Singapore MRT Station Clustering

Chanon Krittapholchai

7 May 2020

1. Introduction

1.1 Background

Singapore or SG for short is one of the place that travelers would like to visit once. SG has many attractive places to see like Singapore flyer, USS, Marina Bay and many. The most impressive in SG is their public transportation system (MRT) that could link between every parts of SG.

1.2 Problem

Some travelers who have been to SG would have same problem after visited attractive places, have some time left but they don't know where to go in that time. So, from author's opinion, It'd be great to have something that could recommend us where to go from other preference.

But since SG has a good public transport system, so , I'd like to focusing on which MRT station should users go.

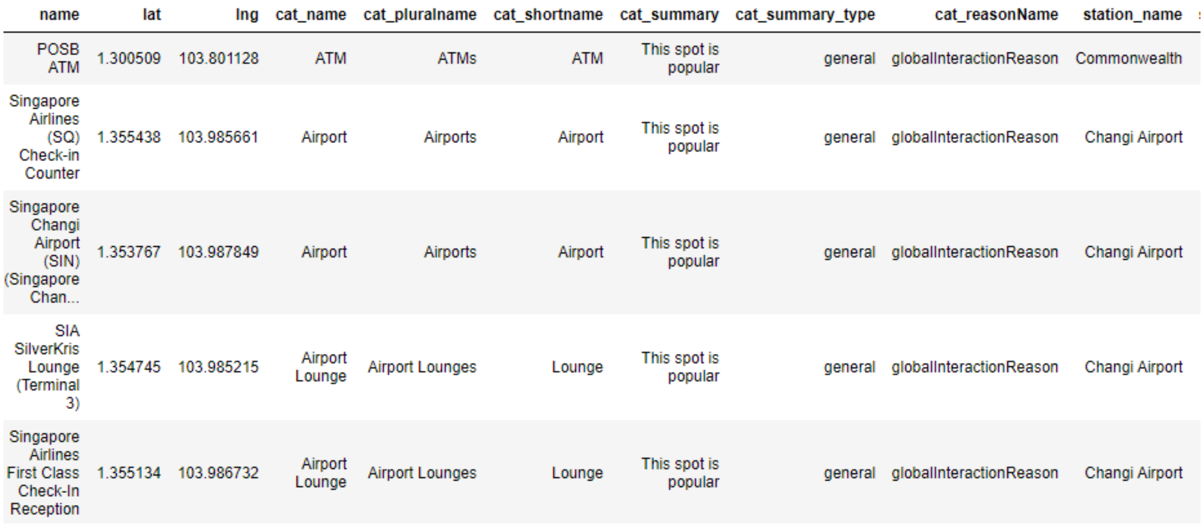
2. Data acquisition and cleaning

2.1 Data sources

In this study, I used data from 2 sources.

1. Kaggle's SG MRT coordinate, contributed by Lee Yu Xuam *https://www.kaggle.com/yxlee245/singapore-train-station-coordinates

