**Introduction / Business Problem**

The San Francisco Bay Area is a popular destination for people who want to live in a diverse cosmopolitan location with significant job opportunities and a comfortable climate. With years of development and growth, many cities have sprung up around San Francisco, contributing to a suburban sprawl that offers vastly distinctive options for potential residents. However, with the sheer number of cities, it is also increasingly difficult to research and find the best city or cities around which to focus a new home search. Data Science can be used collect and analyze data from disparate sources to arrive at a short-list of cities based on potential homeowner preferences.

**Data**

This analysis uses data from the following two sources.

**Wikipedia** – This is a good source of reference data compiled by an extensive user base. The data is often displayed in table format from a page written in html. A number of web-scraping techniques can be used to extract the data required for analysis. This analysis uses the html parser BeautifulSoup for this purpose.

1. Cities in the San Francisco Bay Area (sample screenshot below) (<https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area>)

A screenshot of a cell phone

Description automatically generated

1. Crime rates for cities in the San Francisco Bay Area (sample screenshot below) (<https://en.wikipedia.org/wiki/California_locations_by_crime_rate>)

A screenshot of a cell phone

Description automatically generated

**FourSquare** – This is a good source of information about venues of multiple types located around a specified latitude/longitude (sample screenshot below).

A screenshot of a cell phone

Description automatically generated

Potential residents are likely to be interested in knowing about neighborhood venues when they decide where to relocate, so this information can be used to characterize each Bay Area city and to help focus a search on cities with desirable venues.

**Methodology**

The list of cities in the Bay Area was retrieved from Wikipedia, and the coordinates (latitude and longitude) for each city was retrieved using the Nominatim call from the geopy.geocoders library.

*Venue Analysis*

The city coordinates were subsequently used to retrieve the available venues for each city from FourSquare. The original list from Wikipedia had 101 cities but the resulting list of venues from FourSquare only had 94 cities because there were 7 cities for which FourSquare did not have any venue information. This is not surprising since the list of Bay Area cities includes fairly far-flung rural cities which are not typical destination hotspots.

Since there were 310 categories of venues, the next step involved determining the top venue categories for each city as a way to characterize the profile of the city. The list of venues for all the cities obtained from FourSquare was transformed using one-hot encoding, then grouped by city to produce a list of cities along with the average number of venues in each category for each city. This list was then sorted to identify the top venue categories for each city.

Let’s assume potential residents are interested in cities with a number and variety of restaurants. Searching for cities with top venue categories that contained the string “Restaurant” identified the cities for which the top 1 through 5 venue categories are restaurants.

This analysis can easily be replicated for any other venue categories of interest to potential residents.

*Crime Analysis*

Potential residents are likely to want to know about crime rates in cities, with a natural preference for areas with low crime rates. Sorting the data by either violent crime rates or property crime rates resulted in lists of top 10 cities with low crime rates.

However, cities with low violent crime may not have low property crime, and vice versa. Looking for the intersection of the low violent crime and low property crime lists would be more informative in providing a list of cities with both low violent crime as well as low property crime.

*Combined Venue-Crime Analysis*

Building on the previous analyses of restaurants and crime, the analysis was extended to identify cities with the attractions of both many restaurants and low crime. A combined dataset was constructed by merging the list of cities with at least one restaurant category in its top 5 venues with the complete list of cities with their associated crime rates. This dataset was then sorted by violent crime to get the top 10 cities with many restaurants and low violent crime, and subsequently sorted by property crime to get the top 10 cities with many restaurants and low property crime. The intersection of these 2 top 10 lists resulted in the list of cities with many restaurants and low crime.

As a variant on this approach, the starting list for the merge was taken as the list of cities with either low violent crime or low property crime (once again, acknowledging that these are different lists), and merging with the list of cities with many restaurants. The intersection of these two lists was used to identify the cities with low crime and many restaurants.

*Cluster Analysis*

The cluster analysis used the original venue dataset from FourSquare, i.e. not just the subset of cities with many restaurants, and the original crime data set with all cities. In preparation for clustering, the numerical values representing instances of venues and city crime rates were normalized using the StandardScaler object to minimize the potential for skewing effects of large or small numbers.

k-means clustering was used on the combined dataset of venue information and crime rates. Different numbers of clusters were attempted, ranging from 3 to 10, to see if a particular number of clusters would yield more informative results.

**Results**

The following results were obtained from the analysis of cities with restaurants.

|  |  |  |
| --- | --- | --- |
| **Number of top 5 venue categories that are restaurants** | **# Cities** | **Cities** |
| 4 or 5 | 0 |  |
| 3 | 14 | Benicia, Brisbane, Fairfield, Hayward, Larkspur, Los Altos, Mill Valley, Millbrae, Milpitas, Morgan Hill, Napa, Newark, Oakland, Orinda, Pleasant Hill, Pleasanton, San Bruno, San Carlos, San Rafael, South San Francisco, Tiburon, Vacaville |

In total, the number of cities with at least 1 category of restaurant in its top 5 most common venues is 79.

The following results were obtained from the crime analysis.

|  |  |
| --- | --- |
| **Cities with Low Violent Crime** | **Cities with Low Property Crime** |
| *Monte Sereno* | Ross |
| *Hillsborough* | *Monte Sereno* |
| Tiburon | *Los Altos Hills* |
| Orinda | Moraga |
| Los Altos | *Hillsborough* |
| *Los Altos Hills* | Saratoga |
| San Ramon | Windsor |
| *Clayton* | St. Helena |
| Danville | Cotati |
| Atherton | *Clayton* |

The cities that show up on both top 10 lists, i.e. have low violent as well as property crime, are Monte Sereno, Hillsborough, Los Altos Hills, and Clayton.

