Final Report

Author: Robert Ferdinand

Email: UOFLOUISIANA@GMAIL.COM

Table of Contents

Project Title: Suitable Toronto Neighborhood(s) for Ethnic Cuisine Restaurant Chain

Section 1: Introduction/Business Problem

A medium-priced, international foods restaurant franchise chain IPBS Foods Inc., specializing in south Asian and other international cuisines, based out of MyTown, MyState, USA, is planning to open a new franchise unit in Toronto, Canada. Since this is IPBS' first foray into Toronto and its first foray outside the United States, it needs to know where to place the new unit optimally in Toronto so that the franchise unit is able to:

(a) Attract sufficient amounts of customers during lunch and dinner hours in order for them to be profitable and successful.

(b) Be successful in attracting local people for employment at the franchise unit in positions such as waiters and food deliverers to local businesses as well as residences.

(c) Provide outreach to the local-area community via sponsorship of fund-raising events for service organizations for example.

(d) Contribute to the Toronto economy as a successful business unit and thereby bolster its financial standing in the United States.

In order to investigate suitable neighborhoods in Toronto, where such a franchise unit will be successful, one would need to analyze each neighborhood of Toronto. Such an analysis will include what neighborhoods host restaurants that serve international foods, especially south Asian, on their menus that may be evident from their names or categories. The presence of such clusters of restaurants would lean towards indicating that the new IPBS' franchise unit will be successful in that neighborhood(s).

Hence, IPBS decides to recruit a United States-based Data Science Company, DataTechWiz, Inc. to help them decide what would be the most optimal location for their new franchise unit in Toronto. To get the ball rolling, IPBS business personnel will meet with their business and technical counterparts at DataTechWiz to discuss their needs and expectations as regards the finished product needed. The primary objective of the project will be to find the top 2-3 neighborhoods in Toronto where the franchise unit may be started to get most optimal results. In return the total cost of the project to IBPS, payable to DataTechWiz, Inc. upon completion, will also be negotiated.

Once hired, DataTechWiz will use the services of data pertinent to the neighborhoods of Toronto as well as tools such as the Foursquare API location data, implemented via the PYTHON programming language, on the IBM Watson Studio to be able to assist them with their analyses.

During the analyses, DataTechWiz will provide IPBS a weekly update on their progress on the project. Upon completing their analyses, DataTechWiz will commit to a presentation to be attended by IPBS business personnel. In this presentation, DataTechWiz will present their final report and an outline of their methodology. Upon satisfactory completion and submission of the project by DataTechWiz, IPBS will be under contract to compensate DataTechWiz for their services.

As a side note, IPBS commits that if the new franchise unit is net profitable in their first year of business, then IPBS agrees to pay DataTechWiz an additional 10% of their annual profit for that year as an add-on bonus.

Section 2: The Data

To being its investigation into the project, DataTechWiz, Inc. embarks upon obtaining Toronto neighborhoods' data, pertinent to the project. DataTechWiz, Inc. will avail of the use of technology tools such as the Foursquare API location data via the PYTHON programming language on Watson IBM Studio.

The first data set consists of a table of postal codes for each Toronto borough. Each borough in turn resides in one or more Toronto neighborhoods taken together as "one" neighborhood, owing to their close geographical proximity. This data set is obtained from the city of Toronto and its authenticity is verified. If there is an unnamed borough, the postal code will be disregarded in the analysis. On the other hand if a neighborhood(s) is unnamed, it will assume the name of its respective borough. Twenty-five rows of this data are presented as a sample, in Data Set 1 in the Appendix section at the end of this manuscript. Initially Data Set 1 is in comma-separated-version (CSV) format. Later it will be read into a PYTHON pandas Data Frame to help facilitate the ensuing analysis. One may note that there may be several rows in this data set. Also this data would need to be thoroughly "cleaned" before any analysis follows.

Since DataTechWiz plans to use the Foursquare API location data, one needs the latitudes and longitudes for each neighborhood in Data Set 1. This data is not available in tabular format. However, data on the latitudes and longitudes of each Toronto postal code can be procured in a table format. To that end, Data Set 2 is used. This tabular data set contains the latitude and longitude for each postal code present in Data Set 1. This data set is also obtained from the City of Toronto and its accuracy is

guaranteed. Once again, as a sample, we only present twenty-five rows of this data in Data Set 2 contained in the Appendix section at the end of this manuscript. This data is in CSV file format. Like for Data Set 1, this CSV file will need to be read into a pandas Data Frame using the PYTHON programming language for any analysis to follow. Also the data may need to "cleaned" for any errors or missing values and so on.

Finally we would need to merge Data Set 1 and Data Set 2 into a new data set to be able to proceed further, so that the Neighborhoods in Data Set 1 are correctly mapped or related to their respective latitudes and longitudes in Data Set 2. This so we can use the Foursquare API location data to gather needed information about each neighborhood. This process as well as the processing of Data Sets 1 and 2 will be described in further detail, in a following section.

