Capstone Report

Carlos Restrepo
carlos.restrepo@live.ca

October 2018

1 Introduction/Business Problem

The target goal is to find the best areas to establish a tutoring service in Ontario, Canada. The idea is to find areas with the schools with the greatest enrolment as well as the fewest tutoring services offered nearby. The presence of schools signifies existence of clients, the lack of offered tutoring services a lack of competition for our prospective business owner. Additionally we consider schools with a lower percentage of low income families or a higher percentage of parents with university education as more valuable in the sense that it indicates that parents can afford the services. The process is very easy to specialize, filtering for Elementary or Secondary School, cities, etc.

2 Data

The data we will use is publicly available from the government of Ontario at https://www.ontario.ca/data/school-information-and-student-demographics. It provides info about almost 5,000 schools all over ontario, most importantly their location coordinates and students enrolled.

Using the location data along with foursquare, a venue search engine, it is possible to search for tutoring services near each of the schools. We will categorize these according to the number of services nearby. We can also collect information about nearby tutoring services if useful or necessary, however we are not for this study.

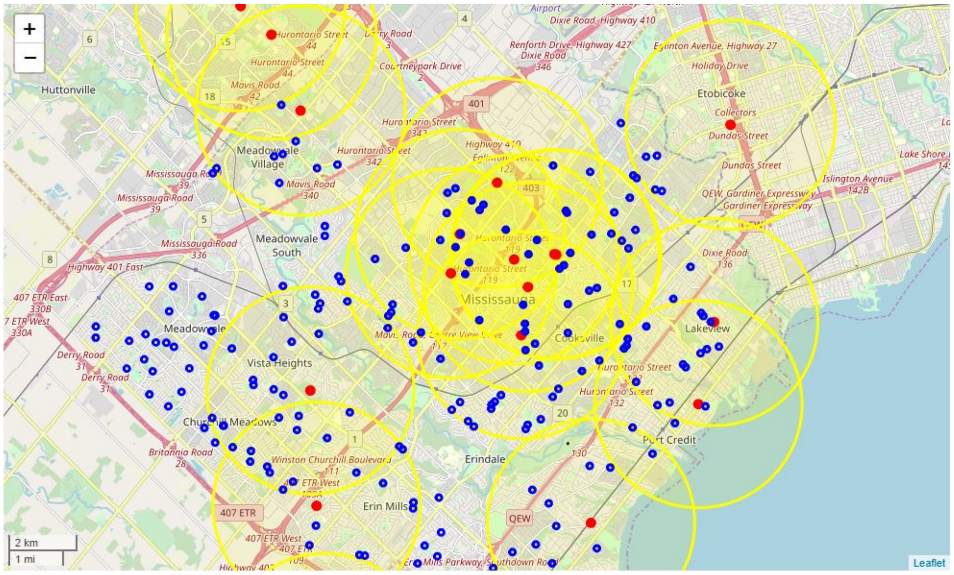
3 Methodology

Once the data is acquired we must reduce the columns to the ones we will use in our study. We need the school name, location coordinates and enrolment as well as the percentages of low income families and parents with university education. Entries with null school names, coordinates or enrolment values are ignored. Null values in the remaining columns are replaced with the average of the non-null values. We will use the schools in the city of Mississauga to show the process applied.

Using the Mississauga school coordinates we search the Foursquare database for 'tutor', 'math', and 'learning' venues at the location of the school. A search radius of 3km was chosen for this example. The results often contained unrelated venues such as those containing the word 'Matheson', a street name. A filter to only include venues with certain subtrings such as 'mathematics' and then exclude those with substrings we do not want, for example 'adult' learning centres. From the resulting venues we collect the number of venues returned for each school as well as a complete list of unique venues returned along with their coordinates for mapping purposes.

To get a rough idea of what our areas look like we can plot markers for both the schools and the tutoring services found.

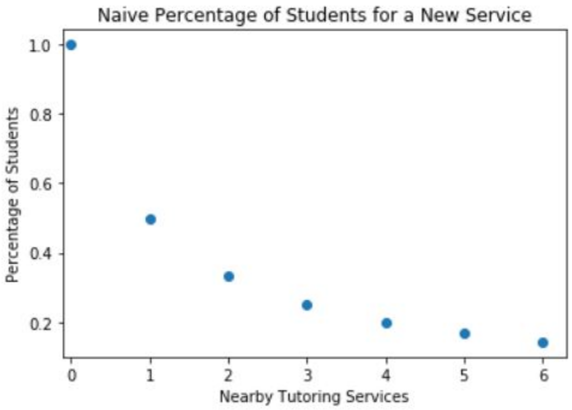
- Blue markers are schools.
- Red markers are tutoring services each with a yellow circle for the 3km search radius used.



We define a naive measure of available students per school. This is calculated for each school by the following equation, the 1 in the denominator represents including our projected tutoring service, the measure gives each nearby tutoring service an equal share of the enrolled students for the school:

$$available\_students = \frac{enrolled\_students}{nearby\_tutors + 1}$$

This measure is important because it emphasizes the priority of schools with few nearby tutoring services by the nature of $f(x) = \frac{1}{x+1}$.



