# Capstone Project - The Battle of Neighborhoods

# Data science analysis report

**Author: Dmitry V.**

**Date: 19.06.2019**

## Contents

# Launch delivery Startup Company Analysis report

## Introduction section

Nowadays we see developing and expansion of new technologies for delivery services. Online stores experiments with copters equipped with grippers to deliver items ordered by customers. It could potentially give big benefits for customers and companies. Following this trend, a new Startup Company has been established to piloting new delivery technologies for lunch delivery to city offices.

The idea is to find best place to start business in Canada. For that it is obvious we should consider major cities: Toronto, Montreal and find best place there.

**The assumption** is that we will make best profit in places where there are a lot of offices and their density is most high

## Business problem section

We selected major Canada cities: Toronto, Montreal. In order to get best start we would like to have an answer to questions below:

**Question 1:**

Which selected city is better in terms of potential revenue, meaning having greater number of customers?

**Question 2:**

After question 1 is answered, what would be the best area in chosen city to locate delivery parking facility?

Data science project has launched to address both questions and get presentation and report to support business decisions.

### *Stakeholders*

Below are the stakeholders and their benefits

*New Startup Company.*

- To succeed pilot project, get maximum revenue to further extend new technologies of delivery to conquer market and increase revenues

*Restaurants.*

- Fast delivery can attract more customers and help with competition

*Customers.*

- Having launch delivery faster will make customers happier

## Data section

For this project we need the following data:

1. Toronto, CA, Montreal, CA data that contains list of Boroughs along with their latitude and longitude.

- Data source : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, https://en.wikipedia.org/wiki/Boroughs_of_Montreal and get latitude, longitude from https://cocl.us/Geospatial_data
- Description: This data set contains the required information. And we will use this data set to explore various boroughs of selected cities.

2. City offices in each borough of selected cities.
    - Data source : Foursquare API
    - Description: By using this API we will get all the venues in each borough. We can filter these venues to get only offices.
3. Geo location data
    - Data source : Generating GeoJSON File for Toronto FSAs https://medium.com/dataexplorations/generating-geojson-file-for-toronto-fsas9b478a059f04
    - Description: By using this geo space data we will get the Toronto, Montreal Borough boundaries that will help us with visualization using Choropleth map.

Collected data was then prepared and cleaned before analysis using data frames and mapping techniques using Folium. These processes established which city has a greater number of potential customers. For the second part of the Data Science problem, potential customers within the selected city were then segmented and clustered. The results overlaid on a map of the city. The data maps and clustering results established the ideal borough within the city.

