Capstone Project: The battle of Neighborhoods in Orange County

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)

	Place	Type[40]	Population[41]	Per capita income[37]	Median household income[citation needed]	Median family income[39]
0	Aliso Viejo	City	47037	$44,646	$99,095	$113,183
1	Anaheim	City	335057	$23,109	$59,330	$63,180
2	Anaheim Hills	City	55036	$52,195	$123,260	$148,360
3	Brea	City	38837	$36,195	$81,278	$98,159
4	Buena Park	City	80214	$23,470	$64,809	$68,872

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).

	City	Median Housing Price
0	Aliso Viejo	653900
1	Anaheim	644100
2	Anaheim Hills	816363
3	Brea	755695
4	Buena Park	632600

Table 2.2 Orange County Median Housing Price 2019

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

	City	Population	MHI	Median Housing Price	Latitude	Longitude
0	Aliso Viejo	47037	99095	653900	33.576138	-117.725812
1	Anaheim	335057	59330	644100	33.834752	-117.911732
2	Anaheim Hills	55036	123260	816363	33.844408	-117.777386
3	Brea	38837	81278	755695	33.917044	-117.888856
4	Buena Park	80214	64809	632600	33.870413	-117.996217

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).

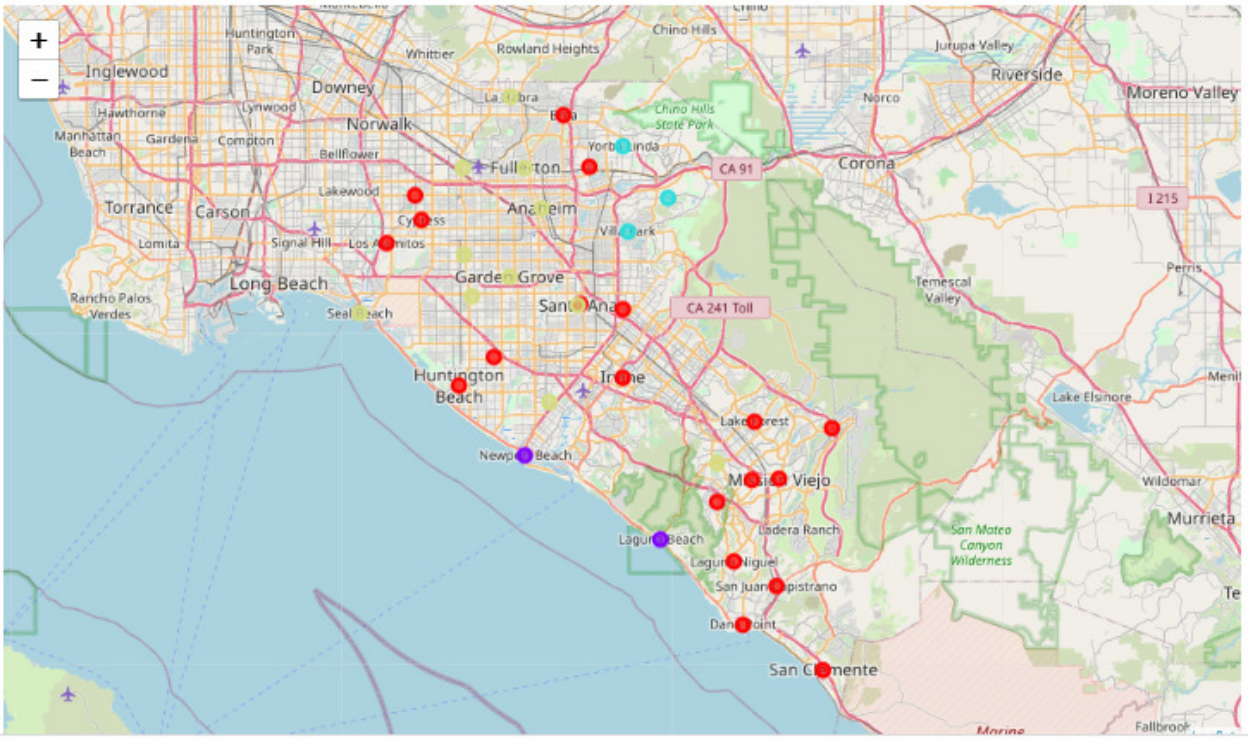

Figure 3.1 Clustering of cities based on median household income and median housing price

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).

