Elementary Schools in Santa Clara County as Clusters

Introduction

Problem
Compare schools based on various characteristics such as performance, gender, economic criteria, ethnicity of students and cluster similar schools together.

Interest
Grouping schools with similar characteristics will help in understanding the influencing criteria. This will help in identifying similar schools in a different neighborhood which may be an easier commute or may have more affordable housing or may be in a community of choice for any reason.

Data Acquisition and Cleaning

Data Sources
California Department of Education website has links to statewide public school data as research files that are downloadable as "Research Files"
Link: https://caaspp.cde.ca.gov/sb2018/ResearchFileListCAA#accurate-results
Here we filtered our selection to Santa Clara county and Year 2018 and downloaded the school performance file. File downloaded as "sb_ca2018_all_43_csv_v3.csv"

There is an entity file(csv) available for all schools which has the County, District, and School entity names and codes for all entities existing in the administration year selected. This file can be used as lookup to the performance data file. File downloaded as "sb_ca2018entities.csv"

Data Processing

A. Entities File
Entities file has data for Counties, Districts and Schools. It has the names and codes of all these entities and their zip-codes for state-wide public schools.

A.1 Data Cleansing
First filtered data to Santa Clara County (County Code=43) as the target county for analysis
Next, deleted rows that had details of school districts and counties(i.e. rows with no School Name) because we need address details of individual schools.

This gave us a clean data set with no blank rows. Result is a dataset of 436 rows.

A	B	C	D	E	F	G	H	I	J
County Code	District Code	School Code	Filler	Test Year	Type Id	County Nam	District Name	School Name	Zip Code
0	0	0		2018	4	State of California			
1	0	0		2018	5	Alameda			
1	10017	0		2018	6	Alameda	Alameda County	Office Of Educ	
1	10017	112607		2018	9	Alameda	Envision Acaden	Envision Acade	94612
1	10017	123968		2018	9	Alameda	Community Sch	Community Sc	94606
1	10017	124172		2018	9	Alameda	Yu Ming Charter	Yu Ming Chart	94608
1	10017	125567		2018	9	Alameda	Urban Montessc	Urban Montes	94619
1	10017	130401		2018	7	Alameda	Alameda County	Alameda Cour	94578
1	10017	130419		2018	7	Alameda	Alameda County	Alameda Cour	94544
1	10017	131581		2018	9	Alameda	Oakland Unity M	Oakland Unity	94605
1	10017	136101		2018	9	Alameda	Connecting Wat	Connecting W	95386
1	10017	136226		2018	10	Alameda	Alameda County	Opportunity A	94601
1	10017	6001788		2018	9	Alameda	Cox Academy	Cox Academy	94603
1	10017	6002000		2018	9	Alameda	Lazear Charter A	Lazear Charter	94601
1	31609	0		2018	6	Alameda	California School	For The Blind	
1	31609	131755		2018	7	Alameda	California Schoo	California Sch	94536
1	31617	0		2018	6	Alameda	California School	For The Deaf-	
1	31617	131763		2018	7	Alameda	California Schoo	California Sch	94538
1	61119	0		2018	6	Alameda	Alameda Unified		
1	61119	106401		2018	7	Alameda	Alameda Unifiec	Alameda Scier	94501
1	61119	111765		2018	7	Alameda	Alameda Unifiec	Ruby Bridges E	94501
1	61119	119222		2018	9	Alameda	Nea Community	Nea Communi	94501

Fig 1: Entities file with sample data

A.2 Get School Location Data

School Name can be used to get location data i.e. Latitude, Longitude. Foursquare did not have this data unfortunately so I used google maps Geocoding API to get the required location data.
The location data was merged with the entities data. This formed the basis of mapping visualizations later after clustering.

Schools Performance Data

B. Schools Data file

The data file contains 2018 test score and stats for each test and rows for each test, school, Grade and student subgroup. These are explained below.

Students take multiple tests such ELA, Math which have unique Test Ids.
Each school administers tests for different grades that it offers. The data at school level is represented by Grade Id 13.
Data for each Student Subgroup is available as well as for all students tested. Subgroups identify sub-totals by Gender, Ethnicity, Economic status, English Language Fluency, Parents Education level, Immigration status etc. Fig.2 show the list of subgroups.

