

**A machine learning approach to municipal planning in London**

by

Hanlun Chen

Project for the Profession Certificate from IBM data scientist

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

Submitted April, 2019

# 1. Introduction

## Background

London is the political, cultural and economic center of the UK and Europe. In 2016, the Greater London metropolitan area was 1,577 square kilometers with a population of approximately 8.9 million and a total production value of $553.5 billion (National Statistics, 2017). In recent years, London’s compatibility ranks at the head among the global cities, which make it a truly international metropolis like Toronto and New York. London's past development has not only affected the UK, but even has far-reaching effects on the world, we have reason to believe that the future direction of London will continue to affect countries in worldwide.

## Problem

London's administrative scope includes 10 Boroughs and 241 Neighborhoods, but the level of infrastructure construction varies widely among different administrative districts and communities. For example, tourist areas such as Westminster have a large number of 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 travel facilities result in longer commutes. 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.

# 2. Data

## 2.1. 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 zip 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 Neighborhood 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 zip code, Borough and Neighborhood 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, Neighborhood 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**.

## 2.2. 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.



Figure 1: the cleaned London dataset.

## 2.3. Feature selection

Based on the exercises at the Coursera Lab, we identified several key variables, including Postcode, Borough, Neighborhood, 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.

# 3. Methodology

## 3.1. Philosophical paradigm

This research held that the real life could be measured by the data, there should be the logical relationship between data inputs and machine learning modelling outputs. In order to build more reliable and robust models, this research also simplifying complex differences among cities' neighborhoods to averaging clusters. It is no doubt that the principle of machine learning is data-based, but the clustering algorithms are relatively strong at explanatory to other machine learning approaches (James et al., 2013), so there should be some inductive methodologies. Therefore, as shown in Figure 2, the ontology of this research is realism, the theory of nature is adaptor, the epistemology is positivism, and the methodology is mainly based on quantitative deduction.

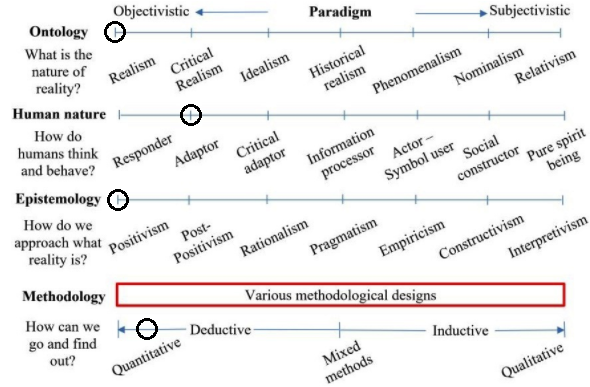


Figure 2: the philosophical assumptions for this program.

## 3.2. 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.

## 3.3. Statistical analysis

## Similar to 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.

## 3.4. 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.

# 4. Results

## 4.1. London clusters

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

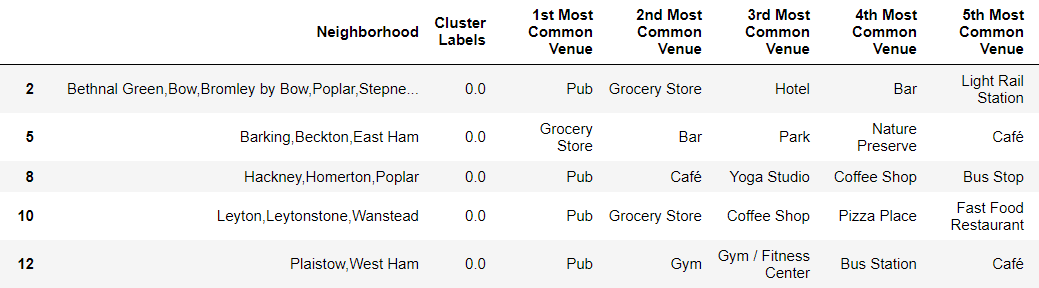


Figure 3: London cluster 0

The cluster 1 has only one subclass, which is especially special compared to other neighbourhoods. By observing the top 5 popular places, we found that this block is mainly composed of parks with some functional sites.

Figure 4: London cluster 1

Similarly, we have reason to believe that the main function of the London cluster 2 is commodity exchange, because this kind of neighbourhoods are mainly consisted of grocery stores, supermarkets, or other kinds of sales markets, hence, we can define it is a supermarket block.



Figure 5: London cluster 2

The composition of London cluster 3 is more complicated. There are pubs, supermarkets, restaurants, and even gymnasiums. We have found many well-known tourist areas by observing the names of the neighbourhoods, so the relatively complete infrastructure is understandable.

