Data Science Capstone Final Project

Fire Safety in Montreal:

Selecting the optimal location for a new fire station on the island of Montreal

December 18, 2019

Alex Cohen

Andreea Florean

Oliver Foster

Natalia Zader

Table of Contents

Problem Statement

The goal of this project is to select the optimal location for a new fire station on the island of Montreal. This analysis is done in two parts: i) predicting the risk of an incident occurring in each area in the city as well as the features that can lead to high risk; and ii) quantifying the readiness of each area of the city by determining the distance to the nearest fire station.

Data Sources

The data sources used in this project are summarized in the table below. Most of the data comes from the Montreal Open Data website or the Statistics Canada website. Note that in this table, FSA represents Forward Sortation Areas, i.e. the first three letters of the postal code.

Dataset	Description	Granularity	Years	Source
Firefighter data	Fire department incidents and interventions	Street Intersection (Latitude and Longitude)	2005-2019	Montreal open data website
	Fire station geospatial territories: Longitude and Latitude	Territory spatial polygon	2019	Montreal open data website
Property assessment	Types of buildings, construction year, number and type of dwellings and location on the map	Building	2019	Montreal open data website
Montreal geospatial territories	Boundary polygons by FSA region	FSA	2016	Statistics Canada
	Boundary polygons by borough	Borough	2019	Montreal open data website
Population Census	Total population & number of dwellings	FSA	2006, 2011, 2016	Statistics Canada
	Population demographics: family size, marital status and age	FSA	2011, 2016	
Traffic data	Free-flow traffic speeds	Latitude & Longitude	2019	TomTom API developer portal for free-flow
Income data	Based on tax filing information for each household	FSA	2015 2016	Canada Revenue Agency

Data Exploration and Cleaning

Firefighter data

The fire incidents data is composed of all incidents on the island of Montreal between 2005 and 2019. This includes all incidents that firefighters respond to, not only fire incidents. Each incident has a description, an incident category, the responding fire station, the number of fire trucks that were sent, as well as information about the location (latitude and longitude, borough, city). The actual location of the incident is obfuscated to the nearest intersection for privacy reasons.

The data was contained in two files, one with incidents from 2005 to 2014, and another with incidents from 2016 onwards, which were appended. In each of these, the description type (called Description group) had different names for the same categories, so we had to determine which categories in each dataset corresponded to each other (for example, “Premier répondant” and “1-REPOND” are the same category, and refer to incidents where first responders were required). This was done in a new column called “Description Groupe Clean”.

Property Assessment data

We used the following columns as features: construction year, number of dwellings within the building, type of building (residential, commercial, touristic, etc.), property category (regular or condominium), property surface, and the geometry, which included spatial polygons for each building.

This dataset had significant missing information: the number of floors and building surface had approximately 30% of data missing. These columns were not used in our analysis. However, some columns were maintained even if they had some nulls: the number of apartments - 10% missing - and year of construction - 3% missing. Also, the few buildings with year of construction in 1600 were removed, due to historical inaccuracy.

Condominiums that were listed as individual properties but were in the same building were grouped by address. The properties that were listed in the type of building column as parking spaces and storage spaces were removed from the dataset.

Geographic data for the Island of Montreal

Spatial polygons for the FSA territories on the island of Montreal were extracted from the Statistics Canada website. Spatial polygons for the Montreal boroughs, and the firefighter administrative territories were taken from the Montreal open data website.

Census data

Census data were based on the various census topics, such as family characteristics, age categories, households and marital statuses. The number of dwellings and total population was available for 2006, 2011 and 2016. However, the demographics: marital status, family size and age of the residents were only available for the 2011 and 2016 datasets.

Traffic data

Traffic data was extracted for each intersection in the road database (geobase) and sent via API to TomTom's developer portal. TomTom provided us with free-flow traffic speeds for the roughly 25 thousand road locations sent.

