Clustering of the Island of Montreal

Noémie Rohel

February 19, 2021

Abstract

This paper presents the clustering of the Island of Montreal at a Forward Sortation Area (FSA) level based on the environment (density of venues), the population and the housing. Data available through the Foursquare API, the open data portals of the City of Montreal and the Quebec government, the Walk Score API and the 2016 census data from Statistics Canada are used. Three clustering approaches are considered and implemented: K-Means, DBSCAN and agglomerative clustering. The optimal results for each algorithm are obtained using an appropriate parameters selection. After analysis, the K-Means model is identified as the best for this project. The resulting clusters are described and analyzed. We conclude by discussing the choice of a clustering model and the possible improvements for this project.

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 analyzed.

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 [1]

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 [2]

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 [3]

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 [4]

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 [5]. 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 [6]

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.

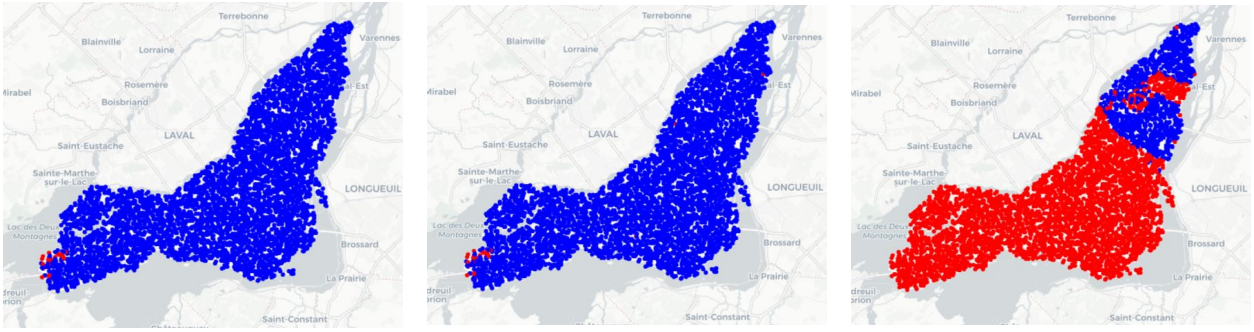
- Preschool, primary and secondary education grouped in PPS
- College and university grouped in PostSecondary
- Adult education and professional training grouped in ProfessionalTraining-AdultEducation

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 is 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 [7], 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

3 Exploratory data analysis

3.1 Environment dataset

The Figure 2 shows a choropleth map of the walk scores in Montreal. They are grouped within the following categories defined on the Walk Score website.

- **0-25 - Car-dependent:** Almost all errands require a car.
- **25-50 - Car-dependent:** Most errands require a car.
- **50-70 - Somewhat walkable:** Some errands can be accomplished on foot.
- **70-90 - Very walkable:** Most errands can be accomplished on foot.
- **90-100 - Walker's paradise:** Daily errands do not require a car.

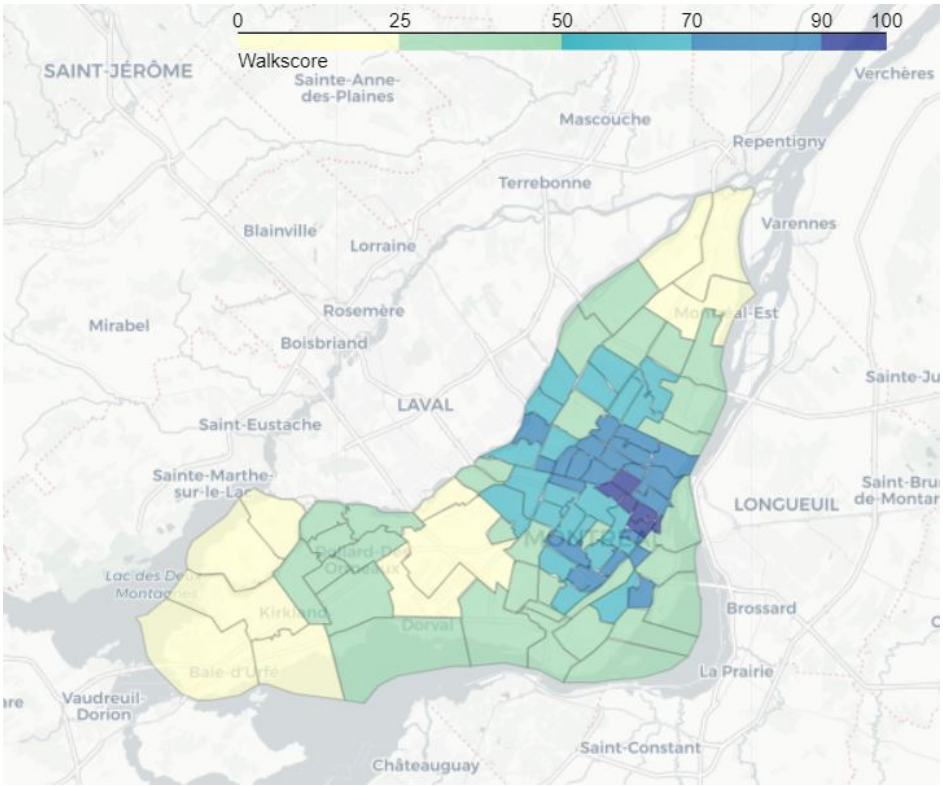


Figure 2: Choropleth map of the walk scores in the Island of Montreal

The highest walk scores are concentrated in the center of the Island and decreases rapidly with distance.

The Figure 3 shows a pareto chart of the density of restaurants and bars. The distribution is not uniform. Three FSAs (H3A-H3B, H2Z and H2Y) have a much higher density than

the others and represent 26% of the total (for 3% of the FSAs). They are located in Montreal Downtown. H3A-H3B reaches a density of 104 restaurants and bars per km^2 while the median is 5 and the mean is 9.

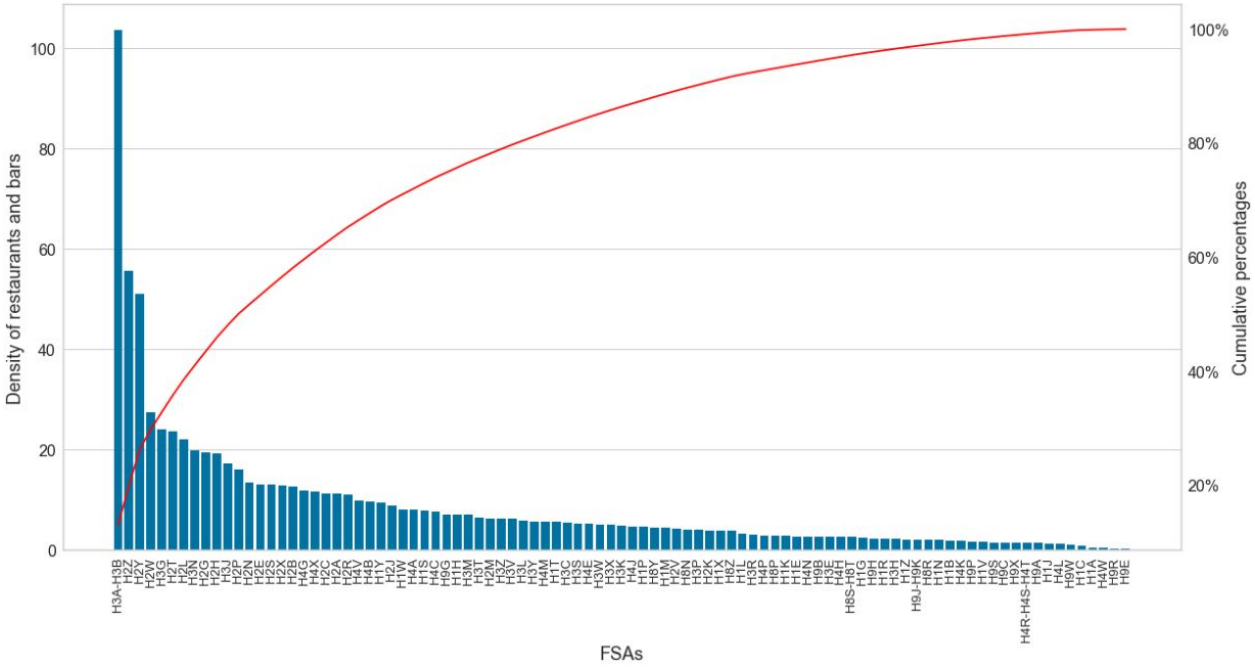


Figure 3: Pareto chart of the density of restaurants and bars in the Island of Montreal

Finally, the Figure 4 shows a histogram of the density of Preschools, Primary and Secondary schools (PPS) in the Island of Montreal. The majority of the FSAs has a density between 0 and 4 PPS per km^2, while there are few ones with a density between 13 and 15.

