Exploring Singapore's food culture at MRT Interchanges

Introduction

Background

Singapore's Mass Rapid Transport (MRT) is well connected, and is one of the main transport modes in the country. In 2019, the average daily ridership is at 3.384 million, which is about half the population of Singapore.

Like any country, going out for food and drinks is common and is a part of the culture where people spend time catching-up with each other, or just trying out new places. Similarly, businesses are also changing, and new ones are appearing.

Problem area

Singapore is a small country and business competition is high. For both business owners and customers, there are too many options to choose from, or compete against.

For this project, we will determine the top 10 food and drinks places located around major MRT interchanges, and then cluster these places together to determine what makes these area different.

Target Audience

Hopefully, this observation can help new business owners understand the landscape just a little more before they startup, as well as to help customers pick their next outing.

Scope

For the purpose of this project, MRT interchanges are the main train networks that connect 2 or more other MRT networks, and excludes those that connect to Light Rail Transits stations, reason being they are mostly located at residential areas, which is not our target geography.

Data

The datasets required for this analysis are:

1. List of MRT interchanges and its latitude and longitude. The source can be found in data.world and has the features required:
    a. Stn_name
    b. Stn_no
    c. Latitude
    d. Longitude

2. Data for busiest MRT interchanges. mytransport.sg provides monthly ridership statistics, however the data needs to be understood further before using, as it records the entry and exits of the station gantries instead of the passenger count.

3. Nearby venues from Foursquare. We will be using the `explore` API endpoint to source for venues near each of the MRT interchange

Methodology

The data exploration is divided into a few stages:

1. Identifying MRT Interchanges
2. Cleaning data for busiest interchanges
3. Getting nearby venues from Foursquare
4. Clustering

Identifying MRT Interchanges

Data Source: Data.world

For the first step, we are using the data retrieved from data.world, which has the following data structure:

	OBJECTID	STN_NAME	STN_NO	X	Y	Latitude	Longitude	COLOR
0	12	ADMIRALTY MRT STATION	NS10	24402.1063	46918.1131	1.440585	103.800998	RED
1	16	ALJUNIED MRT STATION	EW9	33518.6049	33190.0020	1.316433	103.882893	GREEN
2	33	ANG MO KIO MRT STATION	NS16	29807.2655	39105.7720	1.369933	103.849553	RED
3	81	BAKAU LRT STATION	SE3	36026.0821	41113.8766	1.388093	103.905418	OTHERS
4	80	BANGKIT LRT STATION	BP9	21248.2460	40220.9693	1.380018	103.772667	OTHERS
5	153	BARTLEY MRT STATION	CC12	33168.3039	36108.7003	1.342828	103.879746	YELLOW
6	115	BAYFRONT MRT STATION	DT16	30867.0093	29368.6250	1.281874	103.859073	BLUE
7	115	BAYFRONT MRT STATION	CE1	30867.0093	29368.6250	1.281874	103.859073	OTHERS
8	140	BEAUTY WORLD MRT STATION	DT5	21598.1665	35931.2359	1.341223	103.775810	BLUE
9	37	BEDOK MRT STATION	EW5	38757.9520	34024.7048	1.323980	103.929959	GREEN

The data is cleaned up to keep only the features we want to keep:

	STN_NAME	STN_NO	LATITUDE	LONGITUDE	COLOR
0	ADMIRALTY MRT STATION	NS10	1.440585	103.800998	red
1	ALJUNIED MRT STATION	EW9	1.316433	103.882893	green
2	ANG MO KIO MRT STATION	NS16	1.369933	103.849553	red
3	BAKAU LRT STATION	SE3	1.388093	103.905418	others
4	BANGKIT LRT STATION	BP9	1.380018	103.772667	others

This dataset contains 187 records, and we need to determine which of them are MRT Interchanges.