This measure can (and will soon) be extended by instead aquiring the distance between a given school and each of the nearby services, and assigning a portion of the enrolled students to each service depending on how close they are to the school.

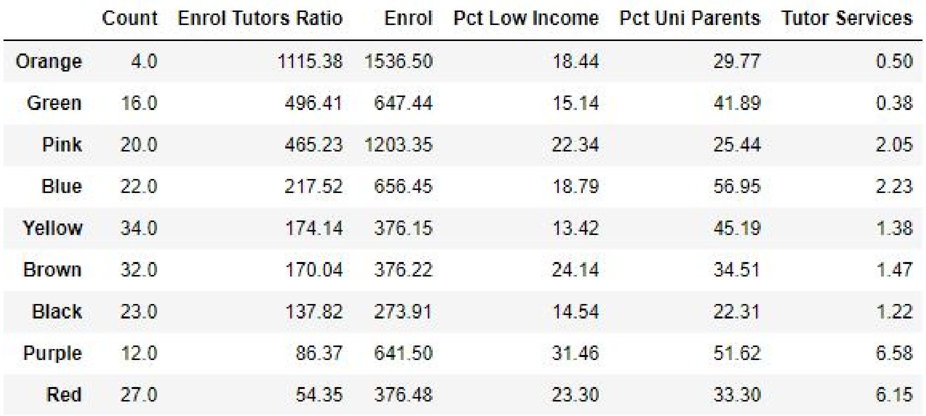
Take for example School X which has 100 students, Service A 1km away and Service B 2km away. We have a ‘total service distance‘ of 3km we want Service A to obtain twice as many students as it is half as far, we calculate $1 - \frac{1km}{3km} = 0.67$ and $1 - \frac{2km}{3km} = 0.33$. As desired Service A is assigned 67 students and Service B is assigned 33 students.

This can be further augmented by only assigning the portion of the students which are not low income or assigning more importance to the percentage of parents with university education.

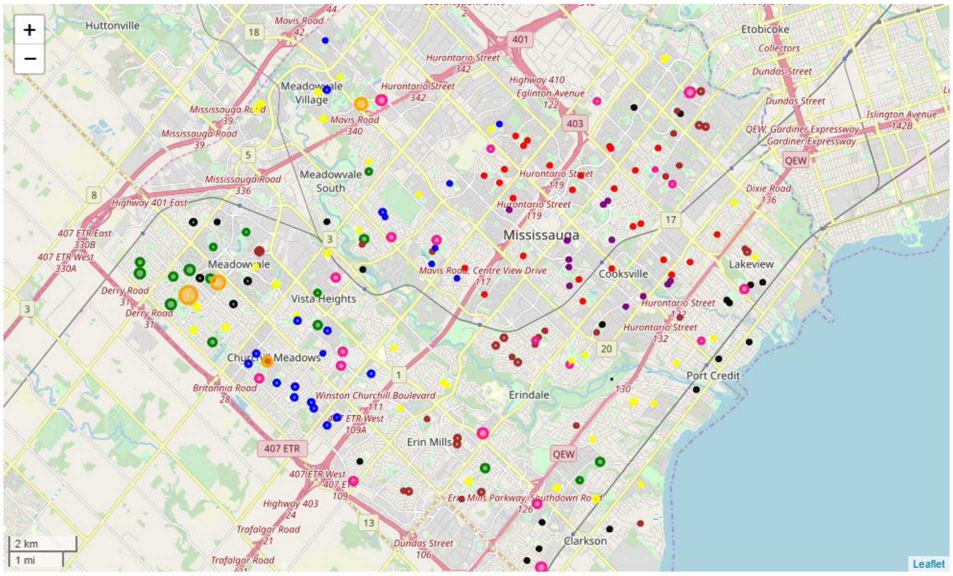
We have some data to categorize our schools by profitability now. We will group our schools into clusters through a KMeans algorithm. We would like to have classes for low/medium/high levels of our three features: enrolment, number of nearby services and economic status so we look to define 9 clusters.

4 Results

The clusters received reveal the schools with the least competition and highest available students. The areas interesting to us are the orange and green markers, orange markers are high enrolment schools:

