Venue Location Analysis of Singapore



Lars Mueller

December 29, 2020

**Table of Contents:**

1. Introduction 2

1.1 Background 2

1.2 Problem Statement 2

2. Data 3

2.1 Data Sources 3

2.2 Data Acquisition and Wrangling 3

2.2.1 Neighbourhood Areas 3

2.2.2 Coordinates for each Neighbourhood 4

2.2.3 100 Most Recommended Venues within each Neighbourhood 5

3. Data Analysis 7

3.1 Exploratory Data Analysis 7

3.2 Clustering 9

3.2.1 K-Means Clustering 9

3.2.2 DBSCAN Clustering 10

4. Results 11

4.1 DBSCAN Vs. K-Means Clusters 11

Cluster 0 – Parks 11

Cluster 1 – Japanese Restaurants 12

Cluster 2 – Coffee Shops 12

4.2 Cluster Visualisation 13

4.3 Interactive Folium Grouping Visualisation 14

5. Discussion 14

6. Conclusion 15

7. References 16

# 1. Introduction

## Background

Located in South East Asia, Singapore has an area of 728km2 with a population of 5.64 million and growing (Plecher, 2020). It is one of the world’s most prosperous nations with an annual GDP of USD 579 billion in 2018 (Plecher, 2020). It is no surprise that tourism is a major contributor to Singapore’s economy; attracting 18.5 million visitors and totaling SGD 27.1 billion in 2018 ("Singapore Tourism Statistics 2020", 2020).

As Singapore is the world leader in infrastructure and continues to invest heavily in new building projects (Brown, 2020), there is an overwhelming amount of activities, and places to explore as a tourist.

In addition to many parks, reservoirs, landmarks, and museums, Singapore is also South East Asia’s leading food and nutrition hub (Scattergood, 2020) with over 4.5 thousand restaurants in 2018 (Hirschmann, 2020).

Writing blogs has become increasingly popular, with approximately 409 million people viewing more than 20 billion pages each month. The average travel blogger charges an average of $200 per sponsored post and 94% of travel blogs sell advertising. However, it is becoming more difficult to get traffic via Facebook and Google due to increased saturation in the market ((Ouellette, 2020).

## Problem Statement

With the hopes of Singapore opening its boarders to tourists in 2021, a friend of mine is wanting to travel to Singapore with the intention of expanding her travel blog. The focus of her travel blog will be on the following five categories; coffee shops, scenic lookouts, Japanese restaurants, parks, and bubble tea shops.

She is hoping to visit as many venues and places within those categories as possible to write more balanced and contrasted blog posts about her experiences. Never having been to Singapore before, she is seeking guidance on where to go for those specific venue types.

Although Singapore is small in size, aimlessly wandering around to find places to review will likely cause a significant amount of time to be lost during her one week stay. She does not want to base her itinerary on other blogs or articles, as she worries that it will cause her posts to be less original.

# 2. Data

## 2.1 Data Sources

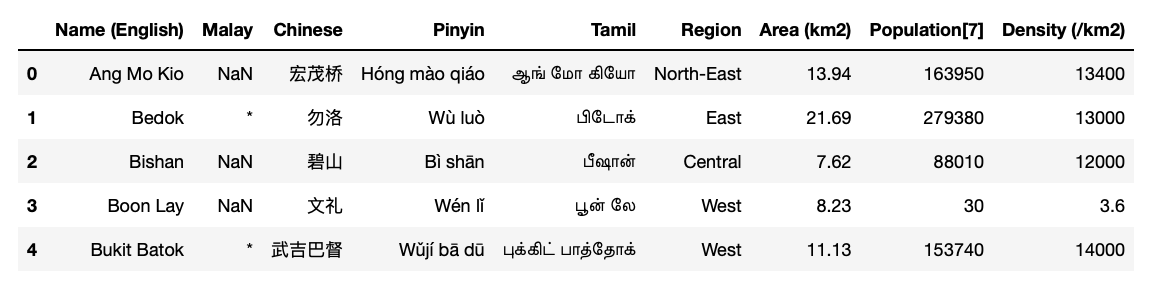
Data used for this report was gathered form a variety of sources as listed below. Once retrieved, it was loaded into a Pandas data frame for cleaning and processing.

1. The list of different neighbourhoods in Singapore is scraped from a Wikipedia article found at <http://en.wikipedia.org/wiki/Planning_Areas_of_Singapore> using Beautiful-soup. This source was used solely to retrieve the different neighbourhoods in Singapore as identified by the Urban Redevelopment Authority of Singapore.
2. The coordinates for each neighbourhood were obtained via the geocoding web-service. This data was used to create markers for each neighbourhood in Singapore.
3. The Foursquare API was used to retrieve data for the 100 most popular venues within a 2.5km radius of each neighbourhood marker.