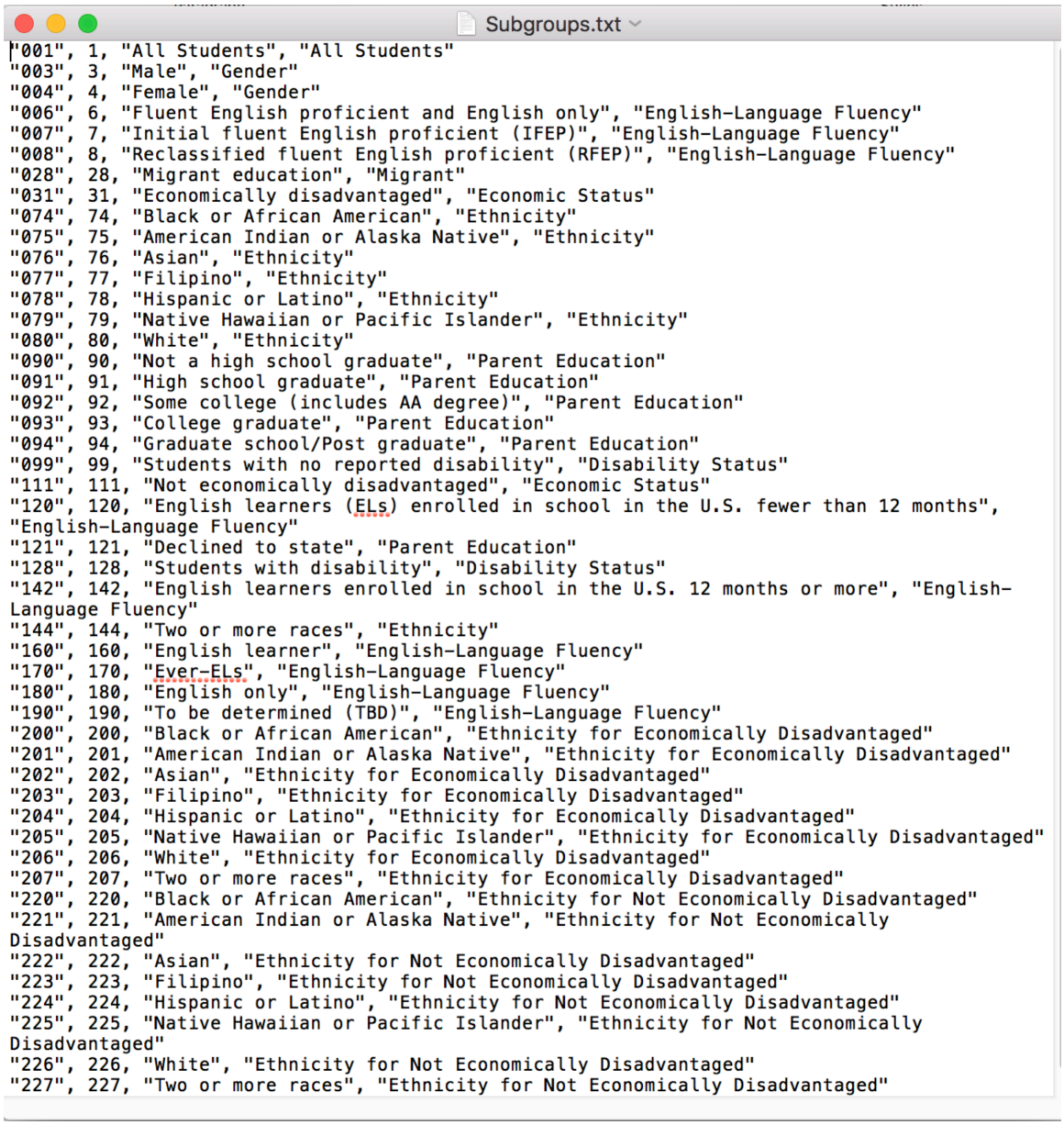

Fig 2: Student Subgroup Definitions

The file also contains rows of data for aggregates at School Total, District Total, County Total for each Test id

The data is suppressed where the population of any row is less than 5.

B.1 Data Cleansing

Dropped 8 columns of scores at detail level

Dropped all rows with summary values above school level, i.e. rows with School Code as Null

Filtered data to limit to Grade 5 and Math test to reduce the data size and get a representative sample

Dropped Columns that are not needed i.e. County Code (which is 43 always), Test Year(2018), Test type(B), Total tested at entity level(which is a summary), Total tested with score(summary),CAASSP reported enrollment
This resulted in a dataset with 9106 rows and 12 columns.

B.2 Approach

Target is to get one row for each school with features(columns) as Students Tested, Mean Score and Percentages above and below Passing grades, with counts of students per Subgroup by Gender, Economic Status, Ethnicity, Immigration Status, English Fluency level. Want to measure the impact of these numbers on total score, so did not need scores per Subgroup, only the count of students.
Other sub-groups that breakdown ethnicity further into economic status, as well as Parents education levels, disability status were dropped from analysis at this time to understand impact of selected features. These could be added back if the cluster definition is not clear enough later.

