throughout the city.

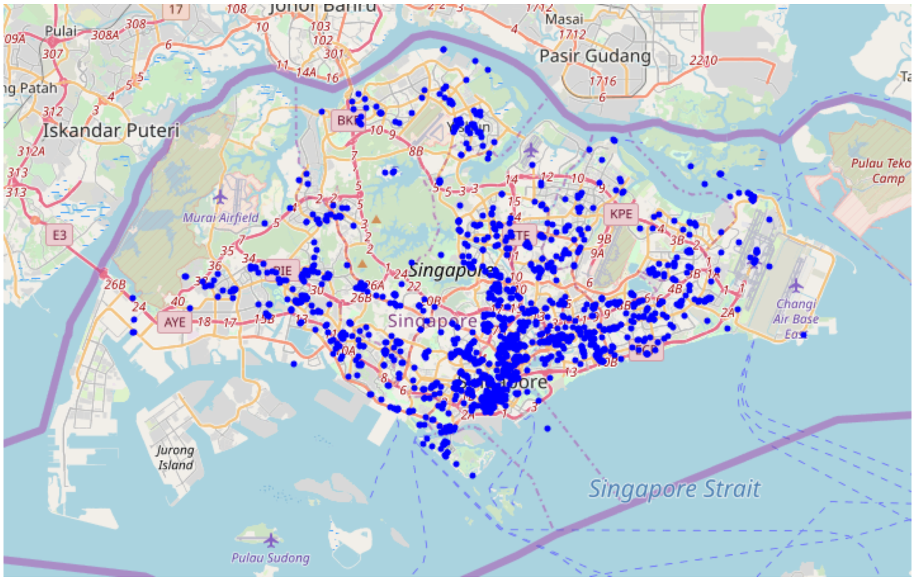


Figure 3: Map showing entries

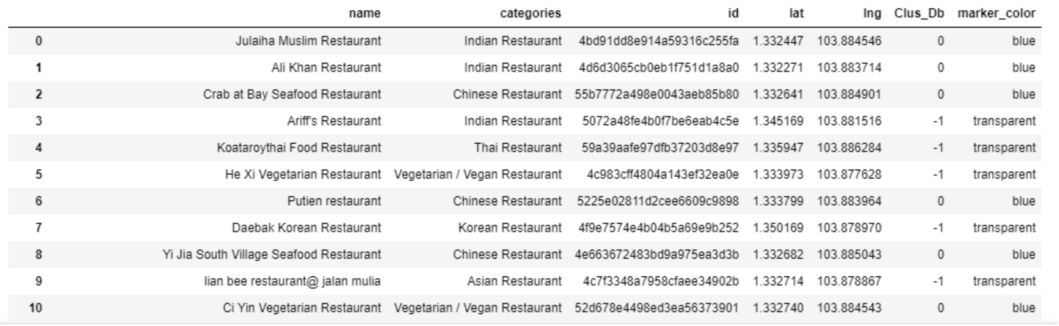
We observe the data to appear correct, where there is a higher density of data points around the Marina Bay area, and there being very few around the military and park areas.

3 Data Modelling

3.1 DBSCAN

To establish which areas to visit when looking for a specific style of cuisine, we need to use density-based clustering techniques. We will begin by using the entire dataframe and conducting a dbscan, where we find the areas of Singapore with a high density of restaurants.

Using dbscan, we assign the entries to a cluster number, or if considered an outlier (most here are), then we assign them to '-1'. We then assign the clusters to colours and outliers to transparent in preparation for the folium visualisation.



