Comparison of quality of life of the neighborhoods around top high schools in Massachusetts, USA

Radislav Sendersky

1 Introduction

Moving to a new state is always a big hustle. One needs to understand the safety of towns or cities in the state, overall amenities or venues, entertainment, children development activities and so on. One of the major task for families with children is to understand the landscape of public schools. Families strive to compare cities to choose the best schools. However, the quality of life should not suffer because of a choice of schools.

In this Project, we will compare the quality of life in the cities of Massachusetts, USA that have the best public schools in terms of SAT scores for all students . To do that we will compare the venues in the schools' zip codes.

2 Data

The data for this project will come from several sources. The data on SAT scores and school districts will come for the Department of Education and can be downloaded from http://profiles.doe.mass.edu/. Specifically, we will use the average SAT scores for 2017-18 schools year. The data can be downloaded from http://profiles.doe.mass.edu/statereport/sat.aspx. The data on school addresses can be found http://profiles.doe.mass.edu/search/search.aspx?leftNavId=11238.

Note : for this project the data files were stored on local drive and are as follows:

SAT Scores	sat_performance.csv
School Locations	SchoolLocSearch.csv

The venue information was obtained using FourSquare API.

3 Methodology

In this section we will discuss the methodology used to address the question stated in the Introduction section.

3.1 Data preprocessing

3.1.1 Cleaning the CSV data

The data files downloaded from the above mentioned sources requires some cleaning and processing. All the data munging is performed in the Jyputer notebook, so as to preserve the data files in the original format.

In Fig 1, you can find a sample data from sat_performance.csv. The data has to be read into the pandas dataframe skipping first 2 rows. Also note that in the original file, the extra column "Writing" has no information and has to be dropped. Additionally, the last row of the data file shows information that is irrelevant to our analysis

2017-18 SAT Performance Report - All Students					
District Name	District Code	Tests Take	Reading / Writing	Writing	Math
Abby Kelley Foster Charter Public (District)	4450000	119	536		526
Abington	10000	139	550		535
Academy Of the Pacific Rim Charter Public (District)	4120000	2			
Acton-Boxborough	6000000	419	656		684
Adams-Cheshire	6030000	76	522		522
Advanced Math and Science Academy Charter (District)	4300000	156	631		676
Woburn	3470000	381	551		552
Worcester	3480000	1,891	478		481
State Totals	0	70,155	550		552

Fig 1: Sample data from sat_performance.csv

In Fig 2 a sample of data from SchoolLocSearch.csv is shown. The data file contains a lot of information that is not needed for our analysis. The columns we need are "Org Name", "Org Code" "Address 1", "Town" and "Zip". The main data for our analysis is "Orc Code" and "Zip" while the "Address1" and "Town", "State" are needed for random check on returned indo from FourSquare for debugging purposes.

Org Name	Org Code	Org Type	Function	Contact Name	Address 1	Address 2
Abby Kelley Foster Charter Public (District)	4450000	Charter District	Charter School Leader	Brian Haas	10 New Bond Street	
Abington	10000	Public School District	Superintendent	Peter Schafer	1071 Washington St.	
Academy Of the Pacific Rim Charter Public (District)	4120000	Charter District	Charter School Leader	Spencer Blasdale	1 Westinghouse Plaza Bldg B	

Town	State	Zip	Phone	Fax	Grade
Worcester	MA	01606	508-854-8400	508-854-8484	K,01,02,03,04,05,06,07,08,09,10,11,12
Abington	MA	02351	781-982-2195	781-982-0053	PK
Abington	MA	02351	781-982-2160	781-982-0061	09,10,11,12

Fig 2: Data sample from SchoolLocSearch.csv

Note: In order to preserve the leading zeros in “Org Code”, “District Code” and “Zip” data, the we must pass extra arguments to pd.read_csv as follows:

df_data_1 = pd.read_csv('~/sat_performance.csv', header = 1, dtype = {'District Code':str})

df_data_2 = pd.read_csv('~/SchoolLocSearch.csv', dtype = {'Org Code':str, 'Zip':str})

The data is stored in the pandas DataFrames where: *sat_scores* is the dataframe for the SAT perrmance and *school_loc* data frame holds data on schools’ addresses

3.1.2 Obtaining the FourSquare data

The data from FourSquare is obtained using their API and the standard endpoints. However, the approach taken in our test is to use the “box” coordinates instead of longitude, latitude and radius around the point of interest. In other words we use the following construct:

url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&sw={},{}&ne={},{}'.

The point of interest is the zipcode of the school and we use FourSquare to explore the venues around that zip code. The variables we pass to form the url are “sw” and “ne”, where “sw” specifies the south bound of the region and west bound of the region corresponding to the zip code and “ne” specifies north and east bonds correspondingly. In order to find those bounds we use a special free package, named uszipcode easily installed through pip as: pip install uszipcode

3.1.3 Implementation of the model

The implementation of the model can be outlined in several steps.

