BATTLE OF NEIGHBORHOODS

Shlok Desai

July 06, 2020

# Introduction

## Background

Toronto, home to banking companies, corporate headquarters, high-powered legal and accounting firms, insurance companies and stockbrokers. In turn, the presence of so many decision-makers has brought advertising agencies and marketing companies. The banks have built large office towers, much of whose space is leased to these companies. Manhattan, the Financial District encompasses roughly the area south of City Hall Park in Lower Manhattan but excludes Battery Park and Battery Park City. The former World Trade Center complex was in the neighbourhood. the neighbourhood Includes the successor One World Trade Center.

## Problem

With rapidly evolving environment and new emerging opportunities, With the help of four-square API and tools and techniques of data science, this challenging business problem can be resolved. which city would be preferable for a new business to thrive? Should the business consider factors like more unique neighbourhoods? Which neighbourhood will be optimal for a business to thrive? What kind of business will be suitable for a neighbourhood?

## Interest

Optimal locations to test market for a new business especially restaurants/Café. Optimal locations for old already established restaurants to dominate already established market.

# Data acquisition and cleaning

## Data sources

1)'https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M' , Geospatial\_Coordinates.csv Data on postal codes and information on city of Toronto.

2)Foursquare Location data of Toronto and New York. Description: Includes neighbourhoods, Preferable venues.

3) <https://cocl.us/new_york_dataset> Data of New York

Based on definition of our problem, factors that will influence our decision are:

* number of existing restaurants in the neighbourhood (any type of restaurant)
* number of and distance to restaurants in the neighbourhood if any

We decided to use regularly spaced grid of locations, cantered around city center, to define our neighbourhoods.

Following data sources will be needed to extract/generate the required information:

* number of restaurants and their type and location in every neighbourhood will be obtained using **Foursquare API**

## Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values, because of lack of record keeping. I decided to remove rows having empty values.

Data used by me did not have coordinates in it so, I downloaded csv file in python and merged two columns with defining postal code as primary key.

The data frame will consist of three columns: Postal Code, Borough, and Neighbourhood. Multiple entries existed for neighborhood

Only processed the cells that have an assigned borough.

More than one neighborhood existed in one postal code area. For example, in the table on the Wikipedia page, you will notice that **M5A** is listed twice and has two neighborhoods: **Harbourfront**and **Regent Park**. These two rows will be combined into one row with the neighborhoods. Cleaned the data.

If a cell has a borough but a **Not assigned**neighborhood, then the neighborhood is assumed same as the borough.

After fixing these problems, I checked for outliers in the data. I found there were some extreme outliers, mostly caused by some types of small sample size problem.

## Feature selection

After data cleaning. Upon examining the meaning specific feature, it was clear that there was some redundancy in the features. For example, features were specific to neighborhoods and grouped the data set according Borough.

Boroughs had high entropy and it looked like **it was necessary to cluster neighborhoods based on minimizing inter-distance between cities in same cluster and maximizing intra-distance between cities in different cluster.**

Using matplot lib, visualized the data to see relevance of clusters location to better understand the data (**figure 1 and figure 2**).