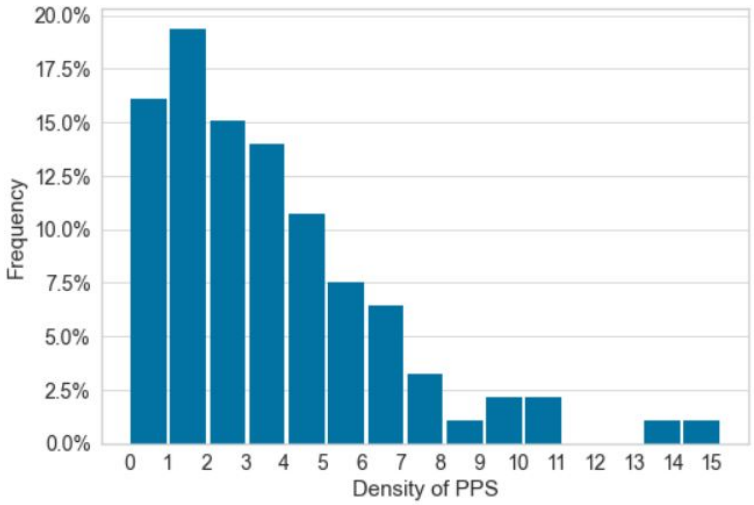


Figure 4: Histogram of the density of Preschools, Primary and Secondary schools (PPS) in the Island of Montreal

3.2 Population dataset

The Figure 5 shows a pareto chart of the density of population in the Island of Montreal. The distribution is not uniform with 18% of the FSAs (the top 17) that represents 35.6% of the population. The H3N FSA has a much higher density of population than the other FSAs, with a value of 16 973 people per km^2, while the median value is 5 614 and the mean value 5 974. On the contrary, some FSAs have a density around 300 people per km^2.

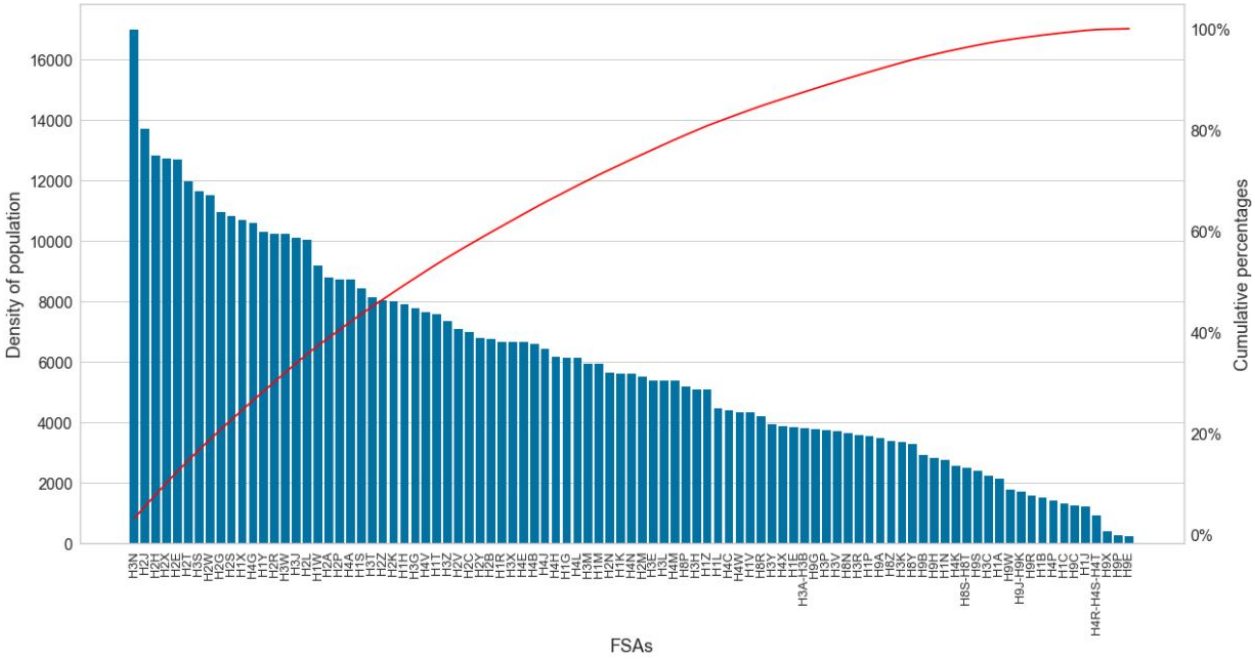


Figure 5: Pareto chart of the density of population in the Island of Montreal

The Figure 6 shows a histogram of the median age of the population in the Island of Montreal. The majority of the FSAs (around 40%) has a median age between 41 and 45 years of age. There are around 3% of the FSAs that have a quite young population, with a median age between 28 and 30 years of age.

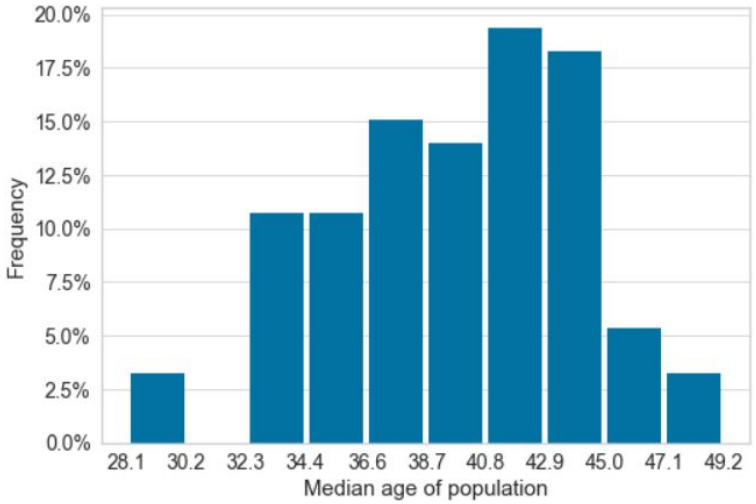


Figure 6: Histogram of the median age of population in the Island of Montreal

3.3 Housing dataset

The Figure 7 shows a pareto chart of the average value of the properties in the Island of Montreal. Again, the distribution is not uniform. The H3N FSA alone represents 3.7% of the total value of the properties, with an average value of 1 500 249 $, which is much higher than the other FSAs. The median average value is 400 532 $ and the mean average value is 436 497 $. 4% of the FSAs (the top 4) represent 10% of the total value.

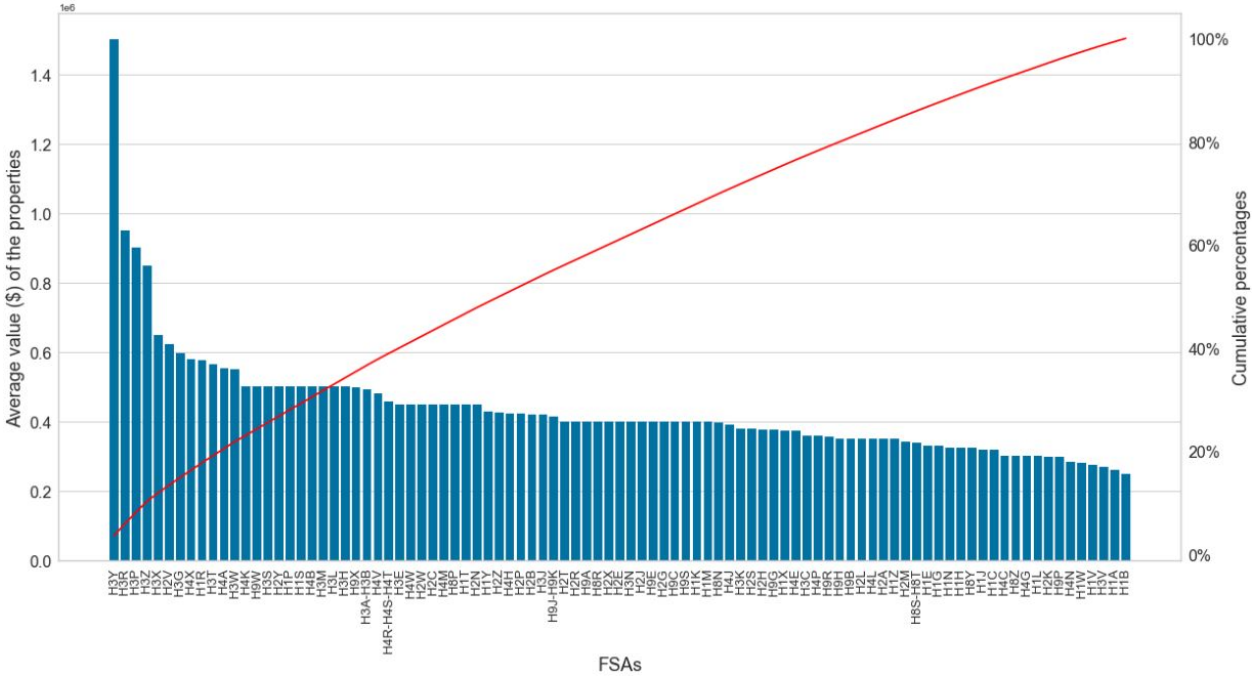


Figure 7: Pareto chart of the average value ($) of the properties in the Island of Montreal

Finally, the Figure 8 shows a histogram of the average number of rooms in the dwellings of the Island of Montreal. This value is between 3 and 8.4 across all the FSAs and the majority of the FSAs' dwellings (more than 35%) has between 4.1 and 4.6 rooms on average.