## 2.2 Data Acquisition and Wrangling

### 2.2.1 Neighbourhood Areas

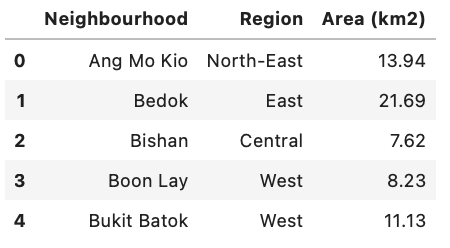
The pd.read.html() method from the Pandas library was used to read the data from the Wikipedia website into the following data frame:



As can be seen that there are many columns that are not relevant to this report. In fact, only the following columns were required;

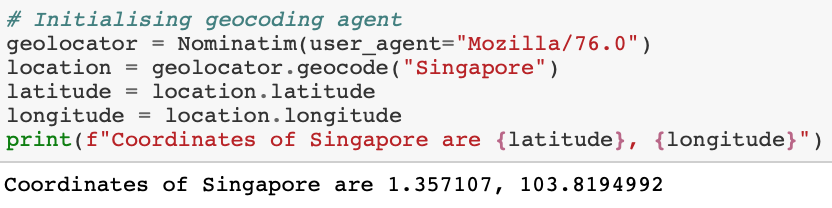
* Neighbourhood (renamed from ‘Name (English)’)
* Region
* Area (km2)

This returned the below data frame, with only the relevant information included.

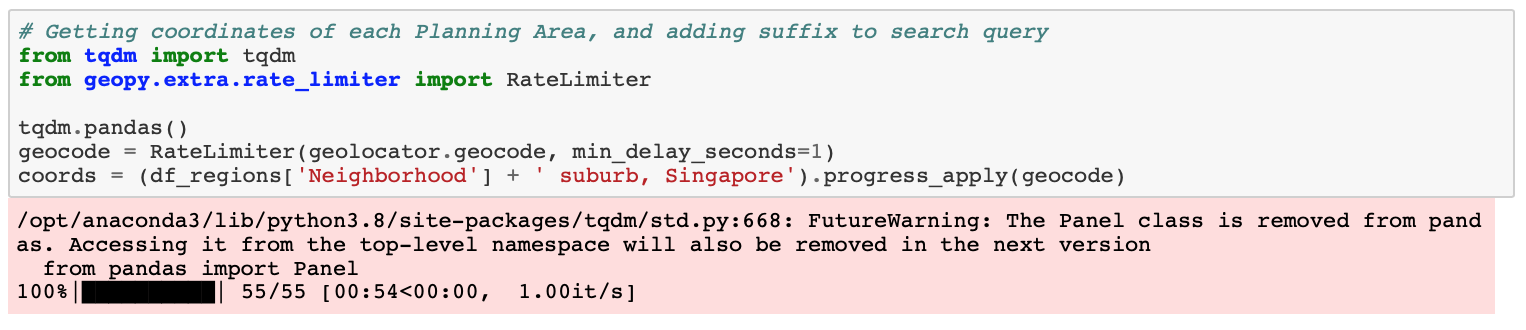


### 2.2.2 Coordinates for each Neighbourhood

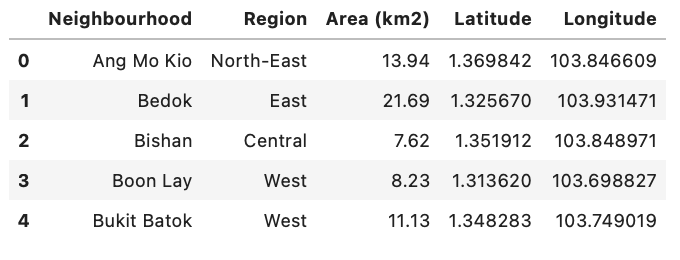
Geocoding is the process of converting addresses or locations into geographical coordinates (Longitude and Latitude). Geocoding was used to obtain the coordinates of Singapore, which is the starting point for the Folium map visualisations used.



