Segmenting and Clustering Neighborhoods in NYC, USA and Toronto, Canada to effectively predict the best possible location for the new store of Melaleuca Inc.

Arsalan Aziz

April 18, 2019

1. Introduction

1.1 Background

Melaleuca Inc. was founded In 1985, a multi-level marketing company that sells environmentally friendly nutritional supplements, cleaning supplies, and personal-care products, centered in Idaho Falls, Idaho, and Knoxville, Tennessee. Customers buy directly from Melaleuca's website or retail locations. The company has large international operations, and 25 percent of its revenue comes from Taiwan, Korea, Japan, Australia, New Zealand, and the United Kingdom.

Melaleuca wants to expand its business on a national and international level and because the company deals with health and wellness they wanted to open a retail store in one of the major metropolitan cities in North America. The cities that were chosen were New York or Toronto. This will be a huge breakthrough for them because the population density is high and there is great amount of awareness about health and Wellness in those areas. Executives want to open a store in such an area where they want to be close to customers so that they can engage with customers more often and understand their needs. Considering in mind if this initiative goes well melaleuca plans on expanding their business in one of those cities and perhaps they would open a research center as well. Thus it is very critical for executives to understand the geographical location of those cities and pick up an appropriate location for their store.

1.2 Problem

Data that might contribute to selecting the best neighborhood in one of those cities (New York or Toronto) might include demographic data. Zip code and location might give us a basis for understanding the potential of market also special settings concerning the location of other shops and businesses should be taken into account as well as retailers network analysis data will also help us find the solution to select the best location between the two. This project aims to predict which particular city and neighborhood inside the city will be the best location for the new store of melaleuca that will help the retailer grow the business in that location and city.

1.3 Interest

Obviously, Melaleuca Executives would be very interested in accurate prediction of best possible location and neighborhood in one of those cities for the competitive advantage and business value. Others who are interested in melaleuca such as investors and loyal customers to the brand will be very interested to know the expansion of the business and the planning and preparation behind it. The online digital data suggests that there is a lot of customers from New York and Toronto who are extremely brand loyal and buy their health and wellness products online, those customers will be extremely excited to know that physical store location opening in one of those areas.

2. Data acquisition and cleaning

2.1 Data sources

Before we can built a predictive model we basically wanted to perform an overall analysis of the two cities (NYC and Toronto) to get the general idea of the population, average income, weather, criminal offenses and the overall economy of the cities. After all opening a new store in those cities will require us to examine the surrounding and understand the business and people living in those areas. Once we are familiar with surrounding neighborhoods of NYC and Toronto the next step will be to start utilizing the Foursquare API to explore the neighborhoods and segment them and after analyzing and modeling the data we will be able to predict the best location to open new store.

The demographics related data like population and average income in different boroughs for New York City were picked up from Wikipedia source https://en.wikipedia.org/wiki/Demographics_of_New_York_City. In order to understand the overall economy of the city we analyzed the worth of some of the big companies in the state using the source https://en.wikipedia.org/wiki/Economy_of_New_York_City. The weather and the number of crimes recorded in New York data were basically obtained from kaggle.

Similarly, for Toronto we used the same metric of analyzing the city based on the five major prospects in the city to open a new business like population, average income in different boroughs, and economy of the city, weather and number of crimes recorded in the city. The demographic data like population and average income were picked up from Wikipedia https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods and the weather related data was picked up from https://www.wunderground.com/history/monthly/ca/toronto/CYTZ/date/2019-1. Similarly, the number of crimes committed in Toronto was picked up from http://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-boundary-file- and to analyze the economy, the data related to businesses were picked up from https://en.wikipedia.org/wiki/List_of_largest_public_companies_in_Canada_by_profit.
The data that we received from those sources required extensive amount of preprocessing before we could use it for analysis.

In general the NYC Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood, this dataset exists for free on the web and to find this dataset we used the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

Similarly, Toronto Neighborhood has a total of 11 boroughs and 208 neighborhoods, In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 11 boroughs and the neighborhoods that exist in each borough. This dataset was easily available on in the link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2.2 Data cleaning

The data we obtained from the sources mentioned above (section 2.1) require a lot of preprocessing before we could do the analysis and modeling, there were some records and columns that we had drop because of their insignificance nature in the analysis of cities. Data downloaded or scraped from multiple sources were combined into one table. For example the weather data for Toronto was pulled for each month individually and then combined into single table in order to do the analysis conveniently.

There were several problems with datasets. First, there were some business that were common in both the cities and we had to change the dimension by which we were analyzing the economics of the city similarly we had to clean and remove business operating in the rest of the country to focus specially on the cities that we are analyzing in this project.

Secondly, in some cases that we were analyzing we noticed that the data was not available at the borough level so we had to pick up the data from the granular level of neighborhoods in the cities and group them together into a borough in the city.

Third, there were some cases where the data was available in different features all together like for the organizations in the cities the data for Toronto was available in terms of yearly revenue and growth as compare to New York where the data is available as a total revenue of the companies. This particular case was resolved by extracting the particular feature like the company worth that would give us the holistic view of how the overall city is performing from economical perspective.

After fixing all these problems I checked the outliers in the data. I found that there were some extreme outliers mostly caused by the small sample data size. For example, some of the companies recorded revenue was very small number as compared to the large organizations and we had to drop those companies because they had an insignificant effect on the results. Obviously, large companies will have more opportunities for the neighborhood so they will have a major impact on the economics in the city.

