Capstone Project

California Zoning

Coursera IBM Certified Data Scientist

# Business Problem

A department from the municipality of California wants to propose new classifications to various areas within the state. The team wants this classification to be based on housing and venues within each zone. They are not sure what venue data is available, but they have access to a dataset containing each zone and housing information.

This problem can be converted into a clustering problem. The clusters will be based on the housing features and number of each venue type. The resulting clusters will then be the new zone classification the department will use in the future.

# Data

There are two data sources that will be used to solve his problem. The first data set come from a paper Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297 (<https://github.com/ageron/handson-ml/tree/master/datasets/housing>). This data set contains around 20k zones and their housing information. The features for each zone are:

* longitude
* latitude
* housing median age
* total rooms
* total bedrooms
* population
* households
* median income
* median house value
* ocean proximity

longitude and latitude will be used as key to match the data with the second data set, which comes from Foursquare. The data will consist of the different types of venues near a set of coordinates.

Once combined together, the two data sets will contain enough information to cluster the zone into related groups based on the housing information and nearby venues. Since the Foursquare API limits us to 950 calls a day, the scope of this project will be limited to the first 500 zones in the data.

# Methodology

Since this is a clustering project, the work done is broken down into two stages. The first stage is enriching the housing data set with nearby venue categories and then clustering the locations based on the new features. After that, some analysis was done to identify what features mostly affect the grouping of zones.

## Feature Engineering

Initially, the data containing the housing data is loaded with the following features:

* longitude
* latitude
* housing median age
* total rooms
* total bedrooms
* population
* households
* median income
* median house value
* ocean proximity

Then based on every latitude and longitude pair, a call to the foursquare venue explore API. Each call returns a list of venues near a particular location. For performance reasons, the call was limited to 100 venues per location within a 500m radius. Then for each list of nearby venues, a one-hot coded representation is obtained, containing each latitude and longitude pair and the count for each venue type found. After iterating through all the 500 zones, there was found a total of 308 unique venue types. Resulting shape of the data is now 500 rows by 320 columns.

## Clustering

With this enriched data set, K-means clustering was used to cluster the data into groups. Using sum of squared distances (SSE) between clusters for different number of clusters, it was possible to extract the ideal number of clusters for k-means. Figure 1 shows the elbow at around 3 or for clusters, where there is significant reduction in SSE between the clusters. Thus the number of clusters for K-means is taken to be 4.

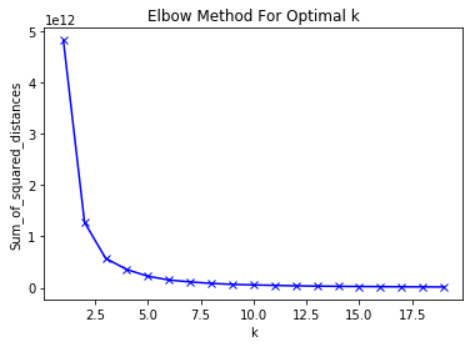


Figure 1 SSE for different number of K in K-Means

Figure 2 shows the distribution of clusters, where it shows that cluster 0 has the most members and cluster 3 has the least. However, all the clusters have enough members and none of the clusters have a significantly smaller number of members.

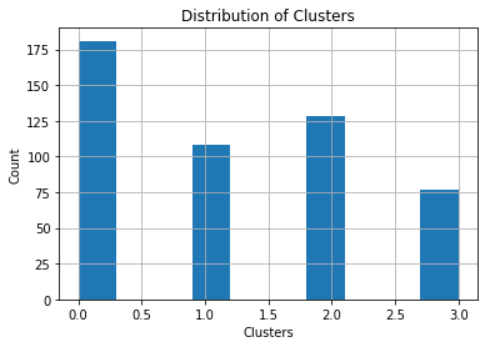


Figure 2 Distribution of Clusters

# Results

With clusters obtained, the clusters on the 500 zones were plotted on a map, showing a clear pattern in the data. As seen in Figure 3, the map shows an apparent pattern in the cluster distribution. Cluster 3 is closest to the bay, then cluster 1 and 2 are further inland, with cluster 0 being the closest to the canyons and furthest from the bay.

