**Juhwan Kim**

**Best location to start a new veterinary clinic in California**

**1. Introduction**

1.1 Background

The united states (US) has the greatest number of pet animals in the world [1]. Based on American veterinary medical association’s 2012 statistics, there are ~ 155 millions of pets in the united states, and each household has ~2 pets [2]. Total number of cats and dogs in the US are more than the sum of 2nd and 3rd ranked countries [1]. As the demand of providing the good quality of life to companion animals are growing, the veterinary services in the US industry has grown by 5.4% annually between 2014-2019 [3, 4, 5]. However, starting a private veterinary practice is risky, and needs to consider many factors such as location, real estate, population, living of cost, and competition between the clinics. Therefore, it is advantageous to accurately predict which neighborhood is the good candidate to start the business.

1.2 Problem

This project aims to find the best location candidates for opening veterinary clinics in California. Data that might contribute to determining the best location for opening new practice including number of clinics, median household income, population in the neighborhoods in California was used.

**2. Data acquisition and cleaning**

2.1 Data sources

A dataset containing coordinates, Zip code, city name, population of the US [6]. Housing data of California containing coordinates and median house values were found from Kaggle dataset [7]. I collected 2012 IRS income by zip code of the US from Data.world [8]. To create the choropleth map of the cities of California, I downloaded California cities GeoJSON shapefile [9].

2.2 Data cleaning and preparation

I sliced California data from the first dataset containing coordinates, zip code, city name, and population of the US. I used this dataset for searching veterinary clinics in California.

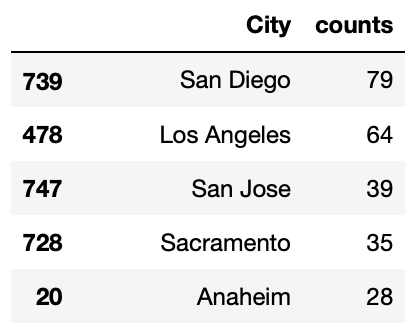
For finding the best neighborhood in California, I merged this dataset with the second dataset containing zip code and house value. I cleaned the minor cities sharing the same coordination and left the primary cities to prevent redundant location data. This process also cleared the duplicated population values since the minor cities with same coordinates of the major cities share same population values. I merged this data with the third dataset containing average income by zip code and removed the cities which contain NaN value. The final dataset contains

coordinates, zip code, city name, population, house value, income. House value and income are described as thousand dollars in the dataset. To get the name of veterinary clinics in California, I used Foursquare API [10].

**3. Methodology**

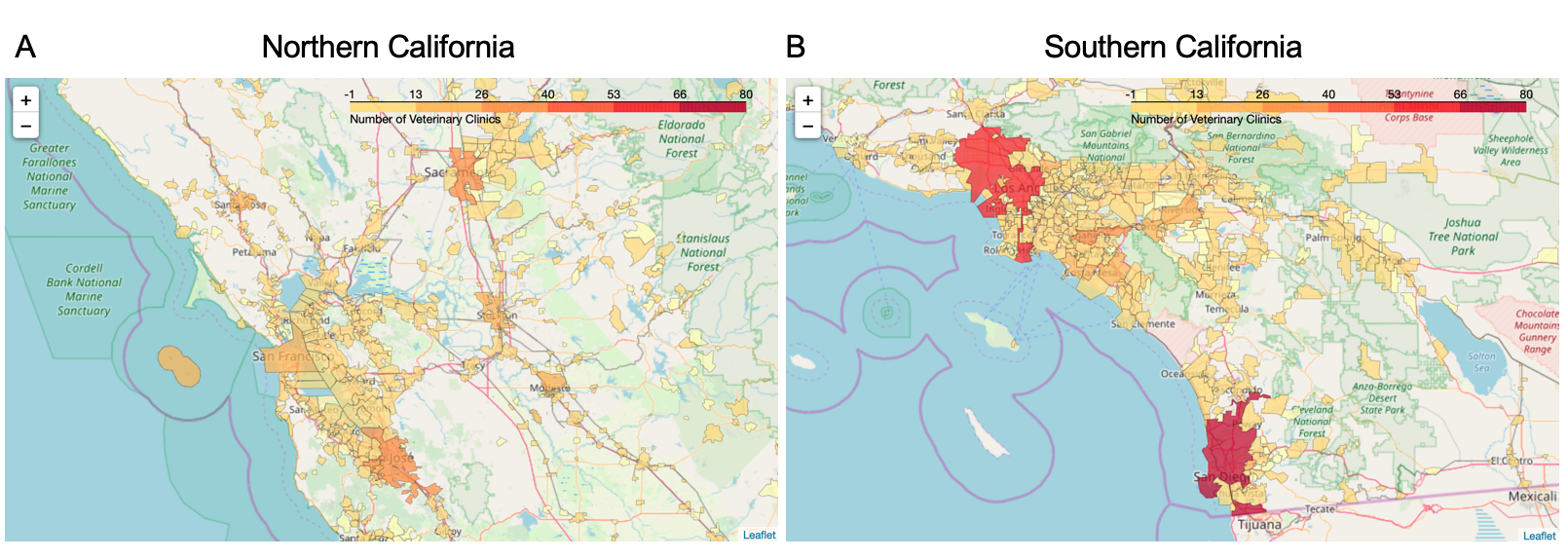
3.1 Distribution of veterinary clinics in California

