Classifying MRT stations in Singapore to assist tourists with their visit

Neil L

15th February 2020

1. Introduction

1.1 Background

There are presently 124 MRT stations in Singapore, which form the backbone of the transport infrastructure for the country. The network is modern, clean, reliable and well connected. At the same time, Singapore weather is typically hot and humid which causes most people to focus their movements and activities to those places that are easy to get to and minimises walking outside. Consequently, the location of these MRT stations is both a consequence of, and a contributing factor to, the types of venues and activities available nearby to their locations. Invariably there will be something of note nearby to any MRT station, however not all stations are made equally. This is of importance to a variety of people, in order to understand the relative significance of each station and why one might visit, live or run a business nearby.

1.2 Problem

One particular example - and the focus of this project - is regarding tourists coming to Singapore, who will be interested to know where might be a good area to stay and which MRT station to be based nearby to. By utilising location data relating to MRT stations, hotels and other places of interest, this project aims to identify those stations that would be best suited to tourists.

1.3 Interest

This analysis will be of immediate interest to tourists coming to Singapore, so that they can best decide on where to stay in order to maximise their enjoyment according to what it is that they are looking for in Singapore. At the same time this analysis may also be of secondary interest to hotel business owners who are looking for pockets of opportunity in Singapore where for example a number of tourist-related businesses are present but hotel presence is not high.

2. Data

2.1 Data sources

A list of MRT stations (both present and planned) can be found on Wikipedia here. Foursquare provided location-based venue information which was used to explore areas adjacent to each active MRT station. Google maps was used to determine geo-location (longitude, latitude) data for MRT stations.

2.2 Data cleaning

Firstly, the MRT station data was scraped from Wikipedia into one table. Stations that are not active yet were removed from the list to leave only those currently working. I have decided to exclude non-active stations because a) the exact location is not always known for the new stations b) I intend the results to be of immediate use with the ability to update later.
One problem with the MRT station data is the lack of an address or geo-location data. I attempted to find this online but could not find a definitive source. Therefore, I used Google Maps to obtain geo-location data for each station and then amalgamated this with the MRT station data extracted from Wikipedia.

2.3 Feature selection

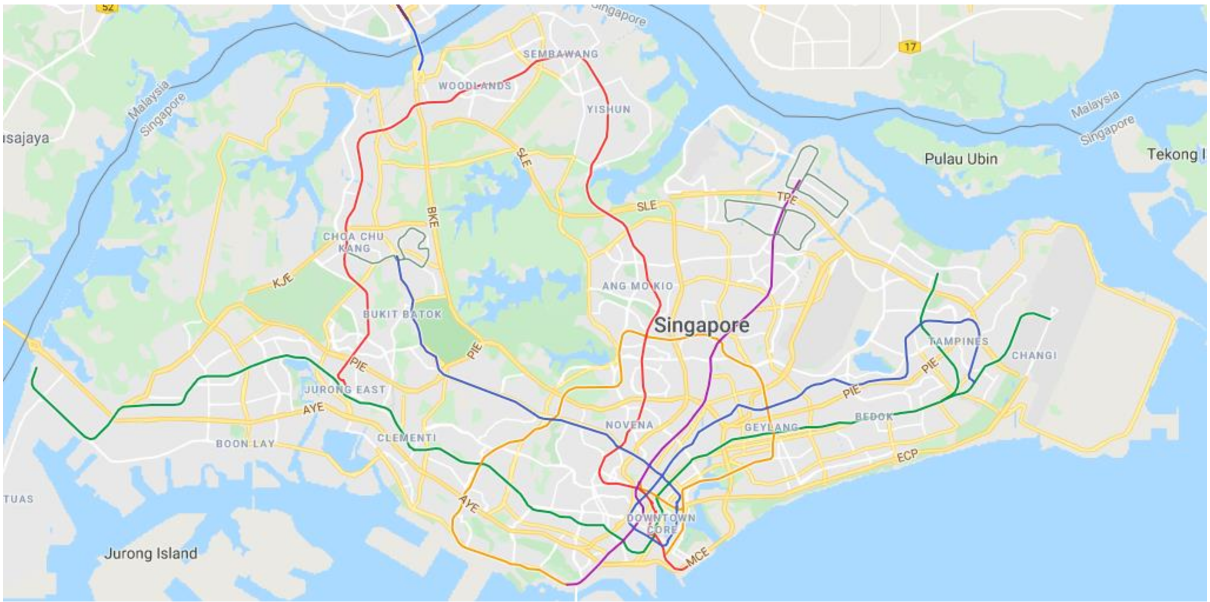
Some MRT stations appear more than once in the list because they cross multiple lines (e.g. Downtown line, Circle line and so forth). Duplications were removed because this project is agnostic to the lines that a station serves, and instead focuses purely on location. Additionally, some stations (e.g. Botanic Gardens) are listed with a dual name including some characters that caused problems when attempting to load the data into a Dataframe; therefore, such characters were removed from the name. In the end the stations table included a list of MRT names along with Longitude and Latitude references.

When selecting data from Foursquare, the query was done via API. The query was based upon the geo-location data of each MRT station and the results included a lot of data. However, given this analysis was focussed on identifying the types of venue around each MRT, I extracted the component of the results that included the venue category, name and geo-location data for each venue.

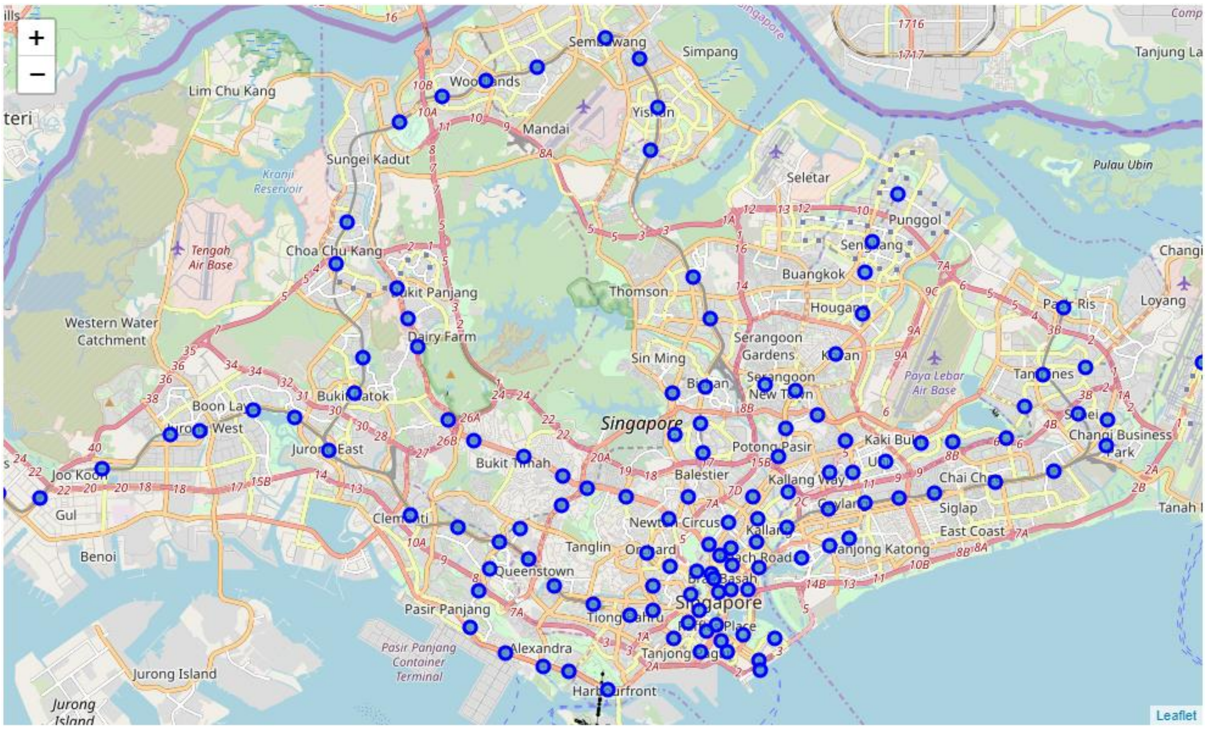
3. Methodology

3.1 Accuracy of the MRT station geo-location data

By creating a map of the MRT stations using the geo-location data it is possible to see the validity of the results. The following is a screenshot from Google Maps, with station lines added for reference:



