

# How to choose the perfect neighborhood for a new dental office in Toronto, Canada



## Problem

Imagine you are a young aspiring dentist. You studied and worked hard and now ready to go into business on your own. It is a big decision and very expensive too. The bank is willing to lend you money but wants to make sure that you can earn enough to cover the loan payments.

How do you decide on the area? There are a lot of dentists out there, in every neighbourhood. How would you even compare one neighbourhood to the other? These are the questions that I will try to tackle in this project.

## Hypothesis

I'm starting with a hypothesis that it is possible to assess and rank neighbourhoods based on the level of competition. In other words, the number of dental offices operating in each neighborhood is mostly driven by the total number of residents or households in that area. The more people live in the neighborhood, the more dentists will be required.

## Interest

Location, location, location. Whether you are a dentist, accountant, mechanic or a banker assessing the location for a future business you need to know how to do it right. You need to know what is important and what is not.

# Data acquisition and cleaning

Data sources:

1. [https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
2. [https://cocl.us/Geospatial\\_data](https://cocl.us/Geospatial_data)
3. <https://Foursquare.com>
4. <https://www.yellowpages.ca/search/si/1/Dentist/North+York+ON>
5. <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/File.cfm?T=1201&SR=1&RPP=9999&PR=0&CMA=0&CSD=0&S=22&O=A&Lang=Eng&OFT=CSV>

To get started with my research I needed to gather information about the city of Toronto and its neighborhoods. There was no relevant dataset I could find online so I decided to build it myself by scraping postal code and neighborhood data off the Wikipedia page (1). Postal codes without matching neighborhood names were ignored at this point.

Toronto has 10 boroughs consisting of 98 distinct neighborhoods. North York is the largest borough, consisting of with 24 neighborhoods. The next step, was to download geospatial information for each postal code (2).

As was just mentioned, North York is the largest borough of Toronto and so I decided that we are going to limit our research to this area of the city. There was no need to keep the entire dataframe so I reduced it to include only the information pertaining to North York.

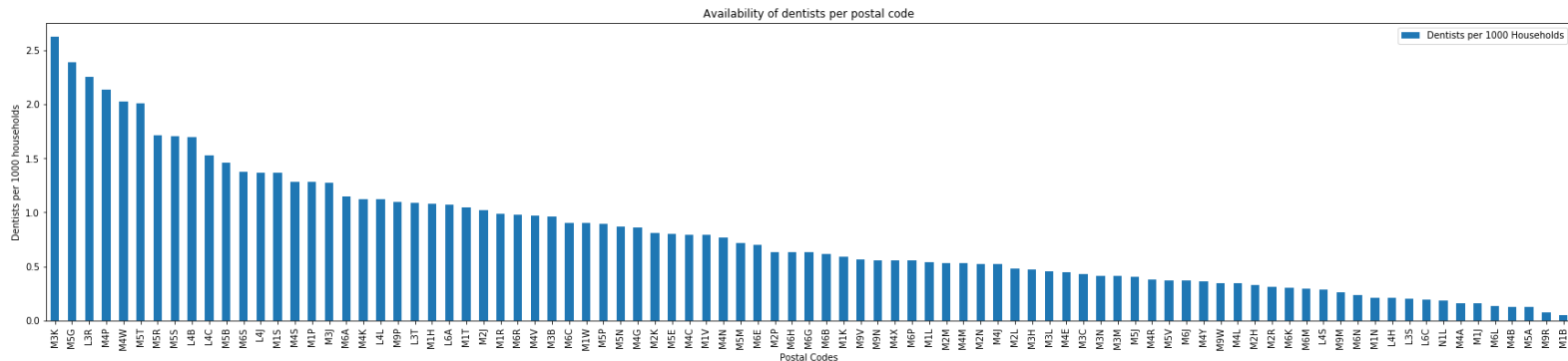
Once I had all neighborhoods and their respective geospatial data I used Foursquare API (3) to download all records for dentists in North York. Unfortunately Foursquare had only very few records and I had to search for another way to get the list of dentists in North York. I researched several publicly available datasets but the only source with reliable data I could find was yellowpages.com. There was no public API available to download necessary data so I used web scraping again.

