IBM Data Science Capstone Project

December 2019

boon siong tan

**1. Introduction**

Home purchase is one of the most important decisions Singaporeans make in their life. Due to limited land resources, flat prices in Singapore are expensive and flat buyers very often take many years to fully pay off the mortgages tied to the house. Very often, this decision will impact how the individual or family live for the rest of their life.

When purchasing a home, buyers are faced with a wide range of questions and these can be questions on the flat location such as proximity to public transit or questions on the housing development project such as neighbourhood theme and developer reputation.

**2. Problem**

Home buyers have many questions to consider and it is often daunting and time consuming to proceed with this enormous task.

This project seeks to assist home buyers with their home purchase by providing more information on the neighbourhood clusters and helping them find their preferred home in the most suitable neighbourhood.

**3. Data**

**3.1 Data Source**

Foursquare

Foursquare is a technological platform that provides information on a location base.

For this project, Foursquare is used to find out the surrounding amenities of a neighbourhood and the categories of these amenities.

Foursquare is accessed through an API call.

*Limitation*

The foursquare API is limited by a maximum of 50 returns per API call and limited by a maximum of 950 API calls per day as this project uses a free account. As a result, each neighbourhood only shows 50 venues even though this number can be much larger.

Google Coordinates

Google Search Engine is used to retrieve coordinate data of each neighbourhood. These coordinates are manually stored in a CSV file.

Singapore Land Authority (SLA)

The SLA provides a free to use data platform called OneMap. OneMap provides a wide variety of population data for each neighbourhood in Singapore. At the point of writing the report, the latest data available is of year 2010 and this is used as an approximate for the clusters.

For the purpose of this project, only data on education, ethnicity, age, religion and industry are used.

OneMap is accessed through an API call.

Wikipedia

Wikipedia is an online encyclopaedia that contains vast online of unstructured information. This project leverages the data source and extracts information on neighbourhood population size and neighbourhood land size.

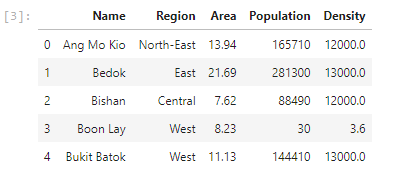
Information from Wikipedia is extracted via HTML.

**3.2 Data Pre-processing**

Neighbourhood Data

Neighbourhood data is extracted from Wikipedia via HTML and is stored in a data-frame named “sg\_area\_table”.

When first extracted, this table contains 9 columns, of which 4 are later dropped. Only “Name”, “Region”, “Area”, “Population”, “Density” columns are kept. Below is a snippet of the data-frame.



Coordinate Data

Coordinate data is extracted from a CSV file and is stored in a data-frame named “coord\_table”.

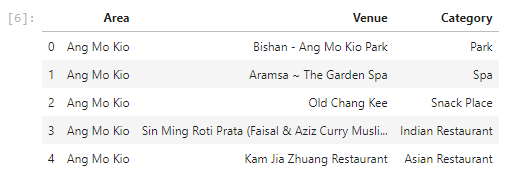
When first extracted, this table list both the latitude and longitude under the same column and this is later separated in the pre-processing process. Below is a snippet of the data-frame.



Venue Data

Venue data is extracted from the Foursquare API and is stored in a data-frame named “venue\_df”.

The data is collected by iteratively calling the Foursquare API using a list of neighbourhoods and each time the API is called, only the “Area”, “Venue” and “Category” information are selectively chosen to be appended into the data-frame. Below is a snippet of the data-frame.



Population Data

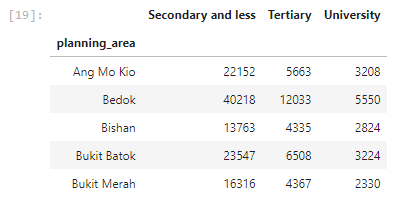
Population data is extracted from OneMap via API call and stored in 5 data-frames, “edu\_df”, “ethnic\_df”, “ind\_df”, “age\_df” and “reli\_df”.

a. Edu\_df

Information on education in the following neighbourhoods are not available and are therefore excluded from analysis.