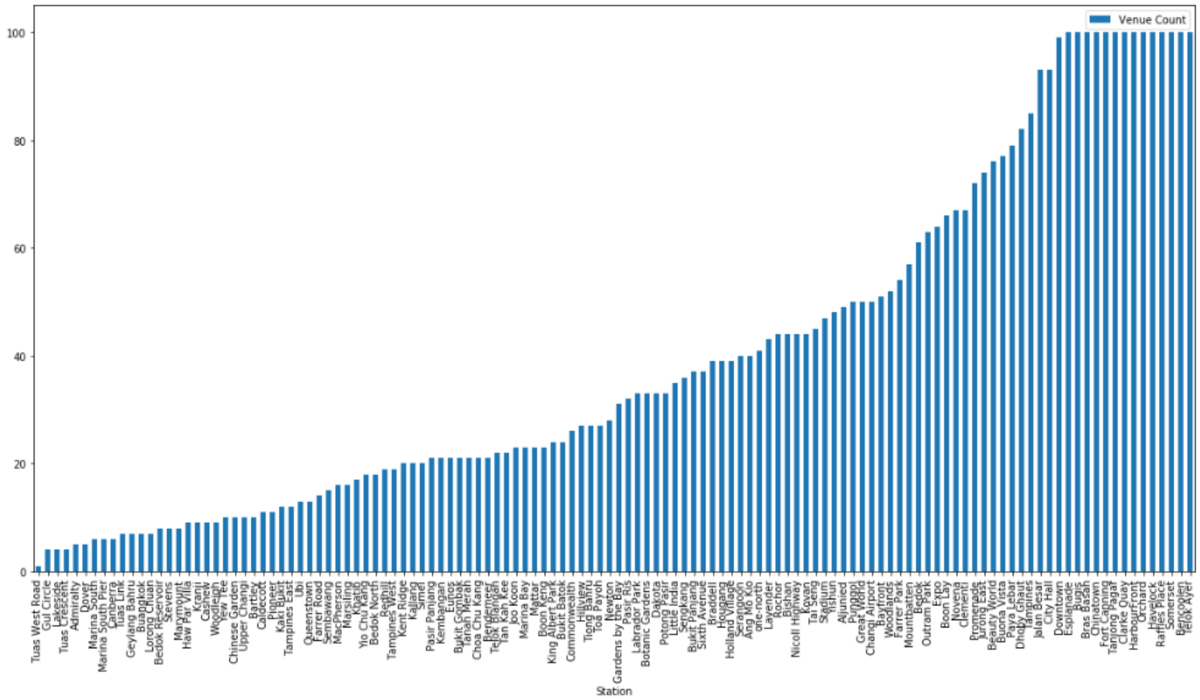
When compared to the station geo-location data plotted onto a map via Folium (see image below), it is possible to see that the locations consistently follow the lines in the map above.



Consequently, it can be surmised that the station location information is accurate and therefore acceptable for use when calling the Foursquare API for venue data.

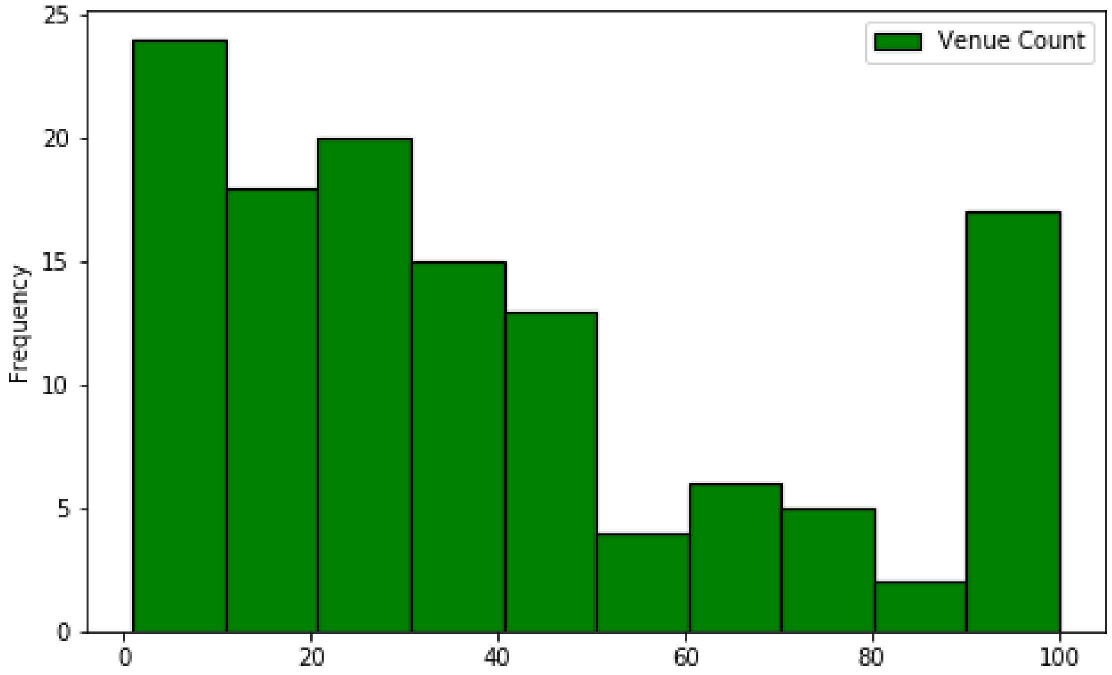
3.2 Foursquare venue list

I limited the query results to 100 venues per location, and within 500m of the station. The query returned nearly 5,000 venues nearby to the various MRT stations. It is also clear from the data that the density of venues varies from station to station.

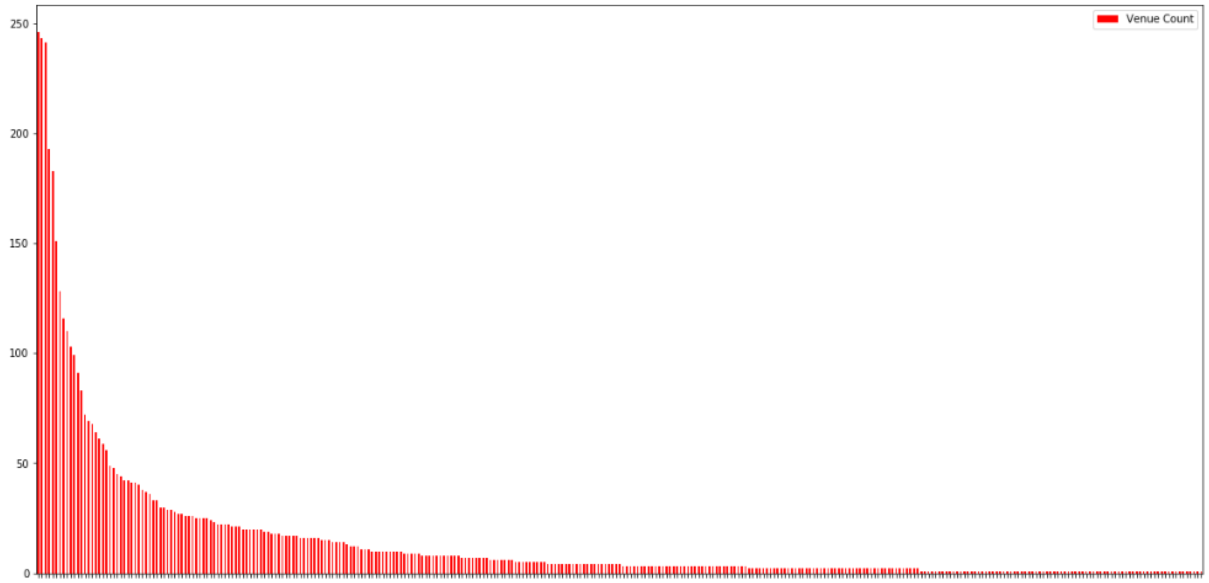


(It is noted from the chart above that some stations have more than 100 venues nearby, however this analysis has capped venues at 100).

When viewed as a histogram (see below) it is apparent that stations typically support either a modest number of venues (< 50) or are densely surrounded (100+).

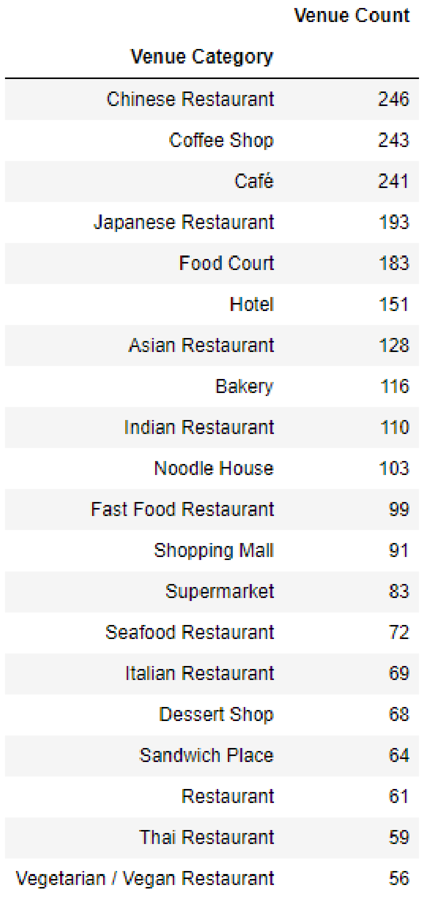


These venues were further spread over 325 different venue categories, and as can be seen below there is large concentration of venues around certain categories, and a very long tail.

