**Which Neighborhood to Live in Toronto?**

**Stacey Lee**

**Feb 14, 2020**

1. **Introduction/Business Problem**

My company has expanded our business to east coast of North America and I was asked to move to Toronto for a year to take care of the on-site implementation of our solution and training in customer environments. Given that my role is required to travel 50%, I have better options in terms of choosing which neighborhood to live. However I want to be able to easily access the coffee shops, cafés, and restaurants and very importantly I want to be close to yoga studio and gym for the workdays and the parks and/or natural scenes on the weekends. With these criteria in mind, I need to do neighborhoods analysis. I looked into location technology platform and found Foursquare provides the Places API that can enable neighborhood location discovery.

This project aims to leverage the Foursquare location data to explore and compare neighborhoods of Toronto. Just like most of problems data scientists target to solve, it meant to help a group of stakeholders, the audiences can include the long term travelers or business visitors and by applying the same methodology, when data sources are available, it can be for people who want to move to major city in the world.

1. **Data Sources**

The data used for this project includes the followings:

* The data set on Toronto neighborhood data is available on Wikipedia page. It is a list of postal codes in Toronto, Canada (NOTE: Postal codes beginning with M are located within the city of Toronto in the province of Ontario.) This wiki page will be performed web scrapping to extract the necessary information: Postcode, Borough and Neighborhood into a postal codes table
* List of Coordinates data set is required in order to utilize the Foursquare location data, the latitude and longitude coordinates of each postal code are obtained from List of Coordinates (<http://cocl.us/Geospatial_data>)
* Foursquare data set: Foursquare provides venues and their assigned categories information. Foursquare enables developers to obtain these data points via Foursquare Place API; the Foursquare API is used to explore and segment the neighborhoods

References:

1. Toronto neighborhood data on Wikipedia page: https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M
2. List of Coordinates: <http://cocl.us/Geospatial_data>

This link provides geographical coordinates of each postal code

1. Determining the number of clusters in a data set (elbow method): <https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set>
2. Foursquare API endpoints from Foursquare Developers: <https://developer.foursquare.com/docs/api/endpoints>
3. A view on cities - Toronto:

<https://www.aviewoncities.com/toronto/queenspark.htm>

1. Wiki Toronto:

<https://en.wikipedia.org/wiki/Toronto>

1. Cluster Analysis

<https://en.wikipedia.org/wiki/Cluster_analysis>

1. K-means (decide the optimal number of clusters):

<https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set>

1. Feature learning for machine learning:

<https://en.wikipedia.org/wiki/Feature_learning>

1. Google maps:

[https://www.google.com/maps](https://www.google.com/maps/@37.2817103,-122.0494399,15z)

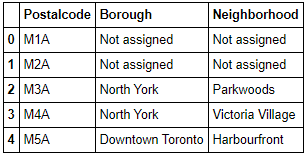
1. **Methodology**

This project aims to analyze different neighborhoods of Toronto by exploring Foursquare location data, I use the Foursquare API to explore neighborhoods in the city of Toronto and group them into similar clusters, i.e. use the explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters, apply the k-means clustering algorithm to complete this task. Finally, use the Folium library to visualize the neighborhoods in Toronto and their emerging clusters.

* 1. **Web Scrapping and Data Wrangling**

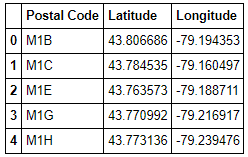
First import libraries required to get the data in structured format by performing web scrapping, i.e. pass out postal code Wiki URL into **BeautifulSoup (BS)** which analyzes the HTML content text and identify the HTML elements from which to extract postal codes table.

The next step, data wrangling is to transform and map the data from the cells in postal codes table into a pandas dataframe. Here are the first 5 rows of this dataframe:



In order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighborhood. Geocoder Python package is usually used to get the geographical coordinates, however given that this package can be very unreliable, use a link to a CSV file that has the geographical coordinates of each postal code: <http://cocl.us/Geospatial_data>, a dataframe is built from this geospatial data CSV file.

Here are the first 5 rows of this dataframe:



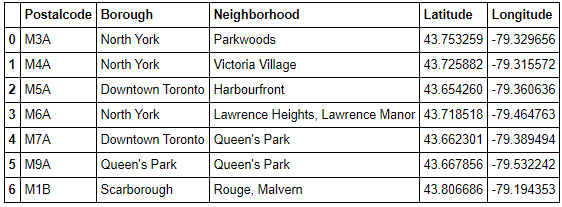
* 1. **Data Cleaning**

At data cleaning phrase, the following steps are performed to postal codes dataframe (that was built from web scrapping):

* Remove the row data with borough that is "Not assigned".
* More than one neighborhood can exist in one postal code area, combine these neighborhoods into one row with the neighborhoods separated with a comma
* If a cell has a borough but a Not assigned neighborhood, then make the neighborhood the same as the borough.

