Coursera Capstone project

Classifying MRT stations in Singapore to assist tourists with their visit

Neil L

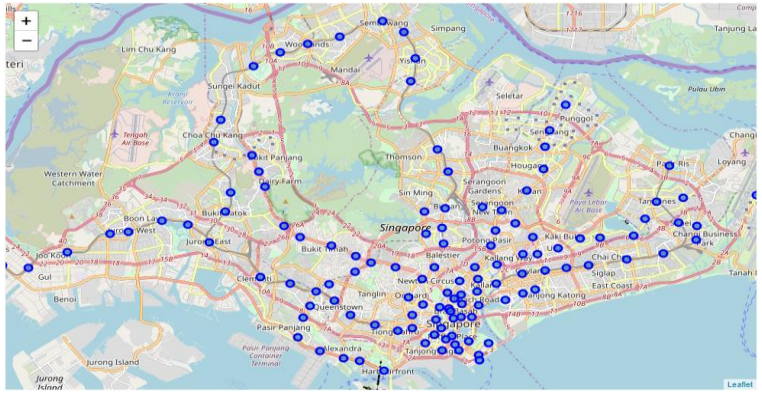
15 February 2020

Singapore has lots to offer to tourists, choosing a good base is essential

- Singapore's weather is humid and hot, especially for tourists
- Singapore's MRT system is modern, clean, reliable and well connected
- MRT stations are typically surrounded by a number of venues
- Being based close to an MRT is essential to maximise the visit
- However not all MRT stations are made equal
- Tourists will benefit from knowing which MRT stations best support their needs:
  - Types of activities being sought whilst in Singapore
  - Prevalence of hotel options nearby

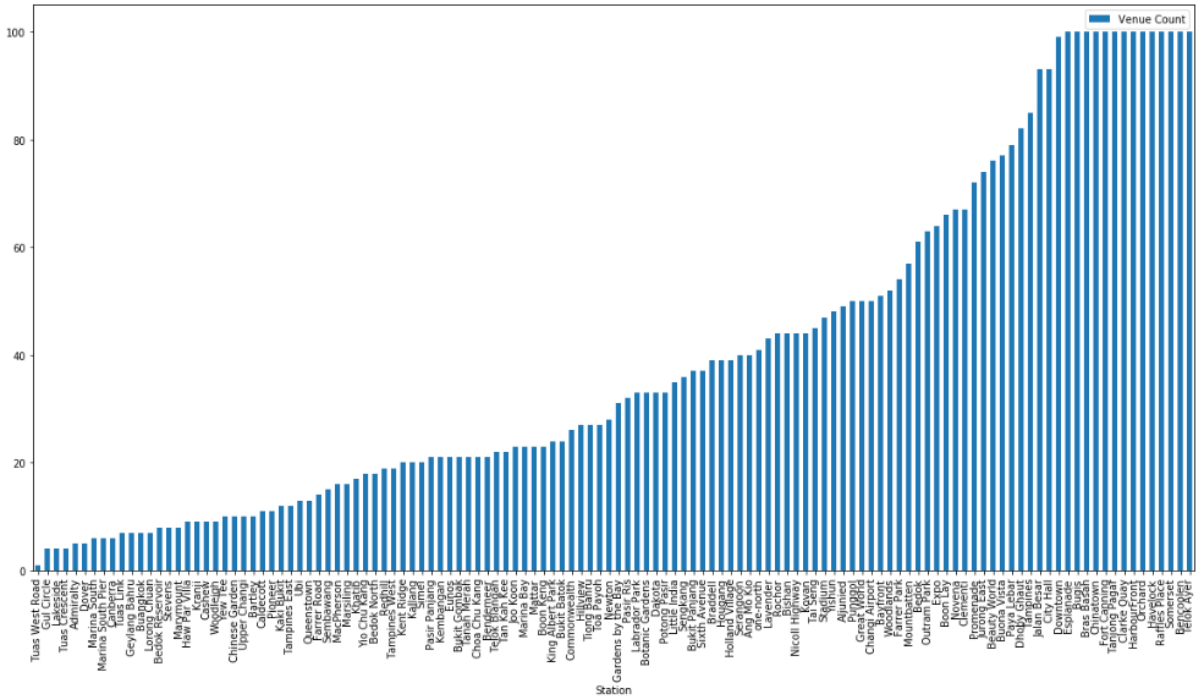
Data acquisition and prep

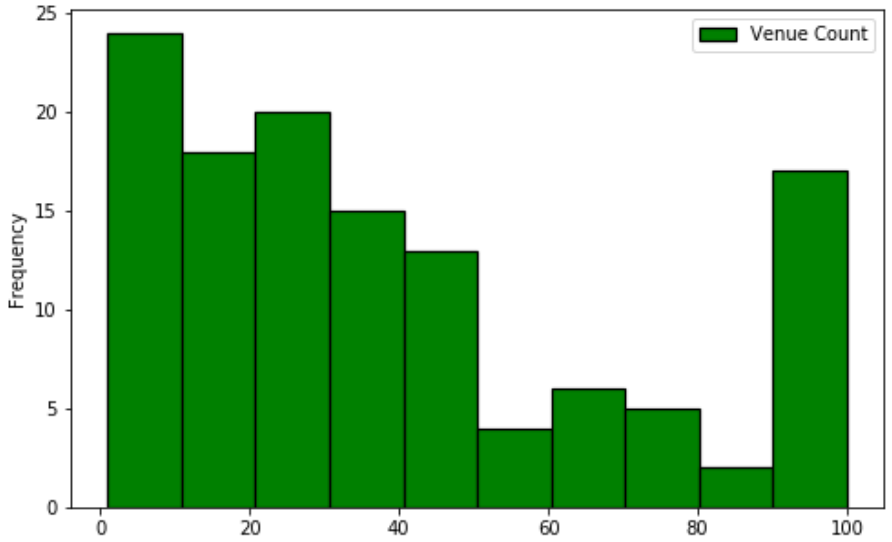
- Singapore MRT station list as at January 2020, scraped from Wikipedia
  - 124 stations in active use as at January 2020
  - Data cleaned for duplicates and stations not yet active
- Geo-location data obtained from Google Maps
- Venue information obtained from Foursquare via API
  - ~5,000 unique venues
  - 325 categories (music store, Italian restaurant etc)