Figure 6: London cluster 3

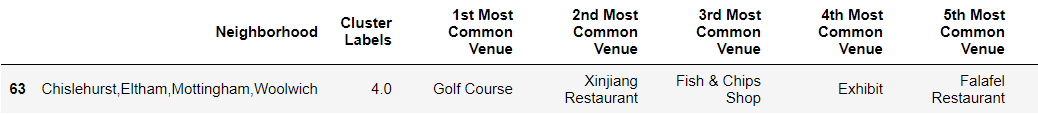
Similar to London cluster 1, London cluster 4 has only one set of data and is a special block. By observing the data composition, we found that the London cluster 4 might be a huge golf course.

Figure 7: London cluster 4

## 4.2. Toronto clusters

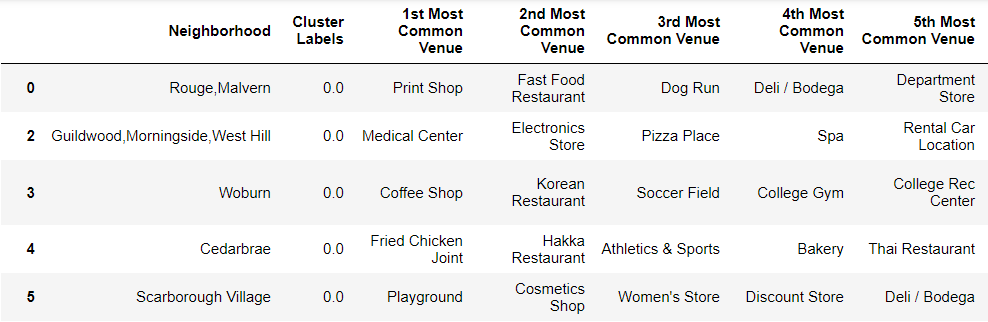
Toronto cluster 0 is similar to London cluster 3, which may be the tourist area because the infrastructure is completed enough.

Figure 8: Toronto cluster 0

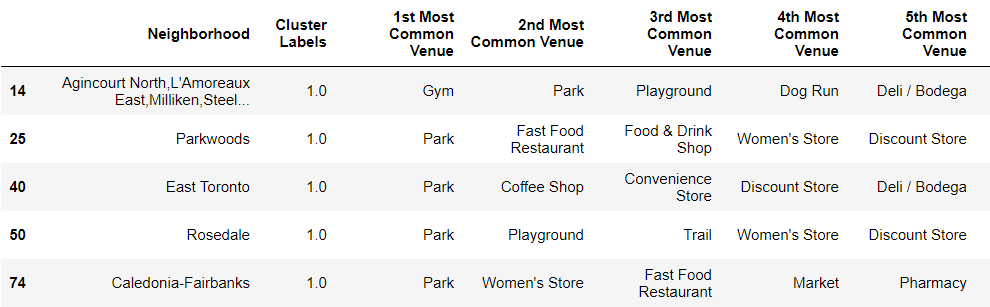
Obviously, Toronto cluster 1 is a park area, similar to London cluster 1.

Figure 9: Toronto cluster 1

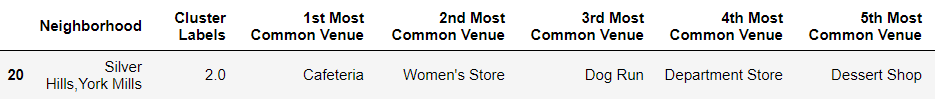
It is difficult to determine what Toronto cluster 2 is, because characteristics of the most common venues are not prominent.

Figure 10: Toronto cluster 2

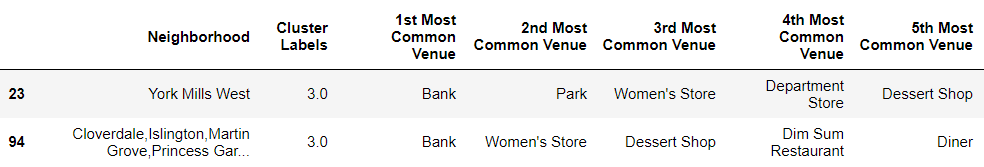
Toronto cluster 3 is banking areas with some storefronts, the functions are quite unique among all clusters.

