Recommending Business to Open in Tampines, Singapore

Eugene

May 2020

1.   INTRODUCTION

A variety of shops exists in each neighbourhood in Singapore to cater to the needs of the residents living in them. As a prospective business owner looking to open a shop in Tampines, it would be wise to identify the gaps (untapped potential) in the needs served by the existing range of shops in the neighbourhood before deciding the type of business to start. To help provide clarity on the type of business (e.g. restaurant, provision store, café) that one should open, Resident Household data from the Department of Statistics Singapore can be used to first identify other neighbourhoods that are most similar to Tampines. Following this, check-in data from Foursquare can be used to gain insights into the range of shops that are available in these similar neighbourhoods, and compare it with the existing range in Tampines. Through this comparison, the underserved areas can be identified, and the prospective business owner can avoid entering a niche that is already well served and hence more competitive.

2.   DATA

2.1. Data Sources

Two main sources of data were used in the analysis:

1)   To identify the neighbourhoods that are most similar to Tampines, data pertaining to the Basic Demographic Characteristics in Singapore and Resident Households published by the Singapore Department of Statistics was used. From these datasets, the distributions of the residents by age, type of dwelling, and household size were extracted and analysed.

2)   Once the most similar neighbourhoods had been identified, Foursquare location data was used to explore these neighbourhoods and extract the list of shops that were operating in each of them. The data was analysed to   identify   the   distribution   of   business   categories   in   these

neighbourhoods and derive an expected distribution of shops in Tampines. The ten most underserved areas were subsequently identified as potential areas for the prospective business.

2.2. Data Cleaning

The data imported from the Singapore Department of Statistics had two main issues: 1) they were filled with *NaNs* and 2) the neighbourhoods were broken down into their respective subzones. As the analysis was not expected to require such resolution, the subzone columns were dropped, i.e. only the total values for each neighbourhood were retained, along with the rows and columns that contained *NaNs*. The data imported from the two links were eventually combined into a single dataframe for subsequent analysis.

3. METHODOLOGY

3.1. Visualisation of neighbourhood locations

To check that the datasets obtained from the Singapore Department of Statistics covered the majority of the neighbourhoods in Singapore, the coordinates for each neighbourhood were extracted using the *Nominatim* package and subsequently visualised using the *folium* package.

3.2. Identifying the most similar neighbourhoods

The identification of neighbourhoods that are most similar to Tampines entails the clustering of neighbourhoods based on the characteristics of the residents that live in them. The K-Means Clustering algorithm is suitable for this purpose and was used to split the neighbourhoods into three different clusters. The data was scaled before fitting was performed. The results were subsequently visualised using the *folium* package.

3.3. Extracting information on businesses in each neighbourhood

Once similar neighbourhoods were identified, the Foursquare API was used to explore each of them to identify the types of businesses that were operating within 1km of the center of the neighbourhood. Coordinates found using *Nominatim* were used in the *Explore* requests. Results for one of the neighbourhoods, Bedok, were visualised using the *folium* package.

3.4. Identifying potentially underserved areas in Tampines

Once the categories of the businesses in these similar neighbourhoods were identified, the median frequency of each type of business was calculated. By comparing the median frequencies with the frequency of each type of business found in Tampines, the categories with the largest differences suggested potentially underserved areas that could be worthwhile to start a business in.