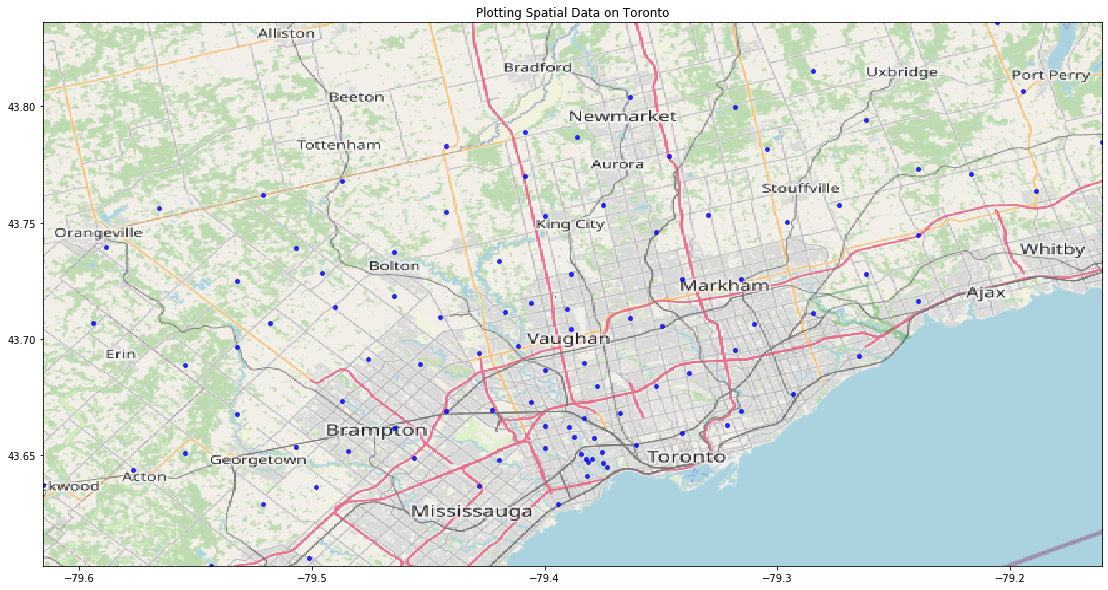


Fig 1 Map of Toronto

A picture containing text, map

Description automatically generated

Fig 2 Map of New York city plotted in python

# Exploratory Data Analysis

3.1. Using four-square API called best venues and coordinates different neighborhood in New York city.

Called data by using credentials in python where locations were normal calls whereas, favorite venues were premium calls. Checked number of unique venue and neighborhood to see how many restaurants already existed in that neighborhood.

Clustered neighborhoods by using Euclidean distance formula and K-Mean clustering. And segmented neighborhood based on nearness with each other.Cluster with least number of restaurants were preferable for new comers as they will face very less competition initially whereas, clusters which were highly dense were preferable for well established restaurants.

## Methodology

In this project we will direct our efforts on detecting areas of Toronto and New-York city with restaurant density, particularly those with low number of restaurants.

In first step we have collected the required **data: location and type (category) of every restaurant from Toronto and New-york**.(according to Foursquare categorization).

Second step in our analysis will be calculation and exploration of '**restaurant density**' across different areas cities of Toronto and Manhattan in New York by clustering. we will use **K-means clustering** to identify a few promising areas close to each other with low number of restaurants in general and focus our attention on those areas.

In third and final step we will focus on most promising areas and within those create **clusters of locations that meet some basic requirements** established in discussion with stakeholders: we will take into consideration location we want locations **restaurants at most prefered venue**. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

## 5 Results and Discussion

Our analysis shows that although there is a great number of restaurants in Toronto and New york (~500 in our initial area of interest), there are pockets of low restaurant density fairly close **cluster 2 and 4** in Manhattan. Highest concentration of restaurants was detected in **cluster 0 and 1** of manhattan, so we focused our attention to those areas , but our attention was focused on **cluster 0** which offer a combination of popularity among restaurants, which can be an indicator of high demand. **cluster 2 and 4** still have less density of restaurants in manhattan and so if a restaurant business wants to start a thriving business, after evaluating demand, it can be best location.

Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all these zones containing largest number of potential new restaurant locations based on number of and distance to existing venues restaurants in general. This, of course, does not imply that those zones are actually optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas close to Berlin center but not crowded with existing restaurants (particularly Italian) - it is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition, but also other factors taken into account and all other relevant conditions met.

**Segmentation based on k-Means clustering is shown in Jupyter notebook**

## 6 Conclusion

Concluding, for Manhattan, **cluster 2 and 4** are good for someone new to restaurant business as there is less competition and a good market to test. whereas **cluster 0, 1, 3** are a good indicator of people fond of eating and already established restaurants can enter these clusters to dominate the market.

for Toronto, segmentation has very less entropy and restaurants should enter **cluster 0,1,3,4** as that cluster has a large network of neighborhood and diverse market.