2.3 Feature selection

After data cleaning there were 487,336 samples and 32 features in the data. Upon examining the meaning of each feature it was clear that there was some redundancy in the features. For examples, the feature of the growth of company and the revenue of the company basically gives us the same picture about the performance of the companies and therefore those two features were either combined or one of them was dropped.

Similarly some features were highly correlated like average temperature, mean temperature, medium and high temperatures in the cities are highly correlated so only one was kept and others were dropped. Features of criminal data like misdemeanor, robbery, theft and violence etc. were combined because we were only interested to know the overall crime recorded in the city that will give us a general idea of the neighborhood.

The table below shows the simple feature selection process during the data cleaning.

City	Kept Features	Dropped Features	Reason for dropping features
New York City	Borough Company Revenues Population Mean Temperature MedianHouseholdIncome Map Neighbourhood	World rank Employees Industry group Gross Domestic Product Land area Density County Estimate Percent-age in Poverty precipitation snow_fall snow_depth	We are interested in the holistic view of the city so we picked only those features that will help us achieve that and they should commonly exist in both cities like Toronto and NYC. Like For Example: FM column was representing the symbol of boroughs in Toronto and we replaced it with actual borough names. similarly, rates and percentages were replaced with count and numbers to give a consistent numerical value for analysis.
Toronto	Population Borough Company Revenues Population Average Income Total_2018_Statistics Map Neighborhood	FM Census Tracts Land area (km2) Density (people/km2) % Change in Population since 2001 Transit Commuting %% Renters Second most common language (after English) by name Second most common language (after English) by percentage Rank Company 5-year growth (%) 2017 revenue ($mil) City Province Employees, 2017	

3. Methodology

3.1 Exploratory Data Analysis

Venues category in the neighborhoods of both NYC and Toronto are not in the feature of dataset that we have used in our analysis and had to be calculated and discovered. There we had to utilize Foursquare API to get the neighborhood in NYC and Toronto and then find the latitude and longitude values of each neighborhood. Then we got the top 100 venues around each neighborhood within the 500 meters for both NYC and Toronto. By calculating the number of venues and their categories for NYC we found out that there were 3325 different venues around NYC which can be categorized into 333 unique categories. Similarly for Toronto we found out there were 1710 different venues around Toronto which can be categorized into 240 unique categories.

This will help us in finding out how many times the particular venue occurs by taking the mean of the frequency of occurrence of each category and eventually it will be helpful in clustering the neighborhoods of each city depending on the occurrence of each venue categories.

3.2 Relationship between NYC and Toronto Populations

We plot the population of NYC Boroughs and observed that the population of Brooklyn was much higher as compared to other boroughs. Similarly we also plotted the population of Toronto boroughs and observed that population of North York, Old City of Toronto and Scarborough were almost the same but when we compared the top NYC boroughs with the top Toronto boroughs, there was a major difference between them as we observed that the top NYC boroughs has the population almost double the size of top Toronto boroughs populations. In general the population of NYC is higher as compare to Toronto. The population of NYC is around 8.623 million and the population of Toronto is around 2.732 million, which is also evident from our analysis as well but we wanted to do the population comparison down to the boroughs level to get the better understanding of picking the appropriate place for attracting more traffic to the new store location of Melaleuca.

