**Introduction**

I live in the San Francisco Bay Area. Like any other areas, there are variations location to location in demographics and economy, among others. It would be handy to have an easy by-zip code view of the basic information for the bay area. This would give researchers, government agencies, and business owners direct and quantitative insights into each zip code. As a result this will help facilitate understanding, planning, and business decisions such as where is the best location to open a small business.

**Data collection**

From online search, I have collected the following datasets:

|  |
| --- |
| * All zip codes of the bay area |
| * Latitudes and longitudes for all zip codes in California |
| * Population by zip code for California (2019) |
| * US nation-wide monthly housing price data by zip code : US (1996-2020) |
| * California’s by zip code tax data from IRS (2017) |
| * By zip code GeoJson file for California (coordinates of the boundaries of each zip code) * Foursquare data via API calls to Foursquare.com |



Data

**Data cleaning**

**Future refinement**

The model can be further refined with more data incorporated. For example, the population distribution by ethnicity varies from zip code to zip code and from city to city. I tried to search for by-zip-code ethinicity data from census but could not get it. This would allow better clustering to fit certain needs. For instance, it would make more sense to open an Asian restaurant in an area with more Asian population. Other data that can be added include location based crime rate, traffic data etc.

Another area of opportunity is that zip codes are not necessarily overlapped with city limits, as it was designed by USPS for their mail delivery purpose and thus some zip codes include very rural areas [1]. Further refining into neighborhood level would improve the accuracy of the clusters. Such data wasn’t found from publicly available data, but can be purchased.

[1] <https://www.unitedstateszipcodes.org/>