Where to open a bar in the Bay Area

Introduction:

In recent years, the Bay Area has experienced a growth in population due to the tech boom of the mid to late 2000's. This creates an increased demand for local services that can be covered by local businesses. Bars are one place that would experience an increased demand. As such, the focus of this investigation is to determine where would be the best place in the Bay Area to open a new bar.

Business problem:
While the demand increases with population size, it is still unwise to open a business in areas that are already sufficiently serviced by a particular business. Data science and methodology can be applied here to determine which cities in the Bay Area are the best to open a new bar based on the prevalence of other bars in each city. Another factor that should be taken into consideration is the size of the city in terms of population.

The Data:

Data will come from two primary sources. They are Wikipedia for the cities and towns in the Bay Area and Foursquare for more in depth location data. The city and town data will be sourced through web scraping and the Foursquare data will be accessed through API calls. Through web scraping the town and city data, a table containing the relevant information can be created through the pandas module in python. The geocoders module will also be used to obtain the coordinates of each city. The Foursquare data will then be used to obtain the venue data for each city. Machine learning will then be applied through the use of K-means clustering to determine where the best places to open a bar are.

Thus the required data is:

- City names
- Population size
- Coordinates
- Top venues
- Prevalence of bars

Methodology:

The first step in determining where to open a bar is to obtain the names and locations of all the cities in the Bay Area. This is done through web scraping with the beautiful soup module. The relevant information can be found on Wikipedia (https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area).

Using the geocoders module, the coordinates of each city can be obtained by looping through the list of cities. Towns will not be included due to computational cost.

The second step is to obtain location data from Foursquare. This is done by making API calls to obtain the top 50 venues for each city. Unfortunately 50 is the limit for a free account. The Foursquare data allows us to see the unique categories of venues for the whole Bay Area.

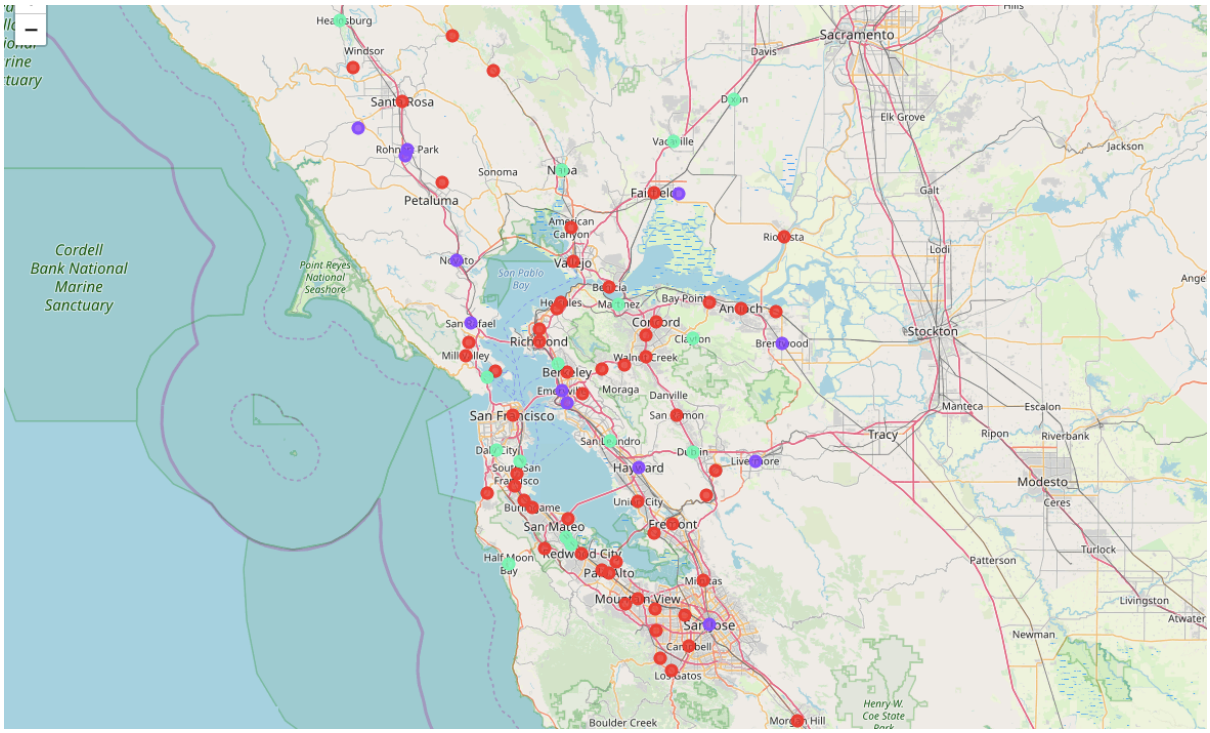
The third step is to utilize K-means clustering. A k-value of three was chosen for this step. After applying K-means clustering, the data can be visualized on a folium map to give an easily presentable visual of where to best open a bar in the Bay Area. K-means clustering is a very simple and popular method of unsupervised machine learning and is well suited to the problem.

The fourth step is to analyze and determine where is the best place to open a bar.

Results:

The clusters were grouped as such:

Cluster 0: the cluster with the least bars (red)
Cluster 1: the cluster with the most bars (purple)
Cluster 2: the cluster with a moderate amount of bars (green)



The mean populations in clusters 0, 1, and 2 were 67501, 151,393, and 38,215 respectively.

Cluster 0 had 11 major cities, cluster 1 had three major cities, and cluster 2 had just one major city.

Discussion:

Interestingly, the cluster with the least amount of bars was the one with the highest number of major cities. The average population is not reflective of that. This is likely due to the fact that we were limited to the top 50 venues by the Foursquare API. Common sense would dictate that the bars be built in the areas with the most demand, but due to the API limitations that may not be the case in this instance. However it can also be argued that since the bars in the major cities did not crack the top 50, the bars in those cities are not very good and perhaps are fertile grounds for a good bar. Another reason for the bars not being in the top 50 is that the major cities in cluster 0 are diluted with bars. It is inadvisable to open a bar in the cities in cluster 1 as they have a higher prevalence of bars than other cities in the Bay Area. Cluster 2 has a moderate amount of bars and thus may or may not be a suitable place to open a new bar. Cluster 2 has the smallest mean population size by a large amount of the bunch suggesting that there may be a high bar to person ratio. It is in my opinion that these areas also be avoided. Thus, it is advised that prospective bar owners open a bar in the areas of cluster 0, but perhaps not in the major cities.

Conclusion:

This project utilizes data sourced from the internet and Foursquare API to determine where is the best place to open a bar in the Bay Area. Through the use of K-means clustering it was determined that areas in cluster 0 are the best places to open a bar, but due to API limitations major cities may not be accurately represented here. Thus the final recommendation is to open a bar in the cluster 0 cities that have a population under 100,00.