Section 3: The Methodology

In this section we aim to use Data Set 1 and Data Set 2 that were described previously in Section 2 above and combine the data sets with a leveraging of the FOURSQUARE API location data technology that was introduced in the course.

So we start with Data Set 1 and perform the following steps:

(a) Import data set and read the .CSV file as a pandas Data Frame.

(b) Rename columns of the Data Frame as shown in the data set exhibit.

(c) Strip columns in Data Frame for any extra white spaces in front or back. This to allow for easier manipulation of the data in future steps.

(d) Check Data Frame for missing data having the Python default value NaN. If there are missing data, the respective rows are deleted from the Data Frame.

(e) Drop rows of Data Frame that do not have any Boroughs assigned to them. If a row has a Borough name but not a neighborhood name, the neighborhood is given the borough name.

This results in giving us a nice, clean Data Frame containing Data Set 1 that we can work with. We display the first 20 rows of this Data Frame to illustrate our work and note that each postal code in Toronto has a borough and neighborhood associated with it, uniquely. Let's name this Data Frame TORONTO_DF1 as reference for this report.

Moving onto Data Set 2, the following similar steps are performed as well:

(a) Import data set and read the .CSV file as a pandas Data Frame.

(b) Rename the columns of the Data Frame as shown in the data set exhibit.

(c) Strip columns in Data Frame for any extra white spaces in front or back. This to allow for easier manipulation of the data in future steps.

(d) Check Data Frame for missing data having the Python default value NaN. If there are missing data, the respective rows are deleted from the Data Frame.

This results in giving us a nice, clean Data Frame containing Data Set 2 that we can work with. We display the first 20 rows of this Data Frame to illustrate our work and note that each postal code in Toronto has unique Latitude and Longitude coordinates mapped to it. Let's name this Data Frame TORONTO_DF2 for this report.

Next, to get a mapping of each neighborhood in Data Set 1 with its corresponding latitude and longitude coordinates in Data Set 2, we merge the TORONTO_DF1 with TORONTO_DF2 on the common column key 'PostalCode'. As a side note, this is highly reminiscent of joining two or more tables in a Sequential Query Language such as SQL or ORACLE. Let's call this Data Frame TORONTO_DF3. Awesome! This new Data Frame gives us each neighborhood, borough, postal codes and corresponding coordinates (latitude and longitude) in tabular fashion. One might take a look at Data Set 3 in the Appendix section for an illustration of this Data Frame.

So now with the Data Frame TORONTO_DF3 in hand, we're all set with applying the FOURSQUARE API location data technology to search/explore corresponding to the geographical coordinates (latitude/ longitude) of each neighborhood as regards eating places (or any other kinds of businesses for that matter).

To that end we set up our FOURSQUARE API credentials and then create an API request, get the respective URL and then retrieve the venues we need using TORONTO_DF3. The retrieved venues are then used to create a new Data Frame whose columns include neighborhoods, their latitudes and longitudes as well as the restaurant venue category name retrieved that lies in that neighborhood. In this report we name this Data Frame as TORONTO_DF4.

Now, it's time for some exploratory data analysis on TORONTO_DF4. Using pandas, we group the Data Frame by neighborhoods to find the number of different categories of venues in each neighborhood. Further, we also find the number of unique different categories of venues. The numbers appear substantial enough to support the application of an unlabeled and unsupervised Machine Learning technique such as the K-Means Clustering Algorithm.

As the supplied IPYTHON notebook with the code will illustrate, further exploratory data analysis is performed on TORONTO_DF4 prior to applying K-Means. This includes presenting the top 20 most common restaurant venue categories in tabular form for each neighborhood.

To get TORONTO_DF4 ready for K-Means it needs more work as follows:

(a) Perform "one-hot encoding" on the venue categories via the pandas "get_dummies" function. This adds numeric values for the categorical values, for all neighborhoods and their respective venue categories. The result is yet another Data Frame that we can call TORONTO_DF5. The neighborhood names are re-inserted into TORONTO_DF5.

(b) Group TORONTO_DF5 on the 'Neighborhood' key. Then take the mean or average on the one-hot encoded data from part (a) directly above. This given us a Data Frame called TORONTO_DF6 in this report.

Now we are ready to apply the K-Mean Clustering algorithm on TORONTO_DF6. One may note here that the K-Means Clustering Algorithm is an unsupervised and unlabeled Machine Learning Technique.

This algorithm fits our "Clustering" problem well as we are looking for cluster(s) of neighborhood(s) in which there may be substantial numbers of residents or office goers or tourists or anyone that lean towards preferring ethnic food, particularly south Asian.

Section 4: The Results

Application of the K-Means clustering algorithm on the Data Frame TORONTO_DF6, using the K-Means model on PYTHON with k = 5 clusters, provides us with an effective clustering of the neighborhoods into five clusters. The clustering is based on restaurant venue categories. The five clusters with its respective most common restaurant venue categories are presented in the IPYTHON notebook on GitHub.

