AN ANALYSIS OF ETHICAL CLUSTERING IN SINGAPORE

An Restaurant Approach

Archer

3 Jan 2020

Zhu Archer

1 Introduction

This paper presents an analysis of neighborhood clustering in Singapore and further apply the conclusion in choosing a best place for opening a themed restaurant.

Singapore is a small country (of only 1,067 km^2) with diversified ethnic groups such as Chinese(76.2%), Malays(15.0%), Indians(7.4%) as well as Indonesian, Peranakan and Western. And religious dietary strictures do exist, Muslims do not eat pork and Hindus do not eat beef and there is a significant group of vegetarians/vegans. It is nature to think of understanding Singapore ethic group clustering through restaurants around each neighborhood.

Furthermore, it's also possible to find the best location of choosing the right neighborhood using the result of above analysis.

Thus, this report would consist of mainly two parts:

1) A general analysis of clustering through machine learning with restaurant data;
2) an application of above neighborhood clustering result, in specific, finding the right neighborhood for opening a Japanese restaurant;

Business problem:

The second part of report is an actual problem: finding the best location of opening a Japanese restaurant. Generally Japanese restaurants are high-ended, targeting at the wealthy. Thus, an intuitive answer is downtown, but we will see if we can get same answer through quantitative analysis.

Normal approach of tackling this problem is starting by analyzing target customer, finding relevant characters of a place with relevant customers. But we will apply a classic assumption of efficient market, where the current existence of market represents the reasonable solution. In short, everything real is rational. This assumption should work well in Singapore since it's a highly developed country with high average education and information flows fast and fluent here.

As a result, the problem can be further transferred into: finding the neighborhood which has a most significant feature of Japanese restaurant.

Target audience of this report:

- Officials who wants to understand geographic clustering of neighborhood in a quantitative way may find this report useful.
- Investors who wants to open a new restaurant may find this report useful in providing methodology in choosing location of opening a new restaurant.
- Market analyst may find this report useful in understanding the link between

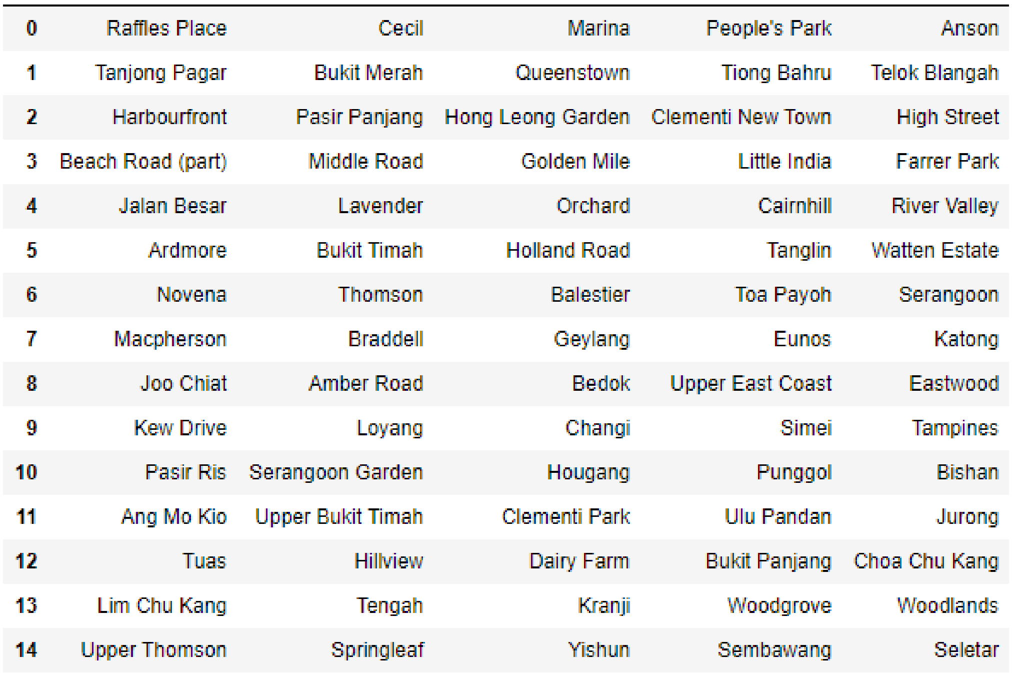
2 Data

2.1 Data Source

Following data would be used for solving the problem.