Income data

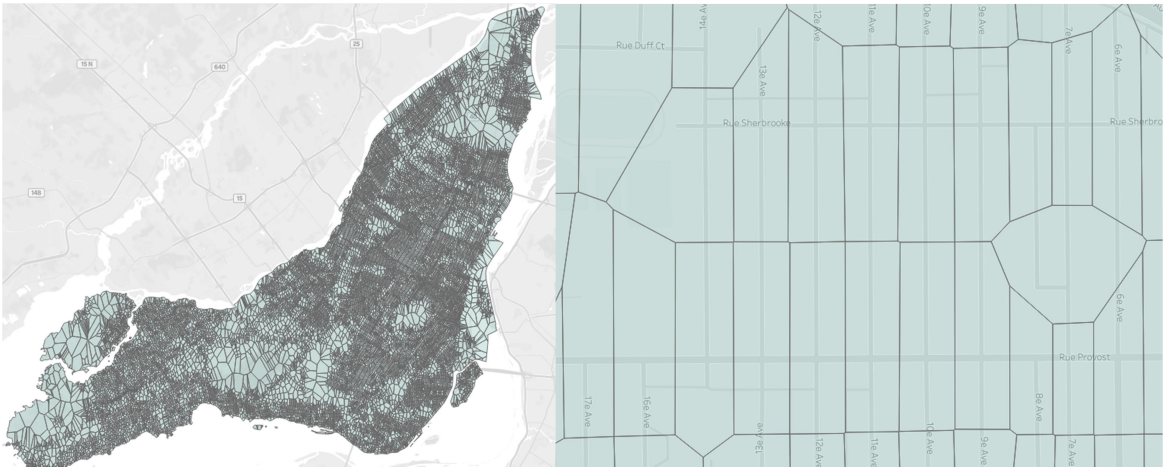
The income data found on the CRA website contained the number of residents per each income group in $5000 increments per FSA region. As this was too granular, the data was grouped into three columns: $0 - $34 999, $35 000 - $69 000, and $70 000 and over. Finally, for this data the percentage of population that fits into each category for each FSA region was also calculated. However, since the data available according to geocode was only for 2015 and 2016, we created an average for both years and used this feature as a constant.

Feature Engineering

Turning intersections into Voronoi polygons

Since the fire incident data was obfuscated to the nearest intersection, it was not possible to determine in which building the incident occured. However, we hypothesized that having information about the buildings in which the incidents occurred (construction year, number of units, etc.) would have strong predictive power on the risk. Therefore, we decided to create small regions around each intersection using Voronoi Polygons, and find statistics about the buildings in these regions.

First, we identified all the intersections in the Montreal Geobase dataset by selecting points that occured on more than one line (street). Then we used the Voronoi function from the Scipy Spatial library to create polygons around each point. Finally, we snapped the edges of the polygons to the border of the island of Montreal, using the spatial data from the administrative territories of each fire station. The figure below shows the island of Montreal segmented into these polygons, as well as an enlarged view of a specific area showing the polygons relative to the underlying streets.



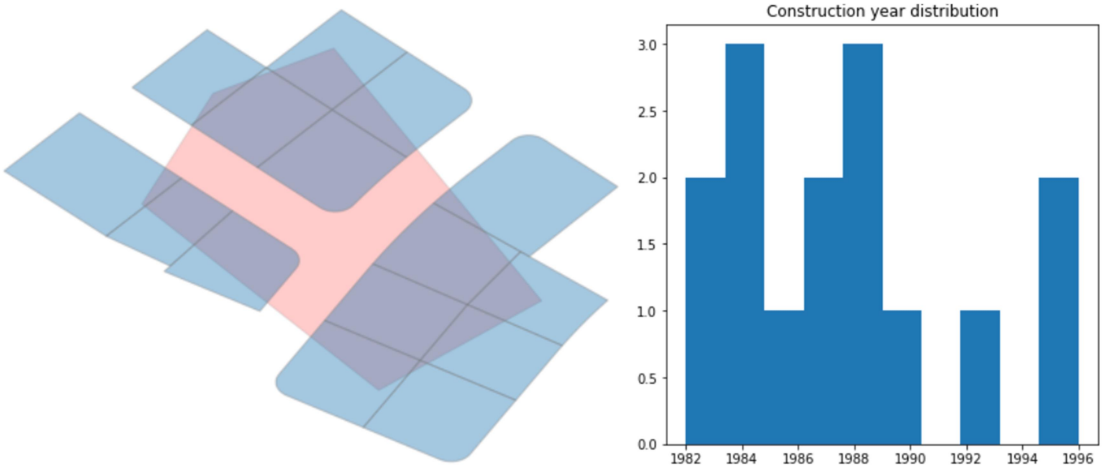
Adding FSA and borough

Because some of the data was at the FSA level, we had to assign each polygon to its corresponding FSA and borough. To do so, we did a Geopandas Spatial join between the polygon data and the Postal Code spatial data, which has a polygon for each FSA. The spatial join was done using a “within” criteria, meaning the intersection polygon had to be within the FSA polygon to be joined. A similar technique was used to find the corresponding borough, using the Boroughs Administrative Borders from the city of Montreal data.

