Childcare Options in the Greater Toronto Area

Ana Caprita

IBM Data Science Capstone Project

Report

Abstract

This project aims at providing a better child care center selection option in York Region, Ontario by leveraging data science methods and exploratory data analysis.

Contents

1 Introduction

This project was motivated by my personal struggles as a young working mother looking for childcare options for her one year old in a pre-pandemic world.

The purpose of the project is to help myself and young families moving into York Region of Ontario, Ontario in exploring and much better informed options for early learning and child care. My project aims to create an analysis of comprehensive features around childcare for families coming to the region from within or outside Canada, in order to help them make smart and efficient decision on what's best for them and their young family members.

2 Business Problem

As I mentioned above, the reasoning behind me choosing this subject to explore and analyze was my personal experience. My husband and I decided to move to Markham, Ontario from another city in the province while I was still pregnant so that we have time- pre baby- to find the necessary facilities, services, amenities and know our surroundings before the baby comes. As we are working parents with no grandparents to pitch in, we knew that once my maternity leave ended we had to have a plan in terms of child care. We found it insanely difficult to find and choose, more so that waiting lists are very long.

While York Region, of which the city of Markham is part, has a very comprehensive website and resource list, everything is very static and time consuming and quite some extensive analysis had to be done to find a list of child care options for our child in a pre-pandemic world. It was very hard to find information on daycare and other childcare options in the city in such a way that it makes the decision process easy, efficient and result driven.

The main aim of my project analysis is to help the decision making process less stressful and easier by providing a comprehensive, easy to follow list of child care options in Markham, Ontario, that can be used by new families either already living in the city or just moving in the city. The end result will need to have details on learning and care options, ratings, locations, reviews, programs included and fees, other amenities and dependencies.

3 Data

Based on the business problem this project will use data from various sources in order to achieve its desired result. The main data sources will be:

- Wikipedia list of Postal Codes in Canada, in order to identify cites and neighborhoods in the Grand Toronto Area https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_L

- Geographical coordinates from https://gist.github.com/488fbbc998d0babd500b45f709661c15 in order to find specific location of each city and neighborhood
- Foursquare API for data on childcare locations and user reviews
- City data on child care options : https://insights-york.opendata.arcgis.com/datasets/childrens-service/data
- Provincial data on child care options in the province of Ontario : https://data.ontario.ca/dataset/licensed-child-care-facilities-in-ontario

Secondary data sets will be used for analysis purposes or partial scrapping where other data is not available. For example:

- National data on child care options: https://www150.statcan.gc.ca/n1/pub/11f0019m/11f0019m2006284-eng.pdf
- York region child care programs and market rates : https://ww6.yorkmaps.ca/Html5Viewer24/Index.html?viewer=ChildrensServices.YorkMaps

4 Methodology

A major part of this project is based on exploring and analyzing data from a multitude of sources. Each data source indicated in the previous chapter has been scraped, cleaned, mapped and analyzed in order to get to a point where it can be used as basis for a decision.

4.1 Exploratory Data Analysis

4.1.1 Postal code data

As I mentioned before, getting to a clean and good set of data for the specific subject I chose and the area I chose, it took a lot of data processing, trial and error, troubleshooting as well as data cleaning.

I focused on York Region in the province of Ontario, Canada as this is the region I live in and where I faced so many issues finding a daycare for my daughter about a year ago, specifically the city of Markham and it's surrounding cities as part of the Greater Toronto Area.

In order to retrieve an initial list of postal codes in the York Region I scraped Wikipedia on all postal codes starting with L as these are equivalent to the region I was interested in.