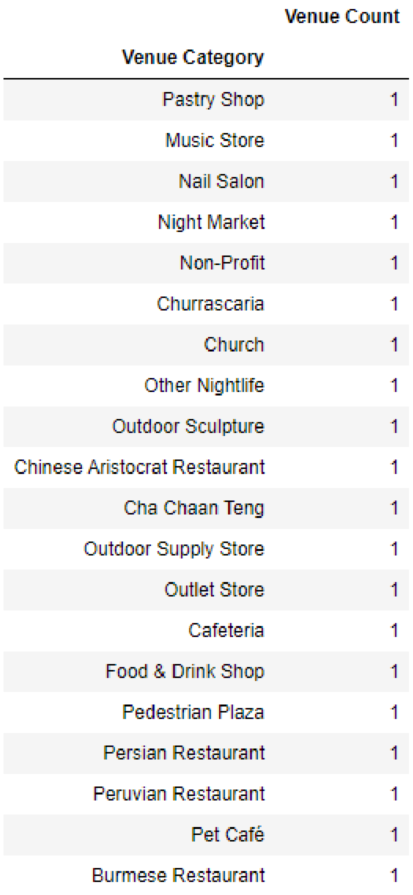


Venue Count by Category

Isolating the top 20 and bottom 20 venue categories (by number of venues tagged to that category) it can be seen in the following tables that even at both ends of the scale a large number of the categories are variations on a similar theme (e.g. restaurants). This presents a challenge for meaningful classification.

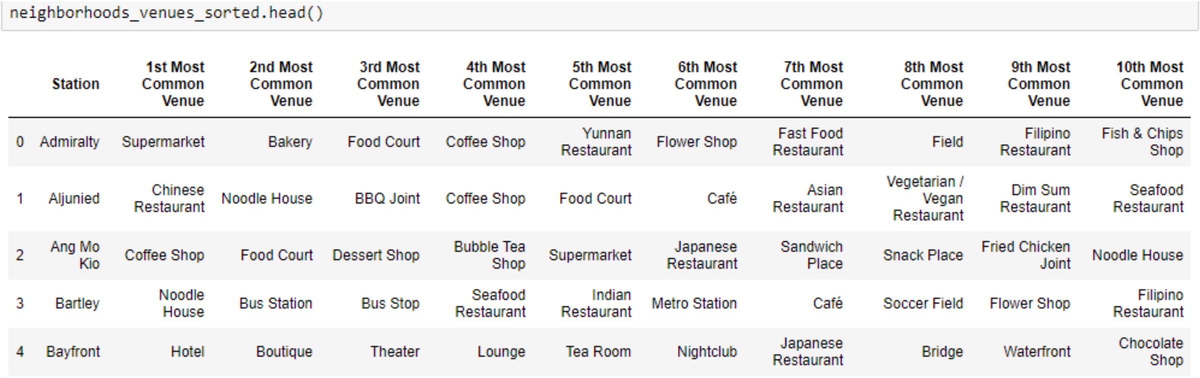


	Venue Count
Venue Category	
Chinese Restaurant	246
Coffee Shop	243
Café	241
Japanese Restaurant	193
Food Court	183
Hotel	151
Asian Restaurant	128
Bakery	116
Indian Restaurant	110
Noodle House	103
Fast Food Restaurant	99
Shopping Mall	91
Supermarket	83
Seafood Restaurant	72
Italian Restaurant	69
Dessert Shop	68
Sandwich Place	64
Restaurant	61
Thai Restaurant	59
Vegetarian / Vegan Restaurant	56



	Venue Count
Venue Category	
Pastry Shop	1
Music Store	1
Nail Salon	1
Night Market	1
Non-Profit	1
Churrascaria	1
Church	1
Other Nightlife	1
Outdoor Sculpture	1
Chinese Aristocrat Restaurant	1
Cha Chaan Teng	1
Outdoor Supply Store	1
Outlet Store	1
Cafeteria	1
Food & Drink Shop	1
Pedestrian Plaza	1
Persian Restaurant	1
Peruvian Restaurant	1
Pet Café	1
Burmese Restaurant	1

The challenge of number of categories becomes even more apparent when the categories are aggregated at a station level and then ranked (for each station) according to number of instances. As can be seen from the following screenshot, from the first 5 stations alone it is clear that a number of categories appear which may be considered similar for the purposes of the analysis being performed here.



```
neighborhoods_venues_sorted.head()
```

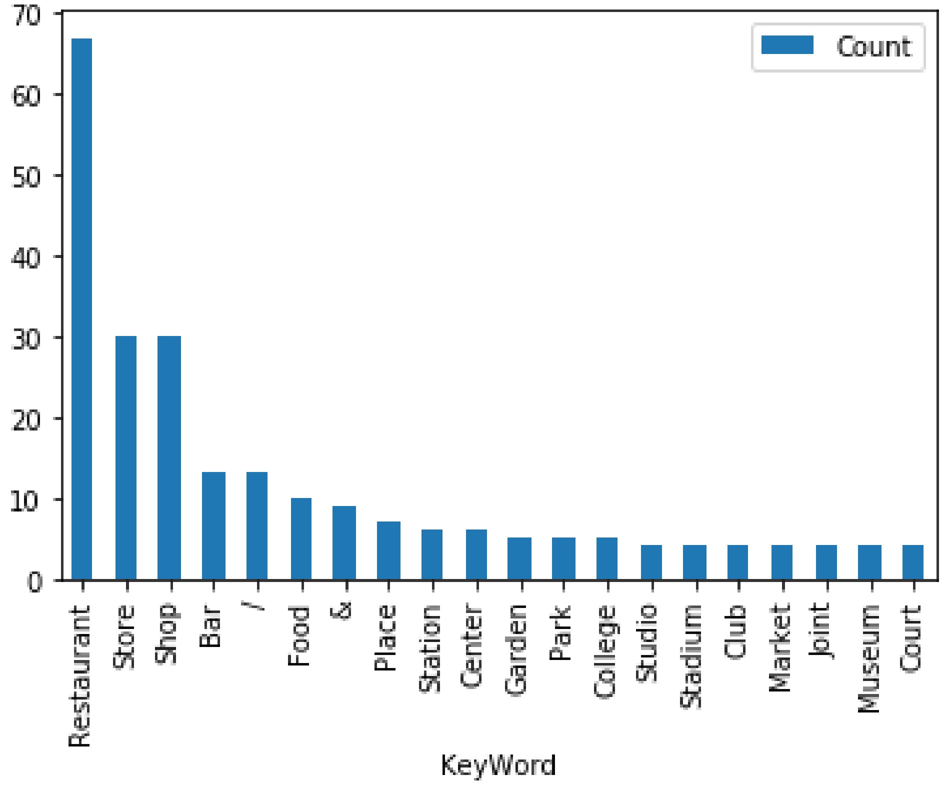
	Station	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Admiralty	Supermarket	Bakery	Food Court	Coffee Shop	Yunnan Restaurant	Flower Shop	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop
1	Aljunied	Chinese Restaurant	Noodle House	BBQ Joint	Coffee Shop	Food Court	Café	Asian Restaurant	Vegetarian / Vegan Restaurant	Dim Sum Restaurant	Seafood Restaurant
2	Ang Mo Kio	Coffee Shop	Food Court	Dessert Shop	Bubble Tea Shop	Supermarket	Japanese Restaurant	Sandwich Place	Snack Place	Fried Chicken Joint	Noodle House
3	Bartley	Noodle House	Bus Station	Bus Stop	Seafood Restaurant	Indian Restaurant	Metro Station	Café	Soccer Field	Flower Shop	Filipino Restaurant
4	Bayfront	Hotel	Boutique	Theater	Lounge	Tea Room	Nightclub	Japanese Restaurant	Bridge	Waterfront	Chocolate Shop

For example, near to Aljunied station there are a number of restaurants appearing separately in the most common venues (e.g. Chinese Restaurant, Asian Restaurant, Dim Sum Restaurant) when in fact we may want to consider these in aggregate under the broader category of Restaurants. This is particularly true when considered in light of the target audience (i.e. tourists) who are likely interested in the wider availability of restaurants as opposed to one particular cuisine type, although this would be an interesting area for further analysis subsequent to this project.