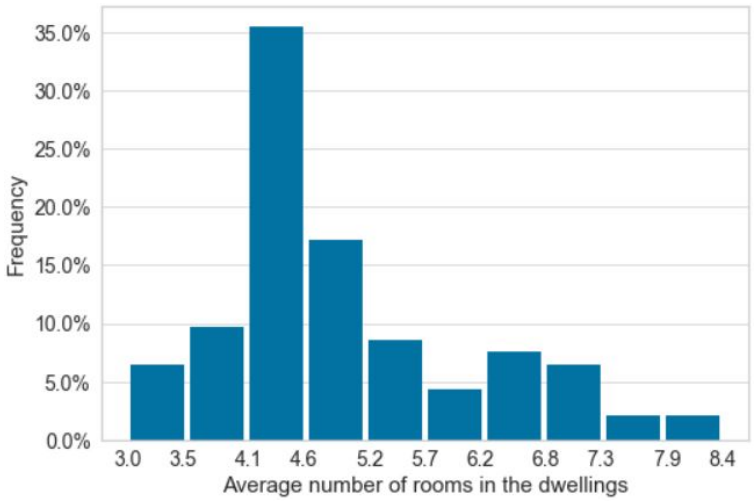


Figure 8: Histogram of the average number of rooms in the dwellings of the Island of Montreal

4 Clustering models

Three different clustering models (K-Means, DBSCAN and agglomerative clustering) are used and compared for this project. For each of them, a parameters selection is performed. Before applying any clustering model, a data standardization is required. It scales the data with zero mean and unit variance.

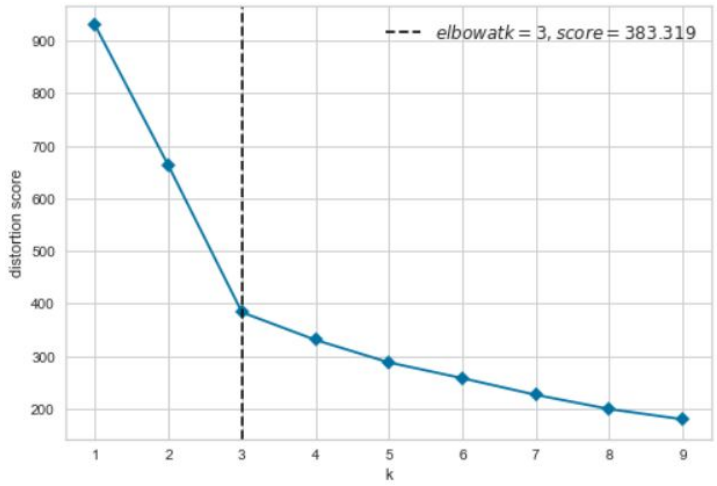
4.1 K-Means

K-Means model is an iterative algorithm that divides the data into k non-overlapping subsets (clusters) without any cluster-internal structure. It identifies k number of centroids (arithmetic mean of all the data points that belong to a cluster) and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. Data points within a cluster are very similar while also being very different across different clusters.

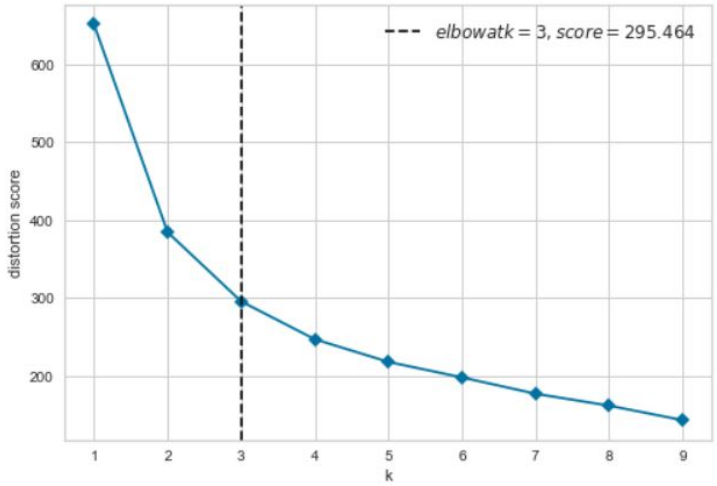
The number of clusters k has to be given as a parameter for this model. A common way to choose the optimal number of clusters is to use the elbow method. It consists in running the clustering across different values of k and look at the graph of distortion versus k. The right value for the number of clusters is for the elbow point (where the rate of decrease sharply shifts). The KElbowVisualizer from Yellowbrick, which implements the elbow method, is used and gives the results shown in Figure 9 for the three datasets. Finally, using the number of clusters identified with the elbow method, the K-Means approach gives the clusters described in Table 5.

Dataset	**Environment**	**Population**	**Housing**
Number of clusters	3	3	4
Number of FSAs per cluster	51	36	62
	40	30	20
	2	27	7
	-	-	4

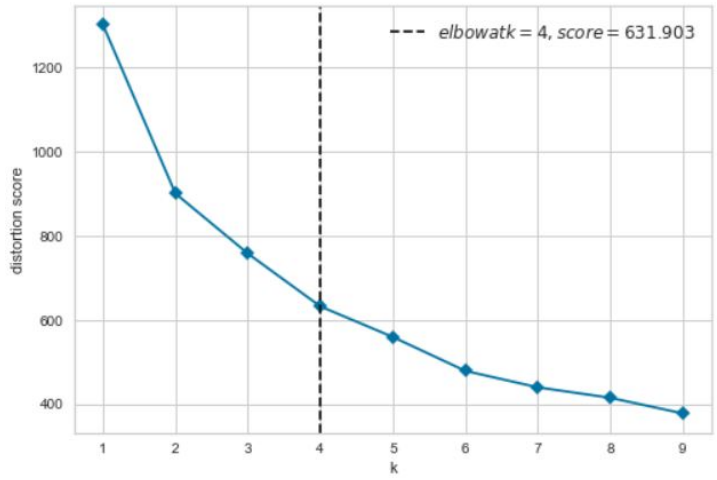
Table 5: Description of the clusters from the K-Means approach for the three datasets