While reviewing the dataset I noticed that some dentists were listed twice for those who paid to have their listing at the top of the page. I removed such duplicates. Some dental offices have more than one dentist working there. Convenience matters and so, even if there is a large dental office within a 5-10 minute drive from your home you may prefer to go a solo practitioner who is within a walking distance from you. At this point I've decided to narrow my dataset to include only dental offices.

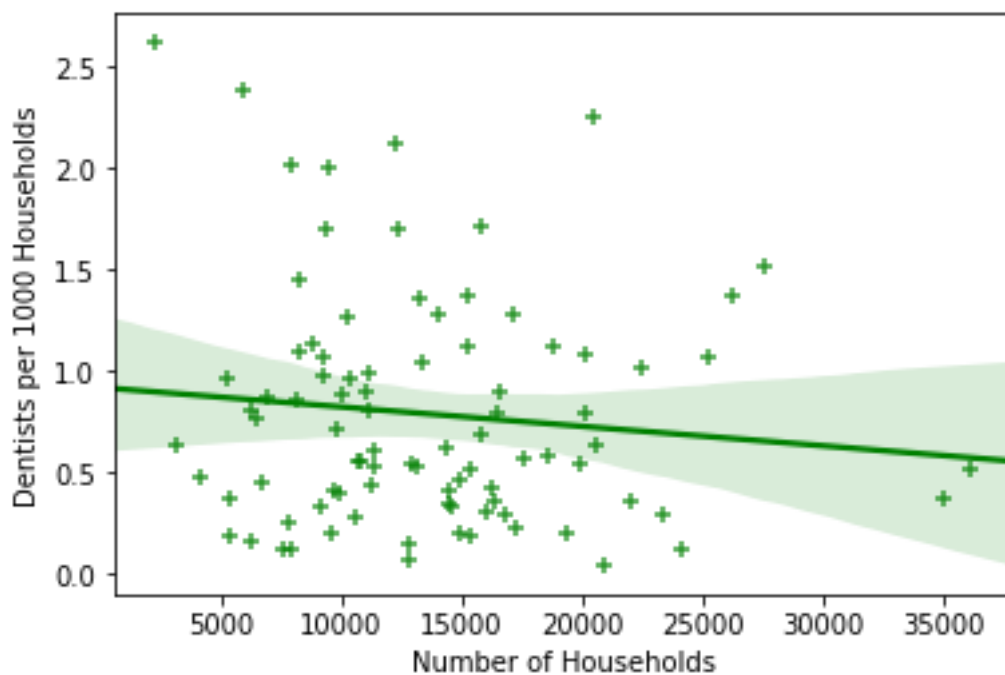
The next step was to get the total number of residents living in North York. For that purpose I used the dataset prepared by Stats Canada (5).

# Exploratory Data Analysis

While reviewing the dataset's stats (`df.describe()`) I noticed that there were some outliers in our dataset skewing the averages which had to be eliminated. I then sorted the table by Dentists per 1000 Households column and plotted it to see if there was a clear pattern in the data.



Unfortunately there was no clear pattern observable. Checking if there is a linear relationship between the number of dentists and the total number of households in each postal code.



At this point we can disprove our hypothesis that the number of dental offices in each neighborhood is dependent on the total number of households in the area.

Just to be absolutely sure, I decided to check if there was a linear relationship between the total number of dental offices, total number of residents and total number of households. We assume that population density (Number of residents / number of households) may be dependent on whether the area has more apartment buildings vs single homes. Single homes should have higher density per household.

A simple multiple regression was used. It was fitted using the entire dataset but it was not able to explain the bulk of the variation, producing the  $R^2$  of 0.24.

## Conclusion

My initial hypothesis that the number of residents/households can predict the total number of dental clinics in the neighborhood was successfully disproved leaving us with the question of how else we could assess suitability of any given area. The variance in the number of dentists per 1000 residents in each area tells us that there are significant factors which influence the number of dentists in the area. It could be availability of commercial real estate available for lease, attractiveness of the area based on the average wealth of its residents etc.