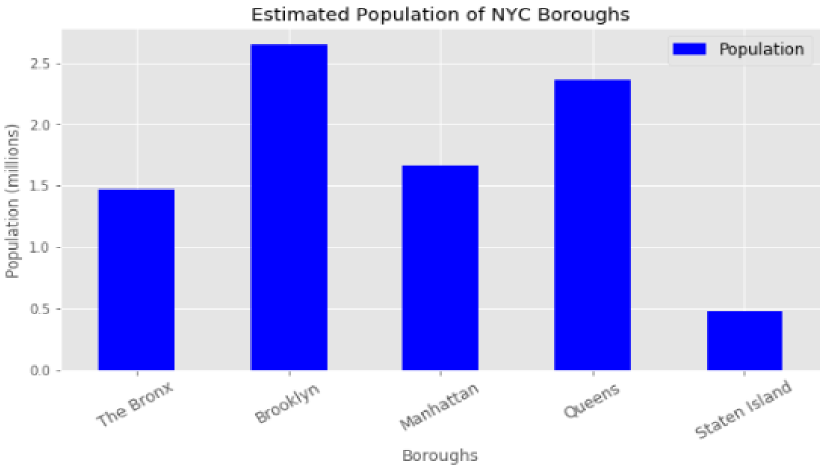

Figure 1. Estimated Population of NYC Boroughs

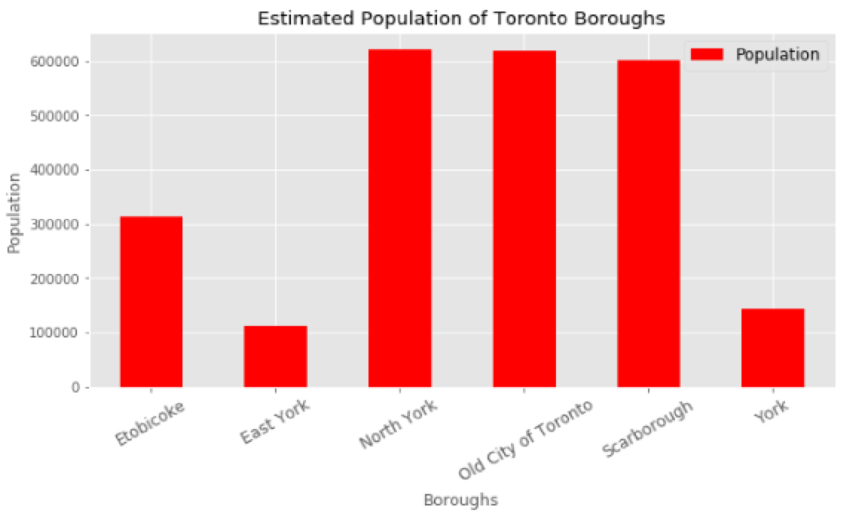

Figure 2. Estimated Population of Toronto Boroughs

3.3 Relationship between NYC and Toronto Average Income

We analyzed the Median household Income of NYC Boroughs and observed that the Median household Income of Staten Island and Manhattan boroughs were higher as compared to other boroughs. Similarly we plotted the Average income of Toronto boroughs and observed that the average income of Old city of Toronto, North York and East York were higher as compare to other boroughs. When we compared the top NYC boroughs with the top Toronto boroughs there is a major difference, the average house hold income for the top NYC boroughs has the income more than 60 thousand dollars as compared to the top boroughs of Toronto where the average household income is around 50 thousand dollars. This shows there is more opportunity for businesses around NYC area to acquire customers because the population over there is wealthier and they are more open to try new brands and products thus creating good environment for businesses like Melaleuca to invest and grabs customer attention.

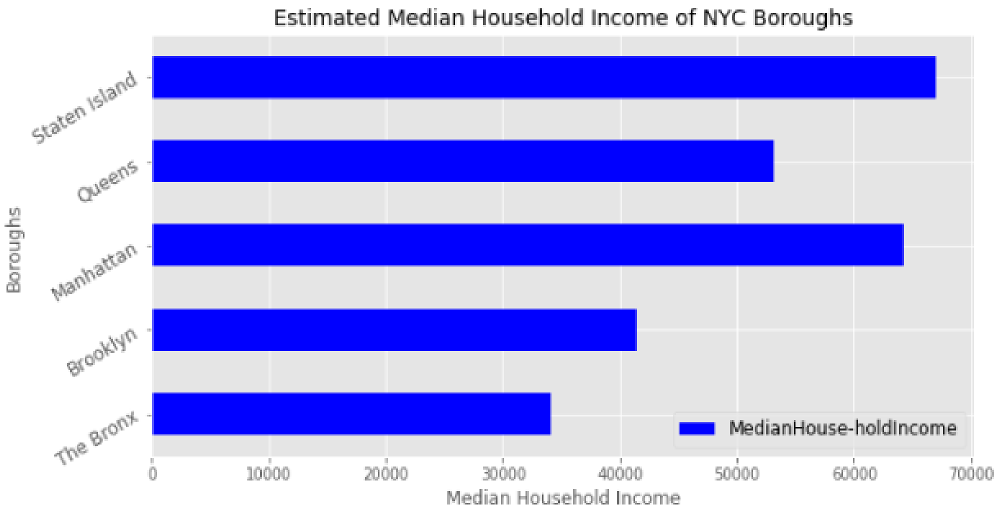

Figure 3. Estimated Median Household Income of NYC Boroughs

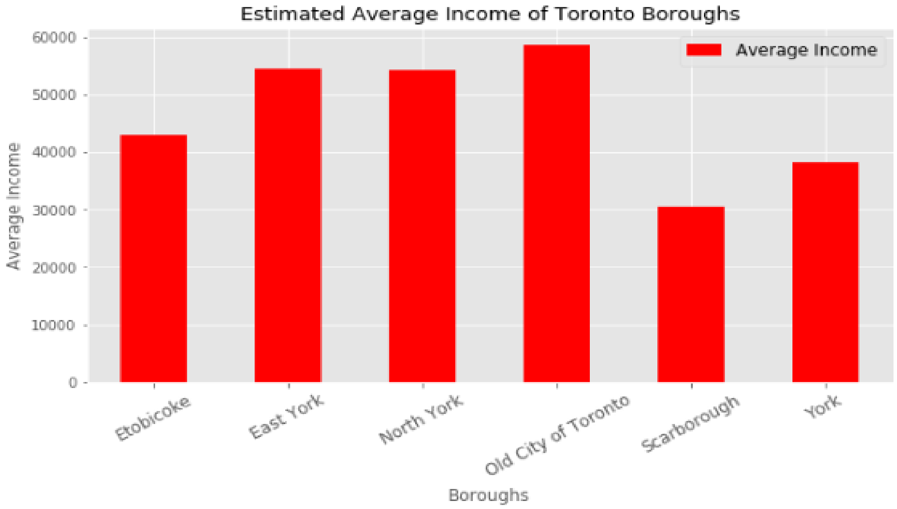

Figure 4. Estimated Average Income of Toronto Boroughs

3.4 Relationship between NYC and Toronto Weather conditions

We also analyzed the Weather of NYC and observed that the average temperature of NYC remains high 80(F) in summers and low 30(F) in winters. There is not a major fluctuation between minimum and maximum temperatures but we can conclude from the plot that the temperatures are generally cold in winter and with possibility of snow days as well. On other hand if we can check the Toronto weather we can observe that the temperature reaches around 80(F) in summer but in winter it can go extreme cold and stays below 20(F) sometimes below 0(F). Toronto also has very high possibility of snow days as well as compare to NYC.

When we compare NYC with Toronto we can see that Toronto has extreme weather conditions as compare to NYC so there is a high chance that the weather condition will have impact on traffic to stores and possibly will affect the business of Melaleuca. NYC on other hand has high probability that more traffic will be attracted to stores with good favorable weather conditions.