*Boon Lay, Lim Chu Kang, North-Eastern Islands, Orchard, Paya Lebar, Pioneer, Seletar, Sungei Kadut, Tengah, Tuas, Western Islands*

When extracted, there are a total of 7 columns. These are grouped into 3 columns in the pre-processing stage, “Secondary and less”, “Tertiary” and “University” to streamline the number of columns for analysis. Below is a snippet of the data-frame.



b. Ethnic\_df

Information on ethnic in the following neighbourhoods are not available and are therefore excluded from analysis.

*Boon Lay, Lim Chu Kang, North-Eastern Islands, Orchard, Paya Lebar, Pioneer, Seletar, Sungei Kadut, Tengah, Tuas, Western Islands*

Below is a snippet of the data-frame.

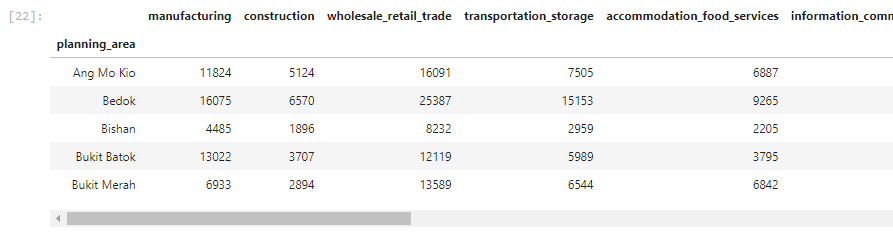


c. Ind\_df

Information on industry in the following neighbourhoods are not available and are therefore excluded from analysis.

*Boon Lay, Lim Chu Kang, North-Eastern Islands, Orchard, Paya Lebar, Pioneer, Seletar, Sungei Kadut, Tengah, Tuas, Western Islands*

When extracted, there are 4 columns which have 0 for all rows. These are removed as they can be no insights derived from clustering. Below is a snippet of the data-frame.



d. Age\_df

When extracted, there are a total of 20 columns and this makes it difficult to analyse. In the pre-processing stage, these are grouped into 4 columns “Pre-school”, “Student”, “Worker” and “Elderly”

Definition

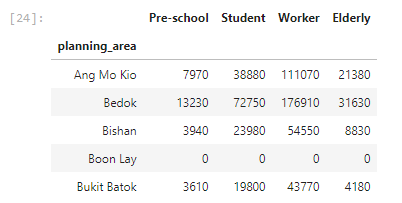
Pre-school: 0-4 years old

Student: 5-24 years old

Worker: 25- 64 years old

Elderly: 65 years old and above

Below is a snippet of the data-frame.

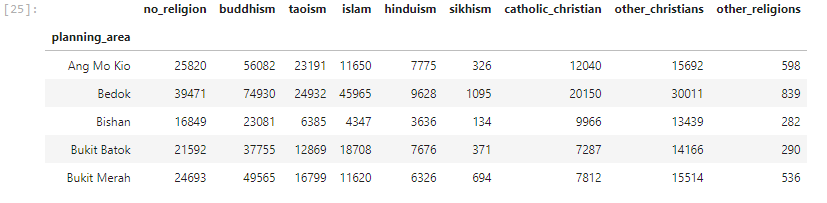


e. Reli\_df

Information on religion in the following neighbourhoods are not available and are therefore excluded from analysis.

*Boon Lay, Lim Chu Kang, North-Eastern Islands, Orchard, Paya Lebar, Pioneer, Seletar, Sungei Kadut, Tengah, Tuas, Western Islands*

Below is a snippet of the data-frame.