(a) Environment dataset



(b) Population dataset



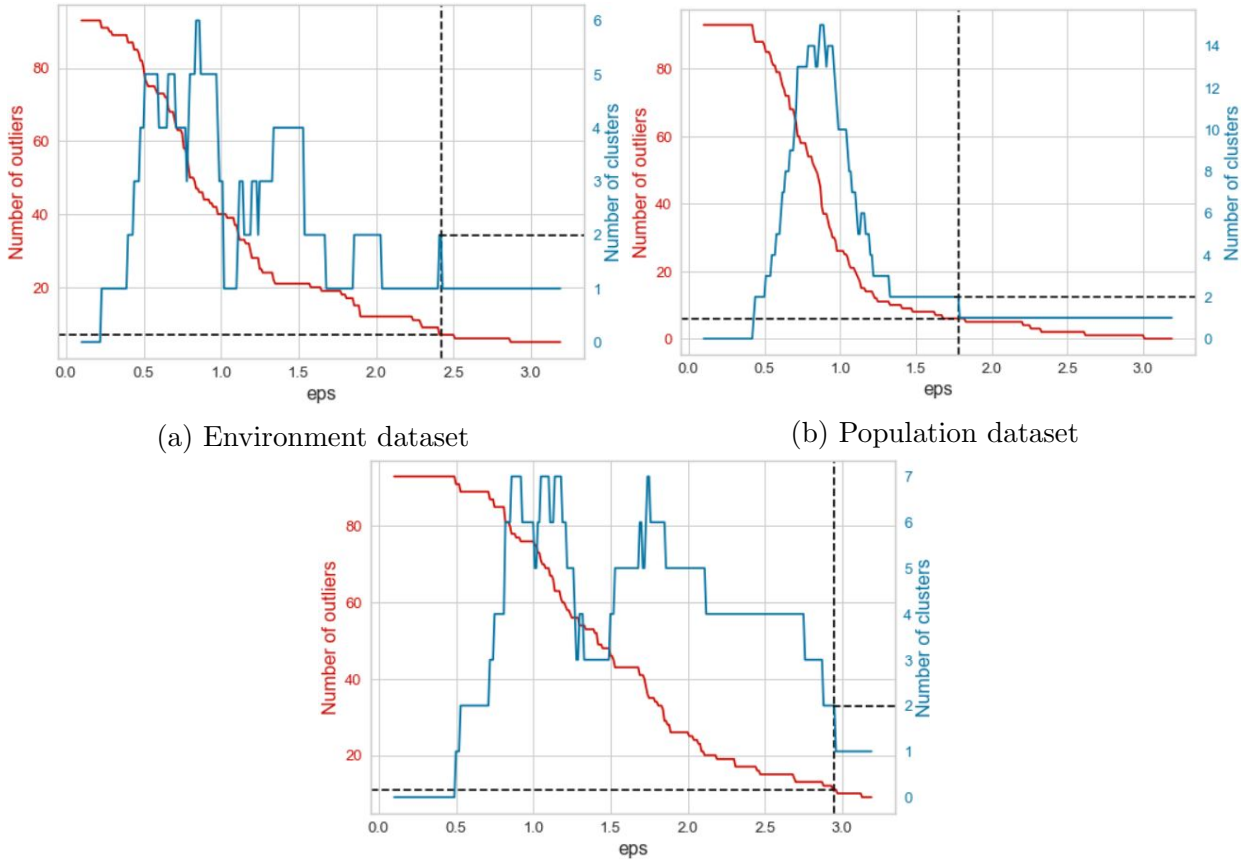
(c) Housing dataset

Figure 9: Distortion score elbow for K-Means clustering

4.2 DBSCAN

DBSCAN model is a density-based spatial clustering of applications with noise. It locates regions of high density, separated from one another by regions of low density, and separates outliers. It gives arbitrary-shaped clusters and not spherical ones, so there can be clusters within clusters.

The maximum distance between two samples for one to be considered as in the neighborhood of the other is an input of this algorithm. It is represented as the radius of the neighborhood and if there are enough points within this distance, it is a dense area. Another input of this algorithm is the number of samples in a neighborhood for a point to be considered as a core point. This is the minimum number of neighbors to define a cluster. This second parameter is set to 2 because having a cluster with two FSAs is not a problem for this project. On the other hand, having outliers should be avoided because all FSAs must belong to a cluster. So the first parameter (eps) is chosen as the value that gives the minimum number of outliers and more than one cluster, as shown in Figure 10. Finally, the DBSCAN approach gives the clusters described in Table 6.



(a) Environment dataset

(b) Population dataset

(c) Housing dataset

Figure 10: Number of outliers and clusters for different values of eps for DBSCAN clustering

Dataset	Environment	Population	Housing
Number of clusters	2	2	2
Number of FSAs per cluster	84	84	80
	2	3	2
Number of outliers	7	6	11

Table 6: Description of the clusters from the DBSCAN approach for the three datasets

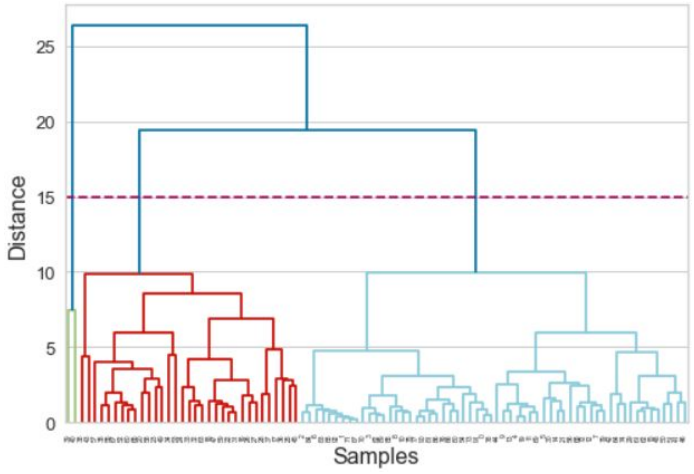
4.3 Agglomerative clustering

Agglomerative clustering is one of the two types of hierarchical clustering. Hierarchical clustering algorithms build a hierarchy of clusters where each node is a cluster, constituted of the clusters of its daughter nodes. Divisive clustering has a top-down strategy because it starts from a large cluster with every data points to smaller pieces of one data point. Agglomerative clustering, on the contrary, has a bottom-up strategy. Its starts with each observation in its own cluster to merge them into larger clusters and finally get a cluster containing all observations. The agglomerative clustering constructs a distance matrix and merge the two closest clusters at each step.

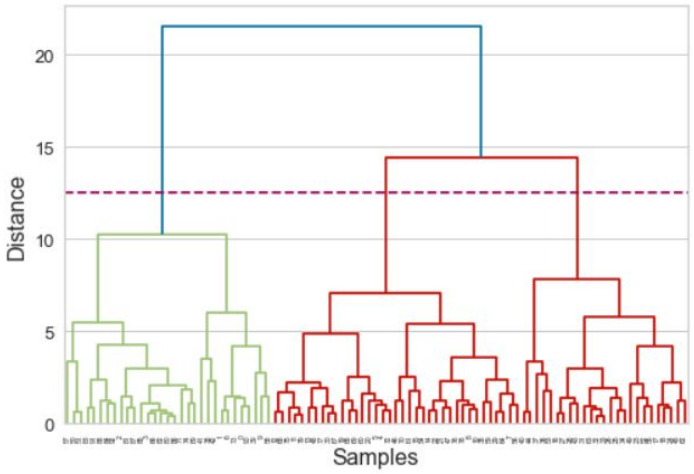
A good visualization of hierarchical clustering is dendrograms. It shows each merge as a horizontal line and its y-coordinate shows the similarity of the two clusters that were merged. The larger the vertical lines in the dendrogram, the further away the clusters are from each other. To get the optimal number of clusters, we generally set the threshold in such a way that it cuts the tallest vertical line. This is the method we apply here for the three datasets [Figure 11]. Finally, using the number of clusters identified with the dendrograms, the agglomerative clustering approach gives the clusters described in Table 7.

Dataset	Environment	Population	Housing
Number of clusters	3	3	2
Number of FSAs per cluster	58	37	59
	33	31	34
	2	25	-

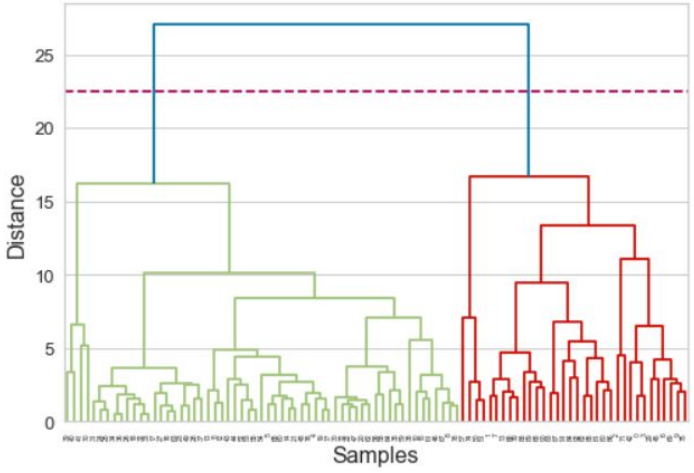
Table 7: Description of the clusters from the agglomerative clustering approach for the three datasets



(a) Environment dataset



(b) Population dataset



(c) Housing dataset

Figure 11: Dendrograms with the identification of the optimal number of clusters for agglomerative clustering

4.4 Comparison between the models

To evaluate and compare the models, metrics like the silhouette score or the Calinski-Harabaz index can be used. The silhouette score, or silhouette coefficient, measures how well samples are clustered with other samples that are similar to each other. It is calculated using the mean intra-cluster distance, i.e the average distance between each point within a cluster, and the mean nearest-cluster distance, i.e the average distance between all clusters. The silhouette score varies between -1 and 1. A score of 1 means the cluster is dense and well-separated than other clusters; a score of 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters; a value of -1 means the clusters are assigned in the wrong way. The Calinski-Harabaz index, also known as the variance ratio criterion, is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. The score is not bounded. The higher the score is, the better the clustering is. The Table 8 shows the scores for the different datasets and models.