3.5 Relationship between NYC and Toronto Crime

We analyzed crime statistics of NYC Boroughs and observed that the number of crimes committed in Brooklyn were higher as compared to other boroughs. For Toronto we used the Toronto's neighborhood data and observed that the number of crimes committed around church-Yonge corridor neighborhood was much higher as compare to other neighborhood. Overall we observed that the number of crimes recorded in NYC is much higher as compare to Toronto in NYC for one full year the number of crimes recorded were 478,579 where in Toronto the number of crimes recorded were 34,930.

Safety of store employers for melaleuca or any other business is of primary importance to the executives and this analysis have grave importance when considering to pick the right location of opening a new store.

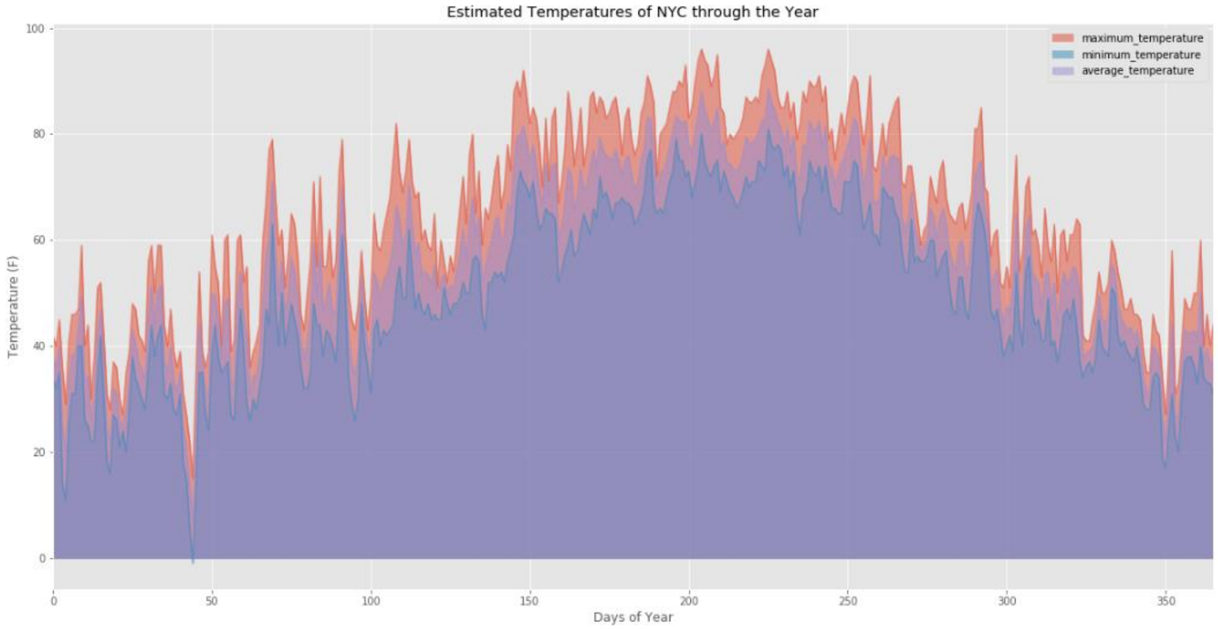

Figure 5. Estimated Temperature of NYC through the entire Year

