# Coursera Capstone Final Project – Bay Area Info

Daniel Bai

May 15, 2020

# Motivation

- Facts
  - San Francisco Bay Area – home of the Silicon Valley
  - Median housing prices of the area are a few times the national average
- Questions
  - How is one place (e. g., zip code) in the area different than another?
    - Population, income, housing price, school, ethniticy distribution, crime rate etc
  - For someone new to the area, where is the best place to live?
  - For a business owner, where is the best place to open a business?
- This project attempts to address these questions, using a subset of the data

# Data

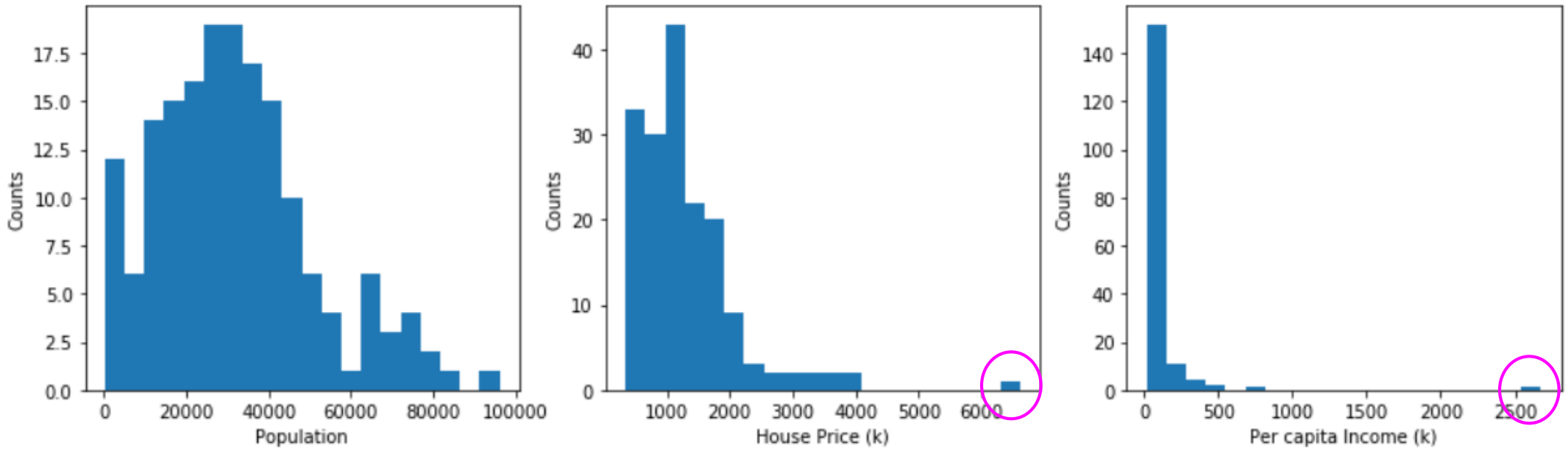
- All zip codes of the bay area
  - https://data.sfgov.org/Geographic-Locations-and-Boundaries/Bay-Area-ZIP-Codes/u5j3-svi6
- Latitudes and longitudes for all zip codes in California
  - https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/
- Population by zip code for California (2019)
  - https://www.california-demographics.com/zip_codes_by_population
- US nation-wide monthly housing price data by zip code : US (1996-2020)
  - https://www.zillow.com/research/data/
- California's by zip code tax data from IRS (2017)
  - https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2017-zip-code-data-soi
- Foursquare data via API calls to Foursquare.com
  - https://developer.foursquare.com