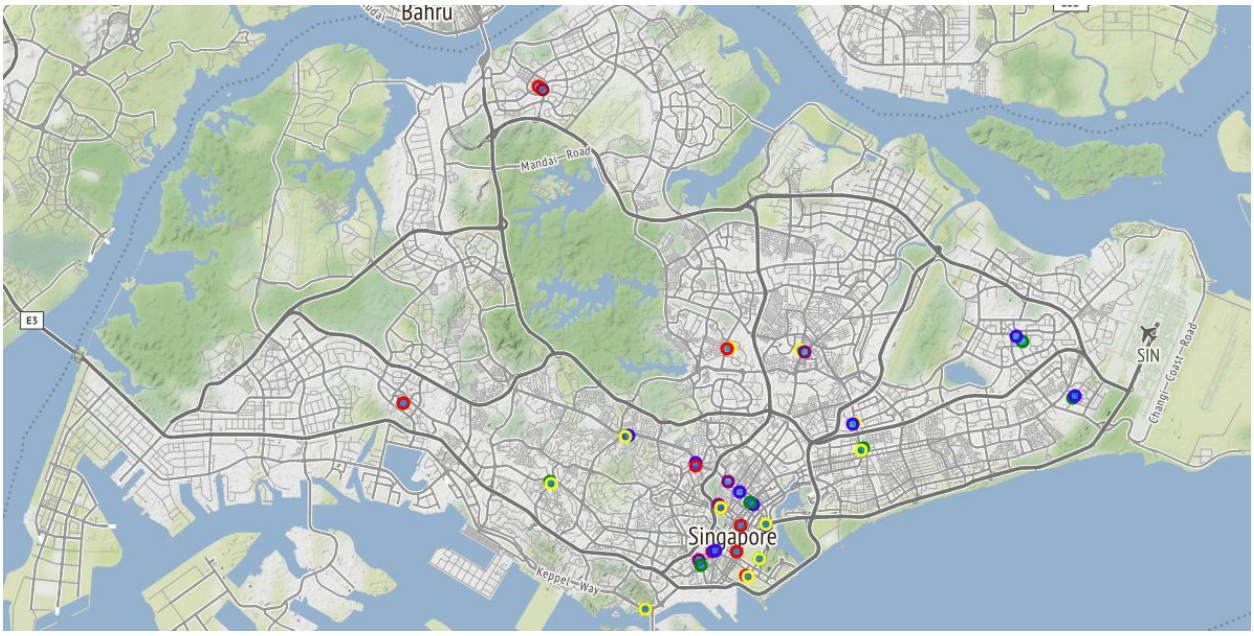
In Singapore, MRT interchanges are connected to 2 or more MRT stations, and they share the same MRT Station name. In the example below, the **Dhoby Ghaut MRT Station** is the same name used in 3 different MRT networks, indicating this is an interchange.

	STN_NAME	STN_NO	LATITUDE	LONGITUDE	COLOR
51	DHOBY GHAUT MRT STATION	NS24	1.298701	103.846112	red
52	DHOBY GHAUT MRT STATION	NE6	1.299705	103.845485	purple
53	DHOBY GHAUT MRT STATION	CC1	1.298843	103.846236	yellow

Similarly, we can group the dataset to view the interchanges as well. In total, there are **22 MRT Interchanges** based on the dataset.

STN_NAME	TOTAL_STNS
DHOBY GHAUT MRT STATION	3
BOTANIC GARDENS MRT STATION	2
SERANGOON MRT STATION	2
HARBOURFRONT MRT STATION	2
RAFFLES PLACE MRT STATION	2

Let's map these stations. We can see that each station has overlapping MRT Stations - the different colors indicate different MRT operating networks.



Grouping MRT Stations into Interchanges

As noted earlier, the dataset provided consists of individual MRT Stations. We now need to change the dataset such that it contains MRT Interchanges only. To do this, we will create a new dataframe with the mean Latitude and Longitude created from the Latitudes and Longitudes of the MRT stations with the same.

For example, Dhoby Ghaut MRT Station has 3 MRT stations connected.

