2020

Applied Data Science Capstone Report: The Battle of Neighborhoods

Bernardo Vimpi

Contents

INTRODUCTION

This project aims at conducting a study, using data, to evaluate the right neighborhood to open a new business. Selecting the correct neighborhood to open a new business is not an easy task. Several factors need to be taken into consideration including, but not limited to location and population density. In this capstone project I analyze and compare data between two major cities, New York City (USA) and Toronto (Canada). To make the final decision as to which of the cities is best fitted for a business to open a new site, requires an in-depth and throughout evaluation of data in order to arrive to the desired conclusion. Thus, I use data science methodology and tools to properly analyze this issue.

BUSINESS PROBLEM

In this section I present the main business problem which is the core of this project. STYLE Pro is a cloth retail business that was founded in 2000 in the city of Chicago, USA. Over the past 20 years the company has grown in both revenue and customer acquisition due to its unique products that are tailored to working professionals. Due to its exponential growth and increased revenue, new ventures for the company have attracted more investors over the past 5 years. As such, the company sees the impending need to expand its business in a metropolitan area which has the potential to continuous business growth hence, high Return on Investment (ROI). The leadership of the company has two major cities in consideration: New York City and Toronto. However, the business executives must decide which of these cities is best to open a new branch for STYLE Pro.

AUDIENCES

Specific Audience:

- The specific audience for this project are the executives, employees, current stakeholders and investors for STYLE Pro. They are the primary reason for the desired decision to open a new branch location. The executives and investors of STYLE Pro need to be fully informed of the ways in which I arrived to the conclusions in this project and based on that conclusion they need to make a decision on the feasibility of opening a new branch based on which of the two cities offers the best business potential, growth and return on investment for STYLE Pro.

General Audience:

- The general audiences of this project are future investors. In order to attract potential investors, STYLE Pro needs to have a clear vision of the company growth, revenues and profit. One of the best approaches to invite new investors in near future is to present the current findings for this project to investors and explain the extent to which expanding the business to either New York City or Toronto will accelerate the company's expansion and growth in customer acquisition, retention and increase profits and revenue, which ultimately will translate into return on investment for stakeholders in the short and long run.

DATA SOURCE AND DESCRIPTION

Describe the data that you will be using to solve the problem or execute your idea. Remember that you will need to use the Foursquare location data to solve the problem or execute your idea. You can absolutely use other datasets in combination with the Foursquare location data. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using, even if it is only Foursquare location data.

- **FOURSQUARE API Geolocation Data:** The data from all the neighborhoods from New York and Toronto will be generated in different tables. However, we need to have a geo special mapping of the actual data in order to locate the neighborhoods and trace them on the map. Thus, I use FOURSQAURE API Geolocation Data to achieve this purpose and map each of the cities and their respective data in the Map.

- **New York City Neighborhood Data Profile:** The New York City Neighborhood Data Profile from New York University Furman Center contains robust demographic data for the major neighborhood cities in the New York City area. The data is also mapped and provides insights into key demographic indicators such as population density and services around each neighborhood. This is relevant to my study since the data will be used to address the main objective of this study.

- **New York City Demographics Data via Wikipedia:** New York City is the largest city in the United States. The data from Wikipedia is segmented into different categories such as population, ethnicity, income, businesses and more. One important element of this data is the fact that its segmented based on each neighborhood. This is relevant to my research because I'll be analyzing which of these neighborhoods, if any, is suitable to open a new branch for STYLE Pro business.

- **Postal Code from Canada:** In order to evaluate the feasibility of each neighborhood business prospects we need to generate each address for major neighborhoods using zip codes or postal code. This is important differentiator because sometimes different geographical locations may have the same or similar names and other attributes, such as population composition and other demographic indicators. However, to differentiate each neighborhood, we use their respective zip or postal code.

- **Boroughs of New York City:** New York City is composed of five major administrative divisions called Boroughs. They are The Bronx, Brooklyn, Manhattan, Queens and Staten Island. This is important data because each administrative division will give us substantial insights at to which of the areas, if any, within the New York City, provides us the best possible opportunity to open a new branch for STYLE Pro and compare that data from the administrative divisions in Toronto, Canada to arrive to a conclusion**.**

- **Demographics of Toronto:** The data from Wikipedia shows the demographics of Toronto for more than a decade. This data supplies robust demographics indicators such as population, ethnicity, cities and more. These data is essential in determining which of the neighborhoods, if

any, provides the best possible option and opportunities for STYLE Pro to open a new branch. We then take the results and compare them with New York to Determine which of the two cities is has the best potential for business opportunity for STYLE Pro.

