Finding the Ideal Town for Home Purchase

Capstone Assignment

Prepared towards partial fulfillment of requirements for

IBM's Data Science Professional Certificate Program

Vikas Sharma

July 2020

Section 1 – Introduction

The Singapore of yore painted a very different picture from its image today as a first-world, bustling metropolis with high-rise residential buildings dotting the skyline. In 1960, almost one-third of Singaporeans lived in dilapidated slums and squatter-dwellings with inadequate sanitation. Soon after gaining independence in 1965, Singapore's policymakers decided to encourage home ownership in view of various 'positive externalities' associated with it (lowering of public health costs, crime-rates, vandalism, drug abuse etc.; greater community activism, political stability and enforced financial discipline). Today, Singapore has among the highest home-ownership rates globally, with 90% of resident households owning homes.

Purchasing their first apartment is a major life milestone and a significant financial commitment for a young family in Singapore. This purchase decision often involves time-consuming research on not just price, but on nearby amenities such as schools, supermarkets, shopping malls, train stations, and such. This research could benefit from a data science approach to whittle down the candidate options and choose the optimum one, and therefore, **prospective home buyers in Singapore (and possibley elsewhere) are the audience for this paper**.

This paper is an attempt to illustrate that by using the example of a fictitious prospective home-owner named 'John'. John's family consists of his wife Jane, and his 2 year old son Joseph. **John has approached us for assistance in deciding on the most suitable township for his future home**. He has provided us the following **specifications** for the ideal township for his family to buy a home in:

1) The town must be 'mature'. John defines it as a town having above-median population. He believes that such mature towns will have high human traffic, good transportation facilities, and ample existing amenities.

2) There are certain amenities/venues that John's family is especially particular about and wants the ideal town to have within a walking distance of 1 km (1000m). At the same time, they understand that it may be difficult to find a town that fulfills this criterion for all their desired amenities/venues. So, they agreed to list their top 5 desired amenities/venues and put a weightage (a measure of importance) on each of them. The higher the weightage, the more important the amenity/venue is to John's family. These weightages are shown below:

Amenity/Venue	Assigned Weightage
Multiplex Cinema	40%
Supermarket	25%
Playground	20%
Bubble Tea Shop	10%
Shopping Mall	5%

Section 2 – Data

1) **Data.gov.sg**: To facilitate the analysis, we needed name and population information for the various towns in Singapore (as defined by the Housing and Development Board (HDB)). This data was procured from Data.gov.sg, which is the Singapore government's one-stop portal for all publicly-available datasets from over 70 government agencies. The exact URL for the data is provided below:

https://data.gov.sg/dataset/estimated-resident-population-living-in-hdb-flats?resource_id=b29c1af8-e11a-4e61-b813-933db9f69633

Data was downloaded in a CSV file and saved on disk as 'SGTownData.csv'. The data file consisted of three columns – 1) Financial Year 2) Town 3) Population. This data was processed (see Methodology section) later.

2) **Foursquare API (from developer.foursquare.com):** The Foursquare API provides detailed and comprehensive location data that can be used to get information on nearby/popular venues, and details about those venues in a given area. This data was used to obtain information on the preferred venues/amenities (indicated by our client John in the specifications) for each of the towns in Singapore. This data was extracted through calls to the API within the Jupyter Notebook, and received as a JSON file. The JSON file was manipulated to convert it into a pandas dataframe for further processing and analysis.

Section 3 – Methodology

The overall methodology/approach was divided into **SIX distinct steps** as shown below. Each of these steps is expanded upon later in this section.

1. Identify and import relevant Python libraries
2. Find data on towns in Singapore from Singapore government's official data portal (https://data.gov.sg/)
3. Explore and clean the township data
4. Obtain geographical coordinates for the towns
5. Use the Foursquare API to get information on venues and amenities in the towns
6. Analyze the venue data in light of client specifications, and arrive at the optimum town selection

Step1: Identify and import relevant Python libraries

In light of the study specifications and the kind of data that would be used, relevant Python libraries were imported to the Jupyter Notebook. The screenshot below shows each imported library and its intended use.

```
import pandas as pd # library for data analsysis
import numpy as np # library to handle data in a vectorized manner
import matplotlib.pyplot as plt # library to plot visualizations
import folium # map rendering library
import json # library to handle JSON files
import geocoder # to get coordinates
import time # to use a timer
from IPython.display import Image # to render images of the maps created
import requests # library to handle requests
```

Step2: Find data on towns in Singapore from Singapore government's official data portal

Data was downloaded from **https://data.gov.sg/** in a CSV file and saved on disk as 'SGTownData.csv'. The data file consisted of three columns – 1) Financial Year 2) Town 3) Population, and 296 rows.