# Master DataFrame

- Data from multiple online sources are combined into the master pandas dataframe, which contains the zip code, city, county, and state name, longtitude and latitude, and population, housing price, and per capita income for total of 171 zip codes in the bay area

| | ZIP | STATE | Population | City | Latitude | Longitude | Metro | CountyName | House Price 2020 (k) | Total AGI (k) | AGI per capita (k) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 95620 | CA | 21854 | Dixon | 38.427208 | -121.81348 | Vallejo-Fairfield | Solano County | 454 | 705675 | 32 |
| 1 | 95476 | CA | 36792 | Sonoma | 38.277147 | -122.47058 | Santa Rosa | Sonoma County | 823 | 1834068 | 49 |
| 2 | 95140 | CA | 115 | Mount Hamilton | 37.388718 | -121.63845 | San Jose-Sunnyvale-Santa Clara | Santa Clara County | 1517 | 26968 | 234 |
| 3 | 95134 | CA | 27224 | San Jose | 37.412539 | -121.94461 | San Jose-Sunnyvale-Santa Clara | Santa Clara County | 1015 | 1889468 | 69 |
| 4 | 95035 | CA | 77562 | Milpitas | 37.436451 | -121.89438 | San Jose-Sunnyvale-Santa Clara | Santa Clara County | 1104 | 3892073 | 50 |

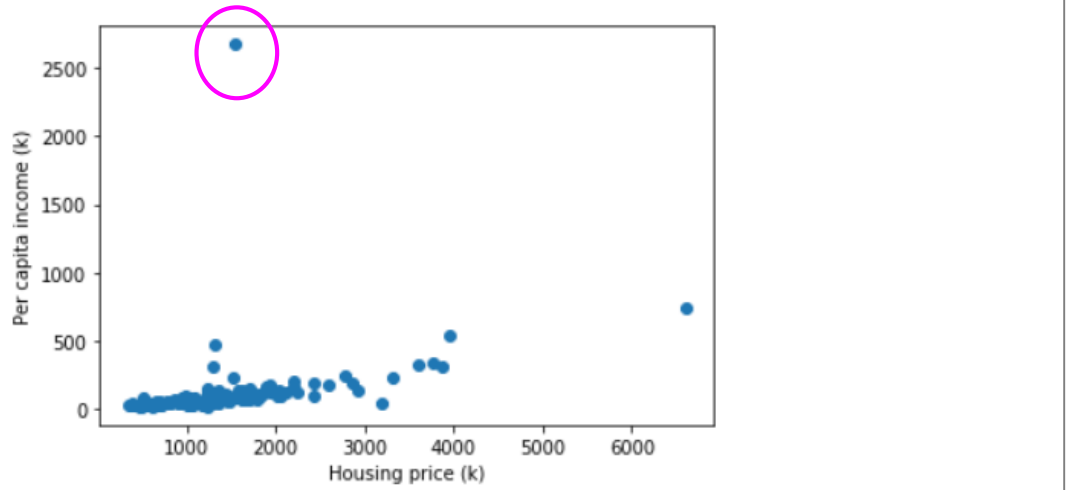
# Exploratory Data Analysis




- There is an outlier for both housing price and per capita income
- They are the same zip code 94027, which is among the wealthiest zip codes of the area and nation wide

