Classification of Chicago Schools

Using Foursquare Data and Machine Learning to Identify Patterns in Chicago Schools

By Zac Yauney

1. Introduction and Statement of Problem

Not all schools are the same. That much seems like it should be common sense, and yet when we try to help schools improve performance, we tend to offer generalized solutions. While it is true that there are guiding principles for what makes a school better, the neighborhood in which a school is embedded makes a difference in what interventions are most successful.

While it is outside the scope of this exercise to prescribe specific interventions to improve school performance, it should be possible to classify the different types of neighborhoods in which the schools are located. This can be done by combining Chicago Census data and Chicago Public School data with geographic venue data from the Foursquare API. We can list all the different types of venues near a school and then use an unsupervised machine learning algorithm (kmeans clustering) to segment the schools.

This segmentation data should allow policy makers to develop individualized approaches to providing aid to schools in need.

2. Data Sources

Chicago Public Schools – Chicago Data Portal

This dataset from the Chicago Data Portal shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year. More importantly, it gives latitude and longitude coordinates for each school that will allow us to use the Foursquare API to find nearby venues.

Foursquare API

The Foursquare API gives access to Foursquare's huge database of location related data. The piece of data that we are going to be using for this project is their venue data. We will use a request that returns the top 100 venues located withing 500 meters of every school.

In addition, each venue is assigned a category that describes the general function of the venue. The varied distribution of venue categories near each school will form the basis of the segmentation of schools.

3. Methodology

Clustering

The Foursquare API was used to list the 100 most common venues near each school, resulting in a dataframe that looked like this.

	School	School Latitude	School Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abraham Lincoln Elementary School	41.924497	-87.644522	Swirlz Cupcakes	41.923668	-87.646759	Cupcake Shop
1	Abraham Lincoln Elementary School	41.924497	-87.644522	Insomnia Cookies	41.923177	-87.645636	Dessert Shop
2	Abraham Lincoln Elementary School	41.924497	-87.644522	Rickshaw Republic	41.924123	-87.646898	Indonesian Restaurant
3	Abraham Lincoln Elementary School	41.924497	-87.644522	Philz Coffee	41.924879	-87.647094	Coffee Shop
4	Abraham Lincoln Elementary School	41.924497	-87.644522	Potbelly Sandwich Shop	41.923126	-87.645914	Sandwich Place

Then we used one-hot encoding to encode information about what venue types were near each school. The records were grouped by school, with the result being one record for each school that associated a number between 0 and 1 for each (school, venue category) pair. The sample here is spare, but that is natural since there are over 300 categories of venue.

	School	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Antique Shop	Arcade	...	Vineyard	Warehouse	Warehouse Store	Weight Loss Center	Wh
0	A.N. Pritzker School	0.0	0.03	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.0	0.0	0.0	
1	Abraham Lincoln Elementary School	0.0	0.00	0.0	0.0	0.0	0.0	0.01	0.0	0.0	...	0.0	0.0	0.0	0.0	
2	Adam Clayton Powell Paideia Community Academy ...	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.0	0.0	0.0	
3	Adlai E Stevenson Elementary School	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.0	0.0	0.0	
4	Agustin Lara Elementary Academy	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.0	0.0	0.0	

