Finding the Ideal Town for Home Purchase

Capstone Assignment –

Data Section

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 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.