# Clustering of the Island of Montreal

Noémie Rohel

February 14, 2021

## Contents

# 1 Introduction

The neighborhoods of a city, especially the large ones, may be very different; the population, the real estate and the accessible venues can vary a lot even within the same area. However, most of the neighborhoods share a lot of characteristics with others. Therefore to have a quick and easy to understand description of a city, it would be interesting to identify the different kinds of neighborhoods and to associate all of them to a specific profile. This project aims to help people who want to move in to select the best place or people who want to start a business to select the best location. This study is addressing this issue for Montreal.

The project aims to cluster the Island of Montreal into similar neighborhoods based on the environment (density of restaurants, bars, cultural activities, schools, ...), the population (density of population, age, income, ...) and the housing (value, type of dwellings, ...). For each of these categories, a set of specific features is selected and a clustering method is used to gather the similar neighborhoods. Three clustering approaches are considered and implemented: K-Means, DBSCAN and agglomerative clustering, and the results are compared to identify the best method.

First, the data sources used for this project are presented, as well as the required data cleaning and preprocessing. An exploratory data analysis is conducted to highlight some preliminary observations about the collected data. Then for the three clustering approaches considered, a short description and a parameters selection is performed and the results are used to compare the methods. Finally, the clusters obtained with the best approach are described and analysed.

# 2 Data collection

In the context of this project, the considered subdivision of the Island of Montreal is the Forward Sortation Area (FSA), the first three characters of a postal code. The different datasets and APIs used to get relevant data about each FSA are described in this section.

## 2.1 Data sources

### 2.1.1 Foursquare API [3]

The Foursquare API is used to obtain all venues in a specific FSA. For each venue, the following information is available.

- Id, a unique identifier
- Name
- Address
- Latitude and longitude
- City, state, country and country code
- Category id and category name

Using this API, a dataset containing the information of the venues of all FSAs is built.

### 2.1.2 Portail données ouvertes Montréal [4]

Portail données ouvertes Montréal is the City of Montreal's open data portal. It is used to access the map of bus and metro routes of the STM, the transportation company of Montreal. Using QGIS software, we isolate the name and coordinates of all metro stations. Then we use the reverse function of the Nominatim GeoPy's geocoder to get the address and the FSA of the metro stations from their coordinates. Finally, a dataset with the name, coordinates and FSA of the metro stations is built.

### 2.1.3 Partenariat Données Québec [5]

Partenariat Données Québec is Quebec government's open data portal. It is used to get the following datasets.

#### 2.1.3.1 Daycares

The dataset of all daycares in Quebec is available. Only the ones with a postal code beginning with a FSA of the Island of Montreal are kept.

#### 2.1.3.2 Educational institutions

The selected datasets related to the educational institutions available on the Quebec government's open data portal are the following.

- College educational institutions
- University educational institutions
- Government educational institutions for preschool, primary and secondary education, and professional and adult training
- Private facilities for preschool, primary and secondary education, and professional and adult training
- Public schools for preschool, primary and secondary education, and professional and adult training

Only educational institutions with a postal code beginning with a FSA of the Island of Montreal are kept.

### 2.1.4 Walk Score API [6]

A map of the Island of Montreal's FSAs is created with QGIS software using the 2016 Census Boundary file available on the Statistics Canada website [7]. Then random points are generated with QGIS software for each FSA. The number of generated points depends on the size of each FSA to guarantee an equivalent point density. Using the API, the walk score, transit score and bike score are computed for each generated points and finally a dataset containing the following information is built.

- FSA
- Latitude and longitude of the point
- Walk score and walk score description
- Transit score and transit score description
- Bike score and bike score description

### 2.1.5 2016 Census Profile [8]

The 2016 census data from Statistics Canada provides the following information for each Island of Montreal's FSAs.

- Housing:
  - Average number of rooms per dwelling
  - Median value of dwellings ($)

  - Number of private households by tenure (owner, renter or band housing)
  - Number of private dwellings by period of construction

- Family
  - Number of occupied private dwellings by structural type of dwelling
  - Number of couple census families in private households by number of children
  - Number of lone-parent census families in private households by number of children
  - Number of non-census-family households by number of persons

- Income
  - Median total income in 2015 among recipients ($)

- Population
  - Population in 2016
  - Median age of the population

## 2.2 Data cleaning

### 2.2.1 Correspondence between the FSAs of 2016 and 2020

The map of the FSAs has evolved between 2016 and 2020, the boundaries of some FSAs have shifted and new FSAs have been created. As we are using data from 2016 and 2020, we need to create FSA groups to have similar territories for all datasets. The considered groups are presented in Table 1.

| **FSA 2016** | **FSA 2020** | **FSA groups** |
|---|---|---|
| H3A | H3A | H3A-H3B |
| H3B | H3B | H3A-H3B |
| H8S | H8S | H8S-H8T |
| H8T | H8T | H8S-H8T |
| H9J | H9K | H9J-H9K |
| H9K | H9K | H9J-H9K |
| H4R | H4R | H4R-H4S-H4T |
| H4S | H4S | H4R-H4S-H4T |
| - | H4T (new) | H4R-H4S-H4T |
| - | H4Y (new) | H9P |
| - | H5B (new) | H2Z |
| - | H5A (new) | H3A-H3B |
| - | H4Z (new) | H3A-H4B |

Table 1: Modified FSAs with their corresponding FSA groups

### 2.2.2 Foursquare dataset

First, the duplicated venues are dropped using the venue id which is a unique identifier. Then for the need of this project, the venues must be gathered in categories. Foursquare already provides a category for the venues, but it is too specific. That is why we create the following general categories and link them to the Foursquare categories.