DATA ANALYSIS & METHEDOLOGY

DATA ANALYSIS

In this section, I imported the python libraries used for this project and displayed the data sets presents in each of the data sources. Examples of this are data that displays the Demographics of the City of Toronto.

Table 1: Demographics of the City of Toronto

	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters
0	Toronto CMA Average	NaN	All	5113149	5903.63	866	9.0	40704	10.6	11.4
1	Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0...	44577	12.45	3580	4.6	25750	11.1	5.9
2	Alderwood	E	0211.00, 0212.00	11656	4.94	2360	-4.0	35239	8.8	8.5
3	Alexandra Park	OCoT	0039.00	4355	0.32	13609	0.0	19687	13.8	28.0
4	Allenby	OCoT	0140.00	2513	0.58	4333	-1.0	245592	5.2	3.4

Another example of data analysis is the Number of borough and neighborhood in New York: There are 5 borough and 306 neighborhoods. Here, I display one of the boroughs which is The Bronx.

Table 2: The Borough of Bronx

	Borough	Neighbourhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

DATA CLEANING

I used the Json libraries to clean the data set and display the needed data for the project. For example, the coordinates of Manhattan in New York (Table 3)

Table 3: Coordinates of Manhattan

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Another example of data cleaning using Json libraries are the coordinates of downtown Toronto which its respective venues and venue categories.

Table 4: Coordinates of downtown Toronto

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rosedale	43.679563	-79.377529	Rosedale Park	43.682328	-79.378934	Playground
1	Rosedale	43.679563	-79.377529	Whitney Park	43.682036	-79.373788	Park
2	Rosedale	43.679563	-79.377529	Alex Murray Parkette	43.678300	-79.382773	Park
3	Rosedale	43.679563	-79.377529	Milkman's Lane	43.676352	-79.373842	Trail
4	Cabbagetown,St. James Town	43.667967	-79.367675	Cranberries	43.667843	-79.369407	Diner

Another technique and tool used was the FOURSQUARE API GEOLOCATION, in combination with Json libraries, to clean and display data both from Toronto and New York Cities. Example is the Foursquare API Geolocation data of downtown Toronto.

Table 5: Foursquare API Geolocation for downtown Toronto

	name	categories	lat	lng
0	Rosedale Park	Playground	43.682328	-79.378934
1	Whitney Park	Park	43.682036	-79.373788
2	Alex Murray Parkette	Park	43.678300	-79.382773
3	Milkman's Lane	Trail	43.676352	-79.373842

METHODOLOGY

Exploratory Data Analysis (EDA)

- In order to understand the distribution of the variables for the data sets for Manhattan and Downtown Toronto, we need to visualize the data by generating tables for the neighborhoods in the boroughs, the downtown of Toronto and Manhattan. To do so we use, folium library which will give a visualization of the spacial distribution of these boroughs and neighborhoods.
- To further analyze the correlation and the trending of the data and gain additional insights on the neighborhoods and boroughs compositions, we employ further exploratory data analysis (EDA)
- Upon completing the step described above, we analyze the data via exploratory data analysis (EDA) utilizing Folium library, Seaborn and Matplotlib libraries to visualize the data distribution, gain further insights of the data.
- Another method employed in this study is k-means clustering in order to group the neighborhoods from New York and Toronto based on their similar characteristics, such as population density, neighborhoods compositions and more.

