FINAL REPORT - The Battle of the Neighborhoods

1. Introduction & Business Problem

Problem Background:

Toronto is Canada’s largest city. Its large population of immigrants from all over the globe has also made Toronto one of the most multicultural cities in the world. It provides lots of business opportunities and business friendly environment. It has attracted many different players into the market. This means that the market is highly competitive. As it is a highly developed city so the cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analyzed carefully. The insights derived from analysis will give a better understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk. And the return on investment will be reasonable.

My client is a successful Entrepreneur in Europe. It's only been 3 years since he started his business. But now, in 2020, He has 16 Fried Chicken Restaurants in big cities of Europe like Paris, Berlin, Brussels, Amsterdam etc. And now, he wants to expand his business beyond the Europe. He has a particular interest in Canada. So, he wants to open a new restaurant in Toronto.

Problem Description:

In this project we will try to find the optimal locations for a new **Fried Chicken Restaurant** in Toronto, Canada.

Since there are lots of restaurants in Toronto, we will try to detect locations that are not already crowded with venues, especially restaurants. We are particularly interested in a potential neighborhood with no Fried Chicken Restaurant in vicinity. We would also prefer locations as close to the city center as possible to attract more customers, assuming that the first two conditions are met.

We will use some data science methodology and K-means clustering machine learning technique to generate a few most promising neighborhoods based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by my client.

Target Audience:

Specifically, this report will be targeted to my client who wants to find the optimal location to open a new Fried Chicken Joint in Toronto. But the other stakeholders interested in the same kind of opportunity can also benefit from it.

2. Data

Data 1 :

In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. So we will scrape the data that contain neighborhoods names and their postal code from the following Wikipedia page: 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M' (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M')

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Data 2 :

Then, we will merge it with the data that contain all the geographical coordinates of the neighborhoods thanks to the following csv file: "https://cocl.us/Geospatial_data" (https://cocl.us/Geospatial_data%E2%80%9D)

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Data 3 :

Finally, to get the locations(latitude and longitude) and other information about various venues in Toronto, we will use **Foursquare's API**.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Parkwoods	43.753259	-79.329656	TTC stop - 44 Valley Woods	43.755402	-79.333741	Bus Stop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

3. Methodology and Analysis

We will try to find the possible locations that have a reasonable density of restaurant and other types of venues, in addition to that they should not have any Fried Chicken restaurant nearby.

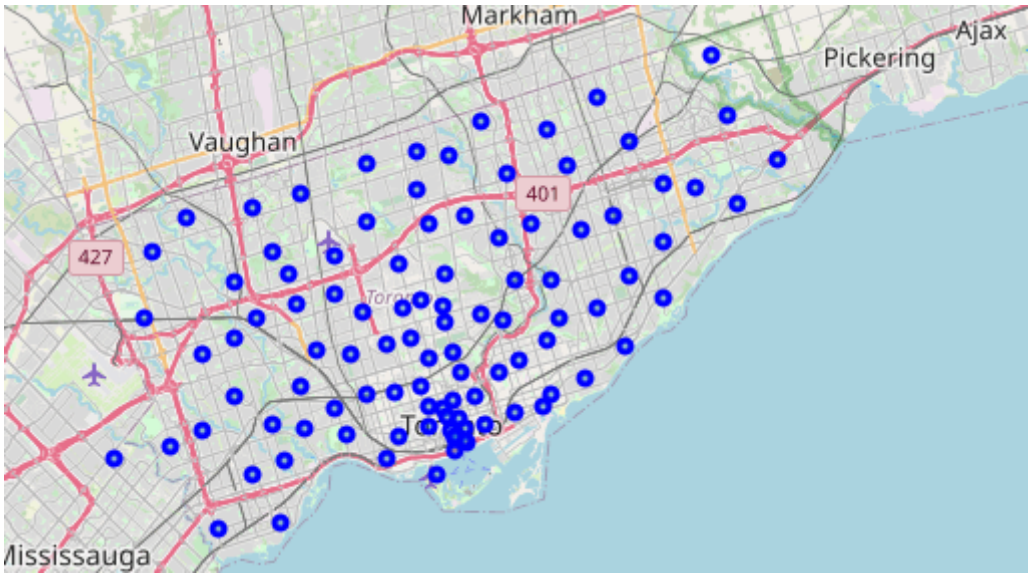
- Firstly, we need to get the list of neighborhoods in the city of Toronto. Fortunately, the list is available in the following Wikipedia page ('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M' (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M')).
- We will do **web scraping** using Python **requests and beautifulsoup packages** to extract the list of neighborhood data.

```
url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
page = urllib.request.urlopen(url)

# parse the HTML from our URL into the BeautifulSoup parse tree format
soup = BeautifulSoup(page, "lxml")

table = soup.find('table', class_ = 'wikitable sortable')
```