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Category
0	Aliso Viejo	33.576138	-117.725812	Trader Joe's	33.577510	-117.724516	Grocery Store
1	Aliso Viejo	33.576138	-117.725812	Wahoo's Fish Taco	33.575093	-117.725684	Taco Place
2	Aliso Viejo	33.576138	-117.725812	Nekter Juice Bar	33.575438	-117.726253	Juice Bar
3	Aliso Viejo	33.576138	-117.725812	Kanpai Sushi	33.575449	-117.724856	Sushi Restaurant
4	Aliso Viejo	33.576138	-117.725812	Opah Restaurant	33.575239	-117.725065	Seafood Restaurant

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.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aliso Viejo	Grocery Store	Bakery	Pizza Place	American Restaurant	Juice Bar	Park	Mexican Restaurant	Sushi Restaurant	Breakfast Spot	Gymnastics Gym
1	Anaheim	Mexican Restaurant	Theme Park Ride / Attraction	Brewery	Coffee Shop	Fast Food Restaurant	Ice Cream Shop	Pizza Place	Taco Place	Burger Joint	Indian Restaurant
2	Anaheim Hills	Sandwich Place	Pizza Place	Trail	Fast Food Restaurant	Playground	Mexican Restaurant	Juice Bar	Burger Joint	Seafood Restaurant	Asian Restaurant
3	Brea	American Restaurant	Bakery	Pizza Place	Mexican Restaurant	Grocery Store	Burger Joint	Steakhouse	Juice Bar	Sandwich Place	BBQ Joint
4	Buena Park	Korean Restaurant	Coffee Shop	Fast Food Restaurant	Bakery	Pizza Place	Steakhouse	American Restaurant	Sandwich Place	Vietnamese Restaurant	Sushi Restaurant

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).

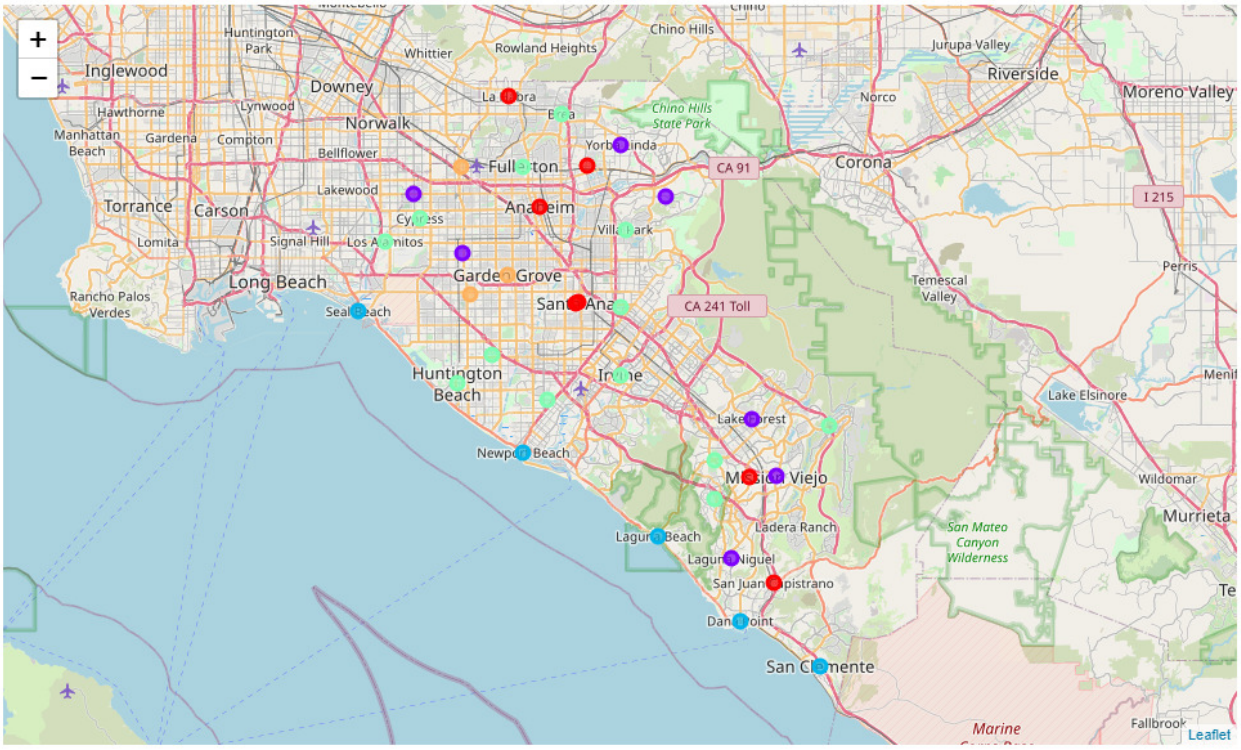

Figure 3.2 Clustering of cities based on Top 10 Most Common Venues

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).

