**Capstone Project: Final Report**

**Helping home buyers choose which neighborhood will best meet their needs.**

**Robert M. Taylor, PhD**

1. **Introduction**

This project is intended for real estate agents who frequently work with out of town clients who may be unfamiliar with the area. Please imagine that you are a real estate agent in Toronto and often have a common problem when helping out of town clients. Many of your clients, who are unfamiliar with the Toronto area, will narrow down their choices to 2 or 3 houses that they liked. In making a final decision, your clients often tell you what types of venues and/or businesses they would like in their ideal neighborhood. The clients then often ask you to help them decide which of the 2 or 3 houses they are looking at are in a neighborhood with their ideal “wants.” You have been learning data science, in the hopes of getting an upper hand in your industry and decide that you can use your newly acquired data science skills to help your clients.

You have a new client that has recently landed a lucrative data science job at a university in Toronto. You have shown the client and their partner several houses and they have narrowed their choice down to 3 lovely homes. Each home is in a different neighborhood in Toronto and, as usual, the clients have asked for your help in deciding which neighborhood would fit their needs the best. Your clients have picked 3 homes in the neighborhoods of 1) Berczy Park, 2) Queen’s Park, and 3) Rosedale. The clients have told you that they MUST have 3 things in the neighborhood they choose that are extremely important to them. They tell you that these 3 things are 1) a park, 2) coffee shops, and 3) a gym. Your project is to now utilize your new skills in python, data analytics, and Foursquare to determine which neighborhood is best for your clients.

1. **Data**

To solve this problem, we will use 3 data sources. First, we will use the following Wikipedia page, <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>. We will scrape this page to get the postcode data from the table and then transform it into a dataframe. The resulting dataframe will contain 3 columns for ‘Postcode’, ‘Borough’, and ‘Neighborhood.’ We will clean and analyze this data and use Folium to visualize the home locations and neighborhoods of interest. Next, we will use the geospatial data for Toronto. This data is accessed with: <http://cocl.us/Geospatial_data>. Finally, the Foursquare location data source to determine what venues (specifically those that our clients are interested in) are in the neighborhoods. We will then be able to utilize this data to make a final determination of which neighborhood is the best choice for our clients.

1. **Methodology**

We began by scraping the Wikipedia page, <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> to get the postcode data from the table. We then transformed the table into a pandas dataframe. The resulting dataframe (‘wiki’) contained 3 columns for ‘Postcode’, ‘Borough’, and ‘Neighborhood.’ We examined the resulting dataframe using ‘head’ and ‘shape.’ We next dropped any “Not assigned” boroughs and also corrected the spelling of the column “Neighborhood.”

We next made a new dataframe called ‘borough’ that only contained the postcode and borough information. We removed duplicate postcode entries from the borough column. Similarly, we then made a new dataframe called ‘neighborhood’ that only contained the “neighborhood” and the “postcode” features. This allowed us to then combine multiple neighborhoods with the same postcode into a single row for the postcode. This was easier to do without the borough data being included. We also changed any “Not assigned” neighborhoods to the name of the borough.

Next we loaded in the Toronto geospatial data (<http://cocl.us/Geospatial_data>) as a pandas dataframe and examined the resulting dataframe that we call “geo.” We also renamed the column “Postal Code” to “Postcode” to match the other two borough and neighborhood dataframes, respectively. We were then able to add the latitude and longitude features from the geo dataframe into the neighborhood dataframe. Finally, we also added back in the borough feature into the neighborhood dataframe. This resulted in a “neighborhood” dataframe that contained all of the information for postcode, borough, neighborhood, latitude, and longitude.

We next pulled out data for only the three neighborhoods of interest (i.e. Berczy Park, Queen’s Park, and Rosedale). We were then able to utilize the geopy library to get the latitude and longitude of Toronto so that we could then create folium maps with the neighborhoods superimposed on top. Following this, we were then able to utilize the Foursquare location data to answer the question of which neighborhood best fit the client’s needs.

To use Foursquare, we first defined our credentials and then located the 3 neighborhoods of interest in our dataframe to locate the latitude and longitude again for each neighborhood. We then began analysis with the first neighborhood (Berczy Park). The general method flow for analysis of each neighborhood with Foursquare was as follows:

1. Get latitude and longitude for neighborhood
2. Get URL request with a limit of 100 and radius of 500.
3. Send GET request
4. Write function to extract venues
5. Clean json and structure into panda dataframe
6. Narrow search down to only client venues (park, coffee, gym)
7. Review results
8. Repeat process for other 2 neighborhoods

Next, to get all the Foursquare results from each neighborhood into a single dataframe, we created a function to do all 3 of the previous Foursquare calls in a new dataframe. We then ran this function to create the new dataframe called “Toronto\_venues.” We then eliminated all venue from this dataframe EXCEPT the client venues (park, coffee, gym). This new dataframe was called “WishList.” Folum maps were generated to visualize all of the park, coffee, and gym venues within the 3 neighborhoods. One hot encoding was then used to generate new features from the different 3 venues and then grouped and the mean taken, to determine the frequency of each venue occurrence in each of the 3 neighborhoods. Finally, K-means was used with a k=3 and clustering was used with subsequent visualizations using folium. At the end, we ended up again reviewing the “wantlist\_grouped” dataframe which showed the frequency of the client venues in each neighborhood.

1. **Results**