**Analysis of Schools’ Neighborhood in Toronto and their Influence on Student’s Education**

Subin Augustine Manickath

November 10, 2019

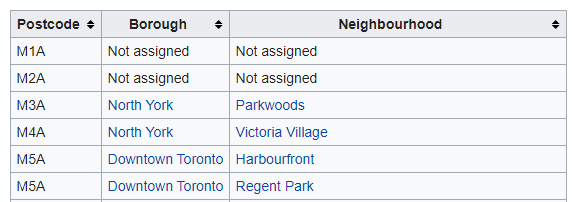
1. **Introduction**
   1. **Description and Discussion of the Background**

While choosing a public school for a child’s education, there are many factors that parents will consider. A key aspect is the rating of the school and the quality of the education it provides. But is the neighborhood and venues in the surrounding area an important factor and an influence on a child’s educational development? Could they be a distraction or a positive influence on the student? Could having more cafés or a movie theatre be a distraction or having more parks and libraries a positive influence? Some parents may give due consideration to the neighborhood where the school is located, and give a higher weightage to schools having more parks in the vicinity than cafes and restaurants. The project seeks to evaluate and ascertain whether the venues surrounding the school have an effect on the student’s educational development and whether they are an influence or detriment to a student’s performance in school.

In order to ascertain this, the project considers a sample set of 104 schools in Toronto. As a representation for the quality of education in a school, the percentage of students passing the Ontario Secondary School Literacy Test in the first attempt is considered. The Ontario Secondary School Literacy Test (OSSLT) measures whether students are meeting the minimum standard for literacy across all subjects up to the end of Grade 9. Successful completion of the literacy test is one of the requirements to earn an Ontario Secondary School Diploma. All students across the province write this test on the same date, usually in late March each year.

1. **Data**
   1. **Toronto’s location information**

Toronto is chosen as the observation area. Information on the boroughs, the neighborhoods and the postcodes were scraped from [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). A snippet of the data downloaded from Wikipedia is as below:



* 1. **Schools information and Test results**

List of schools and the OSSLT results were downloaded from Ontario’s data catalogue, which is available as an excel spreadsheet. The file includes details of all schools within Ontario, including:

* board information
* school information
* grade 3 and 6 EQAO student achievements for reading, writing and mathematics
* grade 9 EQAO academic and applied student achievements
* grade 10 OSSLT student achievement
* student demographic percentages on student parents, special education, first language spoken and new students to Canada

School information includes the location information in longitude and latitude, postcodes, various other data pertaining to the school and students demographics. [Link](https://stage.data.ontario.ca/dataset/school-information-and-student-demographics).

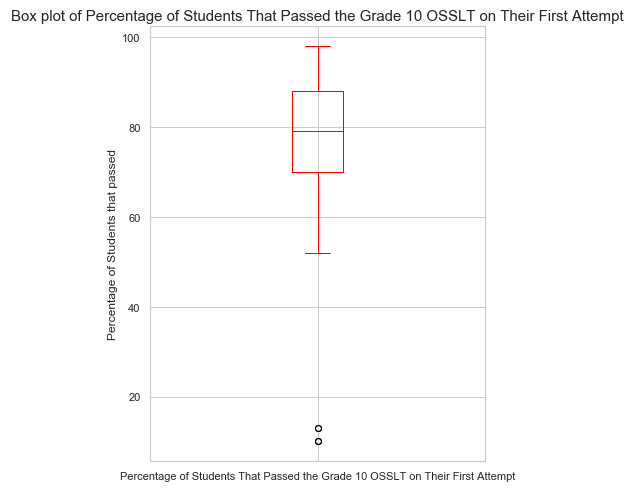
* 1. **Venues in a particular location**

To get the list of venues surrounding a school,Foursquarelocation data was utilized. Foursquare is a technology company that built a massive dataset of location data. They crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers, and this number is only growing.