	STN_NAME	STN_NO	LATITUDE	LONGITUDE	COLOR
51	DHOBY GHAUT MRT STATION	NS24	1.298701	103.846112	red
52	DHOBY GHAUT MRT STATION	NE6	1.299705	103.845485	purple
53	DHOBY GHAUT MRT STATION	CC1	1.298843	103.846236	yellow

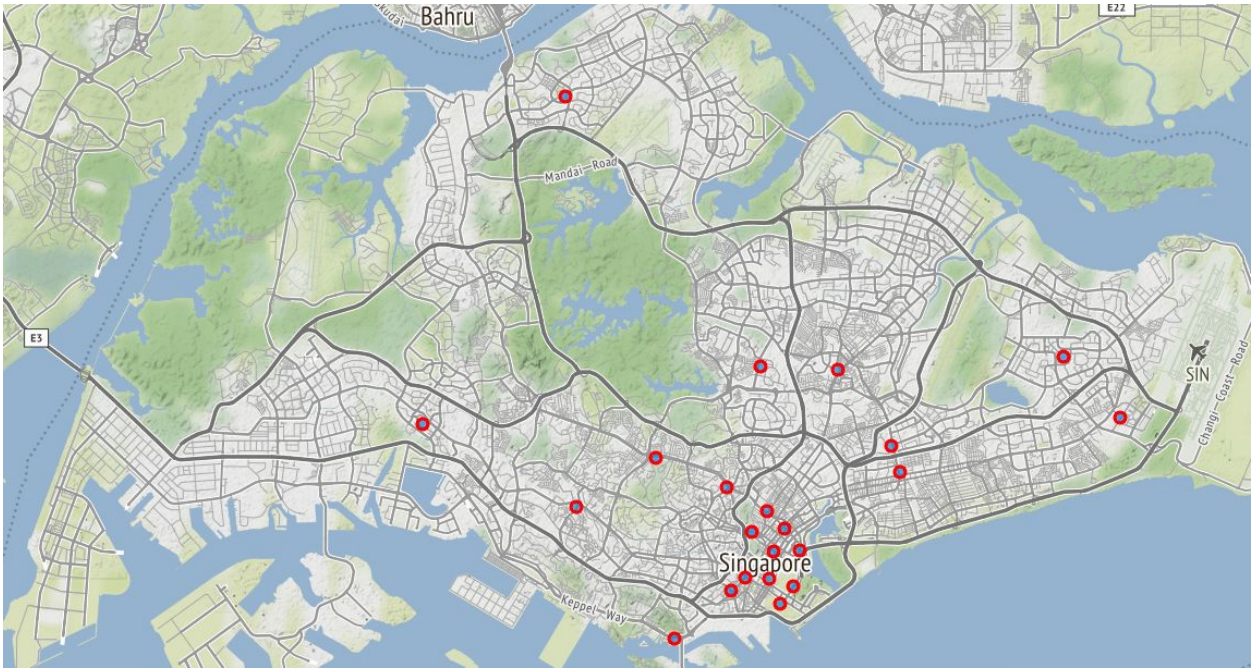
This will be merged into 1 row with a mean Latitude and Longitude.

	STN_NAME	MEAN_LATITUDE	MEAN_LONGITUDE
7	DHOBY GHAUT MRT STATION	1.299083	103.845944

Here is the resulting dataframe.

	STN_NAME	MEAN_LATITUDE	MEAN_LONGITUDE
0	BAYFRONT MRT STATION	1.281874	103.859073
1	BISHAN MRT STATION	1.351074	103.848645
2	BOTANIC GARDENS MRT STATION	1.322267	103.815562
3	BUGIS MRT STATION	1.300008	103.856281
4	BUONA VISTA MRT STATION	1.306838	103.790434

Let’s map these interchanges again.



Cleaning data for busiest interchanges

Data Source: Mytransport.sg

Next, we will be using public data provided by Singapore's Land Transport Authority. In the link above, they provide dynamic data on "Passenger Volume by Train Stations" up to the recent month. For this project, we will be using data from October 2020.

The data structure is as follows.

