**Capstone Project: The battle of Neighborhoods in Orange County**

A view of a beach with palm trees and a body of water

Description automatically generated

Image from <https://www.acg.org/occ>

# **1 Introduction**

## **1.1 Problem**

Every year, at least 10% of American population move to a new place.[1] People move for different reasons: wanting a better/bigger/cheaper house, shorter commute, a new job that requires relocation, better schools for school-age kids, to be closer to extended families, etc.

California is among the top 3 states that Americans are moving to.[1] Located south of Los Angeles county, Orange County is one of the fastest growing counties in California in terms of economy and population, and has become a hot spot for people to move into.

Before people decide on which city to move to, it is always a good idea that they do their homework and gather as much information as possible. Some information such as housing price and school districts, can be directly obtained from real estate websites like Zillow, Redfin and GreatSchools; while some other information is not easily accessed due to the fact that these information is subject to personal preferences. For example, some enjoy vibrant neighborhoods with lots of bars and coffee shops, while others like to live close to gyms/yoga studios/hiking trails. Some like a neighborhood with a variety of fast food options so they can grab a quick bite, while others prefer a neighborhood with lots of ethnic restaurants (Orange County has a high proportion of immigrant population).

## **1.2 Solution**

To help people make an informed decision that balance their budget and lifestyle preference, in this report, we explore the cities in Orange County and cluster them 1) based on household incomes and housing prices and 2) based on most common venues. This way, people can have a better picture of a city and choose a neighborhood that fits their budget as well as their lifestyle.

# **2 Data collection and wrangling**

We scrape demographic data for Orange County (city names, population and median household income) for Wikipedia.[2] Due to the fact that US Census is conducted every 10 years, the latest US Census data is from the 2010 US Census; it is recommended that this part of the data is updated after the 2020 Census is completed. (Table 2.1)

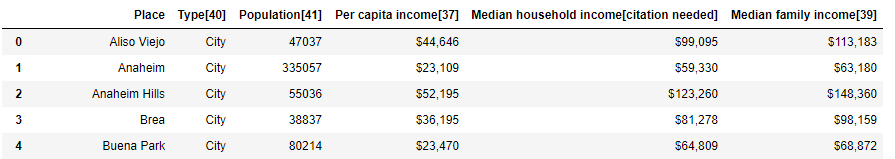


Table 2.1 Orange County Demographics Data

We downloaded housing data from Zillow.[3] The data was cleaned up in Excel and the latest data (year 2019) is used for this report (Table 2.2).

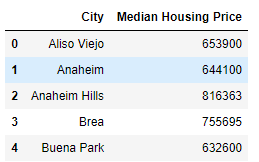


Table 2.2 Orange County Median Housing Price 2019

The coordinates of each city are obtained using API Nominatim (Table 2.3).

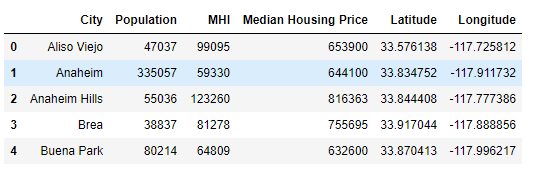


Table 2.3 Latitudes and Longitudes of Cities in Orange County

# **3 Methodology**

## **3.1 Clustering of cities based on median household income and median housing price**

k-means clustering is used and the number of clusters is set to 4 (elbow point). The clusters are visualized using folium map (Figure 3.1).

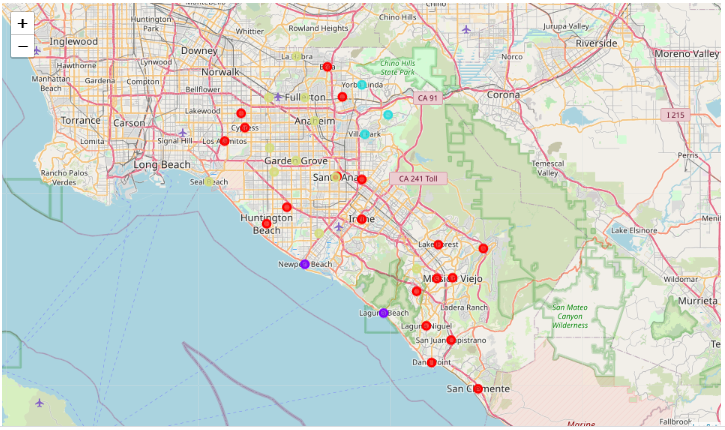


Figure 3.1 Clustering of cities based on median household income and median housing price

## **3.2 Exploring venues and clustering cities based on top 10 most common venues**

Using the longitudes and latitudes from Table 2.3, we can explore the venues within a 3000 meter radius of each city by calling Foursquare API (my free account only returns a limit of 100 venues). This return a json file that is converted to a pandas dataframe. The data frame contains information of venues for each city (Table 3.1).

