Clustering Schools in Santa Clara County

Coursera Capstone Project

May 2019

Objective of the Project

- **Problem:** Find Elementary schools with similar characteristics in Santa Clara County
- **Interest:** Groups of similar schools can be used to determine which neighborhoods are ideal to choose to live in

Data acquisition and cleansing

- **Data Source:** California Department of Education Research files
  - https://caaspp.cde.ca.gov/sb2018/ResearchFileListCAA#accurate-results
  - Schools data for Santa Clara County for Year 2018: sb_ca2018_all_43_csv_v3.csv
  - Entities file: sb_ca2018entities.csv
- **Filters**: School Performance data filtered
  - Dropped columns with detailed scores per subject area and took summary score per Test
  - Dropped columns with same value in all rows like County, Test Type, Test Year
  - Limited data to Grade 5, Math test

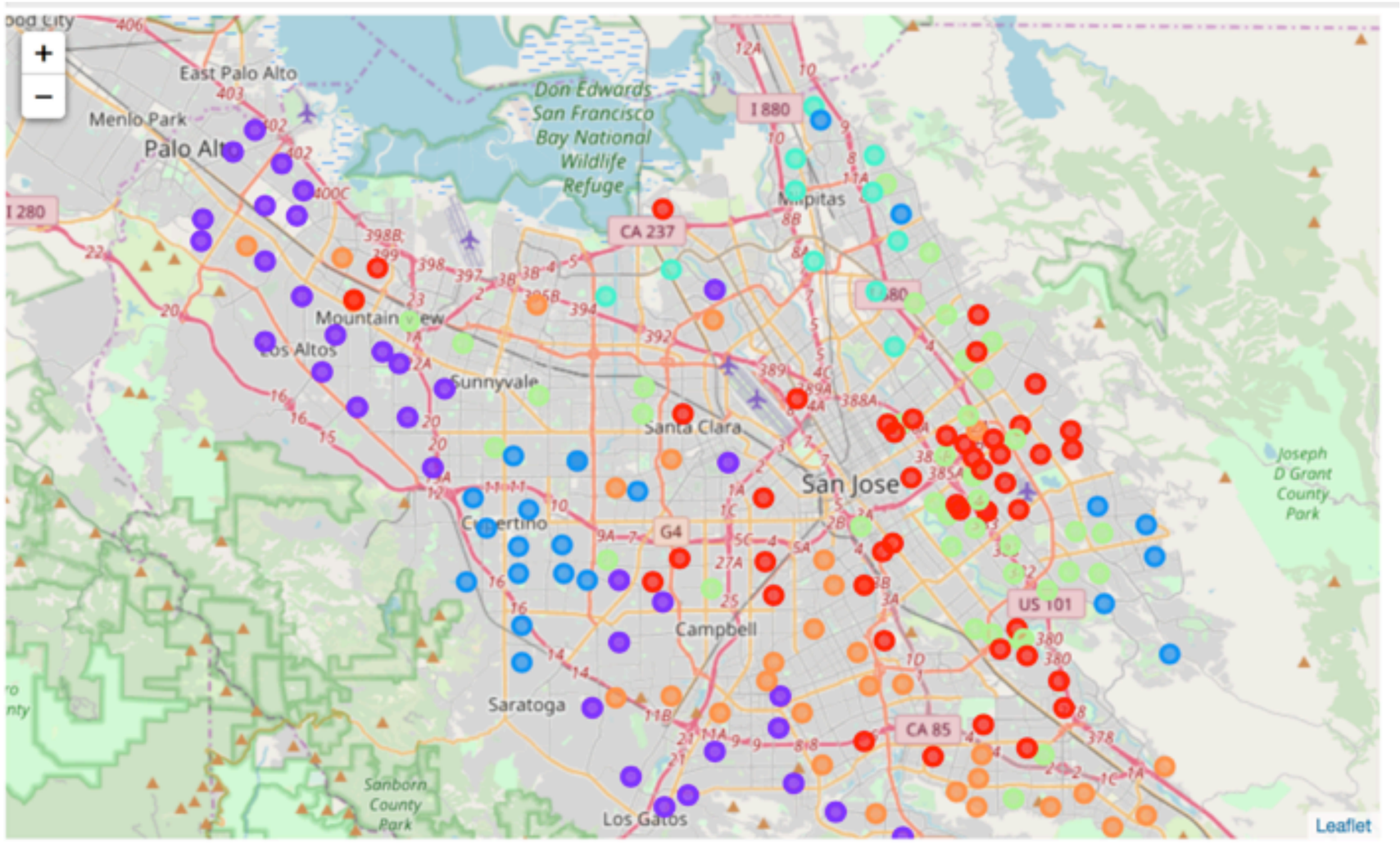
Approach

- **Data Transformation**: 265 Schools.One row per school with all features and location
  - Converted Subgroups from rows to columns with values as count using data pivot
  - Merged location and identifier data as columns
  - Added Cluster Label as column
- **Data Normalization:**
  - Mean Scale Score: Normalized using Main-Max Feature Scaling
  - Subgroup counts: Expressed as percentage of Total Students tested
  - Applied SK Learning StandardScaler normalization
- **Clustering**: K-Means Clustering with 24 features, expressed as 6 Clusters
  - Explored clusters with Mean Values of features in each cluster
  - Visualization with Folium map with School location, Cluster as color

Clusters and characteristics

Cluster Features	Cluster 0: Red	Cluster 1: Indigo	Cluster2: Blue	Cluster 3: Moss	Cluster 4: Green	Cluster 5: Orange
Mean Scale Score	Moderate	High	Lowest	Moderate	Highest	Moderate
Economic Status	Disadvantaged	Wealthy	Disadvantaged	Low-Middle income	Wealthy	Middle income
Ethnic Mix	Majority Hispanic	White majority	Majority Hispanic	Multi-ethnic: Asian, Hispanic, Filipino	Asian majority	Hispanic/White
English Fluency	Low	High	Low	Moderate	High	Moderate
Other features		Crowded Substantial Multi-race population			Crowded	Lower Female Likely rural

Where are those schools?



What’s next

- Take ELA scores and compare the clusters with Math clusters
- Filter on different grades and compare clusters
- Add Parent education levels as features and do the exercise again
- Overlay commute distances for various schools within a selected cluster from a certain point
- Analyse schools within each cluster with median home prices for the zip code and find attractive neighborhoods