	name	categories	id	lat	lng	Clus_Db	marker_color
0	Julaiha Muslim Restaurant	Indian Restaurant	4bd91dd8e914a59316c255fa	1.332447	103.884546	0	blue
1	Ali Khan Restaurant	Indian Restaurant	4d6d3065cb0eb1f751d1a8a0	1.332271	103.883714	0	blue
2	Crab at Bay Seafood Restaurant	Chinese Restaurant	55b7772a498e0043aeb85b80	1.332641	103.884901	0	blue
3	Ariff's Restaurant	Indian Restaurant	5072a48fe4b0f7be6eab4c5e	1.345169	103.881516	-1	transparent
4	Koataroythai Food Restaurant	Thai Restaurant	59a39aafe97dfb37203d8e97	1.335947	103.886284	-1	transparent
5	He Xi Vegetarian Restaurant	Vegetarian / Vegan Restaurant	4c983cff4804a143ef32ea0e	1.333973	103.877628	-1	transparent
6	Putien restaurant	Chinese Restaurant	5225e02811d2cee6609c9898	1.333799	103.883964	0	blue
7	Daebak Korean Restaurant	Korean Restaurant	4f9e7574e4b04b5a69e9b252	1.350169	103.878970	-1	transparent
8	Yi Jia South Village Seafood Restaurant	Chinese Restaurant	4e663672483bd9a975ea3d3b	1.332682	103.885043	0	blue
9	lian bee restaurant@ jalan mulia	Asian Restaurant	4c7f3348a7958cfaee34902b	1.332714	103.878867	-1	transparent
10	Ci Yin Vegetarian Restaurant	Vegetarian / Vegan Restaurant	52d678e4498ed3ea56373901	1.332740	103.884543	0	blue

Figure 4: Dataframe showing assignment

Then using this dataframe, we can build a visualisation using folium. Here, we can observe on the map the areas in which there are high densities of restaurants within Singapore. We set the parameters for the dbscan such that a cluster forms if there exists a circular area of radius 350m contains a minimum of 10 entries.

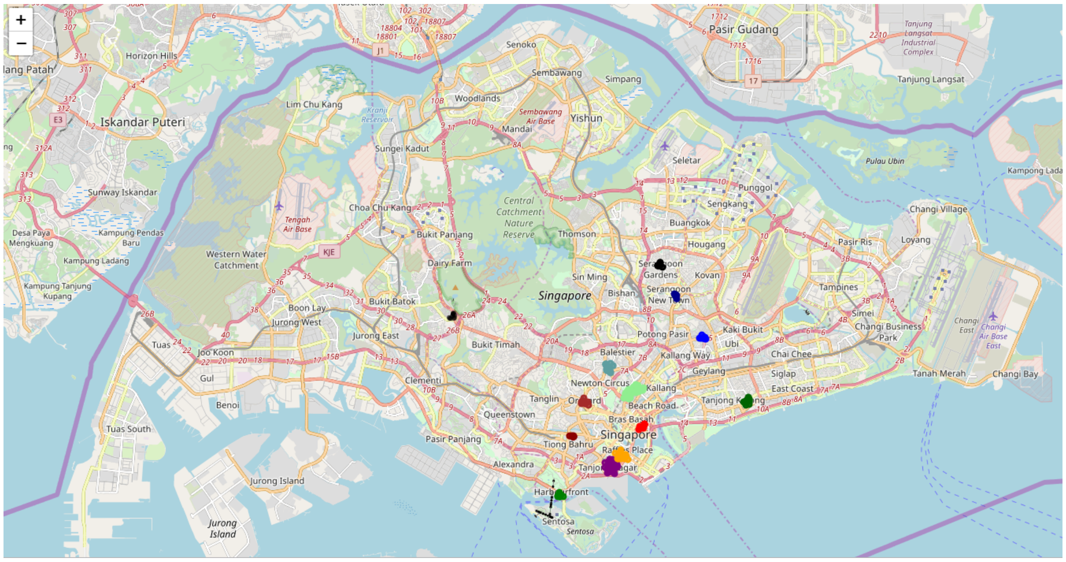


Figure 5: Clusters of all restaurants

3.2 DBSCAN category dependent

Within this section, conduct a dbscan for each category of restaurant, the parameters slightly differing to produce effective results and then combine the clusters into one combined visualisation. However, we first observe for Chinese restaurants only.