4. RESULTS

4.1. Visualisation of neighbourhood locations

The resulting map (Figure 1) shows that the dataset did indeed cover the majority of Singapore.
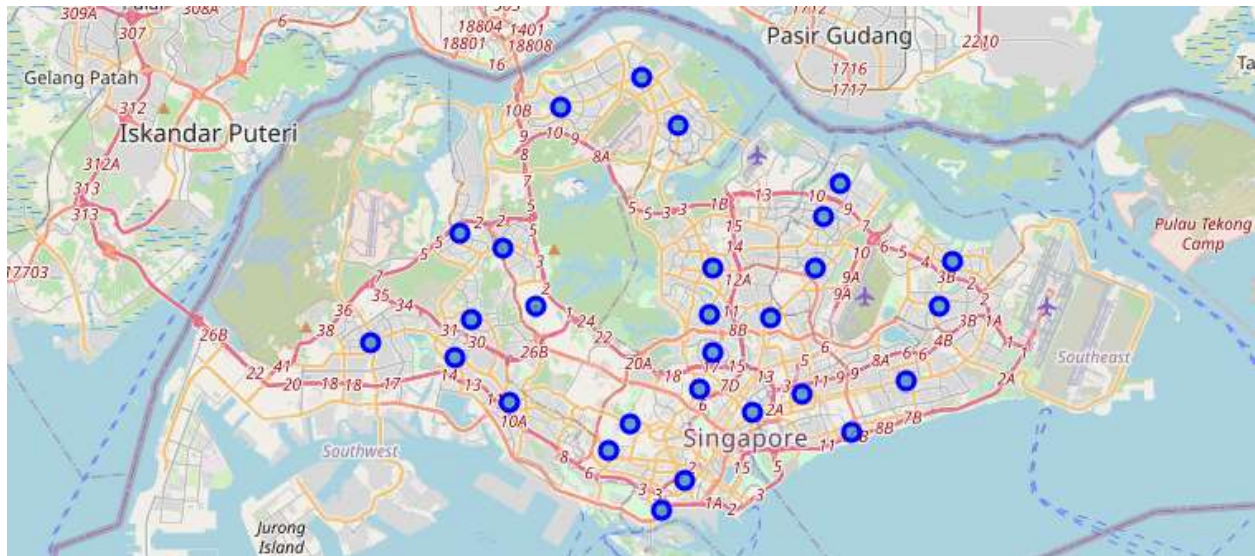

Figure 1: Neighbourhoods extracted covered the majority of Singapore

4.2. Identifying the most similar neighbourhoods

The results visualised using the *folium* package (Figure 2) showed that each neighbourhood identified was assigned a cluster, and that there were indeed a total of three clusters. Six other neighbourhoods (in purple) were assessed to be similar to Tampines.
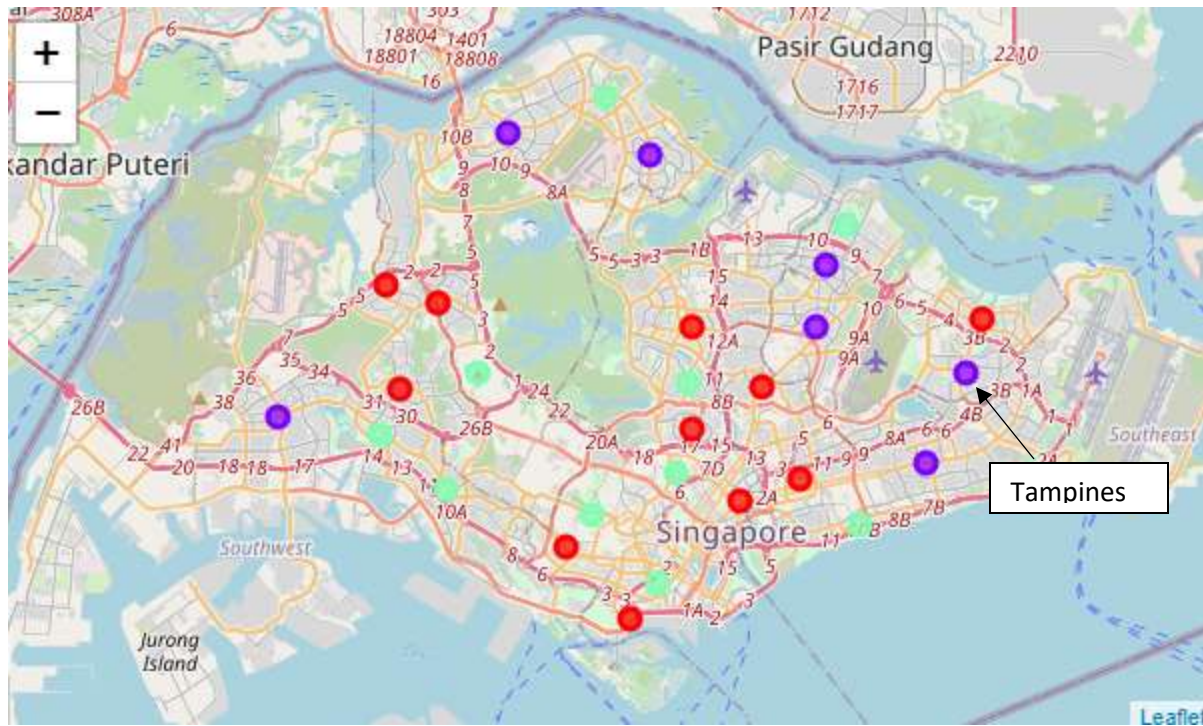
Figure 2: Neighbourhoods split into three clusters