The following results were obtained from the combined venue-crime analysis.

|  |  |
| --- | --- |
| **Cities with Many Restaurants and**  **Low Violent Crime** | **Cities with Many Restaurants and**  **Low Property Crime** |
| *Tiburon* | *Orinda* |
| *Orinda* | *Los Altos* |
| *Los Altos* | *Tiburon* |
| *Mill Valley* | *Mill Valley* |
| *Pleasanton* | *Morgan Hill* |
| *Benicia* | Napa |
| *Morgan Hill* | *Pleasanton* |
| Milpitas | *Benicia* |
| Pleasant Hill | *South San Francisco* |
| *South San Francisco* | Newark |

8 out of the top 10 cities are the same in both lists – Tiburon, Orinda, Los Altos, Mill Valley, Pleasanton, Benicia, Morgan Hill and South San Francisco.

The alternative approach to this analysis, constructing the list of cities with low violent and property crime as well as many restaurants by starting with the low crime lists, yielded three cities – Tiburon, Orinda, and Los Altos – all low in violent crime. In fact, there are no cities from the low property crime list that have many restaurants; only cities with low violent crime can have many restaurants.

The following results were obtained from the cluster analysis, using k-means clustering.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster Description** | **# Cities** | **Violent Crime1** | **Property Crime1** |
| 1. Moderate violent crime, high property crime | 5 | 7.77 – 8.65 | 38.72 – 53.03 |
| 2. Low violent crime, low-moderate property crime | 34 | 0.0 – 1.62 | 8.48 – 20.77 |
| 3. Extremely high property crime, high violent crime | 2 | 7.98 – 10.66 | 146.1 – 180.31 |
| 4. Extremely high violent crime, very high property crime | 1 | 16.85 | 59.43 |
| 5. Low-moderate crime | 30 | 1.56 – 4.86 | 9.29 – 30.2 |
| 6. Low-moderate violent crime, high property crime | 14 | 1.1 – 4.71 | 28.06 – 50.58 |

1 incidents per thousand residents

The clusters can be described in terms of crime rate levels, either for violent or property crime or both. They are listed and highlighted in the map below.

*Cluster 1*

A screenshot of a cell phone

Description automatically generated

*Cluster 2*

*A screenshot of a cell phone

Description automatically generated*

*A close up of a piece of paper

Description automatically generated*

*Cluster 3*

**

*Cluster 4*

**

*Cluster 5*

*A close up of a piece of paper

Description automatically generated*

*A screenshot of a cell phone

Description automatically generated*

*Cluster 6*

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

**Discussion**

There are two limitations in using FourSquare data that were encountered in this project.

1. The quality of the venue analysis is the degree of dependence on the labels used for each venue category. For instance, in searching for cities with restaurants, the analysis excluded categories like “Café”, “Deli”, and “Pizza Place” because these labels did not contain the term “Restaurant” although arguably they should have been included as dining establishments.
2. FourSquare data is live and updated frequently so conclusions drawn from one set of results may change over a short period of time.

The analyses of venues and crime rates illustrate typical trade-offs faced by potential residents as they are forced to prioritize multiple desirable attributes of cities. If low crime is of paramount importance while having many restaurants in the neighborhood is a strong preference, recommended cities to consider are (in descending order):

1. Tiburon, Orinda, and Los Altos – cities with low crime and many restaurants
2. Mill Valley, Pleasanton, Benicia, and Morgan Hill– cities with many restaurants that are relatively low on crime
3. The 27 remaining cities in cluster 2 (“low crime”) of the cluster analysis since this is the cluster shared by the cities listed above in 1. and 2.

If having many restaurants in the vicinity is the most important attribute and crime rates matter less, then also consider the cities that came up on top in the restaurant venue analysis – Brisbane, Fairfield, Hayward, Larkspur, Millbrae, Milpitas, Napa, Newark, Oakland, Pleasant Hill, San Bruno, San Carlos, San Rafael, South San Francisco, and Vacaville.

Despite the normalization of all the numeric values used in clustering, the crime data still seems to have carried more weight than the venue data in the clustering process. This may be because there were so many venue categories (310) that the instance data was spread too thin across many possible values (categories) relative to the crime data which had only two values (violent or property).

**Conclusion**

Data Science can be applied to readily-available locational information, e.g. cities, venues within cities, and crime rates by city, in order to arrive at short lists of cities based on different criteria. In this analysis, the focus was on potential residents looking to settle in cities with desirable characteristics. It was demonstrated that multiple different analyses could be used to enable alternative perspectives on the information available, with the ultimate objective of providing insights on which to base decisions.

Further work in this area could include the development of a front-end user interface that would allow the selection of criteria, in this case venue categories of interest, so that alternative custom analyses can be conducted. Another potential area for expansion could be the inclusion of more data sources to complement the venue and crime data and that may be relevant for different potential residents, e.g. information on the housing market by city (house prices, % buyers vs. renters, etc.), information on schools within each city, and so on.

The tools of Data Science today enable effective data manipulation and analysis by applying the power of computing to massive quantities of information.