	0	1	2	3	4	5	6	7
0	L1APort Hope	L2AFort Erie	L3ANot assigned	L4AStouffville	L5AMississauga(Mississauga Valley / East Cooks...	L6AVaughan(Maple)	L7ABrampton (West)	L8A
1	L1BBowmanville(East)	L2BNot assigned	L3BWelland(East)	L4BRichmond Hill(Southeast)	L5BMississauga(West Cooksville / Fairview / Ci...	L6BMarkham(Cornell, Ontario / Box Grove, Ontario)	L7BKing City	L8B
2	L1CBowmanville(West)	L2CNot assigned	L3CWelland(West)	L4CRichmond Hill(Southwest)	L5CMississauga(West Creditview / Mavis / Erind...	L6CMarkham(Berczy Village / Cachet / Angus Glen)	L7CCaledon(Caledon East)	L8C
3	L1ECourtice(Bowmanville)	L2ENiagara Falls(Central)	L3ENot assigned	L4ERichmond Hill(Oak Ridges / Lake Wilcox / Te...	L5EMississauga(Central Lakeview)	L6EMarkham(Wismer Commons)	L7EBolton	L8E Park
4	L1GOshawa(Central)	L2GNiagara Falls(Southeast)	L3GNot assigned	L4GAurora	L5GMississauga(SW Lakeview / Mineola / East Po...	L6GMarkham(Downtown Markham / Markham Centre)	L7GGeorgetown	L8G Nortl

As you can see, not even Wikipedia has a one format for everything and depending on the region or starting letter of the postal code the information had different structures. Each cell of this data frame had one postal code, City and Neighborhood all in one word so in order to get to a structured and correct data frame I concatenated all columns into one, split the resulting column based on word character position and them by delimiter:

	PostalCode	Borough	Neighborhood
0	L4A	Stouffville	None
1	L4B	Richmond Hill	Southeast)
2	L6B	Markham	Cornell, Ontario / Box Grove, Ontario)
3	L6C	Markham	Berczy Village / Cachet / Angus Glen)
4	L4C	Richmond Hill	Southwest)
5	L6E	Markham	Wismer Commons)
6	L4E	Richmond Hill	Oak Ridges / Lake Wilcox / Temperanceville)
7	L4G	Aurora	None
8	L6G	Markham	Downtown Markham / Markham Centre)
9	L4J	Thornhill	West)
10	L9N	East Gwillimbury	Holland Landing / River Drive Park)
11	L9N	East Gwillimbury	Holland Landing / River Drive Park)
12	L3P	Markham	Central)
13	L3R	Markham	Outer Southwest)
14	L3S	Markham	Armadale, Ontario / Milliken, Ontario)
15	L4S	Richmond Hill	Central)
16	L3T	Thornhill	East)
17	L3X	Newmarket	Southwest)
18	L3Y	Newmarket	Northeast)

4.1.2 Geolocation data

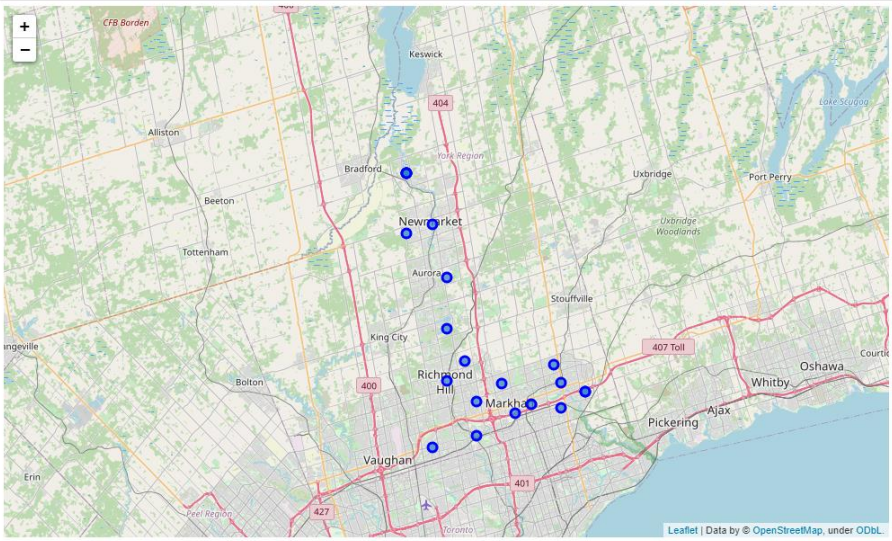
I attempted to retrieve latitude and longitude data for the above postal codes by using geocoder, however it returned just a partial list after several tries and

decided in the end to take the same approach as in the course lab, I retrieved the data separately, per postal code, mainly from Google maps and loaded it in a csv file in Github from where I loaded it in my notebook. End result is:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	L4A	Stouffville	NaN	43.8687	-79.2254
1	L4B	Richmond Hill	Southeast)	43.8584	-79.3916
2	L6B	Markham	Cornell, Ontario / Box Grove, Ontario)	43.8687	-79.2254
3	L6C	Markham	Berczy Village / Cachet / Angus Glen)	43.8783	-79.3522
4	L4C	Richmond Hill	Southwest)	43.8815	-79.4366