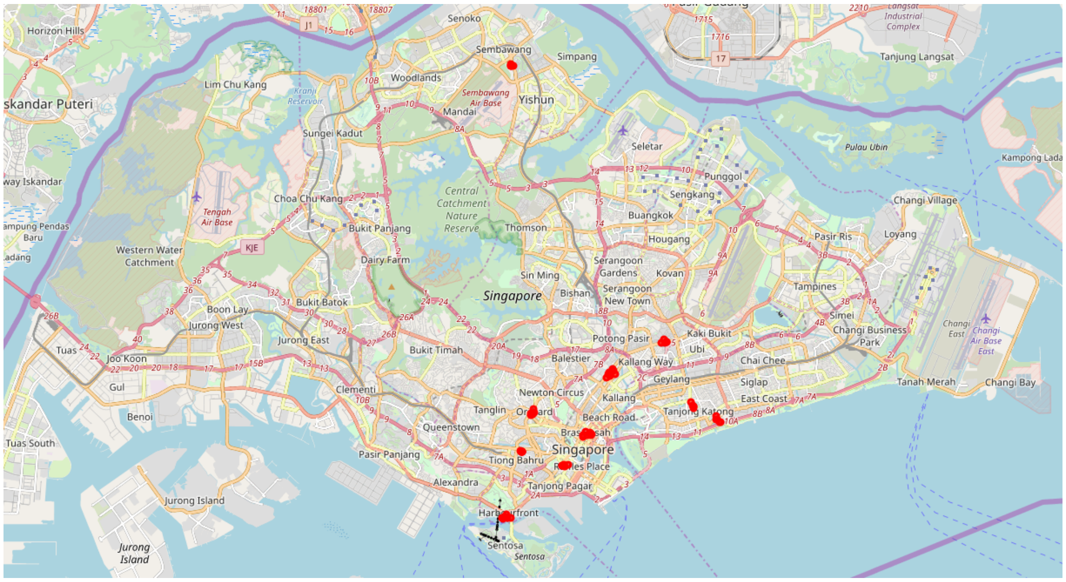


Figure 6: Density clustering of Chinese restaurants

We then produce a visualisation for density clusters of the top 3 categories. These categories being Chinese, Indian and Asian food.

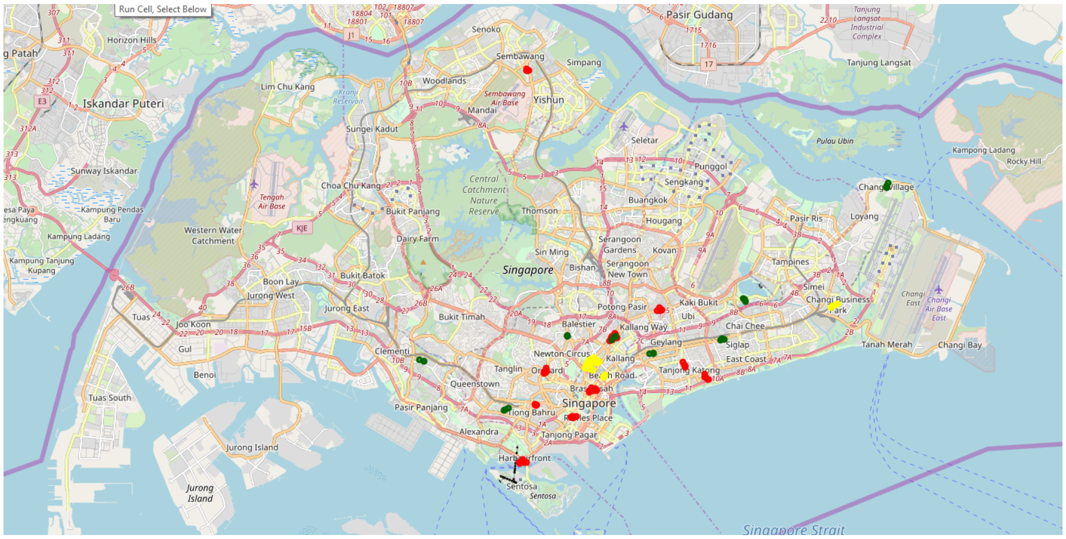


Figure 7: Density clustering of Top 3 categories

Finally, having conducted a dbscan for every included category of restaurant, we produce mapped visualisation of all the different density clusters.