To obtain the number of veterinary clinics in each city, I utilized Foursquare API. I used ‘Veterinary’, ‘Pet Hospital’, ‘Pet Clinic’, ‘Animal Hospital’, ‘Animal medical’, ‘DVM’ as search query to get the name of veterinary clinics in each neighborhood. I made the dataframe by appending the name of the clinics to the dataframe containing coordinates, zip codes, city names. Big cities like Los Angeles and San Diego have many zip codes in my combined dataset (i.g. Los Angeles, 63; San Diego, 57). Given the size of the Los Angeles is 1300km2, theoretically 2564 meters radius can cover whole area of Los Angeles. Therefore, I applied 3000m as radius of search area for each Zip codes and removed the duplicates name of the clinics in the results. The rows don’t contain the search keywords were removed. I dropped the rows don't contain the keywords in the results. I grouped the results by the coordinates and counted the number of the clinics. I appended these data to the dataframe containing coordinates, zip codes and city names. I merged each neighborhood by the name of the cities and summed the counts of the clinics. Therefore, the final dataframe contains the counts of the clinics in each city in California. Based on the search results, San Diego has the most clinics in California and followed by Los Angeles (Table 1.)



**Table 1. Top 5 cities has the most clinics in California.**

I used Folium package to create a choropleth map of the number of the veterinary clinics in California (Figure 1).



**Figure 1. Distribution of veterinary clinics in (A) northern California and (B) southern California.**

3.2 Explore the best neighborhoods

To open the new business, I considered three factors, affordable rent of the clinic, high household income and high volume of customers available in the neighborhood to be the most important. I standardized the values of population, Income and median house value by Standardscaler and used KMeans unsupervised technique to cluster the neighborhoods by the optimal k number (Cluster numbers, 4) by the Elbow method (Figure 2.)



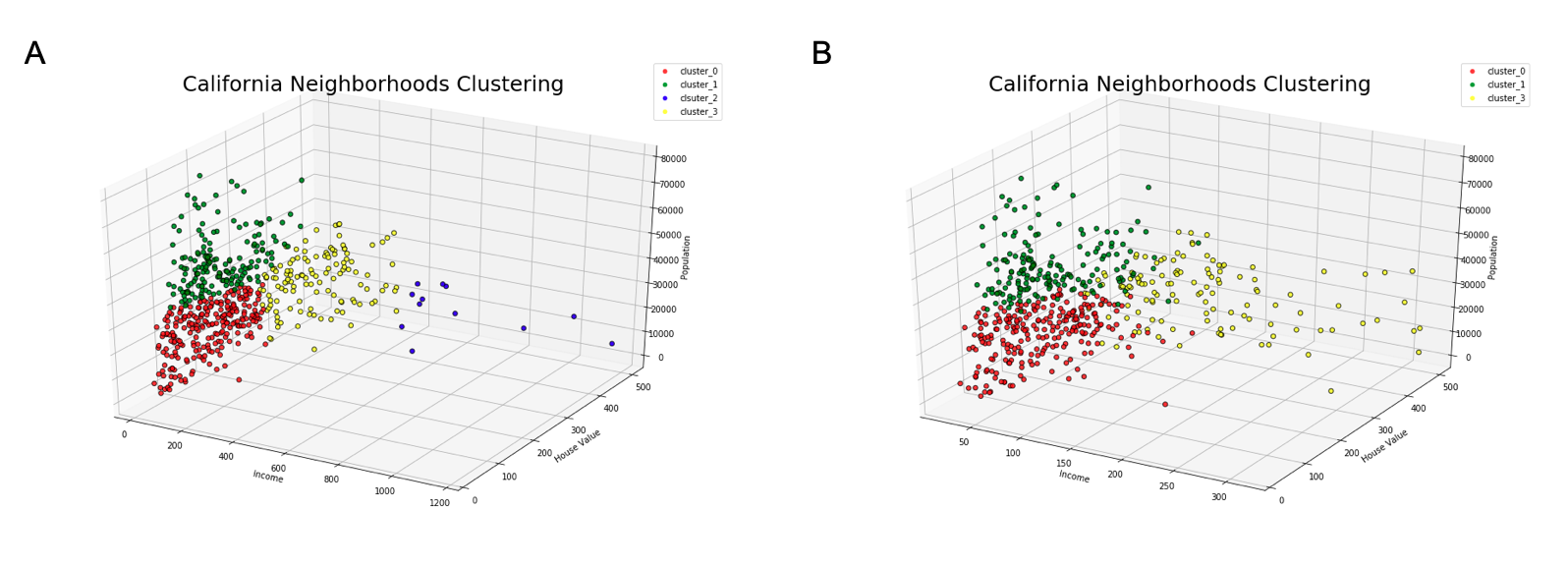
**Figure 2. Elbow method to find the optimal k value for KMeans clustering.**

