The Battle of Neighborhoods

Jon T. Brodin

December 7, 2020

1. Introduction

1.1 Background

San Francisco, California has a population of nearly 900 thousand people; a 11% increase since the 2010 US Census. The city occupies a total land area of 47 square miles resulting in a population density of 19 thousand per square mile. The rapid growth in population is largely attributed to San Francisco's focus on shaping itself as a center for technology and culture.

1.2 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. Real estate agents can capitalize on technology to combine data sources to paint a picture for prospective home buyers visualizing the quality of the life in terms that are important to families seeking new home ownership.

1.3 Interest

I selected San Francisco as a city as it is unfamiliar to me. Additionally, the technology industry is critical to the city's economy, so the city provides an excellent opportunity for continuing to support my current objective in establishing a career as a data scientist. To visualize an answer to our housing problem, we can create a map and chart to show potential home buyers the quality

of life available in neighborhoods of San Francisco, CA. While this project focused on one city, the process is repeatable and can be applied to any location globally provided access to appropriate data sources.

2. Data acquisition and cleansing

2.1 Data sources

Quality of life involves a standard of health, comfort and happiness experienced by an individual or group. Applying this definition to a family seeking employment in an unfamiliar location, I chose to focus on the variables of public safety, quality schools, affordable housing, and local amenities or venues that provide the cultural experiences a family may desire. Using a public application programming interface (API), part 1 crime data was extracted from a local public source. School ratings data was downloaded from California School Ratings as comma delineated text and reformatted to a comma separated value (csv) file [2]. The Foursquare API was used to discover the most common venues for each neighborhood to demonstrate the cultural or social aspects of quality of life. Housing costs were extracted from Zillow Research drawing on 2019 and 2020 sales data for single family residences by zip code [3].

2.2 Data Cleansing

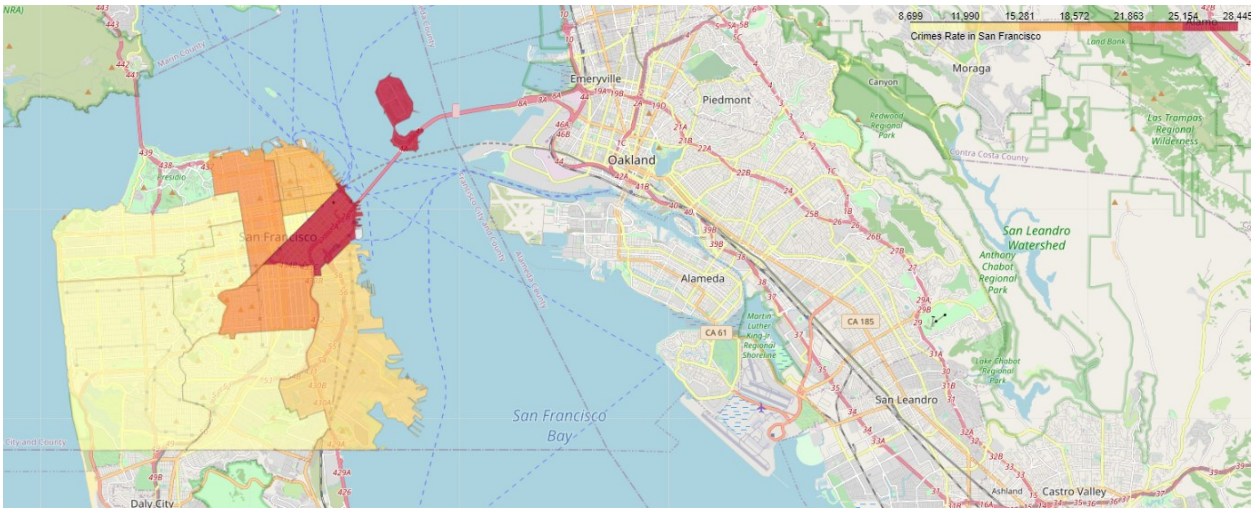
Crime data missing locations were deleted from the data set, then all reports were grouped by neighborhood and counted providing a neighborhood crime density report. School ratings reports were complete, and no cleansing was needed. However, an additional column (Summary) was created as a concatenation of the Name, CSR Rating, and CSR Percentage columns. Housing sales data for the previous two years was available in twenty-four monthly

sales averages by zip codes. An additional column was added to obtain the mean housing sales price by zip code across the two-year period.