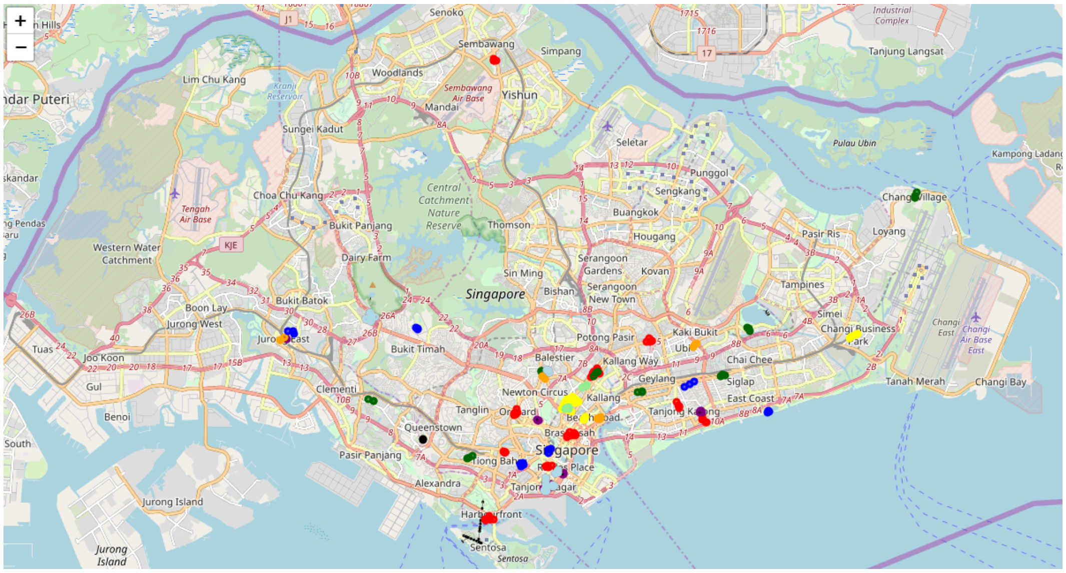
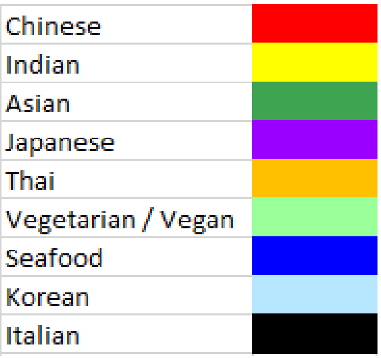


Figure 8: Density clustering for all categories

With the following colourscheme corresponding to the visualisations.



Chinese	🟥
Indian	🟨
Asian	🟩
Japanese	🟪
Thai	🟧
Vegetarian / Vegan	🟩
Seafood	🟦
Korean	🟦
Italian	⬛

Figure 9: Colourscheme of above visualisation

3.3 Conclusion

Within this study, I have identified the locations where high densities of specific restaurant types operate and have therefore given an answer to the proposed question. Using the dbscan density-based clustering technique, I have clustered restaurants of specific types into clusters of high density, indicating the areas to visit if looking for a specific type of restaurant. Having build visualisations of these clusters using Folium maps, individuals with interest (see introduction) can quickly and precisely observe the areas they may want to visit to find a restaurant.

3.4 Further endeavors

To take this further, we could take into account the ratings of the restaurants. This may allow us to create a ranking for these clusters, or include or remove clusters in the findings. We may also further classification methods in which similar styles of restaurant (Korean and Japanese for example) may be grouped together and as a result form new clusters and differ the areas recommended to a client.