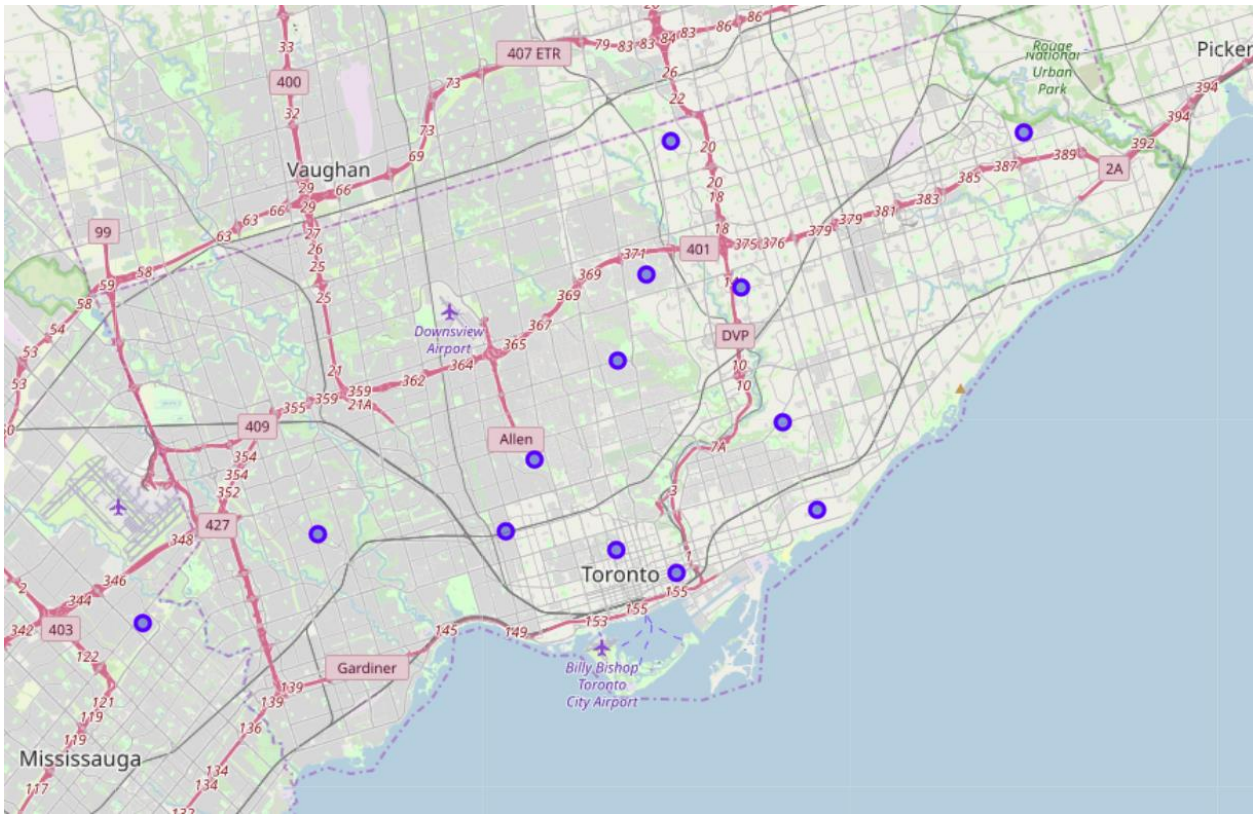
## Methodology/Approach/Analyze section

For **the** Question 1 the process for each city was to data scrape from Wikipedia the boroughs and postcodes of each city location importing them as a .csv file before creating a data frame.

Longitude and Latitude co-ordinates for each borough were also imported into a data frame. Both data frames were then merged together (see data frame below).

| PostCode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| M3A | North York | Parkwoods | 43.753259 | -79.329656 |
| M5A | Downtown Toronto | Harbourfront | 43.654260 | -79.360636 |
| M7A | Queen's Park | Queen's Park | 43.662301 | -79.389494 |
| M9A | Etobicoke | Islington Avenue | 43.667856 | -79.532242 |
| M1B | Scarborough | Rouge | 43.806686 | -79.194353 |

Geopy library was utilized to get the latitude and longitude values of each city which was then used along with the merged Borough data frame to create a map with the boroughs superimposed on top.




The Foursquare API was then used to create a .json file of potential customers within a radius of 750 meters of the center of each city. This radius was chosen because of Foursquare API limitation but it also can be chosen for the fastest delivery time. The resulting output was achieved by using a specific category ID within the foursquare URL, in this case ‘Offices’ the foursquare category included subcategories of ‘ Co-working Space, Tech Startups, Advertising Agency, Campaign Office, Corporate Amenity’. The results were then imported into its own data frame. Making particular note of the data frames size each time.

| | name | categories | lat | lng |
|---|---|---|---|---|
| 8 | Kronos Canadian Systems | Office | 45.495478 | -73.618936 |
| 9 | Virtuo360 Virtual Tour | Advertising Agency | 45.496909 | -73.608657 |
| 10 | Montreally | Advertising Agency | 45.496304 | -73.608806 |
| 11 | MontréAgence | Advertising Agency | 45.497300 | -73.610158 |
| 12 | et voila canada | Advertising Agency | 45.497774 | -73.611016 |

```
nearby_offices_montreal.shape
```

```
(13, 4)
```

Following the same process for each city location, resulted in a clear choice the city of Toronto to explore further using data analysis for Question 2.