2. FoursquareAPI on 750 meter radius from MRT coordinate

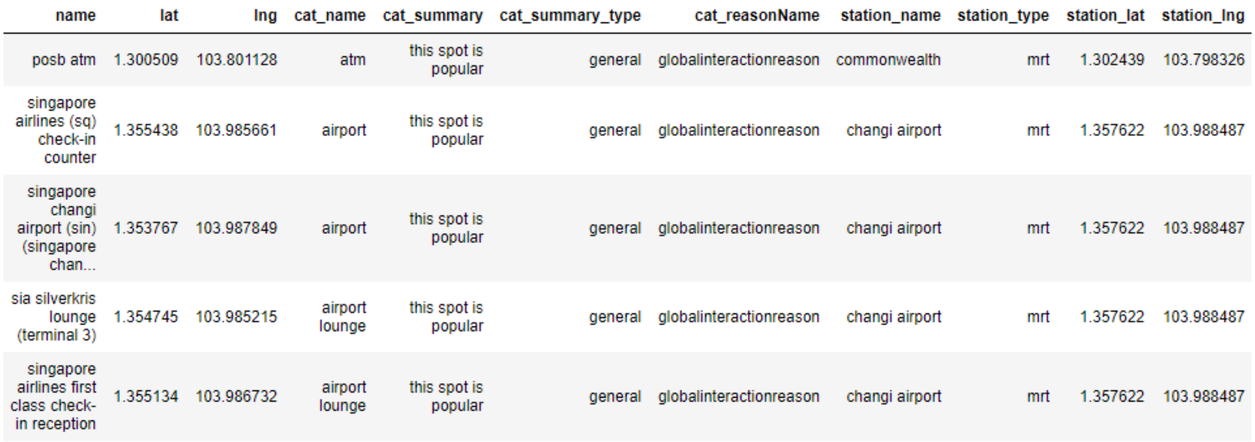


name	lat	lng	cat_name	cat_pluralname	cat_shortname	cat_summary	cat_summary_type	cat_reasonName	station_name
POSB ATM	1.300509	103.801128	ATM	ATMs	ATM	This spot is popular	general	globalInteractionReason	Commonwealth
Singapore Airlines (SQ) Check-in Counter	1.355438	103.985661	Airport	Airports	Airport	This spot is popular	general	globalInteractionReason	Changi Airport
Singapore Changi Airport (SIN) (Singapore Chan...	1.353767	103.987849	Airport	Airports	Airport	This spot is popular	general	globalInteractionReason	Changi Airport
SIA SilverKris Lounge (Terminal 3)	1.354745	103.985215	Airport Lounge	Airport Lounges	Lounge	This spot is popular	general	globalInteractionReason	Changi Airport
Singapore Airlines First Class Check-In Reception	1.355134	103.986732	Airport Lounge	Airport Lounges	Lounge	This spot is popular	general	globalInteractionReason	Changi Airport

2.2 Data cleaning

A. ‘cat_name’, ‘cat_pluralname’ and ‘cat_shortname’ have same information, use only cat_name in this project.
B. From the extracted information, there’re some part that should be correct first
   a. Remove the last ‘s’ letter from every cell in column ‘cat_name’
   b. Lowercased all letter from every cell in column ‘cat_name’ to avoid error from case sensitive

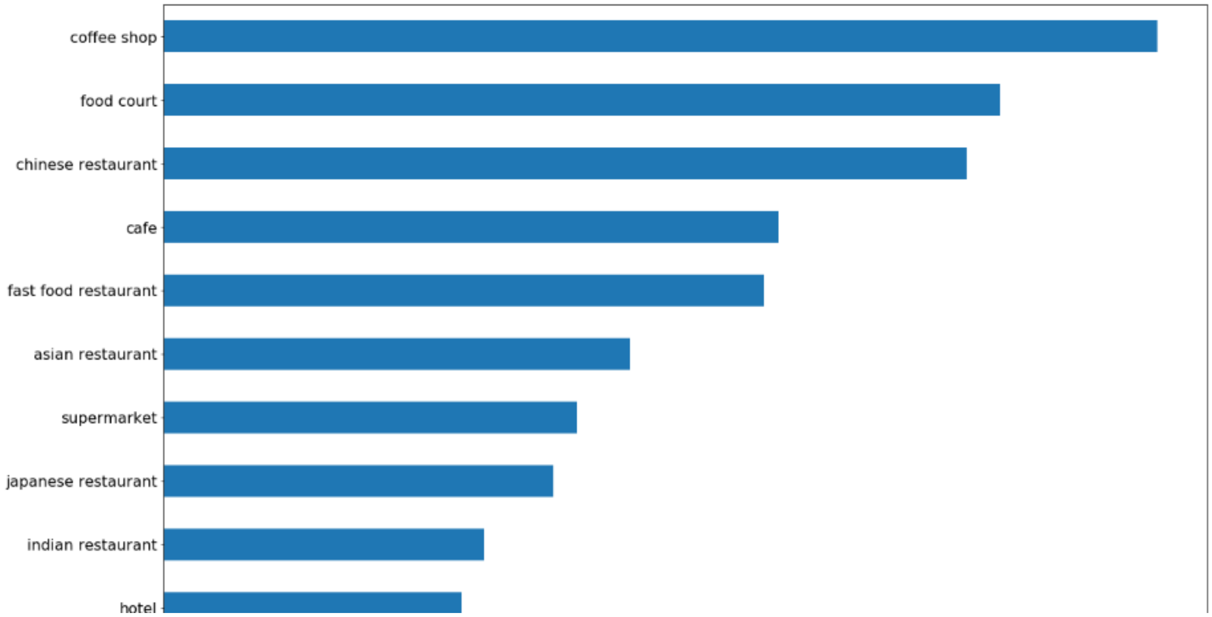
There result after data cleaning process is shown as picture below ;



