**Capstone Project the Battle of Neighbourhoods**

Kimberly Farah

Prof. Alex Aklson, Ph.D.

June 21, 2020.

PROJECT DESCRIPTION

A recent graduate 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 their business will likely succeed; that is, a location selected based on population and taking into consideration the locations of other dental clinics. The graduate has done their research and, through postal code, population, and Foursquare data available online, decided upon an area by calculating a weighted-centroid of the population of Toronto, adjusted based upon existing dental clinics. The scope of this study was limited to the two variables noted; however, considerations such as alternate location calculations and other independent variables, such as rental costs, are discussed.

DATA DESCRIPTION

Information and data for this study was readily available and accessible online; a graduate or entrepreneur, or dentist would have the means to study the problem with extremely precise details and achieve a result down to the precise latitude and longitude coordinates of a proposed enterprise.

Location considerations for this study were built upon postal code data, for Toronto, Ontario, which had been introduced during a previous Capstone Applied Data Science lab; this data was obtained from a Wikipedia website which supplied the boroughs and neighbourhoods belonging to each postal code. Latitude and longitude coordinates, while available online, were provided in the form of a CSV file by the lab instructor. Meanwhile, 2016 census data, listing the population of Toronto, segmented by postal code is also readily available online through the Statistics Canada website. Foursquare was used to collect the names and longitude and latitude coordinates of currently existing dental clinics in the city of Toronto.

The Python code, as will be demonstrated in the associated Jupyter notebook published on Github, scraped all four different sources – Wikipedia, Statistics Canada, CSV file, and Foursquare to obtain and form dataframes of the data. These dataframes were manipulated to normalize, organize, and clean the data for use. The dataframes were then transformed into NumPy arrays to allow for the calculations of population and dental clinic centroids. This provided approximate latitude and longitude coordinates for the centre of the massive Toronto, population and the centre of the cluster of dental clinics in the area.

As it happened, both sets of coordinates were extremely close to each other, providing quite a precise area to situate a new business. However, the mean of both sets of coordinates was obtained to determine, to the degree, what the exact location should be. The result was a prime spot, near a sizable green space and some distance away from its nearest neighbouring dental practice.

The specifics as to coordinates will be provided in the Jupyter notebook and summarized in the conclusion section.