**4. Methodology**

The project seeks to cluster neighbourhoods to help home buyers identify the neighbourhood that is most suitable for themselves.

To do this, 3 sets of clustering is done.

The first clustering identifies the most common type of amenities in the neighbourhood and cluster the neighbourhoods according to the categories of its amenities.

The second clustering identifies the neighbourhoods based on the industries the population living in the neighbourhood is working in.

The last clustering groups the neighbourhood base on the population demographics in the area such as age, ethnicity, religion and education.

**4.1 Amenities Clustering**

The “venue\_df” data-frame is used for this clustering. The data-frame shows venue (amenities) data and the corresponding neighbourhood and category.

The first step is to turn the category column into 0s and 1s for the cluster analysis.

Next, the rows of the data-frame are re-grouped based on the neighbourhood data in each row.

The output of this process is a data-frame that rows of neighbourhood and the degree of categories in each of these neighbourhoods.

A simple exploratory analysis shows that restaurants, hotels and parks are popular choices in the top 5 amenities and further clustering analysis is needed to validate this.

The K-means model is used to cluster the neighbourhoods and a cluster size of 4 is subjectively chosen as it creates clear and identifiable differences between groups.

**4.2 Industry Clustering**

The “ind\_df” data-frame is used for this clustering. The data-frame shows neighbourhood data and the corresponding population in each industry.

At present, the data in the data-frame is hardly comparable between rows and columns and the model will be skewed by the size of the neighbourhood and the industry.

To reduce the effects of size, the data is first turn into proportions by dividing against the total population of each neighbourhood and then normalize through the StandardScaler function.

The resulting data-frame is then modelled through K-means. A cluster size of 3 is subjectively chosen as it creates clear and identifiable differences between groups.

**4.3 Demographic Clustering**

The “edu\_df”, “ethnic\_df”, “age\_df” and “reli\_df” are combined together to create the demographic clusters. The combined data-frame shows neighbourhood data and the corresponding population in education, ethnic, age and religion groups.

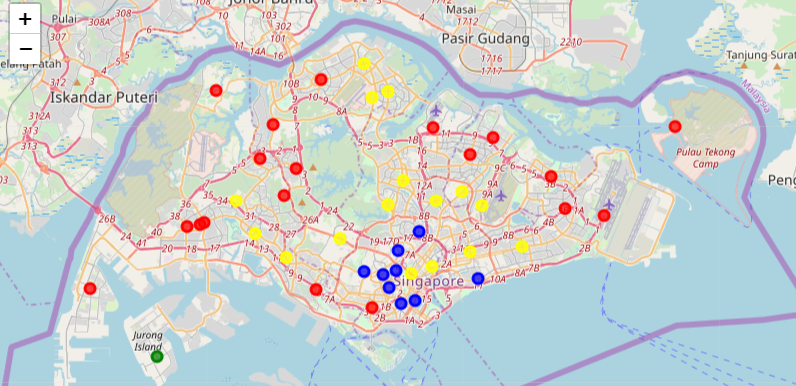
At present, the data in the data-frame is hardly comparable between rows and columns and the model will be skewed by the size of the neighbourhood and each of the columns.

To reduce the effects of size, the data is first turn into proportions by dividing against the total population of each neighbourhood and then normalize through the StandardScaler function.

The resulting data-frame is then modelled through K-means. A cluster size of 7 is subjectively chosen as it creates clear and identifiable differences between groups.