- Restaurant and bar
- Entertainment and culture
- Sports
- Outdoor
- Other, for all the other categories we don't want or need in this project

Finally a dataset containing the following information is built.

- *FSA*: the corresponding FSA or FSAs group
- *Category*: the general category among those developed previously
- *NbVenue*: the number of venues of this category in the corresponding FSA

### 2.2.3 Metro stations dataset

There are no duplicates in this dataset because it is generated from the map in QGIS software in which only one point per metro station is selected. The features of this dataset are the following.

- *FSA*: the corresponding FSA or FSAs group
- *MetroStations*: the number of metro stations in the corresponding FSA

### 2.2.4 Daycares dataset

There are no duplicates in this dataset and the features of the dataset are the following.

- *FSA*: the corresponding FSA or FSAs group
- *Daycare*: the number of daycares in the corresponding FSA

### 2.2.5 Schools dataset

As described in the data sources section, several datasets for the educational institutions are used and are concatenated. There are no duplicates in this dataset but some institutions gather several levels of education that need to be broken down. So the rows for the institutions with multiple levels are duplicated and each one is associated with the corresponding educational level. For this project, the educational levels considered are the following.

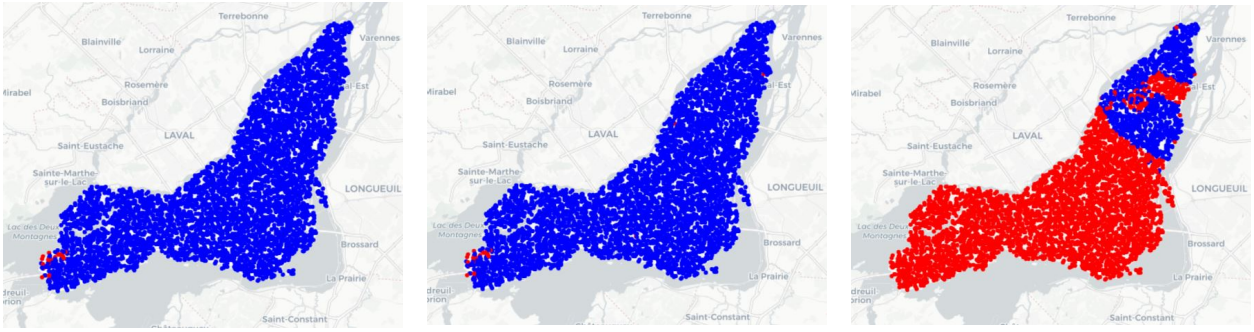
- Preschool, primary and secondary education grouped in PPS
- College and university grouped in PostSecondary
- Adult education and professional training grouped in ProfessionalTraining-AdultEducation

Finally, a dataset containing the following information is built.

- *FSA*: the corresponding FSA or FSAs group
- *SchoolType*: the corresponding educational level
- *NbSchool*: the number of institutions of this school type in the corresponding FSA

### 2.2.6 Walk score dataset

As described in the data sources section, the walk scores, transit scores and bike scores are obtained using the Walk Score API from randomly created points in the Island of Montreal. To ensure the score is available for enough points in the island, the available and unavailable scores are visualized on a map for the walk score, bike score and transit score [Figure 1].




(a) Walk score (b) Bike score (c) Transit score

Figure 1: Available [blue points] and not available [red points] scores in the dataset

The walk scores [Figure 1a] and bike scores [Figure 1b] are available for almost all the generated points, so this information can be used. However the transit scores are missing for most of the island [Figure 1c], so this feature cannot be used for this project. Finally, a dataset containing the following information is built.

- *FSA*: the corresponding FSA or FSAs group
- *Walkscore*: the average walk score of the corresponding FSA
- *Bikescore*: the average bike score of the corresponding FSA
- *Latitude*: the latitude of the corresponding FSA (for future maps)
- *Longitude*: the longitude of the corresponding FSA (for future maps)

#### 2.2.7 Census dataset

The Census dataset does not need any cleaning work and can be used as is.

## 2.3 Data preprocessing

All the data can be grouped in three master datasets:

- Environment dataset: grouping the appeals and amenities
- Population dataset: grouping the population information
- Housing dataset: grouping the information about the housing

To compare the FSAs, the densities must be considered instead of raw numbers. Therefore, the areas of the Island of Montreal's FSAs, available on the Montreal postal codes Cybo webpage [9], are used.

