

by

Ojo Akinwale

Capstone Project for IBM Data Science Professional Certificate

IBM

|  |  |
| --- | --- |
|  |  |

Submitted April 2019

**INTRODUCTION/BUSINESS PROBLEM**

In this project, we will try to identify an optimal location for a coffee cafe in Toronto, the most populous city in Canada. Specifically, we are interested in opening a coffee drink cafe in Toronto.

The competition in the coffee business in Toronto is high and is most likely the business for everyone. The total population of Toronto is more than two million and the coffee cafe market is a large one which allows us to have a target audience and also market share. It is especially important to find the right location. We are particularly interested in areas with less or no coffee cafe in the vicinity. We would also prefer locations with a high volume of people who are lovers of coffee and this is done through feasibility study of the vicinity and with the use of the Foursquare API.

This report will provide data analyst and potential investors important information needed, it will provide some major analysis of the coffee cafe distributions in the neighbourhoods of Toronto and provide some suggestions on the potential good locations.

* **Problem**

London's administrative scope includes 10 Boroughs and 241 Neighbourhoods, but the level of infrastructure construction varies widely among different administrative districts and communities. For example, tourist areas such as Westminster have many cafes, hotels and restaurants, which are convenient for tourists and residents, but in the vast suburbs outside London's 5th district, he quantity and quality of the restaurants, supermarkets, parking lots, etc. are not satisfactory.

* **Interest**

People in the suburbs will be concerned about this problem, because the fewer recreational facilities result in lower quality of life, and poorer service provided reduces the standard of living and there’s no value for money spent. Real estate developers will be concerned about this issue, because the surrounding infrastructure level will directly affect housing prices. Local government officials will be concerned about this issue, because the imbalance of regional development can be guided and re-strategized by policies.

**DATA**

* **Data source**

Three cities' data (London, Toronto and New York) will be horizontally compared in this project, hence, the dataset of this project will be divided into the three parts, and the data sources of each part are different.

We captured all UK postal code, longitude and dimensions from [public sources](https://raw.githubusercontent.com/Gibbs/uk-postcodes/master/postcodes.csv), and captured the basic data such as London's postal code, Borough and Neighbourhood from the government statistical report, In addition, we use the Foursquare API to obtain process data (venues, venues category, venues frequency etc.) based on geographical location. After integrating and normalizing the relevant information, we got the integrated **London data**.

We grabbed basic information like Toronto's postal code, Borough and Neighbourhood from the JS document on the [Wikipedia page](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), and we used the Foursquare API to obtain process data (venues, venues category, venues frequency etc.) based on geographical location. After integrating and normalizing the relevant information, we got the integrated **Toronto data**.

Comparing to the separated datasets of London and Toronto, we can directly download more completed and unified New York dataset from the [web address](https://cocl.us/new_york_dataset) provided by IBM and Coursera. The datasets including postal code, Borough, Neighbourhood and latitude and longitude information. Similarly, we used the Foursquare API to obtain process data (venues, venues category, venues frequency etc.) based on geographical location. After integrating and normalizing the relevant information, we got the integrated **New York data**.

The data scraped from the Wikipedia page shows the Postal code, Borough and Neighbourhood which helped us in analysing the different business in different areas (through postal code) and the neighbourhood each is situated.

For instance, from the data, the analysis showed that The Beaches alone has two coffee cafes; Starbucks and Dip 'n Sip.

Based on the business problem, factors that will influence our decisions are:

(a) The total number of coffee cafe in the neighbourhood.

(b) The percentage of coffee cafe in the neighbourhood.

(c) The populations in the neighbourhood which could be filtered to only those who loves coffee.

* **Data cleaning**

The raw data (London data, Toronto data and New York data) may have some duplicate values, null values, or Unassigned values, the data format has multiple forms such as String, Int, Float etc., so it needs to implement pre-processed methods such as deletion, merging, adjustment, standardization, regularization, and unification are carried out so that it can be substituted into the machine learning algorithm to obtain the model.

A screenshot of a cell phone

Description automatically generated

Fig 1: The cleaned data

* **Feature selection**

Based on the exercises at the Coursera Lab, we identified several key variables, including Postcode, Borough, Neighbourhood, Longitude and Latitude. But unlike the Coursera Lab, this project will focus on the overall planning and comparison among the cities, rather than limit in the core areas of the city (such as Manhattan, downtown Toronto), so the data set should be retained to the greatest extent, and the data bias can be largely avoided in the following Cluster analysis steps.

**METHODOLOGY**

* **Descriptive analysis**

Since only the latitude and longitude of each city dataset are numerical, the explanatory analysis of latitude and longitude indicates the geographical location of most areas of the city. The ranking data obtained through the Foursquare API are String type, which cannot be interpreted analytically. In the clustering step, the interpretative analysis can only be used to count the neighbourhoods contained in different clusters, but such simple counting does not reflect the real situation of the region.

* **Statistical analysis**

Like descriptive analysis, the training set consists mainly of string data, and the digital part does not need to be statistically tested. For example, the correlation tests between the latitude and longitude of the London neighbourhood are meaningless. The latitude and longitude are collected by the point-to-point GPS service, hence, testing its independence, normality or multicollinearity has no practical significance. In summary, the project does not perform statistical tests.