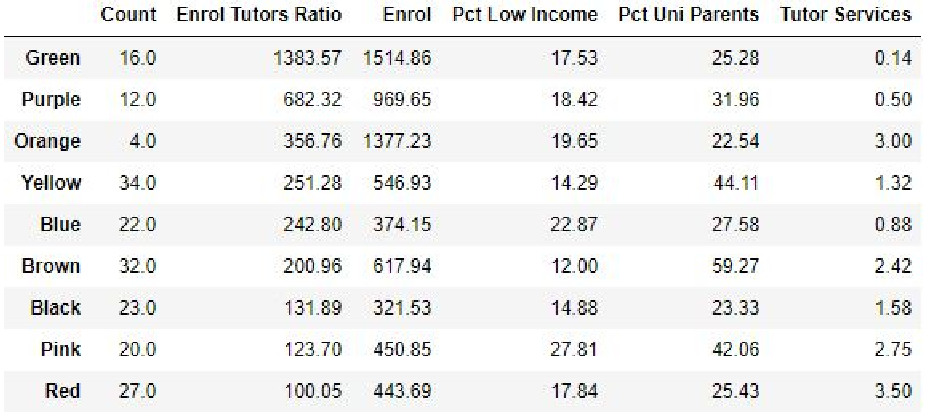


	Count	Enrol Tutors Ratio	Enrol	Pct Low Income	Pct Uni Parents	Tutor Services
Orange	4.0	1115.38	1536.50	18.44	29.77	0.50
Green	16.0	496.41	647.44	15.14	41.89	0.38
Pink	20.0	465.23	1203.35	22.34	25.44	2.05
Blue	22.0	217.52	656.45	18.79	56.95	2.23
Yellow	34.0	174.14	376.15	13.42	45.19	1.38
Brown	32.0	170.04	376.22	24.14	34.51	1.47
Black	23.0	137.82	273.91	14.54	22.31	1.22
Purple	12.0	86.37	641.50	31.46	51.62	6.58
Red	27.0	54.35	376.48	23.30	33.30	6.15

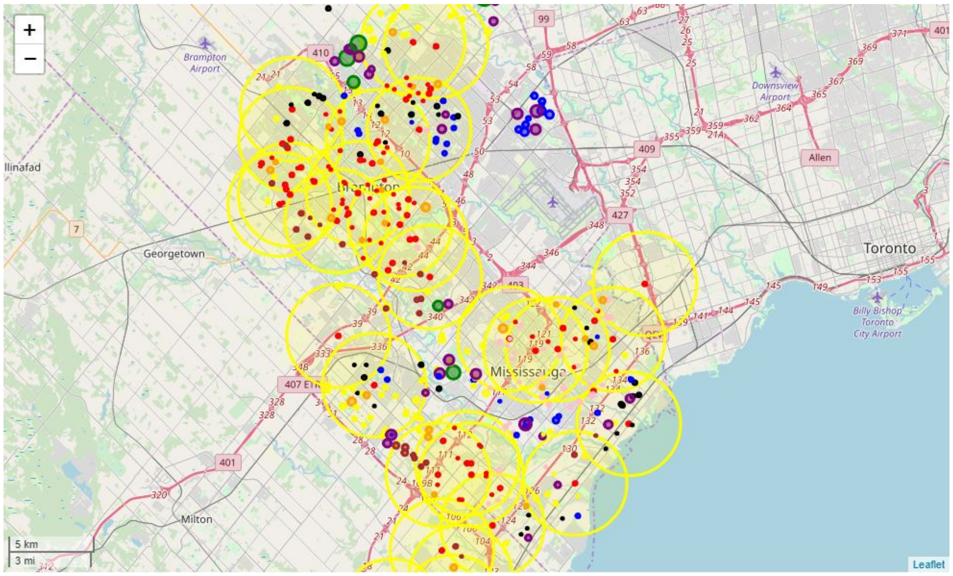


The best area within mississauga seems to be on the west, near Our Lady of Mount Carmel Secondary School and the surrounding schools.

The majority of the code used to generate this report has been generalized for simple use in doing this process for cities in Ontario as well as combinations for cities. For example here is Brampton, Mississauga and Oakville together:



	Count	Enrol Tutors Ratio	Enrol	Pct Low Income	Pct Uni Parents	Tutor Services
Green	16.0	1383.57	1514.86	17.53	25.28	0.14
Purple	12.0	682.32	969.65	18.42	31.96	0.50
Orange	4.0	356.76	1377.23	19.65	22.54	3.00
Yellow	34.0	251.28	546.93	14.29	44.11	1.32
Blue	22.0	242.80	374.15	22.87	27.58	0.88
Brown	32.0	200.96	617.94	12.00	59.27	2.42
Black	23.0	131.89	321.53	14.88	23.33	1.58
Pink	20.0	123.70	450.85	27.81	42.06	2.75
Red	27.0	100.05	443.69	17.84	25.43	3.50



5 Discussion

This study was not taken as far as it could have been. There is much more data that can be used, for example the percentage of students who pass standardized tests. Additionally we are only considering three searches, for better success we may add more.

The biggest limitation in the study is our sole use of foursquare. More search engines could be applied to obtain a complete list of the nearby services.

Please look through the attached Jupyter Notebook on the repository. Do not hesitate to contact at carlos.restrepo@live.ca with any questions or suggestions.

6 Conclusion

The information the data yielded shows us mainly areas which seem unavailable. However there is one clearly optimal area towards the west side of Mississauga. The resulting adviced course of action is for our prospective business owner to establish their tutoring service west of Meadowvale.