name	lat	lng	cat_name	cat_summary	cat_summary_type	cat_reasonName	station_name	station_type	station_lat	station_lng
posb atm	1.300509	103.801128	atm	this spot is popular	general	globalinteractionreason	commonwealth	mrt	1.302439	103.798326
singapore airlines (sq) check-in counter	1.355438	103.985661	airport	this spot is popular	general	globalinteractionreason	changi airport	mrt	1.357622	103.988487
singapore changi airport (sin) (singapore chan...	1.353767	103.987849	airport	this spot is popular	general	globalinteractionreason	changi airport	mrt	1.357622	103.988487
sia silverkris lounge (terminal 3)	1.354745	103.985215	airport lounge	this spot is popular	general	globalinteractionreason	changi airport	mrt	1.357622	103.988487
singapore airlines first class check-in reception	1.355134	103.986732	airport lounge	this spot is popular	general	globalinteractionreason	changi airport	mrt	1.357622	103.988487

2.3 Feature Extration

After finish cleaning, check the content in ‘cat_name’ columns ;

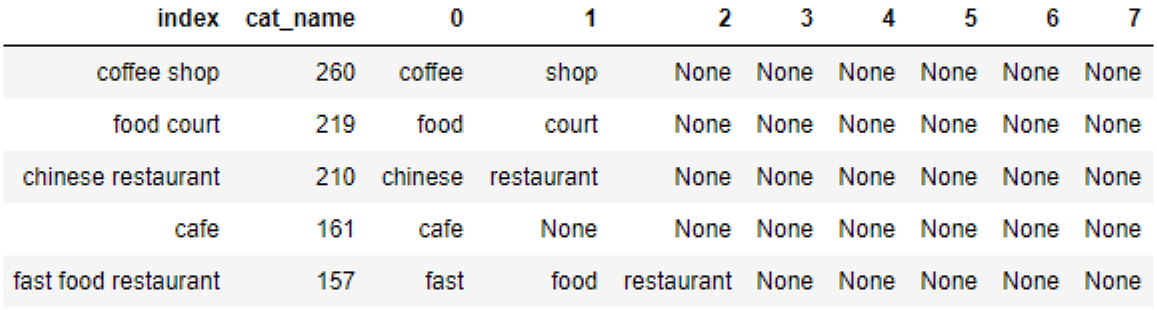


*Full picture could be seen from notebook

I found that data from ‘cat_name’ is not enough information.

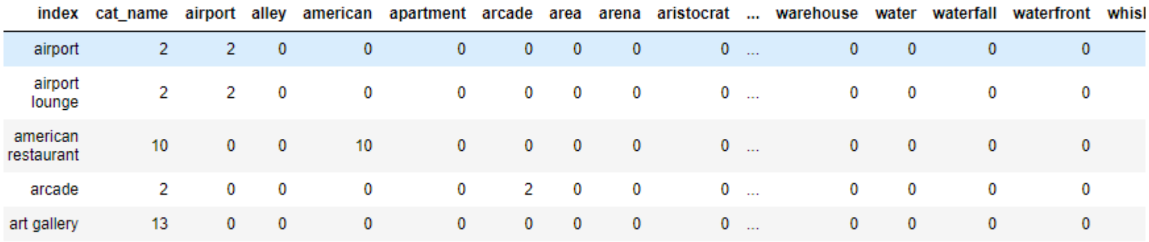
So, I has to extract more information from ‘cat_name’

