The Battle of Neighborhoods

Capstone Project

Applied Data Science

Seif El Kilany

March 2, 2021

Introduction

Background

Montreal is the second most populated city in Canada and the largest city in the province of Quebec (Government of Canada, 2019). Just like all large metropolises, young working families end up leaving the downtown center looking for larger, more affordable homes in less congested areas to grow their families. With working from home now a fact of life thanks to the COVID pandemic, the exodus from the downtown hastens (Hanes, 2020).

Question

The question that these young families looking to move have to answer and that realtors are always trying to address is: where should they be moving to, which neighborhoods they should be looking at.

Data

To help address this question, we need to identify Montreal neighborhoods, popular venues and schools. To do so several sources were used, below is a description of each source used, along with the data wrangling performed on it.

Montreal Postal Codes

Data Sources

We need to look for geographical data of the city of Montreal, and identify each neighbourhood and to do so we use the postal codes. On Wikipedia we can find a list of the city's postal codes and neighborhoods https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_H' .

Data Wrangling

After loading the Wikipedia table with Montreal neighbourhoods in to a dataframe, we can see that it is not clearly listed and that each postal code and neighbourhood name are combined into one cell.

Figure 1
Initial scrapping of the Wikipedia page.

Out[6]:

	0	1	2	3	4	5	6	
0	H0ANot assigned	H1APointe-aux-Trembles	H2ASaint-Michel,East	H3ADowntown Montreal North(McGill University)	H4ANotre-Dame-de-GrâceNortheast	H5APlace Bonaventure	H7ADuvernay-Est	H8A
1	H0BNot assigned	H1BMontreal East	H2BAhuntsicNorth	H3BDowntown MontrealEast	H4BNotre-Dame-de-GrâceSouthwest	H5BPlace Desjardins	H7BSaint-François	H8B
2	H0CNot assigned	H1CRivière-des-PrairiesNortheast	H2CAhuntsicCentral	H3CGriffintown(Includes Île Notre-Dame & Île S...	H4CSaint-Henri	H5CNot assigned	H7CSaint-Vincent-de-Paul	H8C
3	H0ENot assigned	H1ERivière-des-PrairiesSouthwest	H2EVillerayNortheast	H3EL'Île-Des-Soeurs	H4EVille Émard	H5ENot assigned	H7EDuvernay	H8E
4	H0GNot assigned	H1GMontréal-NordNorth	H2GPetite-PatrieNortheast	H3GDowntown MontrealSoutheast (Concordia Unive...	H4GVerdunNorth	H5GNot assigned	H7GPont-Viau	H8G
5	H0HReserved0H0: Santa Claus	H1HMontréal-NordSouth	H2HPlateau Mont-RoyalNorth	H3HDowntown MontrealSouthwest	H4HVerdunSouth	H5HNot assigned	H7HAuteuilWest	H8H
6	H0JNot assigned	H1JAnjouWest	H2JPlateau Mont-RoyalNorth Central	H3JPetite-Bourgogne	H4JCartiervilleCentral	H5JNot assigned	H7JAuteuilNortheast	H8J

So first we use the pandas stack function to output a one leveled list with all the items. Then we strip the postal code and neighbourhood name into a column each from the combined cell.

Figure 2
Wikipedia scrapped table after manipulation.

Out[10]:

	Postal Code	Neighbourhood
1	H1A	Pointe-aux-Trembles
2	H2A	Saint-Michel,East
3	H3A	Downtown Montreal North(McGill University)
4	H4A	Notre-Dame-de-GrâceNortheast
5	H5A	Place Bonaventure
6	H7A	Duvernay-Est
8	H9A	Dollard-des-OrmeauxNorthwest
10	H1B	Montreal East
11	H2B	AhuntsicNorth
12	H3B	Downtown MontrealEast
13	H4B	Notre-Dame-de-GrâceSouthwest

Lastly we drop the original combined column, any 'not assigned' codes to arrive at a clean dataframe with postal codes and neighbourhood names with a shape of (123, 2).