Data1: Neighborhood Names

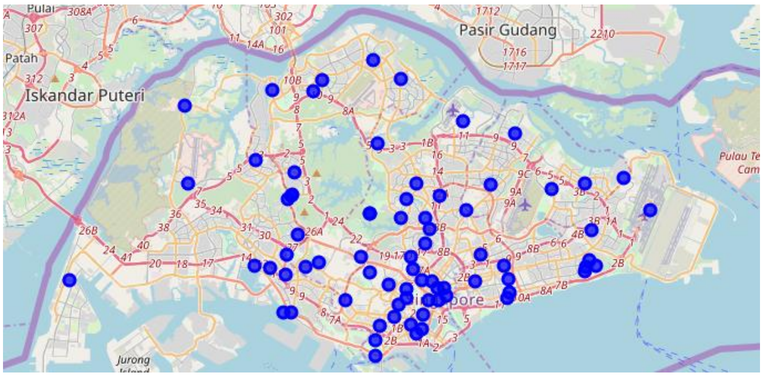
Since the project focuses solely on Singapore, we need neighborhood information about Singapore, which is provided on Wiki[1].
It consists of 75 neighborhoods.



0	Raffles Place	Cecil	Marina	People's Park	Anson
1	Tanjong Pagar	Bukit Merah	Queenstown	Tiong Bahru	Telok Blangah
2	Harbourfront	Pasir Panjang	Hong Leong Garden	Clementi New Town	High Street
3	Beach Road (part)	Middle Road	Golden Mile	Little India	Farrer Park
4	Jalan Besar	Lavender	Orchard	Cairnhill	River Valley
5	Ardmore	Bukit Timah	Holland Road	Tanglin	Watten Estate
6	Novena	Thomson	Balestier	Toa Payoh	Serangoon
7	Macpherson	Braddell	Geylang	Eunos	Katong
8	Joo Chiat	Amber Road	Bedok	Upper East Coast	Eastwood
9	Kew Drive	Loyang	Changi	Simei	Tampines
10	Pasir Ris	Serangoon Garden	Hougang	Punggol	Bishan
11	Ang Mo Kio	Upper Bukit Timah	Clementi Park	Ulu Pandan	Jurong
12	Tuas	Hillview	Dairy Farm	Bukit Panjang	Choa Chu Kang
13	Lim Chu Kang	Tengah	Kranji	Woodgrove	Woodlands
14	Upper Thomson	Springleaf	Yishun	Sembawang	Seletar

Data2: Neighborhood Locations

In order to further locate the neighborhoods on map, we need location data of each neighborhood, which is provided by Geocoder. Below is plotted neighborhoods. Downtown area of Singapore is located at southern part where has a higher density of neighborhoods.



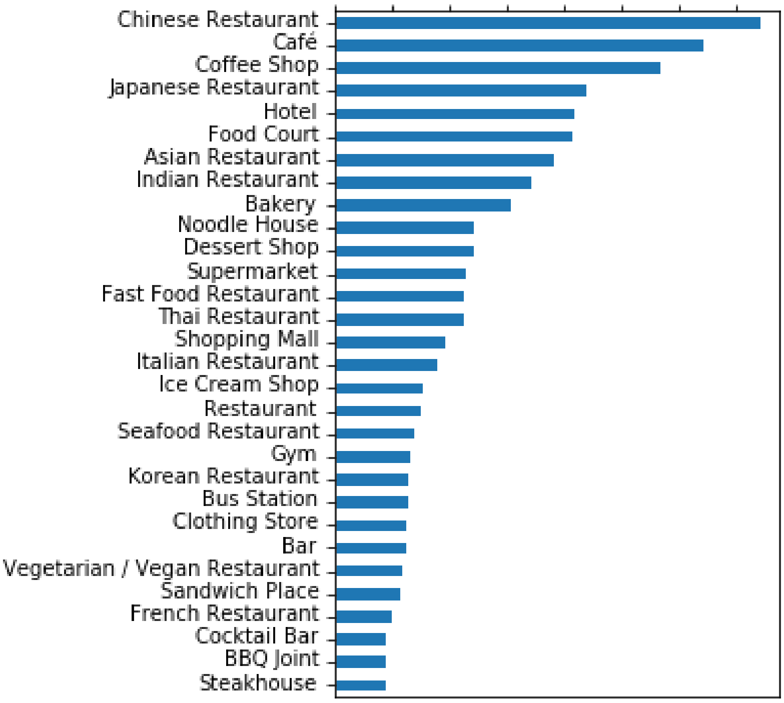
[1] Postal codes in Singapore, Wiki, *https://en.wikipedia.org/wiki/Postal_codes_in_Singapore*