Joining polygons to building data

Once the incidents were summarized at the intersection polygon level, it had to be joined to the buildings data using spatial features. Since we had spatial polygons for each building in our dataset, we were able to join these to our intersection polygons using the Geopandas Spatial join. We used a spatial join with an “intersection” criteria such that a single building can correspond to multiple polygons. This method was selected because i) some buildings spanned more than a single polygon, and ii) the polygons were not perfect, and therefore a building that is at the border of two polygons can just as easily be part of either one.

Below is an example of a polygon (red) and the buildings (blue) that were found to belong to it. In the image, the shape of the street can clearly be observed. A histogram next to it represents a distribution of the year of construction for the 15 buildings in this polygon.



Building statistics by polygon

Since each intersection polygon had multiple buildings, we needed to summarize all the buildings data to get statistics for each polygon. The following features were created:

- Number of buildings in a polygon
- Number of regular buildings, number of condo buildings
- Number of units: total, average, median, and maximum
- Construction year: minimum, maximum, average, and median
- Building area: minimum, maximum, average

Grouping census data

The original dataset contained all our features of interest in one single column which was transposed to create columns. Also, because some of these features in the demographic datasets were too granular, it was modified in the following way:

- Age groups: 0-14, 15-19, 20-24, 25-34, 35-44, 45-54, 55-64 and over 65;
- Family type: couples with children and without children;
- Marital status: married, common-law and single;

In addition, the values in each entry were represented as a percentage of the population that fits into each category for each FSA. Furthermore, since the data available was only for the specific census years, and we wanted to predict year by year, a linear interpolation was done to assume these demographics for every year in between the census years.

Final dataset

Because this model had to be a predictive model, the dataset was structured such that each row contained one polygon for a given year, and a target variable identified how many incidents were in that polygon that year. There were 25 thousand polygons on the island, with incidents data ranging over 15 years (2005-2019).

Tools and Techniques Used

FSA geodata from Stats Canada was extracted using QGIS open source software. The Census data for Montreal was filtered with Alteryx. Changes on the demographic data was done with SQL server.

To handle geographic data the Python library GeoPandas was used to perform complex SQL-style joins based on spatial features in the data preparation process.

In computing distances between points on the road we applied graph theory to the Montreal road database to compute any shortest distance/time metrics (leveraging Dijkstra's Shortest Path First algorithm). This was facilitated through the Python library NetworkX.

Scikit-Learn Pipeline framework was used to construct the estimators that would ultimately score geographic regions in our dataset forecasted for 2020. A Logistic Regression estimator was applied to create this score.

Summary of Modelling Techniques Evaluated

Predicting incident risk

After constructing the dataset it was time to evaluate incident risk in 2020. Here, our goal was to build a predictive model where the inputs would be information about the population,

demographics, and building characteristics of a given polygon in a given year, and the output would be the risk of incident for the following year.

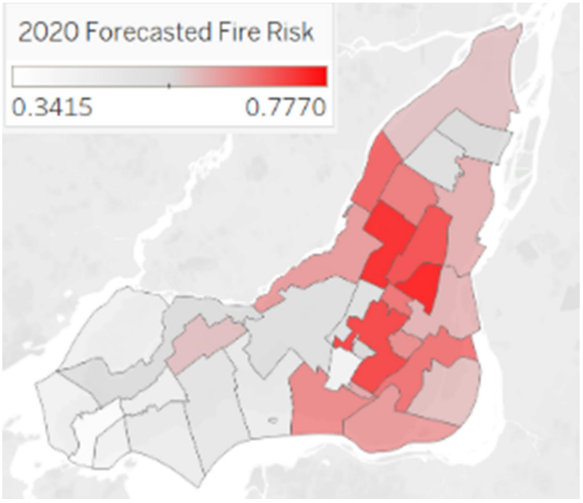
To facilitate this, we took the historical data for each polygon on the grid and, for that year, if there were more than 0 incidents it was labeled as an “incident” area for that year (denoted with the label “1”). Otherwise a “0” label was given.

Once our dataset was labelled we needed to join the input features for the current year to the labels for the next year in the same geographic areas. This way each row in the dataset was comprised of the current-year status and the next-year’s labels.

Before training our models we needed to establish a baseline. We concluded that in order for this model to be useful we needed to be able to surpass the predictive capabilities of saying “if there was an incident this year there will be one next year”. For us to realize this goal we needed to surpass an accuracy of 78.55%.