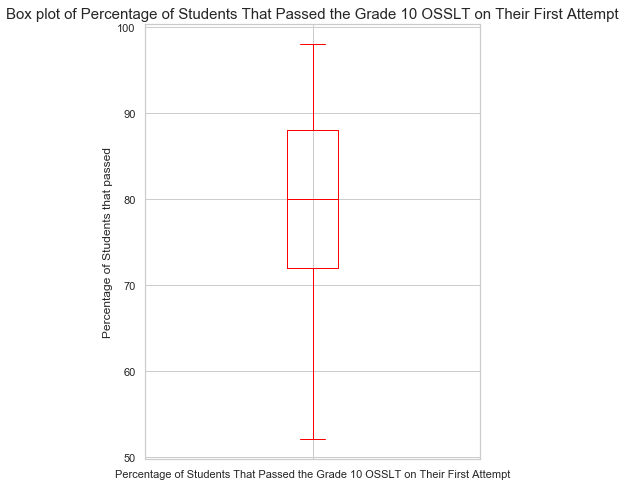
The json file downloaded through the API provides the list of all venues present within a specified radius from the school. For each venue we get its name, unique ID, location, and category.

1. **Methodology**
   1. **Analyzing School Data**

The box plot below shows the distribution of the percentage of students that passed the Grade10 OSSLT on their first attempt (also referred to as the pass percentage of schools in this report). For the purpose of this research, a sample size of 104 schools was considered, all of them located in Toronto.



As we can see, the sample set includes a few outliers, pass percentages for which are below 20%. These could skew our results. Removing the outliers from the data gives us the following boxplot:



Few key observations from the box plot:

* The maximum pass percentage is above 95% while the minimum value is close to 50%, with the median at around 80%
* 75% of the schools had above 72% pass percentage (First quartile).
* 25% of the schools had above 88% pass percentage (Third Quartile)
* 50% of the schools had above 80% pass percentage (Top Half)

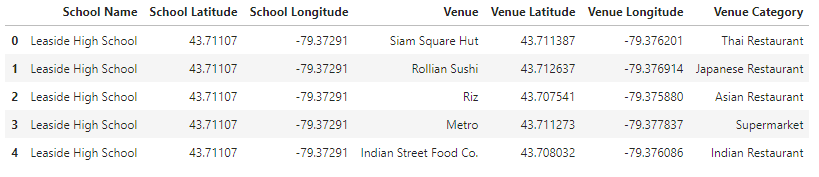
As observed from the plot, there is a significant deviation in the pass percentages of schools. Through this project we’ll try to ascertain whether these deviations are related to the kind of venues surrounding the schools.

The data downloaded from Ontario’ open database listing the schools and the percentages offers accurate and relevant data for our project. Apart from providing the pass percentages for each school, the database provides location information including postcodes and geolocations as longitudes and latitudes. The geolocation data will be used to extract all venues surrounding a school within a certain radius through Foursquare.

* 1. **Analyzing Venues from Foursquare**

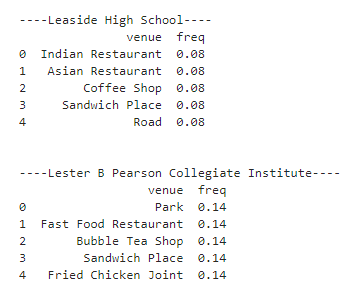
The Foursquare API uses the geolocation data, and the endpoint ‘explore’ to get the list of all venues within a radius from the provided geolocation. The location data of each school will be considered as the center point and all venues within a radius of 500m from each school will be downloaded from Foursquare’s database.

The Foursquare json file downloaded from the API includes a lot of information which will be processed to extract the data which will be relevant to us. The table below shows a snippet of the extracted data post processing:



The table provides information such as the venue name, latitude, longitude and the category of the venue for each school. The number of venues downloaded will be limited to 100.

The data will be further processed to derive the frequency of occurrence of each venue category located around a school. This will be the factor used to classify each school and cluster them according to the similarity in the occurrence of each venue.



