Affordable Housing Project

Capstone Project - Battle of Neighbourhoods

Sangeetha Chinnan - 13 February 2021

1. Introduction

1.1 Background

The San Francisco Bay Area, referred to as the Bay Area, is a populous region surrounding the San Francisco, San Pablo, and Suisun Bay estuaries in Northern California. Bay Area is defined by the Association of Bay Area Governments to include the nine counties that border the aforementioned estuaries: Alameda, Contra Costa, Marin, Napa, San Mateo, Santa Clara, Solano, Sonoma, and San Francisco. Home to approximately 7.75 million people, Northern California's nine-county Bay Area contains many cities, towns, airports, and associated regional, state, and national parks, connected by a complex multimodal transportation network.

1.2 Problem

The Bay Area is the most expensive place to live in the United States. Strong economic growth has created hundreds of thousands of new jobs, but coupled with severe restrictions on building new housing units, has resulted in an extreme housing shortage.With high costs of living, many Bay Area residents allocate large amounts of their income towards housing.Because of the high cost of housing, many workers in the Bay Area live far from their place of employment.

In this project we will try to find an optimal location for affordable housing project. Specifically, this report will be targeted to stakeholders interested in building a housing project in and around the San Francisco Bay Area counties in California, United States.

Since there are lots of housing projects in the Bay Area we will try to detect the locations that has affordable house sales price. And the area should have close proximity to the restaurants or venues. We will be using data science to generate a few most promising cities based on these criteria.

2. Data

2.1 Data Sources

To consider the problem we discussed in the introduction section we can list the required data we needed as below:

- List of counties in California
- List of counties in Bay Area
- Average House sales price per county
- number of existing venues in the city

For the county boundaries, the data available publicly from Open Data Soft portal is used. Folium map is utilised to visualise it.

The dataset set for the list of counties in California and the subset of those counties in Bay Area are downloaded to GitHub repository from DataSF(Office of chief Data Officer) portal

The dataset for the Average housing price are extracted from **Zillow**. For the venue details per city in California Foursquare API is used. The coordinates of each of these cities are obtained using geopy.

2.2 Data Cleaning and Feature Selection

Data related to the Bay Area Counties are downloaded or scraped from multiple sources and are combined into a single table. This project is planned for the Bay Area counties as a result the data from the other counties outside is discarded.

The data extracted from Zillow has average housing price for every month in the past few years. As a result the latest data i.e, the average housing price for the month of December 2020 are only used for this project. In the future the study can be expanded by using the data available from the portal.

