The Battle of Neighborhoods

SAN FRANCISCO, CA

JON T BRODIN

Background

San Francisco, California is a densely populated US city with approximately 19 thousand residents per square mile. Population is growing at approximately 3% per year driving rising housing costs.

Problem: Transitioning to new career opportunities in an unfamiliar area requires potential movers to carefully plan to ensure they are moving into safe neighborhoods with quality schools, reasonable commuting times, and amenities that fit their lifestyle all at an affordable price.

Quality of life components considered for this project:

- Safety
- Quality schools
- Affordable housing
- Social / Cultural Venues

Data Acquisition

Public Safety: Part 1 crime data is exported from a publicly available source through application program interface (API). Crime data without a reported location was excluded from this project.

School Ratings: Data provided by an organization running the website school-ratings.com which compiles ratings for California's public schools based on the California Assessment of Student Performance and Progress (CASPP)

Housing Costs: Data was downloaded in comma separated value (CSV) file from Zillow Research based on a two-year time series of single-family residence sales grouped by zip codes.

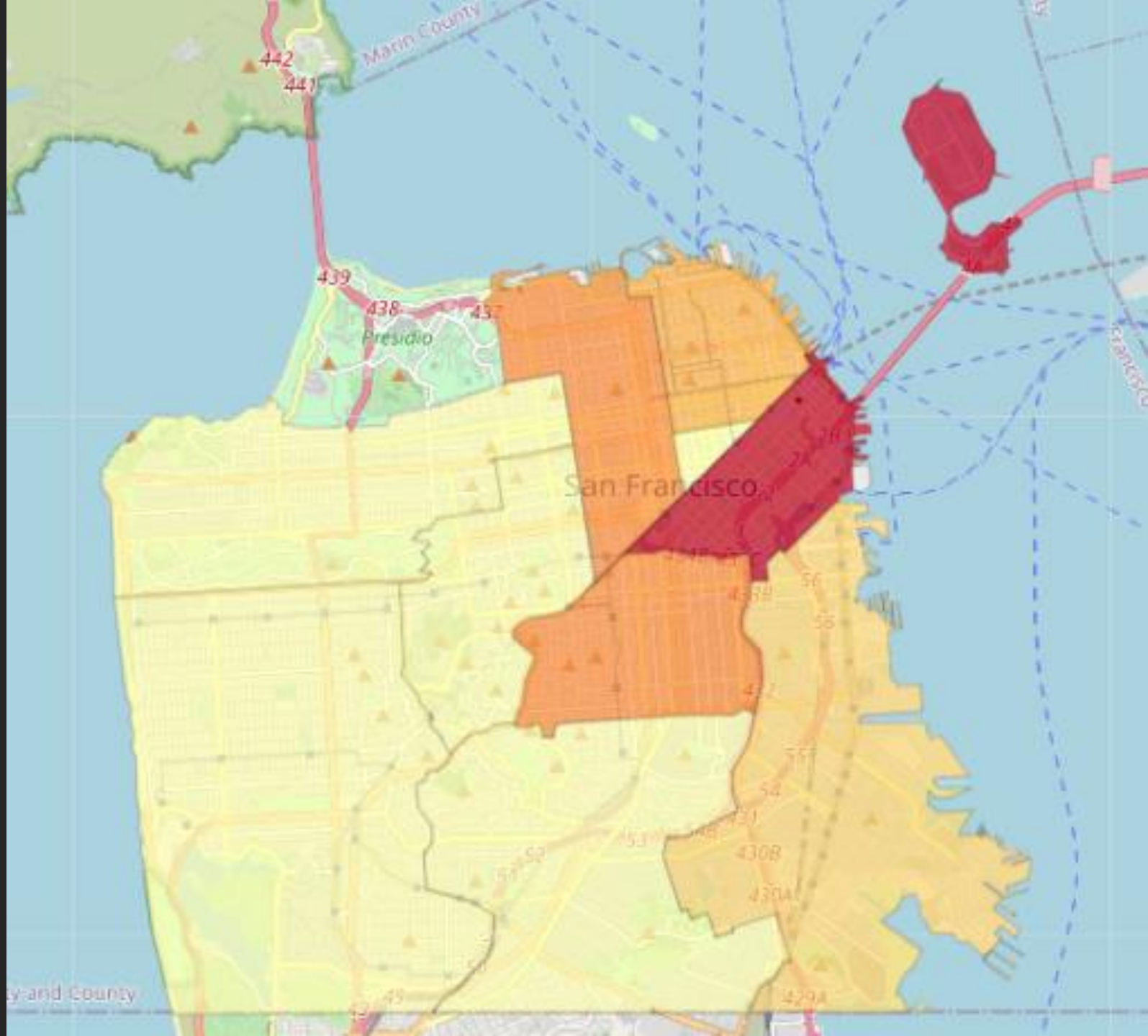
Social / Cultural Venues: The Foursquare API was used to find the most common venue categories within a neighborhood.