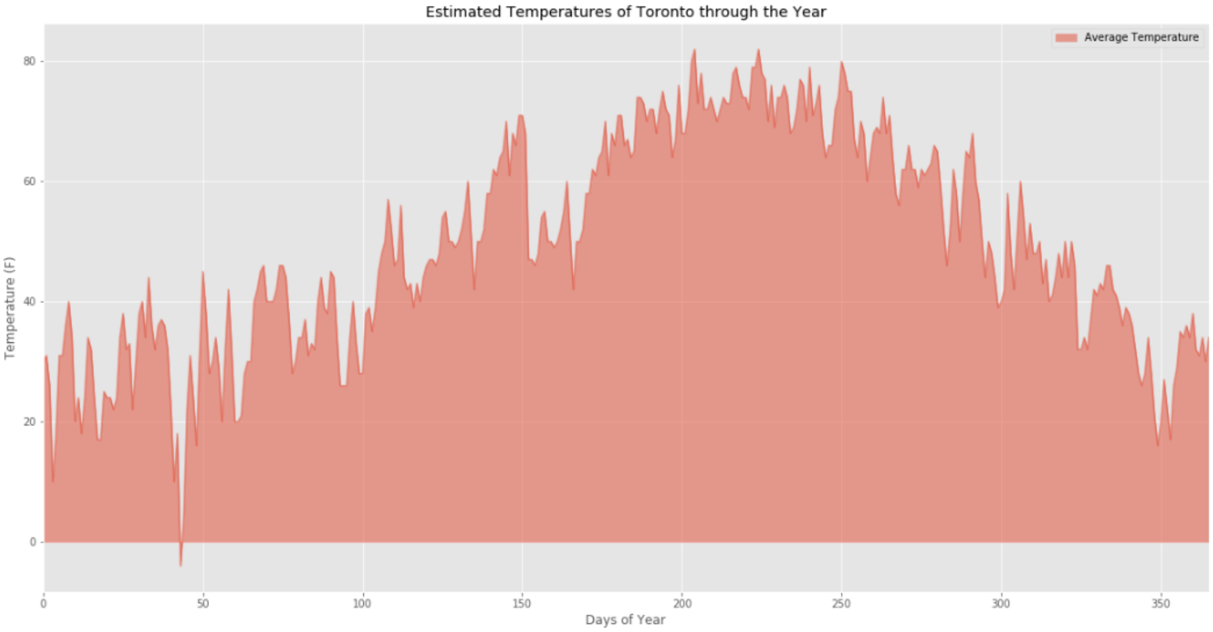

Figure 6. Estimated Temperatures of Toronto through the entire Year

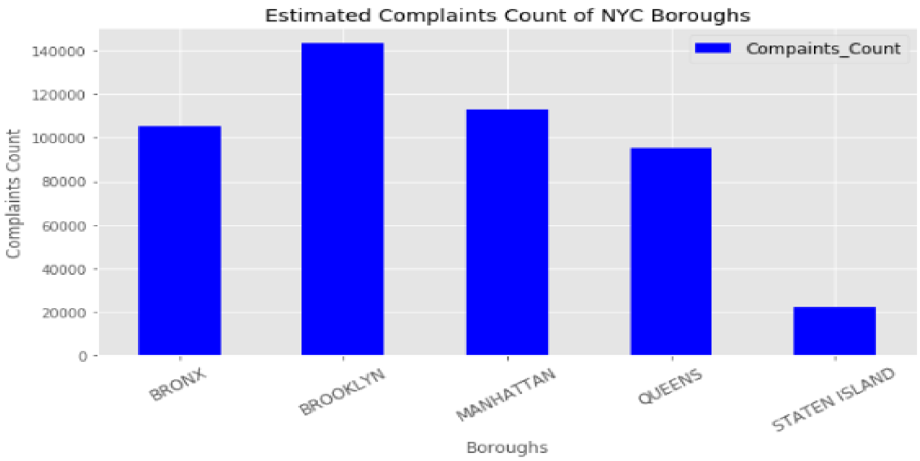

Figure 7. Estimated complaint counts of NYC Boroughs

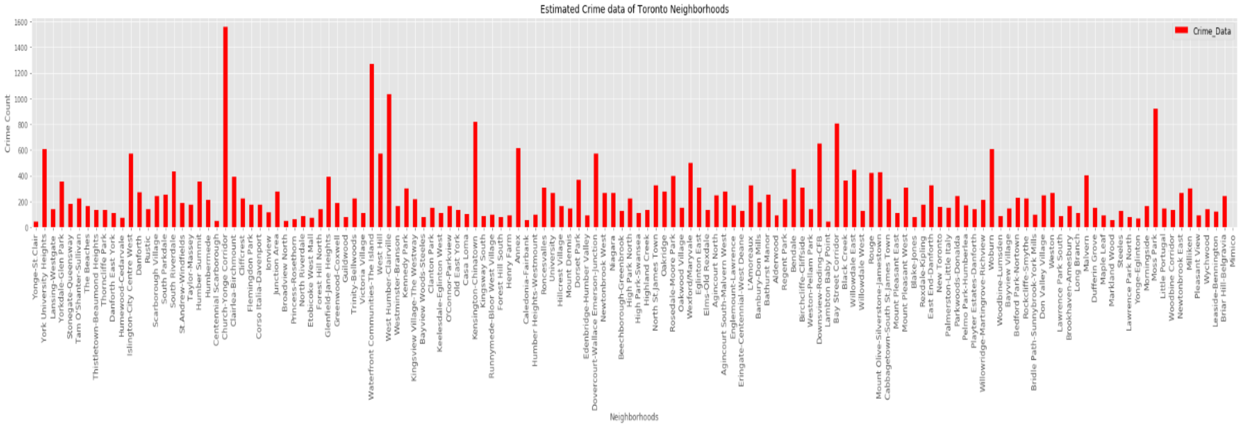

Figure 8. Estimated crime data of Toronto Neighborhoods

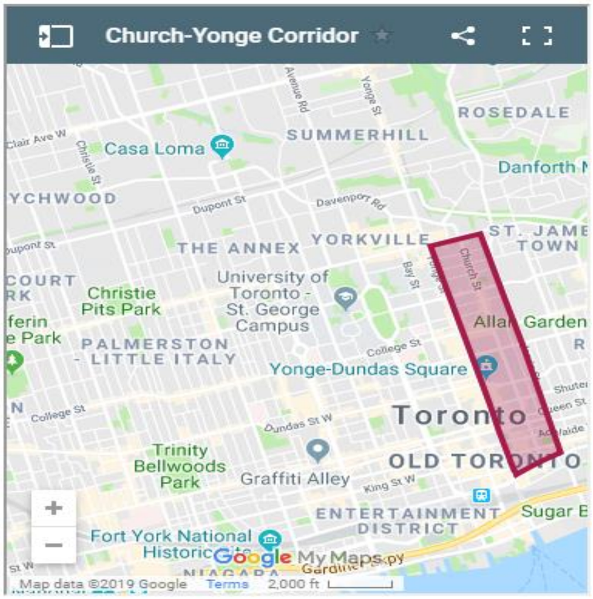

Figure 9. Location of Church-Yonge Corridor Ref: https://www.toronto.com/community-static/4548999-church-yonge-corridor/

3.6 Relationship between NYC and Toronto Economies

We Analyzed the Economic statistics of NYC and observed that NYC has world top companies that has the revenues in hundreds of billions. For the sake of assessment we analyzed the top 5 companies based in NYC that will give us the assessment of how the economy of the city is and it gave us a very good idea that the city is booming with a lot of opportunities for business and career for growth.