- Then, we will **merge** it with the data that contain all the geographical coordinates of the neighborhoods

```
df_coor = pd.read_csv('Geospatial_Coordinates.csv')
df = pd.merge(df, df_coor, on="PostalCode", how="left")
```

- After gathering the data, We need to get the geographical coordinates of Toronto(**Geocoder package**) and we will visualize the neighborhoods in a map using **Folium package**. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Toronto.



- Next, we will use the Foursquare API to get the top 100 venues that are within a radius of 500 meters for each of the neighborhoods.
- We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make **API calls to Foursquare** passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Parkwoods	43.753259	-79.329656	TTC stop - 44 Valley Woods	43.755402	-79.333741	Bus Stop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

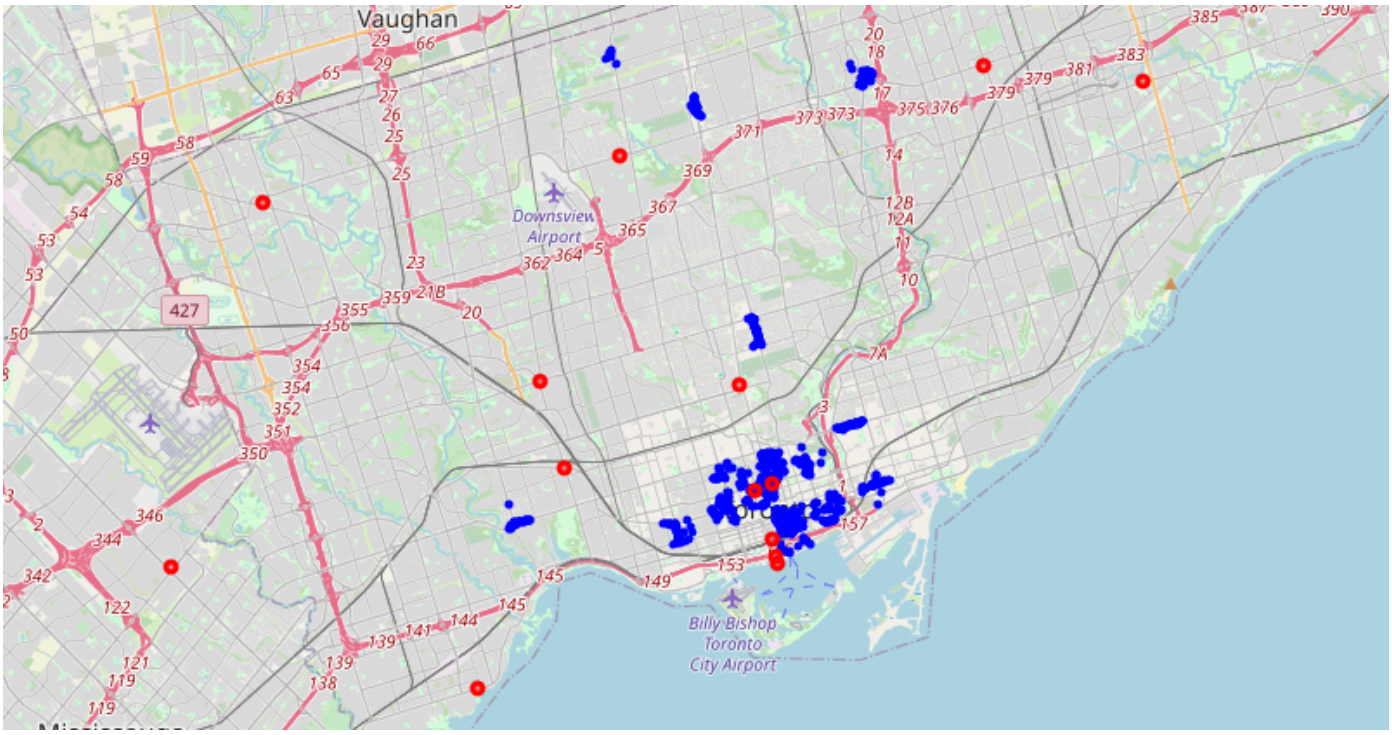
- Now, let's just look at the Fried Chicken Restaurant for further analysis