A. Split cells inside ‘cat_name’ to grain more attributes



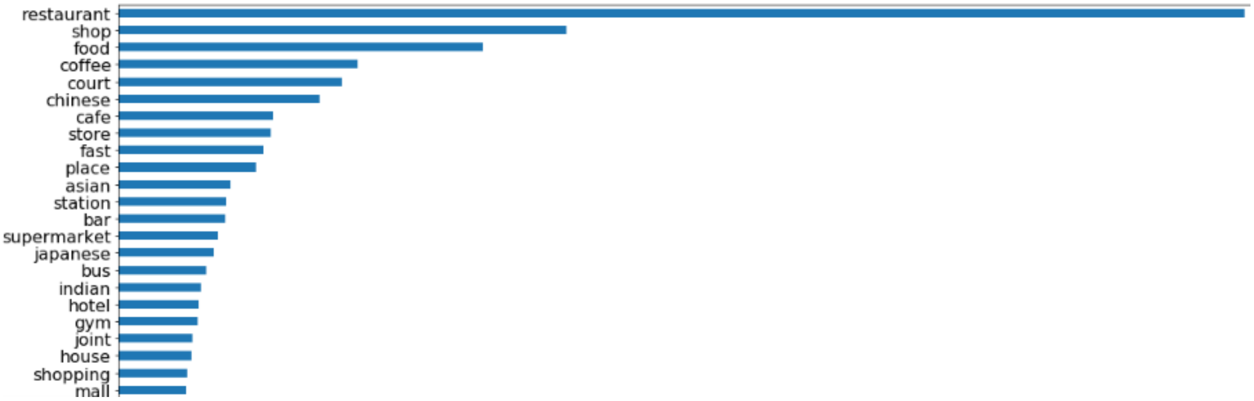
index	cat_name	0	1	2	3	4	5	6	7
coffee shop	260	coffee	shop	None	None	None	None	None	None
food court	219	food	court	None	None	None	None	None	None
chinese restaurant	210	chinese	restaurant	None	None	None	None	None	None
cafe	161	cafe	None	None	None	None	None	None	None
fast food restaurant	157	fast	food	restaurant	None	None	None	None	None

B. Assign attributes to each ‘cat_name’



index	cat_name	airport	alley	american	apartment	arcade	area	arena	aristocrat	...	warehouse	water	waterfall	waterfront	whisl
airport	2	2	0	0	0	0	0	0	0	...	0	0	0	0	
airport lounge	2	2	0	0	0	0	0	0	0	...	0	0	0	0	
american restaurant	10	0	0	10	0	0	0	0	0	...	0	0	0	0	
arcade	2	0	0	0	0	2	0	0	0	...	0	0	0	0	
art gallery	13	0	0	0	0	0	0	0	0	...	0	0	0	0	