Figure 11: Toronto cluster 3

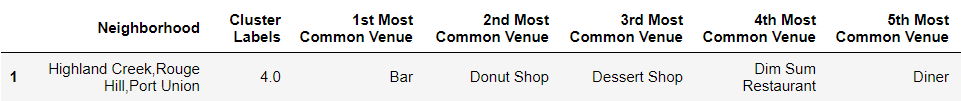
Toronto cluster 4 is the eating area, which based on bars and restaurants. 

Figure 12: Toronto cluster 4

## 4.3. New York clusters

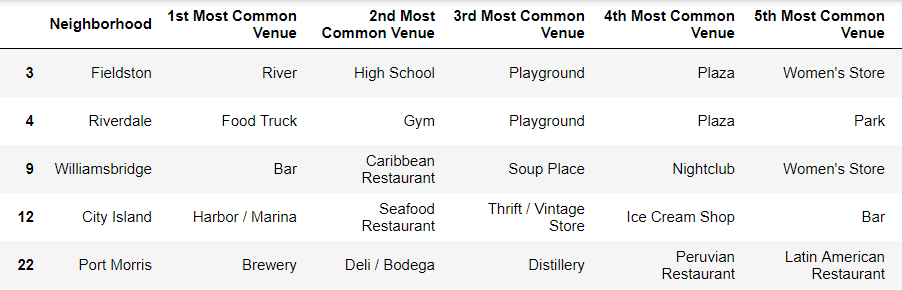
We suspect that New York Cluster 0 is similar to London Cluster 3 and Toronto Cluster 0. It is based on many keywords and is located in the tourist area because the infrastructure is readily available.

Figure 13: New York cluster 0

It is clear that the cluster 1 of London, Toronto and New York are park zone, there are high similarity among such neighbourhoods.

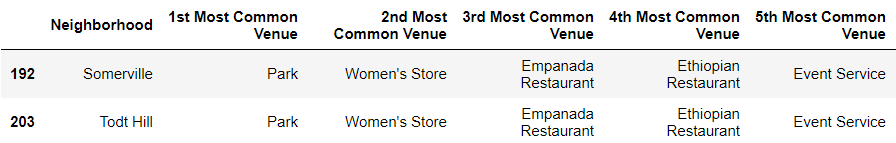


Figure 14: New York cluster 1

The facility foundations of New York cluster 2 are completed enough. However, there are some slight differences between New York cluster 2 and London cluster 0, such as New York cluster 2 have laundromat, metro stations and pharmacy etc., which have more vitality than London’s.



Figure 15: New York cluster 2.

It is hard to define what function the New York cluster 3 has, but we are sure that there are some entertainment facilities nearby.

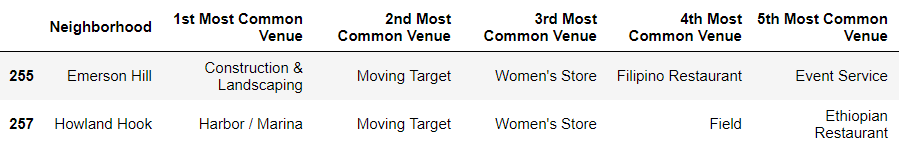


Figure 16: New York cluster 3.

New York cluster 4 is more like a store zone, which is same as London cluster 2. There is no surprised that the bus stops are so popular, because the bus may be the first choice of local residents’ transportation.

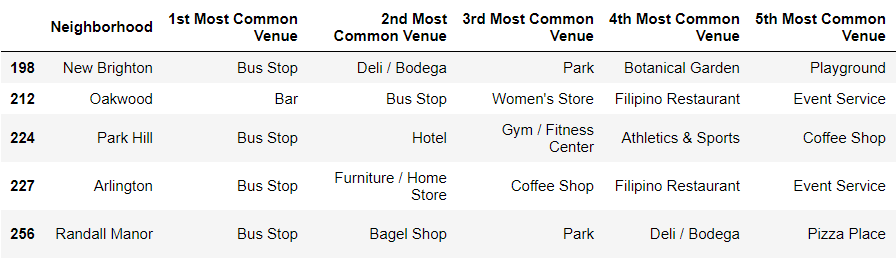


Figure 17: New York cluster 4.

# 5. 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, London Cluster 1 contains only one neighbourhood. If the neighbourhood is to be developed, there is a lack of benchmarking objects in the City of London.

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. For example, clusters 1 in Toronto and New York are park areas, and the park area in London has more reference objects. 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.

# 6. Conclusion

In this study, we selected Borough, Neighbourhood, Postcode, Latitude, and Longitude as reference variables, based on k-means=5 pre-sets we clustering three cities (London, Toronto, and New York), 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.

# 7. 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)