Next, using Folium maps in PYTHON we create a graphical illustration of a Toronto-area map showing the different Toronto area neighborhoods with their respective clusters color-coded. This Cluster map is presented in the Appendix section at the end of this manuscript. The five different colors used for the color-coding are blue, yellow, red, purple and green. We need to present this Cluster map in the Appendix below since the map does not render upon pushing the notebook onto GitHub.

Section 5: Discussion

Application of the pandas libraries in PYTHON produced highly effective results and illustrations for this project. The high quality K-Means Algorithm, implemented via PYTHON, deserves an outstanding recommendation as well.

Some of the neighborhoods, owing to their restaurant venues, could not be placed in a cluster. Hence those neighborhoods needed to be dropped from consideration for a new restaurant. Here the pandas "dropna" function served the purpose well. Also the cluster labels needed to be converted into integer values for the Folium package to work.

From looking at the clusters in the IPYTHON notebook on GitHub we note that the relatively large-sized Clusters 3 and 4 have the highest concentration of restaurants specializing in south Asian food. Clusters 1 and 2 are fairly small clusters that may not be help us predict cuisine preference. Cluster 5 does not demonstrate an exceedingly high preference for customers leaning towards south Asian cuisine.

Neighborhoods that data science would recommend for a new restaurant opening would be as follows:

(a) First Preference: Leaside, Victoria Village, India Bazaar, The Beaches West, Canada Post Gateway Processing Center (Mississauga).

(b) Second Preference: First Canadian Place, Underground City, Downtown Toronto.

Section 6: Conclusion

In concluding this report, the data on the Toronto-area neighborhoods, their postal codes and their geographical coordinates (latitude/longitude) that were obtained from the web were extremely handy. This data was easily read in as pandas Data Frames. Using these Data Frames we could move forward with our analyses.

Then, the pandas library functions, implemented via PYTHON, were used effectively to clean and organize the data as well as convert to format where the FOURSQUARE API location data was leveraged to help us obtain restaurant venue category information from the geographical coordinates of each neighborhood.

The unsupervised, Machine Learning, K-Means algorithm was implemented using the built-in function in PYTHON to cluster restaurant venue data. The clustering gave us results that were highly informative as well as pertinent to the problem

At the end we were able to make suitable recommendation(s) to our client as to what neighborhoods would be beneficial for them to start a new restaurant in.

One should not forget to mention the sophisticated level of mathematics and statistics that is used in algorithms such as K-Means without which none of this could ever come to fruition. Also the effectiveness of Data Science as a business tool cannot be overstated here.

Appendix Section

Data Set 1: Toronto Postal Codes, Boroughs and Neighborhoods (First 25 Rows ONLY)

Postal code	Borough	Neighborhood
M1A	Not assigned	
M2A	Not assigned	
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park / Harbourfront
M6A	North York	Lawrence Manor / Lawrence Heights
M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government
M8A	Not assigned	
M9A	Etobicoke	Islington Avenue
M1B	Scarborough	Malvern / Rouge
M2B	Not assigned	
M3B	North York	Don Mills
M4B	East York	Parkview Hill / Woodbine Gardens
M5B	Downtown Toronto	Garden District, Ryerson
M6B	North York	Glencairn
M7B	Not assigned	
M8B	Not assigned	

Data Set 2: Toronto Area Postal Code Latitudes and Longitudes (FIRST 25 ROWS ONLY)

Postal Code	Latitude	Longitude
M1B	43.8066863	-79.1943534
M1C	43.7845351	-79.1604971
M1E	43.7635726	-79.1887115
M1G	43.7709921	-79.2169174
M1H	43.773136	-79.2394761
M1J	43.7447342	-79.2394761
M1K	43.7279292	-79.2620294
M1L	43.7111117	-79.2845772
M1M	43.716316	-79.2394761
M1N	43.692657	-79.2648481
M1P	43.7574096	-79.273304
M1R	43.7500715	-79.2958491
M1S	43.7942003	-79.2620294
M1T	43.7816375	-79.3043021
M1V	43.8152522	-79.2845772
M1W	43.7995252	-79.3183887
M1X	43.8361247	-79.2056361
M2H	43.8037622	-79.3634517
M2J	43.7785175	-79.3465557
M2K	43.7869473	-79.385975

Data Set 3: Data Set 1 Merged with Data Set 2 on Postal Code Key (FIRST 15 ROWS ONLY)

Postal Code	Borough	Neighborhood	Latitude	Longitude
M3A	North York	Parkwoods	43.753259	-79.329656
M4A	North York	Victoria Village	43.725882	-79.315572
M5A	Downtown Toronto	Regent Park, Harbourfront	43.65426	-79.360636
M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
M3B	North York	Don Mills	43.745906	-79.352188
M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
M6B	North York	Glencairn	43.709577	-79.445073
M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...	43.650943	-79.554724
M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
M3C	North York	Don Mills	43.7259	-79.340923

Cluster Map: Map of Toronto Area with Neighborhoods Clustered According to Restaurant Venue Categories Therein. The Clustering obtains via K-Means Machine Learning Algorithm