4.3. Extracting information on businesses in each neighbourhood

The results visualised using the *folium* package (Figure 3) verified that the search radius was acceptable, i.e. results were still within the neighbourhood.

Figure 3: Search results were within the neighbourhood of Bedok

4.4. Identifying potentially underserved areas in Tampines

The ten business categories that had the largest differences between the current frequencies in Tampines (blue dots) and the median frequencies in the similar neighbourhoods (box plots) are shown in Figure 4.
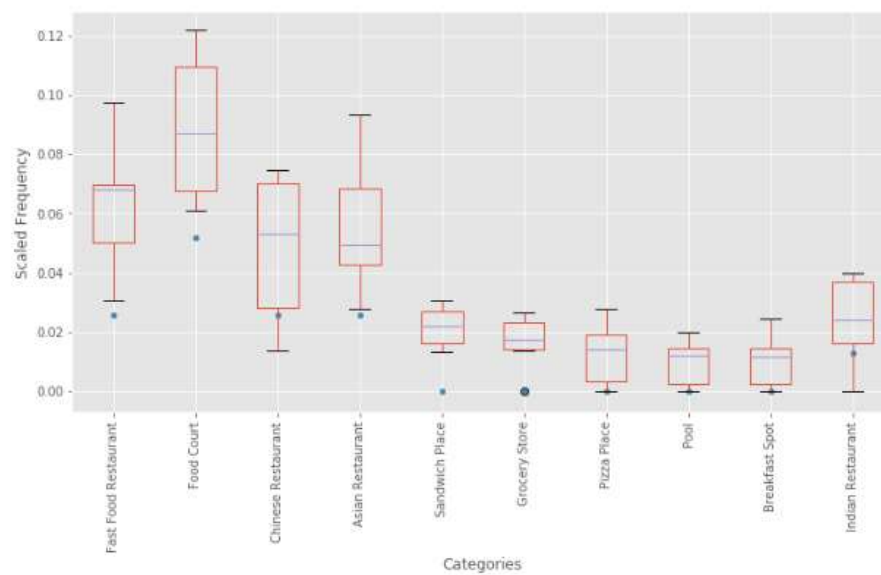

Figure 4: Top 10 business categories to go into in Tampines

5. DISCUSSION

5.1. Recommendation

The business categories that had the largest differences between the distribution in the similar neighbourhoods and Tampines suggested that these categories were underserved. With this, a preliminary recommendation for potential businesses to open in Tampines can be made (Table 1).

Fast Food Restaurant
Food Court
Chinese Restaurant
Asian Restaurant
Sandwich Place
Grocery Store
Pizza Place
Pool
Breakfast Spot
Indian Restaurant

Table 1: List of potential businesses to open in Tampines

5.2. Future Work

The distribution of shops among the top 10 categories appear similar between Tampines and the median of the six other similar neighbourhoods (Figure 4), hence supporting the parameters used for the K-Means clustering of the neighbourhoods and giving confidence to the preliminary recommendations.