With the benchmark defined we one-hot-encoded categorical features and fed the data to our gridsearch to determine the optimal estimator. From the test data we observed an accuracy of 78.78% after the gridsearch, with the best estimator being Logistic Regression (slightly more performant than the baseline).

After establishing our trained model we fed our 2019 data to predict for incidents in 2020. We took as output the probability of incident giving us a score from 0 to 1 (0 meaning no probability of incident and 1 meaning 100% chance of incident). With this data we grouped the results by neighborhood and obtained the following incident-risk heatmap:



This risk analysis has provided us with some insight into the issue of incidents: there seems to be high risk towards the center of the island of Montreal, with moderate risk towards the South and North, while the West of the island is relatively low.

Predicting incident readiness

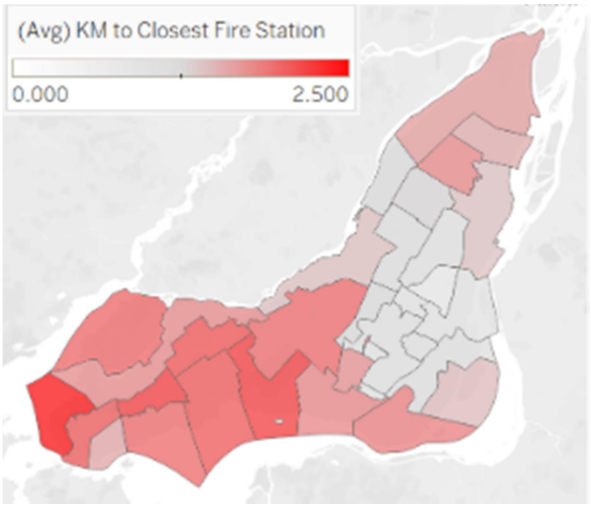
In the previous section it was detailed how the risk of incidents can vary from area to area. In the following section the readiness for a given area to respond to an incident will be evaluated in order to determine if there are areas that not only are at a high risk of an incident, but also lack the ability to respond to such events.

To evaluate the readiness of the city of Montreal to respond to incidents, the average response time to dispatch incidents must be evaluated. Given that the response time data is not released by the city of Montreal, there is no golden source for this data. To overcome this issue graph theory was implemented.

Graph theory is the study of graphs: mathematical structures used to model relationships between nodes via edges between successive nodes. The most common analogy graphs have to our everyday lives happens to be the application we are pursuing: modelling the distance between points on a map via roads. Given that the city of Montreal has released their geobase dataset (a geospatial dataset linking all roadways in the city) we have what is necessary to begin our work.

Approach #1: Minimum Distance to Fire Stations

Our initial (naïve) approach was simple: given we have graph data for every intersection in Montreal, we can calculate the minimum distance a fire truck would need to travel from the closest fire station to that point. To prepare the data the first task was to identify which nodes in the geobase dataset represented the points closest to the fire stations. Taking the Euclidean distance between each node and the fire stations from the fire stations dataset we effectively “snapped” these fire stations to our dataset. With this step complete we now needed to calculate, for every intersection (of which there are roughly 25 thousand), the minimum distance via road to the nearest fire station. This requires iteration through all 25 thousand nodes nested with looping through all 68 fire stations in Montreal. To calculate the shortest distance the data was cleaned into graph form and Dijkstra's algorithm for shortest paths was implemented to find the streetwise distance between the points. The weights of the edges between nodes considered was simply the distance in kilometres between the nodes (hence the naive approach).

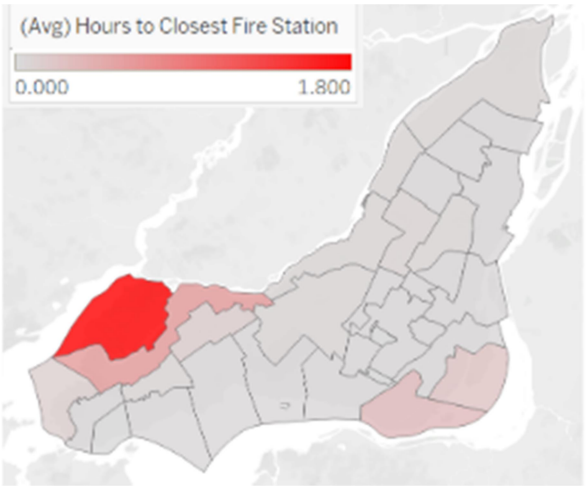


As shown in the previous figure there is a range of travel distances from 2.33km average travel distance in Senneville to 0.74km in the Plateau. Anecdotally the close travel distance being smallest makes sense in the Plateau as this area has the largest population density in all of Canada. In this particular case, the fire stations locations have been planned well.