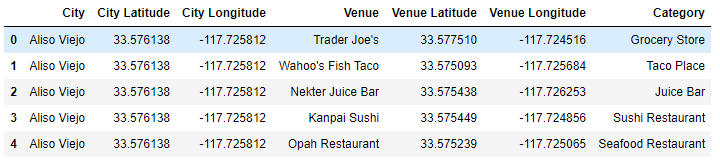


Table 3.1 Venue information from Foursquare API

Since categories are non-numerical, one hot coding was performed to convert them to 0 or 1 numerical values. The categorical values are then grouped by cities and the means for each category were calculated. Finally the top 10 most common venues for each city is displayed in a dataframe.



Table 3.2 Top 10 Most Common Venues per City

k-means clustering is used to cluster the cities based on the top 10 most common venues; the number of clusters is set to 5 (elbow point). The clusters are visualized using folium map (Figure 3.2).

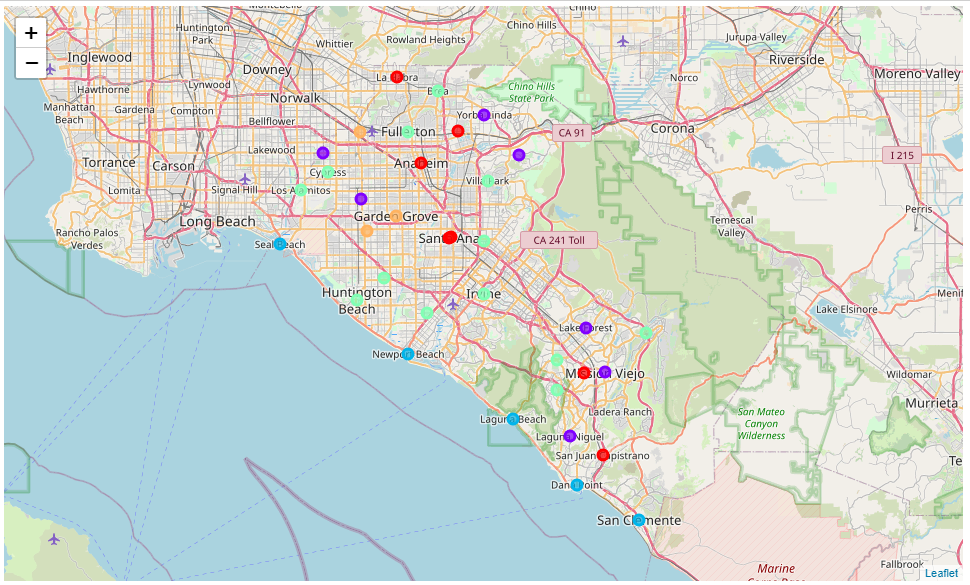


Figure 3.2 Clustering of cities based on Top 10 Most Common Venues

# **4 Results**

## **4.1 Three clusters of cities based on median household income and median housing price**

Cities in cluster 0 have mid-range median household income and mid-range median housing prices (Table 4.1).

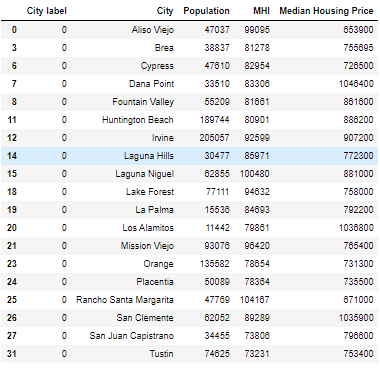


Table 4.1 Cluster 0

Cities in cluster 1 have the highest median housing prices; both are beach cities (Table 4.2).

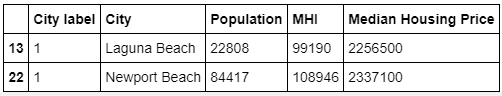


Table 4.2 Cluster 1

Cities in cluster 2 have high median household incomes and high median housing prices; these three cities are also geographically close to each other (Table 4.3).