Step3: Explore and clean the township data

- Data from 'SGTownData.csv' was loaded into a pandas dataframe named '**towndata'**.
- Upon inspecting the data, it was noticed that dataframe had population information on each of Singapore's 26 towns from year 2008 to year 2018. Since we would want to work with only the latest population information, only data for year 2018 was extracted into a new dataframe named 'df'. This dataframe df had 26 rows, with each row representing one town.
- Now that we had the latest town population information in **df,** we dropped the column 'financial_year' and also renamed the column 'town_or_estate' to 'Town' for ease of reference
- As per the specifications given by the client (John) in the Introduction section, we only retained those towns that have above median population for further consideration. 13 of the 26 towns were dropped
- We ended up with our cleaned and processed dataframe (named **df2**) containing names of the 13 candidate towns for further considertaion. The 'population' column was now dropped as well because it was not relevant any longer for further analysis. This cleaned dataframe df2 is shown in the figure next.

```
#our cleaned and proceesed dataframe
df2
```

	Town
0	Ang Mo Kio
1	Bedok
2	Bukit Batok
3	Bukit Merah
4	Bukit Panjang
5	Choa Chu Kang
6	Hougang
7	Jurong West
8	Punggol
9	Sengkang
10	Tampines
11	Woodlands
12	Yishun

Step4: Obtain geographical coordinates for the towns

- Using the ArcGIS system, latitude and longitude coordinates were extracted for each of the 13 towns, and added to two lists named '**latlist**' (for latitude information) and '**longlist**' for longitude information. Since geocoding can be a time-consuming process, a timer was used to check how long the entire geocoding process took.

- The latitude and longitude information extracted above was added to our town dataframe df2 as two new columns, named 'Latitude' and 'Longitude'. The head of df2 is shown below to provide a glimpse at it.

```
df2.head()
```

	Town	Latitude	Longitude
0	Ang Mo Kio	1.37161	103.84546
1	Bedok	1.32425	103.95297
2	Bukit Batok	1.34952	103.75277
3	Bukit Merah	1.28417	103.82306
4	Bukit Panjang	1.37877	103.76977

- Next, to get a visual understanding of where these towns are in Singapore, the information from df2 was plotted on a map of Singapore using the folium library. The resulting map is shown below, with each of the 13 towns marked in red circles.

```
1 SingMap
```

Step5: Use the Foursquare API to get information on venues and amenities in the towns

- In accordance with the client's specifications, the radius of exploration was restricted to 1km (1000m). In addition, the number of venue results from each town was limited to 100

- Next, we looped through all 13 towns, creating a API request URL for each, getting the response from Foursquare API, extracting relevant information (town name, venue name, category) from the response JSON file, and storing this information in a list called '**venuelist**'. A timer was also used to check how long this whole process took.

- Each element in venuelist represented a venue and contained information on the - town, venue's name, and venue's category. We provide a glimpse at the content of 'venuelist' by showing its first 5 elements

```
venuelist[0:5]
```

```
[['Ang Mo Kio', 'NTUC FairPrice', 'Supermarket'],
 ['Ang Mo Kio', 'Face Ban Mian 非板面 (Ang Mo Kio)', 'Noodle House'],
 ['Ang Mo Kio', 'Kam Jia Zhuang Restaurant', 'Asian Restaurant'],
 ['Ang Mo Kio', 'Old Chang Kee', 'Snack Place'],
 ['Ang Mo Kio', 'MOS Burger', 'Burger Joint']]
```

- Checking the length of venuelist revealed that a total of 787 venues across the 13 towns had been added from the Foursquare API

Step6: Analyze the venue data in light of client specifications and arrive at optimum town selection

- The venueslist obtained in Step5 was converted into a pandas dataframe **'mydf'**, with 787 rows (each row representing one venue) and 3 columns (one each for town name, venue name, and category name)

```
mydf.head()
```

	Town	Name	Category
0	Ang Mo Kio	NTUC FairPrice	Supermarket
1	Ang Mo Kio	Face Ban Mian 非板面 (Ang Mo Kio)	Noodle House
2	Ang Mo Kio	Kam Jia Zhuang Restaurant	Asian Restaurant
3	Ang Mo Kio	Old Chang Kee	Snack Place
4	Ang Mo Kio	MOS Burger	Burger Joint

- Upon inspection of the column 'Category', we found 145 unique values, implying that the 787 venues are spread across 145 different venue categories.