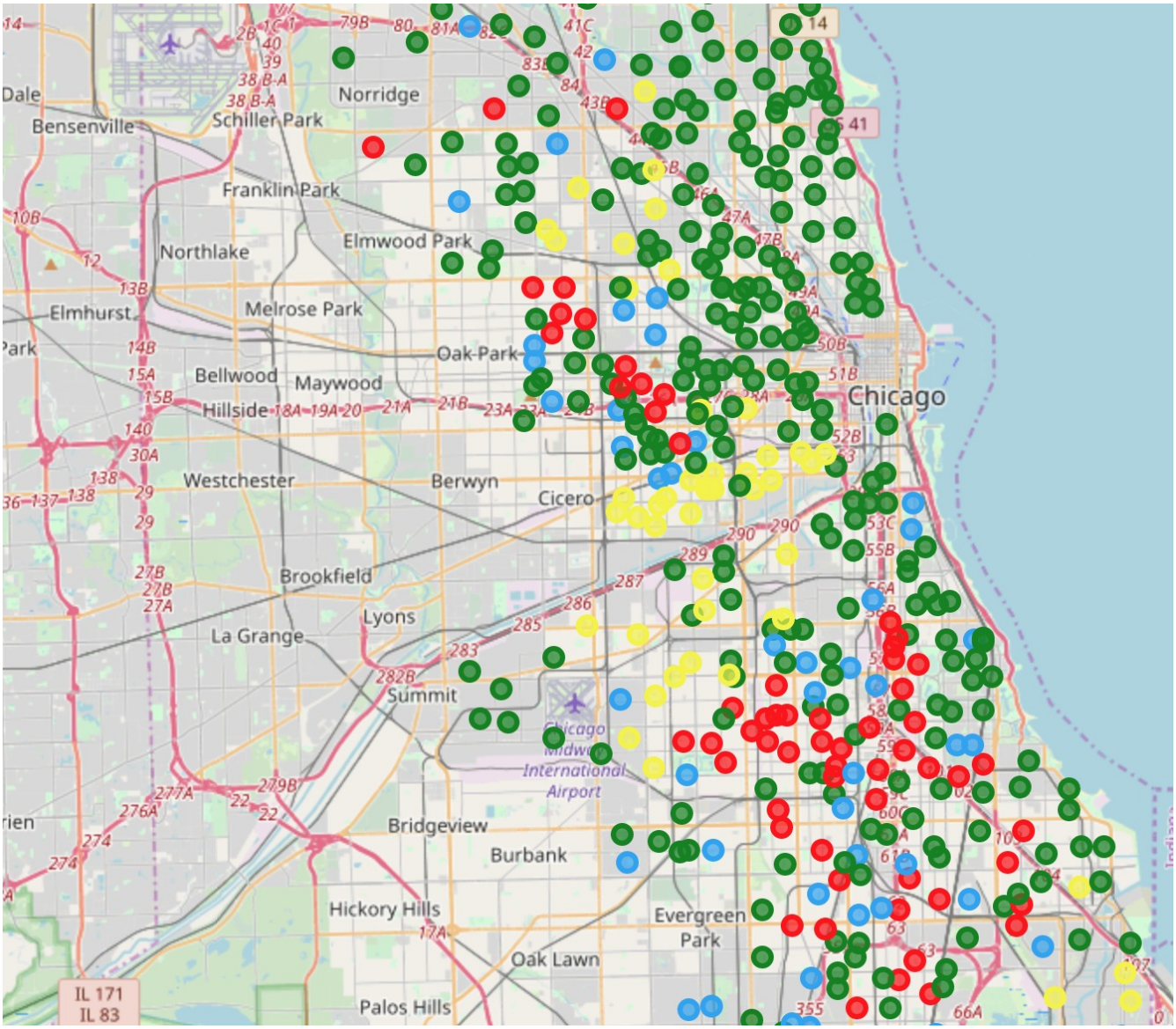
These venue frequency scores were then used to run a kmeans analysis to cluster schools based on frequency of nearby venue categories, the results of which follow.

Clustering Analysis

Once the clusters were created and schools were labeled, schools were grouped by clusters and a few metrics were averaged over cluster members. The hope was to find some variation in metrics between clusters, but overall they were fairly homogeneous.

4. Results

The schools were placed on a map and color coded according to their cluster label, resulting in the following figure:



The cluster labels and the resulting cluster sizes were:

Cluster	Count
1	47
2	62
3	50
4	271

The cluster metrics table is as follows:

Cluster Label	Instruction Score	Teachers Score	AVERAGE_STUDENT_ATTENDANCE	Rate of Misconducts (per 100 students)
0	44.642857	43.785714	95.878571	12.275000
1	53.090909	46.969697	93.357576	29.254545
2	52.368421	46.789474	94.278947	22.863158
3	52.331081	51.601351	94.675676	17.150676

5. Discussion

The clustering was successful, identifying 5 unique clusters. One of the limitations of kmeans clustering is that it is difficult to score the usefulness of the separation, or to figure out how many clusters should be used. An interesting observation is that even though the algorithm ran looking for 5 clusters, only 4 were identified. This does show some natural segmentation in the data, or that the schools fall neatly into four categories.

Recommendations are hard to make based on this data. It would be valuable to policymakers to see if there are any correlations between cluster labels and performance. While the metrics chosen did not show much variance based on cluster, there may be other factors that do.

Overall, further analysis with additional datasets is needed to identify truly compelling trends in this data.

6. Conclusion

We have shown that there is a natural segmentation of Chicago schools based on the surrounding venues, and that there are 4 distinct segments. We have also shown that these surroundings are not a major factor in determining student attendance or in predicting teaching or instruction scores. The rate of misconducts did seem to vary, but that may be due to outliers.