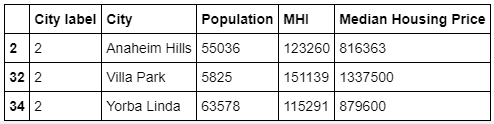


Table 4.3 Cluster 2

Cites in cluster 3 have relatively affordable housing prices for modest income families (Table 4.4).

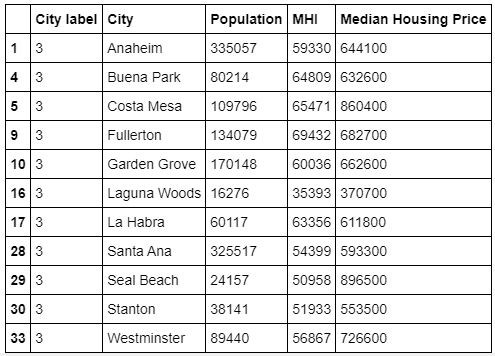


Table 4.4 Cluster 3

## **4.2 Five clusters of cities based on top 10 most common venues**

Cities in cluster 0 have a great varieties of food options (Mexican restaurant, coffee shop, pizza place, etc.); Mexican restaurants are the most popular venues.



Table 4.5 Cluster 0

Cities in cluster 1 have mainly fast food options (sandwich places, fast food restaurant, pizza place, etc.), as well as some outdoor venues like parks and trails.



Table 4.6 Cluster 1

Cities in cluster 2 are all beach cities.



Table 4.7 Cluster 2

Cites in cluster 3 have a variety of restaurant and fast food options, as well as other amenities such as grocery stores, gyms, etc.

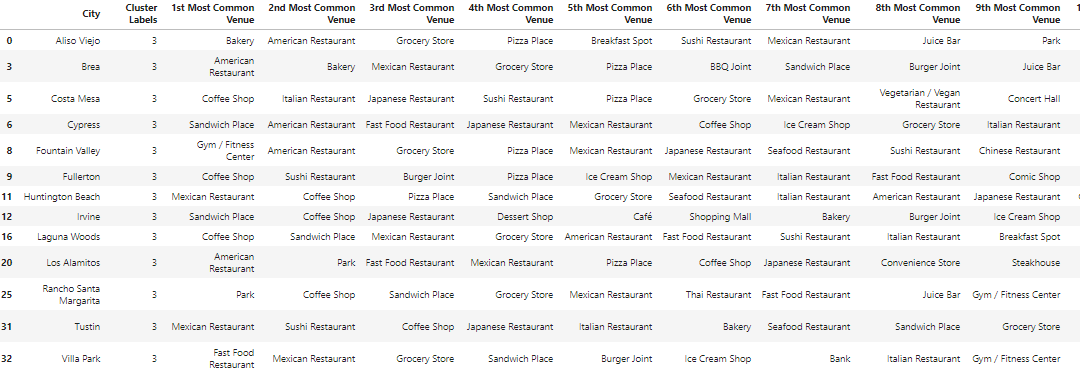


Table 4.8 Cluster 3

Cities in cluster 4 have predominantly Vietnamese and Korean restaurants.



Table 4.9 Cluster 4

# **5 Discussion**

The goal of the project is to help people who want to move to Orange County to make an informed decision that balances their budget and lifestyle preference. For example, people with a big budget can consider beach cities (Table 4.2 Cluster 1) while people who are looking for an afford place can take a look at Table 4.4 Cluster 3. For fashion lover, Irvine has the highest concentration of shopping malls; for people who like outdoor activities, they can look at cities where parks/trails are common; for foodies that are into Asian food, Buena Park/Garden Grove/Westminster have the highest concentration of Korean and Vietnamese food (Table 4.9 Cluster 4). Everyone has a place in Orange County!

# **6 Conclusion**

The project uses web scraping, data wrangling, Foursquare API and folium map to cluster and visualize the cities in Orange County, 1) based on household income and housing prices and 2) based on most common revenues. This way, people can have a better picture of a city and choose a neighborhood that fits their budget as well as their personal preferences. Some data sources can be updated or added in the future to better reflect the current trend: 1) the median household income should be updated after the 2020 US Census data becomes available; 2) Foursquare API “venues/explore” function only returns 100 venues at most; other types of API can be used to obtain more than 100 venues, so we have a more comprehensive understanding of a neighborhood that has a lot more venues.

# **References**

The image of Orange County is from <https://www.acg.org/occ>

[1] <https://www.moving.com/tips/us-moving-statistics-for-2019/>

[2] <https://en.wikipedia.org/wiki/Orange_County,_California>

[3] <https://www.zillow.com/orange-county-ca/home-values/>