* **Algorithm analysis**

There are some basic information labels such as region, GDP, population, etc. in the training set of each city, but these labels are not the required variables in this project and have been deleted in the feature selection process. Furthermore, the training set of this project can be regarded as “unlabelled” data, hence, unsupervised learning algorithm can be used for regression or classification, Clustering algorithm is such a good choice.

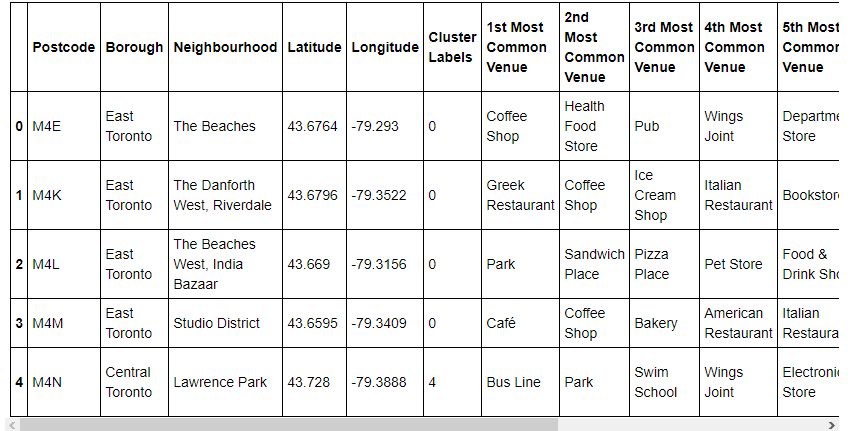
In addition, there is no hierarchical division of the target model, so the top-down or bottom-up hierarchy modelling is not needed; since the regions in the city were gradually established year by year, such as the central London in Zone 1 while the suburbs in high value zones, the density algorithm is not applicable in this case. The K-Means algorithm only needs to formulate the number of clusters, which can directly reflect the different functional areas of the city, so that K-Means will be chosen as the main analytical methodology.

**RESULTS**

* **Toronto Clusters**

Five neighbourhood combinations in cluster 0 are selected for analysis, the keywords in this area are Pub, Coffee Shop, Park etc., so it is clear that these blocks have more entertaining functions.

Cluster 0 looks like a whole bunch of coffee shops and eateries



Toronto Cluster 0

Cluster 1 is set apart by not having as many restaurants around and instead having a gym and playground

A screenshot of a cell phone

Description automatically generated

Toronto Cluster 1

Cluster 2 is unique by having different types of restaurants and more gardens, it has more ethnic food

A screenshot of a cell phone

Description automatically generated

Toronto Cluster 2

Cluster 3 is more outside the city and has more electronics, jewelry, and rural activities

A screenshot of a cell phone

Description automatically generated

Toronto Cluster 3

Cluster 4 has less stuff to do, mostly just bus lines to get into the city and some parks

A screenshot of a cell phone

Description automatically generated

Toronto Cluster 4

**DISCUSSION**

Each area of ​​each city can be divided into a park area, an entertainment area, a tourist area, etc. according to the functions realized, and each area has a corresponding cluster. We thought that in the same city, due to the high similarity of history, culture and economy, the links between the regions will be close, the functions will be similar, and the differences among clusters will be relatively small. Therefore, when a region draws on development experience and benchmarks other regions, it should look more from other regions in the city. The above views are partly correct, but the actual operation is a bit difficult. For example, Toronto Cluster 1 contains only one neighbourhood. If the neighbourhood is to be developed, there is a lack of neighbourhoods to be used as a benchmark.

The results of data analysis prove that the similarity of the same functional areas in different cities is sometimes greater than the similarity of different functional areas in the same city. Moreover, in the results of data analysis, we can find that the regional division function of each city is relatively clear, and the park area, tourist area and store area have their own classification. But we can also find that there are some hard-to-define areas in each city. Compared with other areas, they do not have their own characteristics, and the infrastructure construction is not complete. Therefore, we suggest that these areas first need to identify their own advantages and look for several benchmark objects in other clusters to plan the development path.

# CONCLUSION

In this study, we selected Borough, Neighbourhood, Postcode, Latitude, and Longitude as reference variables, based on k-means=5 pre-sets, and get the five different functional areas of each city, and the model provides a benchmarking object for the development planning of the city block based on its own characteristics. For example, we can choose any neighbourhood in any city, and looking for similar neighbourhood in the same city based on the clusters in which they are located, for cross-regional communication. We can also choose similar neighbourhood in clusters with similar functions in other cities to communicate across countries.

**REFERENCES**

[1] [London Postcode District — Wikipedia](https://en.wikipedia.org/wiki/London_postal_district)

[2] [National Statistics](https://www.ons.gov.uk/)

[3] [Foursquare API](https://developer.foursquare.com/)

[4] [Google Map](https://www.google.com/maps/)

[5] [IBM Cognitive Lab](https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DP0701EN/DP0701EN-3-3-2-Neighborhoods-New-York-py-v1.0.ipynb)