Venue distribution by Station

Venue density varies across the 124 stations

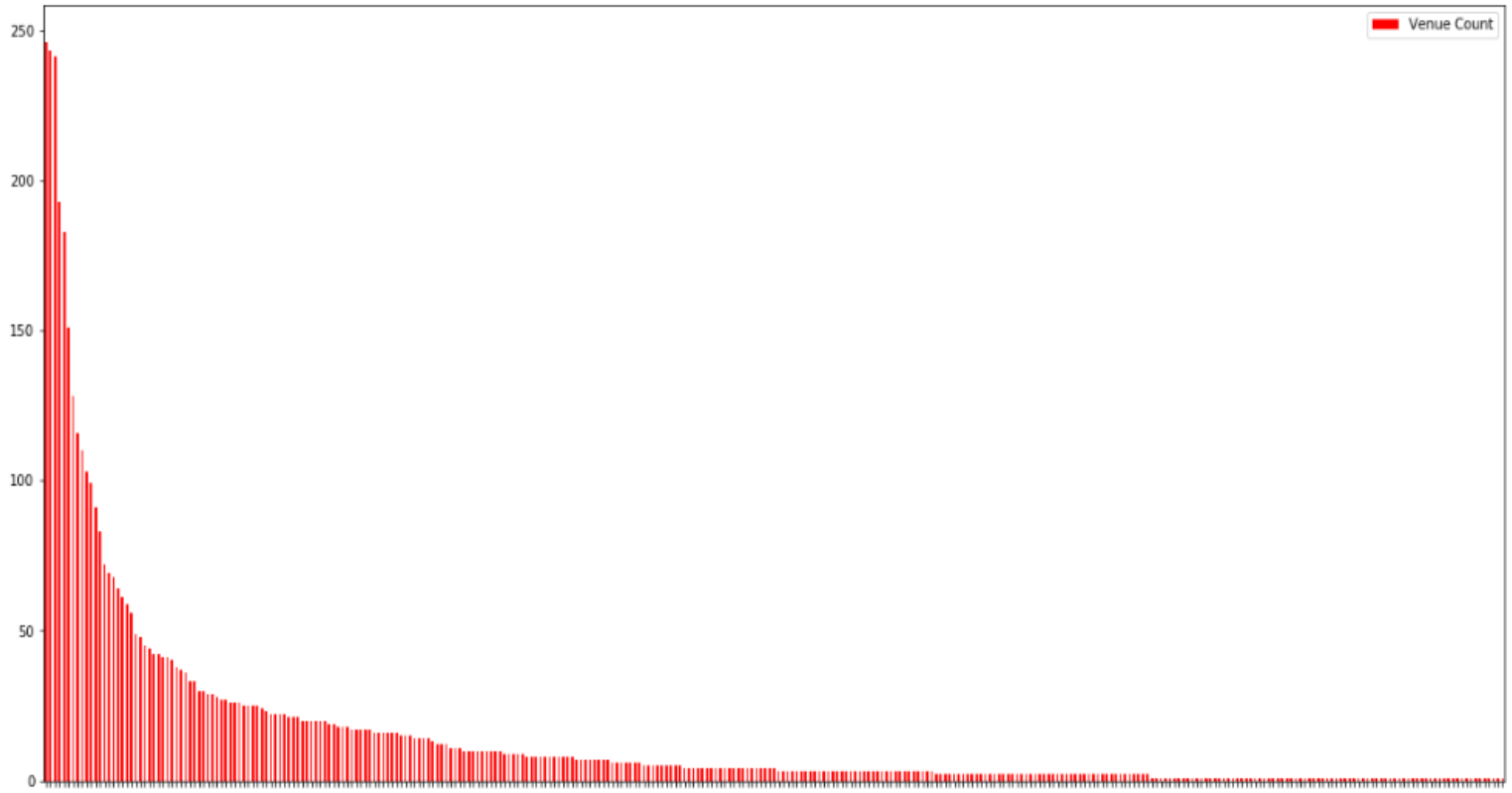




But typically stations support either modest (<50) or a high (100+) number of venues

Venue distribution by Category

325 different categories, there is large concentration of venues around certain categories, and a very long tail



Categories by venue count

Top 20

Venue Category	Venue Count
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

Bottom 20

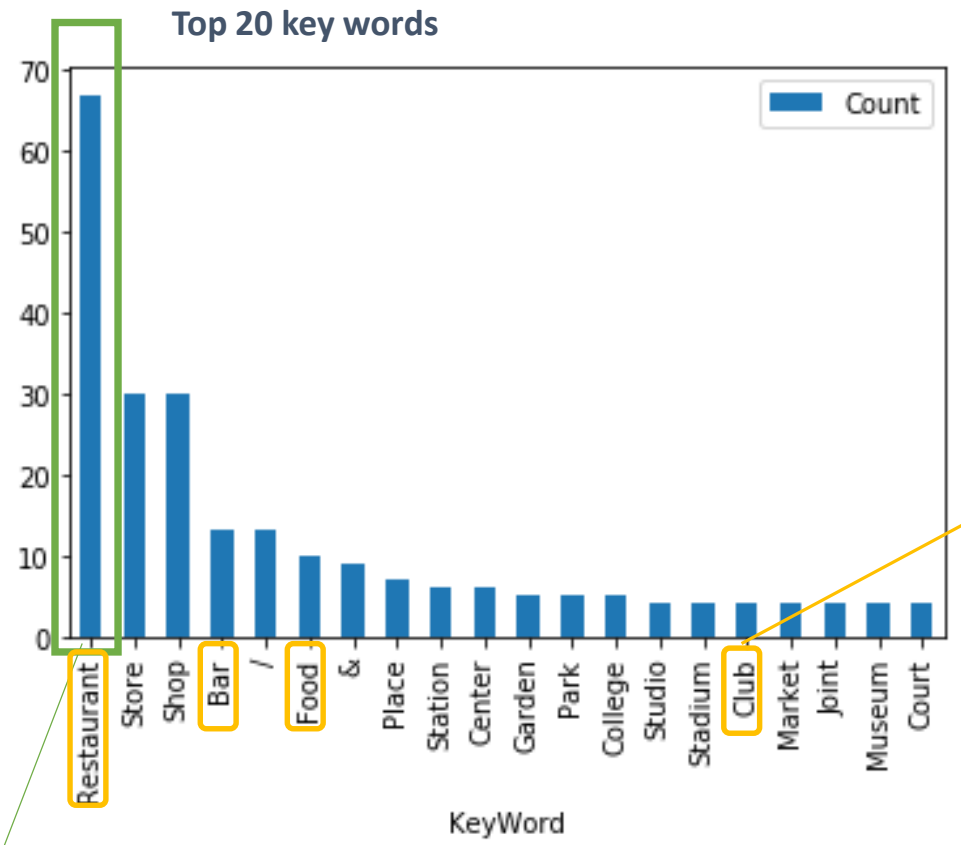
Venue Category	Venue Count
Pastry Shop	1
Music Store	1
Nail Salon	1
Night Market	1
Non-Profit	1
Churrascaria	1
Church	1
Other Nightlife	1
Outdoor Sculpture	1
Chinese Aristocrat Restaurant	1
Cha Chaan Teng	1
Outdoor Supply Store	1
Outlet Store	1
Cafeteria	1
Food & Drink Shop	1
Pedestrian Plaza	1
Persian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

It is clear from the data that a number of categories are a variation on a similar theme e.g. restaurants

This presents a challenge for meaningful classification

Using key words to define broad category groups

By breaking venue category names down into component words, it was possible to identify the frequency of certain words appearing.