Approach #2: Minimum Travel Time to Fire Stations

While the Naïve approach gives a good geographic estimate on incident response readiness it does not take into account how traffic can affect the time it takes first responders to get to the scene. The metric that should really be minimized is the response time.

Starting from the naïve approach we have the travel distance between each node and which nodes are fire stations. To enrich this data we amassed free-flow traffic data from TomTom. TomTom Automotive supports a developer pack which allows you to call their API with a given coordinate pair and it will return the average free-flow traffic speed close to that point. With this information we can take the edge data from our graphs and divide the distance between each node with the average free-flow travel speed at that point – giving us the free-flow travel times between these two nodes. Obviously this approach doesn't mimic the travel time taken by fire fighters exactly, but it's a decent approximation. With our data assembled we can run the iterative approach described in the naïve solution substituting minimizing travel distance with travel time.



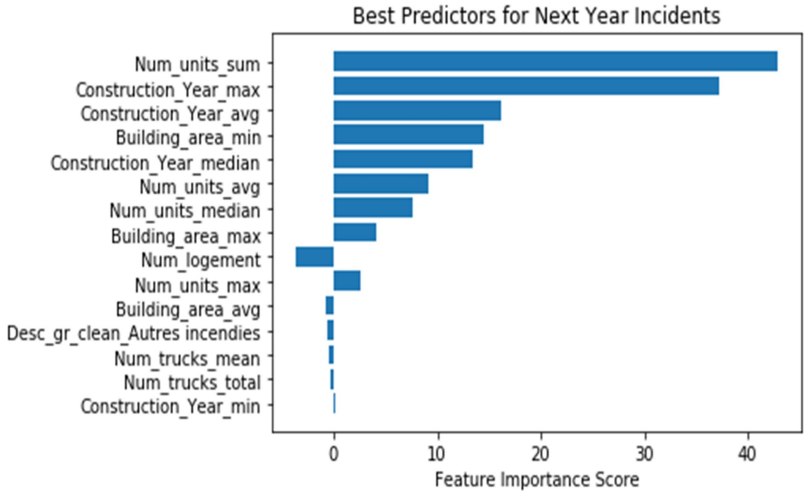
Through this exercise we've demonstrated that there are areas central to the island of Montreal that are far from a fire station - however the free-flow traffic to and from these areas are more efficient. This means that even if they are far from the nearest station they can still quickly travel from one point to the other.

Factoring in travel time has shown that the areas with the least level of readiness are towards the West (Ile Bizard and Pierrefonds-Roxboro) and towards the South (Lasalle and Verdun).

Modelling Results and Insights

Bringing this analysis together requires us to union the readiness level of each neighborhood to handle a fire incident to the potential risk in 2020. The readiness analysis determined that all but the West of the island (Ile Bizard and Pierrefonds-Roxboro) and the South (Lasalle and Verdun) are relatively prepared for an incident. Our forecasted risk in 2020 has highlighted that the center of the island is the most risky with moderate risk in the South and the West. With this in mind we recommend to build a new fire station in the South of Montreal (Lasalle or Verdun) due to its low readiness level and moderately high risk of fires in 2020.

The modelling portion of our analysis also allowed us to expose the most important features in predicting next-year fire incidents. They can be seen in the following graph:



Here, num_units_sum represents the total number of units in the buildings in a polygon. Overall the most important predictors relate to building information in each of the polygons, including construction year (min and max), building area, and total and average number of units per building. This indicates that in order to mitigate the risk of incidents, the most important features to consider are the urban planning and upkeep of all buildings.

Challenges

Overall the largest challenges we faced were related to data preparation and compute-time. In order to join the variety of data sources we needed to settle on a defined grid to join this data to. This required us to upsample and downsample the individual data sets on a geographic key which proved to be a challenging task.

In computing the shortest path (by distance and time) from each intersection to each fire station required a significant amount of resources. This resulted in a nested loop for each fire station and each intersection in Montreal taking roughly 8 days of compute time for the distance and travel time combined.

Conclusion

In conclusion, our analysis has shown that the optimal location to build a new fire station would be in either Lasalle or Verdun. As these neighborhoods are next to each other it would be ideal to pick a single location able to service both efficiently. This analysis has also uncovered that the features that affect fire risk the most significantly are related to the age and size of buildings at the respective intersections. This highlights the need for intelligent urban planning and upkeep to reduce fire risk looking into the future.