	YEAR_MONTH	DAY_TYPE	TIME_PER_HOUR	PT_TYPE	PT_CODE	TOTAL_TAP_IN_VOLUME	TOTAL_TAP_OUT_VOLUME
0	2020-10	WEEKDAY	11	TRAIN	NS7	2353	1912
1	2020-10	WEEKENDS/HOLIDAY	11	TRAIN	NS7	1434	1940
2	2020-10	WEEKDAY	16	TRAIN	SW4	1033	1457
3	2020-10	WEEKENDS/HOLIDAY	16	TRAIN	SW4	514	522
4	2020-10	WEEKDAY	10	TRAIN	CC5	1319	3515
5	2020-10	WEEKENDS/HOLIDAY	10	TRAIN	CC5	628	1336
6	2020-10	WEEKDAY	13	TRAIN	CC23	8021	8550
7	2020-10	WEEKENDS/HOLIDAY	13	TRAIN	CC23	1230	1328
8	2020-10	WEEKENDS/HOLIDAY	22	TRAIN	EW33	249	323
9	2020-10	WEEKDAY	22	TRAIN	EW33	727	547
10	2020-10	WEEKENDS/HOLIDAY	14	TRAIN	CC21	3436	3430
11	2020-10	WEEKDAY	14	TRAIN	CC21	7410	7241
12	2020-10	WEEKDAY	18	TRAIN	EW24/NS1	167094	127122
13	2020-10	WEEKENDS/HOLIDAY	18	TRAIN	EW24/NS1	37152	36741
14	2020-10	WEEKENDS/HOLIDAY	7	TRAIN	NE10	2812	2349

A few observations can be made from the data:

- **DAY_TYPE** - Data is separated by **Weekdays** and **Weekends/Holiday**.
- **TIME_PER_HOUR** - Shows the hourly taps in and out of each station / interchange
- **PT_CODE** - Data for multiple MRT Stations (i.e Interchanges) are combined with a '/' - see the last 2 rows in the screenshot
- **TAP_IN** / **TAP_OUT** - Volume of passengers going in and of each station / interchange

Most notable is this dataset does is missing the MRT Station name (e.g "**Dhoby Ghaut MRT Station**"). We'll need to format this data in a way that it can be joined with the MRT Interchange dataset we created in the previous section.

First, we'll will clean up the data by doing a few things:

- Create a new AVG_**PASSENGERS** column as a mean of the **TAP_IN** and **TAP_OUT** values
- Drop the **TAP_IN** and **TAP_IN** columns after calculating the mean
- Drop the **PT_TYPE** column as we know that this data is for TRAIN only.

	YEAR_MONTH	DAY_TYPE	TIME_PER_HOUR	PT_CODE	AVG_PASSENGERS
0	2020-10	WEEKDAY	11	NS7	2132.5
1	2020-10	WEEKENDS/HOLIDAY	11	NS7	1687.0
2	2020-10	WEEKDAY	16	SW4	1245.0
3	2020-10	WEEKENDS/HOLIDAY	16	SW4	518.0
4	2020-10	WEEKDAY	10	CC5	2417.0

Next, we'll address the MRT Station names by doing the following:

1. Splitting the values by '/', and then duplicate the row with each PT_CODE value
2. Assign the station names using the data from the previous section
3. Remove the duplicates

This is the resulting dataset.

	YEAR_MONTH	DAY_TYPE	TIME_PER_HOUR	AVG_PASSENGERS	STN_NAME
0	2020-10	WEEKDAY	18	147108.0	JURONG EAST MRT STATION
1	2020-10	WEEKENDS/HOLIDAY	18	36946.5	JURONG EAST MRT STATION
2	2020-10	WEEKDAY	11	45338.0	JURONG EAST MRT STATION
3	2020-10	WEEKENDS/HOLIDAY	11	24881.5	JURONG EAST MRT STATION
4	2020-10	WEEKENDS/HOLIDAY	22	15347.5	JURONG EAST MRT STATION

Finally, we can now merge this dataset with data from the previous section to get a complete list of interchange with latitude, longitude, and average passengers.