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	Flock Rotisserie + Greens	43.659167	-79.389475	Fried Chicken Joint
1	Central Bay Street	43.657952	-79.387383	Popeyes Louisiana Kitchen	43.660727	-79.382738	Fried Chicken Joint
2	Cedarbrae	43.773136	-79.239476	Popeyes Louisiana Kitchen	43.775930	-79.235328	Fried Chicken Joint
3	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259	Popeyes Louisiana Kitchen	43.754671	-79.442740	Fried Chicken Joint
4	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752	Joe Bird	43.638204	-79.380355	Fried Chicken Joint

- Like we decide at the beginning of this analysis, we want to find the neighborhoods that don't have too many venues because it may be risky for our new restaurant. In another aspect, we don't want to open our new restaurant in a neighborhood that don't have much potential. So finally, after discussing with my client, we will focus on the neighborhoods that have more than 35 and less than 80 venues that are within a radius of 500m from the center of the neighborhood.

```
most_venues=Toronto_venues.Neighborhood.value_counts().to_frame()
optimal_venues = most_venues[(most_venues.Neighborhood < 80) & (most_venues.Neighborhood >= 35) ]
optimal_neigs = optimal_venues.index.tolist()

df_35_80 = pd.DataFrame()
for neig in optimal_neigs:
    df_35_80 = df_35_80.append(Toronto_venues[Toronto_venues['Neighborhood'] == neig], ignore_index=True)
```

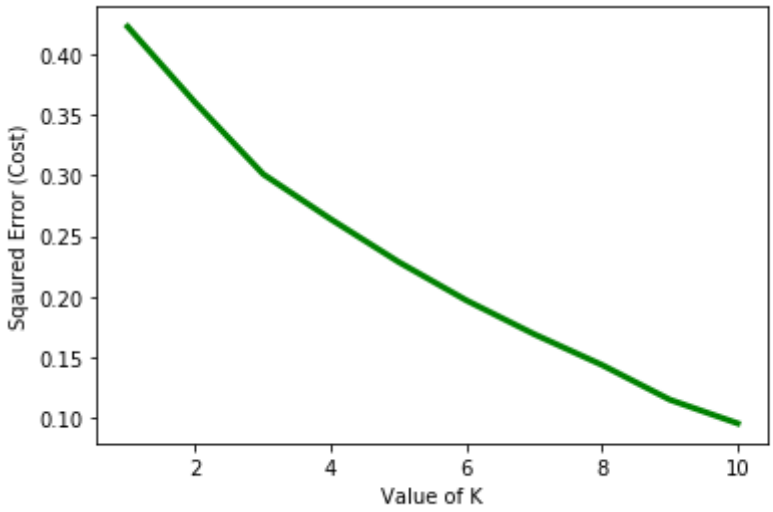
- Visualize the Fried Chicken Restaurant and other venues for determining neighborhoods.



- Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category and let's see the top 10 venues for each neighborhood.