- Next, a pivot table was created with 'Town' as the index and 'Category' as the columns. This pivot table was named **'piv'** and was populated with the number of venues for each category for each town. As expected, this pivot table had 13 rows (one row for each town) and 145 columns (one column for each venue category). The pivot table was then manipulated using the 'droplevel' function to change the columns from multi-level to single-level. This was done for ease of use. A snapshot of the resulting 'piv' is shown below for reference.

```
piv
```

Category	Accessories Store	American Restaurant	Arcade	Arts & Entertainment	Asian Restaurant	Athletics & Sports
Town						
Ang Mo Kio	0	0	0	0	3	0
Bedok	0	1	0	0	2	0
Bukit Batok	0	0	0	1	0	0
Bukit Merah	0	0	0	0	1	0
Bukit Panjang	1	1	0	0	3	0
Choa Chu Kang	0	0	0	0	2	0
Hougang	0	0	0	0	3	1
Jurong West	0	1	0	0	6	0

- From initial specifications given by the client, we knew that the **venues/amenities of interest** to his family are the following - 'Multiplex','Supermarket','Shopping Mall','Playground','Bubble Tea Shop'. So all other venue categories were deleted from the pivot table and only the 5 focus categories were retained. This new pivot table was named 'piv2'. A screenshot of the resulting **'piv2'** is shown below for reference.

```
piv2
```

Category	Multiplex	Supermarket	Shopping Mall	Playground	Bubble Tea Shop
Town					
Ang Mo Kio	1	2	1	0	2
Bedok	0	1	0	1	0
Bukit Batok	1	0	1	0	0
Bukit Merah	0	2	0	0	1
Bukit Panjang	0	1	3	1	0
Choa Chu Kang	0	1	0	1	1
Hougang	0	2	2	1	0
Jurong West	1	1	2	1	1
Punggol	1	3	1	0	1
Sengkang	0	2	1	0	0
Tampines	0	1	1	1	0
Woodlands	0	3	5	0	0
Yishun	1	2	2	0	1

- We also knew that the key decision metric for the client and his family is the presence (or not) of focus venue categories in their selected town, NOT the number of venues of the focus venue categories. With this in mind, a **binary coding** of the pivot table 'piv2' was done whereby - '1'= town has one or more venues of a focus category '0'= town has zero venues of a focus category. Resultant piv2 is shown below

```
piv2
```

Category	Multiplex	Supermarket	Shopping Mall	Playground	Bubble Tea Shop
Town					
Ang Mo Kio	1	1	1	0	1
Bedok	0	1	0	1	0
Bukit Batok	1	0	1	0	0
Bukit Merah	0	1	0	0	1
Bukit Panjang	0	1	1	1	0
Choa Chu Kang	0	1	0	1	1
Hougang	0	1	1	1	0
Jurong West	1	1	1	1	1
Punggol	1	1	1	0	1
Sengkang	0	1	1	0	0
Tampines	0	1	1	1	0
Woodlands	0	1	1	0	0
Yishun	1	1	1	0	1

- Now that we had town-wise information on the presence (or absence) of focus venue categories, the weightage provided by the client (in the Introduction section) was used to come up with a weighted score for each of the 13 towns. The weightages are reproduced below:

Amenity/Venue	Assigned Weightage
Multiplex Cinema	40%
Supermarket	25%
Playground	20%
Bubble Tea Shop	10%
Shopping Mall	5%

- Using the weightages, a new column 'Score' was added to our pivot table 'piv2' to hold the weighted score for each town. The pivot table was then sorted on 'Score' in descending order. The final piv2 with the score column added to it is shown below.

