**CAPSTONE PROJECT: The Battle of Neighbourhoods**

Kimberly Farah

Prof. Alex Aklson, Ph.D.

June 22, 2020.

PROJECT DESCRIPTION

A recent graduated from dental school wants to start their own business in dentistry. They are wanting to open their business in the Greater Toronto Area (GTA) where there is not already a dental clinic. They have done their research and found a centroid of population and existing dental clinics around using the Foursquare location data to determine the best geographical location as well as factoring in the existing dental clinics in those areas already and spacing an equal distance from all others.

INTRODUCTION

The business problem is where to open the new dental clinic in a large metropolis such as Toronto, Ontario. If the dental graduate opens a dental clinic in the wrong geographical location, they will lose money if they cannot adequately penetrate the existing, likely saturated market. If there is already a dental clinic to support the current demand in a nearby neighbourhood, the graduate will not have enough clients to support their business. The aim, therefore, of this project is to obtain open source data from which to discern or calculate an appropriate place to establish a thriving dental clinic.

METHODOLOGY

To conduct this research, Python was the sole resource for digesting, processing, and calculating the data. Sources of data were open source and constituted the ubiquitous Wikipedia, the Statistics Canada website – a website administered by the Government of Canada, and Foursquare – a pioneering service that collates geographic information regarding all sorts of locations, especially businesses, and user-generated information about the locations such as customer ratings and reviews.

Wikipedia was consulted to obtain a list of postal codes for the city of Toronto; this study hinged on postal codes as the link between the myriad data. Statistics Canada provided population data for the different areas of Toronto, namely the postal codes. Incorporating population data into the postal code data identified the population distribution within the city.

A centroid calculation was conducted independent of any data from Foursquare to serve as a control calculation to determine, solely from population, where to locate the clinic.

Then, Foursquare data was loaded into Python and applied to the problem. A Foursquare search for the term “dental”. The resulting data was cleaned and used to obtain a second centroid, this time for a location some distance away from all others.

The mean of the resulting two centroids was calculated as they, quite incidentally, were a very short distance from the other. This mean location was therefore deemed the best for the graduate to situate their enterprise.

CONSIDERED DATA

The data search began by scraping population data by postal code from Wikipedia[[1]](#footnote-2) to cover the entire GTA. Then, the data was converted into a dataframe for manipulation. Wikipedia provided not only postal codes, but the boroughs and neighbourhoods which belonged to them. Quite a few postal codes were marked as “Not assigned” under the Borough and Neighbourhood colums as they codes had not been assigned as of the date of the study. Such “Not assigned” entries were dropped.

Next, Statistics Canada’s website[[2]](#footnote-3) for population data concerning the entire country of Canada was scraped for population numbers as they concerned all the country’s postal codes; this data, dated as it may have been, was gathered from a census conducted during 2016. There were approximately 1,642 entries for postal codes. A dataframe was created from the results of the scrape. The postal codes were then narrowed down to analyze relevant codes only, namely those beginning with “M”; postal codes beginning with a different letter were dropped from the dataframe, leaving only 104. Irrelevant columns were dropped for a more concise view of the information. Finally, the column holding the postal code information had to be renamed “geographic name” to “Postal Code” to simplify the merger of both of the newly created dataframs.

In order to pinpoint the areas denoted by the postal codes and compare said areas to the precise data from Foursquare, this project required latitude and longitude coordinates for each postal code. This desire was fulfilled by a previous lab during the Applied Data Science course offered by Coursera. The data was provided in the form of a CSV file.[[3]](#footnote-4) This was read into Python, converted into a dataframe and also added to the previously merged dataframe. The result was one clear table containing all necessary information independent of Foursquare.

The second set of necessary data, as stated previously, was extracted from Foursquare. A conception of the number and exact location of dental clinics currently in existence was necessary for such a consideration as the one undertaken here. Therefore, a Python get command, intended to call all entries of the word “dental”, was entered into the Jupyter notebook and specified the focal point as the latitude and longitude coordinates of the city of Toronto, latitude 43.65, longitude -79.38. Very large limits were used in the get command to allow for the most precise analysis possible; the radius limit was set at 100 kilometres while the result limit was set at 10,000. It was unknown, with a city as large as Toronto, how many clinics would be found. Further, it may be the case that Foursquare limits the number of results provided per search. The search resulted in 50 clinics located near the downtown Toronto area.

The Foursquare data was requested and received in JSON format. This simplified its transformation into a Python datafame and the dropping of all unnecessary columns once again; only columns relevant to name, category, and location were maintained.

Finally, both the population dataframe and the Foursquare results dataframe were converted into arrays to allow a centroid calculation of both sets of latitude and longitude coordinates. The population centroid was weighted, taking into account the size of the population associated to each set of coordinates. The centre of the population was found to be at latitude 43.71, longitude -79.40. The centroid of all latitude, longitude coordinates of the dental clinics was calculated to place the proposed clinic some distance away from all others. This calculation resulted in the coordinates latitude 43.66, longitude -79.39. Incidentally, both sets of coordinates presented locations approximately 6.5 kilometres away from each other. To refine the entire process, a mean (average) of the two sets of coordinates was calculated resulting in the final location.

RESULTS

After running the weighted population centroid and a centroid of the locations of all 50 dental clinics, an average was taken to determine the best location for the clinic. These final coordinates were entered into Google Maps to determine what surrounded the area. Latitude 43.68, longitude -79.39 positioned the prospective clinic at 10 R Price Street, Toronto, ON, M4W 1Z4. Situated there were a large number of apartment buildings surrounded by the Pricefield Road Playground, the Toronto Lawn Tennis Club, and various well-known franchises and restaurants; this, it must be said, proved to be a very promising location.

DISCUSSION

The mean of both centroids of population and dental clinic location was determined to be the ideal since it took into account the population in that given area and the proximity to other dental clinics. However, it may be the case that, rather than using the centroid of other dental clinics, calculating a point equidistant from all other dental clinics in the area may have provided a more feasible and profitable locale. Foursquare or Google Maps depictions of such a location would provide considerable information towards the graduate’s efforts.

Another factor of consideration for the location of a dental clinic should be the financial ability (household income) of the population surrounding the new business as well as cost of the purchase or rental costs associated to the proposed new project. As the city of Toronto is being considered in this case, it may very well be that that such real estate are prohibitive. These are important factors to determine long term sustainability of the dental clinic.

One final, interesting point of discussion might also be that the determined location is near a park that supports a green space which coincides with the natural serenity of a dental clinic atmosphere and promotes healthy living.

CONCLUSION

Several factors were considered before determining the exact location for a proposed dental clinic to be situated in the city of Toronto: other existing dental clinics in the nearby geographical location and the population to support the enterprise. A specific and promising location was found; however, a cost analysis, which was beyond the scope of this study, would provide critical information for the entrepreneur.

1. https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M [↑](#footnote-ref-2)
2. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&S=22&O=A [↑](#footnote-ref-3)
3. http://cocl.us/Geospatial\_data/Geospacial\_coordinates.csv [↑](#footnote-ref-4)