However later in the data exploration process I find that there is a more accurate way to retrieve the geolocation data, considering the full postal code of each of the child care centers in the region. Non the less the above table is useful at a higher level to determine a generic map of the area we are looking at:

Out[33]:



4.1.3 Childcare data in the region of York

The government of Ontario does offer a lot of Open Source data sets for different services available in the province, including childcare licensed facilities. Initial dataset is quite extensive and does have a lot of unnecessary information, at least for the purpose of this exercise.

Out[34]:

	Report Snapshot Date	Licensee Name	Program Type Desc	Region Display Name	CMSM DSSAB Name	Child Care Center Name	Licence Number	Program Option	Original Issue Date	Licence Status	Closure or Termination Date	Language of Service Desc	Street Number	
0	2020-02-29	002350227 Ontario Inc. (Creative Minds Childre...	Child Care Centre	Southwest Region	City of Brantford	Creative Minds Children Services	0006463	Full Day(6 hours or more in a day)	2013-04-15	Inactive	NaT	English	61	Sh
1	2020-02-29	002350227 Ontario Inc. (Creative Minds Childre...	Child Care Centre	Southwest Region	City of Brantford	Creative Minds children services inc	56690	Full Day(6 hours or more in a day)	2016-11-24	Active	NaT	English	5	wa
2	2020-02-29	002599266 Ontario INC.	Child Care Centre	Central East Region	Regional Municipality of Durham	Durham Montessori School and Daycare	57087	Full Day(6 hours or more in a day)	2018-01-08	Active	NaT	English	200	By
3	2020-02-29	002633409 Ontario Corporation	Child Care Centre	West Region	Regional Municipality of Halton	Western Heights Montessori Academy	57245	Full Day(6 hours or more in a day)	2018-08-10	Active	NaT	English, French	186	Mc Rd
4	2020-02-29	1.2.3. Look At Me Co-operative Nursery School ...	Child Care Centre	Southwest Region	City of Stratford	1.2.3. LOOK AT ME CO-OPERATIVE NURSERY SCHOOL	14085	Half day(Less than 6 hours in a day)	1991-12-05	Active	NaT	English	465	Ma

I had to drop several columns as a result of that and also renamed most columns in order to have uniformity across the board and all datasets that I had to pull in for this analysis. The end goal with this data set was to have a clean list of all childcare units in the region, with details about the programs they offer.

Out[40]:

	ProgramType	ChildCareCenterName	ProgramType	ServiceLang	City	PostalCode
6	Child Care Centre	CUTIES & PATOOTIES	Full Day(6 hours or more in a day)	English	Newmarket	L3X 1V6
7	Child Care Centre	Cuties and Patooties Childcare Center	Full Day(6 hours or more in a day)	English	newmarket	L3X 1V6
9	Child Care Centre	Eh to Zed Preschool Canada Early Learning Academy	Full Day(6 hours or more in a day), Half day(L...	English	AURORA	L4G 8A3
12	Child Care Centre	1043282 Ontario Inc. - Markham Montessori School	Full Day(6 hours or more in a day), Half day(L...	English	Markham	L3P 2B2
18	Child Care Centre	Kids Can Doodle Inc.	Full Day(6 hours or more in a day), Half day(L...	English	Richmond Hill	L4C 2M6

While the data set is clear, it offers generic information about the Program Type so I had to pull in additional information from another regional data source, this time by calling an API resource and parsing the json format data it returned. After processing the resulting data set I got to the following shape and gained an extra column with Program details:

	ChildCareCenterName	ProgramType	Type	ProgramDetails	ServiceLang	City	Phone_no	PostalCode
0	Cuties and Patooties	Full Day(6 hours or more in a day)	Centre Based Child Care	None	English	Newmarket	(905) 252-7833	L3X 1V6
1	Cuties and Patooties	Full Day(6 hours or more in a day)	Centre Based Child Care	None	English	Newmarket	(905) 252-7833	L3X 1V6
2	Eh to Zed Preschool Canada Early Learning Academy	Full Day(6 hours or more in a day), Half day(L...	Centre Based Child Care	None	English	Aurora	(905) 900-0487	L4G 8A3
3	Markham Montessori School	Full Day(6 hours or more in a day), Half day(L...	Centre Based Child Care	None	English	Markham	(905) 294-3373	L3P 2B2
4	Kids Can Doodle	Full Day(6 hours or more in a day), Half day(L...	Centre Based Child Care	* Open from 7:00 a.m. to 6:30 p.m.; located n...	English	Richmond Hill	(905) 770-7417	L4C 2M6

