Where to Relocate in California after the COVID-19?

Dai Li

Introduction/Business Problem

COVID-19 is an infectious disease caused by a coronavirus newly discovered in 2019. It is highly contagious and spreads fast from person to person in close contact. After the World Health Organization (WHO) declared the COVID-19 outbreak a global pandemic on March 11, 2020, many regions and areas issued the stay-at-home order, which means a number of non-essential workers need to work from home.

An interesting result of work-from-home during the COVID-19 is that people start to realize the business goes just fine when employees work from home in pajamas. There are 27 United States companies, most in IT, that have announced permanent work-from-home decision by September 2020. As people can work from anywhere, they try to escape from the areas of sky-high home values and choose a place where they can enjoy time with family (and friends after the pandemic).

California is one of those regions. California, especially the Silicon Valley area, has the clusters of world-leading IT companies and its home value is constantly among TOP 3 in the United States. After those companies announced long-terms work-from-home decisions, many employees then decide to relocate to where they can live in the styles that they always want.

The business problem of this project is to cluster California cities that fit different lifestyles for the people considering relocation.

Targeted Audiences/Stakeholders

The primary targeted audience of this project are those who are looking for a place that fit their lifestyles.

Data

Based on the definition of the business problem, the factors that would affect people’s decision include:

1. The safeness of the city based on crime statistics
2. The education quality of the city based on the public school data
3. The affordability of the home in the city based on real estate data
4. The lifestyles offered by the city based on the most common venue categories

Data Sources and Methods to Extract the Data

The data sources that will be used for this project:

1. [California Offenses Data by County and City:](https://www.kaggle.com/fbi-us/california-crime?select=ca_offenses_by_agency.csv) a dataset containing the crimes submitted either through a state Uniform Crime Reporting program or directly to the FBI’s program.
2. [US Schools Dataset](https://www.kaggle.com/andrewmvd/us-schools-dataset): the dataset taken from the US Department of Homeland Security and contains information of all public and private schools with their geographical distribution.
3. [Housing Data](https://www.zillow.com/research/data/): the dataset downloaded from Zillow Research which contains the seasonally adjusted measures of the typical home value in the 35th to 65th percentile range.
4. [Dataset of the Neighborhoods, boroughs, and the most common venues along with co-ordinates](https://developer.foursquare.com/developer/): This data will be fetched using Four Square API to explore the neighborhood venues.

Methodology

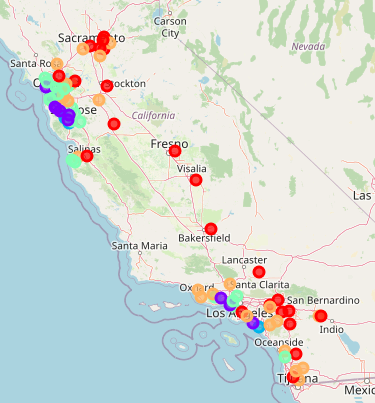
Data cleaning: Four data sets are used for this project and each needs cleaning before it can be used for the final step of clustering.

* For California Crime data, I create a new variable named Crime Density, which is the population of the city divided by the total number of the crimes reported in the data set. The new indicator represents the overall safeness of the city given the population density. The cities with the highest crime density are Los Angeles, San Jose, and San Diego, and the safest city is Industry.
* For U.S. School data, I, again, create a new variable of Student-to-Teacher ratio as an approxy for Elementary school quality. The higher the Student-to-Teacher ratio is, the less the school resource, and therefore the education quality. Grouped by the cities, the average Student-to-Teacher ratio of all elementary schools in the given city is used. The city with the highest Student-to-Teacher ratio is Yuba, where one teacher teaches 52 students. The city with the lowest Student-to-Teacher ratio is Yreka, where one teacher teaches only 13 students.
* Zillow Housing Data is collected at the neighborhood level. However, because the purpose of the project is to find the relocation city, the neighborhood is a level too fine for the project. Also, because I will exceed the daily access quota when requesting Four Square location data if I do at the neighborhood level, I decide to aggregate the whole dataset to the city level. Like I do for the school data, the average Single-Family House value recorded by Zillow sales data in October 2020 of the city is used. Tehachapi shows the lowest SFH value of a bit higher than $200K, and Malibu, unsurprisingly, is the city of highest SFH value of over 4.7 million US Dollars.
* After merging the above three datasets together, I have 85 cities in the final data before I request their venue data in Four Square. As mentioned above, my dataset is at the city level. Therefore, I choose to request the venue categories rather than individual venues for the cities because it is likely the most common venues fall in the similar categories then the dataset will not have enough variation. After I finished requesting and cleaning the Four Square data, I have 108 unique venue categories, and the most common venue categories in my dataset is Salon/Barbershop, Office, and Mexican Restaurants. The final dataset contains 82 cities that is ready for the Machine Learning.

