Schools in San Mateo County as Clusters

1. Introduction: Problem

In this project, we will explore the schools in a neighborhood by comparing them based on various features. Specifically, this exercise is to group similar schools in **San Mateo county in California.**

We will use various features such as **grade level, gender, tests performance, economic criteria, ethnicity of student and parent education** etc., to create the school clusters. There are many different choices of selecting certain combinations of attributes, this grouping exercise of schools may help in identifying the *influencing criteria* of such characteristics.

This study would help in making certain decisions based on individual's choice, such as moving to San Mateo county in California, for either easier commute or may have more affordable housing or ethnic culture or for any other personal and/or business reasons.

2. Data

We need the following data items, based on the above problem description:

- List of schools and locations in county of San Mateo, California
- Tests performance data of the schools for different grades, gender, ethnicity and various other attributes
- The geolocation data i.e. Latitude, Longitude for each school in San Mateo

The following data sources will be needed to extract the required information:

1. The **"Research Files"** at California Department of Education for Smarter Balanced Assessments data can be downloaded as csv files. These research files contain results from the administrations of the California Assessment of Student Performance and Progress (CAASPP) Smarter Balanced Assessments. As per the website, these research files require two primary tables, the *entities* and the *test data*.
2. **Entities:** lists the County, District, and School entity name, code and zip-code for all entities as the existed in the administration year selected. This file must be merged with the research file to join these entity names with the appropriate score data.
3. **Tests data:** comprised of the school, district, county, and state aggregate CAASPP counts and scores.
4. **Other supporting data:** can be downloaded from the above website, the *Student Group ID,* identifies each demographic student group and ID reported in the

CAASPP results. The *Test ID,* each student will take a number of tests and a specific test should be selected during analysis.

5. **Geocoding data:** obtain the school location data i.e. Latitude, Longitude based on the name of the school and zip-code, using Google Maps API geocoding.

Entities - California Schools data

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

County Code	District Code	School Code	Filler	Test Year	Type Id	County Name	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 Education		
1	10017	112607		2018	9	Alameda	Envision Academy For Arts & Technology	Envision Academy For Arts & Technology	94612
1	10017	123968		2018	9	Alameda	Community School For Creative Education	Community School For Creative Education	94606
1	10017	124172		2018	9	Alameda	Yu Ming Charter	Yu Ming Charter	94608
1	10017	125567		2018	9	Alameda	Urban Montessori Charter	Urban Montessori Charter	94619
1	10017	130401		2018	7	Alameda	Alameda County Office Of Education	Alameda County Juvenile Hall/Court	94578
1	10017	130419		2018	7	Alameda	Alameda County Office Of Education	Alameda County Community	94544
1	10017	131581		2018	9	Alameda	Oakland Unity Middle	Oakland Unity Middle	94605
1	10017	136101		2018	9	Alameda	Connecting Waters Charter - East Bay	Connecting Waters Charter - East Bay	95386
1	10017	136226		2018	10	Alameda	Alameda County Office Of Education	Opportunity Academy	94601
1	10017	6001788		2018	9	Alameda	Cox Academy	Cox Academy	94603
1	10017	6002000		2018	9	Alameda	Lazear Charter Academy	Lazear Charter Academy	94601

In this project, we are only interested in San Mateo County in California, which is "County Code = 41". Therefore, after loading the provided "sb_ca2018entities.csv" file, filter for San Mateo county. Next, delete the rows with no School Name (these are of Type ID -> 4=State, 5=County, 6=District), and we need only Schools. The result is a dataset of 197 rows (schools) in San Mateo County.

Geocoding data for San Mateo Schools

Let's find the latitude & longitude for San Mateo Schools, using Google Maps geocoding API. Unfortunately, Foursquare did not have this location data. We can't just use the School Name since the School Name may be repeated in different districts, such as "Hoover Elementary" in "Burlingame Elementary" district and also in "Redwood City Elementary" district, however, they have different zip-codes.

Therefore, for Google Maps geocoding API, use the combination of School Name and Zip-Code to uniquely identify a school, to avoid the above described issue. Alternatively, use the provided "SM_Schools_Geospatial_Cordinates.csv" file, has the School Code, Latitude and Longitude for the 197 schools in San Mateo county.

Now, merge the schools' data frame and the geospatial data frame on "School Code" so that we have the Latitude and Longitude for every school in San Mateo county.