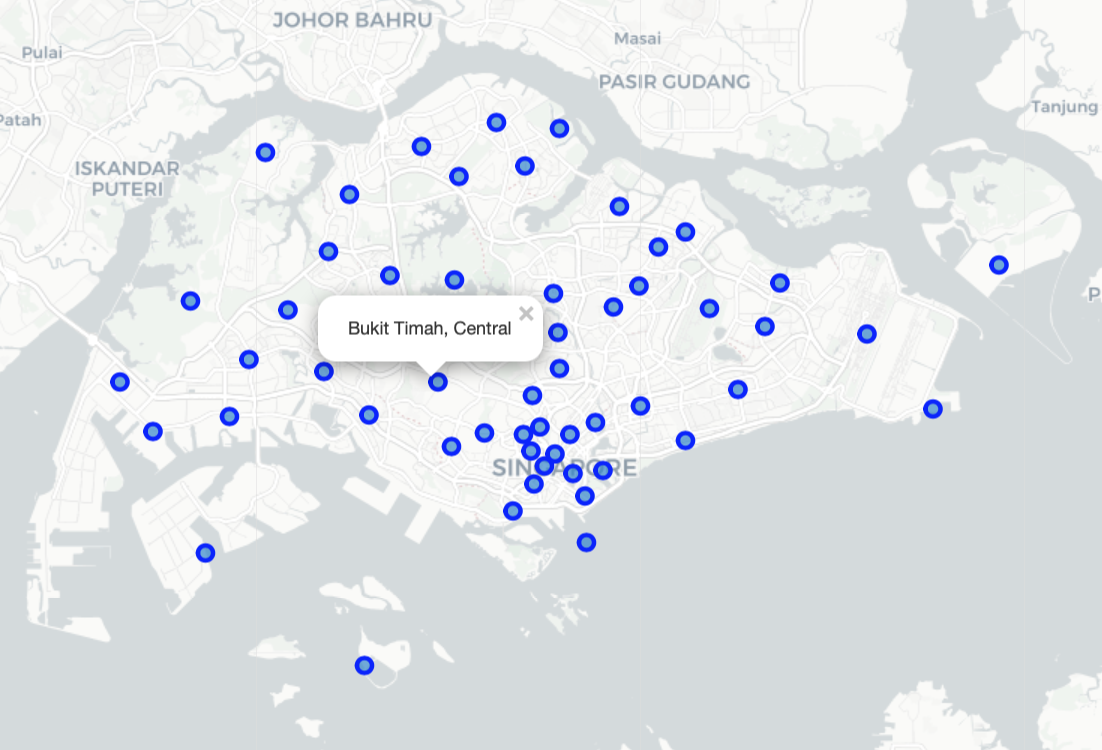
The coordinates for each neighbourhood were obtained via the same method. However, as too many geocoding requests may overload the service, the RateLimiter() method is used to add a 1 second delay between the calls made to the geocoding service. The tqdm() function is used to display a progress bar.



The coordinates retrieved for each neighbourhood were then added to the data frame to allow for further processing. The first five results are shown below.



Using the Folium library and the above data frame, each neighbourhood was plotted as below. The look of the map has been adjusted to ‘CartoDB positron’ as it provides a nicer visual with more contrast.



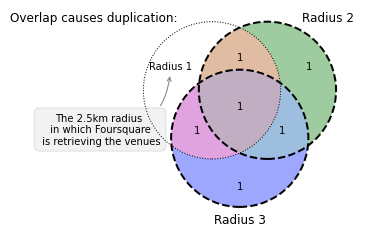
### 2.2.3 100 Most Recommended Venues within each Neighbourhood

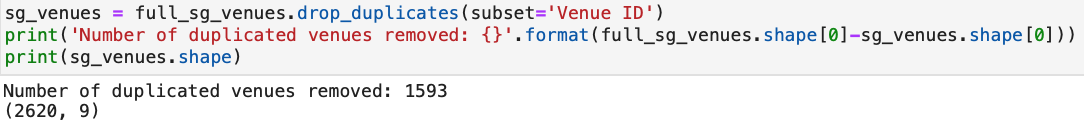
A foursquare developer account is registered and the user-specific credentials are returned, these are required to use the service. When using a personal account, 500 premium calls and 99,500 regular calls can be made per day. For the purposes of this report no premium calls were required to retrieve the required data; venue location, address, ID, and category.

Using the Foursquare ‘RESTful’ API, a radius of 2.5km from each neighbourhood marker is used to retrieve the 100 most recommended venues. If there are less than 100 venues within a given neighbourhood, all venues are returned. This is done using the ‘explore’ function in the API. The ‘Limit’ parameter is set to 100 (number of venues). After running the API, the data is stored in a data frame named ‘full\_sg\_venues’. The first five results are shown below.



As the search radius is of circular nature, duplicates were removed to avoid skewed data. As shown in the below Venn diagram; any overlaps in radius’ would cause certain venues to be counted more than once and affect the accuracy of the results.