Data3: Venues around Neighborhood

Most importantly, neighborhood venue information is needed. Since Singapore is small, a radius of 5 km within each neighborhood is applied. And relevant data is fetched from Foursquare.

A total of 2,466 venues are fetched for 75 neighborhoods. Foursquare provided venue category by default, which distinguished different types of restaurants. Singapore has 268 different venue categories in total.

Below are the Top 30 frequently shown venue types.



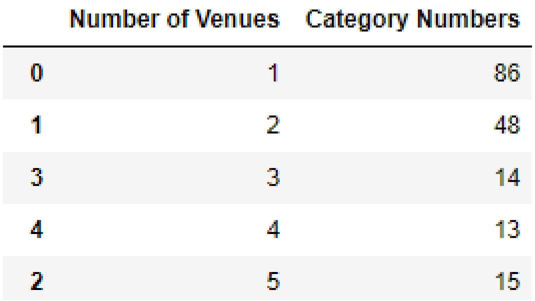
Chinese restaurant ranks first, 148 out of 2,466 venues of all kinds. The result is in consistence with stats that Chinese makes up the majority of population.

Singaporean have a huge fever for Coffee, which also aligns with the rankings of number 2 Café and 3 Coffee Shop.

2.2 Data Cleansing:

Our venue data has two major problems thus require data cleansing,

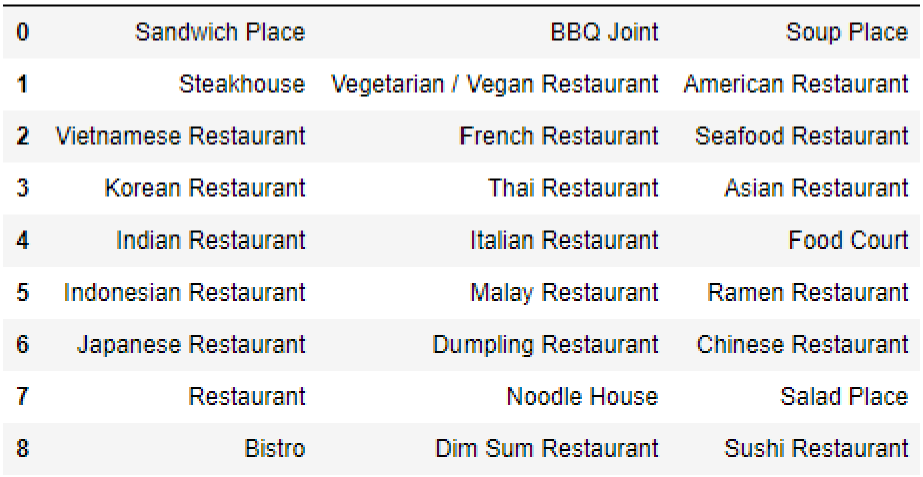
1. 86 Category has only 1 corresponding venue and 48 has only 2 corresponding venues. These venue category feature will serve no purpose in clustering and may lead to over fitting in certain machine learning algorithms.
   Our solution is to delete venue categories with little sample. By applying a threshold level of 10, we have 1,888 samples left(excluded 578 venues).



	Number of Venues	Category Numbers
0	1	86
1	2	48
3	3	14
4	4	13
2	5	15

2. This report aims to analyze the clustering of ethical neighborhoods through restaurant data, so we need to use only food relevant venue categories.
   We split food relevant venues into 4 kinds using category keywords:
   - Restaurant: all kind of restaurants including food court, Indian restaurant, etc.
   - FastFood: only contain fast foods such as pizza, fried chicken, etc.
   - Snack: bakery, dessert, etc.
   - Drink: Café, Bar, etc.

Below are categories of Restaurant.