B.3 Data Normalization

In order to equalize the impact of magnitude differences between various features, the following normalization techniques were used.

a. Mean Scale Score: Normalized using Min-Max feature scaling
b. All Subgroup counts were expressed as percentage of Students tested

In addition, the SK Learning StandardScaler was applied to normalize as per best practice before K-Means Clustering

B.4 K-Means Clustering

K-Means clustering was applied with 6 Clusters (after trying from 2-8 clusters) with 24 features and for 265 Schools. We have so far worked with data which has only school codes and have not identified school names, districts, zip codes etc. Wanted to add these at the end as I had a hypothesis that I had formed (being a resident of this area).

B.5 Exploring Clusters

The resulting clusters are summarized as follows:
Cluster 0(Red): Moderate score, English learners, economically disadvantaged, Hispanic majority
Cluster 1(Indigo): High performance schools with high English proficiency, wealthy, majority White, with multi-race ethnicity. Crowded schools
Cluster 2(Blue): Lowest performing, English learners, economically disadvantaged, Hispanic majority

Cluster 3(Moss): Moderate score, low-moderate income, multi-ethnic (Asian, Filipino, Hispanic)
Cluster 4(Green): Highest performance, well-off, Asian majority, Crowded
Cluster 5(Orange): Moderate score, low female population, moderate economically disadvantaged, Hispanic/White majority. Likely rural

B.6 Cluster visualization

The school names and locations are now merged with the school cluster data and put on a folium map (Fig 3)

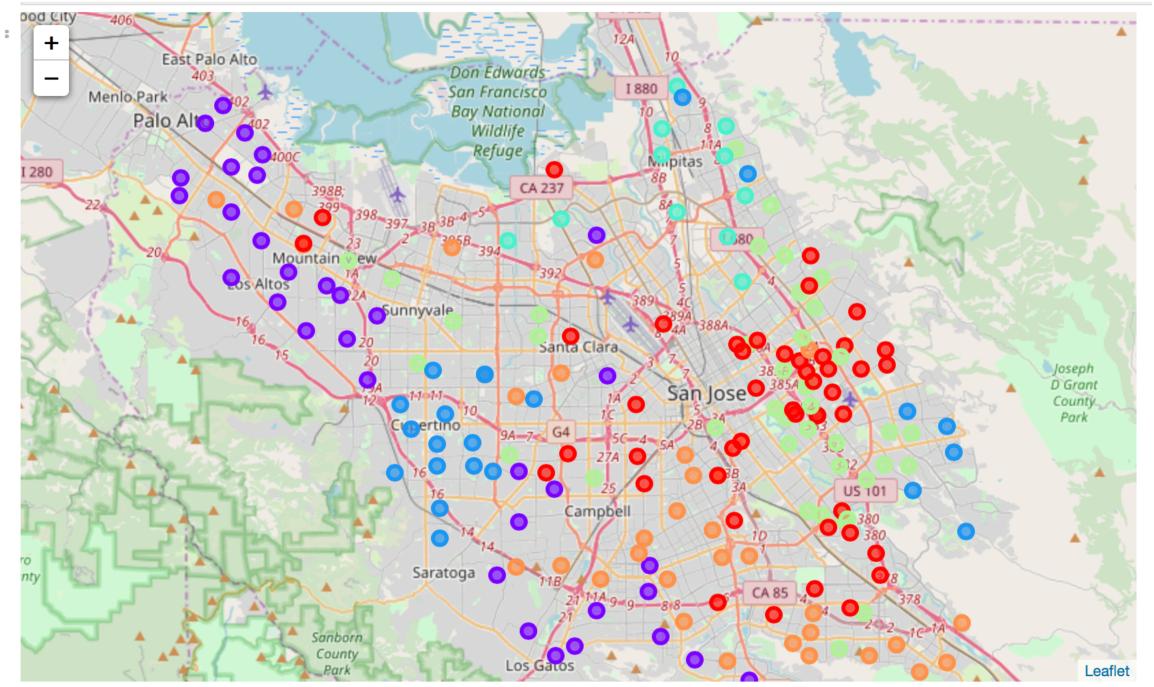

Fig 3: Cluster visualization of schools based on 24 features

Conclusions

Factors that influence clustering most seem to be Mean Scale Score, Ethnicity, English Proficiency, Economic status which may all be inter-related.
Gender, Immigration status don't seem to be influencers with this dataset.

Future directions

Many different directions that can be pursued from here. Some of them are discussed below:

1. Take ELA scores and compare the clusters with Math clusters
2. Filter on different grades and compare clusters
3. Add Parent education levels as features and do the exercise again

4. Overlay commute distances for various schools within a selected cluster from a certain point
5. Analyse schools within each cluster with median home prices for the zipcode and find attractive neighborhoods.