As a result, I set about to further aggregate the categories for the purposes of classification.

3.3 Creating major venue categories for further analysis

By breaking venue category names down into component words, it was possible to identify the frequency of certain words appearing. The top 20 key words are listed below:



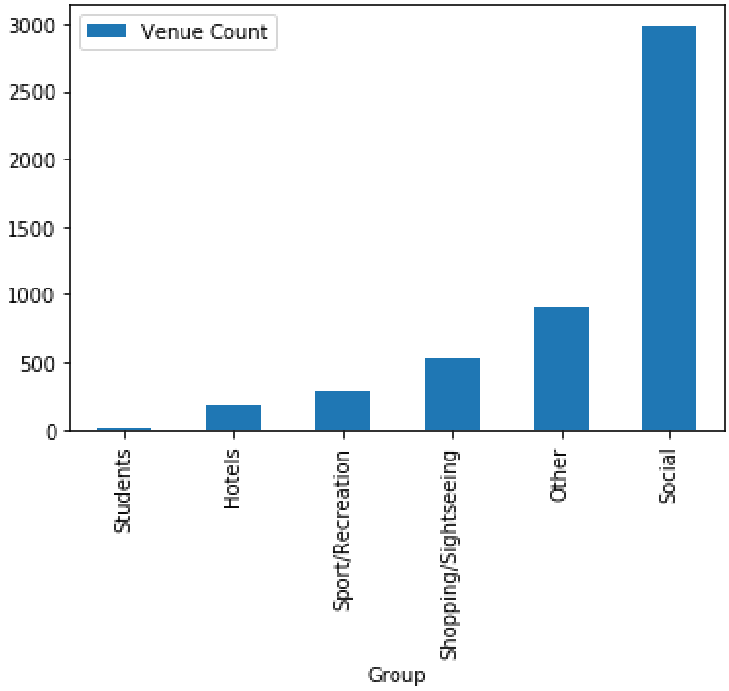
The keyword ‘Restaurant’ is by far the most common, appearing in almost 70 different venue category names. However even from the chart above it is possible to infer an aggregation opportunity inherent in the keywords, because alongside ‘Restaurant’ we can also see ‘Food’, ‘Bar’ and ‘Club’ amongst others. Clearly there is merit in aggregating certain words to form more coherent groupings.

Given the focus of this project and the needs of tourists, I created a set of major groups to which each venue category would be broadly assigned:

- **Socialising** (restaurants, bars, clubs etc)
- **Shopping & Sightseeing** (shopping malls, theatres, major sights etc)
- **Sports & Recreation** (sports facilities, parks etc)
- **Attending courses, studying** (universities, colleges etc)
- **Relaxing at the Hotel** (hotels, motels etc)
- **Other** (any other venues not fitting in to the above categories e.g. car park)

These groups are designed to broadly reflect the main types of activity that a tourist may be in Singapore to undertake.

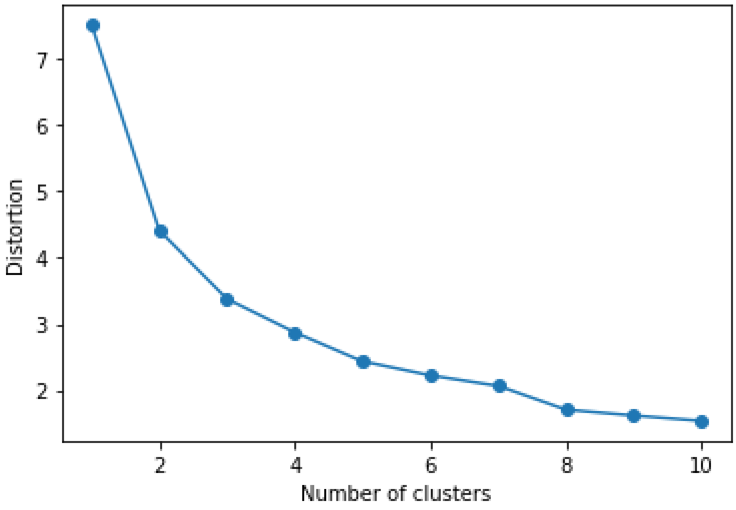
As a result of grouping the venues into these categories it is much easier to see the density of venues and what purposes they serve.



With this dataset it was much better to cluster the stations according to the activity groups they support. As can be seen above, social venues are the most popular in Singapore and tourists looking for this type of activity are very well served. It is noted that an extension of this project could be to analyse at a 'sub-group' level e.g. 'Social' could be broken into restaurants, bar, clubs and so forth.

3.4 K-means clustering

In order to cluster the stations, I decided to use the k-means algorithm which is one of the most common methods of unsupervised machine learning for clustering. Further, because of its unsupervised nature I utilised the elbow method to determine an optimum number of clusters for the algorithm to solve for.

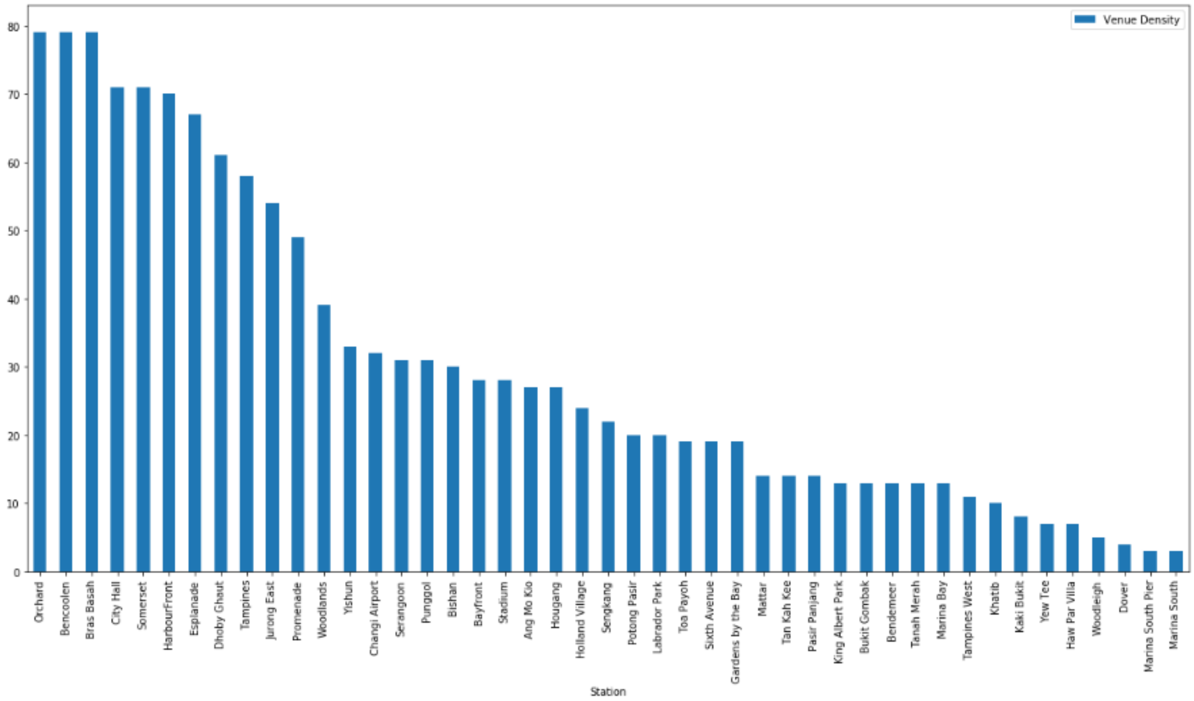


The plot above shows the results for cluster (k) values from 1 to 10. There is no single point at which the elbow is dominant, however I decided to take a k value of 3 as the gradient of decrease in distortion appears to drop notably from this point.