Score	K-Means	DBSCAN	Agglomerative clustering
Silhouette score	0.333	0.338	0.343
Calinski-Harabaz index	64.2	18.5	61.8

(a) Environment dataset

Score	K-Means	DBSCAN	Agglomerative clustering
Silhouette score	0.282	0.262	0.252
Calinski-Harabaz index	54.0	9.1	48.2

(b) Population dataset

Score	K-Means	DBSCAN	Agglomerative clustering
Silhouette score	0.395	0.242	0.315
Calinski-Harabaz index	31.5	7.6	35.7

(c) Housing dataset

Table 8: Silhouette score and Calinski-Harabaz index for the three clustering models and the different datasets

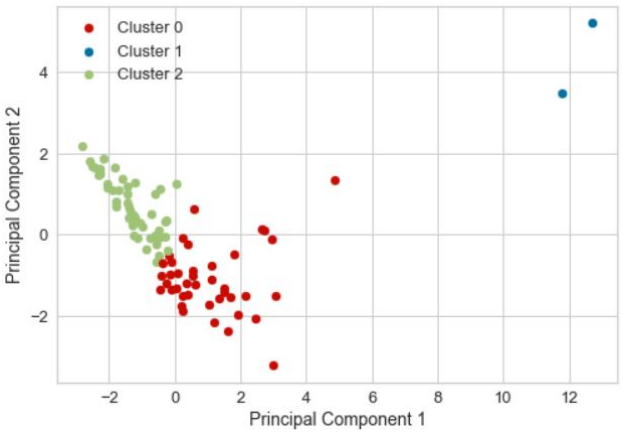
The metrics show, especially with the Calinski-Harabaz index, that the DBSCAN algorithm seems to give poorer results than the two other algorithms, for all the three datasets. This is in line with the observations made at the time of the parameters selection. Indeed, the DBSCAN model identifies outliers, whereas the problem here requires all the FSAs to be clustered to have a portrait of the Island of Montreal as a whole. The DBSCAN algorithm is not suitable for this problem, therefore only the K-Means and the agglomerative clustering are considered in the following sections.

The silhouette scores and Calinski-Harabaz indexes for K-Means and agglomerative clustering are quite similar. For the environment dataset, the silhouette score is the best for the agglomerative clustering, while the Calinski-Harabaz index is the best for K-Means. For the population dataset, both the silhouette score and Calinski-Harabaz index are the best for K-Means. Finally, for the housing dataset, the silhouette score is the best for K-Means, while the Calinski-Harabaz index is the best for the agglomerative clustering. Both scores do not really help to determine the best algorithm in our context, except for the population dataset where the K-Means approach seems to be more efficient.

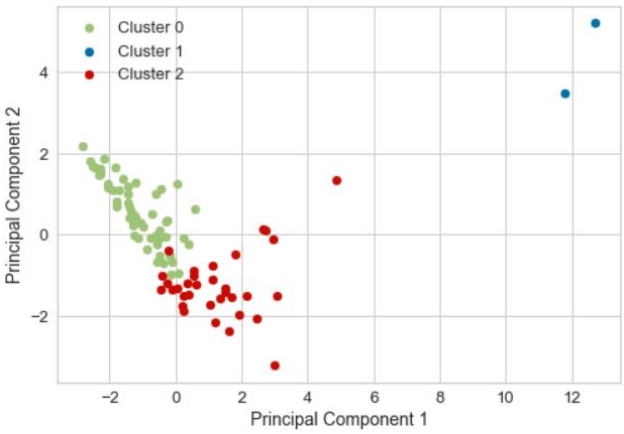
To have a better understanding of the performances of the two approaches, the results of the clustering are visualized. To do so, a linear dimensionality reduction is performed using Singular Value Decomposition of the data. It provides a projection in a lower dimensional space and allows to visualize the results in two dimensions. The results for the three datasets and the two models are presented in the Figure 12.

The two dimensions projection for the environment dataset highlights that the points from the two biggest clusters obtained with the agglomerative clustering are quite mixed, whereas the clusters obtained with the K-Means method are fairly distinct. For the population dataset, the same observation can be made. As the number of clusters built by the two approaches is different for the housing dataset, it is difficult to compare the resulting clusters. However, according to the two dimensions projection they seem well separated in both cases.

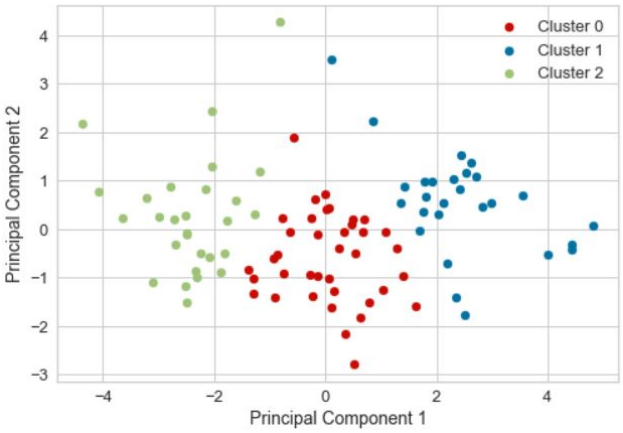
In conclusion, in the context of this project the K-Means model and the agglomerative clustering provide good results, whereas the DBSCAN approach is not able to build suitable clusters. In addition, as the clusters obtained with the K-Means model are well separated, this method is preferred to answer to our problematic.



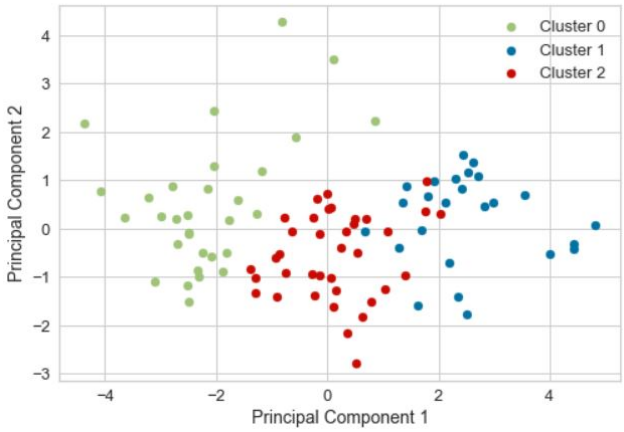
(a) Environment - K-Means
Information carried: 73.23%



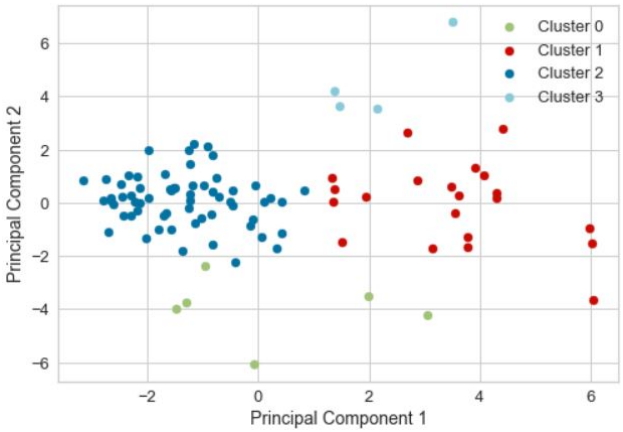
(b) Environment - Agglomerative clustering
Information carried: 73.23%



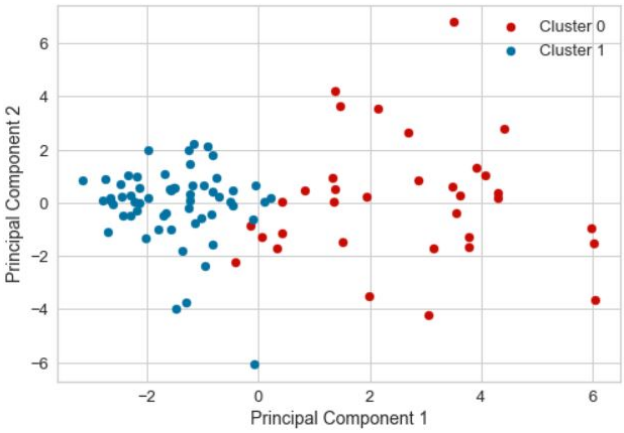
(c) Population - K-Means
Information carried: 79.88%



(d) Population - Agglomerative clustering
Information carried: 79.88%



(e) Housing - K-Means
Information carried: 61.61%



(f) Housing - Agglomerative clustering
Information carried: 61.61%

Figure 12: Two dimensions projection showing the clustering results for the three datasets and the two models (K-Means and agglomerative clustering), with the percentage of information carried by the dimension-reduced data

5 Clustering results

In this section, the clusters obtained with the K-Means model are analyzed in details in order to identify the specifics of each cluster for the three datasets.

5.1 Environment dataset

The Table 9 shows the main descriptive statistics of the environment dataset and its three clusters.

The descriptive statistics confirm that the three clusters have different profiles. The cluster 1 has really high walk score and bike score and a very large density of each venue, except for PPS schools that is lower than cluster 0. The cluster 0 has medium walk score and bike score and a density of venues always higher than cluster 2 but much lower than cluster 1. According to these observations, the clusters can be described as follows.

- Cluster 0: Medium environment
- Cluster 1: Very rich environment
- Cluster 2: Poor environment

The Figure 13 shows the clustered FSAs of the Island of Montreal based on the environment dataset.

