Classification of Chicago Schools

Introduction and Statement of Problem

- Schools vary based on the neighborhood that they are embedded in
- It should be possible to use machine learning to segment schools based on surrounding venues
- This classification information should help policymakers tailor iad to different schools based on different needs

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.

Data Sources

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 within 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.

Methodology

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

One Hot Encoding and Group Means

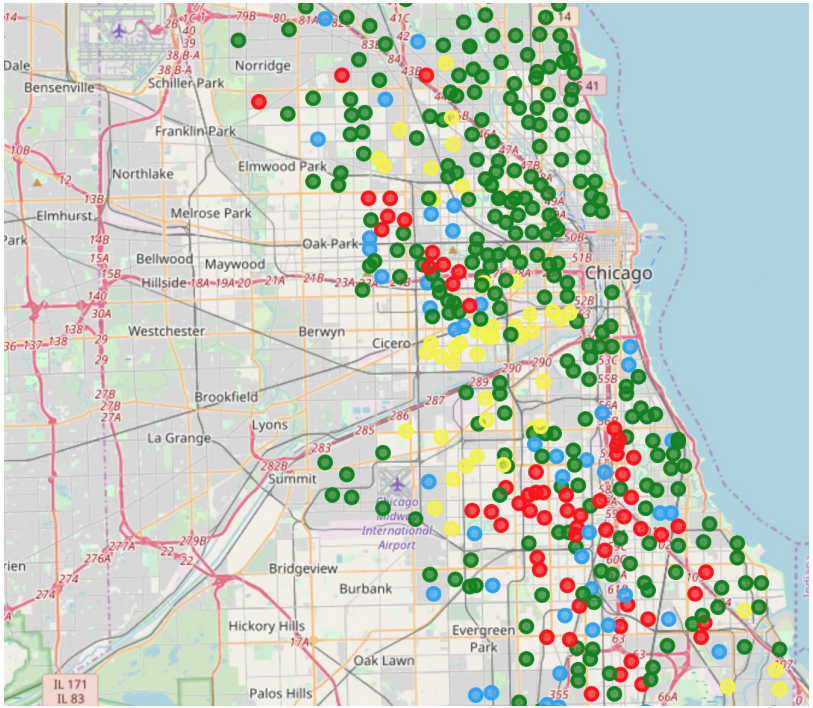
	School	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Antique Shop	Arcade	...	Vineyard	Warehouse	Warehouse Store	Weight Loss Center	Wh
0	A.N. Pritzker School	0.0	0.03	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.0	0.0	0.0	
1	Abraham Lincoln Elementary 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	

Results

A map was generated showing the locations of schools, colored according to their cluster labels.

Also, the schools were grouped by clusters and a table was created with the cluster means for various school performance metrics

Schools, Colored By Cluster



Cluster	Count
1	47
2	62
3	50
4	271

School Performance Metrics By Cluster

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

Discussion

- There was a nice distribution among the clusters. One cluster dominated , but the others were well distributed. That can be taken as evidence that there was some meaningful differentiation.
- The mean metrics by cluster were not as differentiated as might be hoped. This means that the amount of actionable insight gained is also minimal.
- It would be good to take this clustering data and analyze it by combining it with other datasets.