Tests Data - San Mateo Schools Smarter Balanced tests performance data for 2018

The data file contains Smarter Balanced tests performance scores and stats for 2018 for each test and rows for each test, school, Grade and student subgroup etc., for San Mateo Schools. Students take multiple Smarter Balanced tests such as English Language Arts (ELA), Mathematics which have unique Test Ids. Each school administers tests for different grades that the school offers. The data at school level is represented by Grade Id 13. The file also contains rows of data for aggregates at School Total, District Total, County Total for each Test id. In order to protect student confidentiality, no scores are reported (or included in the research files) for any group of 10 or fewer students.

County Code	District Code	School Code	Filler	Test Year	Subgroup ID	Test Type	Total Tested At Entity Level	Total Tested with Scores	Grade	Test Id	CAASPP Reported Enrollment	Students Tested	Mean Scale Score	Percentage Standard Exceeded	Percentage Standard Met	Percentage Standard Met and Above	Percentage Standard Nearly Met	Percentage Standard Not Met	Students with Scores	Area 1 Percentage Above Standard	Area 1 Percentage Near Standard	Area 1 Percentage Below Standard	Area 2 Percentage Above Standard	Area 2 Percentage Near Standard
41	0	0		2018	1	B	48680	48586	3	2	6894	6810	2460.2	34.53	27.71	62.24	19.19	18.57	6806	47.3	30.22	22.49	39.96	38.44
41	0	0		2018	1	B	48406	48245	3	1	6894	6736	2447.1	35.77	22.82	58.59	20.93	20.47	6731	34.66	43.25	22.09	31.48	43.58
41	0	0		2018	1	B	48406	48245	4	1	7157	6999	2490.7	37.19	22.74	59.93	16.31	23.76	6996	34.7	44.44	20.86	33.85	42.55
41	0	0		2018	1	B	48680	48586	4	2	7157	7058	2498.7	31.17	26.27	57.45	24.5	18.05	7057	43.33	27.63	29.04	35.1	41.41
41	0	0		2018	1	B	48680	48586	5	2	7124	7041	2523.4	32.88	17.95	50.82	22.82	26.36	7038	39.33	26.14	34.53	32.45	39.29
41	0	0		2018	1	B	48406	48245	5	1	7125	6978	2523.8	32.82	28.12	60.94	16.36	22.7	6974	35.26	42.25	22.49	40.17	37.53
41	0	0		2018	1	B	48406	48245	6	1	7376	7215	2542.9	26.23	31.9	58.13	20.4	21.47	7210	30.43	41.12	28.45	34.4	40.75
41	0	0		2018	1	B	48680	48586	6	2	7376	7271	2541.7	30.24	19.86	50.1	22.57	27.33	7266	36.52	29.27	34.22	30.4	39.31
41	0	0		2018	1	B	48680	48586	7	2	7182	7058	2567.5	32.23	21.1	53.33	22.19	24.48	7030	39.93	28.96	31.12	33.95	40.2
41	0	0		2018	1	B	48406	48245	7	1	7183	7017	2575.7	25.65	36.77	62.42	19.43	18.15	6932	34.26	41.25	24.49	40.34	42.93
41	0	0		2018	1	B	48406	48245	8	1	7261	7085	2594.2	26.76	36.63	63.4	19.98	16.62	7081	35.45	40.98	23.57	38.48	42.89
41	0	0		2018	1	B	48680	48586	8	2	7261	7119	2587.4	35.35	18.03	53.38	19.75	26.87	7109	39.7	28.93	31.37	37.13	40.22
41	0	0		2018	1	B	48680	48586	11	2	6965	6323	2601.5	22.04	21.8	43.84	21.37	34.79	6280	32.99	25.36	41.64	25.9	41.21
41	0	0		2018	1	B	48406	48245	11	1	6967	6376	2616.5	34.79	28.89	63.68	18.15	18.18	6321	39.08	41.81	19.11	43.24	35.71
41	0	0		2018	1	B	48406	48245	13	1	49963	48406		31.22	29.76	60.98	18.8	20.22	48245	34.76	42.15	23.09	37.35	40.91
41	0	0		2018	1	B	48680	48586	13	2	49959	48680		31.33	21.78	53.11	21.79	25.1	48586	39.93	28.11	31.96	33.64	40
41	0	0		2018	3	B	24854	24764	3	1	3572	3479	2437.2	32.49	22.43	54.92	20.93	24.15	3478	31.68	43.19	25.12	27.61	43.63
41	0	0		2018	3	B	25021	24969	3	2	3572	3526	2461.5	36.26	26.18	62.44	18.47	19.09	3525	49.26	28.09	22.64	41.99	36.34
41	0	0		2018	3	B	25021	24969	4	2	3686	3636	2499.1	32.43	25.23	57.66	23.55	18.79	3635	45.13	24.96	29.91	36.07	39.56