Public Safety

Data is read into a dataframe from a dataset on the Cognitive Class website: https://cocl.us/sanfran_crime_dataset

The data is cleansed to eliminate all crime data without a reported location. It is then grouped by neighborhood and counted.

The data is then presented as a crime density map (by neighborhood) using the folium choropleth map.



School Ratings

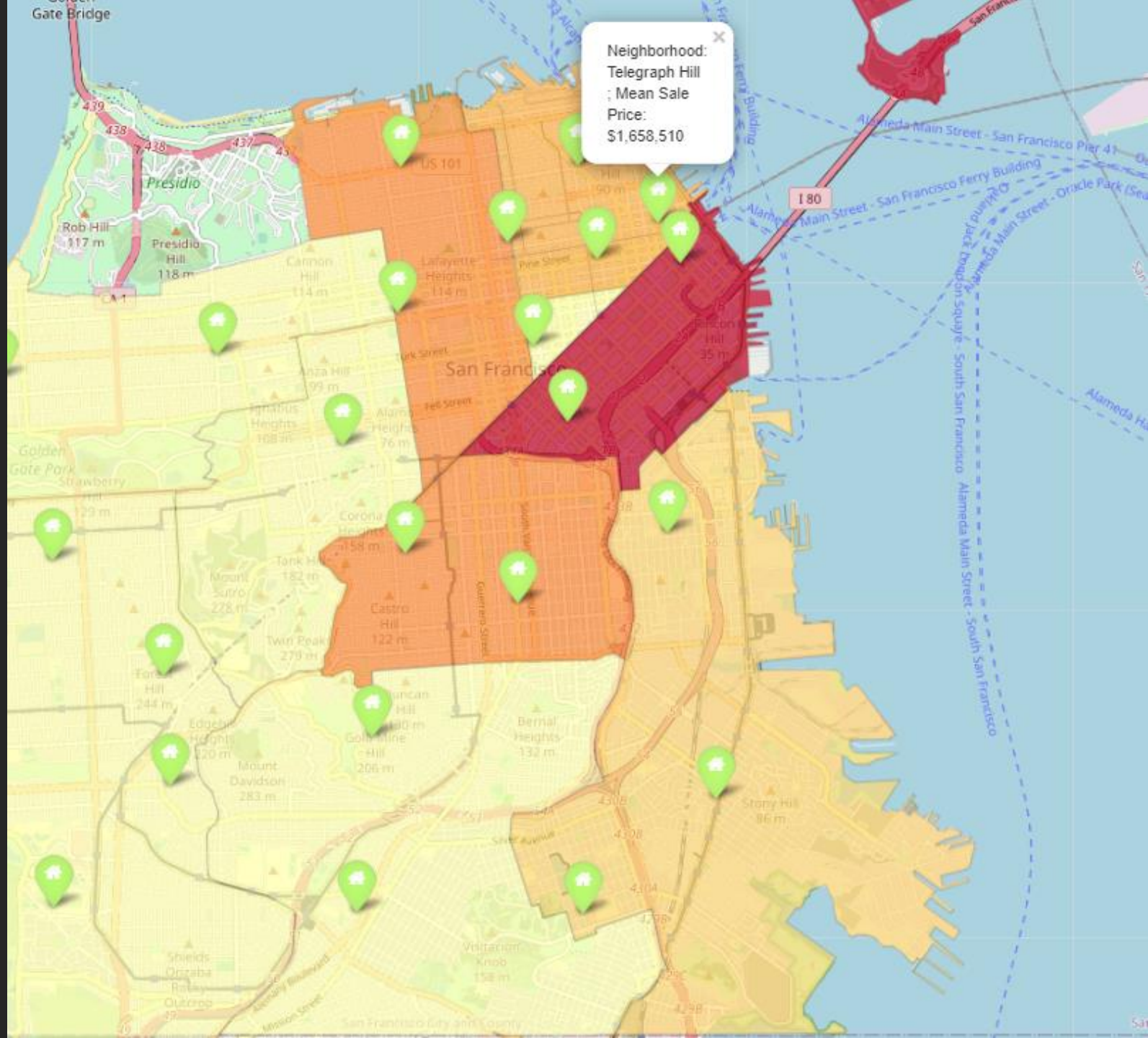
Data is extracted in csv format from https://school-ratings.com. A single "Location" column consisting of Latitude-Longitude data is split into two columns. Each record consists of one school and provides location, CSR Rating, and CSR Percentage.