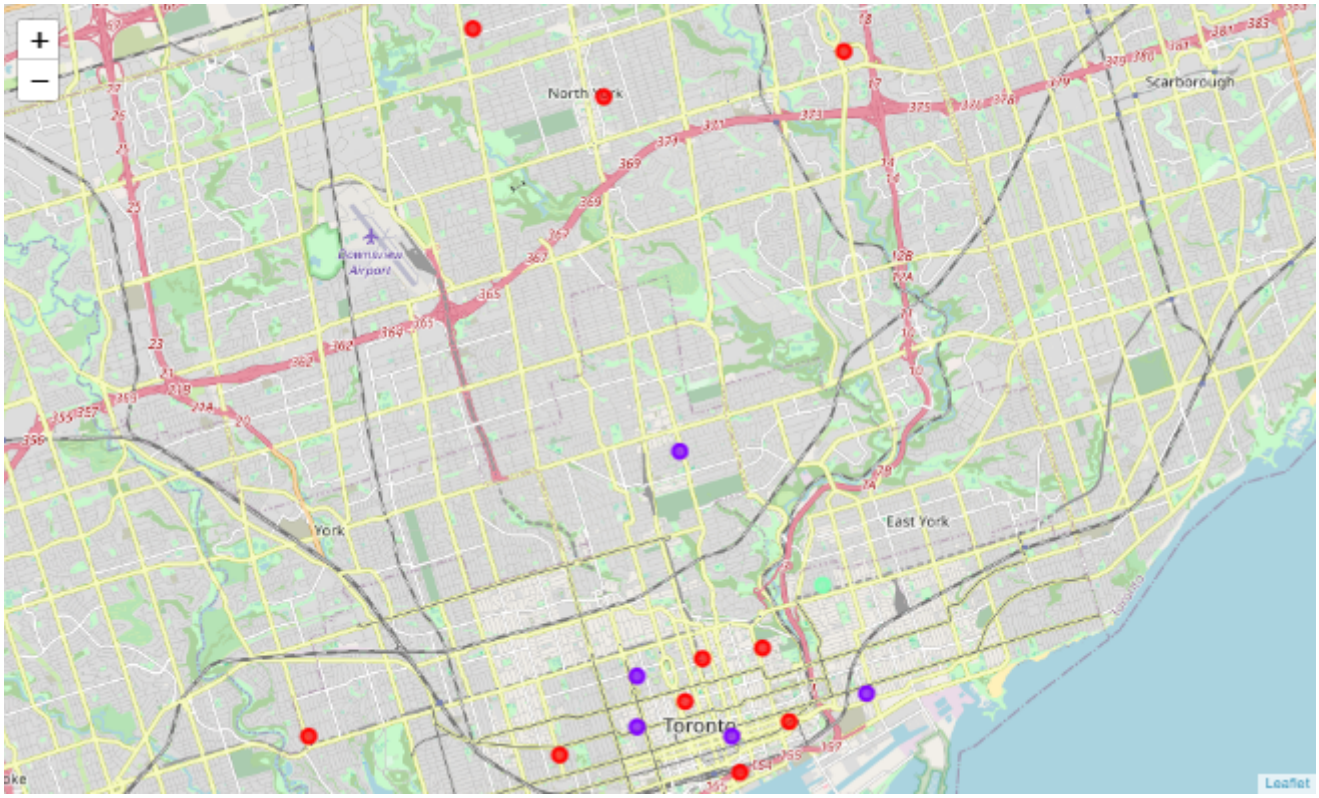
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Cocktail Bar	Bakery	Café	Cheese Shop	Seafood Restaurant	Beer Bar	Restaurant	Hotel	Eastern European Restaurant
1	Central Bay Street	Coffee Shop	Italian Restaurant	Café	Sandwich Place	Thai Restaurant	Bubble Tea Shop	Burger Joint	Bar	Japanese Restaurant	Ice Cream Shop
2	Church and Wellesley	Coffee Shop	Japanese Restaurant	Sushi Restaurant	Restaurant	Yoga Studio	Men's Store	Mediterranean Restaurant	Gastropub	Pub	Gay Bar
3	Davisville	Dessert Shop	Pizza Place	Café	Sandwich Place	Gym	Italian Restaurant	Sushi Restaurant	Coffee Shop	Indian Restaurant	Farmers Market
4	Fairview, Henry Farm, Oriole	Clothing Store	Coffee Shop	Fast Food Restaurant	Restaurant	Japanese Restaurant	Bank	Shoe Store	Toy / Game Store	Bakery	Boutique

- Lastly, we will perform clustering on the data by using k-means clustering. **K-means clustering** algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.



4.RESULTS

The results from the k-means clustering show that we can categorize the neighborhood into **3 clusters** The results of the clustering are visualized on the map below with cluster 1 in red color, cluster 2 in purple color, and cluster 3 in a mint green color.