3. Methodology

3.1 Crime data

Crime data was read into a Pandas dataframe and then grouped by neighborhood. Using the count feature, the data was organized to provide crime density by neighborhood. This allowed for visualization using a folium choropleth map to depict the crime density throughout neighborhoods in San Francisco.



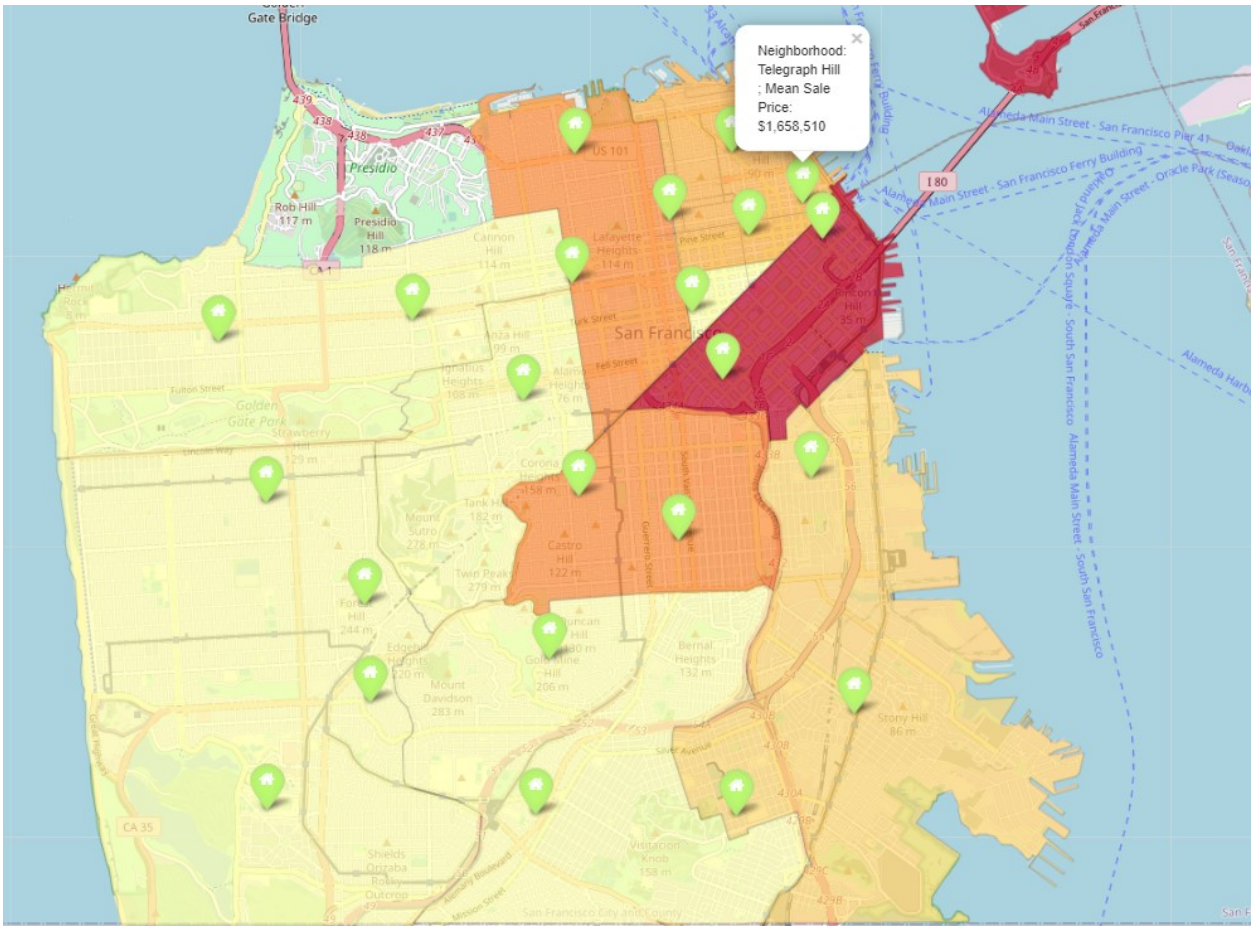
3.2 Single family residence sales data

Sales data for all single family residences from the previous two years was extracted from Zillow Research as a csv file. The mean sales price for each record was calculated in order to show a smoothed average over time and eliminate any biases that may cause for an elevated or depressed sales price.