Figure 3
Confirmation of shape.

```
In [11]: montreal_df.shape
Out[11]: (123, 2)
```

Montreal Coordinates

Data Sources

The Wikipedia page does not provide coordinates, we first try to get them from geocoder, but unfortunately after letting it run for over an hour with no outcome, an alternate source was necessary. We were able to find a list of all Canadian postal codes and coordinates from GeoNames

http://download.geonames.org/export/zip/

Data Wrangling

At GeoNames I was able to find coordinates for all Canadian postal codes in a text file that was in a compressed Zip folder. We use the ZipFile and BytesIO libraries to unzip the folder and read the contents: readme.txt and CA.txt . We read the coordinates table into a dataframe, label the columns as per readme file and drop the unnecessary columns to arrive at a clean dataframe of all Canadian postal codes, neighbourhood names, province and coordinates. Since we are only interested in Montreal, we merge the Montreal postal codes dataframe from earlier with this Canadian coordinates dataframe on 'Postal Code' to get a new dataframe that has all Montreal postal codes, neighbourhoods and coordinates for mapping.

Figure 4
GeoNames data before manipulation.

Out[17]:

	0	1	2	3	4	5	6	7	8	9	10	11
0	CA	T0A	Eastern Alberta (St. Paul)	Alberta	AB	NaN	NaN	NaN	NaN	54.7660	-111.7174	6.0
1	CA	T0B	Wainwright Region (Tofield)	Alberta	AB	NaN	NaN	NaN	NaN	53.0727	-111.5816	6.0
2	CA	T0C	Central Alberta (Stettler)	Alberta	AB	NaN	NaN	NaN	NaN	52.1431	-111.6941	5.0
3	CA	T0E	Western Alberta (Jasper)	Alberta	AB	NaN	NaN	NaN	NaN	53.6758	-115.0948	5.0
4	CA	T0G	North Central Alberta (Slave Lake)	Alberta	AB	NaN	NaN	NaN	NaN	55.6993	-114.4529	6.0
5	CA	T0H	Northwestern Alberta (High Level)	Alberta	AB	NaN	NaN	NaN	NaN	57.5403	-116.9153	6.0
6	CA	T0J	Southeastern Alberta (Drumheller)	Alberta	AB	NaN	NaN	NaN	NaN	50.9944	-111.4632	6.0
7	CA	T0K	International Border Region (Cardston)	Alberta	AB	NaN	NaN	NaN	NaN	49.4721	-112.2408	6.0

Figure 5
GeoNames data after manipulation.

Out[20]:

	Postal Code	Neighbourhood	Place Name	Province	Province Code	Latitude	Longitude
0	H1A	Pointe-aux-Trembles	Pointe-Aux-Trembles	Quebec	QC	45.6753	-73.5016
1	H2A	Saint-Michel,East	Saint-Michel East	Quebec	QC	45.5618	-73.5990
2	H3A	Downtown Montreal North(McGill University)	Downtown Montreal North	Quebec	QC	45.5040	-73.5747
3	H4A	Notre-Dame-de-GrâceNortheast	Notre-Dame-de-GrÔce Northeast	Quebec	QC	45.4717	-73.6149
4	H5A	Place Bonaventure	Place Bonaventure	Quebec	QC	45.4992	-73.5646
5	H7A	Duvernay-Est	Duvernay-Est	Quebec	QC	45.6739	-73.5924
6	H9A	Dollard-des-OrmeauxNorthwest	Dollard-Des-Ormeaux Northwest	Quebec	QC	45.4948	-73.8317
7	H1B	Montreal East	Montreal East	Quebec	QC	45.6320	-73.5075

Montreal Venues

Data Sources

To explore the neighbourhoods, we use the Foursquare API to get points of interest/venues in each Montreal neighbourhood. Since we use a free account with limitations and for simplicity's sake, we limit each neighbourhood to one hundred venues within a five hundred meter radius of the coordinates.

Data Wrangling

The retrieved data required no manipulation to use (at this point) and consists of neighbourhood name and coordinates, venue name, category and coordinates that we load into a dataframe.

Figure 6
Foursquare Montreal venues data.