3.5 Clustering results

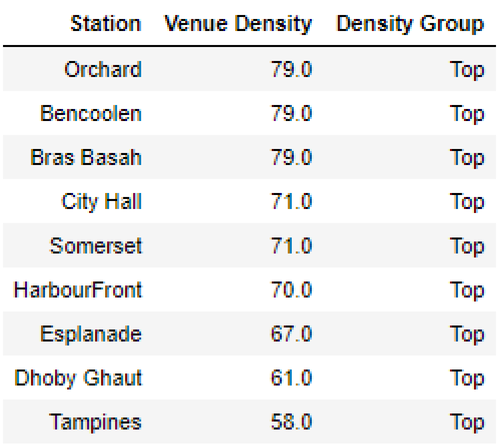
By assessing the results of the clustering, I was able to label the clusters as follows:

Cluster 1: Social, Shopping & Sightseeing



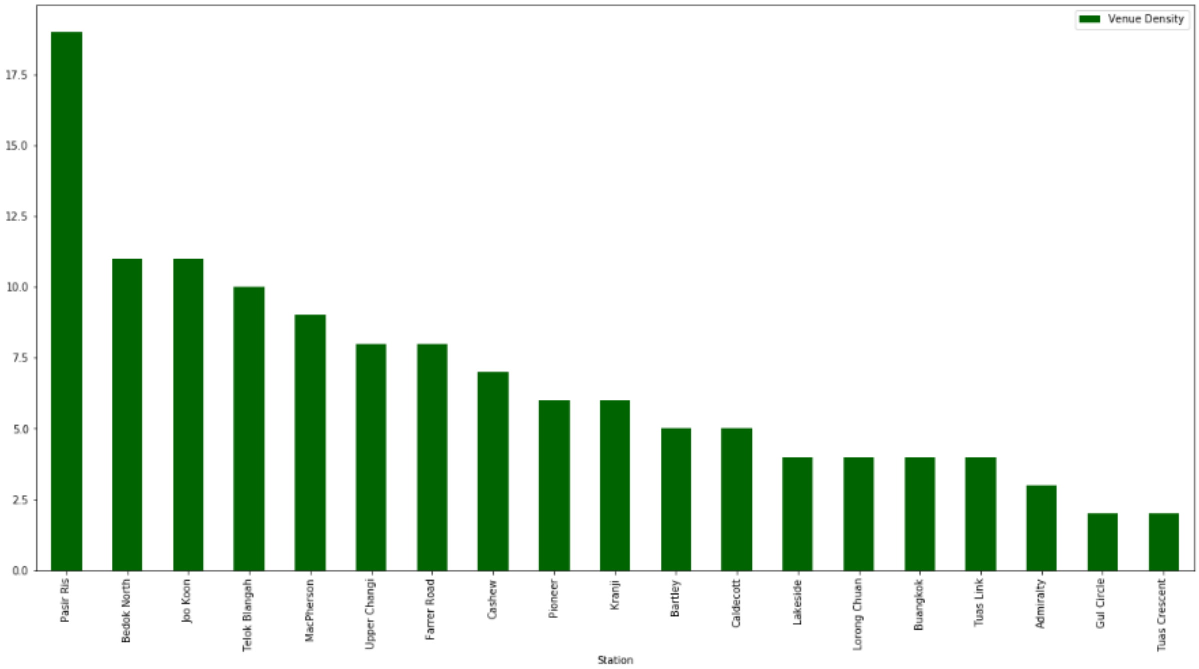
The station in cluster 1 with most venues was a tie between Orchard, Bencoolen and Bras Basah with 79 venues each. Using this number of 79 I categorised any venue in cluster 1 with > 70% of this number (70% of 79) as being a 'Top' station for venue density in that cluster.

Those categorised as 'Top' for cluster 1 are as follows:

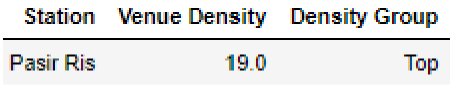


Station	Venue Density	Density Group
Orchard	79.0	Top
Bencoolen	79.0	Top
Bras Basah	79.0	Top
City Hall	71.0	Top
Somerset	71.0	Top
HarbourFront	70.0	Top
Esplanade	67.0	Top
Dhoby Ghaut	61.0	Top
Tampines	58.0	Top

Cluster 2: Sports, Recreation & Other

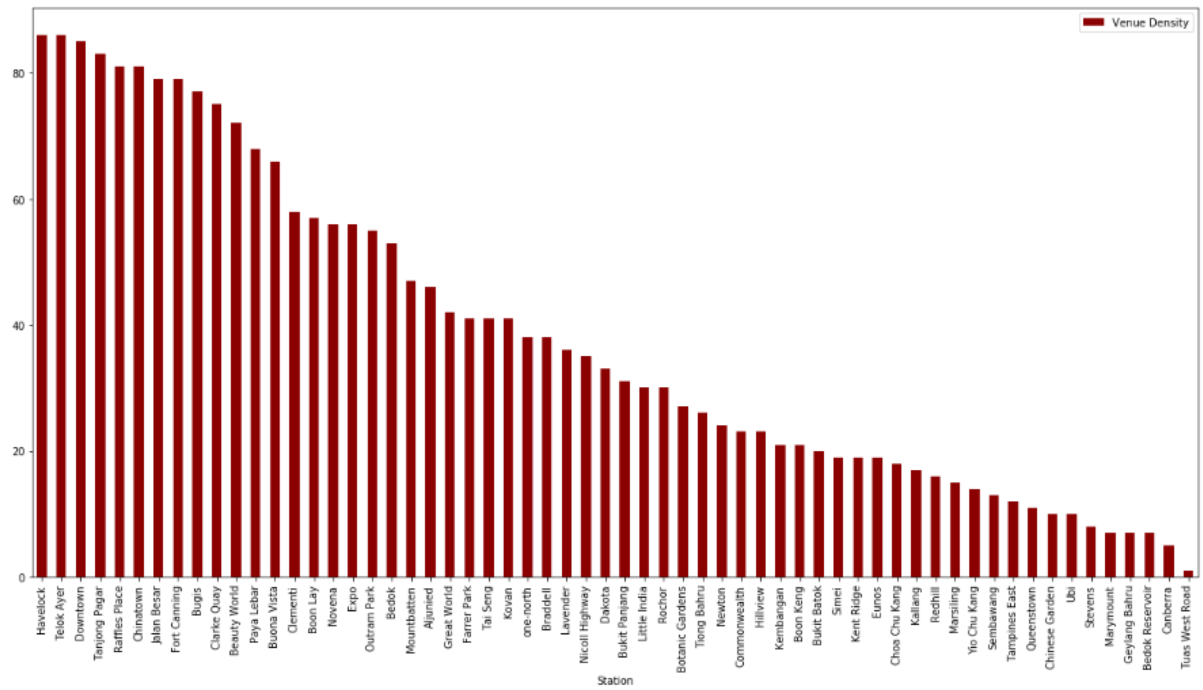


The station in cluster 2 with most venues was Pasir Ris with 19 venues. Using a similar approach as cluster 1, there are no other stations with > 70% of this number. Therefore, only Pasir Ris is categorised as a ‘Top’ station for venue density.



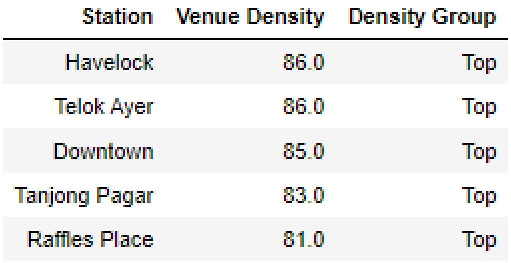
Station	Venue Density	Density Group
Pasir Ris	19.0	Top