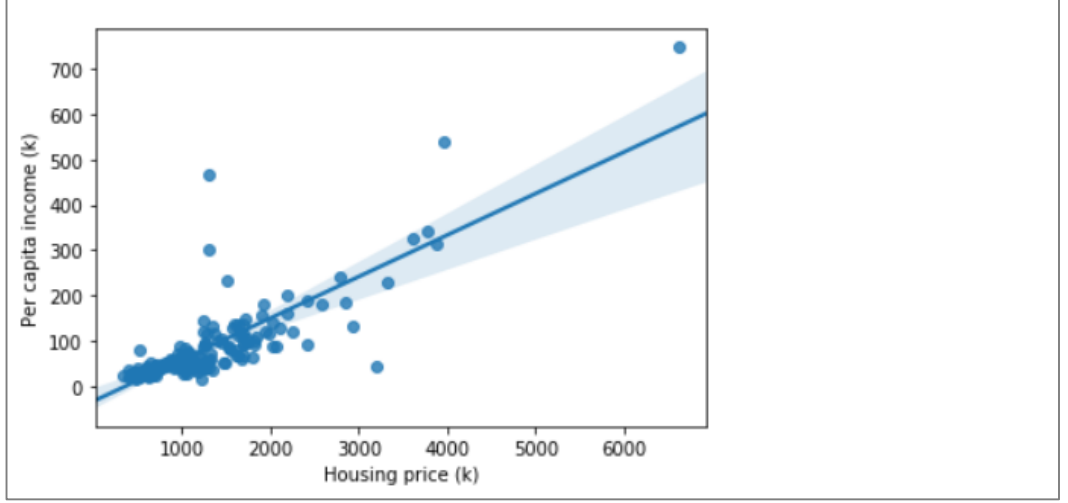
# Exploratory Data Analysis Cont'd

| | Population | House Price 2020 (k) | AGI per capita (k) |
|---|---|---|---|
| Population | 1.000000 | -0.228067 | -0.249686 |
| House Price 2020 (k) | -0.228067 | 1.000000 | 0.363845 |
| AGI per capita (k) | -0.249686 | 0.363845 | 1.000000 |




Outlier removed

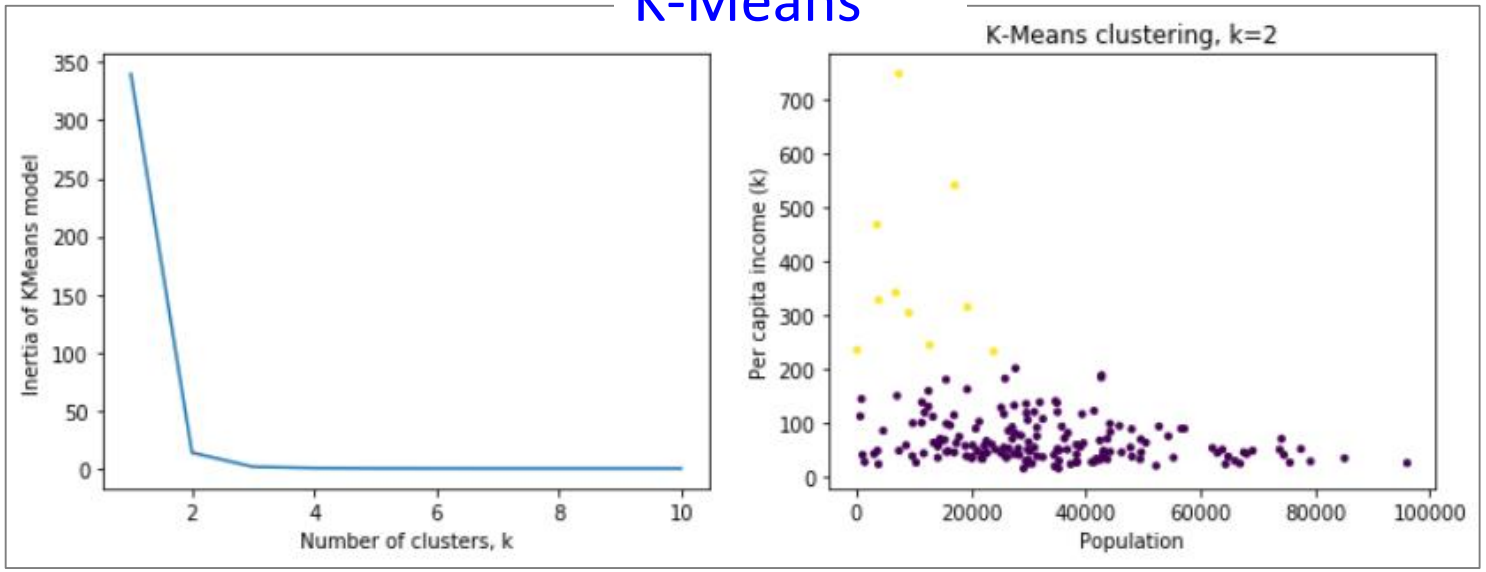
| | Population | House Price 2020 (k) | AGI per capita (k) |
|---|---|---|---|
| Population | 1.000000 | -0.226810 | -0.330512 |
| House Price 2020 (k) | -0.226810 | 1.000000 | 0.820047 |
| AGI per capita (k) | -0.330512 | 0.820047 | 1.000000 |