Out[29]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Pointe-aux-Trembles	45.6753	-73.5016	Parc-nature de la Pointe-aux-Prairies	45.678834	-73.501162	Park
1	Pointe-aux-Trembles	45.6753	-73.5016	AMT Gare Pointe-aux-Trembles	45.674882	-73.504908	Train Station
2	Pointe-aux-Trembles	45.6753	-73.5016	Parc Yves-Thériault	45.678675	-73.502037	Park
3	Saint-Michel,East	45.5618	-73.5990	Bar Zoe	45.559673	-73.597542	Karaoke Bar
4	Saint-Michel,East	45.5618	-73.5990	Marché Aux Puces Saint-Michel	45.562502	-73.605079	Flea Market
5	Saint-Michel,East	45.5618	-73.5990	STM Station Saint-Michel	45.559425	-73.599749	Metro Station
6	Saint-Michel,East	45.5618	-73.5990	Restaurant Kim Hour	45.561836	-73.605112	Chinese Restaurant
7	Saint-Michel,East	45.5618	-73.5990	Petro-Canada	45.560984	-73.602396	Gas Station

Montreal English Schools

Data Sources

Since families are a big stakeholder here, we need the location of all English elementary and high schools in each neighborhood. We can get an up to date list of elementary and high schools from

the website of the English Montreal School Board https://az184419.vo.msecnd.net/emsb/emsb-website/en/docs/2020-2021/list-of-schools-20-21.pdf .

Data Wrangling

The elementary and high school data are in two separate tables in a pdf file. First we use the tabula library to read the pdf file. This gives us a list of all the tables in the pdf document. The first scrapped table contains elementary schools without headings as they were in text in the pdf.

Figure 7
Pdf scrapped data.

Out[80]:

	0	1	2	3	4	5	6	7
0	BancroftPK	B	1001	4563 St. Urbain H2T 2V9	514.845.8031	514.845.4352	Dorothy Ostrowicz	NaN
1	CarlylePK	E&IB	1002	109 Carlyle, TMR H3R 1S8	514.738.1256	514.738.0373	Dina Vourdousis	NaN
2	Cedarcrest	I	1003	1505 Muir, St. Laurent H4L 4T1	514.744.2614	514.744.3310	Elena Zervas	NaN
3	CoronationPK	E&I	1045	4810 Van Horne H3W 1J3	514.733.7790	514.733.7701	Mike Talevi	NaN
4	Dalkeith\rPK	E	1004	7951 Dalkeith, Anjou H1K 3X6	514.352.6730	514.352.0243	John Wright	NaN
5	DantePK	B	1005	6090 Lachenaie, St. Léonard H1S 1P1	514.254.5941	514.254.6697	Joseph Schembri	NaN
6	Dunrae GardensPK	I	1006	235 Dunrae, TMR H3P 1T5	514.735.1916	514.735.7051	Despina Michakis	NaN
7	East HillPK	I	1007	10350 Perras, RDP H1C 2H1	514.494.3202	514.494.3153	Liboria Amato	Cynthia Canale

So we take that and put it in a dataframe, rename the column heads correctly and add a column 'Type' to label these schools as elementary for our records.

Figure 8
Elementary school dataframe after manipulation.

Out[81]:

	School	Prog	Ext	Address	Tel	Fax	Principal	Vice Principal	Type
0	BancroftPK	B	1001	4563 St. Urbain H2T 2V9	514.845.8031	514.845.4352	Dorothy Ostrowicz	NaN	Elementary
1	CarlylePK	E&IB	1002	109 Carlyle, TMR H3R 1S8	514.738.1256	514.738.0373	Dina Vourdousis	NaN	Elementary
2	Cedarcrest	I	1003	1505 Muir, St. Laurent H4L 4T1	514.744.2614	514.744.3310	Elena Zervas	NaN	Elementary
3	CoronationPK	E&I	1045	4810 Van Horne H3W 1J3	514.733.7790	514.733.7701	Mike Talevi	NaN	Elementary
4	Dalkeith\rPK	E	1004	7951 Dalkeith, Anjou H1K 3X6	514.352.6730	514.352.0243	John Wright	NaN	Elementary
5	DantePK	B	1005	6090 Lachenaie, St. Léonard H1S 1P1	514.254.5941	514.254.6697	Joseph Schembri	NaN	Elementary
6	Dunrae GardensPK	I	1006	235 Dunrae, TMR H3P 1T5	514.735.1916	514.735.7051	Despina Michakis	NaN	Elementary
7	East HillPK	I	1007	10350 Perras, RDP H1C 2H1	514.494.3202	514.494.3153	Liboria Amato	Cynthia Canale	Elementary

We now repeat the process for the high school which is the second table in the pdf tables list.

Figure 9
High school dataframe after manipulation.