Cluster 3: Social & Other



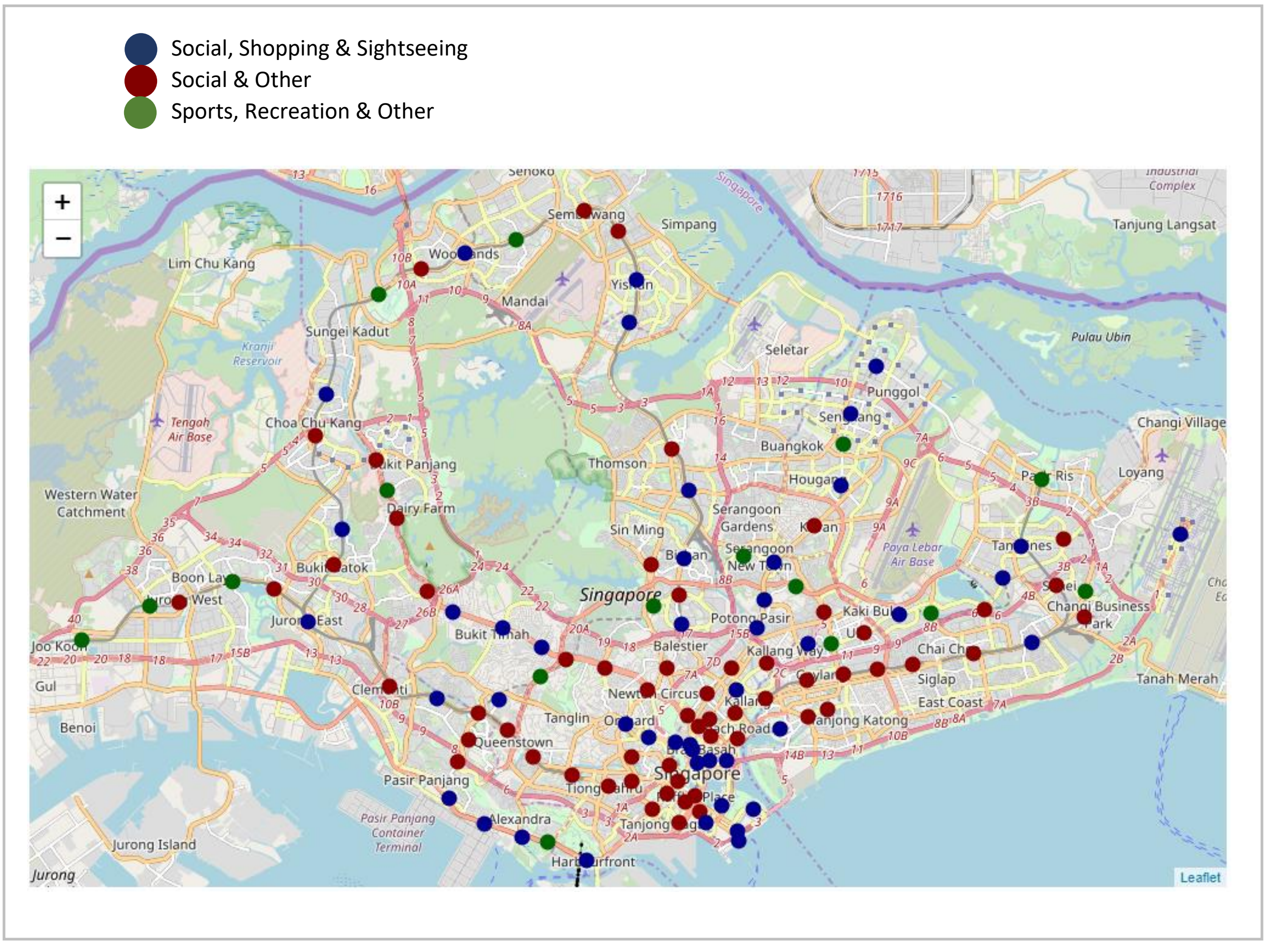
The station in cluster 3 with most venues was a tie between Havelock and Telok Ayer with 86 venues each. Using this number of 86 I categorised any venue in cluster 1 with > 70% of this number (70% of 86) as being a 'Top' station for venue density.

Those categorised as 'Top' are as follows:

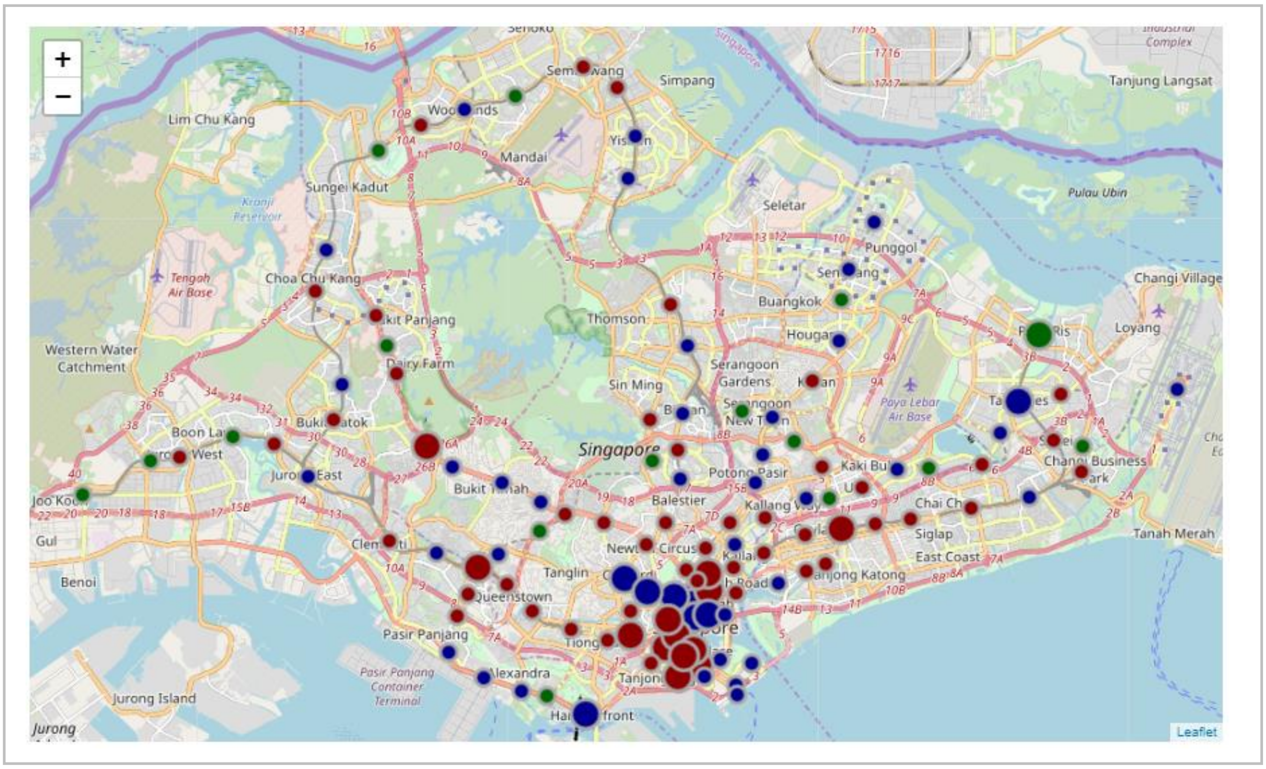


Station	Venue Density	Density Group
Havelock	86.0	Top
Telok Ayer	86.0	Top
Downtown	85.0	Top
Tanjong Pagar	83.0	Top
Raffles Place	81.0	Top

When the above results are aggregated and plotted onto a map of Singapore it becomes much more apparent the geographical distribution of the clusters. As explained by the key on the chart below, the colour of each marker corresponds to the cluster type:

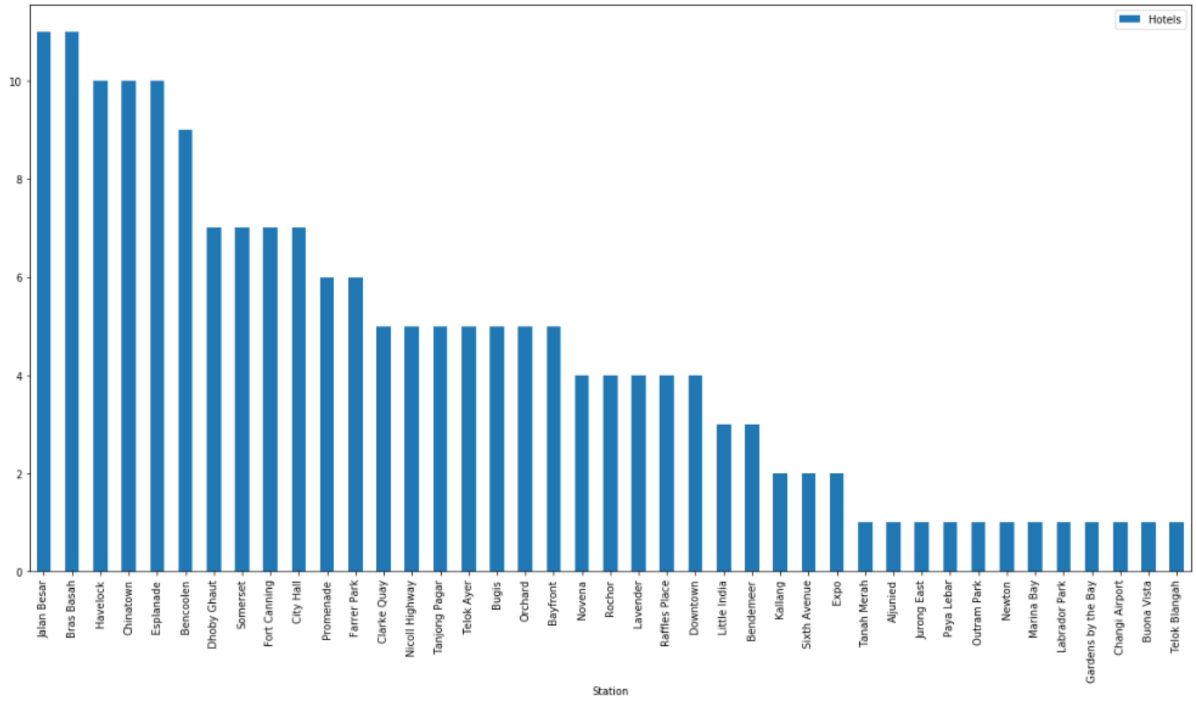


Further, it was possible to overlay the venue density tags (‘Top’ or ‘Other’) so that it is possible to see the location of those stations that have the highest density of venues for each cluster. In the map below, the top stations are designated by a large icon, the others with small icons.