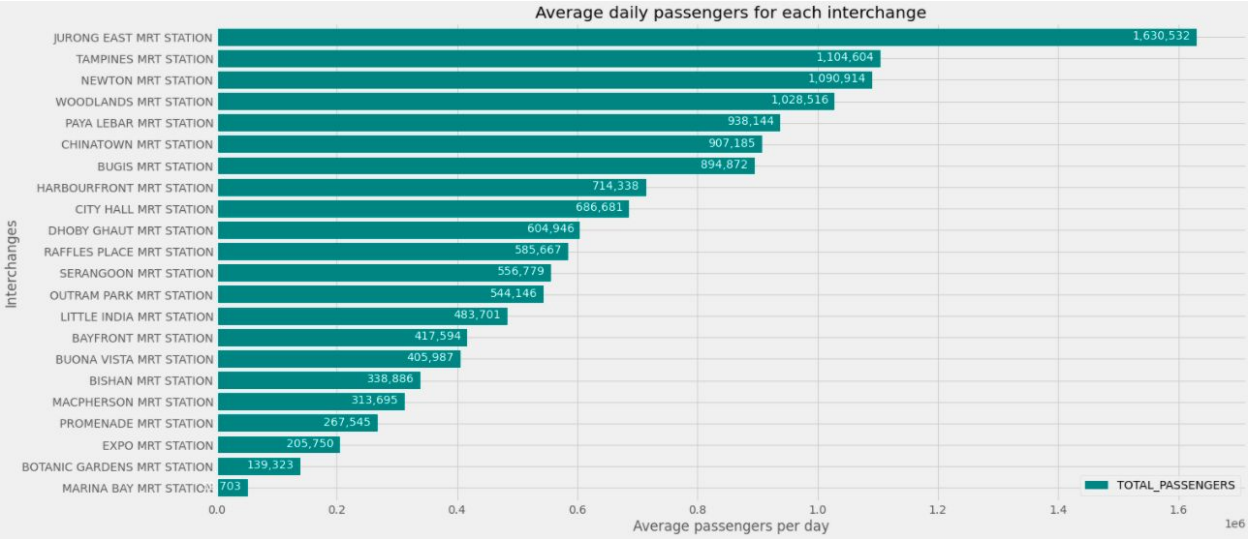
	YEAR_MONTH	DAY_TYPE	TIME_PER_HOUR	AVG_PASSENGERS	STN_NAME	MEAN_LATITUDE	MEAN_LONGITUDE
0	2020-10	WEEKDAY	18	147108.0	JURONG EAST MRT STATION	1.333153	103.742311
1	2020-10	WEEKENDS/HOLIDAY	18	36946.5	JURONG EAST MRT STATION	1.333153	103.742311
2	2020-10	WEEKDAY	11	45338.0	JURONG EAST MRT STATION	1.333153	103.742311
3	2020-10	WEEKENDS/HOLIDAY	11	24881.5	JURONG EAST MRT STATION	1.333153	103.742311
4	2020-10	WEEKENDS/HOLIDAY	22	15347.5	JURONG EAST MRT STATION	1.333153	103.742311

Prepare data for plotting

To rank the busiest interchanges, we'll the **TIME_PER_HOUR** column in the multiple columns (from 0 - 23), and sum the **AVG_PASSENGERS** to get the **TOTAL_PASSENGERS** per station:

STN_NAME	WEEKDAY_TOTAL	WEEKENDS_TOTAL	TOTAL_PASSENGERS
JURONG EAST MRT STATION	1190389.5	440142.5	1630532.0
TAMPINES MRT STATION	815781.0	288823.0	1104604.0
NEWTON MRT STATION	799639.5	291275.0	1090914.5
WOODLANDS MRT STATION	769262.5	259253.5	1028516.0
PAYA LEBAR MRT STATION	649757.5	288387.0	938144.5

And the resulting bar plot:



From the chart, we have the busiest **Jurong East** as the busiest interchange with **1.6million** daily passengers. Jurong East has a high concentration of public housing, which could explain the high number, and **Marina Bay** being the least busiest with only **52.7K** daily passengers - the reason could be that this interchange is tucked far away from residential and office buildings.

Getting nearby venues from Foursquare

Recall that we have 22 MRT Interchanges in total. We will use Foursquare's Explore API with the follow parameters:

1. Latitude / Longitude of each Interchange
2. 1KM radius of each interchange
3. "Food" category only

4. 100 venues per interchange

After running the API, we get the results like below. In total, **2,013** venues across 22 MRT Interchanges were returned.