Then data aggregation is performed based upon common column, postal code, by merging two dataframes: postal codes dataframe and geospatial dataframe into a new 5 columns dataframe.

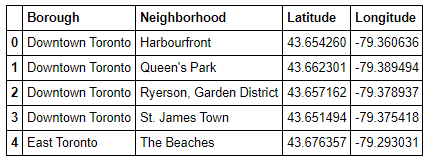
Here are first 7 rows of new dataframe shown below:



* 1. **Exploratory Data Analysis (EDA)**

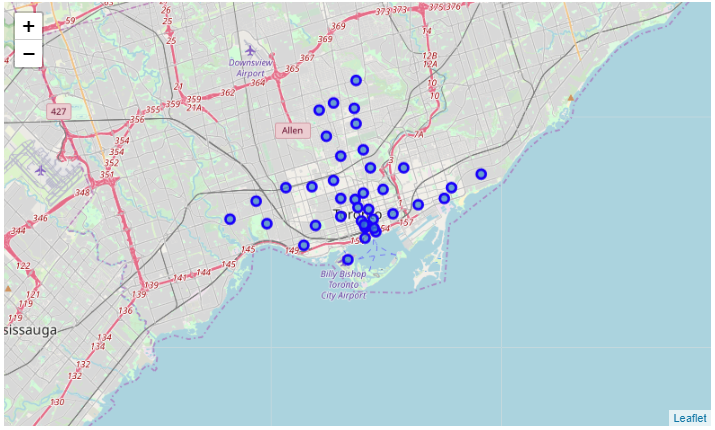
With data cleaned as expected I need to analyze data sets in order to summarize the main characteristics of neighborhoods in Toronto in order to make choice the neighborhood that best fit my interest. EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. I use the Foursquare API to explore neighborhoods in Toronto. I utilize the Foursquare API to explore the neighborhoods and segment them, i.e. use the explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters, use the k-means clustering algorithm to complete this task. Finally use the Folium library to visualize the neighborhoods in Toronto and their emerging clusters.

NOTE: for the simplicity I only work with boroughs that contain the word "Toronto". Here is the first 5 row of dataframe for Toronto geo data with boroughs that contain the word “Toronto”:



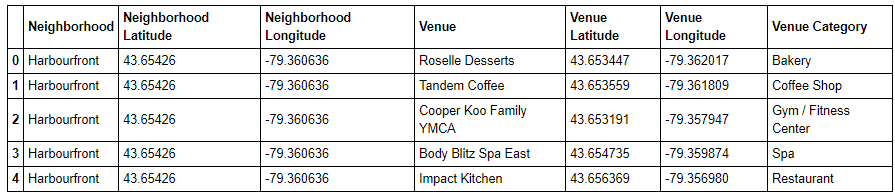
* + 1. **Data Visualization**

I have used Folium to create map of Toronto using latitude and longitude values, here is the Toronto map with the neighborhoods in it:

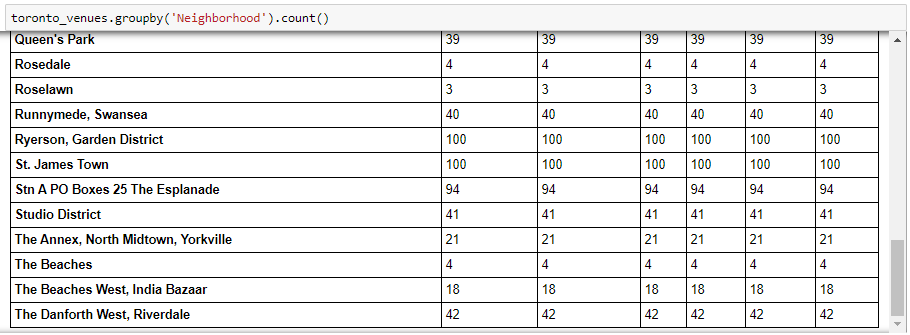


* + 1. **Clustering**

For the process of utilizing Foursquare API to explore Toronto’s neighborhoods, I first work on one neighborhood (the very first in the dataframe) and make sure things work as expected: explore via GET request: <https://api.foursquare.com/v2/venues/explore>, to get the top 100 venues that are in the first neighborhood within a radius of 500 meters, I filter the columns I am interested in from the results, i.e. venue name, venue categories, venue latitude and venue longitude. Also found out the number of venues returned by Foursquare for the first neighborhood. I then repeat the same process to all the neighborhoods in Toronto and create a new dataframe called toronto\_venues, here are the first 5 rows:



Also how many venues returned for each neighborhood is checked (via count method), here are some outputs:



* 1. **Machine Learning and Inferential Statics Testing**

For the purpose is to discover patterns in Toronto venue data and group the inputs into categories, as in feature learning, I have chosen centroid-based clustering model for analyzing venue categories type of data that Foursquare provides; the unsupervised learning algorithm, K-means is used to find the structure in the clustering of data points.