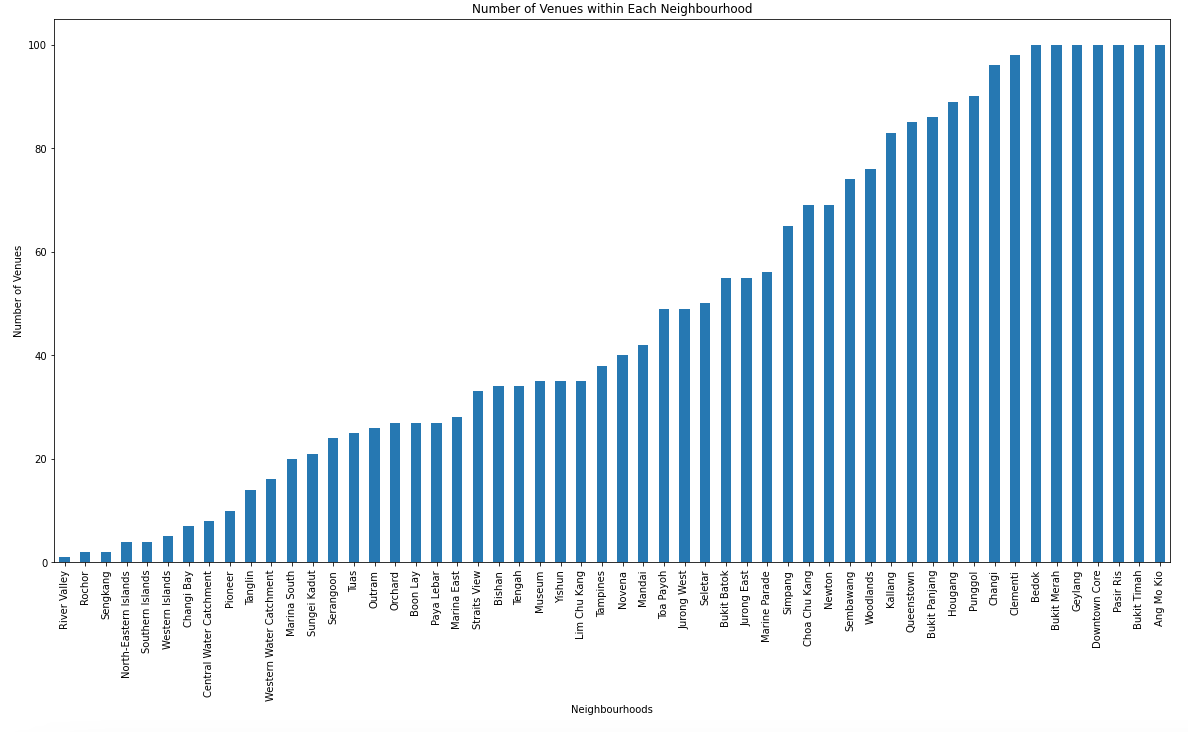




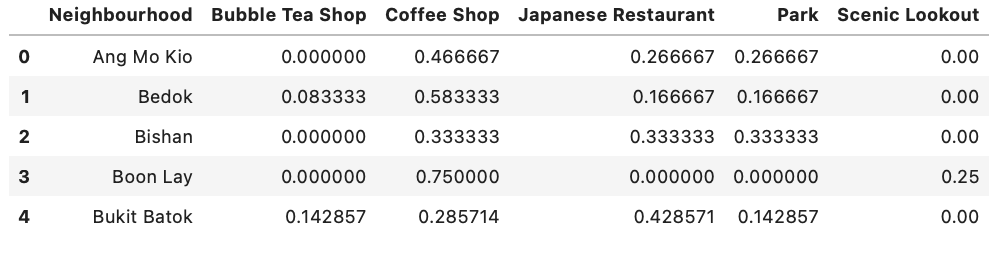
# 3. Data Analysis

## 3.1 Exploratory Data Analysis

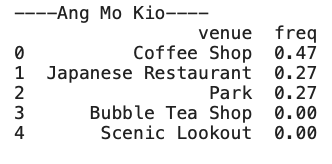
To visualise the dataset, a bar chart is used to display the number of venues within each neighbourhood.



To convert categorical features to numerical values, which will then be used for the clustering, ‘one-hot-encoding’ is performed on the dataset. The first five results are shown in the data frame below.



The above data frame is then used to calculate the respective frequency of the selected venue categories in each neighbourhood. The first neighbourhood is shown below.