**Results**

I visualized all 4 clusters by 3d scatter plot (Figure 3A). I found neighborhoods in cluster 2 are very rich based on their income and house value. Since these neighborhoods also have very low population, this cluster may not be suitable to start the new business. Furthermore, it is unclear to see the differences of income values since neighborhoods have approximately 10-20 times higher income. Therefore, I removed this cluster to get better visualization of other clusters (Figure 3B). Although neighborhoods in cluster 0 have higher income values, house value are relatively higher, and population is lower than other two clusters. Cluster 3 has low income value and low population. Cluster 1 has broad distribution of house value and income. Noticeably, these neighborhoods has higher population than others. Therefore, I chose cluster 1 as the best neighborhoods in California to start new business. I sliced Cluster 1 to find the veterinary clinics near these neighborhoods and obtained 170 neighborhoods (Table 2).

For the next steps, I utilized Foursquare API to find the name of veterinary clinics. I used ‘Veterinary’, ‘Pet Hospital’, ‘Pet Clinic’, ‘Animal Hospital’, ‘Animal medical’, ‘DVM’ as search query to get the name of veterinary clinics in each neighborhood.

I made the dataframe by appending the name of the clinics to the dataframe containing coordinates, zip codes, city names. I also applied 3000 meters as radius of search area for each zip codes and removed the duplicates name of the clinics in the results. The rows don’t contain the search keywords were removed. I grouped the results by the coordinates and



**Figure 3. California Neighborhoods clustering.** (A) Visualization of all 4 clusters. (B) Visualization of cluster 0,1, and 2.



**Table 2. Top 20 neighborhoods in cluster 1.**

counted the number of the clinics. I appended these data to the dataframe containing coordinates, zip codes and city names. I sorted the data by counts of the clinic in each neighborhood, and acquired top 50 neighborhoods have the most clinics, and top 50 neighborhoods have the least clinics (Figure 4-5).