C. Make a visualize to gain more information



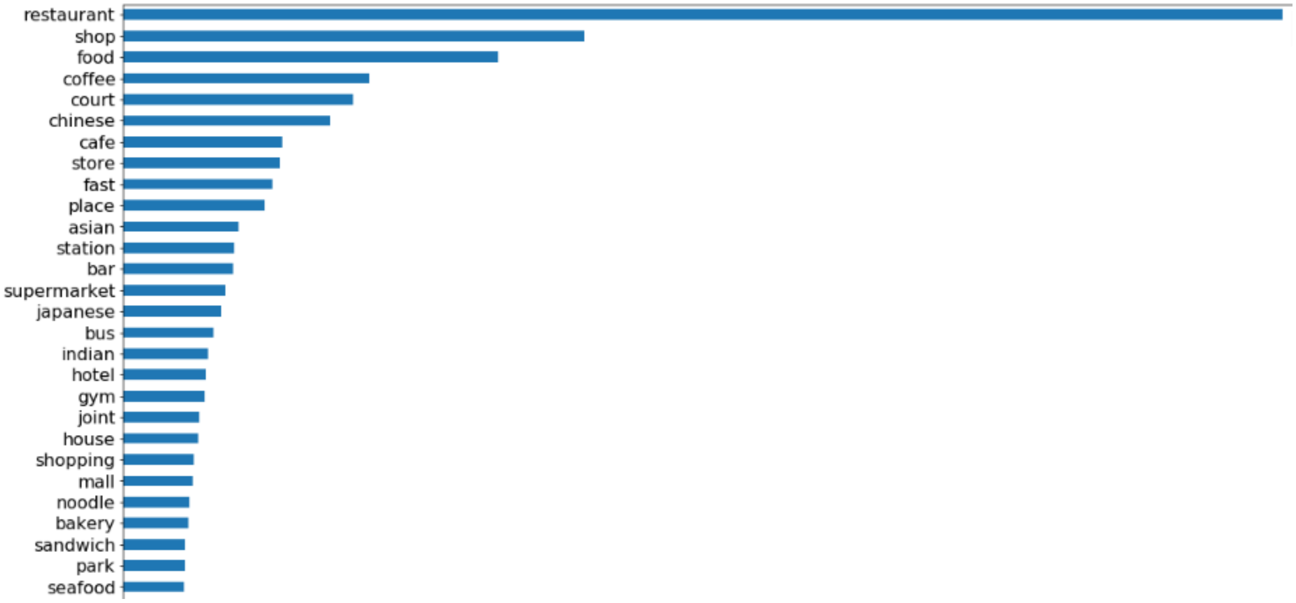
*Full picture could be seen in notebook

D. Some attributes may not have clear or could have different meaning. So, I has to correct them, for example
   a. sometime ‘shop’ could mean ‘café’
   b. sometime ‘house’ could mean ‘restaurant’

E. Grouping some world with the same meaning, for example ‘shop’ and ‘store’

3. Exploratory Data Analysis

From attribute data, I could see more inside from ‘cat_name’



*attributes of each station could be seen in notebook

After review results from attributes of each stations, I could divide that there’re 6 main attributes of each stations as listed below ;

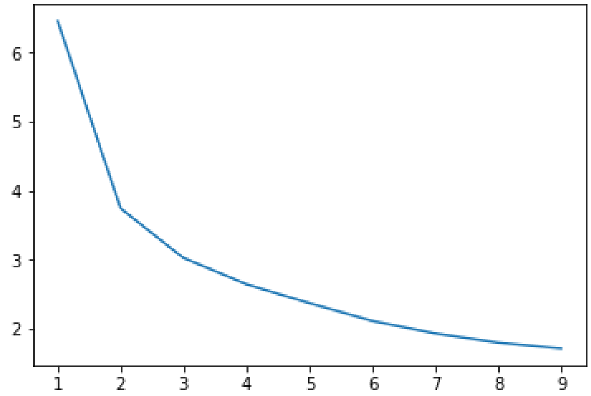
1. Restaurant
2. Café
3. Bar
4. Shop
5. Gym
6. shopping