The venues are then sorted into the 10 most common venue categories in each neighbourhood, and subsequently this is repeated to sort and display the same data frame for the five selected categories. Below are the first twenty-five rows:



## 3.2 Clustering

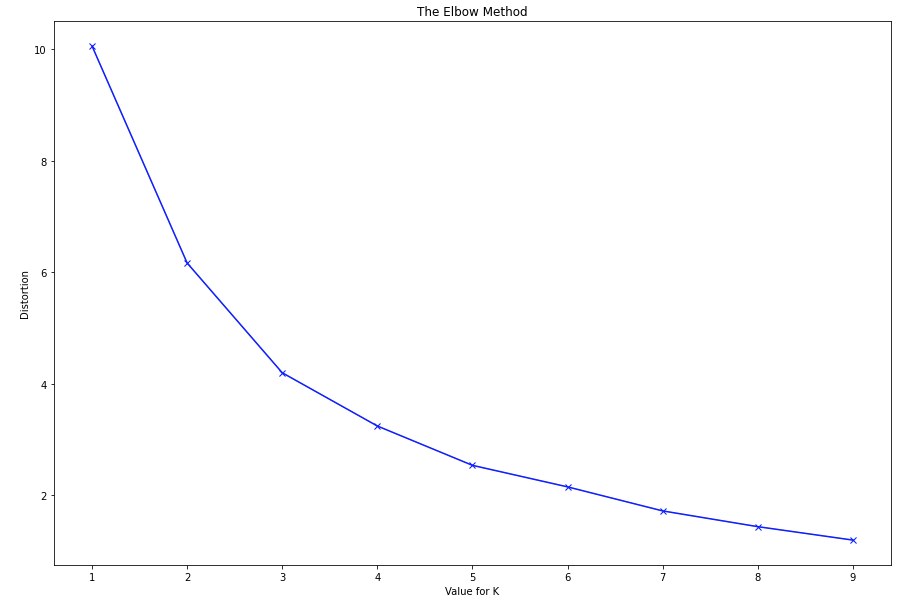
Clustering is the process of grouping a set of unlabelled objects according to their similarity in such a way that objects within the same cluster will be more similar to each other than those in other clusters.

K-Means and Density Based Spatial Clustering of Applications with Noise (DBSCAN) are two of the most popular unsupervised machine learning clustering algorithms.

### 3.2.1 K-Means Clustering

K-Means is a type of unsupervised partitioning clustering algorithm, it divides the data into k non-overlapping subsets or clusters without any cluster internal structure or labels. Objects within a cluster are very similar and objects across different clusters are very dissimilar. The algorithm tries to minimise the intra-cluster distances and maximise the inter-cluster distances.

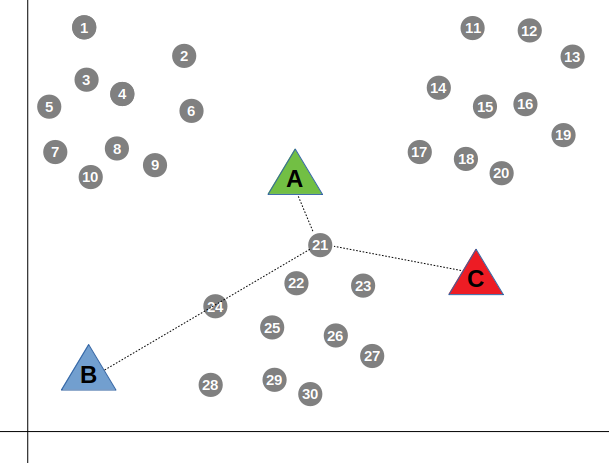
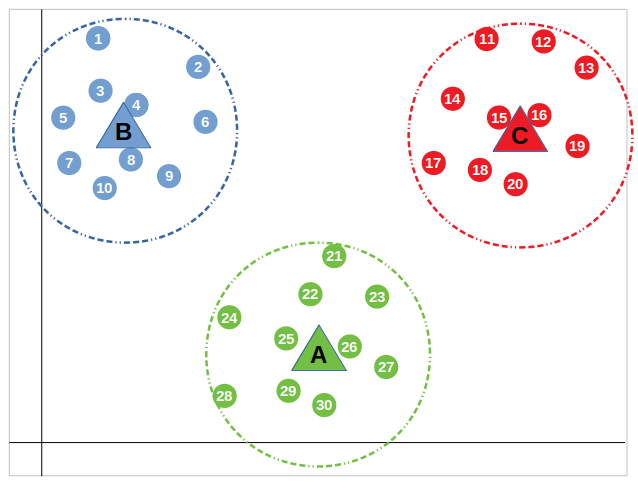
A Centroid represents the data point that is at the center of the cluster and the total distance of each point from its centroid is the error of K-Means.

For this algorithm to work, the number for k must be specified and the optimal value can be determine using the elbow method as shown below.

In this method, the value of k should be selected where the rate of decrease in error sharply shifts. In this instance, the optimal value for k is 3.

Once the three centroids have been placed in the dataset, the distance of each point to the centroid is measured and the centroid is adjusted to become the mean of all the data points that are closest to it – its cluster. This process is repeated until the centroids no longer move and there are three distinct clusters. An example of this is shown in the below diagrams.