By analyzing the data in this data set, I can definitely see that most data is complete but for the Program Type. It seems that more facilities than not aren't reporting what type of programs they offer in the York region. However, by pulling data from multiple sources I managed to get another column in with Program details which seems to have no data missing. As such we will replace the missing values in ProgramType with the info from Program details. I will not get rid of Program details as a column as I consider it a complement to Program type and has more useful information. I think this is the best approach as we would miss the point of the project if we just drop the missing data or replace with frequency of others.

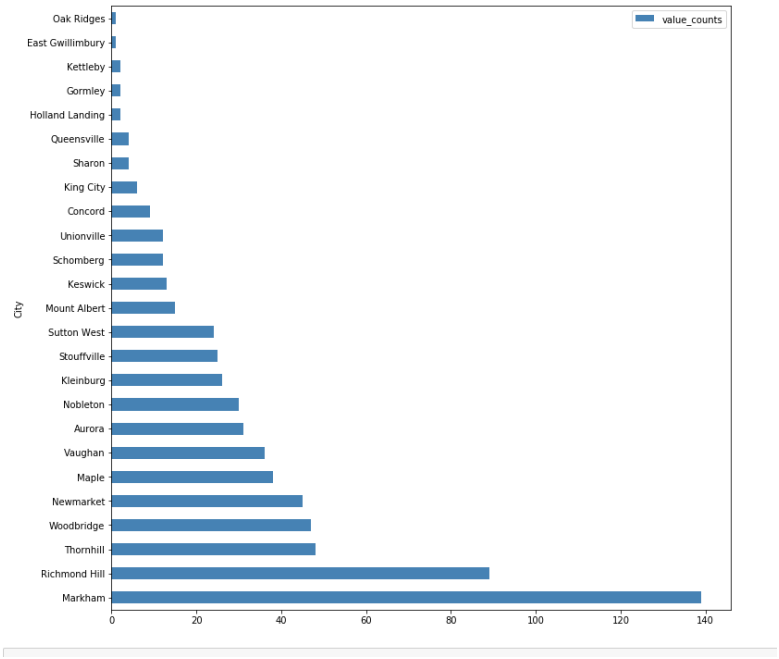
```
: #identify the number of missing values
for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")
```

```
ChildCareCenterName
False    661
Name: ChildCareCenterName, dtype: int64

ProgramType
False    661
Name: ProgramType, dtype: int64

Type
False    661
Name: Type, dtype: int64

ProgramDetails
True     409
False    252
Name: ProgramDetails, dtype: int64

ServiceLang
False    661
Name: ServiceLang, dtype: int64

City
False    661
Name: City, dtype: int64

Phone_no
False    661
Name: Phone_no, dtype: int64

PostalCode
False    661
Name: PostalCode, dtype: int64
```

There are 661 licensed childcare centers in York Region, spread across 28 cities (boroughs) offering services in 7 languages.

	ChildCareCenterName	ProgramType	Type	ProgramDetails	ServiceLang	City	Phone_no	PostalCode
count	661	661	661	252	661	661	661	661
unique	494	17	3	172	7	28	350	411
top	Learning Jungle School - Vaughan South Campus	Full Day(6 hours or more in a day)	Centre Based Child Care	Due to COVID-19 program is supported by virtua...	English	Markham	(905) 552-0610	L0G 1N0
freq	7	165	644	35	634	138	33	30

Out of these 661 childcare centers with active licenses, 644 are Center Based and only 10 are Home based while 7 of them are Nurseries. Most of the centers are located in Markham (139) followed by Richmond Hill, Thornhill, Woodbridge and New Market.



As I do live in Markham this insight is very reassuring as there is a wide variety to choose from.

While this extra data source was useful it still has limited data regarding features and details for each center.