- **Cluster 1 :** Most of the neighborhoods fall into this cluster. There are mostly business areas with coffee shops, pizza places, restaurants, bar, etc.. There are also social activity venues. Some of the neighborhoods are close to the University of Toronto. So they are at the center of Toronto. So it means high cost high gain for a new business. Some of the neighborhoods are far away from the center.

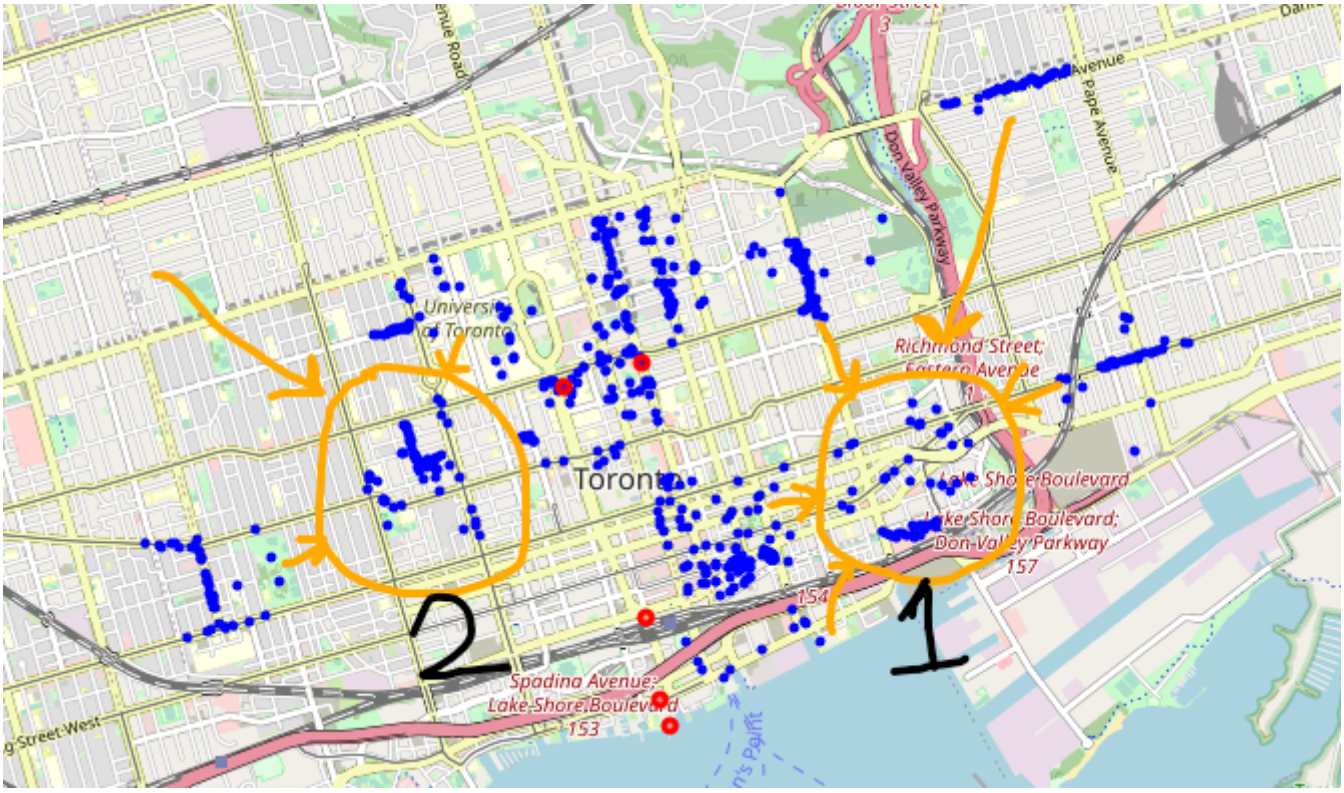
- **Cluster 2 :** 40% of neighborhoods are in this cluster. There are mostly business areas with cafe, restaurants, bar, etc.. The neighborhoods are a little bit far from the center of Toronto. So it means high cost, high gain for a new business
- **Cluster 3 :** There are generally restaurants, coffee shops, etc. The neighborhoods are near the center of Toronto.

