Coursera - Applied Data Science

Capstone Project

Best Tourist Spots in Singapore

An exercise in Machine Learning using Clustering Algorithms

By – Ashwin Mohananath Syamala

Date: January 2020



# Introduction:

Singapore, an island city-state off southern Malaysia, is a global financial center with a tropical climate and multicultural population. Singapore is much more than the sum of its numerous attractions. It’s constantly evolving, reinventing, and reimagining itself, with people who are passionate about creating new possibilities.

It’s where foodies, explorers, collectors, action seekers, culture shapers, and socialisers meet―and new experiences are created every day.

There’s something for everyone―world-class theme parks, iconic waterfront attractions, the world’s first night safari, hidden gems to explore in charming neighbourhoods, lush greenery. Discover how a City in the Garden is rooted in history and heritage. Stroll along the colourful shophouses in Haji Lane. Or trek off the beaten path and walk atop trees at MacRitchie Reservoir. If you’re passionate about discovering places, here’s where nature lovers, history buffs, backpackers, city trippers, trekkers, discoverers, sightseers―meet.

# Business Problem:

Given the diverse range of activities, historical sites and places to visit, a traveller could easily be overwhelmed by the huge number of choices that are available and will make it very hard for them to organise an itinerary for their travel. Apart from the actual venues, often the number of places travellers could stay is also a key metric that is considered when making travel plans.

K-Means Clustering, a Machine Learning Algorithm, is implemented to analyse the various neighbourhoods in Singapore and group them into various clusters based on the categories of venues in those neighbourhoods and the number of available Airbnb rental properties that are close to the venues in those categories.

The objective of this specific use of K-means clustering algorithm is to identify places of interest from all the neighbourhoods of Singapore and group them into clusters and visualise the results in a map for better understanding.

The main venue categories that have been included in this analysis are purely tourist destinations such as monuments, historic places, adventure areas etc. Place such as restaurants and other categories have been excluded from the analysis.

# Target Audience:

Travel agents and tour organisers can make use of a model such as the one described above to create itineraries and tour packages for potential travellers.

# About the Data:

As the task involves identifying venues and their categories in each neighbourhood of Singapore, as well as the number of rental properties in those neighbourhoods, the following data have been obtained from the sources listed below.

**Source: Inside Airbnb**

**Link:** <http://insideairbnb.com/get-the-data.html>

## **Files Used:**

|  |  |
| --- | --- |
| **Name** | **Description** |
| neighbourhoods.geojason | GeoJSON file of neighbourhoods of the city. |
| Listings.csv | Summary information and metrics for listings in Singapore. |

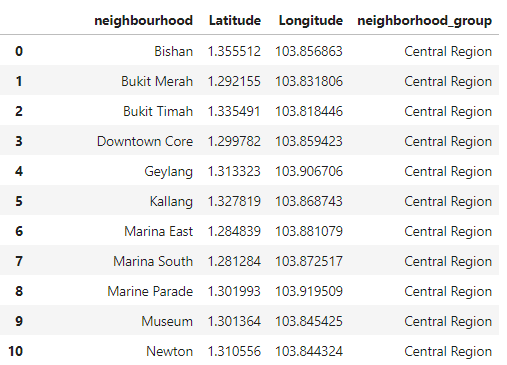
## **Details used from Source Files:**

**neighbourhoods.geojason**

|  |  |  |
| --- | --- | --- |
| **s.no** | **Column name** | **Description** |
| 1 | Neighbourhood | Name of the neighbourhoods. |
| 2 | Neighbourhood\_group | The Neighbourhood group the actual neighbourhood belongs to. |
| 3 | Coordinates | The Latitude and Longitude information that can be parsed from the neighbourhoods coordinates. |

The file is read using the library JSON in Python. With this library, the file can be read and parsed to extract the above information listed in the table. The file will contain all the Neighbourhoods, their groups and the coordinates of those neighbourhoods. By looping through all of them using a for loop, the results are stored in a Pandas Dataframe.

The screenshot containing a sample of the resulting Dataframe is shown below for reference.



**Listings.csv**

|  |  |  |
| --- | --- | --- |
| **s.no** | **Column name** | **Description** |
| 1 | Neighbourhood | Name of the neighbourhoods. |
| 2 | Neighbourhood\_group | The Neighbourhood group the actual neighbourhood belongs to. |
| 3 | Latitude | Latitude of the Airbnb listing |
| 4 | Longitude | Longitude of the Airbnb listing |

The Pandas Dataframe is used again to read the listings.csv file to download the data from the file into a dataframe.

A screenshot of the resulting dataframe with all the columns (including the ones mentioned in the above table) have been shown below.



## **Foursquare Venue Data:**

In addition to the above files, calls will be made to Foursquare API using a URL to explore the different neighbourhoods using the geographical coordinates.

By using a custom defined function, all the neighbourhoods from the geojason file which has already been loaded into a dataframe are passed through the function which creates a URL to explore the venues and sent to the Foursquare API to get the venues surrounding those neighbourhoods in a json format.

The results from the Json are parsed and the venue names, latitude and longitude and the venue category are stored as a separate Pandas Dataframe.

A screenshot of the sample Dataframe has been displayed below for reference.