I downloaded extra data sets with Program types and market rates from York region's main website. As the website is not as well structured and formatted it was impossible to scrape data directly, however they did offer download options in csv or pdf format based on a specific query. The response time of the website however was and is very slow and you have to have a very specific search on small areas (km wise) to get a response hence it took a lot of small searches and downloads to get the full data I needed.

This was one of my initial challenges a year ago when actually looking for a child care option.

After downloading all the data I needed I uploaded it on github and loaded it from there.

The program details data set offers information on each type of program a child care center offers, the capacity, vacancies and waiting list numbers for each.

	Unnamed: 0	GCX_OID	Program	Available	Capacity	Vacancies	Waitlist
0	NaN	0	Infant	No	0	0	0
1	NaN	0	Toddlers	No	0	0	0
2	NaN	0	Preschool	No	0	0	0
3	NaN	0	Kindergarten	No	0	0	0
4	NaN	0	Primary/Junior School Age	Yes	54	22	0

The main index of the data set is GCX_OID, which is the unique ID of each child care center and the main relation between the Program, Market rates and a more extensive dataset with Child care center names I will be using later in the analysis s process.

The data set has information of programs up to school age however I am only interested in Infant and Toddler programs. As such, I can see there are 294 centers offering Toddler programs and 134 Infant programs.

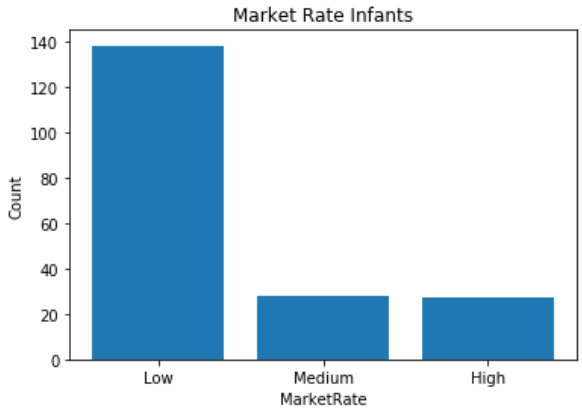
In a similar way I load the Market rates data for the above centers:

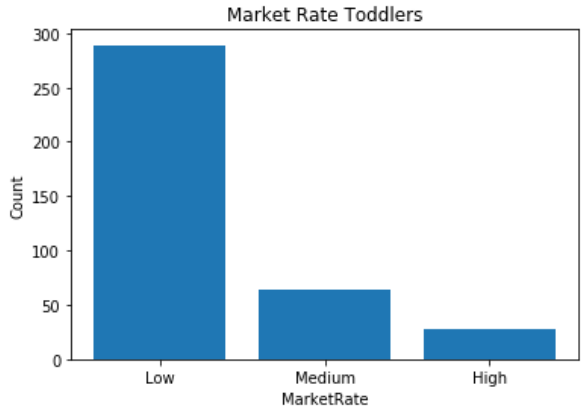
	Unnamed: 0	GCX_OID	CareDescription	AgeGroup	MarketRate
0	NaN	0	School Age After School	School Age	$14.21
1	NaN	0	School Age Before & After	School Age	$15.26
2	NaN	0	School Age Before School	School Age	$7.13
3	NaN	0	School Age Full Day	School Age	$46.00
4	NaN	1	Kindergarten After School	Preschool	$21.00

After cleaning the data and just filtering by the same areas of interest: toddler and infant, I get to a data set offering me insight on different types of programs for each of these two and prices for them. There are half and full day programs for each of them as well as emergency care programs for each, which again is very reassuring.

```
Infant Full Day                 132
Infant Half Day                  19
Emergency Infant Full Day        11
H Infant Full Day                 9
Emergency Infant Extended Day     8
H Infant Half Day                 7
H Infant Part Day                 7

Toddler Full Day                275
Toddler Half Day                 63
Emergency Toddler Full Day       11
H Toddler Full Day                9
Emergency Toddler Extended Day    8
H Toddler Part Day                7
H Toddler Half Day                7
```

From a price perspective I grouped the market rates in Low medium and high for both Infants and Toddlers.



