Schools in San Mateo as Clusters

Applied Data Science Capstone by IBM/Coursera

4/5/2020

Objective

- Problem: Group the schools with similar characteristics in San Mateo County

- Purpose: This grouping can be used to determine to choose the neighborhoods that are ideal to live in based on individuals’ choices

Data

- The “Research Files” at California Department of Education for Smarter Balanced Assessments data can be downloaded as csv files.

- Entities: ”sb_ca2018entities.csv” lists the County, District, and School entity name, code and zip-code for all entities for year 2018.

- Tests data: ”sb_ca2018_all_41_v3.csv” comprised of the tests and scores for each school in San Mateo County.

- Other supporting data: ”CAA_ca_Subgroups.csv” identifies student groups and Subgroup IDs.

- Google Maps API to get the Latitude and Longitude for each School

Methodology

- **Extract Data:**
  - Limited data to Grade 8, Math test and Students Tested more than 10
  - Selected Student Subgroups based on
    - Gender
    - Economic Status
    - Ethnicity
    - Parent Education

- **Data Transformation:**
  - Converted Subgroups from rows to columns with values as count

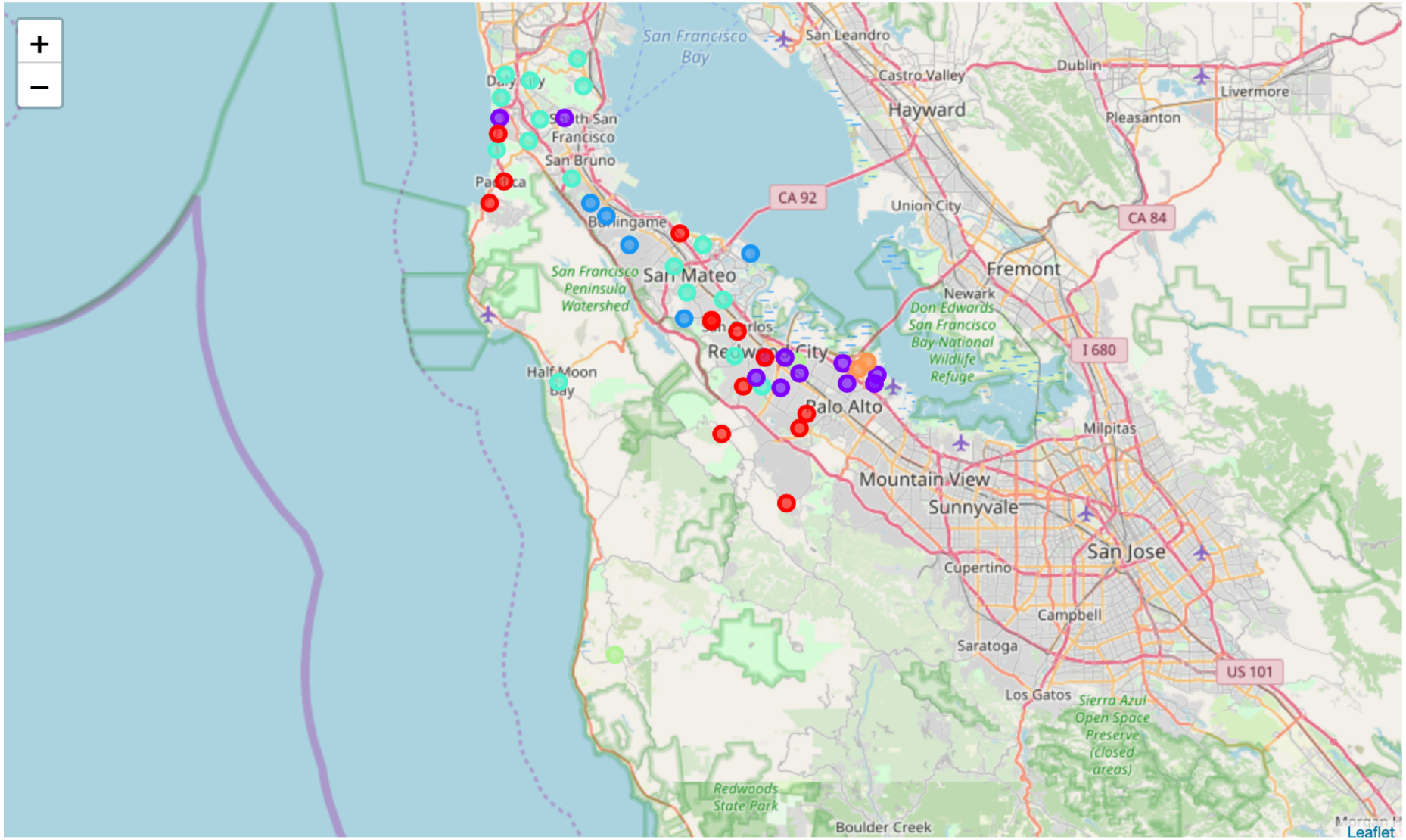
- **Data Normalization:**
  - Normalized data using Main-Max Feature Scaling and for Subgroup counts, as percentage of Total Students tested
  - Applied SK Learning StandardScaler normalization

- **Clustering**:
  - K-Means Clustering with 22 features, expressed as 6 Clusters

Explore 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

Visualization of Clusters



Conclusion & Future

- This project dealt with 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

- 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.