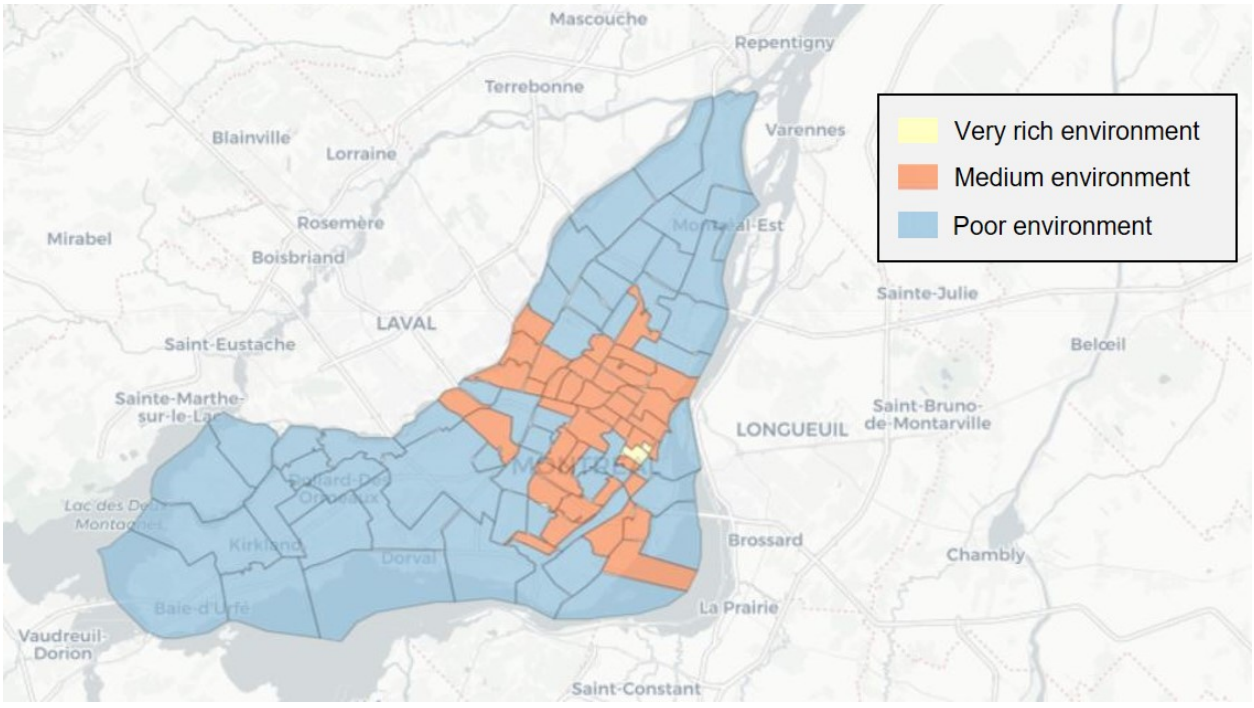


Figure 13: Map of the clustered FSAs of the Island of Montreal based on the environment dataset

Feature	Dataset	Mean	Median	Standard deviation	Minimum value	Maximum value
Walk score	Overall *(n=93)*	54	54	24	1	97
	Cluster 0 *(n=40)*	73	73	13	46	95
	Cluster 1 *(n=2)*	96	96	1.5	95	97
	Cluster 2 *(n=51)*	37	37	16	1	69
Bike score	Overall *(n=93)*	70	68	15	39	99
	Cluster 0 *(n=40)*	82	85	10	62	99
	Cluster 1 *(n=2)*	85	85	5	81	88
	Cluster 2 *(n=51)*	60	59	11	39	85
Restaurants & bars	Overall *(n=93)*	8.6	4.7	13.4	0.2	103.5
	Cluster 0 *(n=40)*	11.9	10.3	9.1	1.2	51.0
	Cluster 1 *(n=2)*	79.5	79.5	33.9	55.6	103.5
	Cluster 2 *(n=51)*	3.2	2.6	2.2	0.2	9.8
Entertainment Culture	Overall *(n=93)*	3.3	1.3	7.2	0.0	39.2
	Cluster 0 *(n=40)*	4.6	1.8	7.3	0.0	39.2
	Cluster 1 *(n=2)*	37.4	37.4	1.1	36.6	38.2
	Cluster 2 *(n=51)*	1.0	0.7	0.8	0.0	3.3
Sports	Overall *(n=93)*	1.7	1.0	2.7	0.0	20.8
	Cluster 0 *(n=40)*	1.6	1.2	1.5	0.0	5.0
	Cluster 1 *(n=2)*	17.3	17.3	5.0	13.7	20.8
	Cluster 2 *(n=51)*	1.1	0.8	1.0	0.0	5.5
Outdoor	Overall *(n=93)*	1.4	1.0	1.4	0.0	6.9
	Cluster 0 *(n=40)*	1.7	1.1	1.4	0.0	5.2
	Cluster 1 *(n=2)*	5.8	5.8	1.7	4.6	6.9
	Cluster 2 *(n=51)*	1.0	0.8	0.9	0.0	3.7
Metro stations	Overall *(n=93)*	0.3	0.0	0.9	0.0	6.9
	Cluster 0 *(n=40)*	0.5	0.4	0.6	0.0	1.9
	Cluster 1 *(n=2)*	4.8	4.8	3.0	2.7	6.9
	Cluster 2 *(n=51)*	0.1	0.0	0.1	0.0	0.5
Daycares	Overall *(n=93)*	4.1	3.4	3.2	0.0	13.9
	Cluster 0 *(n=40)*	6.3	5.4	2.4	2.8	11.0
	Cluster 1 *(n=2)*	13.8	13.8	0.1	13.7	13.9
	Cluster 2 *(n=51)*	1.9	1.5	1.5	0.0	7.2
PPS schools	Overall *(n=93)*	3.6	3.0	3.0	0.0	15.2
	Cluster 0 *(n=40)*	5.6	5.2	3.2	0.0	15.2
	Cluster 1 *(n=2)*	2.7	2.7	3.9	0.0	5.5
	Cluster 2 *(n=51)*	2.2	2.0	1.8	0.0	10.9
Post Secondary schools	Overall *(n=93)*	0.7	0.0	2.0	0.0	14.7
	Cluster 0 *(n=40)*	0.9	0.3	1.3	0.0	4.5
	Cluster 1 *(n=2)*	12.5	12.5	3.0	10.4	14.7
	Cluster 2 *(n=51)*	0.2	0.0	0.5	0.0	2.9

Table 9: Descriptive statistics of the environment dataset and its three clusters

5.2 Population dataset

The Table 10 shows the main descriptive statistics of the population dataset and its three clusters.

Feature	Dataset	Mean	Median	Standard deviation	Minimum value	Maximum value
Population density	Overall *(n=93)*	5 974	5 614	3 526	231	16 973
	Cluster 0 *(n=38)*	5 986	6 139	2 010	1 402	10 236
	Cluster 1 *(n=27)*	9 300	10 240	3 549	2 242	16 973
	Cluster 2 *(n=28)*	2 750	2 686	1 689	231	6 775
Median age	Overall *(n=93)*	39.8	39.9	4.5	28.1	49.2
	Cluster 0 *(n=38)*	40.1	40.0	3.0	34.0	48.1
	Cluster 1 *(n=27)*	35.1	34.8	2.9	28.1	42.3
	Cluster 2 *(n=28)*	44.0	44.5	2.5	37.1	49.2
Median income ($)	Overall *(n=93)*	31 445	30 566	7 935	16 811	66 500
	Cluster 0 *(n=38)*	28 643	27 850	4 570	21 258	40 843
	Cluster 1 *(n=27)*	28 537	28 670	7 127	16 811	51 548
	Cluster 2 *(n=28)*	38 051	36 489	8 499	28 969	66 500
Percentage of childless couples	Overall *(n=93)*	20.9	21.0	3.4	14.0	28.8
	Cluster 0 *(n=38)*	19.3	19.9	2.2	14.0	23.2
	Cluster 1 *(n=27)*	19.4	19.2	3.0	15.0	26.9
	Cluster 2 *(n=28)*	24.6	23.9	2.2	21.0	28.8
Percentage of families with children	Overall *(n=93)*	35.9	36.4	13.2	10.3	62.3
	Cluster 0 *(n=38)*	38.8	38.6	7.1	25.8	52.8
	Cluster 1 *(n=27)*	20.9	20.1	7.2	10.3	39.0
	Cluster 2 *(n=28)*	46.5	49.7	10.9	12.0	62.3
Percentage of people living alone	Overall *(n=93)*	37.7	38.2	11.6	14.7	62.7
	Cluster 0 *(n=38)*	37.2	36.9	6.2	23.6	46.1
	Cluster 1 *(n=27)*	50.0	49.9	6.0	37.8	62.7
	Cluster 2 *(n=28)*	26.6	24.7	9.5	14.7	55.4
Percentage of people living with other non-family members	Overall *(n=93)*	5.4	4.5	3.5	1.3	15.8
	Cluster 0 *(n=38)*	4.7	4.6	1.5	2.4	9.1
	Cluster 1 *(n=27)*	9.7	9.0	3.0	4.5	15.8
	Cluster 2 *(n=28)*	2.3	2.0	0.9	1.3	4.5

Table 10: Descriptive statistics of the population dataset and its three clusters

The descriptive statistics confirm that the three clusters have different profiles. The cluster 1 has a high density of population, the cluster 0 a medium one and the cluster 2 a low one. The cluster 1 has also the youngest population, while the cluster 2 has the oldest. About the median income, the clusters 0 and 1 are quite similar, but the cluster 2 has a median income much higher than the two others. The cluster 1 has a majority of people living alone, about half of the population, and more people living with other non-family members than

the other clusters. In addition, it has less families with children. The cluster 2 has a majority of families with children, almost half of the population, and the most of childless couples. It has the least of people living alone or with non-family members. Finally, the cluster 0 has almost as many people living alone as families with children, which are the majority of the population. It has less couples without children. According to these observations, the clusters can be described as follows.