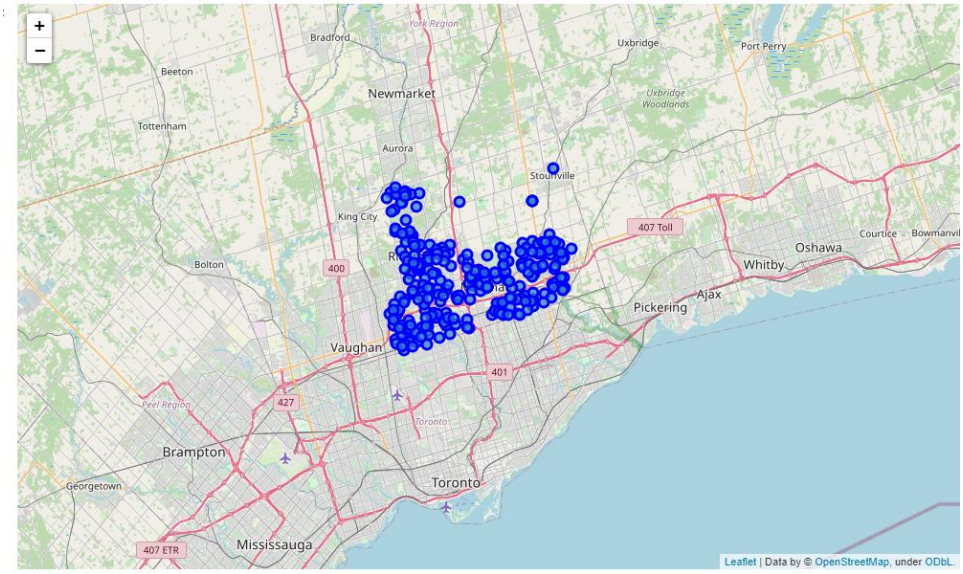


This insight gave me reassurance that the vast majority of centers do have affordable pricing and we will see later on in the analysis data set that most of the also offer fee subsidy which is another plus.

After loading and cleaning the main data set I mentioned above

	Name	Organization	ServiceCategory	Type	Address	Unit	Neighboourhoods	Municipality	MainIntersection	PostalCode	Cor
GCX_OID											
0	Kids Connection @ Beverly Acres	Kids Connection Care & Education	Child Care Centre	Centre Based Child Care	283 Neal Drive	NaN	Richmond Hill	Richmond Hill	Bayview Avenue & Major MacKenzie Drive	L4C 3L3	Mel Taw
1	Windham Ridge YMCA Before & After School Age P...	YMCA of Greater Toronto	Child Care Centre	Centre Based Child Care	32 Red Cardinal Trail	NaN	Richmond Hill	Richmond Hill	Yonge Street & Bloomington Road	L4E 3Y4	Mor Ters
2	Vaughan City Hall YMCA Child Care	YMCA of Greater Toronto	Child Care Centre	Centre Based Child Care	2141 Major Mackenzie Drive	NaN	Maple	Vaughan	Major Mackenzie Drive West & Keele Street	L6A 1T1	Mar Dia
3	John McCrae Child Care	Upper Canada Creative Child Care Centres of On...	Child Care Centre	Centre Based Child Care	565 Fred McLaren Boulevard	NaN	Markham	Markham	Major Mackenzie & McCowan Road	L6E 1N7	Ste Mur
4	Armitage Village Child Care Centre	Upper Canada Creative Child Care Centres of On...	Child Care Centre	Centre Based Child Care	125 Savage Road	NaN	Newmarket	Newmarket	Yonge Street & Mulock Drive	L3X 1R1	Cor