The keyword 'Restaurant' is by far the most common, appearing in almost 70 different venue category names

There is merit in aggregating certain words to form more coherent groupings (e.g. Restaurant, Bar, Food, Club etc.. could all be aggregated)

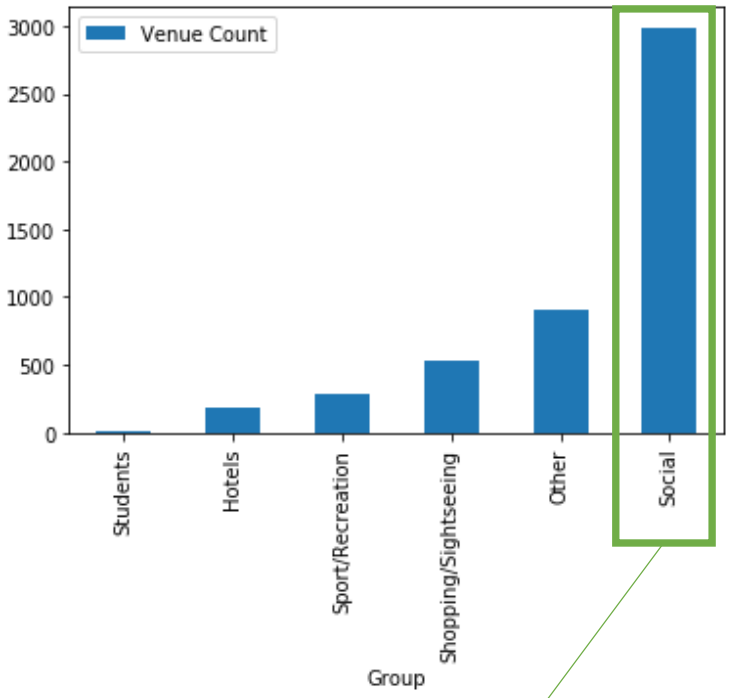
I created a set of major groups to which each venue category would be broadly aligned

- ✓ **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.

Clustering of new category groups

With the aggregated category groups it is much easier to see the density of venues and what purposes they serve.

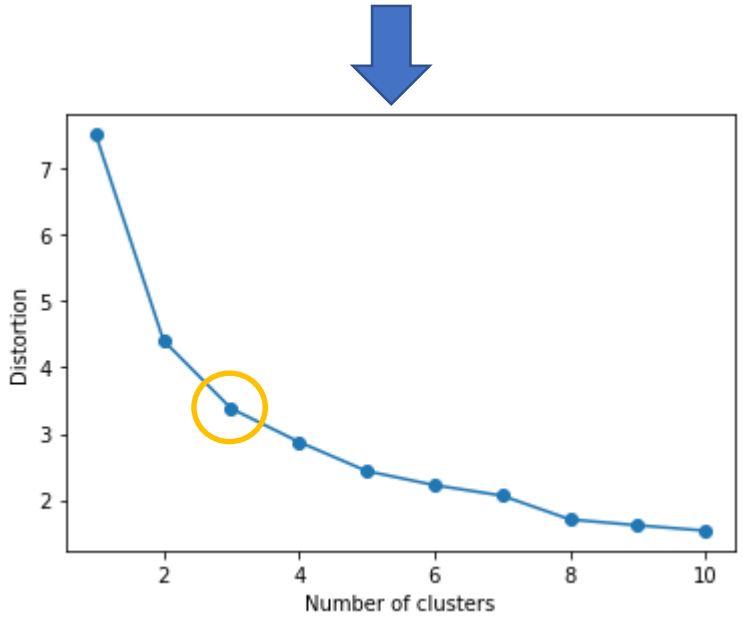


Social venues are the most popular in Singapore and tourists looking for this type of activity are very well served

It is noted that with such a high number of instances in 'Social', an extension of this project could be to break this down into further sub-groups for analysis