Out[84]:

	School	Ext	Address	Tel	Fax	Principal	Vice Principal	Type
0	F.A.C.E.	1147	3449 University H3A 2A8	514.350.8899	514.350.2612	Marilyn Ramlakhan	Jennifer Harriet	High
1	James Lyng	1101	5440 Notre Dame W. H4C 1T9	514.846.8814	514.846.3006	Lino Buttino	Andrea Dillon	High
2	John F. Kennedy	1102	3030 Villeray H2A 1E7	514.374.1449	514.374.2224	Otis Delaney	Vito Campbell-\rGuerriero	High
3	John Grant	1117	5785 Parkhaven, Cote St. Luc H4W 1X8	514.484.4161	514.484.4969	Jennifer Le Huquet	NaN	High
4	L.I.N.K.S.	1109	9905 Papineau H2B 1Z9	514.723.2845	514.723.2666	Maria Calderella	NaN	High
5	Laurenhill Academy	1104	2505 Cote Vertu, St. Laurent H4R 1P3	514.331.8781	514.331.7145	Donna Manos	Rea Limperopoulos	High
6	Laurenhill Jr. Campus	5662	2355 Decelles H4M 1C2	514.331.8019	514.331.0205	Alexander Kulczyk\rMireille Tehbelian	NaN	High

To simplify our data for clustering and segmenting later on we need to clean up the schools data, so we start off by putting both school dataframes together. Since both dataframes have the same columns except for one, we just drop the extra column from the elementary dataframe and use the pandas append function to combine them. As we can see the postal codes (needed for coordinates), in the 'Address' field, so use string manipulation to strip out the postal code from the 'Address' and put it in its own 'Postal Code' column. We don't really need all the school info we have so we can drop six columns from the dataframe to leave us with just school name, type and postal code.

Figure 10
Schools dataframe after basic cleaning and combining.

Out[88]:

	School	Type	Postal Code
0	BancroftPK	Elementary	H2T
1	CarlylePK	Elementary	H3R
2	Cedarcrest	Elementary	H4L
3	CoronationPK	Elementary	H3W
4	Dalkeith\rPK	Elementary	H1K
5	DantePK	Elementary	H1S
6	Dunrae GardensPK	Elementary	H3P
7	East HillPK	Elementary	H1C

To finalize a school dataframe ready for mapping later (Figure 11 below), we groupby 'Postal Code' to get the number of schools in a postal code, then add coordinates and neighbourhood names from the Montreal postal codes dataframe from Figure 5.

Figure 11

Schools dataframe after manipulation.

Out[106]:

	Postal Code	School	Type	Neighbourhood	Place Name	Province	Province Code	Latitude	Longitude
0	H1C	1	1	Rivière-des-PrairiesNortheast	Rivière-des-Prairies Northeast	Quebec	QC	45.6656	-73.5367
1	H1E	2	2	Rivière-des-PrairiesSouthwest	Rivière-Des-Prairies Southwest	Quebec	QC	45.6342	-73.5842
2	H1G	2	2	Montréal-NordNorth	Montreal North North	Quebec	QC	45.6109	-73.6211
3	H1H	1	1	Montréal-NordSouth	Montreal North South	Quebec	QC	45.5899	-73.6389
4	H1K	1	1	AnjouEast	Anjou East	Quebec	QC	45.6097	-73.5472
5	H1N	1	1	MercierSoutheast	Mercier Southeast	Quebec	QC	45.5779	-73.5304
6	H1P	1	1	Saint-LéonardNorth	Saint-Léonard North	Quebec	QC	45.5966	-73.5928
7	H1R	1	1	Saint-LéonardWest	Saint-Léonard West	Quebec	QC	45.5864	-73.6082

References

Government of Canada, S. (2019, April 03). Census in Brief: municipalities in Canada with the largest and fastest-growing populations between 2011 and 2016 CENSUS In Brief: municipalities in Canada with the largest and fastest-growing populations between 2011 and 2016. Retrieved March 02, 2021, from https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016001/98-200-x2016001-eng.cfm

Hanes, A. (2020, May 28). Allison Hanes: Call of the suburbs gets loud during age of covid-19. Retrieved March 02, 2021, from https://montrealgazette.com/opinion/columnists/allison-hanes-call-of-the-suburbs-gets-loud-during-age-of-covid-19