Table 6: EDA for Downtown Toronto

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Adelaide,King,Richmond	100	100	100	100	100	100
Berczy Park	56	56	56	56	56	56
CN Tower,Bathurst Quay,Island airport,Harbourfront West,King and Spadina,Railway Lands,South Niagara	17	17	17	17	17	17
Cabbagetown,St. James Town	47	47	47	47	47	47
Central Bay Street	79	79	79	79	79	79
Chinatown,Grange Park,Kensington Market	86	86	86	86	86	86
Christie	18	18	18	18	18	18
Church and Wellesley	85	85	85	85	85	85
Commerce Court,Victoria Hotel	100	100	100	100	100	100
Design Exchange,Toronto Dominion Centre	100	100	100	100	100	100

Table 7: EDA for Manhattan New York

Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Battery Park City	98	98	98	98	98	98
Carnegie Hill	100	100	100	100	100	100
Central Harlem	46	46	46	46	46	46
Chelsea	100	100	100	100	100	100
Chinatown	100	100	100	100	100	100

Machine Learning: Inferential Statistical Test and K-Means Clustering

In this section, I incorporate K-Mean Clustering to group the different neighborhoods. This is a technique of using Machine learning to encode the data that is part of the analysis. The main goal here is to group all neighborhoods that present similar characteristics, especially demographics. Another important element to emphasize here is the fact that while we get similar groups of neighborhoods, we can also use the information for the differences between or among the neighborhoods to make other inferences for the

data. The end goal here is till to characterize the groups of neighborhoods from downtown and Manhattan, based on our observations.

Table 8: Clustering for Manhattan

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Washington Heights	Café	Grocery Store	Deli / Bodega	Bakery	Spanish Restaurant	Mobile Phone Shop	Donut Shop	Bank	Tapas Restaurant	New American Restaurant
3	Inwood	Mexican Restaurant	Restaurant	Lounge	Pizza Place	Café	Deli / Bodega	Bakery	Spanish Restaurant	Frozen Yogurt Shop	Chinese Restaurant
4	Hamilton Heights	Pizza Place	Deli / Bodega	Coffee Shop	Café	Mexican Restaurant	Yoga Studio	Bakery	Liquor Store	Park	Cocktail Bar
5	Manhattanville	Coffee Shop	Deli / Bodega	Italian Restaurant	Mexican Restaurant	Chinese Restaurant	Park	Seafood Restaurant	Cosmetics Shop	Supermarket	Boutique
7	East Harlem	Mexican Restaurant	Bakery	Latin American Restaurant	Thai Restaurant	Deli / Bodega	Spanish Restaurant	Spa	Beer Bar	Grocery Store	Taco Place
25	Manhattan Valley	Bar	Pizza Place	Indian Restaurant	Coffee Shop	Mexican Restaurant	Playground	Yoga Studio	Furniture / Home Store	Clothing Store	Cosmetics Shop
36	Tudor City	Mexican Restaurant	Park	Café	Greek Restaurant	Diner	Pizza Place	Coffee Shop	Deli / Bodega	Asian Restaurant	Gym

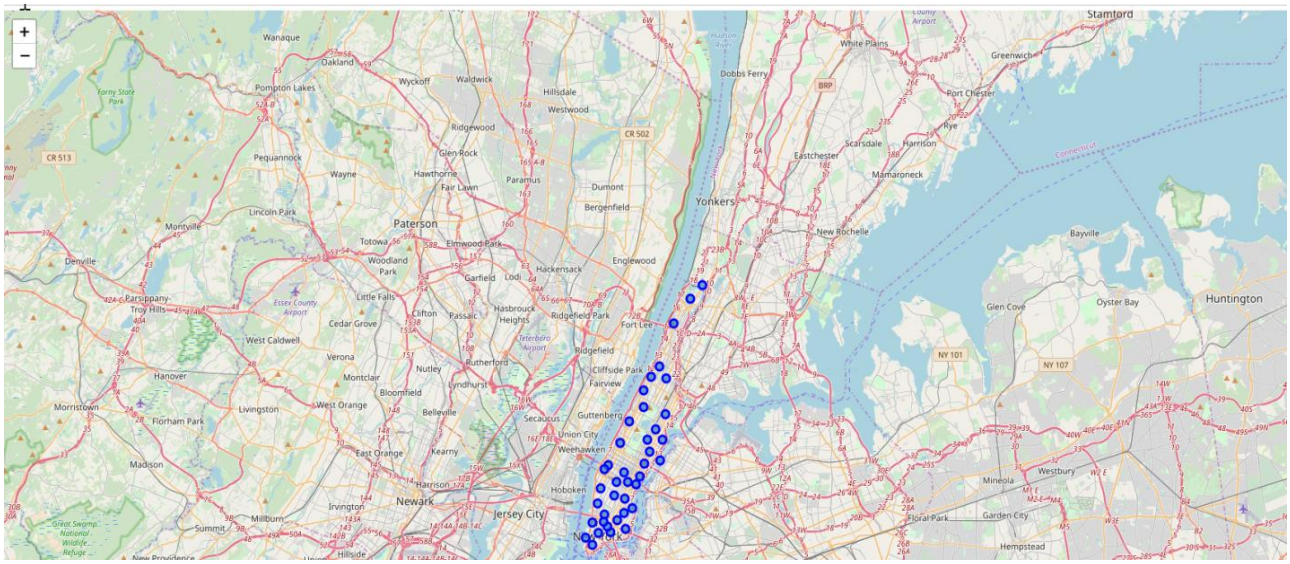
Table 9: Clustering for Downtown Toronto