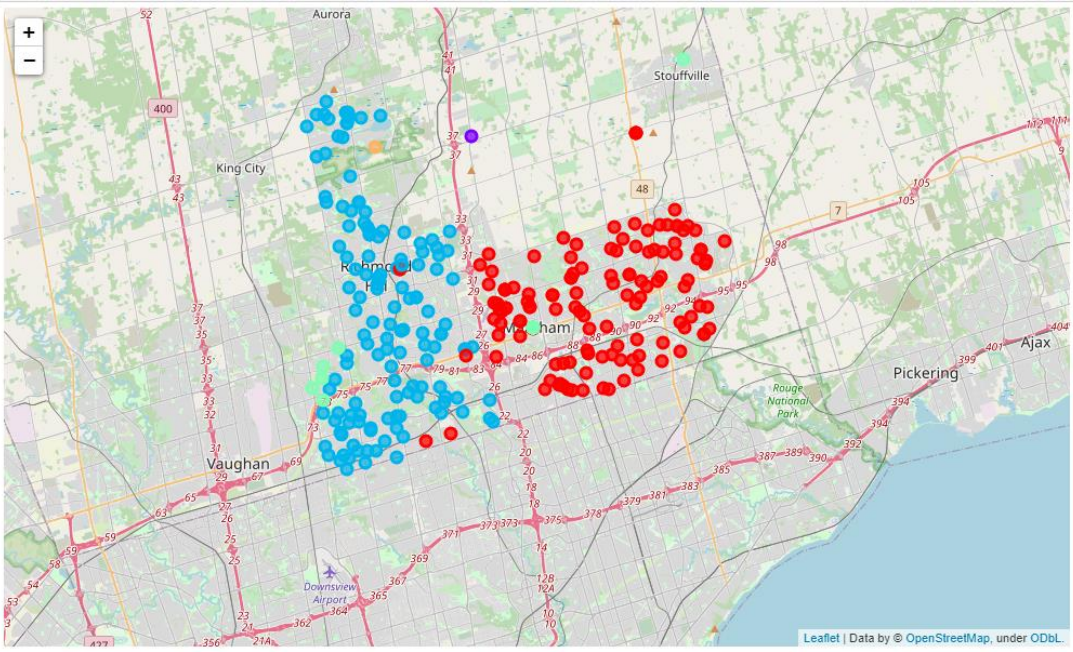
I was able to create and plot a distribution map of all these childcare centers:



One key item to deciding which center to go for is the neighborhood it is in and the type of businesses it has around that center. As such, I decide to analyze a bit the region and each neighborhood by leveraging the Forsquare API, in order to get an initial insight of what type of other businesses is there in the region and in each city/borough and we can see Chinese/Japanese restaurants, bakeries, tea shops as being the most frequent. Insight into this came from one hot encoding for each neighborhood.

4.2 K-means Clustering

I clustered the data by using K-means and made 5 clusters across the region. After which I joined the cluster data with the initial Child care data and plotted on a map the different clusters



What is very obvious for me from the above map and the individual analysis of each cluster, is that the main cities with high density of businesses and child car centers are Markham, Richmond Hill and New Market.

Full details of the whole exploratory data analysis and steps taken to get to clearer set of information can be found in my notebook: https://github.com/bluecare88/Coursera_Capstone/blob/master/DS%20Capstone%20part%202.ipynb

5 Results and Discussion

As a result of going through this process and running this project I now know that Markham as a city has a high density of business and wide variety of them. At the sametime it is the city with the highest number of child care centers and most of them provide center based child care with 50% of these centers being linked to a school environment.

At the sametime it seems that ~75% of the centers have low to medium rates for the programs they offer. The programs themselves seem to cater to full day and half day for all ages and most centers have emergency services as well, in case something unexpected happens for a parent and they need to leave their child in good care.

The main challenge I still see is the fact that to get to these results I had to access and pull in data from a multitude of sources and do a lot of post processing of this data as it comes in different formats, it's often incomplete and is insufficient.

6 Conclusion

All in all I do think my subject was very ambitious and at the end of this exercise I did not get the full result I was looking expecting and attempting. However I do think I have a much better view what my options are in the area I live in and I have gathered a short list of child care centers I will call to get my daughter in for next year (this hoping this pandemic is under control).

The insight I gathered from going through this is extremely valuable to me and is making my life and decision easier.

At the same type a few key takeaways came out of it:

- Province data can be improved and frequency of updating it can be increased
- Childcare centers need a uniform way of sending input and updates on programs and rates they offer
- I have filtered the project data for my own use and decision however the generic datasets this project produced can be used by any parents to look for available childcare options in their cities
- The same exercise can be easily expanded for other regions. For example Toronto has much better data sets and data policies as well as center grading and scoring

7 References

- Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_L
- Foursquare
- York region website: https://insights-york.opendata.arcgis.com/datasets/childrens-service/data
- Ontario Government website : https://data.ontario.ca/dataset/licensed-child-care-facilities-in-ontario
- Stat Canada: https://www150.statcan.gc.ca/n1/pub/11f0019m/11f0019m2006284-eng.pdf

https://childcarecanada.org/sites/default/files/ECEC-in-Canada-2016.pdf