Measuring the distance: Forming the clusters:

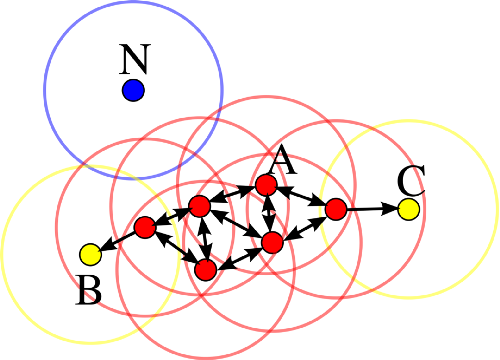
(<https://towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0>)

### 3.2.2 DBSCAN Clustering

DBSCAN locates regions of high density and separates outliers. It can find out any arbitrary shaped cluster without getting affected by noise. There are two parameters that are required for this algorithm; the radius (eps), and the minimum number of neighbours (M).

A Core point has at least M points in its neighbourhood whilst a Border point either has less than M points in its neighbourhood or it is reachable from the same core point. A cluster is formed from at least one core point + all reachable core points + all of their respective border points.

This is illustrated in the below diagram. Core points = A, border points = B and C, and outliers = N.



(<https://en.wikipedia.org/wiki/DBSCAN#/media/File:DBSCAN-Illustration.svg>)

Contrary to the K-Means clustering, DBSCAN does not require the number of clusters to be specified and is robust to outliers.

To do this, DBSCAN uses a metric such as Euclidean distance or Haversine (for coordinates) to determine the distance between data points. The algorithm offers great performance for location-based clustering, which is why it is used for the purposes of this report (Marchienne, 2020).

# 4. Results

## 4.1 DBSCAN Vs. K-Means Clusters

The clusters from both algorithms were added to the same data frame to spot any differences between them. The optimal value of k was identified to be 3 using the elbow method above, and DBSCAN also returned a total of 3 clusters excluding outliers.