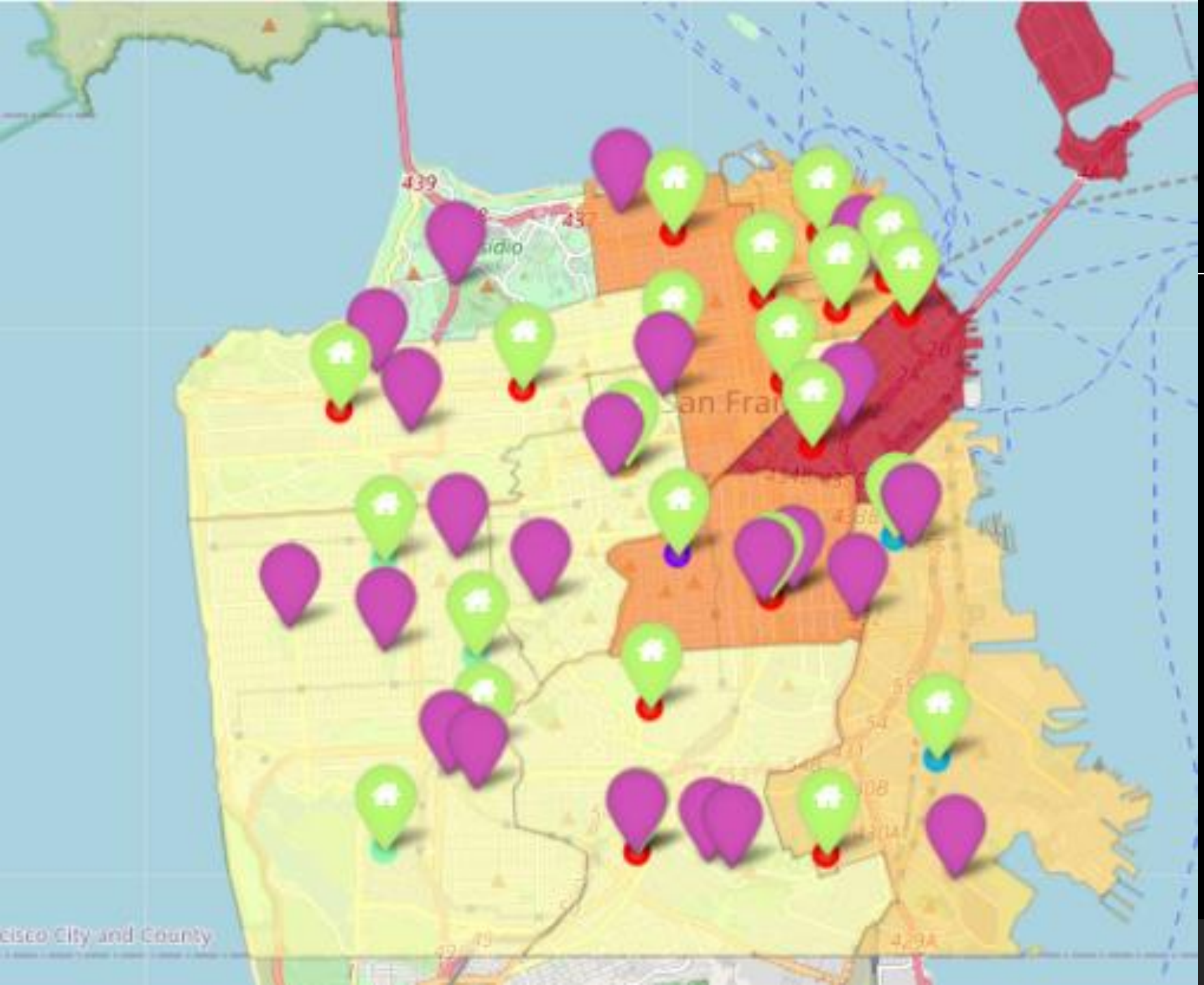
Each school record is added to the San Francisco map to provide location visualization combined with other datasets

Housing Sales

Housing data is extracted from Zillow Research as a two-year time series grouped by zip code.

Each record represents a zip code in San Francisco. A popup feature is added to each folium marker to provide details of the neighborhoods that comprise the zip code and the mean single-family residence sales price for the two-year period.





Social Venues

Leveraging the Foursquare API, venue data was pulled into a dataframe and aligned to the nearest neighborhood. Using onehot encoding, the most common venue categories were discovered for each neighborhood.

Using k-means clustering, these common venues are clustered to reduce the number of markers added to the map to ensure legibility. These cluster markers are added as another layer on top of the previous map to combine all of the data into a common view to depict the data that allows for the visualization of quality of life given these data sources.

Conclusion

The combination of data provided a more complete picture of quality of life within a neighborhood.

Challenged assumptions that more expensive homes are typically in low-crime areas.

The processes used in building this data model is repeatable and can be used in any locality offering access to similar datasets to produce visualization tools.