Step 1: Define the best school.

This step is accomplished by using the data frame holding the sat scores.

	District Name	District Code	Tests Taken	Reading / Writing	Math
0	Abby Kelley Foster Charter Public (District)	04450000	119	536.0	526.0
1	Abington	00010000	139	550.0	535.0
2	Academy Of the Pacific Rim Charter Public (Dis...	04120000	2	NaN	NaN
3	Acton-Boxborough	06000000	419	656.0	684.0
4	Adams-Cheshire	06030000	76	522.0	522.0

We define a “Good School” as the school where average where not only the total SAT score but the Writing/Reading score is greater than 570 and Math score is greater than 570. This is done to avoid situations where the total score looks great, such as 1250, but its components are Math is 750 and Writing/Reading is 500. Two columns are added to the *sat_scores* dataframe to hold the Total Score

and an indicator column Good School that takes values of 0 and 1 for bad and good school respectively based on our criteria.

Step 2: Merge dataframes and choose top 5 schools

In this step we merge the data frames that hold the SAT scores from each school and the data frame that holds the addresses of those schools. The combined data frame is sorted on Total Score and we check that the top 5 schools also satisfy the critera of Good School = 1. The above conditions are satisfied for the following top 5 schools:

	District Code	Address 1	Town	State	Zip	District Name	Tests Taken	Reading / Writing	Math	Total Score	Good School
0	04680000	85 Prescott Street	Worcester	MA	01605	Ma Academy for Math and Science	56	700.0	757.0	1457.0	1
1	06000000	15 Charter Rd	Acton	MA	01720	Acton-Boxborough	419	656.0	684.0	1340.0	1
2	01550000	146 Maple Street	Lexington	MA	02420	Lexington	599	645.0	674.0	1319.0	1
3	04300000	201 Forest Street	Marlborough	MA	01752	Advanced Math and Science Academy Charter (Dis...	156	631.0	676.0	1307.0	1
4	06400000	120 Meriam Rd	Concord	MA	01742	Concord-Carlisle	305	631.0	644.0	1275.0	1

Step 3: Use SearchEngine from uszipcode to define the “box” coordinates:

The code is simple for this part of the model and we get the following coordinates for the 5 zip codes that we identified in the previous step.

	bounds_west	bounds_east	bounds_north	bounds_south
0	-71.838922	-71.755751	42.314981	42.264392
1	-71.498546	-71.384899	42.533992	42.436937
2	-71.255717	-71.177885	42.490033	42.424904
3	-71.625825	-71.475505	42.380632	42.310977
4	-71.430389	-71.289302	42.504550	42.412715

Step 4: Use FourSquare to explore the neighborhoods of each zip.

As it turns out there are about 150 venues combined for each zip code and there 30 venues in each zip code.

Step 5: Apply clustering algorithm to on the combined data.

The data from all 5 zipcodes is combined into one data frame and the zip codes are removed. One hot encoding is utilized to convert the categorical variables into the numerical values. The clustering algorithm is initialized with 5 clusters. The reason for this choice is two fold. First, we have only 5

zipcode areas and, second, we assume that if regions are all different than all the points will lie in 5 different clusters belonging to each zip code.

Step 6: Examine the data after clustering.

4 Results

The results of the methodology described in the previous section show an extremely strong result. The majority of the data belonged to only 1 cluster:

Clusters	0	1	2	3	4
Zip					
01605	26.0	1.0	0.0	1.0	2.0
01720	25.0	1.0	1.0	2.0	1.0
01742	26.0	2.0	1.0	1.0	0.0
01752	22.0	2.0	2.0	3.0	1.0
02420	27.0	1.0	0.0	1.0	1.0

5 Discussion of the results

The result shows a strong gravitation of the data to one cluster. It can be seen that the data belongs to cluster with a label of zero and that each zip code "contributes" the same amount of data to the clusters. In other words, the venues in each zip code contribute almost equally to the corresponding clusters and the data in the same cluster is by definition "similar". Based on this result we can conclude that the results support the idea that the neighborhoods with good school tend to offer the same "quality of life" as measured by: the number of the existing venues in each zip code area.

6 Conclusion

In this project we designed a back of the envelope methodology to answer the following question: Do the neighborhoods of good schools offer the same quality of life as measured by the similarity of the venues in each neighborhood. The methodology just looked at top 5 schools defined by the average SAT scores of all students in the area. All schools are considered are located in Massachusetts, USA area. The results of the model suggest that the regions are the same in terms of our designed metric. When choosing the area, other qualities and or metrics may need to be considered that could potentially differentiate the neighborhoods.