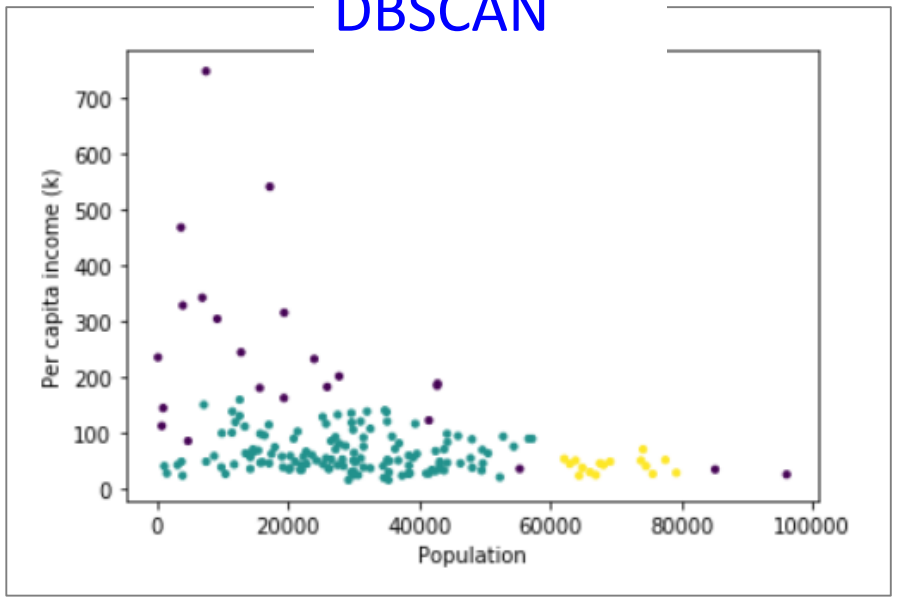

- Cross correlation shows the correlation between housing price and per capita income is low, and it was driven by an outlier data point
- After removing the outlier, correlation improved from 0.36 to 0.82
- For next steps only per capita income is used, together with population

# Clustering: K-Means and DBSCAN

K-Means



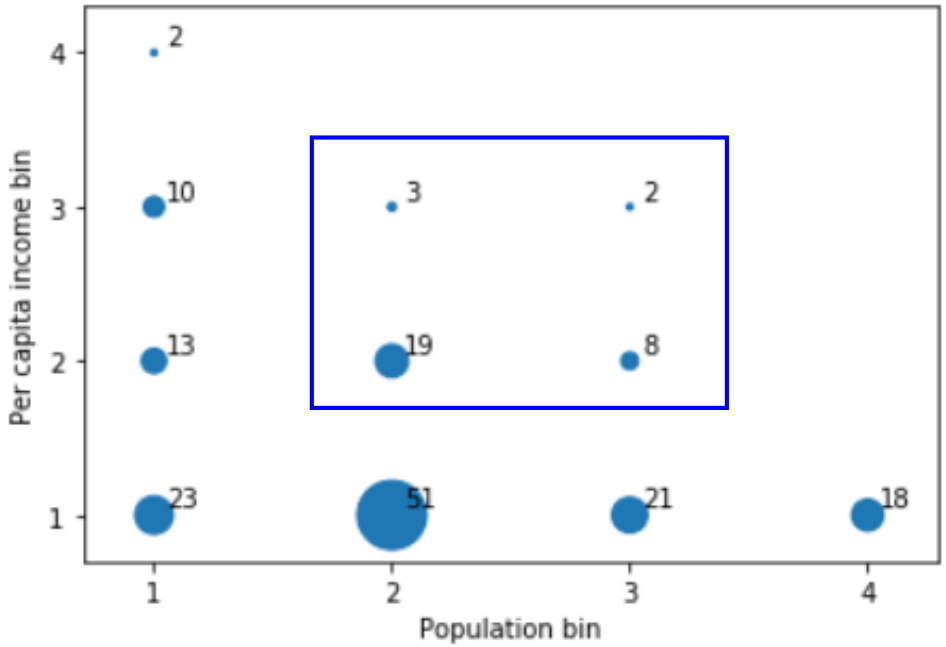

DBSCAN




- Clustering was tried for population and per capita income data using both K-Means and DBSCAN. K-Means used k=2 determined by elbow plot
- Both K-Means and DBSCAN do not show very clear clustering of data, because the data distribution is more continuous rather than clustered