	STN_NAME	STN_LATITUDE	STN_LONGITUDE	VENUE	VENUE_LATITUDE	VENUE_LONGITUDE	VENUE_CATEGORY
0	BAYFRONT MRT STATION	1.281874	103.859073	Spago	1.283615	103.860682	Italian Restaurant
1	BAYFRONT MRT STATION	1.281874	103.859073	Din Tai Fung 鼎泰豐	1.282270	103.857608	Dumpling Restaurant
2	BAYFRONT MRT STATION	1.281874	103.859073	Waku Ghin	1.283977	103.858597	Japanese Restaurant
3	BAYFRONT MRT STATION	1.281874	103.859073	Adrift by David Myers	1.283141	103.860453	Gastropub
4	BAYFRONT MRT STATION	1.281874	103.859073	CUT by Wolfgang Puck	1.285350	103.859440	Steakhouse

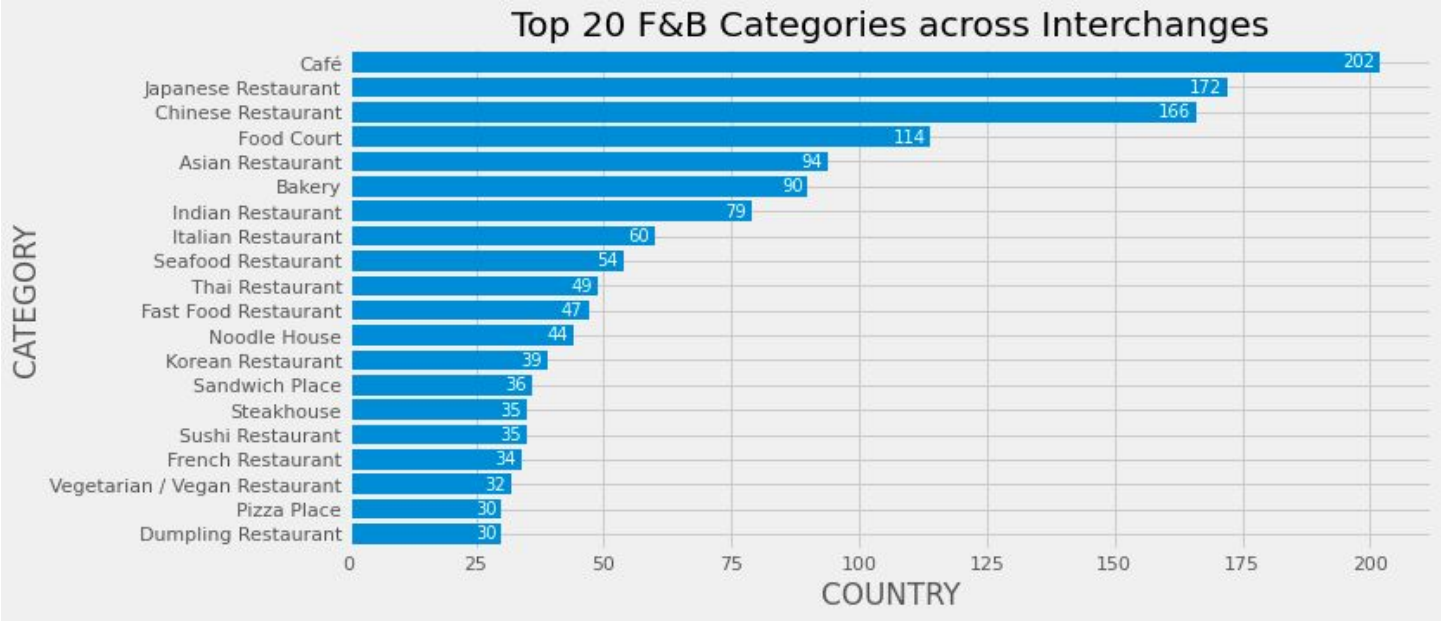
Let's look at the **Top 20 Categories**.