	City label	City	Population	MHI	Median Housing Price
0	0	Aliso Viejo	47037	99095	653900
3	0	Brea	38837	81278	755695
6	0	Cypress	47610	82954	726500
7	0	Dana Point	33510	83306	1046400
8	0	Fountain Valley	55209	81661	861600
11	0	Huntington Beach	189744	80901	886200
12	0	Irvine	205057	92599	907200
14	0	Laguna Hills	30477	85971	772300
15	0	Laguna Niguel	62855	100480	881000
18	0	Lake Forest	77111	94832	758000
19	0	La Palma	15536	84893	792200
20	0	Los Alamitos	11442	79861	1036800
21	0	Mission Viejo	93076	96420	765400
23	0	Orange	135582	78654	731300
24	0	Placentia	50089	78364	735500
25	0	Rancho Santa Margarita	47769	104167	671000
26	0	San Clemente	62052	89289	1035900
27	0	San Juan Capistrano	34455	73806	796600
31	0	Tustin	74625	73231	753400

Table 4.1 Cluster 0

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

	City label	City	Population	MHI	Median Housing Price
13	1	Laguna Beach	22808	99190	2256500
22	1	Newport Beach	84417	108946	2337100

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).

	City label	City	Population	MHI	Median Housing Price
2	2	Anaheim Hills	55036	123260	816363
32	2	Villa Park	5825	151139	1337500
34	2	Yorba Linda	63578	115291	879600

Table 4.3 Cluster 2

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

	City label	City	Population	MHI	Median Housing Price
1	3	Anaheim	335057	59330	644100
4	3	Buena Park	80214	64809	632600
5	3	Costa Mesa	109796	65471	860400
9	3	Fullerton	134079	69432	682700
10	3	Garden Grove	170148	60036	662600
16	3	Laguna Woods	16276	35393	370700
17	3	La Habra	60117	63356	611800
28	3	Santa Ana	325517	54399	593300
29	3	Seal Beach	24157	50958	896500
30	3	Stanton	38141	51933	553500
33	3	Westminster	89440	56867	726600

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.

	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
1	Anaheim	0	Mexican Restaurant	Theme Park Ride / Attraction	Brewery	Coffee Shop	Fast Food Restaurant	Ice Cream Shop	Taco Place	Pizza Place
14	Laguna Hills	0	Mexican Restaurant	Pizza Place	Coffee Shop	Bank	Donut Shop	Fast Food Restaurant	Breakfast Spot	Sushi Restaurant
17	La Habra	0	Mexican Restaurant	Pizza Place	Fast Food Restaurant	Coffee Shop	Grocery Store	Sandwich Place	Convenience Store	Burger Joint
23	Orange	0	Mexican Restaurant	Coffee Shop	Convenience Store	American Restaurant	Bar	Café	Sandwich Place	Taco Place
24	Placentia	0	Mexican Restaurant	Brewery	Pizza Place	Coffee Shop	Fast Food Restaurant	BBQ Joint	Supermarket	Breakfast Spot
27	San Juan Capistrano	0	Mexican Restaurant	Coffee Shop	Restaurant	American Restaurant	Farm	Garden	Grocery Store	Furniture / Home Store
28	Santa Ana	0	Mexican Restaurant	Convenience Store	Coffee Shop	Bar	American Restaurant	Video Game Store	Sandwich Place	Café

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.

	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9t Co Ve
2	Anaheim Hills	1	Sandwich Place	Pizza Place	Coffee Shop	Trail	Fast Food Restaurant	Burger Joint	Donut Shop	Juice Bar	Th Re
15	Laguna Niguel	1	Park	Grocery Store	Pharmacy	Pizza Place	Fast Food Restaurant	Mexican Restaurant	ATM	Sushi Restaurant	Sh St
18	Lake Forest	1	Sandwich Place	Pizza Place	Fast Food Restaurant	Grocery Store	Park	Coffee Shop	Japanese Restaurant	Sushi Restaurant	AT
19	La Palma	1	Sandwich Place	Mexican Restaurant	Fast Food Restaurant	Coffee Shop	Korean Restaurant	Pizza Place	Shipping Store	Pharmacy	Pa
21	Mission Viejo	1	Pizza Place	Coffee Shop	Sandwich Place	Mexican Restaurant	Donut Shop	Convenience Store	Italian Restaurant	Ice Cream Shop	Fa Re
30	Stanton	1	Fast Food Restaurant	Convenience Store	Coffee Shop	Sandwich Place	Pizza Place	Sushi Restaurant	Japanese Restaurant	American Restaurant	Gr St
34	Yorba Linda	1	Sandwich Place	Grocery Store	Mexican Restaurant	Fast Food Restaurant	Pizza Place	Pharmacy	Convenience Store	Gym	Br Sp

Table 4.6 Cluster 1

Cities in cluster 2 are all beach cities.