So, I could make the Machine learning model based on these 6 attributes

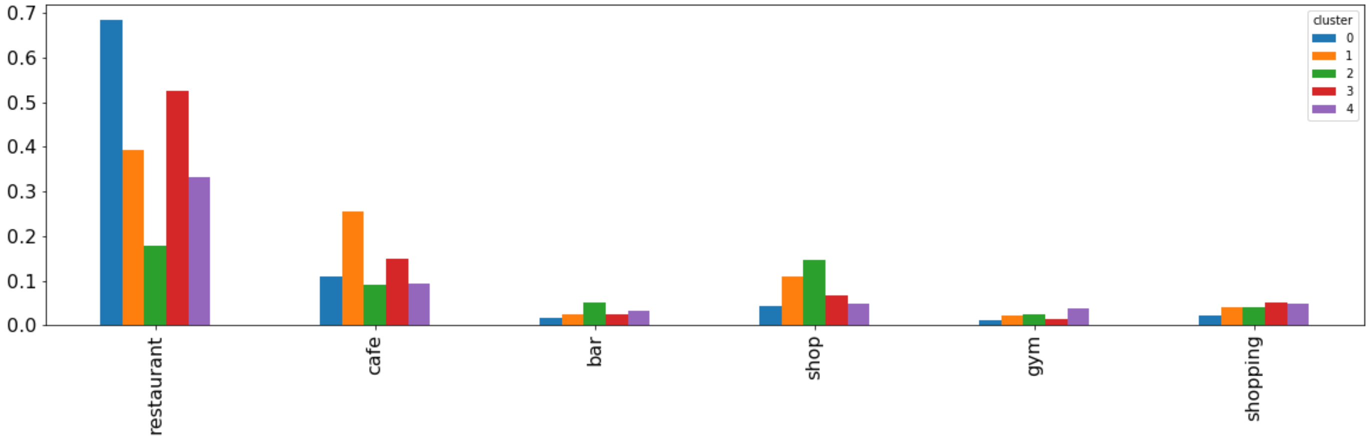
4. Clustering Model

Base on attributes data, Using K-mean to make clusters for each stations

First generate a list of 'K' from 1 – 9 to train models and find the inertia using elbow method.



From the graph, I will use the 'K' as 5 for clustering model and check attributes of each clusters as bar chart below ;

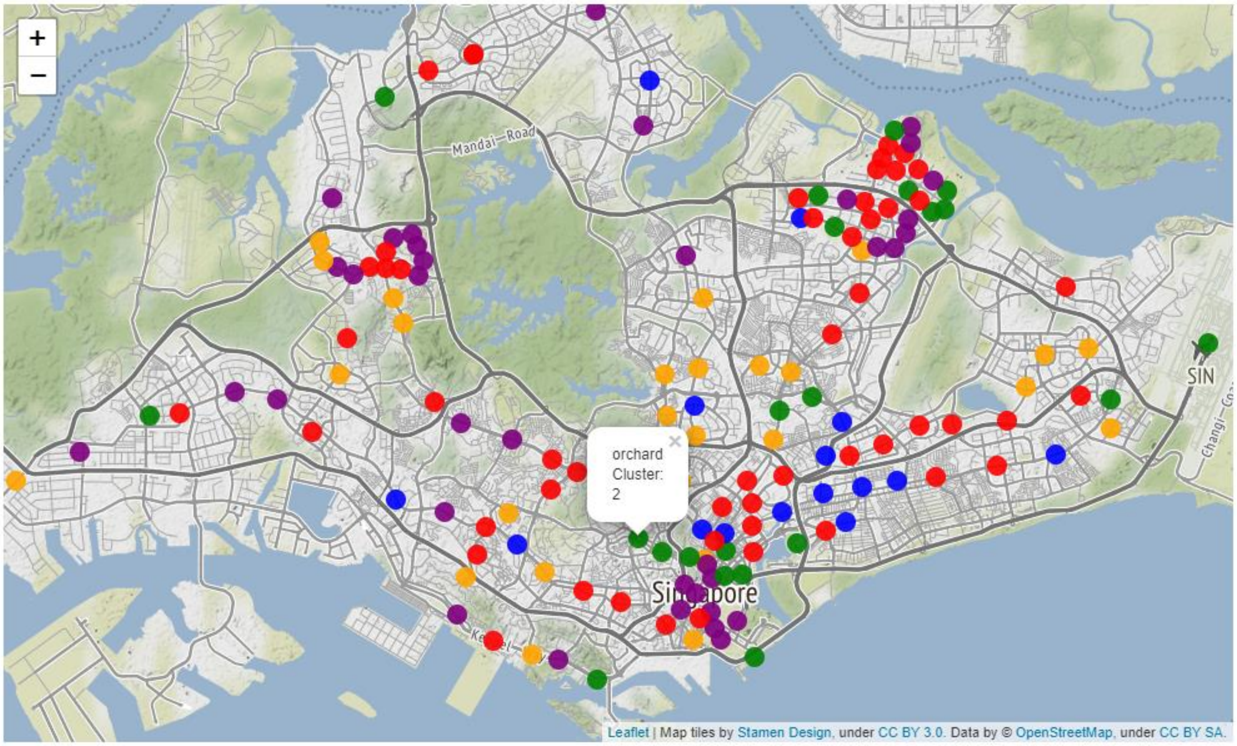


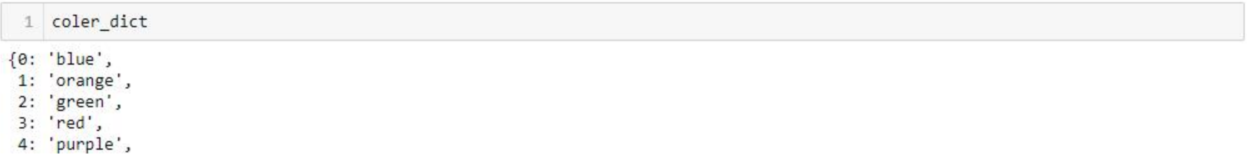
I could define the meaning of each clusters as listed ;