After loading the provided "sb_ca2018_all_41_v3.csv" file, let's delete all the rows with School Code = 0 (these are Summary rows), and I chose to delete the rows (schools) where couldn't get the lat & long using the Google Maps API (as mentioned in the above **"Entities - California Schools data"** section under "No Coordinate Schools"). [Since I am using the provided geospatial coordinates file, all schools have the lat & long]

Also, filtered data to limit to 8th Grade and Math test to reduce the data size and get a representative sample. Additionally, kept only 12 columns of interest.

3. Methodology

In this project, we will cluster the San Mateo schools based on the tests performance data and few selective Student Group attributes.

Step 1: We have collected the required **data: Schools, Tests Scores and location** for all schools in San Mateo county.

Step 2: We can now use this data and select various student sub groups (46 different subgroup Ids) for clustering of the schools, the following section has the details on Student Groups/Subgroups IDs. Our target is to get one row for each school with features (columns) as Students Tested, Mean Score and Percentages above and below Passing grades etc., with counts of students per Subgroups. We want to measure the impact of these numbers on total score, so do not need scores per Subgroup, only the count of students. While, there may be many different ways (methods) to select various sub groups and, I just picked

the following sub group categories and respective sub group Ids, as per my personal interest:

- **Gender**
- **Economic Status**
- **Ethnicity**
- **Parent Education**

Step 3: In the final step, we will use the selected sub group attributes, normalize the data and use K-means clustering, and explore the clusters

Student Group/Subgroups IDs

Data for each Student Group (Subgroup ID) 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., They are grouped in to 10 different categories and each category has different subgroups. For example, **"Ethnicity"** has Student Groups with different (Student Group ID) Subgroup Id, such as "Black or African American"=74; "American Indian or Alaska Native"=75; "Filipino"=77; "Asian"=76 etc., Note that "All Students" has the "Subgroup ID=1", see the sample below

1	1	"All Students"	"All Students"
3	3	"Male"	"Gender"
4	4	"Female"	"Gender"
6	6	"Fluent English proficient and English only"	"English-Language Fluency"
7	7	"Initial fluent English proficient (IFEP)"	"English-Language Fluency"
8	8	"Reclassified fluent English proficient (RFEP)"	"English-Language Fluency"
28	28	"Migrant education"	"Migrant"
31	31	"Economically disadvantaged"	"Economic Status"
74	74	"Black or African American"	"Ethnicity"
75	75	"American Indian or Alaska Native"	"Ethnicity"
76	76	"Asian"	"Ethnicity"
77	77	"Filipino"	"Ethnicity"
78	78	"Hispanic or Latino"	"Ethnicity"
79	79	"Native Hawaiian or Pacific Islander"	"Ethnicity"
80	80	"White"	"Ethnicity"
90	90	"Not a high school graduate"	"Parent Education"
91	91	"High school graduate"	"Parent Education"
92	92	"Some college (includes AA degree)"	"Parent Education"
93	93	"College graduate"	"Parent Education"
94	94	"Graduate school/Post graduate"	"Parent Education"
99	99	"Students with no reported disability"	"Disability Status"
111	111	"Not economically disadvantaged"	"Economic Status"
120	120	"English learners (ELs) enrolled in school in the U.S. fewer than 12 months"	"English-Language Fluency"

Let's get various subtotal records separated, and merge them into another dataframe for final analysis. Finally, use the provided "CAA_ca_Subgroups.csv" file for mapping the Subgroup ID columns to descriptive names.

Data Normalization

In order to equalize the impact of magnitude differences between various features, normalize the data in various subgroup columns, used the following:

- Mean Scale Score: Normalized using Min-Max feature scaling
- 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

4. Analysis

Now that we have the selected features (columns) and data normalized, we can feed in to the K-means clustering algorithm. As I mentioned earlier, there may be many different ways (methods) to select various sub groups and, I just picked the following sub group categories and respective sub group Ids, as per my personal interest
of, **Gender**, **Economic Status**, **Ethnicity** and **Parent Education**

Other sub-groups that breakdown the Immigration Status, English Language Fluency, Disability status were dropped from analysis at this time to understand impact of selected features. These could be added back per the stakeholders' interests and needs or if the cluster definition is not clear.