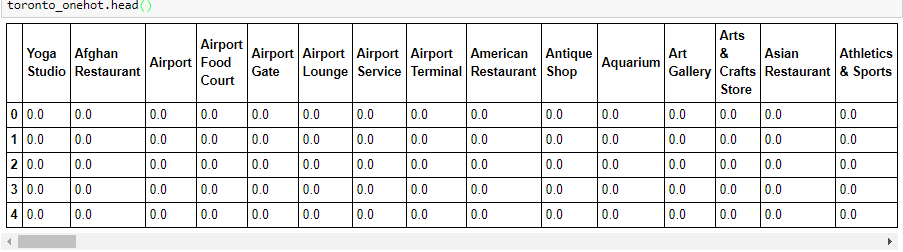
* + 1. **One-hot Encoding**

Many machine learning algorithms cannot operate on label/categorical data directly. They require all input variables and output variables to be numeric. In general, this is mostly a constraint of the efficient implementation of machine learning algorithms rather than hard limitations on the algorithms themselves. This means that I need to convert categorical venue data to numerical data.

Given that categorical venue data has no ordinal relationship, I use one-hot encoding to apply the integer representation.

It refers to splitting the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Each column contains “0” or “1” corresponding to which column it has been placed.

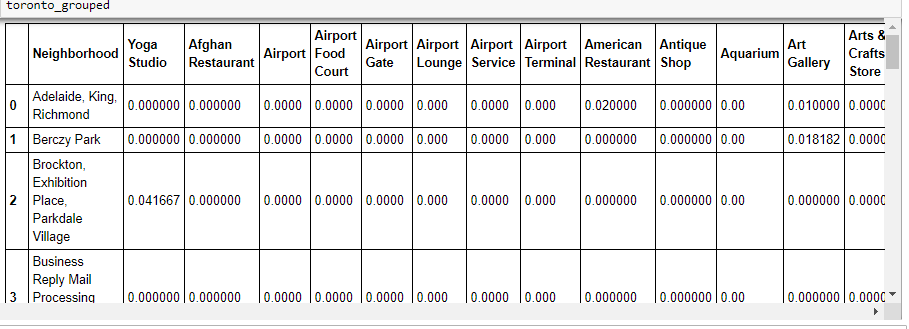
Here are some outputs of one-hot encoding:



* + 1. **Inferential Statics Testing**

In inferential statistics, data are analyzed from a sample to make inferences in the larger collection of the population. The purpose is to answer or test the hypotheses. There are two main areas of inferential statistics: 1) estimating parameters and 2) hypothesis tests.

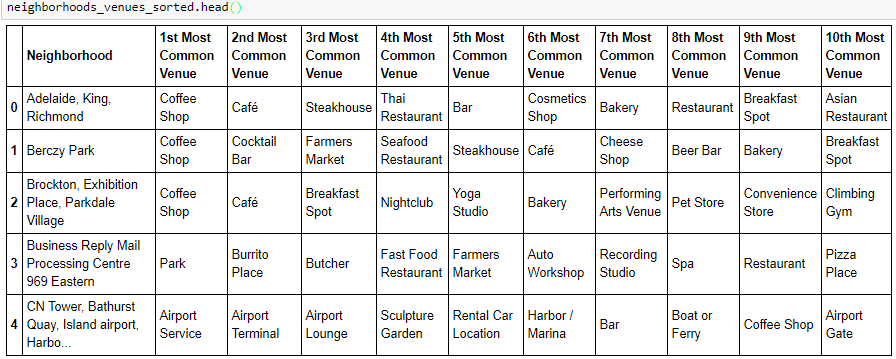
One of the simplest inferential tests used by data scientists is to compare the average performance of two groups on a single measure to see if there is a difference. With this project, only venue category/group based, it is not that applicable to perform hypothesis tests. I have taken “estimating parameters” approach by taking a **sample mean** from the sample data (i.e. venue data), i.e. using the dataframe created from one-hot encoding I have *grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category*. This process creates a new dataframe, toronto\_grouped, here are some outputs of **frequency distribution**:



I then checked and confirmed the new dataframe size has correct number of neighborhoods and number of venue categories with [dataframe].shape:



Finally, from the data of venues and the frequency of occurrence of each category, a new dataframe, neighborhoods\_venues\_sorted, for the top 10 venues for each neighborhood, is created, here are the first 5 rows:

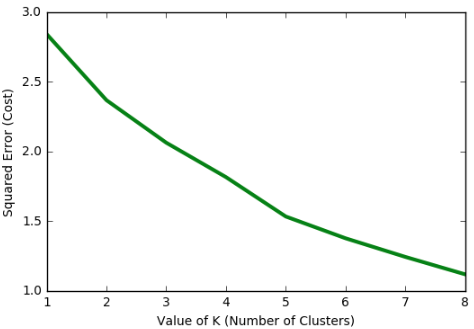


* + 1. **Clustering Analysis**

K-means clustering is the best known unsupervised learning algorithm for clustering analysis. I use the centroid-based clustering to organize and group Toronto neighborhoods into the clusters and use the most widely-used centroid-based clustering algorithm, K-means clustering algorithm to complete this task.

First find out the optimal value for K with popular elbow method, i.e. find the k cluster centers and assign the objects (i.e. ‘neighborhoods’) to the nearest cluster center, such that the squared distances from the cluster are minimized. NOTE: The basic idea behind this method is that it plots the various values of cost with changing K. As the value of K increases, there will be fewer elements in the cluster and therefore average distortion will decrease.

Also I use K-means++ to ensure a smarter initialization of the centroids and improve the quality of the clustering. I then plot the cost against K values (number of clusters), here is the plot:



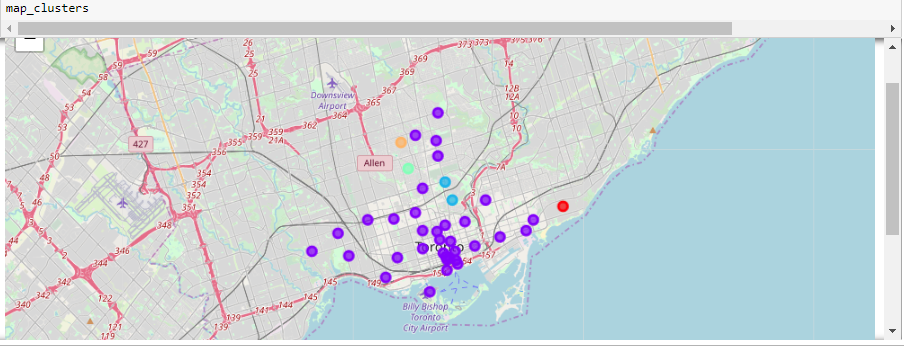
Using the above plot, the choice of K can be 2 or 5; the better choice for the given data is 5 as per error convergence, I therefore run K-means to cluster the neighborhoods into 5 clusters.

I then create a new dataframe, toronto\_merged that includes the clusters as well as the top 10 venues for each neighborhood:



The next step is to visualize the resulting clusters, use the Folium library to visualize the neighborhoods in Toronto and their emerging clusters:

* The majority of neighborhoods are in Cluster 1, 2 neighborhoods in Cluster 2 and one neighborhood in Cluster 0, 3 and 4
* It looks like my most interested neighborhood should be in Cluster 1 but this cluster includes 34 neighborhoods



* + 1. **Feature Learning**

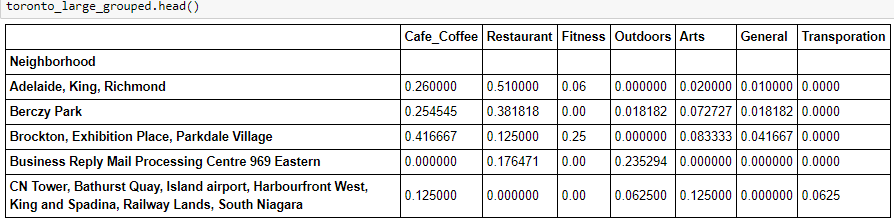
The ideal neighborhood to live in for me (and likely many people) should have the criteria: close to public transportations, easy access to outdoors and arts, in addition close to university and common access to daily life need. After examining each cluster, it is obvious; the best fit neighborhoods should be in Cluster 1 however it is hard to identify them manually, this means I need to apply feature learning.

In machine learning, feature learning is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data and allow a machine to both learn the features and use them to perform a specific task.