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Downtown Toronto	0	Coffee Shop	Restaurant	Pub	Flower Shop	Café	Italian Restaurant	Pizza Place	Bakery	Pharmacy	Park
2	Downtown Toronto	0	Coffee Shop	Japanese Restaurant	Gay Bar	Sushi Restaurant	Restaurant	Mediterranean Restaurant	Hotel	Gastropub	Café	Pub
3	Downtown Toronto	0	Coffee Shop	Pub	Bakery	Café	Park	Theater	Mexican Restaurant	Breakfast Spot	Restaurant	Shoe Store
4	Downtown Toronto	0	Coffee Shop	Clothing Store	Middle Eastern Restaurant	Bubble Tea Shop	Café	Japanese Restaurant	Lingerie Store	Italian Restaurant	Theater	Burger Joint
5	Downtown Toronto	0	Coffee Shop	Café	Restaurant	Hotel	Clothing Store	Bakery	Cosmetics Shop	Beer Bar	Diner	Italian Restaurant
6	Downtown Toronto	0	Coffee Shop	Cheese Shop	Café	Beer Bar	Restaurant	Bakery	Farmers Market	Seafood Restaurant	Cocktail Bar	Lounge
8	Downtown	0	Coffee Shop	Restaurant	Thai	Café	Bar	Sushi	Seafood	Bakery	Burger Joint	Steakhouse