- Cluster 0: Medium density, average age, medium income, mixed types of families
- Cluster 1: High density, young population, medium income, majority (half) of people living alone and the most people living with non-family members
- Cluster 2: Low density, old population, high income, majority (almost half) of families with children and the most childless couples

The Figure 14 shows the clustered FSAs of the Island of Montreal based on the population dataset.

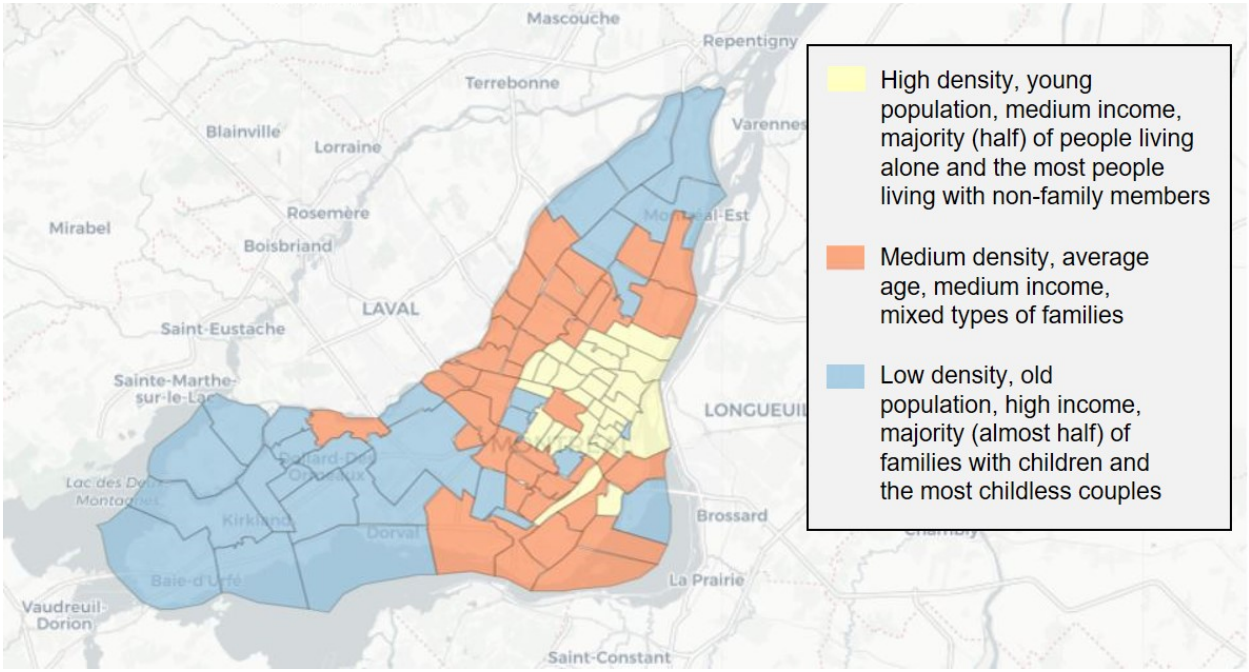


Figure 14: Map of the clustered FSAs of the Island of Montreal based on the population dataset

5.3 Housing dataset

The Table 5.3 shows the main descriptive statistics of the housing dataset and its four clusters.

Feature	Dataset	Mean	Median	Standard deviation	Minimum value	Maximum value
Density of dwellings	Overall *(n=93)*	2 834	2 443	1 899	90	7677
	Cluster 0 *(n=6)*	2 016	1 773	1 695	361	5 000
	Cluster 1 *(n=21)*	892	910	522	90	1 950
	Cluster 2 *(n=62)*	3 665	3 258	1 706	675	7 677
	Cluster 3 *(n=4)*	1 378	1 353	73	1 325	1 483
Median value ($) of dwellings	Overall *(n=93)*	436 497	400 532	165 348	250 531	1 500 249
	Cluster 0 *(n=6)*	424 424	437 756	55 035	358 753	494 159
	Cluster 1 *(n=21)*	381 021	375 639	70 330	259 561	501 665
	Cluster 2 *(n=62)*	421 239	400 631	106 449	250 531	850 402
	Cluster 3 *(n=4)*	982 356	924 728	382 187	579 719	1 500 249
Average number of rooms per dwellings	Overall *(n=93)*	5.0	4.6	1.2	3.0	8.4
	Cluster 0 *(n=6)*	4.0	3.7	0.8	3.3	5.2
	Cluster 1 *(n=21)*	6.4	6.4	0.9	4.7	8.4
	Cluster 2 *(n=62)*	4.4	4.4	0.5	3.0	5.6
	Cluster 3 *(n=4)*	7.2	6.9	0.8	6.7	8.4
Rental percentage	Overall *(n=93)*	56.1	62.3	19.7	10.3	90.5
	Cluster 0 *(n=6)*	53.7	47.2	16.2	38.2	74.0
	Cluster 1 *(n=21)*	28.7	28.3	11.8	10.3	53.4
	Cluster 2 *(n=62)*	67.3	67.8	8.8	50.8	90.5
	Cluster 3 *(n=4)*	28.7	30.6	8.4	16.9	36.9
Percentage of detached houses	Overall *(n=93)*	15.5	5.0	22.4	0.0	86.3
	Cluster 0 *(n=6)*	2.7	1.2	3.5	0.0	8.5
	Cluster 1 *(n=21)*	49.2	53.9	22.8	3.6	86.3
	Cluster 2 *(n=62)*	4.2	1.9	5.1	0.0	21.6
	Cluster 3 *(n=4)*	34.0	34.6	1.8	31.5	35.4
Percentage of attached houses	Overall *(n=93)*	9.1	6.0	9.0	0.1	53.6
	Cluster 0 *(n=6)*	8.6	4.9	9.7	0.2	24.1
	Cluster 1 *(n=21)*	16.2	14.6	8.2	4.3	39.2
	Cluster 2 *(n=62)*	5.3	4.4	4.2	0.1	18.3
	Cluster 3 *(n=4)*	31.4	25.0	14.9	22.1	53.6
Percentage of apartments	Overall *(n=93)*	75.4	86.4	27.5	4.9	99.8
	Cluster 0 *(n=6)*	88.7	92.4	12.7	67.5	99.8
	Cluster 1 *(n=21)*	34.7	30.5	20.2	4.9	81.9
	Cluster 2 *(n=62)*	90.5	93.1	7.7	65.0	99.7
	Cluster 3 *(n=4)*	34.6	40.4	13.1	14.9	42.6

Table 11: Descriptive statistics of the housing dataset and its four clusters (*First part*)