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

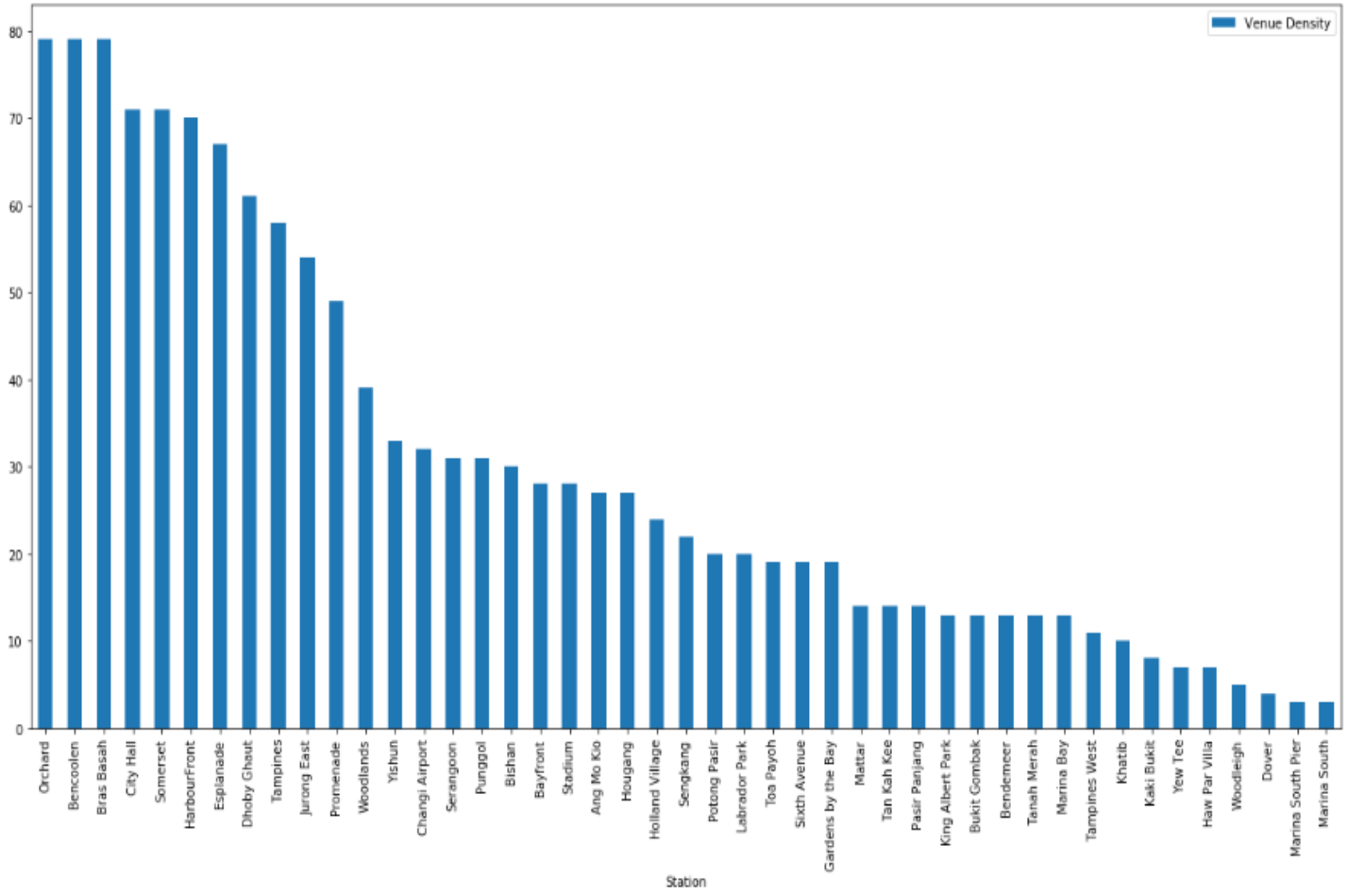
Further, because of its unsupervised nature I utilised the elbow method to determine an optimum number of clusters for the algorithm to solve for.



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.

Cluster 1 results

By assessing the results of the clustering I was able to label cluster 1 as **Social, Shopping & Sightseeing**



I further categorized the stations by highlighting those that had the most instances of venues relevant to the cluster

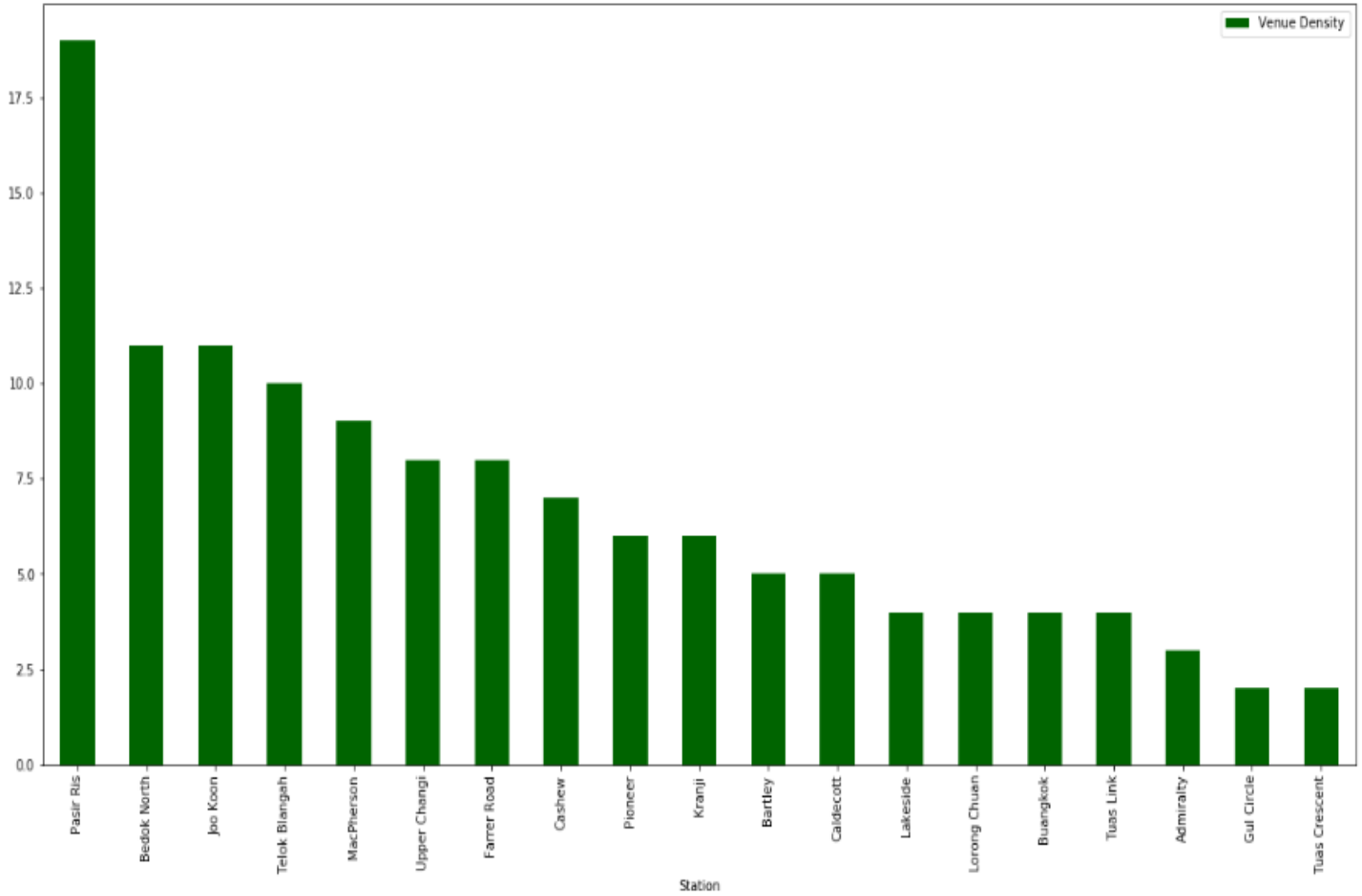
Station	Venue Density	Density Group
Orchard	79.0	Top
Bencoolen	79.0	Top
Bras Basah	79.0	Top
City Hall	71.0	Top
Somerset	71.0	Top
HarbourFront	70.0	Top
Esplanade	67.0	Top
Dhoby Ghaut	61.0	Top
Tampines	58.0	Top