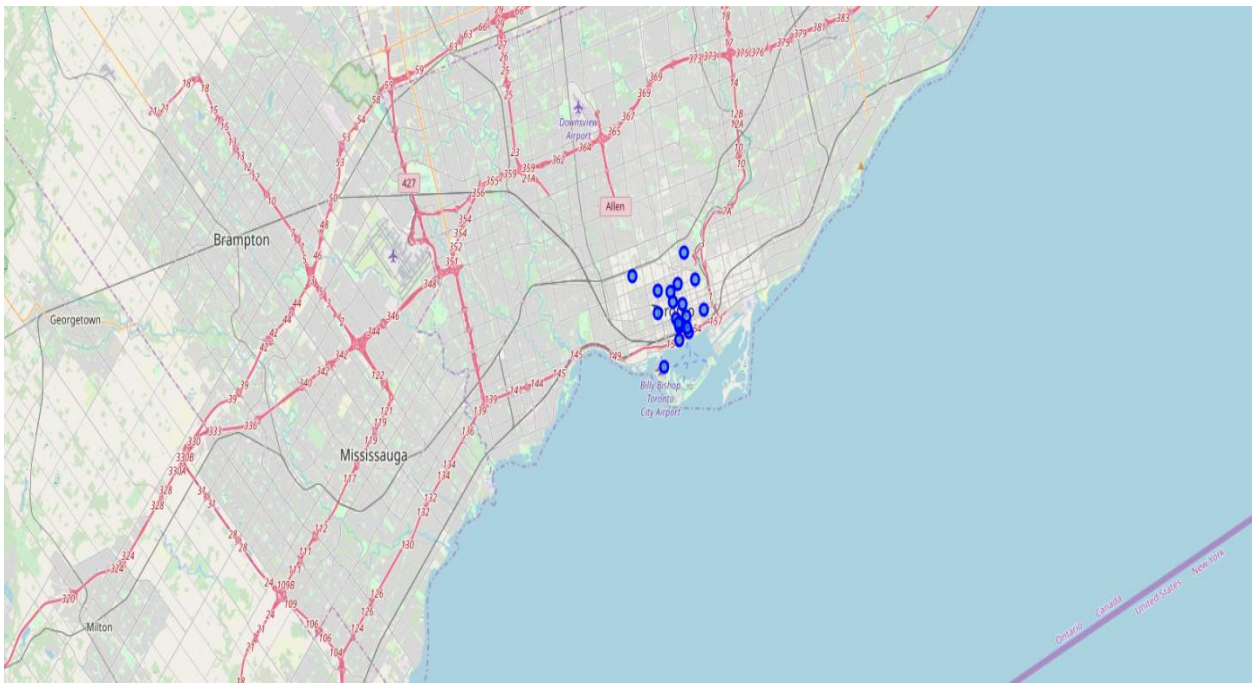
RESULTS

In this section I present the results from Foursquare Api Geolocation. The results are presented in map format both for Manhattan and Downtown Toronto.