Feature	Dataset	Mean	Median	Standard deviation	Minimum value	Maximum value
Percentage of dwellings constructed in 1960 or before	Overall *(n=93)*	37.9	39.9	20.7	2.6	89.0
	Cluster 0 *(n=6)*	12.3	12.0	8.1	2.6	22.2
	Cluster 1 *(n=21)*	20.9	14.2	16.8	3.7	62.5
	Cluster 2 *(n=62)*	43.7	44.8	15.8	9.9	76.4
	Cluster 3 *(n=4)*	76.8	74.1	8.4	69.9	89.0
Percentage of dwellings constructed in 1961-1980	Overall *(n=93)*	32.0	29.5	14.1	4.2	67.6
	Cluster 0 *(n=6)*	19.5	20.8	8.9	4.2	29.5
	Cluster 1 *(n=21)*	37.7	36.9	14.7	9.2	67.6
	Cluster 2 *(n=62)*	32.4	31.8	13.2	7.7	67.5
	Cluster 3 *(n=4)*	15.1	17.2	6.3	6.4	19.7
Percentage of dwellings constructed in 1981-1990	Overall *(n=93)*	11.4	8.7	7.7	1.2	41.2
	Cluster 0 *(n=6)*	11.5	9.2	8.7	3.5	27.5
	Cluster 1 *(n=21)*	19.2	16.7	9.0	7.6	41.2
	Cluster 2 *(n=62)*	9.3	8.3	5.2	2.4	31.0
	Cluster 3 *(n=4)*	3.0	3.0	1.5	1.2	4.8
Percentage of dwellings constructed in 1991-2000	Overall *(n=93)*	6.6	5.2	4.4	0.6	23.4
	Cluster 0 *(n=6)*	8.0	5.6	5.8	2.6	15.7
	Cluster 1 *(n=21)*	10.1	8.2	5.8	2.3	23.4
	Cluster 2 *(n=62)*	5.5	4.9	2.8	1.6	18.5
	Cluster 3 *(n=4)*	2.4	1.2	2.8	0.6	6.6
Percentage of dwellings constructed in 2001-2005	Overall *(n=93)*	3.5	2.6	2.9	0.3	16.1
	Cluster 0 *(n=6)*	9.2	9.7	5.1	0.3	16.1
	Cluster 1 *(n=21)*	5.1	5.0	2.8	1.6	12.0
	Cluster 2 *(n=62)*	2.5	2.4	1.5	0.4	1.9
	Cluster 3 *(n=4)*	1.3	1.1	0.7	0.6	2.3
Percentage of dwellings constructed in 2006-2010	Overall *(n=93)*	4.0	2.9	4.0	0.0	21.7
	Cluster 0 *(n=6)*	15.1	14.6	5.6	9.0	21.7
	Cluster 1 *(n=21)*	4.0	3.3	3.4	0.0	17.0
	Cluster 2 *(n=62)*	3.1	2.7	2.0	0.6	10.7
	Cluster 3 *(n=4)*	0.7	0.8	0.5	0.0	1.1
Percentage of dwellings constructed in 2011-2016	Overall *(n=93)*	4.6	2.7	6.9	0.0	44.4
	Cluster 0 *(n=6)*	24.4	17.9	15.7	9.5	44.4
	Cluster 1 *(n=21)*	3.1	2.5	3.0	0.0	12.7
	Cluster 2 *(n=62)*	3.5	2.7	2.6	0.3	11.9
	Cluster 3 *(n=4)*	0.8	0.6	0.5	0.4	1.5

Table 11: Descriptive statistics of the housing dataset and its four clusters (*Second part*)

As for the two other datasets, the descriptive statistics confirm that the clusters have different profiles. The cluster 1 has a low density of dwellings, while the cluster 2 has a high one, the cluster 0 a medium high one and the cluster 3 a medium low one. The cluster 3 has a very high median value of dwellings, the cluster 1 a low value (more than 2.5 less than the cluster 3) and the clusters 0 and 2 are quite similar, with a medium value. About the average number of rooms per dwellings, the cluster 3 has a very high number of rooms, as would be expected since it is usually correlated with the price. The cluster 1 has less rooms than the cluster 3 but still a quite high number of rooms. The clusters 0 and 2 have a lower number of rooms, with cluster 2 above cluster 0. Then, the clusters 1 and 3 have more owners than tenants, while the cluster 2 has a large majority of tenants and the cluster 0 is pretty balanced. For the type of dwellings, more than 80% of the dwellings are apartments for the clusters 0 and 2, whereas the cluster 3 is equally shared among detached houses, attached houses and apartments. The cluster 1 has almost half of dwellings that are detached houses and then a majority of apartments. Finally, for the year of construction, the cluster 3 has very old properties, the cluster 0 has quite recent ones and the clusters 1 and 2 have a majority of old ones. According to these observations, the clusters can be described as follows.

- Cluster 0: Medium high density, medium value, low number of rooms, almost as many owners as tenants, mostly apartments, and recent properties
- Cluster 1: Low density, low value, high number of rooms, mainly owners, majority of detached houses, and quite old properties
- Cluster 2: High density, medium value, medium number of rooms, majority of tenants, mostly apartments, and old properties
- Cluster 3: Medium low density, very high value, very high number of rooms, mainly owners, equitably distributed types of dwellings, and very old properties

The Figure 15 shows the clustered FSAs of the Island of Montreal based on the housing dataset.

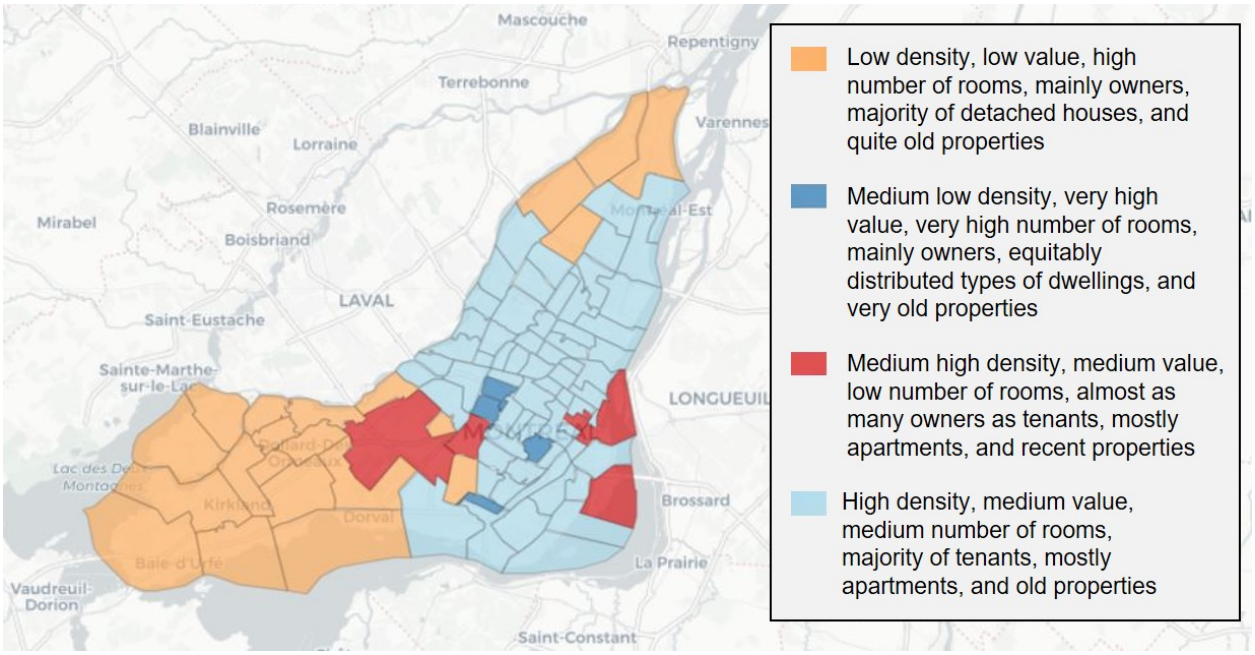


Figure 15: Map of the clustered FSAs of the Island of Montreal based on the housing dataset

6 Discussion

There are several approaches to perform clustering and the most suitable model for a specific project must be selected. In the context of our study, three algorithms are considered: K-Means, DBSCAN and agglomerative clustering. DBSCAN is an algorithm that automatically identifies outliers. It could be useful for some applications but it does not meet our needs for this project. All of the FSAs need to belong to a cluster and none of them must be considered as outlier. The DBSCAN approach could therefore be eliminated according to this observation, without even testing it. The DBSCAN model should then be used only if outliers are possible.

For the two other approaches, the parameters selection allows to have the best results. Both algorithms perform fairly equally but the detailed analysis of the results points out the K-Means model as the best regarding the goal of our study. To select the best model for a project, the choice must not be driven only by the model performance metrics. Indeed, the expected results for the application must also be considered.

7 Conclusion

In this project, homogeneous clusters of neighborhoods of the Island of Montreal are built based on the environment, the population and the housing. The data required is collected from different sources, then cleaned and processed to be used for three clustering approaches: K-Means, DBSCAN and agglomerative clustering. After a parameters selection for each one, the three models and their results are compared and the K-Means approach is identified as the best for this study. The clusters thus obtained are described and analyzed.

This study shows that the clusters obtained regarding the environment, the population and the housing are quite different, many neighborhoods are together for a specific category but not for the two others. However, there is always a distinction between the center and the periphery of the Island and more generally some areas are quite homogeneous regarding the three categories. In conclusion, even if some areas seem to be uniform, neighborhoods at a FSA scale needs to be considered to get a precise understanding of a specific location regarding the environment, the population or the housing.

This project could be improved by exploring some other directions. First, instead of considering three separated datasets, the main features from all the data sources could be considered as a whole and the clustering could be performed with all these features. It would give an overall profile of the Island of Montreal, taking into account the environment, the population and the housing. The data could also be refined. For example, real estate data could be collected to consider the actual sale price of properties, instead of their declared value, and the current number of properties for sale or for rent. Another interesting path to explore could be the study of the evolution of neighborhoods over time. Historical data could be used to build clusters over time. This analysis would show how the neighborhoods' profiles evolves and how quickly it happens.

References

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

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

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

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

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

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

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