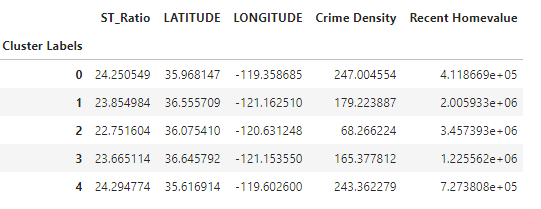
Machine Learning: I used K-Means clustering to find the cities with similar features. My dataset is quite spread-out, and hard to visual the obvious clusters with those variables. I set the number of clusters to be 5.

Results and Discussion

From the Folium map, we can see the cluster markers are centered around the major cities including Sacramento, San Francisco, and Los Angeles. Most cities in the central valley area fall in the first cluster (Cluster 0).



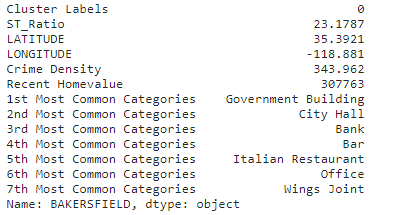
The following table shows the average values of the key features grouped by cluster.



Here are the typical cities and brief description of each cluster.

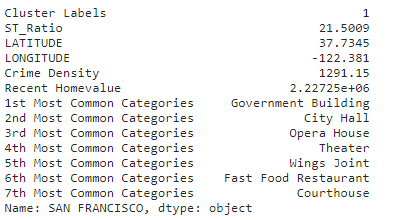
First Cluster (Cluster=0): away from the ocean, affordable homes (in CA), high crime density, good places for pizza lovers

Typical city: Bakersfield



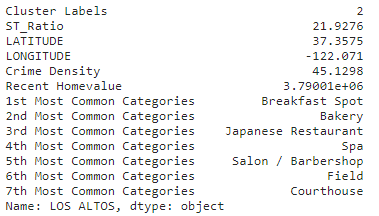
Second Cluster (Cluster=1): by the beach, expensive homes, high crime density, much entertainment

Typical city: San Francisco



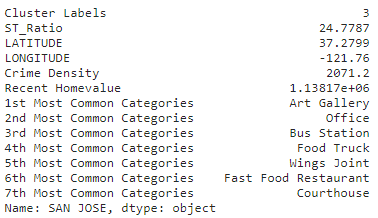
Third Cluster (Cluster=2): close to the beach, most expensive homes, low crime density, high school quality, diverse food and entertainment choices

Typical city: Los Altos



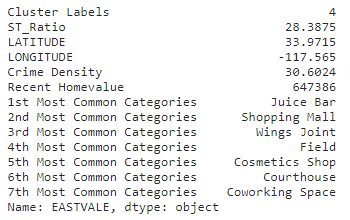
Fourth Cluster (Cluster=3): by the beach, expensive homes, high crime density, good entertainments and food choices

Typical city: San Jose



Fifth Cluster (Cluster=4): close to the beach, still-expensive-but-not-intimidatingly-so homes, high student-to-teacher ratio, and some entertainments

Typical city: Eastvale



Conclusion

This study is to provide some insights for the people who may be considering relocate in California. The K-means clustering is used to put CA cities with similar features into the same clusters. Hopefully, it could guide the consideration for the people trying to find the places that fit their lifestyles.