- Cluster 0 : Focusing on Restaurant
- Cluster 1 & 3 : Focusing on Café
- Cluster 2 : Focusing on shop and bar
- Cluster 4 : Mixture of every categories

To check the result, plot map of SG's station with clustering as color,

I could see that 'orchard' the popular shopping station was clustered as cluster 2, shopping


+
−
Mandai Road
orchard
Cluster:
2
Singapore
SIN
Changi
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under CC BY SA.

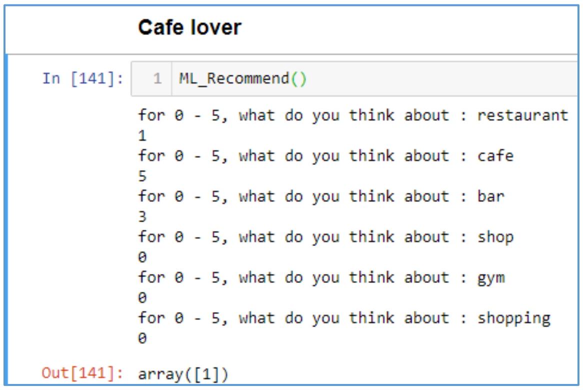


```
coler_dict
```

```
{0: 'blue',
 1: 'orange',
 2: 'green',
 3: 'red',
 4: 'purple',
```

And I also found that most stations at the center of SG are purple or cluster 4, with the mixture of all attributes.

And from this ML model, I could make a recommend ML model based on users' preference of each attributes as shown below ;

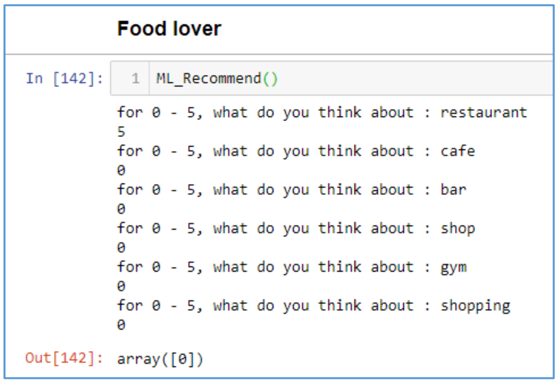


Cafe lover

```
In [141]:  ML_Recommend()

           for 0 - 5, what do you think about : restaurant
           1
           for 0 - 5, what do you think about : cafe
           5
           for 0 - 5, what do you think about : bar
           3
           for 0 - 5, what do you think about : shop
           0
           for 0 - 5, what do you think about : gym
           0
           for 0 - 5, what do you think about : shopping
           0

Out[141]:  array([1])
```



Food lover

```
In [142]:  ML_Recommend()

           for 0 - 5, what do you think about : restaurant
           5
           for 0 - 5, what do you think about : cafe
           0
           for 0 - 5, what do you think about : bar
           0
           for 0 - 5, what do you think about : shop
           0
           for 0 - 5, what do you think about : gym
           0
           for 0 - 5, what do you think about : shopping
           0
Out[142]:  array([0])
```

5. Discussion

From data I analyst so far, there're many other attributes that I'd use to extract event more information, for example the rating of each attributes of each place, the average price per personal, traveling time and many…

Moreover, the biggest job in this project is about cleaning data and extracting more information from data. With more skill in data cleaning, I should be able to get more information.

6. Conclusion

In this project,

I found that we could divide SG's MRT station into 5 clusters base on attributes of places around each stations. And I also can recommend which station the user should go visit during their time in SG base on their preference at some satisficed accuracy by using Data Science technic.