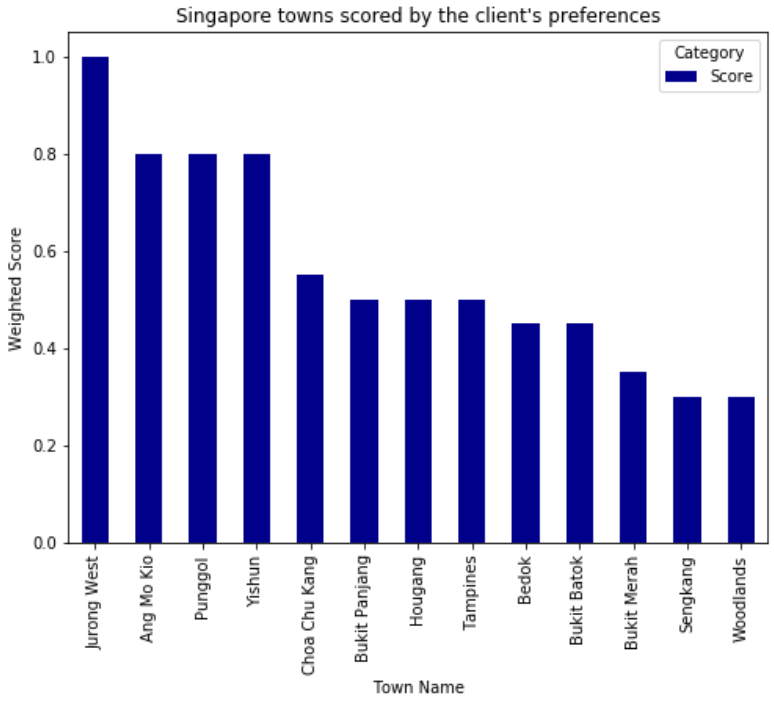
```
piv2
```

Category	Multiplex	Supermarket	Shopping Mall	Playground	Bubble Tea Shop	Score
Town						
Jurong West	1	1	1	1	1	1.00
Ang Mo Kio	1	1	1	0	1	0.80
Punggol	1	1	1	0	1	0.80
Yishun	1	1	1	0	1	0.80
Choa Chu Kang	0	1	0	1	1	0.55
Bukit Panjang	0	1	1	1	0	0.50
Hougang	0	1	1	1	0	0.50
Tampines	0	1	1	1	0	0.50
Bedok	0	1	0	1	0	0.45
Bukit Batok	1	0	1	0	0	0.45
Bukit Merah	0	1	0	0	1	0.35
Sengkang	0	1	1	0	0	0.30
Woodlands	0	1	1	0	0	0.30

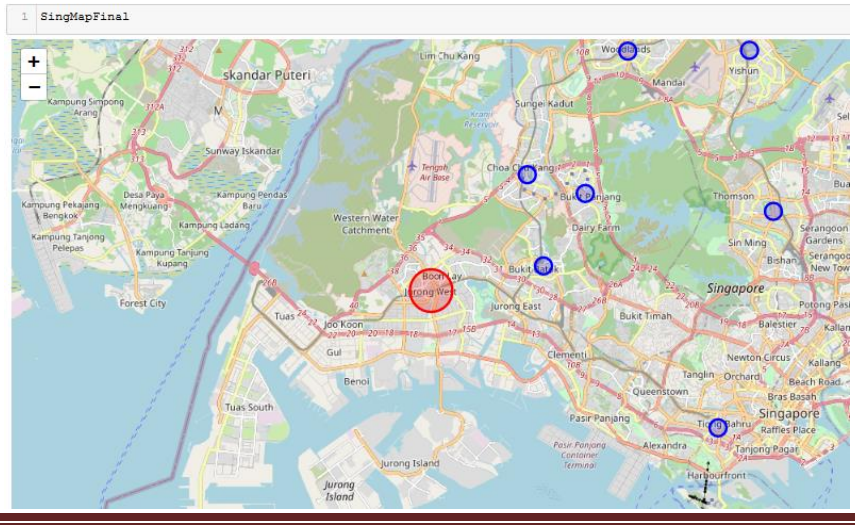
This concludes the Methodology section of this report.

Section 4 – Results

At the end of the methodology section, we had obtained weighted scores for each of the 13 towns under consideration. The higher the score, the better it is as a candidate for selection. The scores for the 13 towns were plotted in a bar graph for visual representation.



As can be seen clearly from the bar chart, the town with the highest score is **'Jurong West'. Hence, our client John and his family should select Jurong West as their preferred town for purchasing their home.** For good measure, we also highlight the optimum town selection by plotting it on Singapore's map, centering the map on the optimum town (Jurong West) and placing a Red marker on it to differentiate it from the other towns.



Section 5 – Discussion

1. The approach described in this paper could be easily adapted for use by propsective home buyers in any other part of the world by tweaking certain parameters such as:
    a. Radius within which preferred amenities/venues should be located
    b. Categories of preferred venues/amenities
    c. Weights assigned to the preferred venues/amenities
    d. Township information for other cities
    e. Venue information for other cities extracted from Foursquare API

2. Instead of trying to score towns on the number of venues they have from the preferred venue categories, this paper scores them on a binary (1 = town contains one or more venues of a preferred category, 0 = town contains zero venues of a preferred category). This binary approach is more accurately reflects real life scenarios of location selection by prospective home buyers.

3. The described approach could be improved further by adding in price (for e.g. median price per square foot) as another factor to score towns on.

Section 6 – Conclusion

Using the case study of a fictitious family looking to purchase their first home in Singapore, this paper described a data-science driven approach to sieve through candidate towns and choose the optimum town that has the best mix of venues/amenities that are of interest to prospective home buyers.