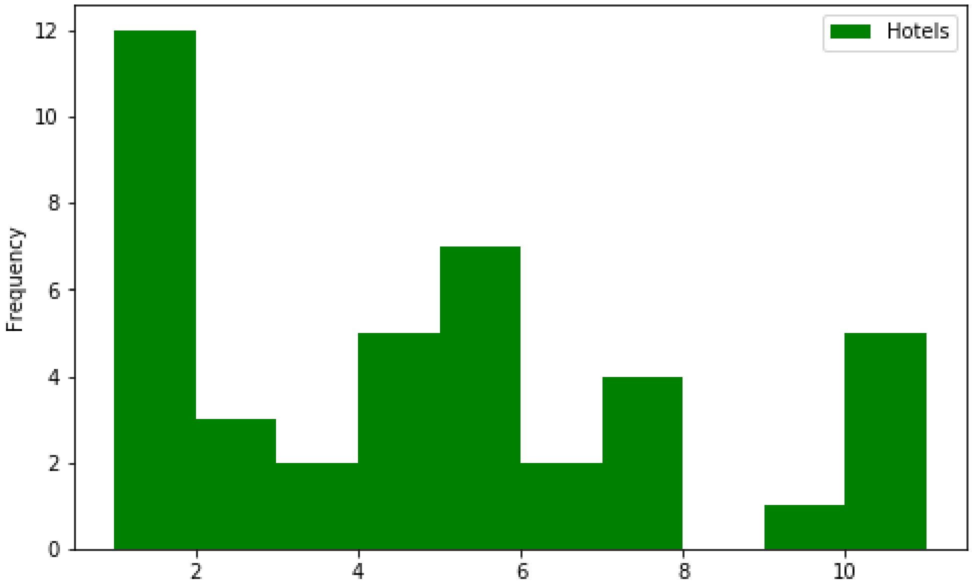


3.6 Hotel density overlay

For visiting tourists, it may also be useful to understand where hotels are located, and one way to assess this is by the density of hotels around each station. I again used a Dataframe to determine the count of hotels near each station and then plotted the results as follows.



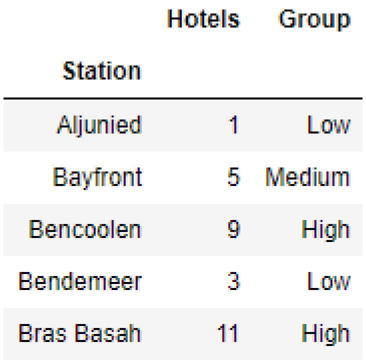
My intention was to group these, so I plotted a histogram as follows.



As can be seen in the above chart, there are 3 rough groupings of density which I have set out as follows:

- **Low:** 0-4 hotels nearby
- **Medium:** 5-8 hotels nearby
- **High:** 9-12 hotels nearby

Therefore, to further enrich the analysis I added these tags to the station data so that it could be overlaid to the resulting map of Singapore. A sample of results as follows:



	Hotels	Group
Station		
Aljunied	1	Low
Bayfront	5	Medium
Bencoolen	9	High
Bendemeer	3	Low
Bras Basah	11	High

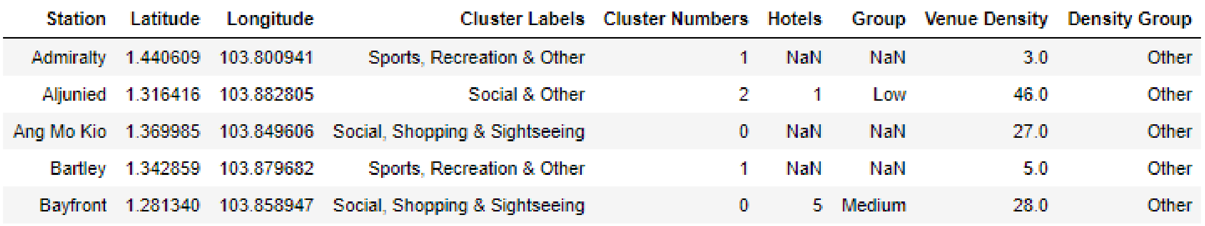
Above is a sample of stations that are now grouped (Low, Medium, High) according to the number of hotels nearby.

4. Results

Bringing this all together, the final dataset was set up to include the following data:

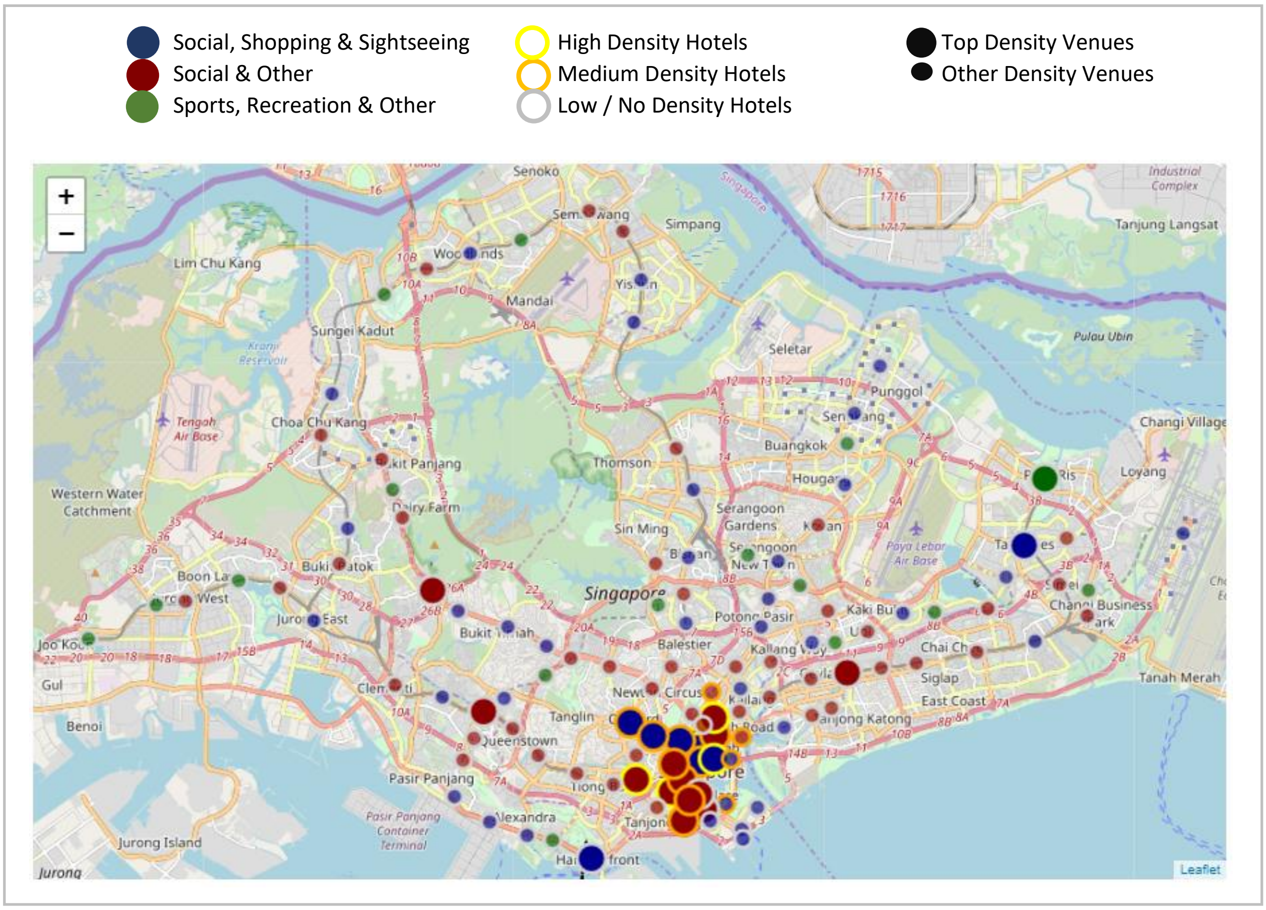
- **Station**: the original station names and latitudes along with the result of the analysis above, namely:
- **Cluster groups**: this is represented by the columns 'Cluster Labels' (e.g. Sports, Recreation & Other) and 'Cluster Numbers' (the numerical assignment of the clusters)
- **Hotel Density**: this is represented by the columns 'Hotels' (which is the number of hotels in the vicinity of the station) and 'Group' (which is the hotel density group i.e. Low/Medium/High)
- **Venue Density**: this is represented by the columns 'Venue Density' (number of venues near the station) and 'Density Group' which is the tag of 'Other / Top' for the purposes of highlighting those stations that are the best for a given activity group.