	VENUE_CATEGORY	STN_NAME	S
0	Café	202	
1	Japanese Restaurant	172	
2	Chinese Restaurant	166	
3	Food Court	114	
4	Asian Restaurant	94	
5	Bakery	90	
6	Indian Restaurant	79	
7	Italian Restaurant	60	
8	Seafood Restaurant	54	
9	Restaurant	50	
10	Thai Restaurant	49	
11	Fast Food Restaurant	47	
12	Noodle House	44	
13	Korean Restaurant	39	
14	Sandwich Place	36	
15	Steakhouse	35	
16	Sushi Restaurant	35	
17	French Restaurant	34	
18	Vegetarian / Vegan Restaurant	32	
19	Pizza Place	30	

We observe that one of the categories (highlighted in blue above), contains a generic name - **Restaurant**. Since we cannot reliably rename each venue to its appropriate category, we will drop this category altogether. This should help improve the clustering later.

The final venue count after dropping the **Restaurant** category is **1,963** venues.

We can now visualize these categories in a bar chart.



Clustering

We are now in the last phase where we will use KMeans to cluster all the categories across the interchanges. The goal is to see what makes these interchanges similar or dissimilar based on food categories.

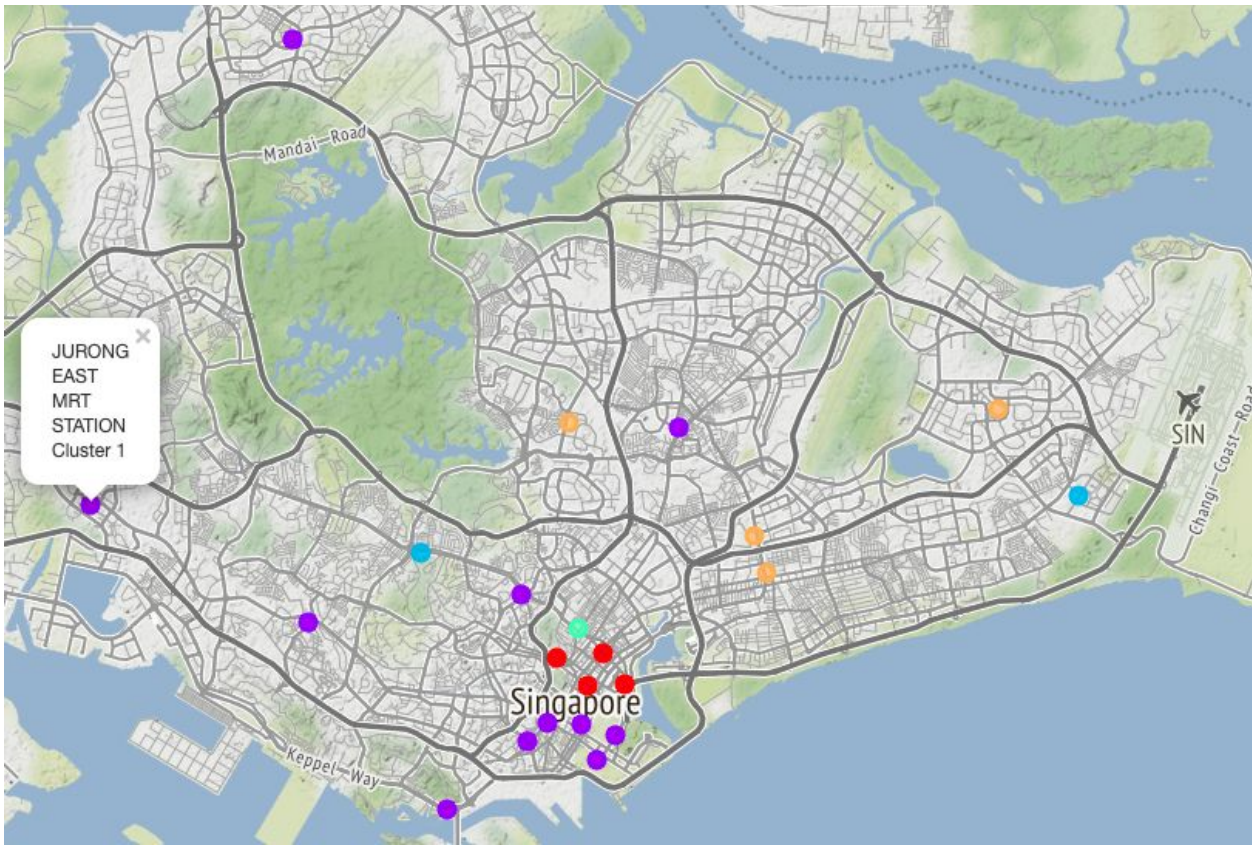
To do this we will:

1. Create one-hot dataframe using the categories
2. Calculate the mean of each category per interchange
3. Run KMeans with 5 clusters - I've tried 4 and 5 clusters, and I found that 5 clusters produce more significant groupings.
4. Rank the 10 most categories per interchange using the mean values from step #2.

Due to space constraints, screenshot only shows the 5 most common category

	STN_NAME	1st Most Common Category	2nd Most Common Category	3rd Most Common Category	4th Most Common Category	5th Most Common Category
0	BAYFRONT MRT STATION	Japanese Restaurant	Café	Sandwich Place	Chinese Restaurant	Food Court
1	BISHAN MRT STATION	Food Court	Chinese Restaurant	Café	Asian Restaurant	Seafood Restaurant
2	BOTANIC GARDENS MRT STATION	Café	Chinese Restaurant	Japanese Restaurant	Asian Restaurant	Thai Restaurant
3	BUGIS MRT STATION	Café	Japanese Restaurant	Chinese Restaurant	Bakery	Indian Restaurant
4	BUONA VISTA MRT STATION	Café	Bakery	Food Court	Indian Restaurant	Japanese Restaurant

Finally, we can visualize the clusters on the map.



Results

The generated clusters are grouped as follows:

Cluster	Color	Description
1	Red	These 4 interchanges are close to the city center. Japanese Restaurants and Cafes are the 2 most common categories, followed by Chinese Restaurant and Bakery being the 3rd most common category
2	Purple	These interchanges are considered busy as well as they connect to other major lines. We can see that it includes the busiest interchange in Singapore, which is **Jurong East** - the purple dot furthest to the left. Japanese Restaurants and Cafes are still 2 most common but they are paired with other categories, such Cafe + Bakery, or, Japanese + Bakery. Additionally, we see the a number of Food Courts as 3rd most common categories
3	Blue	For this cluster, Cafe and Chinese restaurants are the 2 most common category
4	Green	This is perhaps the most popular single cluster. The Little India MRT Station is an interchange and probably has the highest number of Indian restaurants here
5	Orange	The last cluster contains mostly Food Courts, Chinese, and Asian Restaurants as the 3 most common categories. It is likely that these interchanges are further away from the city and closer to residential areas, and food courts are usually common in residentials.

Discussion

In the **Top 20 Categories** chart in the previous, we can clearly see that **Cafes**, **Japanese Restaurants**, **Chinese Restaurant**, and **Food Courts** are highly common in the **1,963** venues we pulled from **FourSquare**.

It's quite clear that Cafe culture is really strong in Singapore - many new young and 'hipster' cafes are popping in the recent years.

However, as a resident in Singapore myself, I'm quite surprised to see that there are so many Japanese restaurants here. Chinese Restaurants are not surprising, seeing that it is quite a common local cuisine here, and lastly, Food Courts are almost a symbol here in Singapore, serving cheap food across multiple cuisines. In general, Asian cuisines (Jap, Thai, Korean, Chinese) are very common around the interchanges.

For western cuisines such as French, Italian, and others, they are available but not as common around these interchanges. Perhaps, a good opportunity for new business owners.

In terms of interchanges - as discussed earlier, Jurong East is the busiest station and it's high public housing concentration could mean we can potentially do a follow-up analysis for Jurong East township, and see what is available in this area. Marina Bay has very low traffic due to its location near the business district, which has a low residency population, and low traffic on weekends, and perhaps new business should stay away from this area unless their business model can support working-hours traffic only.

Conclusion

In this report, we have looked at:

1. The busiest interchanges in Singapore
2. The most common food categories around each interchange

Singapore is still building new MRT stations, and new interchanges will soon appear on the map. We can further refine this report for the future, to allow for new discovery for both business and foodies alike!