On the other hand we also did the analysis of fortune 500 business in Canada and observed that the out of fortune top 500 businesses of Canada, 112 are from Toronto and they show a very high growth where some of the businesses like Eden park and Maropost are showing almost 10,000% growth. This means that the economy is booming and in the future there will a lot more opportunities.

Economic situation of city is very important for opening up a new business because the better the economic situation of the city is the more people will spent their money because of the perception that in future they can always earn more in the city because of the business opportunities it provides and vice versa.

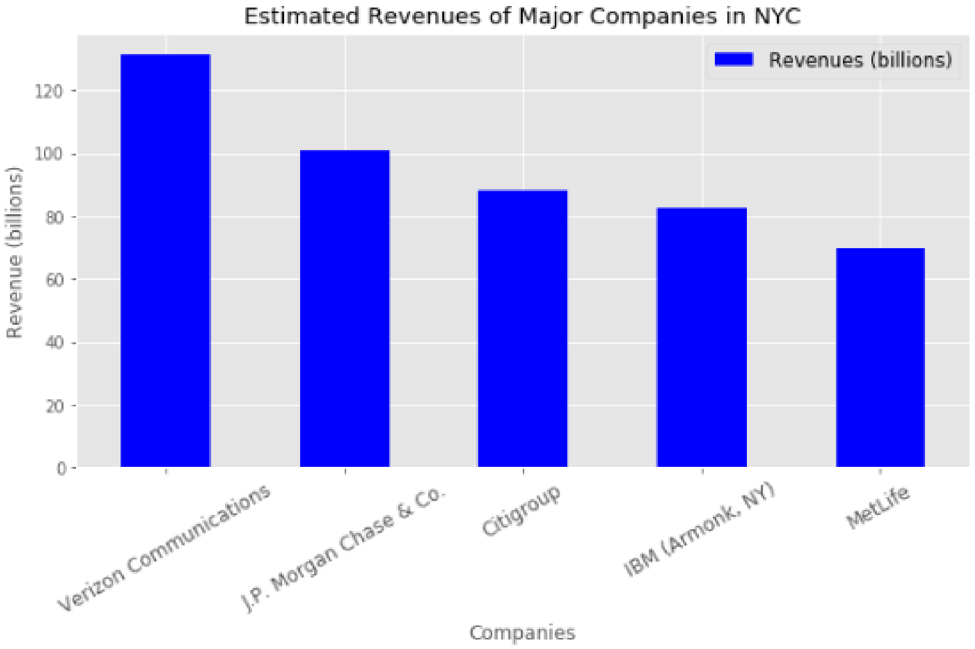

Figure 10. Estimated Revenue of Major Companies located in NYC

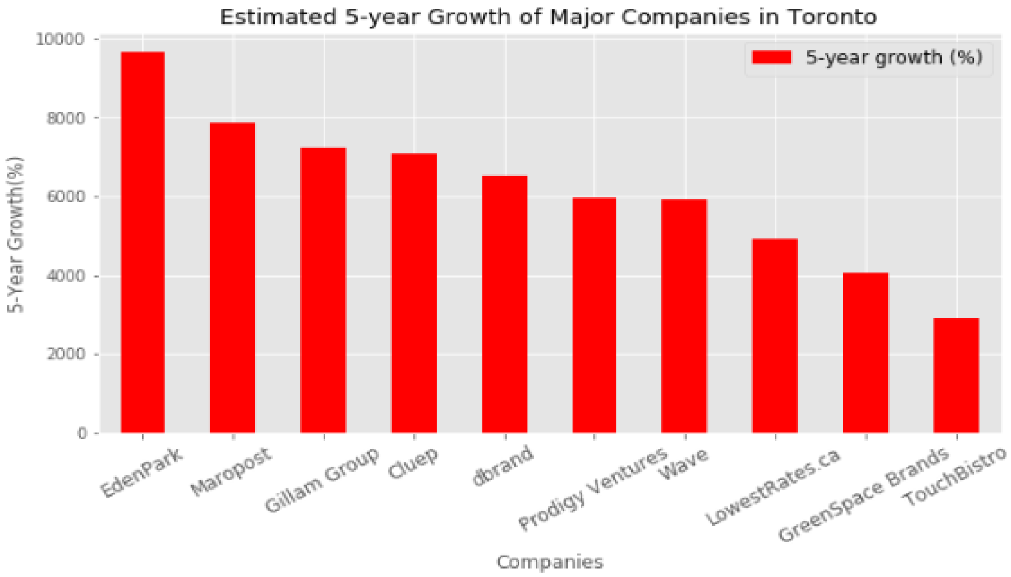

Figure 11. Estimated 5-Year Growth of Major Companies in Toronto

4. Predictive Modeling

This particular case of picking up the best location for opening the new store of Melaleuca is definitely a perfect example of clustering. We will select the number of clusters around a neighborhoods in NYC and Toronto and find out which cluster will be best for the new store opening. Then combining the clustered predictive result with the analytical exploration we can confidently suggest the executives that which is the best city and neighborhood for the new store to open. Other models like regression, classification or support vector machine won't work in this case because this particular example is more related to grouping of similar venue categories around the neighborhood that can give a sense of the surrounding in a particular place and these grouping and categorization can be best done by using K- nearest neighbor modeling approach.