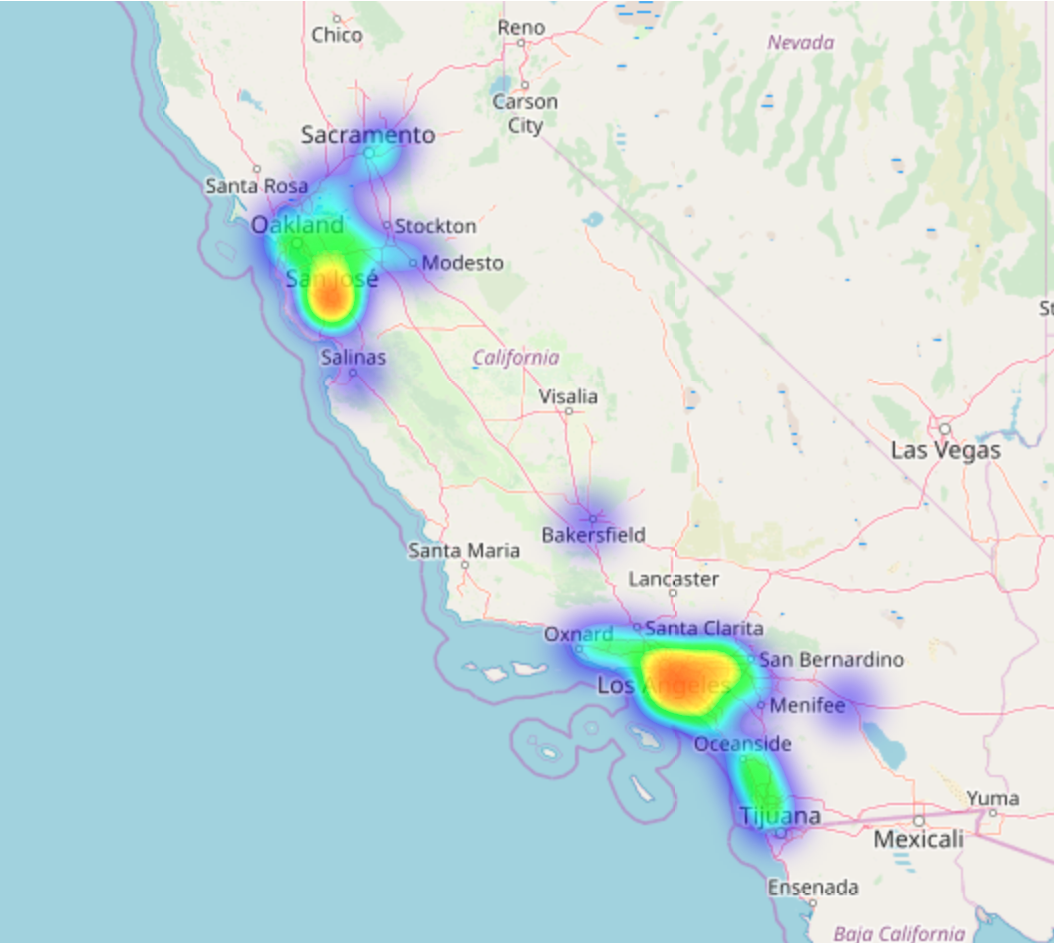


**Figure 4. Top 50 neighborhoods in cluster 1 having the most veterinary clinics.**



**Figure 5. Top 50 neighborhoods in cluster 1 having the least veterinary clinics.**

I visualized the 170 neighborhoods by heatmap to see the distribution of cluster 1 neighborhoods (Figure 6). I marked the City name, Zip code, numbers of clinics in the neighborhoods on the maps (Figure 7).



**Figure 6. Distribution of neighborhoods in cluster 1 in California.** Heatmap indicates frequencies of neighborhoods.



**Figure 7. Best neighborhoods for starting veterinary clinic in California.** Circle markers shows City name, Zip code, numbers of clinics. Color of Circle markers is described as yellow to red color gradient indicating the number of clinics within 3000m radius (yellow = 0, Red = 30).

**Discussion**

Finding a good location is one of the most important steps to be considered when to start a new business. I used the demographic data to predict the best candidates. However, this approach can be overly simplified. In real world, there are many factors for starting the best veterinary clinics location. For example, in the large city like Los Angeles, people might prefer to use public transportation because of the traffic. In this case, people might consider the clinics located near the subway station would be best. Similarly, in case cities don't have enough parking spaces available, people might prefer to the clinics which have ample parking spaces. In the future study, it would be interesting to consider the proximity to the transportation station to find the best location in the big cities like Los Angeles, San Diego, and San Francisco. Furthermore, diverse demographic data like race, age of households and estimated pet numbers in each neighborhood can be used for targeting specific customers.

**Conclusion**

This project shows potential best locations to start the new business in California based on a combination of location-based API and demographic data.

**References**

[1] https://www.stevedalepetworld.com/blog/world-pet-population-data-mixed-bag/

[2] https://www.avma.org/KB/Resources/Statistics/Pages/Market-research-statistics-US-pet-ownership.aspx

[3] https://www.ftc.gov/sites/default/files/documents/public\_events/pet-medications-workshop/daspros.pdf

[4] https://www.aavmc.org/assets/site\_18/files/annual%20reports/v3\_econ\_2016\_report3\_mketvet\_061416.pdf

[5] https://www.ibisworld.com/industry-trends/market-research-reports/professional-scientific-technical-services/professional-scientific-technical-services/veterinary-services.html

[6] http://federalgovernmentzipcodes.us

[7] https://www.kaggle.com/camnugent/california-housing-prices

[8] https://data.world/jonloyens/irs-income-by-zip-code

[9] http://boundaries.latimes.com/sets/

[10] https://foursquare.com