RegionName	Neighborhood	Lat	Long	RegionType	StateName	State	City	Metro	CountyName	1/31/2019	2/28/2019	3/31/2019	4/30/2019	5/31/2019	6/30/2019	7/31/2019	8/31/2019	9/30/2019	10/31/2019	11/30/2019	12/31/2019	1/31/2020	2
94109	Central	37.792511	-122.419922	Zip	CA	CA	San Francisco	San Francisco-Oakland-Hayward	San Francisco County	$2,757,281.00	$2,765,306.00	$2,773,637.00	$2,764,385.00	$2,778,098.00	$2,799,807.00	$2,826,666.00	$2,834,046.00	$2,841,186.00	$2,851,396.00	$2,870,034.00	$2,898,795.00	$2,915,931.00	$2,93
94110	Ingleside	37.754230	-122.418610	Zip	CA	CA	San Francisco	San Francisco-Oakland-Hayward	San Francisco County	$1,696,614.00	$1,695,079.00	$1,693,486.00	$1,677,734.00	$1,674,929.00	$1,671,280.00	$1,667,903.00	$1,648,155.00	$1,638,186.00	$1,649,248.00	$1,673,969.00	$1,698,310.00	$1,709,740.00	$1,71
94122	Taraval	37.758709	-122.481102	Zip	CA	CA	San Francisco	San Francisco-Oakland-Hayward	San Francisco County	$1,485,476.00	$1,483,919.00	$1,481,127.00	$1,473,196.00	$1,481,654.00	$1,486,997.00	$1,494,662.00	$1,483,048.00	$1,472,544.00	$1,472,183.00	$1,480,729.00	$1,499,109.00	$1,502,802.00	$1,50
94112	Ingleside	37.721451	-122.440178	Zip	CA	CA	San Francisco	San Francisco-Oakland-Hayward	San Francisco County	$1,147,206.00	$1,147,349.00	$1,142,636.00	$1,138,395.00	$1,139,896.00	$1,140,745.00	$1,139,809.00	$1,132,837.00	$1,131,087.00	$1,134,520.00	$1,141,591.00	$1,150,926.00	$1,158,195.00	$1,16
94115	Northern	37.785149	-122.434761	Zip	CA	CA	San Francisco	San Francisco-Oakland-Hayward	San Francisco County	$3,656,297.00	$3,666,916.00	$3,680,132.00	$3,674,686.00	$3,718,042.00	$3,750,028.00	$3,781,587.00	$3,770,683.00	$3,765,807.00	$3,785,492.00	$3,822,653.00	$3,877,600.00	$3,905,418.00	$3,92