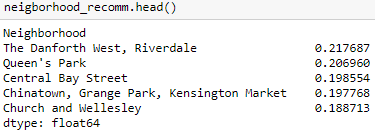
Currently there are 229 venue categories but some of them are similar so for simplicity and clarity group those to be seven higher level categories of interests as follow:

* Cafe\_Coffee - keywords (Coffee Shop|Cafe|Bakery|Breakfast|Sandwich)
* Restaurants - keywords (Restaurant|Sushi)
* Fitness - keywords (Yoga|Gym|Dance Studio|Sports|Fitness)
* Outdoors - keywords (Park|Trail|Field|Garden)
* Arts - keywords (Art|Museum|Gallery|Paintings|Sculpture)
* General - keywords(Supermarket|Grocery|Shopping|Hospital|Clinic|Salon)
* Transportation - keywords (Metro|Bus|Train|Boat)

Here are first 5 rows of this new broad categories grouped dataframe:



I then set the user priority ratings of most interested venues categories in this sequence: 1) Transportation, 2) Fitness, 3) Outdoors, 4) Arts, 5) Cafe Coffee, 6) Restaurants and 7) General. After weighting in these user priority ratings, a new dataframe for neighborhood recommendation, neightborhood\_recomm is created for the user preferred neighborhoods with the weighted average of user priority ratings against the frequency of seven broad categories for each neighborhood. Here are the top 5 recommended neighborhoods:



1. **Results**

After applying Machine Learning with Feature Learning, the top 5 recommended neighborhoods based upon user preferences on ratings priority (with aggregated, broader venues categories), i.e. 1) Transportation, 2) Fitness, 3) Outdoors, 4) Arts, 5) Cafe\_Coffee, 6) Restaurants and 7) General, are discovered. I can now decide the best neighborhood among three options: the first recommendation, The Danforth West or Riverdale and second recommendation neighborhood, Queen’s Park.

I have used three web sites: A view on cities - Toronto, Wiki Toronto and Google maps to further analyzing three options and then decided my preferences in the sequence: Queen’s Park, Riverdale and The Danforth West for the following reasons:

* Queen’s Park is bordered mostly by the buildings of University of Toronto, I can access many resources offered by the college, such as music, arts and other events.
* Comparing with The Danforth West , Riverdale is closer to the parks and gardens and it has better public transportation, Google maps show every 3-5 minutes metro schedule and only about 20 minutes to Queen’s Park

1. **Discussion**

Foursquare provides useful venues and venues categories data for analyzing the neighborhoods in Toronto at the high level however for the large city like Toronto, most common venues are restaurants and café related, the clustering analysis with carefully choosing the optimal number of clusters may not be able to summarize main characteristics of neighborhoods, grouping the neighborhoods in different ways may be more practical than common venues categories grouping. In this project after in- depth study from clustering analysis, the best fit neighborhoods should be in Cluster 1 which includes 34 neighborhoods, further feature learning analysis weighted in user preferences is needed in order to come up with neighborhood recommendation automatically.

While doing the initial investigation for this project, I have faced one of challenges of data analytics, i.e. collecting the useful data can be time consuming. At some point I wanted to do city comparison for San Francisco and Toronto but I could not find the longitude and latitude data associated with neighborhoods for S.F. city. A week later while doing other web searching I found this data in [S.F. Police Department Incident Reports](https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/data) . Through this process I have witnessed collecting good data sources at regular basis is practical approach for the data scientists.

1. **Conclusion**

This project aims to leverage the Foursquare location data to explore and compare neighborhoods of Toronto. In this study of clustering analysis, user preferred neighborhood happened to be in one out of 5 clusters that has the majority of neighborhoods therefore it is hard to come up with meaningful result - the same scenario may be common to the study of the neighborhoods in major cities in the world. After applying additional Machine Learning with Feature Learning - user preferences on ratings priority with aggregated, broader venues categories: 1) Transportation, 2) Fitness, 3) Outdoors, 4) Arts, 5) Cafe\_Coffee, 6) Restaurants and 7) General, the top neighborhood recommendation is then concluded.

There are many other factors to consider when choosing a good neighborhood to live in, such as crime rates, monthly rental or housing price, even community culture is important too. For the purpose of this project, one year stay in Toronto, these additional factors may not be as important comparing with the factors weighted in user ratings priority but they are important for longtime residents, in addition, different users may have their own preferences for the living environment. Ideally this project can be enhanced to target both short and long terms residents and make it more user friendly, for example configurable such that users can specify their own ratings priority preferences, run the analysis real-time and do the adjustment interactively.