Rearranging the table in terms of most common, we get the following:

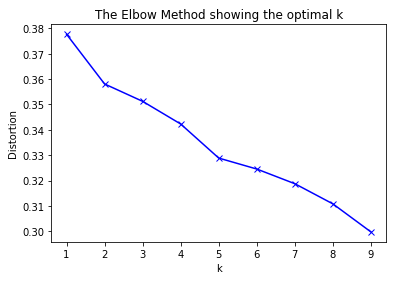


* 1. **K-means Clustering**

The project will use unsupervised machine learning algorithm k-means to cluster each school based on the similarity of their surrounding venues. K-means clustering is a technique in which the dataset is divided through segmentation. The segmentation has to do with complex statistical analysis in which examples within a group are more similar than the examples outside of a group.

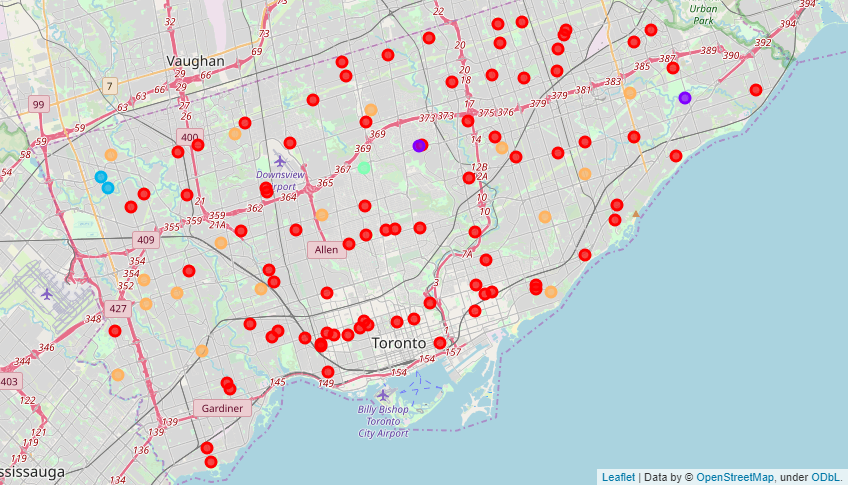
Before we perform a k-means analysis we must specify the number of clusters. For determining the appropriate number of clusters, we’ll use the elbow method.

The elbow method measures the sum of error in each cluster. As the number of clusters increase, this error decreases. However, at a certain point the return on increasing clustering becomes minimal and this is known as the elbow.

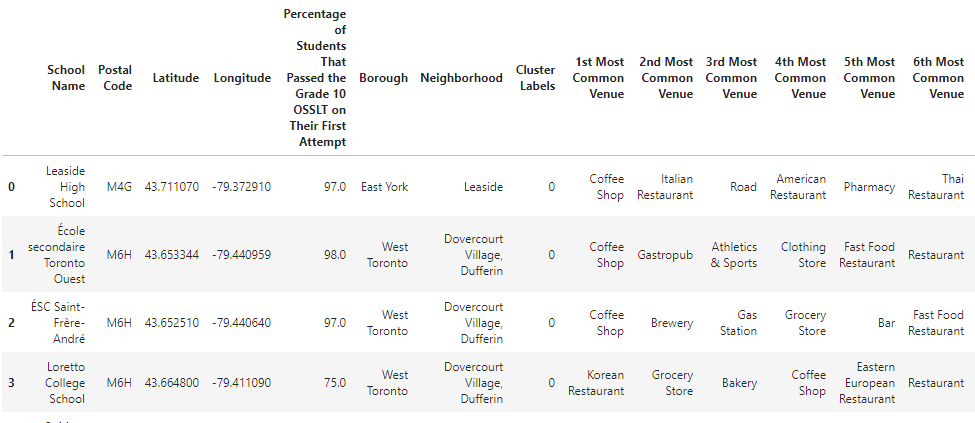


The graph indicates that 5 clusters may be sufficient for this dataset.