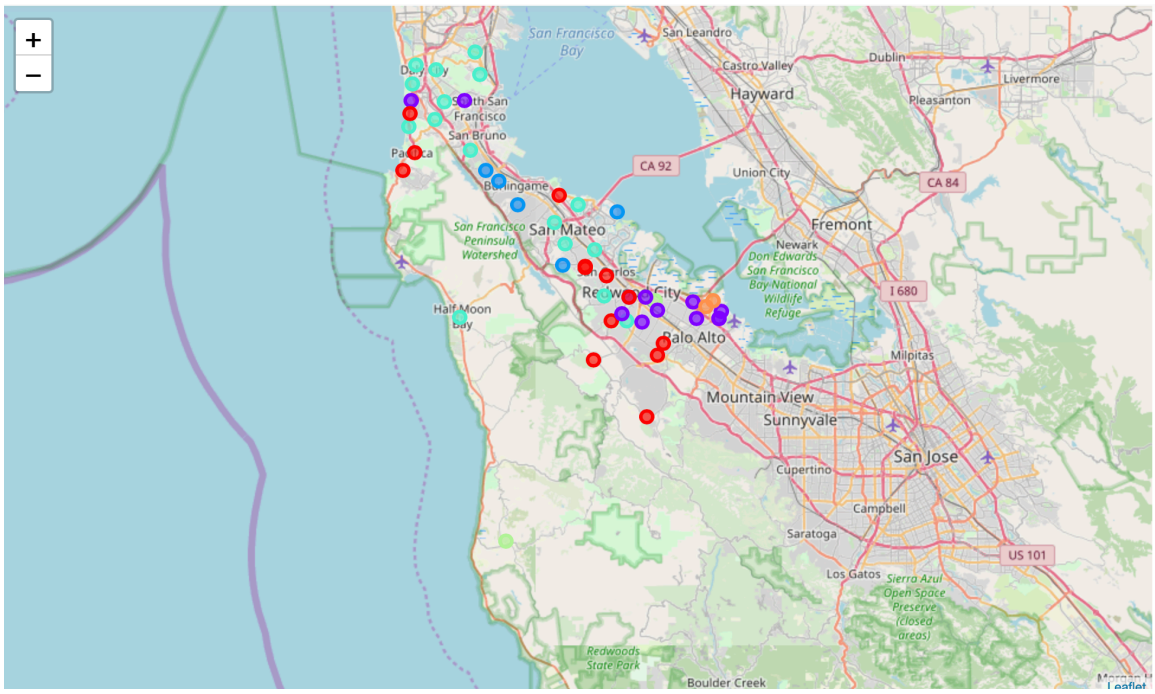
K-Means clustering was applied with 6 Clusters with 22 features (out of 26 columns) and for 51 Schools.

5. Results and Discussion

It is interesting to see clustering of San Mateo county schools based on the 8th grade tests performance and the selected subgroup features of the data, looks like the Ethnicity and Parents education background has influenced these clusters.

- Cluster 0 (Red): High performing, very well doing economically, White majority followed by Hispanic and Asian, with highly educated parents
- Cluster 1 (Indigo): Low performing, highly Economic disadvantaged, Hispanic or Latino majority, with highly uneducated parents
- Cluster 2 (Blue): High performing, very well doing economically, White majority followed by Asian with highly educated parents
- Cluster 3 (Turquoise): Medium performing, economically disadvantaged, Hispanic majority followed by White, Filipino and Asian with college or equivalent educated parents
- Cluster 4 (Madang): Low performing, economically disadvantaged, high female populated, White majority followed by Hispanic

- Cluster 5 (Sunshade): Poorly performing, high male populated, very highly economically disadvantaged, Hispanic majority followed by Native Hawaiian or Pacific Islanders with highly uneducated parents



6. Conclusion

The purpose of study was to help in making certain decisions based on individual's choice, such as moving to San Mateo county in California, for either to be part of their respective ethnic culture, easier commute or may have more affordable housing or for any other personal and/or business reasons. This project in particular does the clustering of the schools in San Mateo for the 8th grade, the factors that influenced this clustering most seem to be Mean Scale Score, Ethnicity, Economic status and Parent's educational level which may all be inter-related.

There are many possible combination of features (Subgroups) can be pursued based on individuals' preferences or interests, such as:

- Immigration status as features and do the exercise again
- English Language Arts (ELA) scores and compare the clusters with Math clusters
- Different grade levels such as 5th or 11th and compare clusters
- Commute distances for various schools from a certain point
- For home buyers, analyze with home prices for the zip-code and find attractive neighborhoods for their choice.