Map.1: Manhattan

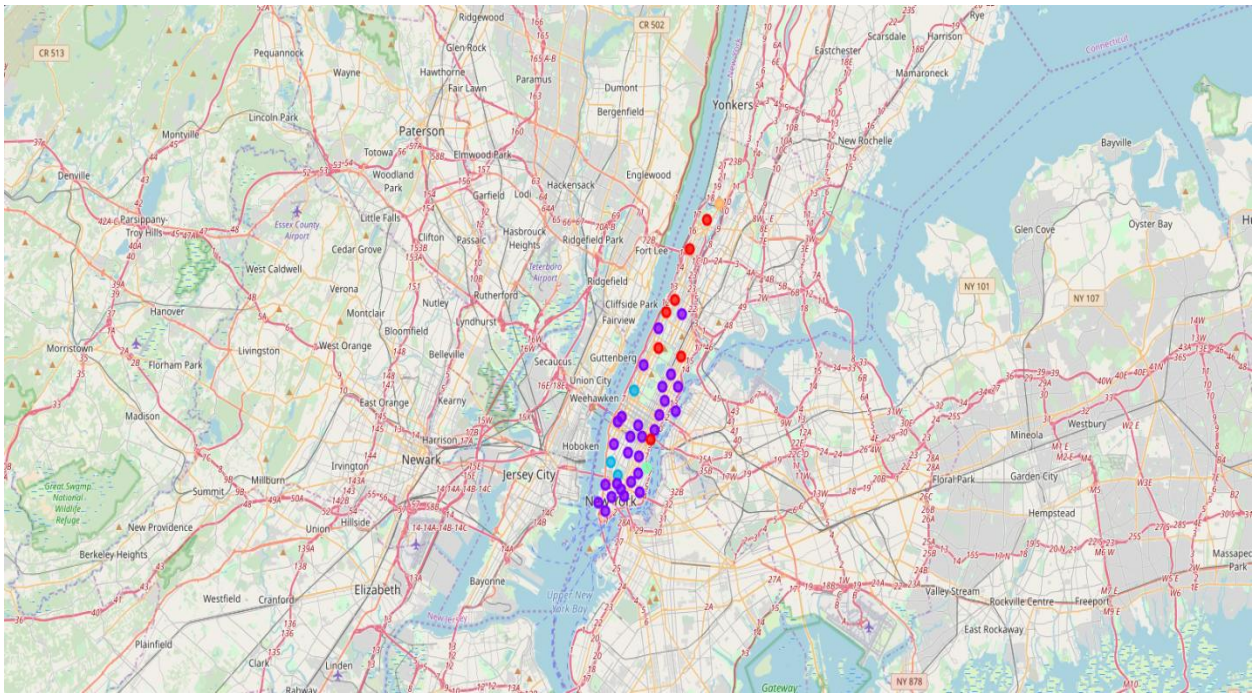


Map 2: Downtown Toronto

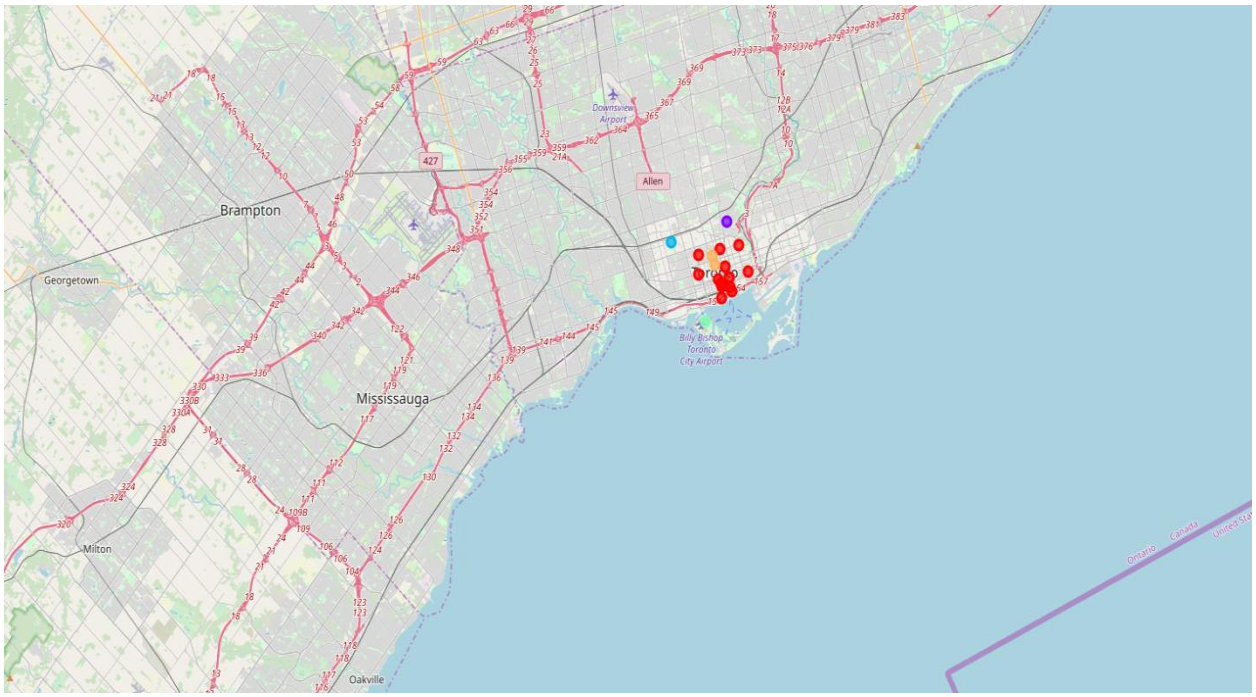


Map Clusters

Map Cluster 1: Manhattan



Map Cluster 2: Downtown Toronto



DISCUSSIONS OF RESULTS

OBSERVATIONS

Based on the data analyzed in this project, it was found that the neighborhoods in Manhattan (New York City) and Downtown Toronto City have significant similarities. On either side we find high population density, retail banks, shopping stores such as markets and farmers markets, waterfall, transportation stations, colleges and Universities, grocery stores.

Although there are a few differences in the neighborhoods between the cities of Toronto and New York, the presence of factors that strike similarities between both cities, as explained above, provides us further insights into the influx of capital and the opportunity for business potential on either side. These similarities also indicate the extent to which these areas have relatively high population density. These are factors, among many, worthy of observation for STYLE Pro in order to open a new branch. These variables not only can serve to attract more investors for STYLE Pro, but also to accommodate the new or transfer employees since either New York City or the City of Toronto, has different amenities and services around which can always be convenient for employees.

RECOMMENDATIONS AND FURTHER OBSERVATIONS

Our research found some limitations on some level of data availability and for that matter, our recommendation will be also limited based on the data that was hereby presented. In the end, when evaluating all the variables that were analyzed here, including the clustering of neighborhoods, population density and availability of other businesses around New York City and Toronto City, this project advocates and recommends STYLE Pro to open a new branch in New York rather than in Toronto. Furthermore, the analysis is supported on the bases that if we look at the data presented by clustering all boroughs in either New York or Toronto, when we factor for Manhattan in New York and Downtown Toronto, it is reasonable to conclude that Manhattan has the highest potential for business opportunities and growth. Thus, we recommend STYELE Pro to consider opening a new business in New York City, precisely Manhattan.

CONCLUSION

To conclude, our research finds that Manhattan has a better prospect for business opportunity when compared to Downtown Toronto. Although both neighborhoods are very similar in many factors that influence geographical location for a given business, yet the two cities are still very dissimilar when it pertains to population density. Population density does have direct influence on market share and the opportunity to acquire and retain new customers. Therefore, based on these factors, we conclude that Manhattan is a better location for a new branch for STYLE Pro.