Appendix

Risk Scores by Neighborhood (2020 Forecast)

NOM	Forecasted Incident Risk Score
Le Plateau-Mont-Royal	0.777029
Villeray-Saint-Michel-Parc-Extension	0.769236
Côte-des-Neiges-Notre-Dame-de-Grâce	0.741717
Rosemont-La Petite-Patrie	0.732069
Montréal-Nord	0.709441
Le Sud-Ouest	0.692389
Outremont	0.690714
Saint-Léonard	0.680919
Lachine	0.667146
Westmount	0.662615
LaSalle	0.647666
Ahuntsic-Cartierville	0.646183
Montréal-Ouest	0.630690
Mercier-Hochelaga-Maisonneuve	0.618059
Ville-Marie	0.613176
Verdun	0.600468
Dollard-des-Ormeaux	0.597773
Rivière-des-Prairies-Pointe-aux-Trembles	0.597204
Hampstead	0.564999
Montréal-Est	0.559622
Pierrefonds-Roxboro	0.557141
Anjou	0.539411
Saint-Laurent	0.536045
Mont-Royal	0.523829
Pointe-Claire	0.499514
Beaconsfield	0.473552
Dorval	0.471689
L'Île-Bizard-Sainte-Geneviève	0.446160
Kirkland	0.435886
Senneville	0.432481
Baie-d'Urfé	0.411398
Côte-Saint-Luc	0.405310
Sainte-Anne-de-Bellevue	0.341513

Approach #1 Readiness Score by Neighborhood

NOM	Average distance to nearest Fire Station (km)
Senneville	2.333275
Dorval	2.123995
Kirkland	2.110505
Dollard-des-Ormeaux	2.008051
Pointe-Claire	1.992911
Sainte-Anne-de-Bellevue	1.973865
L'Île-Bizard-Sainte-Geneviève	1.918367
Beaconsfield	1.892790
Saint-Laurent	1.887281
LaSalle	1.780833
Lachine	1.739057
Anjou	1.728558
Pierrefonds-Roxboro	1.722328
Côte-Saint-Luc	1.685995
Rivière-des-Prairies-Pointe-aux-Trembles	1.615904
Baie-d'Urfé	1.558487
Montréal-Est	1.520652
Verdun	1.434700
Ahuntsic-Cartierville	1.401469
Mercier-Hochelaga-Maisonneuve	1.393917
Saint-Léonard	1.268774
Mont-Royal	1.207525
Montréal-Nord	1.190002
Villeray-Saint-Michel-Parc-Extension	1.115468
Le Sud-Ouest	1.053221
Westmount	1.025236
Côte-des-Neiges-Notre-Dame-de-Grâce	1.021401
Rosemont-La Petite-Patrie	1.020906
Hampstead	0.948456
Ville-Marie	0.929660
Outremont	0.916217
Montréal-Ouest	0.834939
Le Plateau-Mont-Royal	0.744118

Approach #2 Readiness Score by Neighborhood

NOM	Average travel time required to nearest Fire Station (hours)
L'Île-Bizard-Sainte-Geneviève	1.817250
Pierrefonds-Roxboro	0.585997
LaSalle	0.313530
Verdun	0.215166
Dollard-des-Ormeaux	0.117471
Senneville	0.108621
Villeray-Saint-Michel-Parc-Extension	0.054000
Rivière-des-Prairies-Pointe-aux-Trembles	0.044972
Ahuntsic-Cartierville	0.041738
Saint-Laurent	0.040288
Montréal-Nord	0.039405
Kirkland	0.039304
Rosemont-La Petite-Patrie	0.038732
Le Plateau-Mont-Royal	0.038323
Anjou	0.038023
Mercier-Hochelaga-Maisonneuve	0.037973
Outremont	0.037557
Dorval	0.034649
Saint-Léonard	0.033277
Sainte-Anne-de-Bellevue	0.031846
Westmount	0.031803
Ville-Marie	0.030809
Montréal-Est	0.030061
Côte-Saint-Luc	0.028947
Mont-Royal	0.025106
Beaconsfield	0.024226
Côte-des-Neiges-Notre-Dame-de-Grâce	0.023451
Lachine	0.021580
Le Sud-Ouest	0.021405
Pointe-Claire	0.019477
Baie-d'Urfé	0.014864
Hampstead	0.013839
Montréal-Ouest	0.013196