We apply the k-means clustering with k as 5. Superimposing the schools using their geolocation and their cluster number on a map of Toronto gives us the following:



Rearranging the data, adding the cluster labels to school information and most common venues derived from our analysis, we get the table below:



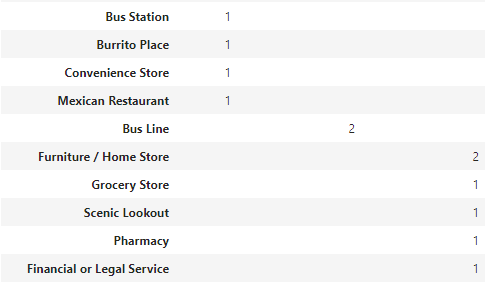
The table provides information on the 1st to 10th most common venue in the vicinity of each school. This information was the primary factor used for clustering the schools. We’ve also added the pass percentage for each school in this table. This information will be used to assess the consistency or the deviation in pass percentages of schools belonging to the same cluster.

* 1. **Analyzing Clusters**

Further analyzing the 1st most common venue in each cluster, we derive the below table.





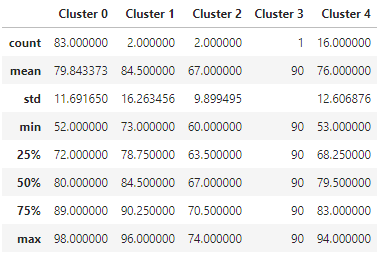


From the table, we can say that:

* The most common venue category surrounding schools in Cluster 0 were coffee shops, restaurants and food outlets. These suggest the neighborhoods in which these schools exist could be social hubs.
* Schools within Cluster 1 have pools as the common factor.
* Distinguishing factor for cluster 2 schools are Bus lines
* The school in Cluster 3 has Pubs as the most common venue
* Cluster 4 is distinguished by the coffee shops and furniture/home store and grocery store as the most common venue. The number of coffee shops and restaurants surrounding schools in cluster 4 are significantly lower than those in Cluster 0.

1. **Result**

The below table summarizes analysis on the pass percentages of schools within each of the 5 clusters. However, it is to be noted the number of schools in each cluster varies a lot. The count in cluster 1, 2 and 3 is not more than five, while in cluster 0 it is 83 and cluster 4 it is 16.



Considering the clusters with maximum sample of schools within them, i.e, cluster 0 and cluster 4, both clusters have similar max and min pass percentages, while the median pass percentages are also the same at approximately 80%. This suggests that although the schools in these two clusters have dissimilar surrounding venues, there are no observed characteristics in the pass percentages of the schools to conclusively suggest one cluster of school has better quality of education than the other.

If we look within cluster 0 or cluster 4, we see a similar trend. Each of the two clusters has a large deviation between the maximum pass percentage (above 90%) and minimum pass percentage (below 55%). The standard deviation (approximately 12%) provides further evidence of the variation in pass percentages of schools within each cluster.

Given the similarity in the two clusters in terms of the pass percentages and the large deviation in pass percentage of schools within each cluster, we can conclude that neighborhood does not have an influence on the education of students of a particular school.

1. **Discussion**

A few aspects of the project that might result in providing a more accurate or in-depth analysis is:

* Although the study covered 104 schools, it should be noted that the cluster size varied significantly, with three of the clusters having less than 5 schools within them. A larger data set may provide a higher number of significant cluster sizes, which would further enhance comparative analysis of the clusters.
* The project uses k-means unsupervised clustering. Using another unsupervised clustering method may provide a different set of clusters which might be a more accurate categorization of the schools based on the venues surrounding their locations.
* The data set can be expanded with more information, such as, on the income status of the student’s households. It could also be further processed, consolidating venues into fewer categories, which may provide a different conclusion.

1. **Conclusion**

Post analysis, we see no observed difference in the pass percentages of schools between clusters. Assessing schools within each cluster, we see significant variation in pass percentage of schools located in identical neighborhood. Therefore, we can conclude that the neighborhood of a school does not have a significant influence over the quality of education of students of a school.

It was observed that in each cluster, pass percentages varied significantly. Similar studies could be performed to understand the root cause of such huge variation in the pass percentages of the school. The factors can be explored with a more comprehensive set of data. Such a study may aid in understanding the root cause of underperformance of a schools and aid in improving the quality of education such that all students have access to high quality education irrespective of the schools that they attend.