Segmentation for a Restaurant Location Planning

Michael Dubem

May 23, 2021

2. Data Acquisition and Cleaning

2.1 Data Sources

The research was focused on the boroughs and neighbourhoods of Toronto, Canada, majorly because this is a developed region and a lot of data about this region is readily available online. For some less popular location, on ground/field survey might be essential for an applicable analysis. A lot of the data needed for the market analysis can be gotten through location APIs. The Foursquare Location API was used in this analysis to get crucial details about the geographical locations under scrutiny. Other APIs and platforms like the Geocoder was used to get the geographical coordinates of the areas under scrutiny and google maps was used to get the approximate land area covered by the respective boroughs of Toronto. The research also covered web scraping Wikipedia: (https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=prev&oldid=926287641) to get data like the list of neighbourhoods in each borough and their respective postal codes.

2.2 Data Extraction and Cleaning

The first data gotten was scraped from the Wikipedia page for list of postal codes in Toronto, and was loaded in to a data frame. Unassigned neighbourhoods and boroughs were cleaned off the table, neighbourhoods with the same postal codes were merged into the same rows and some neighbourhood names were changed. The next step was getting the geographical coordinates of each postal code and loading it in to a data frame and merging it to the previous list of boroughs and neighbourhoods, so each neighbourhood was matched to its respective coordinate. The coordinate for each borough was also gotten and loaded into a data frame, because the boroughs were to be analysed first before analysing the neighbourhoods of the optimal borough. The land area of the respective boroughs was also manually measured and merged into the existing data frame, so that any location call made on each borough would reflect the true size of the borough in relation to others. Instead of having to go through the stress of analysing and visualizing all the neighbourhoods, it seemed like a more efficient approach to analyse the optimal borough, then focus on only the neighbourhoods of that borough. The final part of the data extraction was using the Foursquare API to call up the number of venues in some certain categories that could give us insight into the nature of the demand and supply, as well as the competition in the areas under scrutiny. The categories covered were split into three groups, which was supply, demand and competition. For supply, the count on the number of supermarkets and shopping malls in each borough was extracted. The concept behind these venues is that, proximity to shopping malls and supermarkets means readily available raw materials for operations. There are other sources of supply, but these two were the focus in this research. For demand, the count on the number of colleges and universities, outdoors and recreation centres, hotels, professional complexes, and inhabitant's residences was gotten, due to the fact that these venues give an insight into the level of population and the type of population in the particular borough. Some age grades are more welcoming of the concept of eating out. The final category was competition, and the count for the number of restaurants in each borough was also extracted. After the optimal borough was selected, the same extraction process was repeated for the neighbourhoods in the optimal borough.

2.3 Feature Selection

K-means clustering was used to detect similarities amongst boroughs and neighbourhoods, and the major focus for checking for similarities was the features that fell under the three outlined categories (supply,

demand and competition). Hence, the features under each category were merged using a simple average. The clustering was then carried out on only the numeric columns in the data frame (except the coordinates)