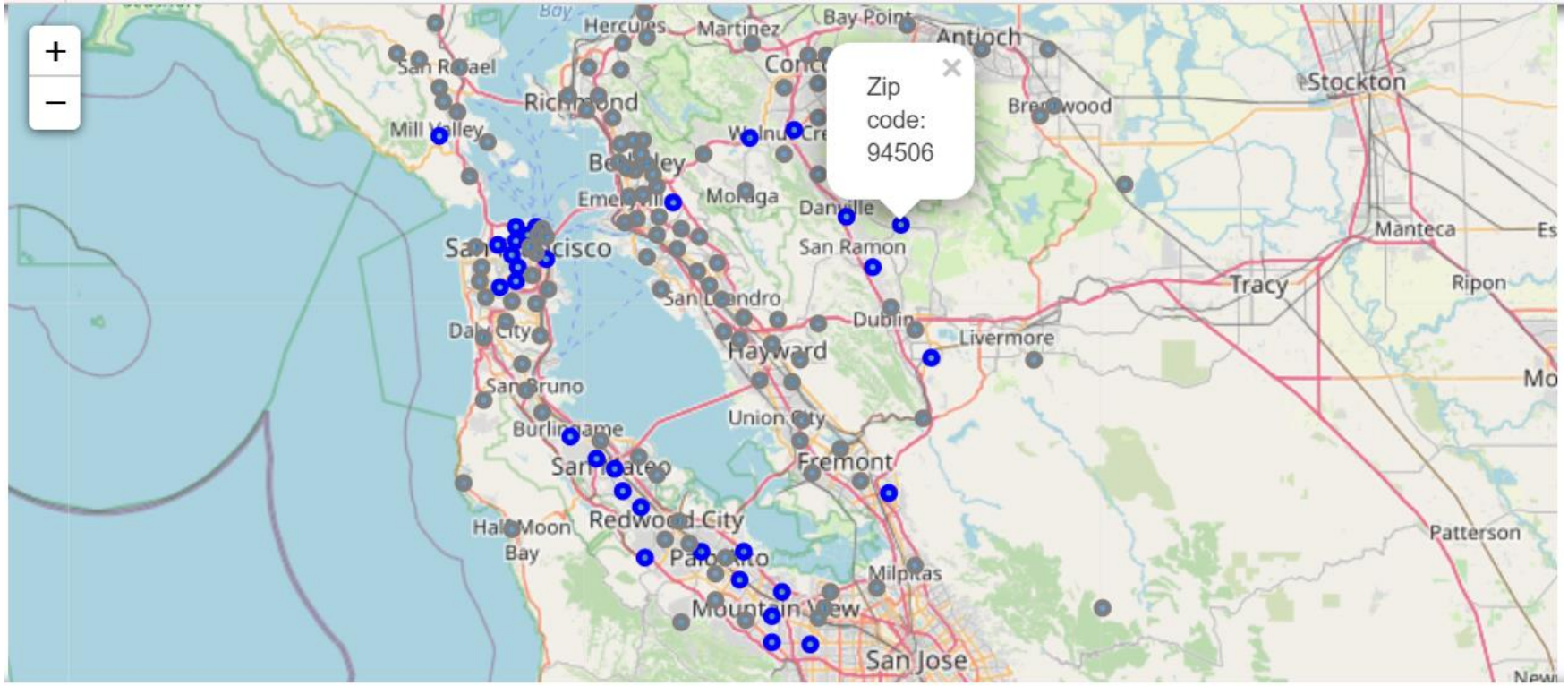
# Binning of Data

| Population Bins | Income Bins (k) | Housing Price Bin (ks |
|---|---|---|
| 0 | 0 | 0 |
| 20000 | 80 | 1000 |
| 40000 | 150 | 2000 |
| 60000 | 500 | 4000 |
| 100000 | 3000 | 8000 |