### Cluster 0 – Parks



### Cluster 1 – Japanese Restaurants

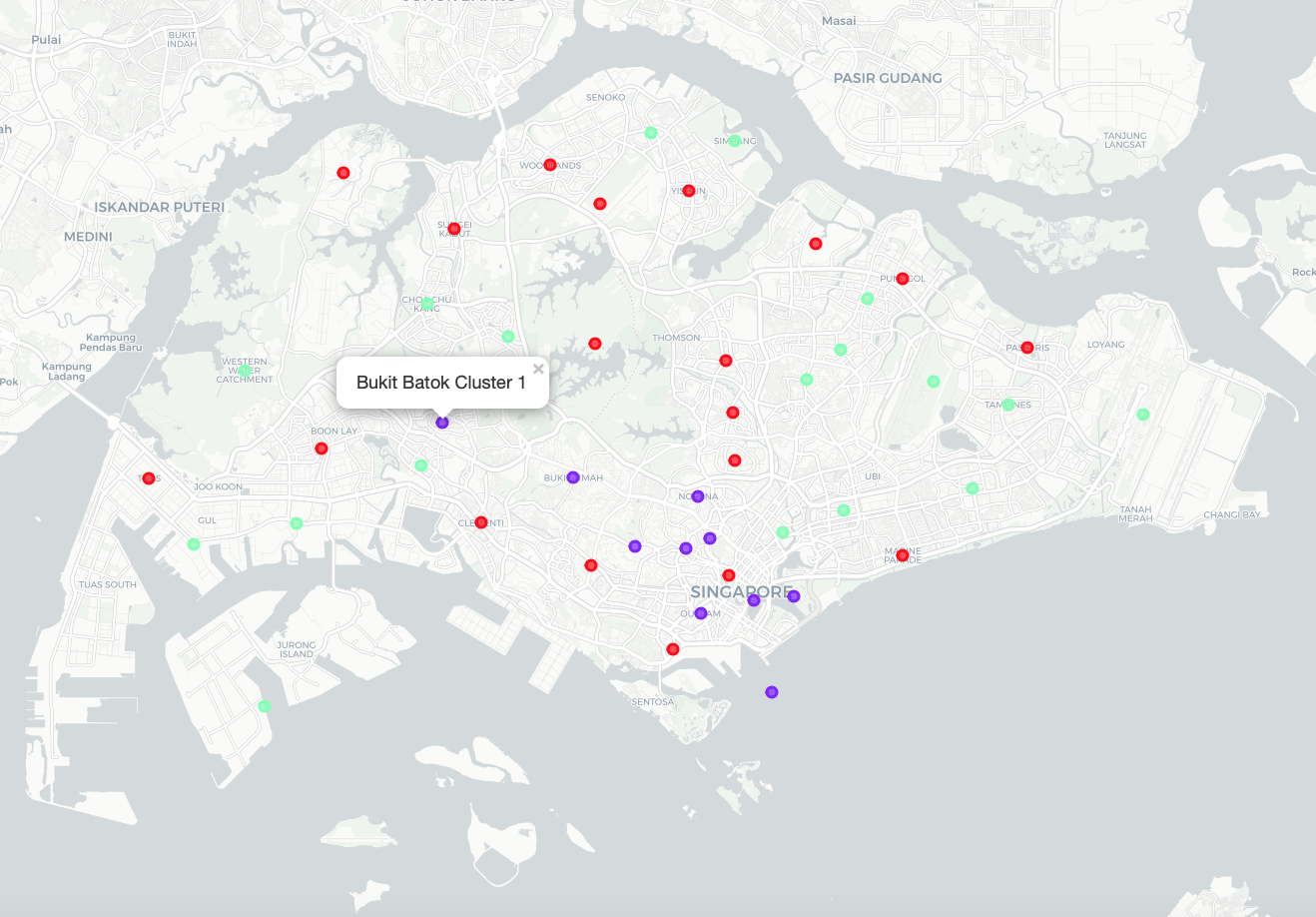


### Cluster 2 – Coffee Shops



## 4.2 Cluster Visualisation

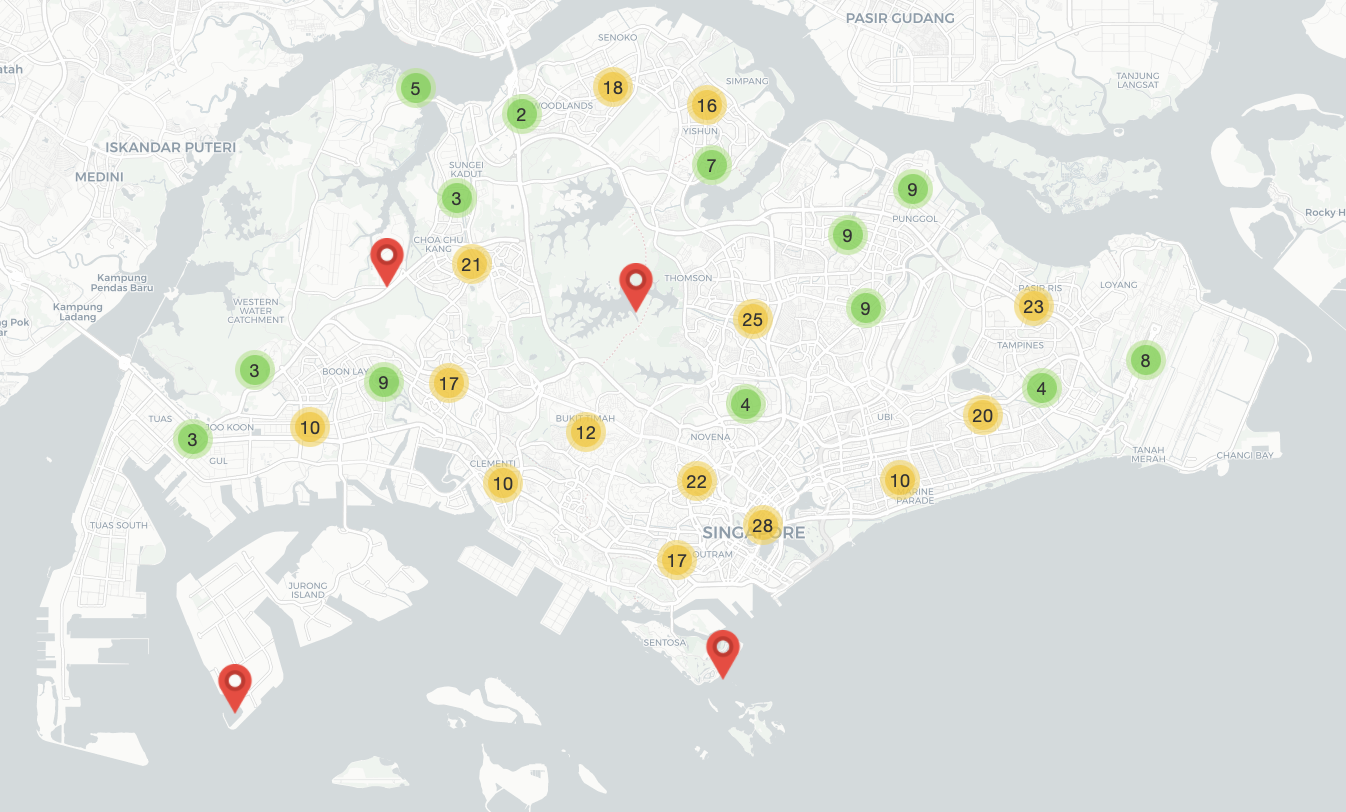
Using Folium, the clusters are colour-coded and plotted as shown below. This gives a quick and comprehensive overview of where a respective venue category is found most frequently.