Those stations with a venue count > greater than 70% of the maximum in this category (79), were tagged as ‘Top’ stations

(Top stations shown above)

Cluster 2 results

By assessing the results of the clustering I was able to label cluster 2 as **Sports, Recreation & Other**



I further categorized the stations by highlighting those that had the most instances of venues relevant to the cluster

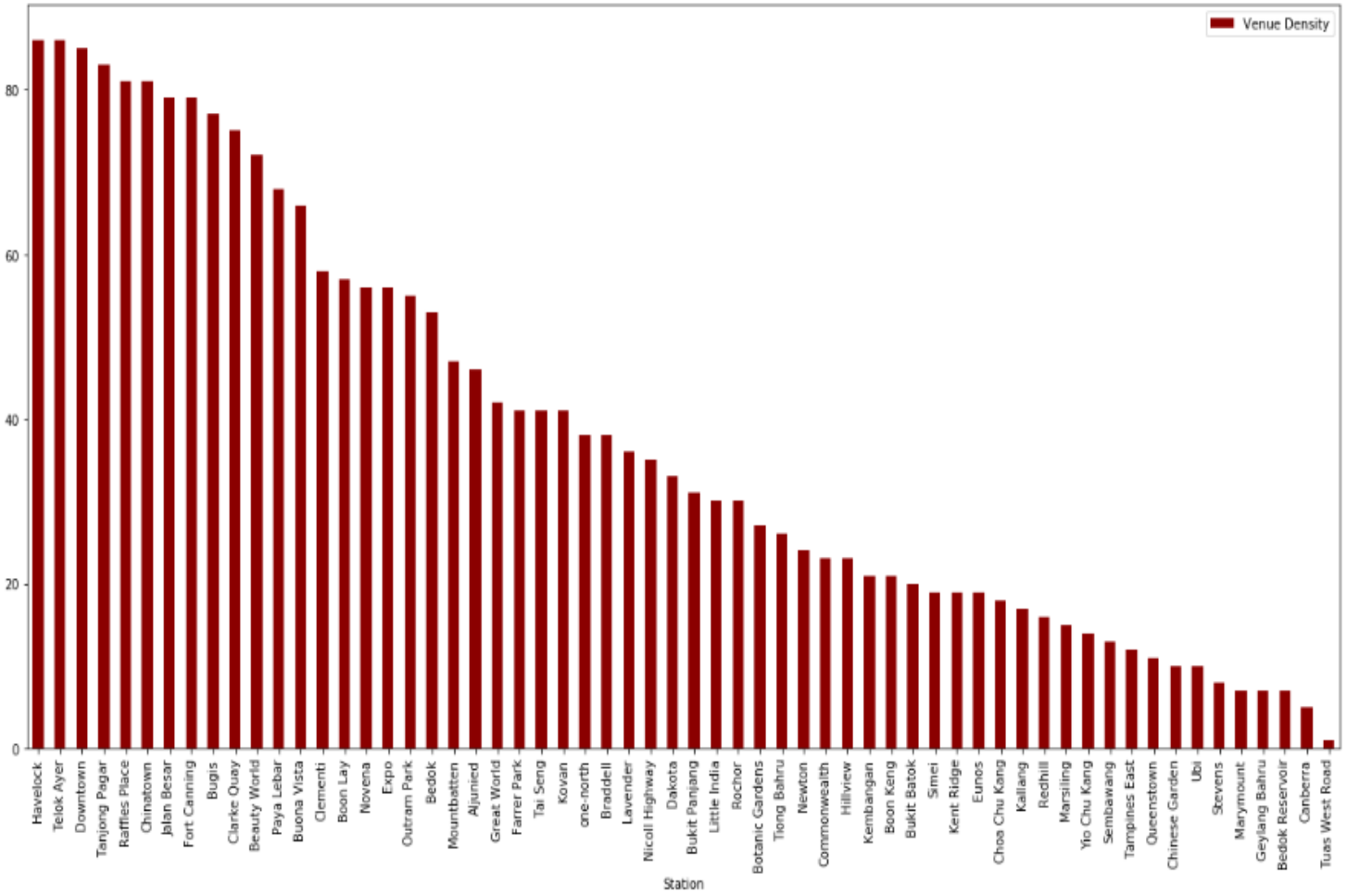
Station	Venue Density	Density Group
Pasir Ris	19.0	Top

Those stations with a venue count > greater than 70% of the maximum in this category (19), were tagged as ‘Top’ stations

(Only Pasir Ris in this category!)

Cluster 3 results

By assessing the results of the clustering I was able to label cluster 3 as **Social & Other**



I further categorized the stations by highlighting those that had the most instances of venues relevant to the cluster

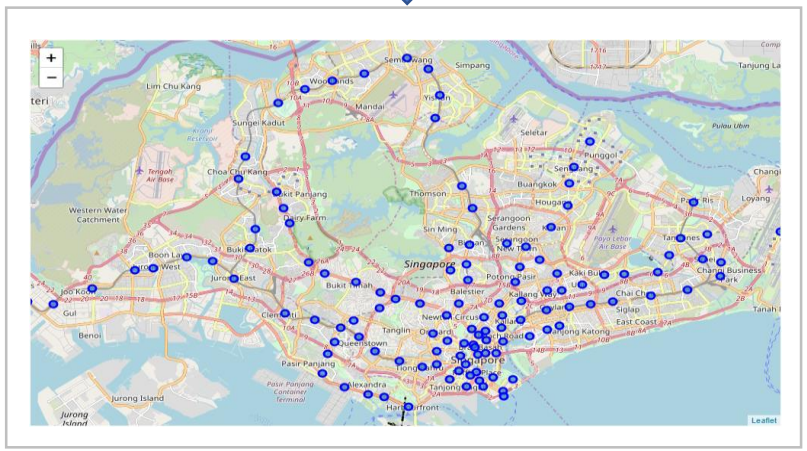
Station	Venue Density	Density Group
Havelock	86.0	Top
Telok Ayer	86.0	Top
Downtown	85.0	Top
Tanjong Pagar	83.0	Top
Raffles Place	81.0	Top

Those stations with a venue count > greater than 70% of the maximum in this category (86), were tagged as ‘Top’ stations