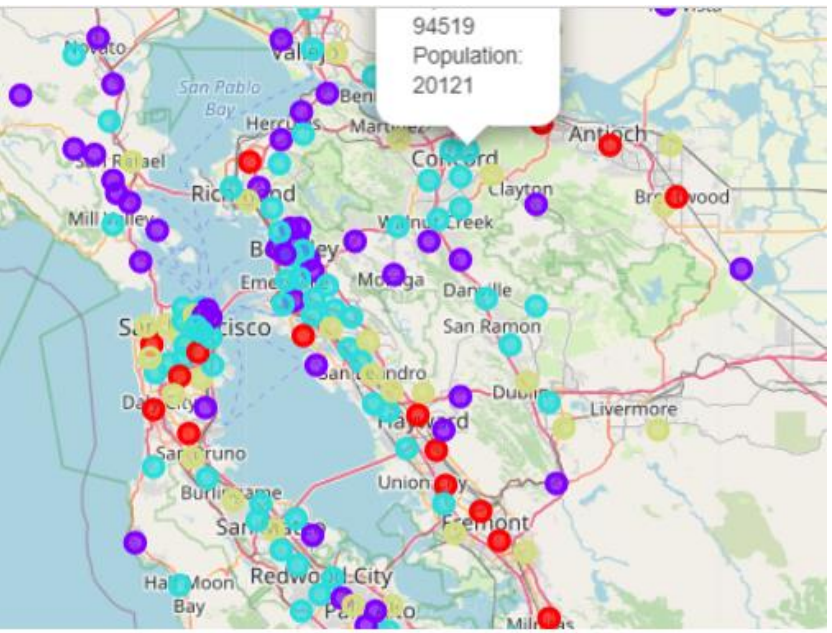

- Population, housing price, and per capita income for each zip code of the bay area are assigned into 4 bins each using the ranges in the table at the left
- Statistics of population and per capita income falling into each of the 4x4 = 16 bins are shown on the plot at the right, with the marker size and the number denoting the total number of zip codes in the bin
- The 32 zip codes in the blue box are of more interest for businesses because they have relatively higher population and per capita income, thus more spending power

# Maps – the 32 Zip Codes of Interest

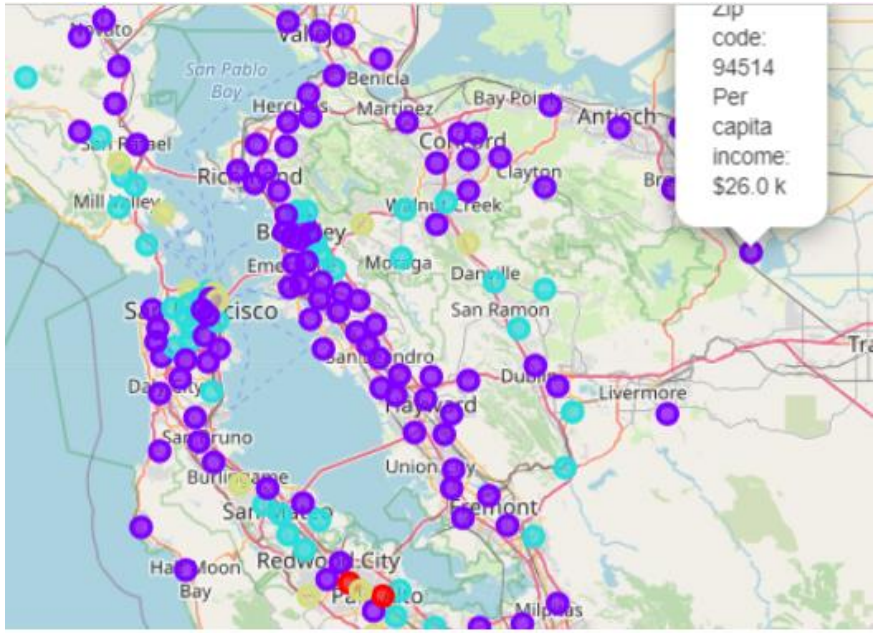



- The 32 zip codes of interest with higher population and per capita income are displayed using folium on the map highlighted in blue, the rest of the zip codes are in gray
- Clickable marker displays the zip code