Legend:

* Red: Predominantly Parks
* Purple: Predominantly Japanese Restaurants
* Green: Predominantly Coffee Shops

## 4.3 Interactive Folium Grouping Visualisation



The number within the circles represent the number of venues within the given area. The red markers represent single venues. The interactive map is available in the Jupyter notebook used for this report: <https://gist.github.com/LarsMueller-DS/367235ee8ace256a09aaf21e6d458687> .

# 5. Discussion

Singapore is a fast-growing city with many different venues scattered across it. It is entirely possible that some venues are not captured in the Foursquare service and thus excluded from this report. However, as the Foursquare API returned 2,618 venues across 301 separate venue categories, it provided enough data for the purposes of this report.

It can be noticed from the bar chart showing the number of venues by neighbourhood included earlier in this report, that certain areas such as River Valley and Rochor are very low. An explanation of this can be that those are mostly residential areas with private housing or condominiums.

The value for k in K-Means clustering can significantly impact the accuracy of the algorithm. The value used in this instance is 3, as evidenced by the elbow method explained earlier in the report. If the value of k were to be increased, more clusters would be formed. This would be appropriate if we were handling a larger dataset.

The resulting clusters, visualised in section 4.2 provide a comprehensive overview of the location of parks, Japanese restaurants, or coffee shops in Singapore. Japanese restaurants appear to be mostly present in the central region of the island, whilst parks and coffee shops are more scattered and present on the outskirts of Singapore.

There were no clusters specifically for bubble tea shops and scenic lookouts, this is because said two venue categories are scarcer in each neighbourhood than restaurants, coffee shops, or parks. However, the Folium clustering map in section 4.3 accurately portrays groupings, showing the number of venues within a given neighbourhood and can be used to supplement the clustering. Furthermore, in the Jupyter notebook used for this project, there is an additional interactive map specifically filtered to bubble tea shops and scenic lookouts.

When comparing the DBSCAN and K-Means clusters, there are no differences in the actual clustering. As previously noted, DBSCAN automatically determines the number of clusters that best fit the data – which was also 3. When comparing the resulting cluster labels, both algorithms returned the same results. As the data was not split into test and training sets due to the nature of this project, calculating accuracy metrics was replaced by referencing the performance of both algorithms side by side – which raised no concerns.

# 6. Conclusion

In this project, the required data was extracted from the internet and the Foursquare API. The data was then wrangled and formatted into Pandas data frames. The processed data was clustered using two unsupervised machine learning algorithms to give further insights into where certain venues are situated across Singapore. The data was visualised using several tools throughout the project to provide a better understanding of the data.

It can be concluded that there are neighbourhoods which are better than others for visiting as many of the specified venue categories as possible to write a travel blog.

# 7. References

Brown, A. (2020). Singapore to remain world leader in infrastructure. Retrieved 25 December 2020, from https://www.khl.com/international-construction/singapore-to-remain-world-leader-in-infrastructure/140306.article

Hirschmann, R. (2020). Singapore: restaurants numbers 2018 | Statista. Retrieved 29 December 2020, from https://www.statista.com/statistics/723478/number-of-restaurants-in-singapore/

Marchienne, S. (2020). Multi-feature geo clustering with DBSCAN. Retrieved 29 December 2020, from https://medium.com/@sylvainma/multi-feature-geo-clustering-with-dbscan-4857c6b932e2

Ouellette, C. (2020). Ultimate List of Blogging Statistics and Facts (Updated for 2020). Retrieved 29 December 2020, from https://optinmonster.com/blogging-statistics/

Plecher, H. (2020). Singapore - Gross domestic product (GDP) 2021 | Statista. Retrieved 29 December 2020, from https://www.statista.com/statistics/378648/gross-domestic-product-gdp-in-singapore/

Plecher, H. (2020). Topic: Singapore. Retrieved 1 December 2020, from https://www.statista.com/topics/2721/singapore/

Scattergood, G. (2020). From zero to hero: How Singapore became SEA's leading food and nutrition hub in 20 short years. Retrieved 29 December 2020, from https://www.foodnavigator-asia.com/Article/2019/02/12/From-zero-to-hero-How-Singapore-became-SEA-s-leading-food-and-nutrition-hub-in-20-short-years

Singapore Tourism Statistics 2020. (2020). Retrieved 27 December 2020, from https://www.budgetdirect.com.sg/travel-insurance/research/singapore-tourism-statistics