0	Sandwich Place	BBQ Joint	Soup Place
1	Steakhouse	Vegetarian / Vegan Restaurant	American Restaurant
2	Vietnamese Restaurant	French Restaurant	Seafood Restaurant
3	Korean Restaurant	Thai Restaurant	Asian Restaurant
4	Indian Restaurant	Italian Restaurant	Food Court
5	Indonesian Restaurant	Malay Restaurant	Ramen Restaurant
6	Japanese Restaurant	Dumpling Restaurant	Chinese Restaurant
7	Restaurant	Noodle House	Salad Place
8	Bistro	Dim Sum Restaurant	Sushi Restaurant

After data cleansing, we have 1460 sample venues and a total of 41 unique venue categories left.

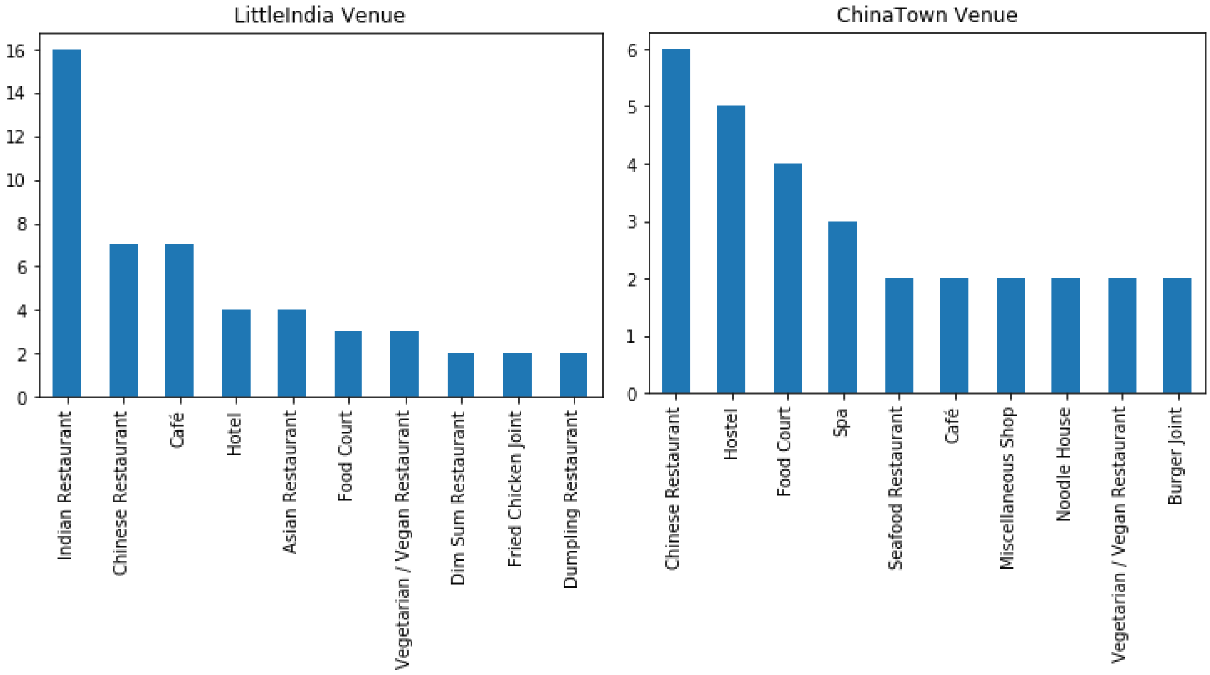
3 Methodology

Business Understanding:
The main goal is to get the optimum location of neighborhood for opening a Japanese restaurant, where, by our assumption, should already have many Japanese restaurants.

3.1 Exploratory Data Analysis:

In this part we will first look at whether venue feature can reflect ethical group in certain traditional ethical regions. There are two perfect samples in Singapore: China Town and Little India. And next, we will look at restaurant venues categories and find if any pattern can be revealed.