The final data now looks like:

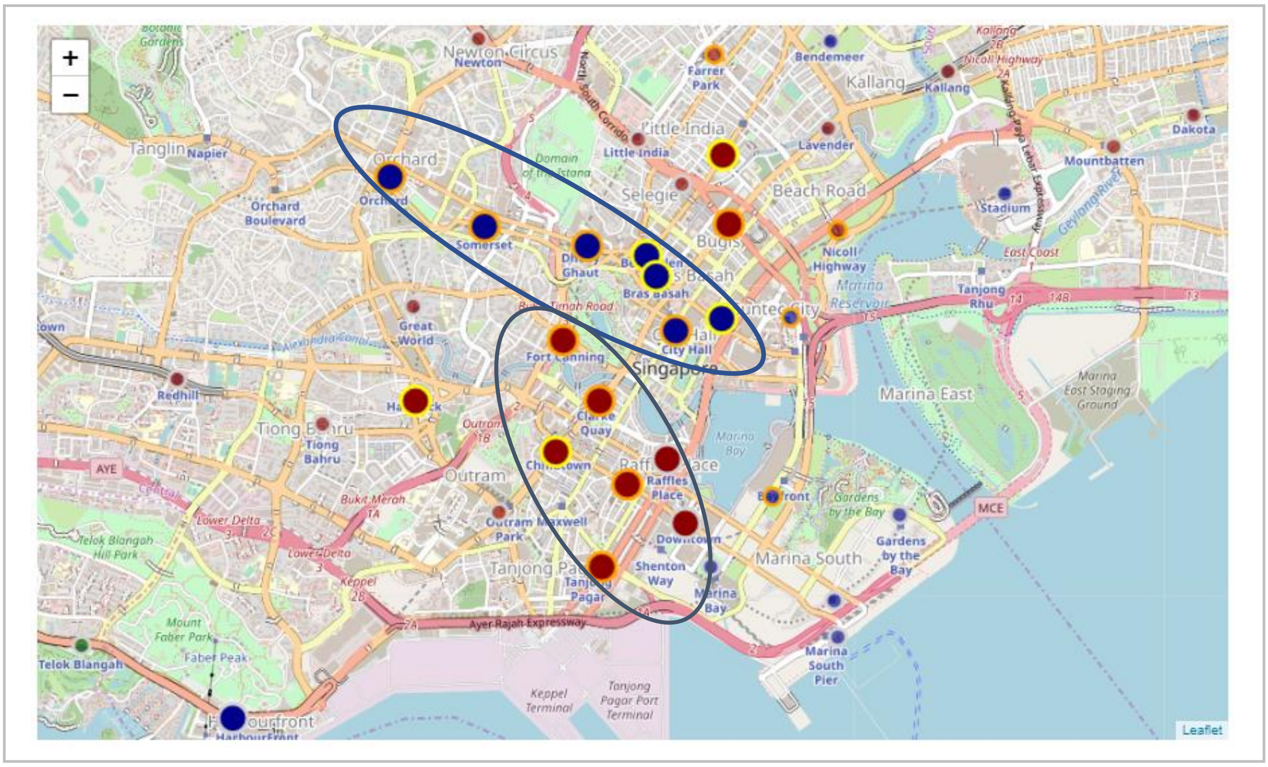


Station	Latitude	Longitude	Cluster Labels	Cluster Numbers	Hotels	Group	Venue Density	Density Group
Admiralty	1.440609	103.800941	Sports, Recreation & Other	1	NaN	NaN	3.0	Other
Aljunied	1.316416	103.882805	Social & Other	2	1	Low	46.0	Other
Ang Mo Kio	1.369985	103.849606	Social, Shopping & Sightseeing	0	NaN	NaN	27.0	Other
Bartley	1.342859	103.879682	Sports, Recreation & Other	1	NaN	NaN	5.0	Other
Bayfront	1.281340	103.858947	Social, Shopping & Sightseeing	0	5	Medium	28.0	Other

By overlaying this dataset onto a Folium map centred on Singapore, we can see visually where tourists may wish to visit/base themselves according to their interests and hotel needs:



Given the concentration of high-density stations in the South-Central it is useful to further zoom in on that area in order to see more clearly what is going on (see map below), which we will consider further in section 5 below.



5. Discussion

5.1 Which stations/areas are best?

It is interesting to see that the types of activities available across the island are well distributed (all activities are available to some degree across the island). However there appears to be a large number of high-density stations located in the South-Central area of Singapore, and in particular the South-Central region appears to contain:

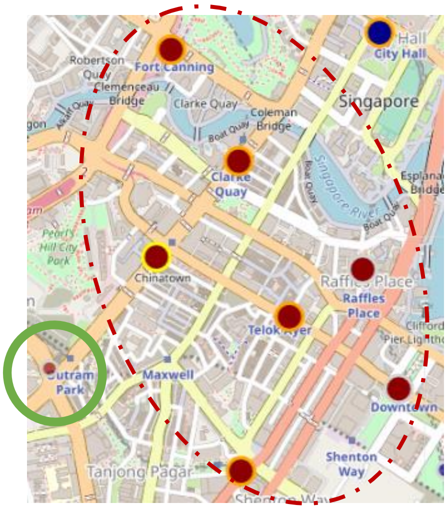
- All of the stations that are surrounded by a high / medium number of hotels
- A large number of 'Top' stations that are particularly good for people looking for Social activities as well as Shopping and Sightseeing

A further detailed look at this area (as seen in the more detailed map above) suggests that there are two clear primary clusters in the South-Central area that are good for tourists:

- The area from **Orchard to City Hall** (upper middle section of the map) is a particularly good area for people looking for **Social, Shopping and Sightseeing activities**.
- The area from **Fort Canning to Downtown / Tanjong Pagar** (lower middle section of the map) is a particularly good area for people looking for **Social activities**.

Having said this, as mentioned above there is a good distribution of activities across the island. Firstly, outside of the areas highlighted above there are other stations that are away from the obvious cluster pockets but still have a high density of venues e.g. Harbourfront (good for Social, Shopping & Sightseeing) in the far South or Pasir Ris (good for Sports & Recreation) in the far East. Therefore, for those tourists looking for places to stay that are away from the crowd but still lots to do, there are definitely options for them that are clear from the analysis in this project.

Further, some people may well be explicitly looking for areas in Singapore where it is possible to experience certain activities but with fewer people (or cheaper hotel prices), and this analysis can potentially help to identify where those places can be found e.g. Outram Park station:



As you can see from the map above, Outram Park is good for Social activities, and is very near to Chinatown / Clarke Quay etc (which are extremely good for Social activities). However, Outram Park itself is not a high-density area in terms of either activities or hotels – for some people this may be a good 'fringe' area where it is possible to enjoy nearby activities without being right in the centre of the action.

5.2 Hotel density

This analysis has demonstrated those areas where hotel density is high, which makes it possible for tourists to select a station/area that meets their needs in terms of both of activity and hotel availability, for example Chinatown has a high number of hotels nearby and is also one of the best stations/areas in Singapore for enjoying Social activities.

Conversely, and as eluded to in the introduction, this could be useful to hotel business owners for identifying areas which have a high density of venues but low number of hotels in the vicinity; for example Harbourfront is densely packed with venues, is particularly good for tourists looking for Social, Shopping & Sightseeing, however the number of hotels nearby is low. This may represent an opportunity worth exploring.

5.3 Further opportunities for analysis

There are certainly multiple opportunities to build upon this analysis, for example:

- It could be interesting to break down the 'Social' category to give more context to sub-activities e.g. Restaurants vs. Bars vs. Clubs.
- Further, it would be interesting to consider the density of different restaurant / cuisine types (e.g. local Singaporean vs. Italian)
- Alternatively, it would be interesting to consider hotel prices to see best value areas
- It would even be possible to widen the objective to other people e.g. residents of Singapore; where is best to live for different job types?

6. Conclusions

The basis of this project was to identify those stations that would be best suited to tourists, and looking at the results it is fair to say that the analysis has been successful.

Activities are generally well distributed across the island (all activities are available to some degree in each region). However, the South-Central region is clearly where the action is at, particularly with respect to Social activities as well as Shopping and Sightseeing. The South-Central is also where tourists should look for a large choice of hotels.

For Sports and Recreation activities there appears to be options across the island (Pasir Ris being the best) although being located in the South-Central region may still be the best choice for access to Hotels and other activities.

It is clear there are many ways to expand upon the analysis here, either by improving the data / analysis itself or by widening / modifying the objective to suit other needs.