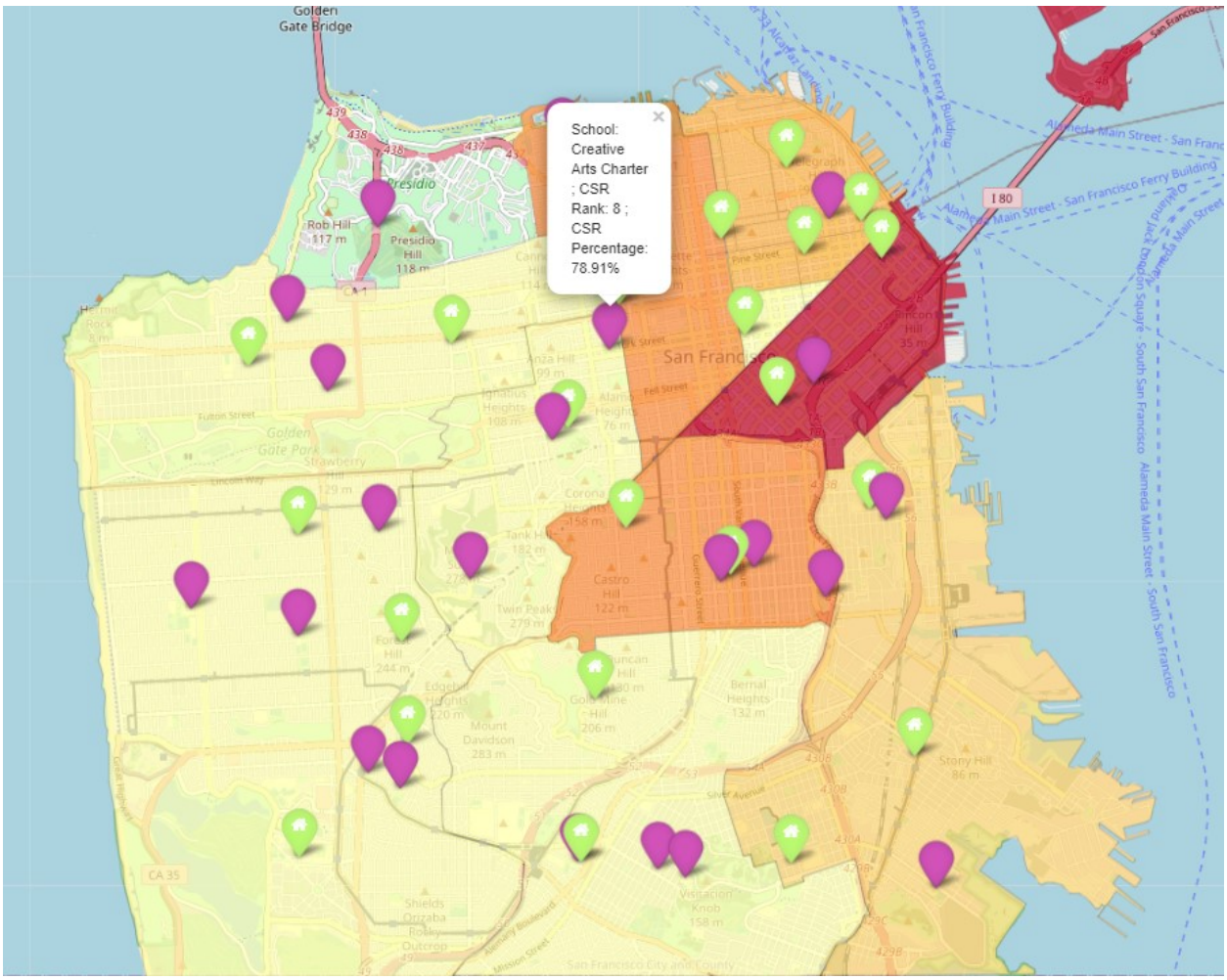
3.3 Single family residence sales data visualization

Using the marker feature in folium maps, each record representing a zip code is added as a layer on the crime density map. The popup feature provides additional information on each marker with the neighborhoods represented in the zip code and the mean sale price for a single family residence.



3.4 School data

Public school data, including publicly funded charter schools, was extracted from a public data source that documents California school ratings (CSR). An additional column was added to the dataframe as a concatenation of the school name, CSR rating, and CSR Percentage. This summary column allowed for enriching information using the popup feature on a folium map marker. Each school record, identified in purple markers, was added as a layer on the housing map.