**5. Results**

**5.1 Amenities Clustering**



Top 5 Amenities

**Blue:** Hotel, Japanese Restaurant, Chinese Restaurant, Bakery, Boutique

**Green:** Boat or Ferry, Cruise, Coffee Shop, Recreation Center, Beach

**Red:** Coffee Shop, Park, Café, Chinese Restaurant, Food Court

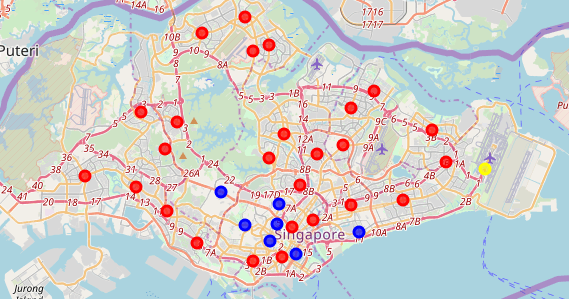
**Yellow:** Chinese Restaurant, Ice Cream Shop, Indian Restaurant, Park, Thai Restaurant

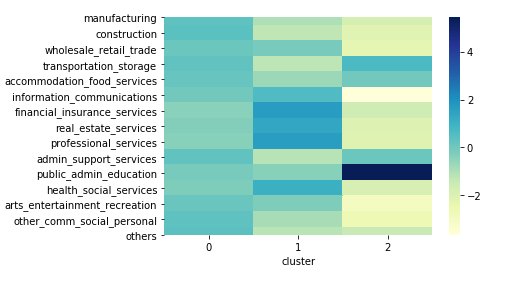
Commentary

Neighbourhoods in the southern region are better represented by hotels. Neighbourhoods in the outer parts of the country are better represented by coffee shops and parks while those in the inner parts of the country are better represented by the different restaurant cuisines.

Jurong island is different from the rest of the country and is represented by water transport features.

**5.2 Industry Clustering**





**Red:** 0

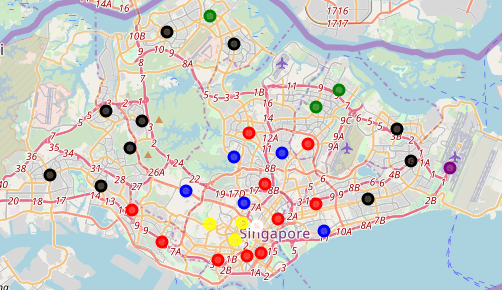
**Blue:** 1

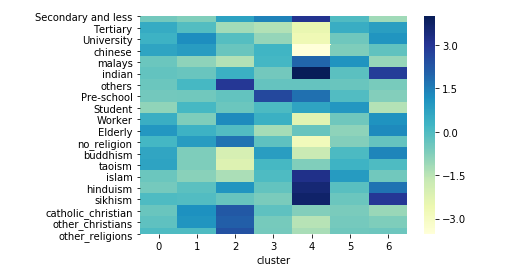
**Yellow:** 2

Commentaries

Population residing in the southern part of Singapore are more likely to work in the financial, real estate and professional services industry while those residing in Changi are more likely to work in the public, admin and education industry.

**5.3 Demographic Clustering**





**Red:** 0

**Blue:** 1

**Yellow:** 2

**Green:** 3

**Purple:** 4

**Black:** 5

**White:** 6

Commentaries

Population residing in the sub-urban (blue) are more likely to be university educated, Chinese and Christians. Population residing in the Orchard area (yellow) are more likely to be Christians of the non-core races. The North and North-Eastern Outskirts (green) have more preschools. The population in Changi (purple) are more likely to be Malays and Indians. The population in Rocher (White) are more likely to be Indians.

**6. Future Directions**

As of writing, the project faces a couple of data limitations namely from Foursquare API and OneMap API.

Foursquare API allows a maximum of 50 rows per call and a total of 950 calls per day, thereby limiting the number of amenities data the project uses.

OneMap does not yet provide up to date data.

Once both of these limitations are removed, there can be room for more relevant and robust data sets, and this would also translate to better insights from the cluster analysis.

**7. Conclusion**

There are indeed clear lines and differences among neighbourhood clusters in Singapore and this is now highlighted by aggregating and finding patterns in data from different sources.

While the study concluded that neighbourhood clusters in Singapore differ from one another in amenities, population industries and population demographics, the reader of this report should also take note of the limitations in this study.