	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mos Comm Venue
7	Dana Point	2	American Restaurant	Beach	Mexican Restaurant	Park	Bar	Coffee Shop	Hotel	Seafood Restaurant	Board Shop
13	Laguna Beach	2	Beach	American Restaurant	Seafood Restaurant	New American Restaurant	Mexican Restaurant	Italian Restaurant	Art Gallery	Sushi Restaurant	Resort
22	Newport Beach	2	Beach	Seafood Restaurant	Coffee Shop	Italian Restaurant	American Restaurant	Bar	Restaurant	Donut Shop	Taco Place
26	San Clemente	2	Mexican Restaurant	Clothing Store	Beach	Pizza Place	Coffee Shop	American Restaurant	Chinese Restaurant	Burger Joint	Sandwi Place
29	Seal Beach	2	Coffee Shop	Mexican Restaurant	Seafood Restaurant	Grocery Store	Sushi Restaurant	Beach	Cosmetics Shop	BBQ Joint	Breakfa Spot

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.

	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Aliso Viejo	3	Bakery	American Restaurant	Grocery Store	Pizza Place	Breakfast Spot	Sushi Restaurant	Mexican Restaurant	Juice Bar	Park
3	Brea	3	American Restaurant	Bakery	Mexican Restaurant	Grocery Store	Pizza Place	BBQ Joint	Sandwich Place	Burger Joint	Juice Bar
5	Costa Mesa	3	Coffee Shop	Italian Restaurant	Japanese Restaurant	Sushi Restaurant	Pizza Place	Grocery Store	Mexican Restaurant	Vegetarian / Vegan Restaurant	Concert Hall
6	Cypress	3	Sandwich Place	American Restaurant	Fast Food Restaurant	Japanese Restaurant	Mexican Restaurant	Coffee Shop	Ice Cream Shop	Grocery Store	Italian Restaurant
8	Fountain Valley	3	Gym / Fitness Center	American Restaurant	Grocery Store	Pizza Place	Mexican Restaurant	Japanese Restaurant	Seafood Restaurant	Sushi Restaurant	Chinese Restaurant
9	Fullerton	3	Coffee Shop	Sushi Restaurant	Burger Joint	Pizza Place	Ice Cream Shop	Mexican Restaurant	Italian Restaurant	Fast Food Restaurant	Comic Shop
11	Huntington Beach	3	Mexican Restaurant	Coffee Shop	Pizza Place	Sandwich Place	Grocery Store	Seafood Restaurant	Italian Restaurant	American Restaurant	Japanese Restaurant
12	Irvine	3	Sandwich Place	Coffee Shop	Japanese Restaurant	Dessert Shop	Café	Shopping Mall	Bakery	Burger Joint	Ice Cream Shop
16	Laguna Woods	3	Coffee Shop	Sandwich Place	Mexican Restaurant	Grocery Store	American Restaurant	Fast Food Restaurant	Sushi Restaurant	Italian Restaurant	Breakfast Spot
20	Los Alamitos	3	American Restaurant	Park	Fast Food Restaurant	Mexican Restaurant	Pizza Place	Coffee Shop	Japanese Restaurant	Convenience Store	Steakhouse
25	Rancho Santa Margarita	3	Park	Coffee Shop	Sandwich Place	Grocery Store	Mexican Restaurant	Thai Restaurant	Fast Food Restaurant	Juice Bar	Gym / Fitness Center
31	Tustin	3	Mexican Restaurant	Sushi Restaurant	Coffee Shop	Japanese Restaurant	Italian Restaurant	Bakery	Seafood Restaurant	Sandwich Place	Grocery Store
32	Villa Park	3	Fast Food Restaurant	Mexican Restaurant	Grocery Store	Sandwich Place	Burger Joint	Ice Cream Shop	Bank	Italian Restaurant	Gym / Fitness Center

Table 4.8 Cluster 3

Cities in cluster 4 have predominantly Vietnamese and Korean restaurants.

	City	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Mo Comm Venue
4	Buena Park	4	Korean Restaurant	Coffee Shop	Fast Food Restaurant	Bakery	Pizza Place	Sandwich Place	Steakhouse	Bar	Vietnar Restau
10	Garden Grove	4	Vietnamese Restaurant	Mexican Restaurant	Coffee Shop	Bar	Ice Cream Shop	Bakery	Korean Restaurant	Bubble Tea Shop	Seafoo Restau
33	Westminster	4	Vietnamese Restaurant	Chinese Restaurant	Korean Restaurant	Coffee Shop	Fast Food Restaurant	Bakery	Seafood Restaurant	Discount Store	Bank

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/