However, a comparison between Tampines and neighbourhoods in a different cluster also yielded results that seemed similar (Figure 5). This suggests the need for refinement in the parameters used for clustering so as to obtain more distinct distributions of businesses in each neighbourhoods, and increase confidence in the recommendation.
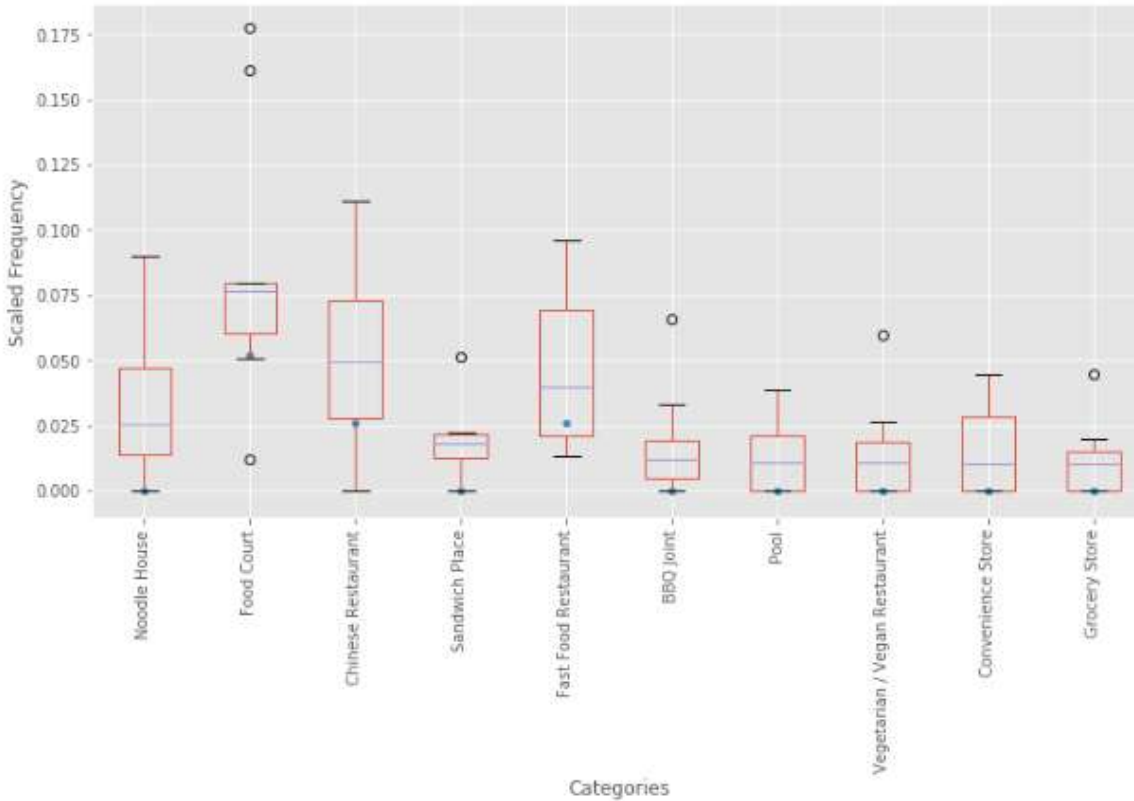
Figure 5: Distributions were similar with a dissimilar cluster, suggesting that refinements to the clustering parameters would be beneficial

6. CONCLUSION

In this work, data on the distributions of the residents by age, type of dwelling, and household size published by the Singapore Department of Statistics were extracted and analysed using K-Means clustering to identify similar neighbourhoods to Tampines, a neighbourhood where a prospective business owner would like to open a shop in. Information on the businesses that operate in the similar neighbourhoods were extracted using the Foursquare API to establish an expected distribution of the various business categories in Tampines, based on that in the similar neighbourhoods. By comparing the expected distribution against the current existing distribution in Tampines, recommendations on potentially underserved business categories can be provided to the prospective business owner.