Question 2

Was focused on the analysis of the chosen city's individual boroughs in order to find the greatest area of potential delivery parking facility placement. Once again Foursquare API was utilized using the specific category ID of 'Offices' but this time for each borough within Toronto.

| | Borough | 1st Most Common Office | 2nd Most Common Office | 3rd Most Common Office |
|---|---|---|---|---|
| 0 | Central Toronto | Office | Tech Startup | Medical Center |
| 1 | Downtown Toronto | Office | Tech Startup | Coworking Space |
| 2 | East Toronto | Advertising Agency | Tech Startup | Office |
| 3 | East York | Office | Tech Startup | Medical Center |
| 4 | Mississauga | Office | Coworking Space | Building |

Each borough was segmented and then clustered using K-means clustering. To find the borough with the highest potential customer numbers and their business type.

| | Latitude | Venue Longitude | Venue Category | Cluster Labels | 1st Most Common Office | 2nd Most Common Office | 3rd Most Common Office |
|---|---|---|---|---|---|---|---|
| 0 | 43.753259 | -79.327941 | Office | 0 | Office | Coworking Space | Tech Startup |
| 1 | 43.753259 | -79.332662 | Office | 0 | Office | Coworking Space | Tech Startup |
| 2 | 43.753259 | -79.333319 | Office | 0 | Office | Coworking Space | Tech Startup |
| 3 | 43.753259 | -79.325866 | Office | 0 | Office | Coworking Space | Tech Startup |
| 4 | 43.753259 | -79.332035 | Coworking Space | 0 | Office | Coworking Space | Tech Startup |
| 50 | 43.806686 | -79.195614 | Office | 0 | Office | Coworking Space | Tech Startup |
| 51 | 43.806686 | -79.189754 | Office | 0 | Office | Coworking Space | Tech Startup |
| 52 | 43.806686 | -79.198802 | Office | 0 | Office | Coworking Space | Tech Startup |
| 53 | 43.806686 | -79.199744 | Coworking Space | 0 | Office | Coworking Space | Tech Startup |

A map was created of Toronto with the clusters overlaid. This and the resulting data

clearly revealed the borough's with the highest business potential for the client and through analysis of the results narrowed the choice to one borough in particular (see Results section).

## Results section

After data collection, cleaning and analysis we can now answer the initial data science problem posed in the introduction of this project.

**Problem/Question 1:** Which out of two possible locations (Toronto and Montreal) has the greater number of potential offices/customers?

```
print('{} Nearby offices in Montreal.'.format(nearby_offices_montreal.shape[0]))
print('{} Nearby offices in Toronto.'.format(nearby_offices_toronto.shape[0]))
```

```
13 Nearby offices in Montreal.
97 Nearby offices in Toronto.
```

**Answer**: Toronto as it offers the greater number of potential offices/customers for starting the business.

**Problem/Question 2:** In the city chosen in answer to Problem/Question 1, in this case Toronto, which city borough would offer the best place for the business?

Further analysis of the boroughs of Toronto revealed a cluster containing more potential offices/customers than the others. **Cluster 3** contained the greatest concentration of offices. Customers in this cluster are in Offices/ Tech Startup Offices.

**Cluster 3** consists of three boroughs which are Downtown Toronto, Queen's Park and Mississauga. We are looking at the one borough location, so we showed each cluster onto a map of Toronto. (**Cluster 3** boroughs are in blue on the map) From this data, the city borough which would offer the greatest place for the business are Downtown Toronto and Queen's Park. he third option while having a large concentration of potential customers is isolated and would reduce opportunities for future business growth. Between first two options Downtown Toronto would be preferable having more potential offices/customers.

**Answer**: The city borough of Montreal which would offer the best place for starting the business is **Downtown Toronto**.

## Discussion section

Through using the Foursquare API we were able to analyze a large data set for each of locations. However, for more detailed analysis the data set could be explored further, down to neighborhood level.

## Limitations section

- The accuracy of data depends purely depends on the data provided by Foursquare and geo location services
- Foursquare limits number of records returned, so we cannot easily get more detailed information. This also limit us from increasing radius of searching

## Conclusion section

Data science project implemented providing all necessary data, analysis and reasoning to make decision on where to better start business and locate launch delivery facilities. It should be noted that there are assumptions and limitations (described in Report) which should be taken into account for further investigation and data analysis refinement