Expansion of Melaleuca Retail Brand in NYC or Toronto

Data acquisition and cleaning

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 biggest companies in the city 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 6 major boroughs and 208 neighborhoods, In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains

the 6 major 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

Data cleaning

The data we obtained from the sources mentioned above required 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 MedianHouse-holdIncome 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	