3.1.1 Neighborhood Venues around Little India and China Town:



Above charts showed obvious correlation between regional trending ethical group and venue category:

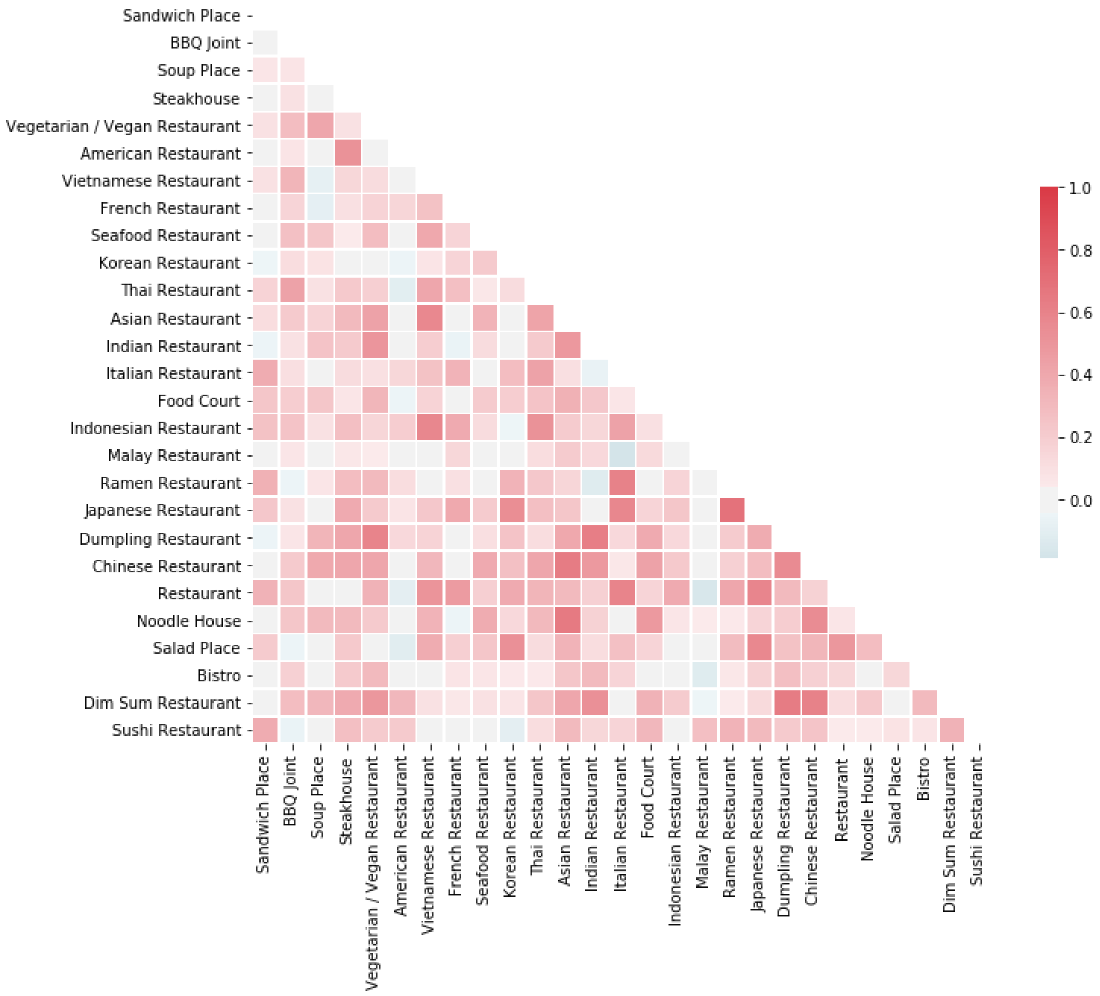
1. The biggest group of venue category is Indian food, which clearly represents regional character. The restaurant venue style for China Town may seem not so obvious, but it's not true. Actually, there's a huge food court serving only Chinese food has a capability of over 500 people, but it's not considered as Chinese restaurant in category.
2. There are some Chinese restaurants in Little India such as Chinese restaurant and Dim Sum restaurants. Although it may seem counter intuitive, but it's also true. Contrary to stereotypes, Little India is not solely an Indian neighborhood. Located in the neighborhood alongside shops that cater predominantly to the Indians are Chinese clan associations, places of worship of different religions, and a variety of different business

ranging from electrical supplies, hardware, second-hand goods alongside traditional spice grinders and grocers.[2]

From above analysis, we confirmed that venue categories features of the two regions can reflect local character.

3.1.2 Location Patterns of Chinese/ Indian/ Malay Restaurants:

The intuition of analyzing these 3 groups are that they constitutes the majority of Singaporean.