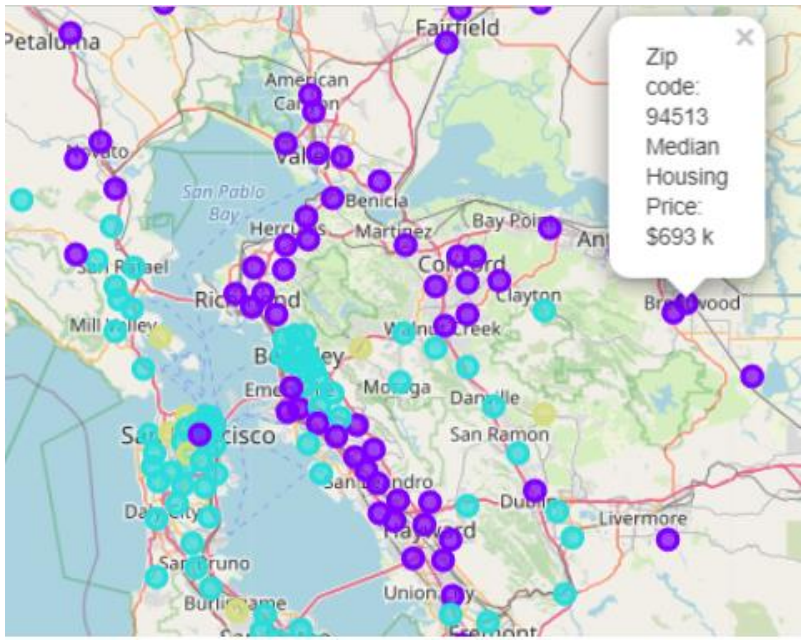
# More Maps




**Population**




**Per capita income**




**Housing Price**

- All 171 zip codes are displayed in three maps, each color coded for the 4 bins for population, per capita income, and housing price

# Explore the Area – Venues

| | ZIP | Total Venues | arts_entertainment | building | education | food | nightlife | parks_outdoors | shops | travel |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 94403 | 100 | 2 | 0 | 1 | 38 | 4 | 3 | 52 | 0 |
| 1 | 94506 | 29 | 2 | 2 | 1 | 12 | 1 | 2 | 9 | 0 |
| 2 | 94526 | 40 | 1 | 5 | 0 | 17 | 0 | 0 | 14 | 3 |
| 3 | 94539 | 6 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 0 |
| 4 | 94549 | 75 | 1 | 1 | 1 | 47 | 1 | 0 | 24 | 0 |
| 5 | 94566 | 74 | 2 | 0 | 0 | 49 | 2 | 1 | 18 | 2 |
| 6 | 94583 | 57 | 1 | 4 | 1 | 27 | 1 | 3 | 18 | 2 |
| 7 | 94596 | 100 | 2 | 2 | 0 | 71 | 8 | 1 | 15 | 1 |
| 8 | 94611 | 15 | 4 | 0 | 0 | 1 | 0 | 8 | 2 | 0 |
| 9 | 94941 | 33 | 1 | 1 | 0 | 15 | 1 | 3 | 12 | 0 |