4.1 Segmenting and Clustering

In order to segment and cluster the neighborhoods we segmented each city (NYC and Toronto) into 5 different clusters using K-Mean clustering than we used folium to visualize the top 10 venues for each neighborhood in the cluster. Depending on the venue we analyzed the cluster to get the better idea what the particular cluster will be good for residents. This will give us the idea of the particular cluster that will work best for the location of new store of Melaleuca. Depending on the analysis that we did and the clustering of venues in each city we can make an effective decision of choosing which neighborhood to open the new store and in fact which particular location as well within that city.

4.2 Examining the clusters of NYC

By Examining the five different clusters in NYC we observed that each cluster can be named based on the businesses that operates in that area. NYC has a tremendous amount of varieties of different venues that it has to offer. For Example: Cluster 1 basically looks like a Shopping Center Cluster with mix of shops like boutiques, clothing shops, Bars, restaurants and Cafe. Cluster 2 is basically more kind of recreational center where there are a lot of venues like Spa's, restaurants, hotels, nightclubs, theaters and bars etc. Most of the venues are related to enjoyment and if someone wants to spend quality joyful time. This cluster is best collection of all those activities.

Cluster 3 is named as public Visiting Park because venues are more related to outdoor activities like Park, Basketball court, baseball Field and boat or ferry etc. It is the best spot for someone who is interested in outdoor activities. Cluster 4 will be great for tourist to stay and travel around in this area that's why we named it as tourism attraction center. The bridge, metro station and scenic lookout for tourists will be great spots to visit plus this cluster has to offer a lot of restaurants, hotels, cafe and theaters etc. which makes this cluster a best place for tourists to stay and visit around conveniently. Cluster 5 more looks like a commercial area where there is mixture of different shops, restaurants, lounges, stores and schools etc. These variety of combination makes this area a good choice to rent a place somewhere close to it so that it's easy for family to send their kids to school and shop and dine in near buy stores and restaurants.

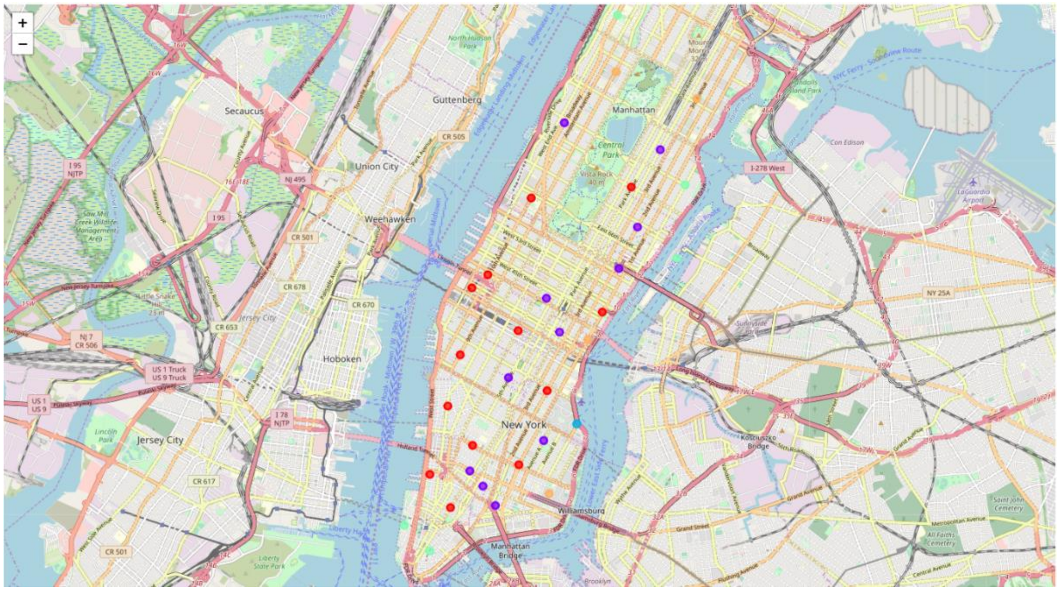

Figure 12. 5-Clusters of NYC Manhattan Neighborhoods identified with 5-different colors

4.3 Examining the clusters of Toronto

By Examining the five different clusters in Toronto we observed that each cluster can be named based on the businesses that operates in that area. For Example: Cluster 1 basically looks like a Toronto Downtown Cluster, this cluster is mainly showing all the business located in a downtown of Toronto. Everything that a downtown can have, is listed in this cluster like restaurants, bars, hotels, Markets etc. The downtown borough has occurred multiple times in this cluster and it’s the largest cluster among all the 5 clusters. Cluster 2 basically looks like a Business Center Cluster with mix of businesses like restaurants, Bars, pet track and medical center etc. with no proper pattern

Cluster 3 is picnic spot cluster because there is park with a trail. There are also multiple restaurants around this cluster so it will be very convenient for the families to have an outdoor activity in this area. Pet lovers will also enjoy the spot because there is a Dog Run venue as well. In short it is very good outdoor venue for families and friends. Similarly, Cluster 4 is given a name of festival cluster because this area is very suitable if someone who wants to arrange an event or a festival. There is an event space with numerous restaurants around it so catering will be very easy with electronic store to support any electronic needs for the businesses. Cluster 5 looks like community center cluster because there is bus line so the transport is very convenient. There is a park, restaurants and stores which makes this place useful of everyone in the community to go in this area and buy food or other useful items.