Chinese Restaurant:

- sample size: 148
- Since there are too many Chinese restaurants in Singapore and the biggest portion of Singaporean ethics group is Chinese, it seems that Chinese restaurant has a relatively mild correlation with everything else.

Indian Restaurant:

- sample size: 68

[2] Little India, Singapore, Wiki, *https://en.wikipedia.org/wiki/Little_India,_Singapore*

- It has a low correlation with **Malay/ Japanese/ French/ Korean/ Italian/ Sandwich Restaurant** of around zero, indicating that relevant ethics group lives in distance with Indian group or that target consumers of these restaurants are in no overlapping with it.
- It also has a relevant high correlation with Chinese restaurants such as Chinese/ Dim Sum/ Dumpling Restaurants, which, as mentioned before is mainly because there are too many Chinese restaurants.

Malay Restaurant:
- sample size: 12
- Astonishingly it has no significant correlation with any other restaurants.
- This low correlation may derive form data error, since basically every food court has a Malay food window serving local foods. But these food courts are not counted as Malay restaurants.

We explored data sets and found correlations between restaurants and ethical groups as well as some other insights. Next, we will try to find nature clusters of groups and find the group that suits most to our business plan.

3.2 Analytic Approach:

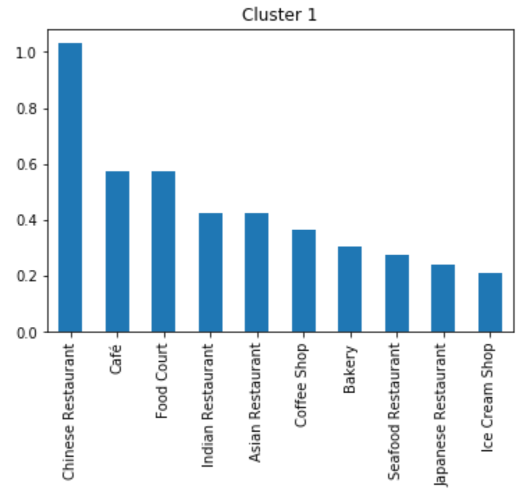
K-Means clustering can be a basic approach of clustering restaurants based on sum occurrence of venue category. It's an unsupervised machine learning algorithm which happens to suit this situation since we do not have 'correct' clustering results for data to fit.

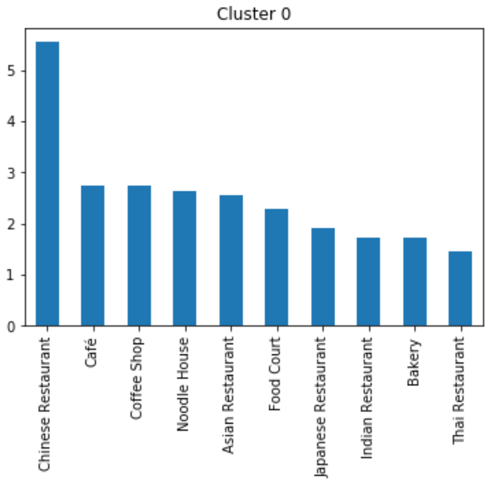
The expected result is that through K-Means clustering, groups with certain features can be split from each other. And through applying ones understanding and intuition, I will be able to judge the performance of the results.

4 Results

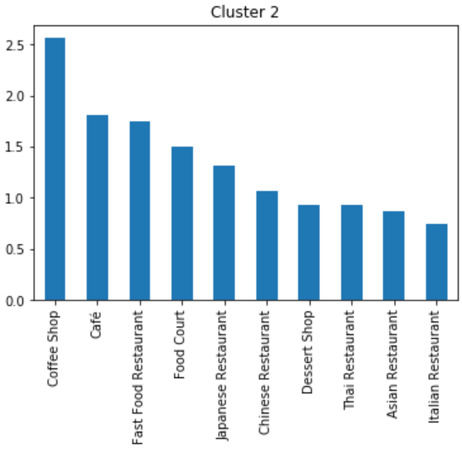
4.1 Cluster basing on food venues

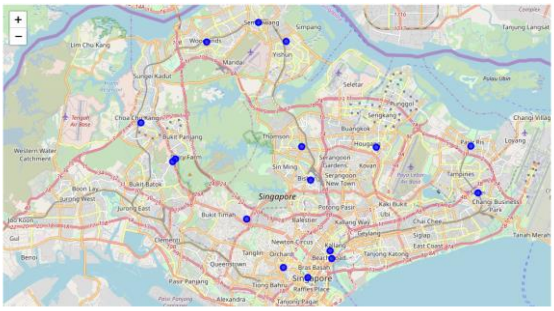
Setting 5 clusters, and using data of restaurants, drinks, snacks and fast food, we obtain below results:



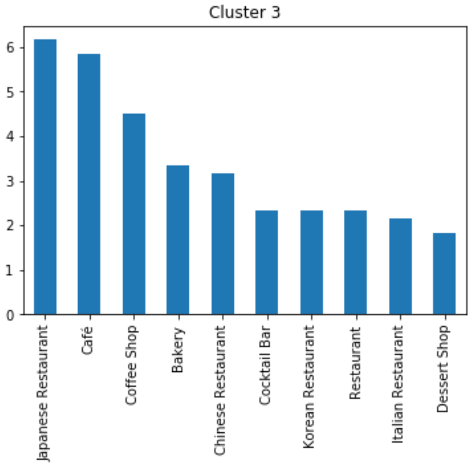


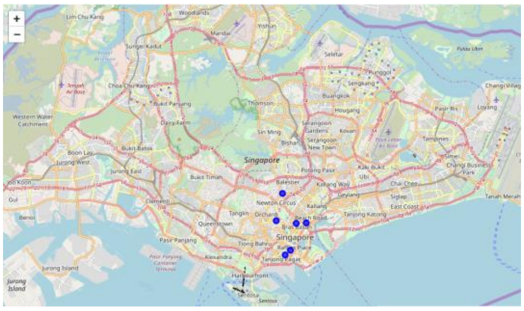
Cluster 0 and 1 are clusters of typical Singapore neighborhoods, with Chinese restaurant dominate in food venues, just like our observation of all data in section 1. The difference lies in that cluster 0 has a higher sum number within certain area while cluster 1 has around 20% of it, indicating that cluster 1 are at barren areas.



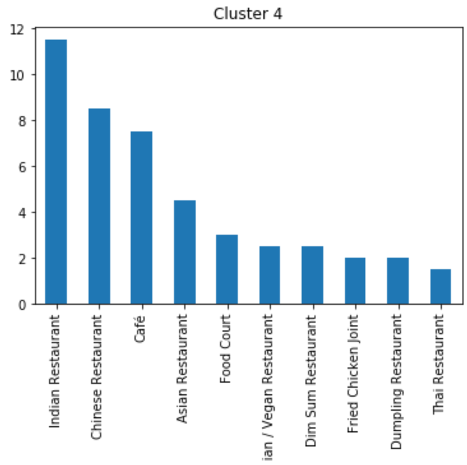


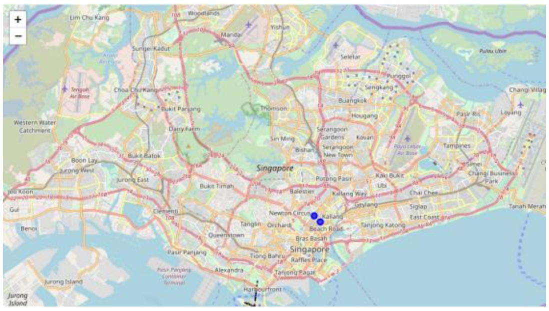
Cluster 2 looks like a typical office places where coffee shops domain the neighborhood.





Cluster 3 is the kind of neighborhood we are looking for. with Japanese food domain in the group, café also accompanies. The result complies with intuition that Japanese restaurants are located at downtown area. Furthermore, we can conclude that these places tend to be places where Italian restaurants and Korean restaurants are and has a lot of coffee shops.





Cluster 4 is a typical Indian cluster, with Indian restaurant domain venue list. The cluster locates exactly around Little India.

The result works great on Cluster 2, 3 and 4 since they clearly represent a kind of neighborhoods. Cluster 0 and 1 represents typical Singapore neighborhoods located in central area and distanced ones.

5 Discussion

In conclusion, we analyzed the relationship between ethic groups and neighborhood venues, and got a conclusion that neighborhood venues can represent group character. We also conformed that currently the common practice of opening a Japanese restaurant is to open it at downtown areas. And lastly, K-Means works well for unsupervised learning, it can classify groups according to input features.