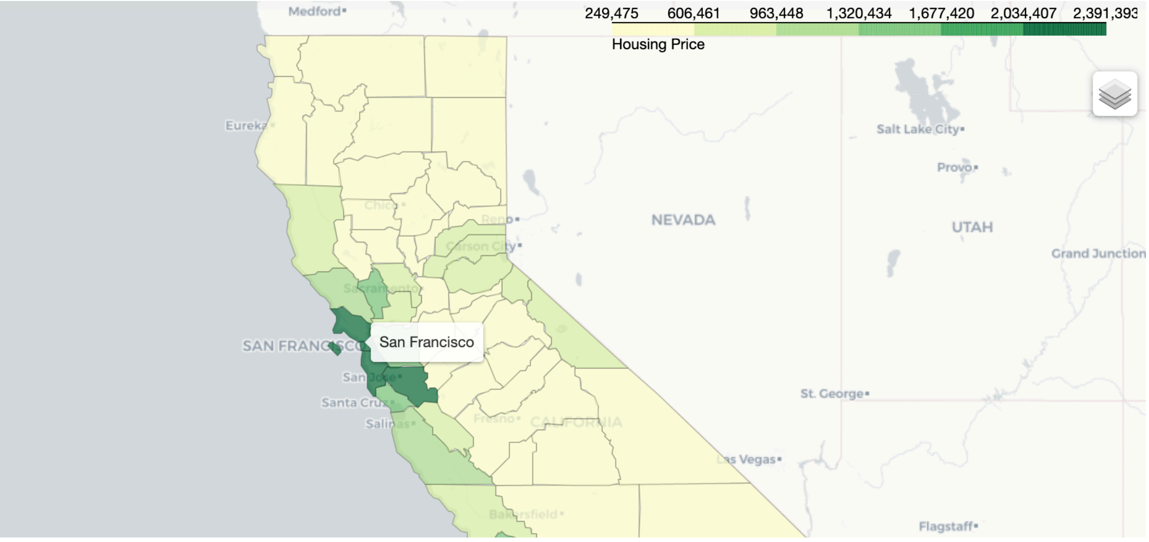
3. Methodology

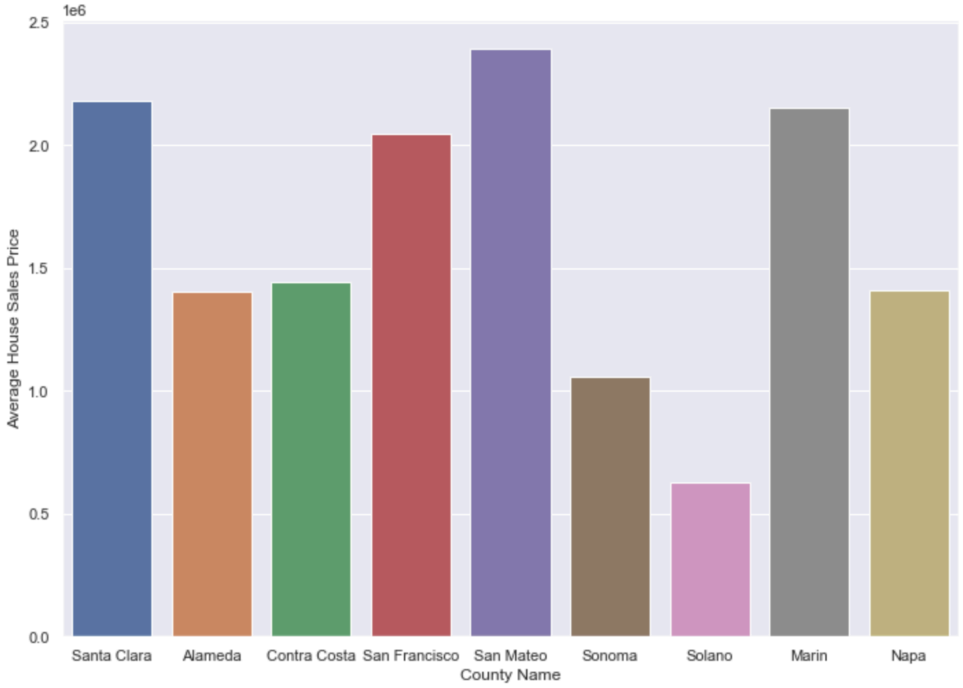
Folium map is used to visualise the geographic boundaries of the counties in California with the help of the geojson file in the GitHub Repository. Choropleth is used to display the average house dales price per county. A choropleth map is a thematic map in which the areas are shaded according to the statistical variable being displayed on the map.

In this project Foursquare API is utilised to get the venue details for each of the Bay Area Counties. Each county has multiple cities. City names are used to extract the venue details from Foursquare. Using k-means clustering algorithm the venues are clustered to identify the better location to start an affordable housing project.

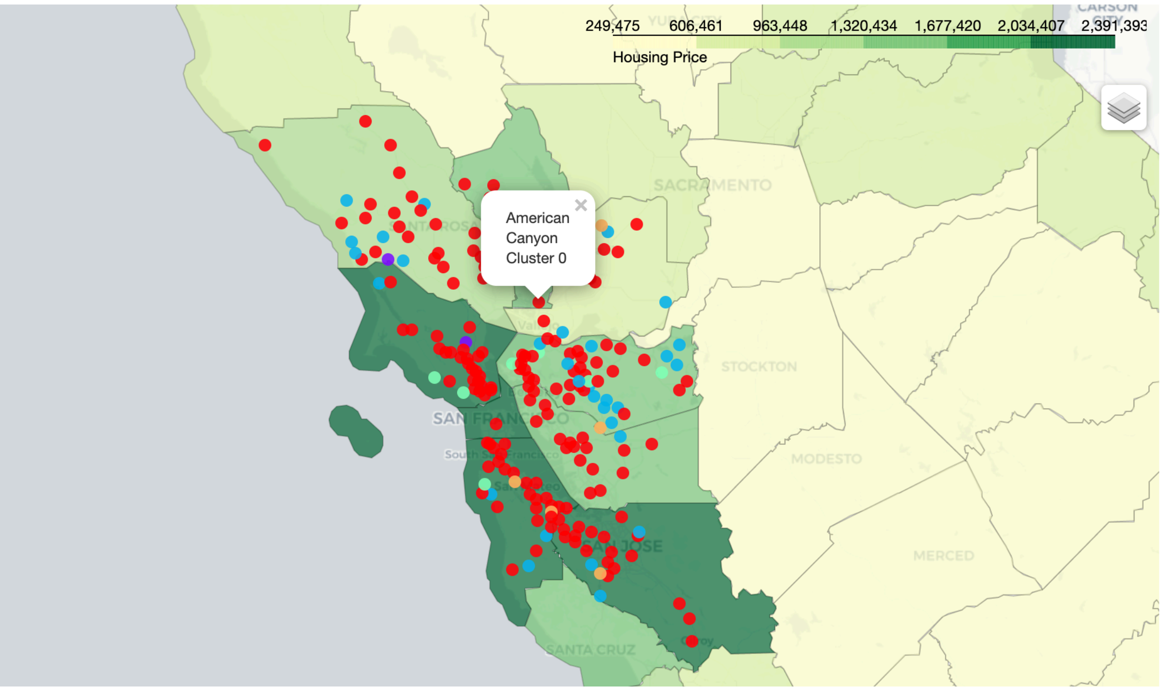
4. Results

One of our target was to visualise the Average Housing Sale Prices for each county with choropleth style map. Bar Plot is also utilised to verify the same. Higher the price stronger the colour.





The data from Foursquare API is then used to super impose the venue details onto the folium map.



5. Discussions

The Bay Area is the most expensive place to live in the United States. Strong economic growth has created hundreds of thousands of new jobs, but coupled with severe restrictions on building new housing units, has resulted in an extreme housing shortage.With high costs of living, many Bay Area residents allocate large amounts of their income towards housing.Because of the high cost of housing, many workers in the Bay Area live far from their place of employment.

As you can see the complexity of the problem a very different approaches can be tried in clustering and classification studies. K-Means algorithm is being used as part of this clustering study. I tested the data with K value as 5. The dataset for the counties that are included in the Bay area are used. You can expand the dataset to include all the counties in California.

Data analysis is also performed through these datasets by adding the coordinates of counties and house sales price averages as static data on GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I concluded the study by visualising the data and clustering information on the California map.

6. Conclusion

The main purpose of this study was to identify the counties that has affordable house sales price, And are in close proximity to the restaurants or venues in San Francisco Bay Area in order to aid stakeholders in narrowing down the search for optimal location for new housing projects.

By using the Foursquare data we were able to narrow down our search to the venue locations within the radius of 1000 meters. Clustering of those locations was then performed in order to find the potential locations.

Stakeholders will decide the final location based on the characteristics of the city and the venues in every recommended cities, taking into consideration additional factors like attractiveness of each location, real estate availability, levels of noise / proximity to major roads etc.