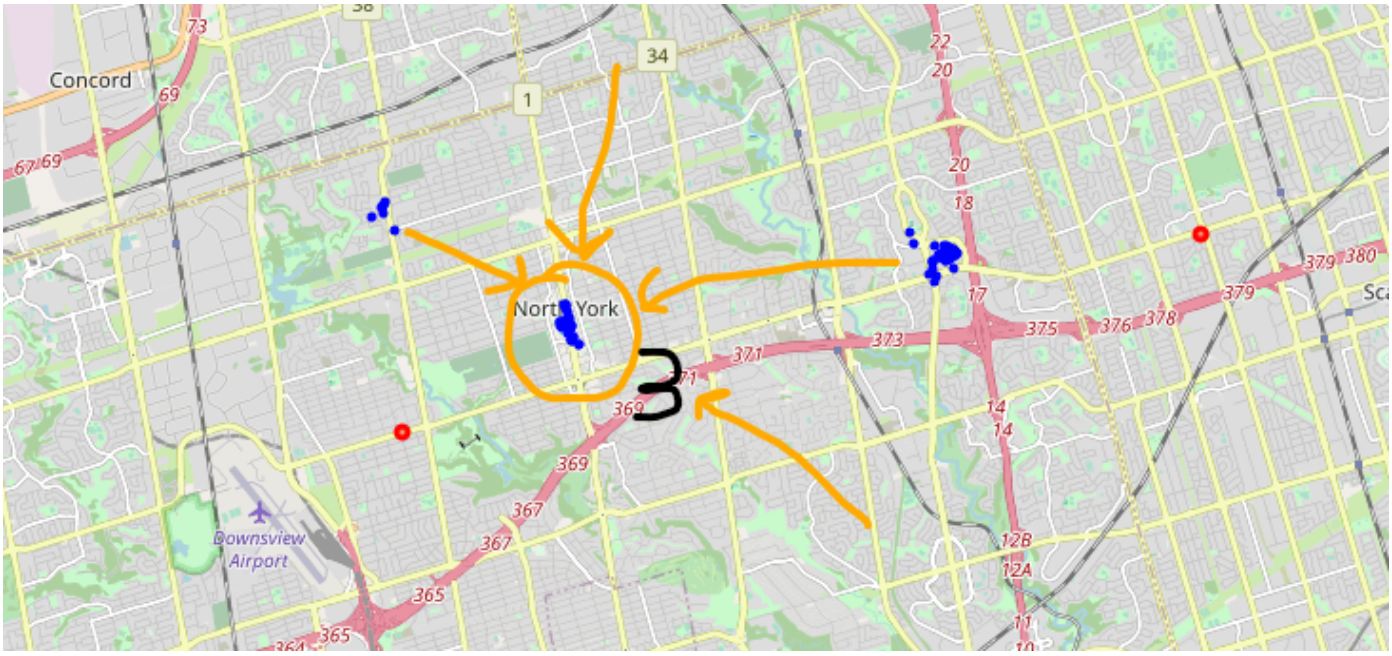
5.DISCUSSION

In this project, we only consider one factor, i.e. frequency of venues, there are other factors such as population, demographics and income of residents that could influence the location decision of a new business. However, to the best knowledge of this researcher, such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Fried Chicken Restaurant.

6.CONCLUSION

The purpose of this project was to identify Toronto areas close to center with reasonable number of restaurants and venues in order to aid my client in narrowing down the search for optimal location for a new Fried Chicken Restaurant. By seeing the density of restaurants and venues from Foursquare data we have identified the borouhgs that don't have a Fried Chicken Restaurant and also have a normal density of venues and restaurants.





1. Regent Park, Harbourfront

    - According to the criteria and results of this analysis, it seems as the best option.
    - The venues and restaurants density is not saturated.
    - There is not other chicken restaurant so close.
    - The neighborhood is close to the center of the city and other neighborhoods that don't have a fried chicken restaurant

2. Kensington Market, Chinatown, Grange Park

    - It may be the best option but there are two other fried chicken restaurants close by.
    - High cost, high gain for a new business

3. Willowdale

    - Good place to start with a new business in a new country to see how it will work.

- Moderate cost, moderate risk.

The final decission on optimal restaurant location will be made by my client based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood, etc.