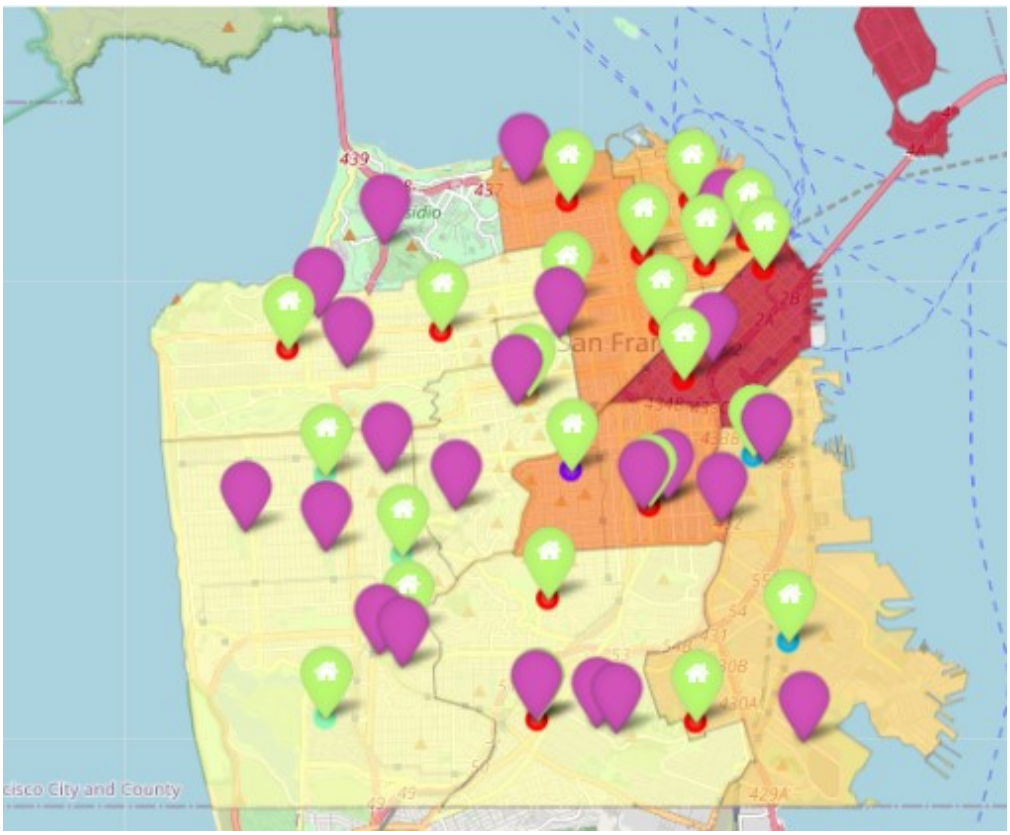


3.5 Venue data

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.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bayview	Café	Mexican Restaurant	Southern / Soul Food Restaurant	Bakery	Park	Grocery Store	Breakfast Spot	Brewery	Light Rail Station	Coffee Shop
1	Central	Coffee Shop	Hotel	Italian Restaurant	Café	Park	Gym	Gym / Fitness Center	Pizza Place	Bakery	Bar
2	Ingleside	Mexican Restaurant	Coffee Shop	Park	Latin American Restaurant	Indian Restaurant	Grocery Store	Bar	Café	Pizza Place	Playground
3	Mission	Gay Bar	Thai Restaurant	Coffee Shop	Yoga Studio	Clothing Store	New American Restaurant	Scenic Lookout	Seafood Restaurant	Indian Restaurant	Deli / Bodega
4	Northern	Gym / Fitness Center	Spa	Cosmetics Shop	Sushi Restaurant	Italian Restaurant	Bakery	Wine Bar	Gift Shop	Grocery Store	American Restaurant

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.



4. Results

The project allowed for the compilation of several disparate data sets to provide a complete picture of quality of life afforded within the neighborhoods comprising San Francisco, California. These data sets take into consideration public safety, quality education, housing costs, and social venues as the most common features responsible for ensuring quality of life.

5. Conclusion

The data challenged a few of my assumptions in that I ventured into this project with the assumption that more expensive housing would be found in the safest or least crime dense neighborhoods. That assumption was not validated by the data in that the most expensive zip code in San Francisco in terms of housing is found in the 94123 zip code that contains the Marina neighborhood which had the third highest crime density within the city. The mean sales price for a single-family residence was $4.4M over the previous two years. Similarly, a few of the least crime dense neighborhoods also enjoy some of the most affordable housing in the city. In most cases, the highest rated schools are in some of the neighborhoods with the most expensive housing. There are a few anomalies in the data as the lowest ranked school is located in the second most expensive neighborhood.

Sources

[1] https://cocl.us/sanfran_crime_dataset

[2] https://school-ratings.com/counties/San_Francisco.html

[3] https://www.zillow.com/research/data/