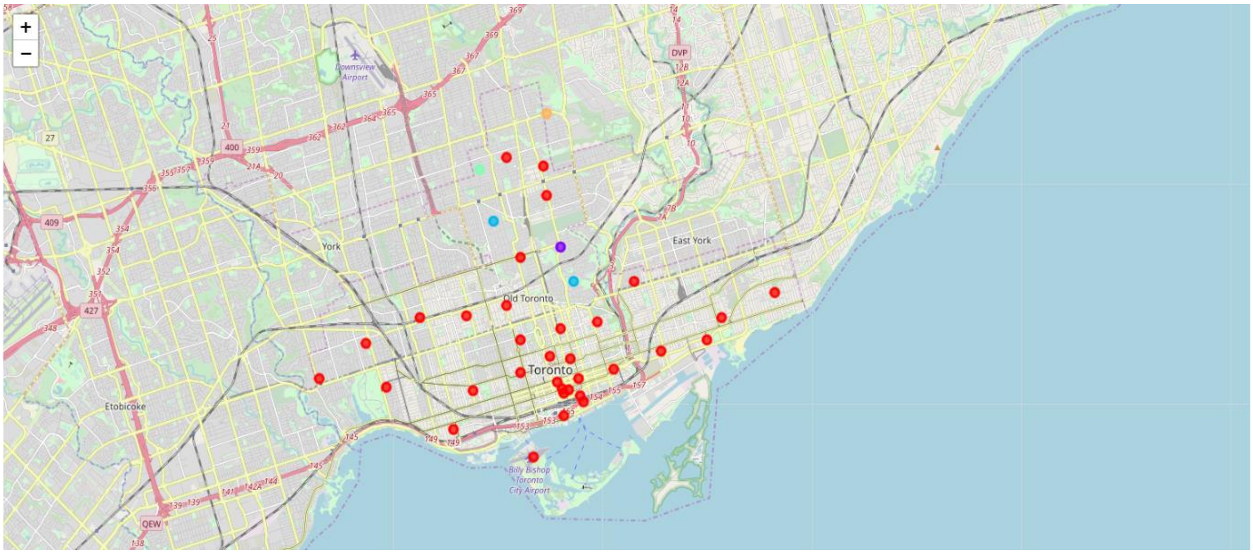

Figure 13. 5-Clusters of Toronto Neighborhoods identified with 5-different colors

5. Results and Challenges

Results of cluster shows that both cities are very vibrant and flourishing with several different types of businesses are operating and running in each city. Different neighborhood has different significance and each cluster explicitly displays the significance of each neighborhood. NYC can be a suitable pick for the new store opening of melaleuca because based on the analysis and the predictive model that we have built we can suggest that the population, average house hold Income and Economic situation are much more higher and better as compare to Toronto City. Also the weather of NYC is very favorable as compare to Toronto for opening a store and hence with all these advantages of NYC over Toronto there is a high possibility that more traffic will be attracted to melaleuca store in NYC as compare to Toronto.

Opening a store in NYC will be a better choice as compare to Toronto but it also comes up with some challenges. Because of high population density in NYC the moving of inventory into the city will be a challenge so special arrangements needs to be made in order to cope up with it and make sure that items doesn't get out of stock which has high demand. Also the crime rate is high in NYC as compare to Toronto. We saw that the crime rate in Brooklyn was very high as compare to other boroughs. We can choose Manhattan borough for the location preference as it will reduce the probability of any mishaps around the store location area. But Manhattan is one of the expenses places in the United States so there will be high upfront expenses on establishing and operating the store. So all these challenges has to be kept in mind before we can select Manhattan as our choice of location but overall based on the analysis and the predictive model we can confidently decide that Manhattan will be the best option for melaleuca to open the new retail store.

6. Discussion and Future Recommendation

In this study we analyzed and build a predictive model using clustering and segmentation method of machine learning that can accurately display the best possible location of opening a store for melaleuca Inc., whose executive wants to expand their business and open a new brick and motor store either in NYC or Toronto area. This twostep approach of analyzing and building a predictive model really helped us in thoroughly investigating the capabilities of two cities based on their population, average household income, economic statistics, weather forecasting and crime rate to find the best location for opening a new store. We can also do deeper analysis by considering the real estate data, the traffic and transit information to get further deeper knowledge of the two cities.

Clusters were built mainly focusing on the top 100 venues surrounding the neighborhood but these clusters can further be enhanced by taking more venues under consideration. Increasing the number of venues also involve some challenges like the amount of data will increase by huge amount and it will take more time to process, also the venue categories will increase so it will be hard to extract a particular category but by increasing the number of clusters we will be able to get more granular information about the neighborhoods in each city.

We can further refine and enhance the model by understanding the customer shopping habits in those cities. This can be achieved by getting different retailers data available on Kaggle or other open data source websites. Also we can check the customer preferences of whether customers in those cities are more interested in online shopping or they are interested in shopping in stores because this will be very helpful in understanding the customers shopping preferences and we can enhance the store shopping experience by introducing more features like easy pickup in store, buy online pick up in store and other shopping options like these to customers.

7. Conclusion

In this study, we analyzed the two major cities in the world and then build a predictive model that can cluster the neighborhoods in each cities and based on those clusters we can identify the best possible city and the best possible location within the city for melaleuca Inc. to open a new retail store location. At first we analyzed each city and then we clustered each neighborhood based on their venues that they have to offer to accurately understand the surrounding of the neighborhood and suggest the best possible location.

This clustering and segmentation model not only show us the city that will work best for melaleuca Inc. to open there new store but also the particular area and neighborhood within the city. These type of granularity is only possible using the machine learning model and the foursquare API to get the geographical data of any particular location on the planet. We can go further granular if we want and create more cluster that are smaller in size to get more detailed venues with in the cluster and understand their overall surrounding and behavior. More clusters will further increase the overall efficiency of the model.