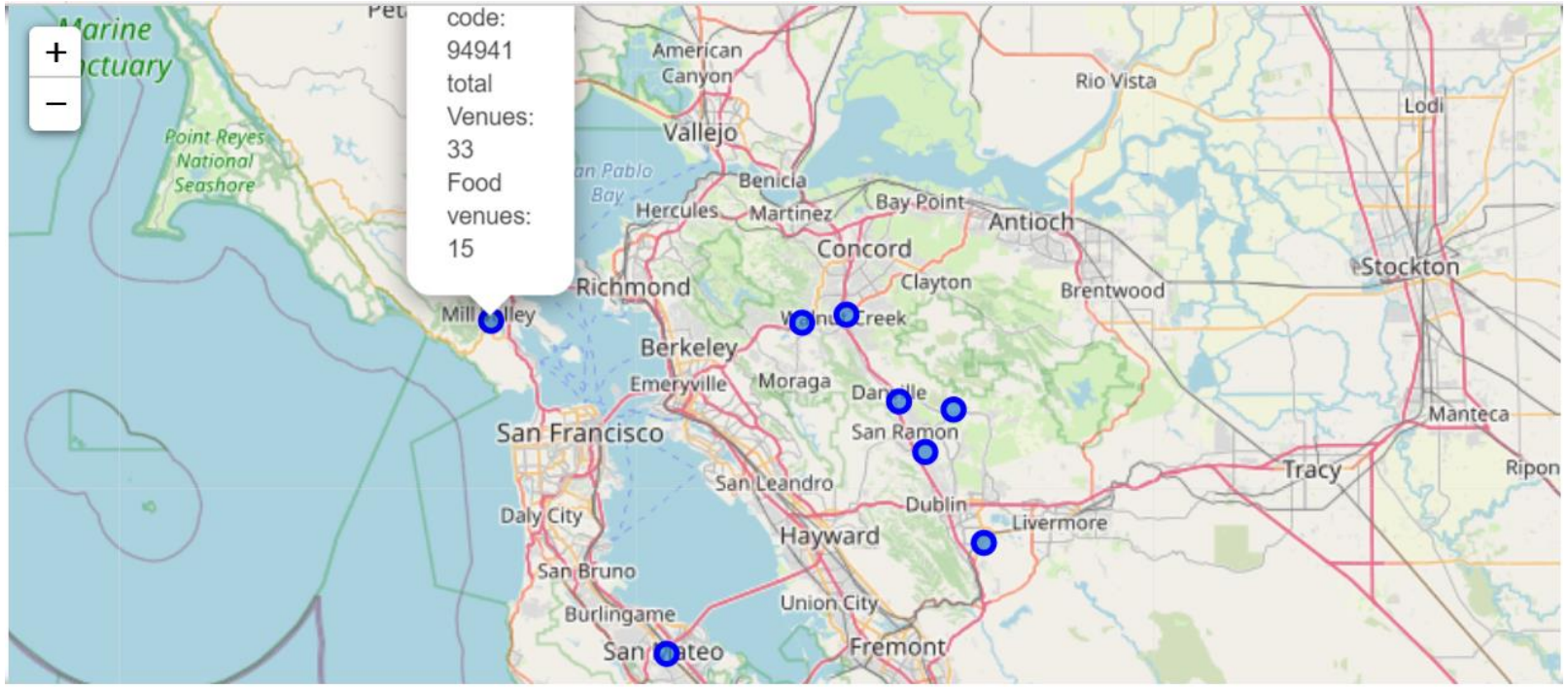
- Using Foursquare API calls, pulled the venues within 1000 meters radius of the center of each zip code, and further separated into high level categories, i.e., food, shops, travel etc

# Top 8 Zip Code with Most Venues

| | ZIP | Total Venues | food | Food Venue % | Latitude | Longitude |
|---|---|---|---|---|---|---|
| 0 | 94596 | 100 | 71 | 71.0 | 37.9 | -122.1 |
| 1 | 94403 | 100 | 38 | 38.0 | 37.5 | -122.3 |
| 2 | 94549 | 75 | 47 | 62.7 | 37.9 | -122.1 |
| 3 | 94566 | 74 | 49 | 66.2 | 37.7 | -121.9 |
| 4 | 94583 | 57 | 27 | 47.4 | 37.8 | -122.0 |
| 5 | 94526 | 40 | 17 | 42.5 | 37.8 | -122.0 |
| 6 | 94941 | 33 | 15 | 45.5 | 37.9 | -122.5 |
| 7 | 94506 | 29 | 12 | 41.4 | 37.8 | -121.9 |

- Top 8 zip codes with most venues are listed, together with the food related venues
- As an example, the top two zip codes have the same number of venues within the search radius, but different number of food related venues, so zip code 94403 may be preferred over 94596 thanks to less competition

# Top 8 Zip Code for Business Opportunities




- The Top 8 zip codes with best business opportunities (using the limited data features) are displayed on the map with zip code, total number of venues within the search range, and food related venues shown by the clickable markers

Thank you!