(Top stations shown above)

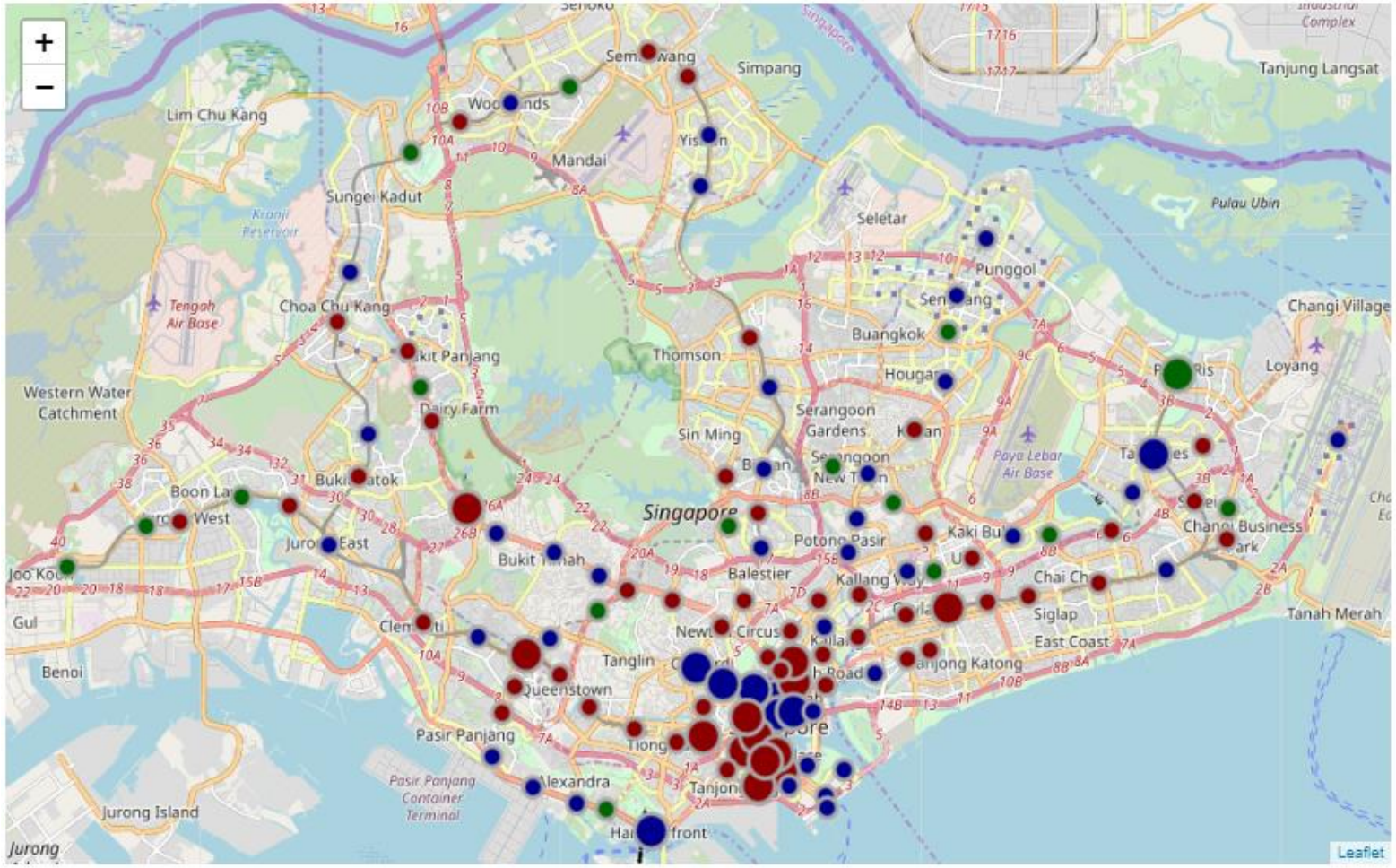
Clustering results

I overlaid the station geo-location data onto a Folium map of Singapore



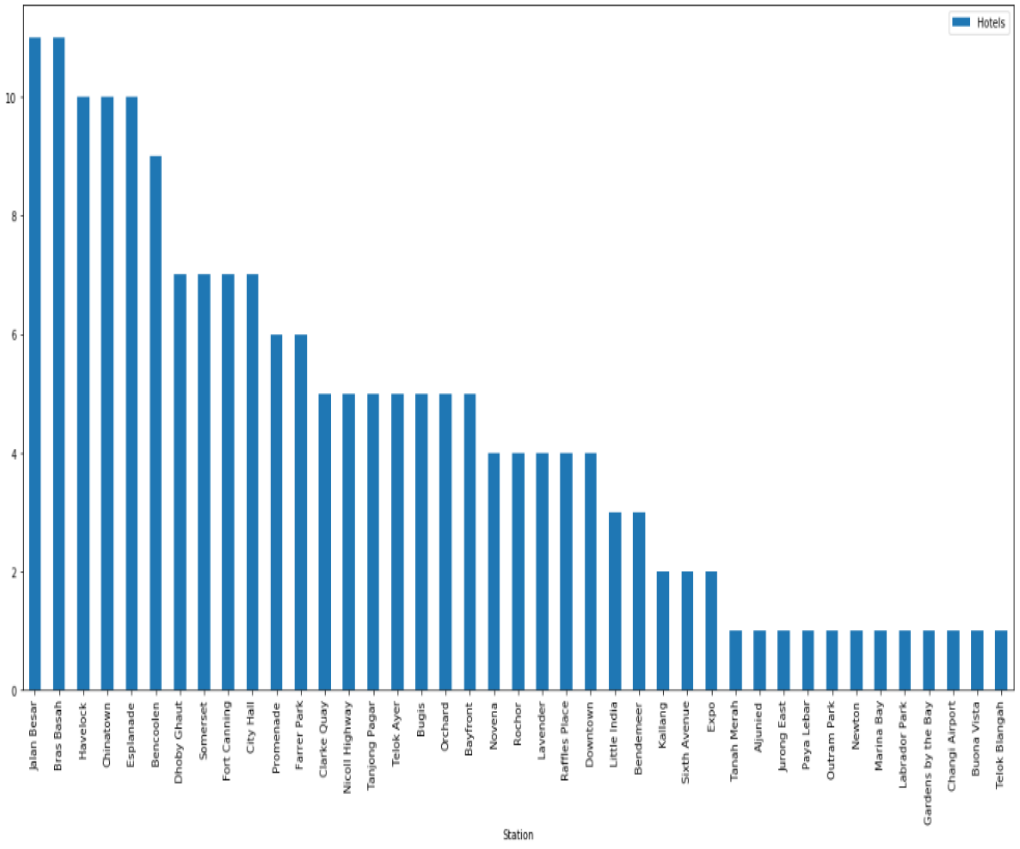
And then further enhanced the markers using:

- ✓ Cluster tags
- ✓ Venue Density tags

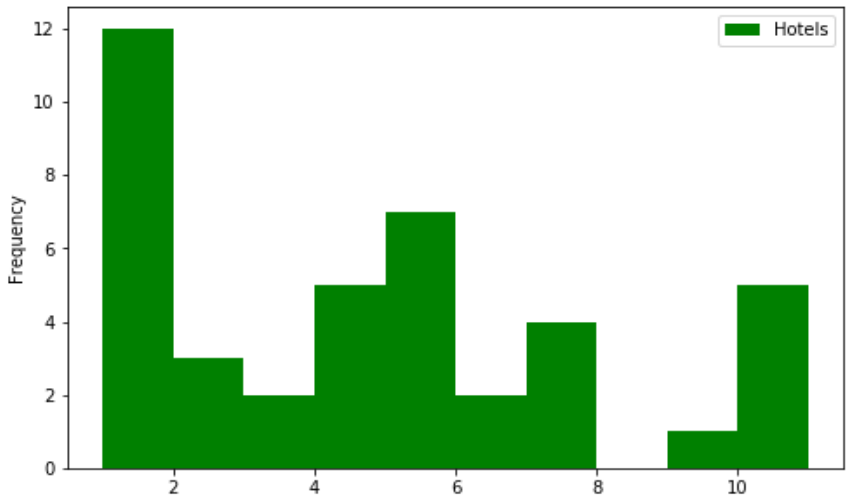


For visiting tourists, it may also be useful to understand where hotels are located

I used a Dataframe to determine the count of hotels near each station:



I further categorized the stations by highlighting those that had the most instances of venues relevant to the cluster



Based on the results above I grouped the hotels according to:

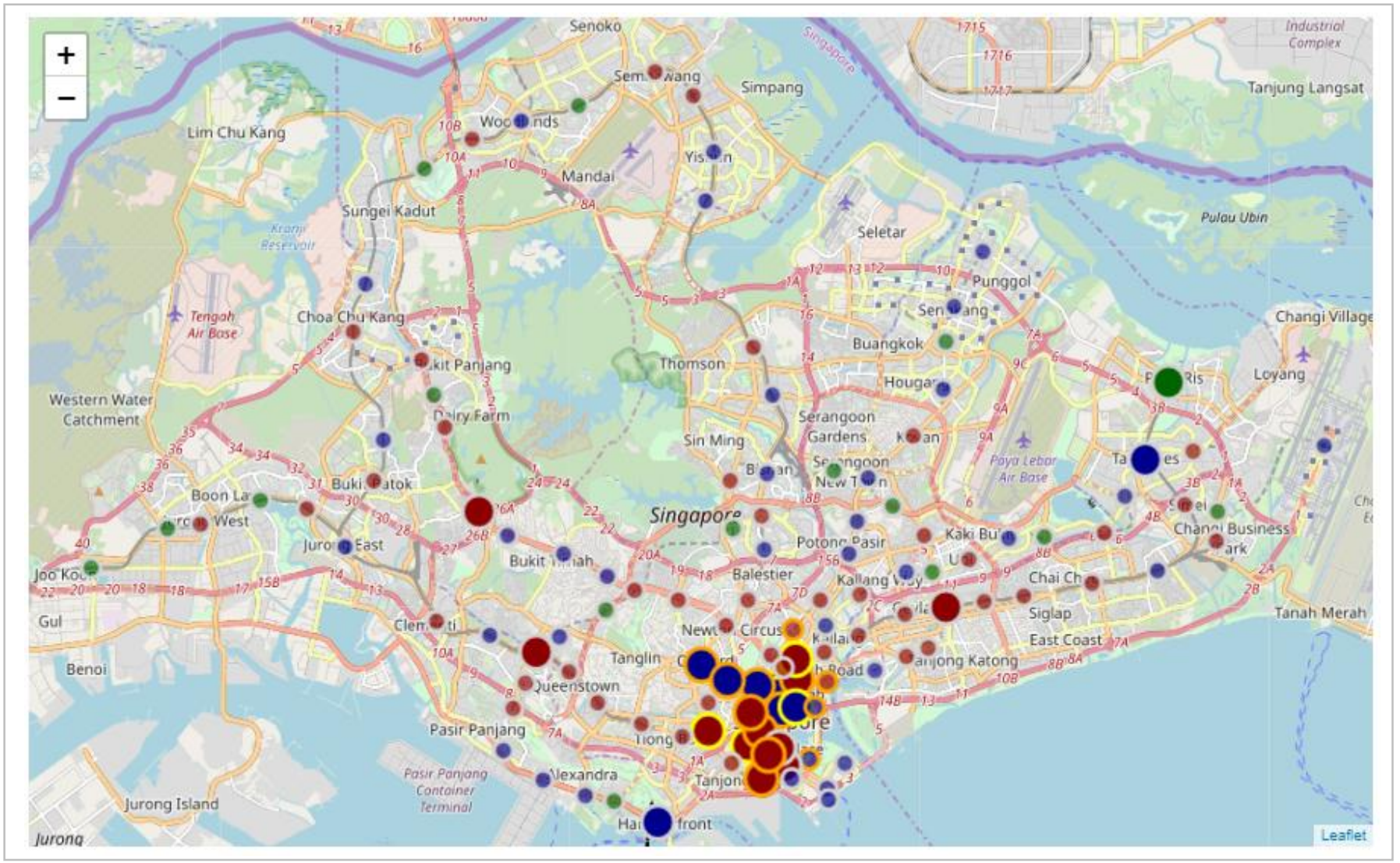
Low: 0-4 hotels nearby

Medium: 5-8 hotels nearby

High: 9-12 hotels nearby

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

Final results mapped onto Singapore

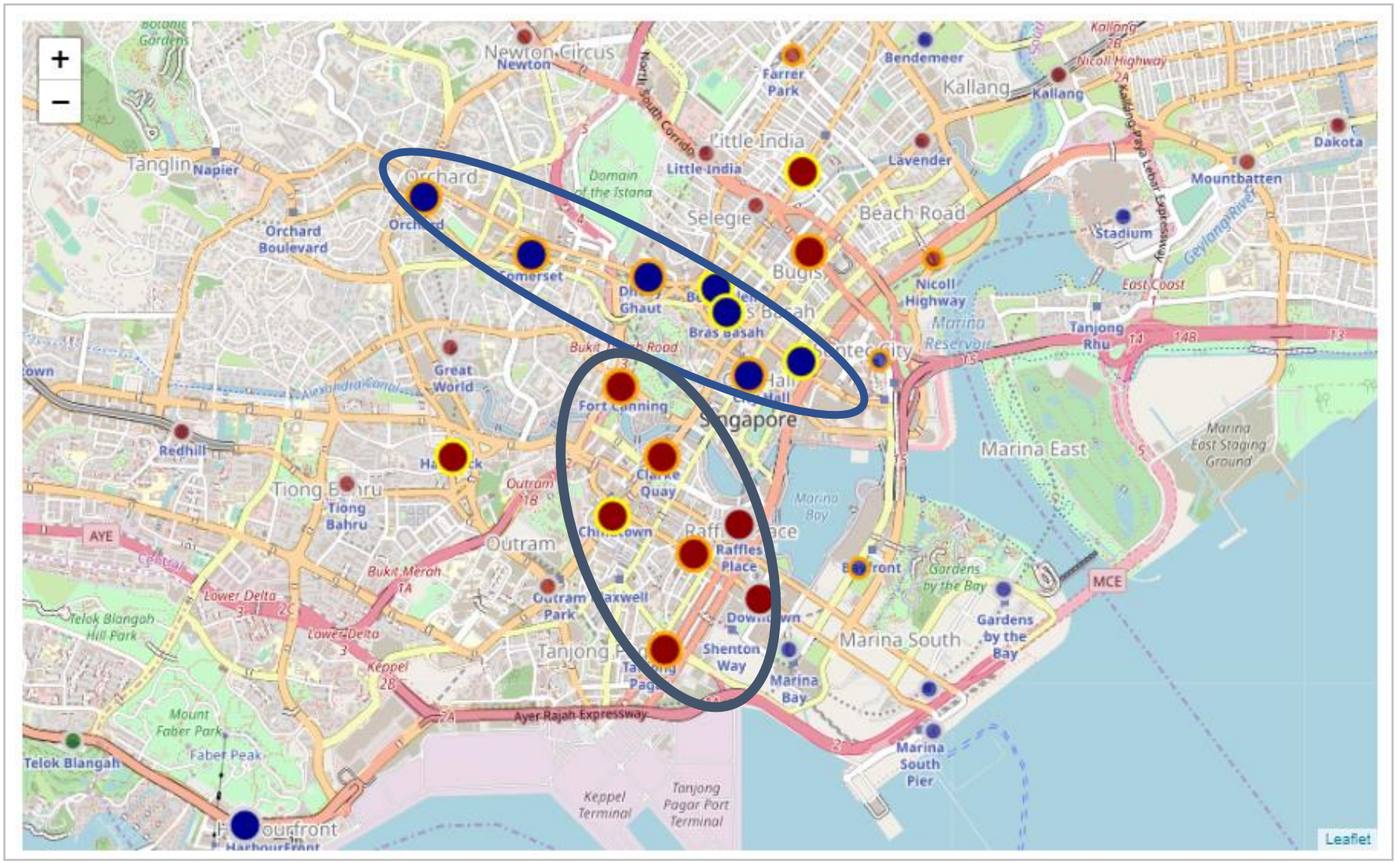


Social, Shopping & Sightseeing
Social & Other
Sports, Recreation & Other

High Density Hotels
Medium Density Hotels
Low / No Density Hotels

Top Density Venues
Other Density Venues

The Central-South region has a large number of high density stations, with two clear primary clusters

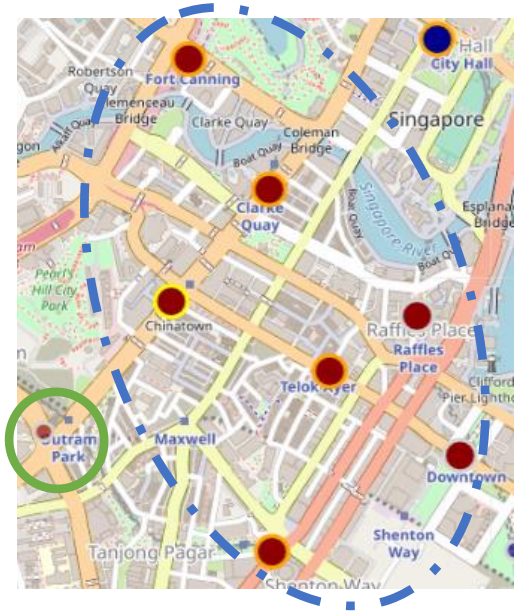


Social, Shopping & Sightseeing
Social & Other
Sports, Recreation & Other

High Density Hotels
Medium Density Hotels
Low / No Density Hotels

Top Density Venues
Other Density Venues

‘Fringe’ areas also becoming apparent e.g. Outram Park



Results and Discussion

- ✓ The types of activities available across the island are **well distributed** (all activities are available to some degree across the island)
- ✓ Number of high-density stations located in the **South-Central area of Singapore**, and in particular:
  - 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
- ✓ 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
- ✓ 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, for example:
  - Harbourfront (good for Social, Shopping & Sightseeing)
  - Pasir Ris (good for Sports & Recreation)
- ✓ It is possible to identify **'Fringe' areas** where it is possible to enjoy nearby activities without being right in the centre of the action
  - For Example Outram Park which is very close to China Town, Clarke Quay etc.., but itself is low density

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, however, the South-Central region is clearly where the action is at, with two clear primary clusters
  - **Orchard to City Hall** for Social, Shopping and Sightseeing activities
  - **Fort Canning to Downtown / Tanjong Pagar** for Social activities
- ✓ For Sports and Recreation activities there appears to be options across the island (Pasir Ris being the best)
- ✓ Being located in the South-Central region be the best choice in all cases for access to Hotels
- ✓ However there are opportunities to stay away from the crowds, or take advantage of 'Fringe' areas if so desired
- ✓ This analysis could be used by other interested parties e.g. hotel business owners looking for opportunities
- ✓ 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. For example:
  - Break down the 'Social' category to give more context to sub-activities e.g. Restaurants vs. Bars vs. Clubs
  - consider the density of different restaurant / cuisine types (e.g. local Singaporean vs. Italian)