#### 2.3.1 Environment dataset

The environment dataset, with 93 observations, contains the information described in Table 2.

| Feature name | Description |
|---|---|
| FSA | The corresponding FSA or FSAs group of the Island of Montreal |
| Walkscore | The average walk score of the corresponding FSA |
| Bikescore | The average bike score of the corresponding FSA |
| RestaurantBar | The density (per $km^2$) of restaurants and bars in the corresponding FSA |
| EntertainmentCulture | The density (per $km^2$) of entertainment and culture places in the corresponding FSA |
| Sports | The density (per $km^2$) of sports venues in the corresponding FSA |
| Outdoor | The density (per $km^2$) of outdoor appeals in the corresponding FSA |
| MetroStations | The density (per $km^2$) of metro stations in the corresponding FSA |
| Daycare | The density (per $km^2$) of daycares in the corresponding FSA |
| PPS | The density (per $km^2$) of PPS (preschool, primary and secondary) schools in the corresponding FSA |
| PostSecondary | The density (per $km^2$) of post secondary institutions in the corresponding FSA |

Table 2: Name and description of the features of the environment dataset

### 2.3.2 Population dataset

The population dataset, with 93 observations, contains the information described in Table 3.

| Feature name | Description |
|---|---|
| FSA | The corresponding FSA or FSAs group of the Island of Montreal |
| Population | The density (per $km^2$) of population in the corresponding FSA |
| Age | The median age of the population in the corresponding FSA |
| Income | The median income ($) of the population in the corresponding FSA |
| CouplesNoChildren(%) | The percentage of childless couples among the different family types in the corresponding FSA |
| FamiliesWithChildren(%) | The percentage of families with children (couples or lone parents) among the different family types in the corresponding FSA |
| OnePerson(%) | The percentage of people living alone among the different family types in the corresponding FSA |
| SeveralNoFamily(%) | The percentage of people living with other non-family members among the different family types in the corresponding FSA |

Table 3: Name and description of the features of the population dataset

### 2.3.3 Housing dataset

The population dataset, with 93 observations, contains the information described in Table 4.

| **Feature name** | **Description** |
|---|---|
| FSA | The corresponding FSA or FSAs group of the Island of Montreal |
| NbDwellings | The density (per $km^2$) of dwellings in the corresponding FSA |
| Value($) | The median value ($) of the dwellings in the corresponding FSA |
| NbRooms | The average number of rooms per dwellings in the corresponding FSA |
| Rental(%) | The rental percentage of the dwellings in the corresponding FSA |
| DetachedHouse(%) | The percentage of detached houses among the different dwelling types in the corresponding FSA |
| AttachedHouse(%) | The percentage of attached houses (semi-detached houses, row houses and other single-attached houses) among the different dwelling types in the corresponding FSA |
| Apartment(%) | The percentage of apartments (in small or large buildings or in duplex) among the different dwelling types in the corresponding FSA |
| 1960OrBefore(%) | The percentage of dwellings constructed in 1960 or before among all the dwellings in the corresponding FSA |
| 1961-1980(%) | The percentage of dwellings constructed between 1961 and 1980 among all the dwellings in the corresponding FSA |
| 1981-1990(%) | The percentage of dwellings constructed between 1981 and 1990 among all the dwellings in the corresponding FSA |
| 1991-2000(%) | The percentage of dwellings constructed between 1991 and 2000 among all the dwellings in the corresponding FSA |
| 2001-2005(%) | The percentage of dwellings constructed between 2001 and 2005 among all the dwellings in the corresponding FSA |
| 2006-2010(%) | The percentage of dwellings constructed between 2006 and 2010 among all the dwellings in the corresponding FSA |
| 2011-2016(%) | The percentage of dwellings constructed between 2011 and 2016 among all the dwellings in the corresponding FSA |

Table 4: Name and description of the features of the housing dataset

# References


[1] A. C. Melissinos and J. Napolitano, *Experiments in Modern Physics*, (Academic Press, New York, 2003).

[2] N. Cyr, M. Têtu, and M. Breton, IEEE Trans. Instrum. Meas. **42**, 640 (1993).

[3] Foursquare API, available: https://developer.foursquare.com/docs/places-api/

[4] Portail données ouvertes Montréal, available: http://donnees.ville.montreal.qc.ca/dataset

[5] Partenariat Données Québec, available: https://www.donneesquebec.ca/

[6] Walk Score API, available: https://www.walkscore.com/professional/api.php

[7] 2016 Census - Boundary files, available: https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm

[8] 2016 Census Profile Web Data Service (WDS) - User Guide, available: https://www12.statcan.gc.ca/wds-sdw/cpr2016-eng.cfm

[9] Cybo - Montreal Postal codes, available: https://postal-codes.cybo.com/canada/montreal/listcodes