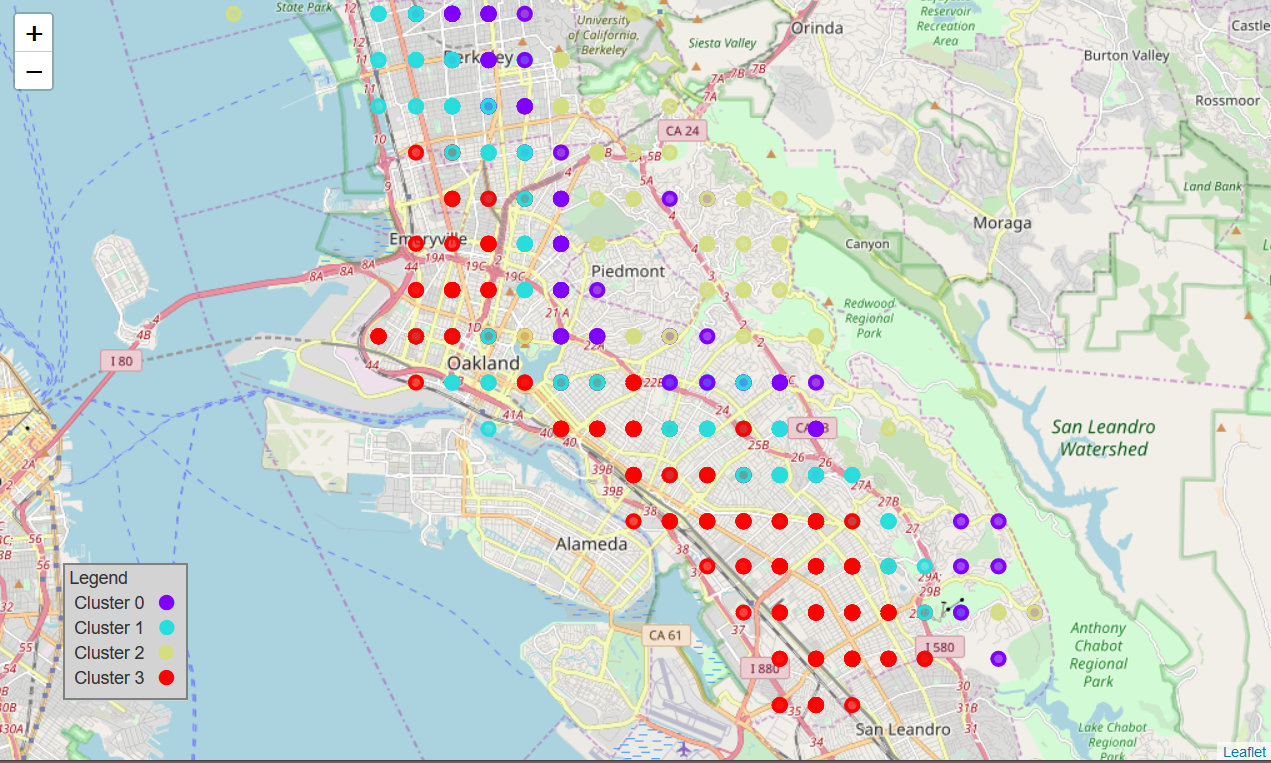


Figure 3 Cluster Map

On top of the map results, there also seem to be some patterns in the correlation between some of the features and the clusters. The median house value seems to have the highest correlation, with cluster 3 and 2 having the higher value. Which makes sense considering they are either closest to the bay or closest to the national park. The highest inverse correlation lies with food. Indicating that zones 0 and 1, which are closer inland but further from the national parks tend to have more food venues. Which is reasonable, considering these zones would define the “downtown” or zones. It is notable that the median house value, median, income, and total rooms (which are all features relating to housing) have higher correlations than count of nearby venue types. This indicates that housing information plays a large role in how similar the zones are.

Table 1 Correlations with the clusters

|  |  |
| --- | --- |
| Feature | Correlation with Clusters |
| median\_house\_value | 0.751223 |
| median\_income | 0.565441 |
| Latitude | 0.471349 |
| total\_rooms | 0.230474 |
| Sushi Restaurant | 0.220535 |
| Gift Shop | 0.204207 |
| Trail | 0.195111 |
| Pool | 0.189966 |
| Bagel Shop | 0.185673 |
| Food | -0.215834 |
| Outdoors & Recreation | -0.140647 |
| Nail Salon | -0.132775 |

# Conclusion

In this report, we discussed a method to group zones in California based on housing data and nearby venues. Based on this analysis, we concluded that we would be able to break down units into 4 clusters. Key differentiators between the clusters is the proximity to the bay, whereas key influencers for each group include the median housing value and median income, and the key inverse correlation is the availability of food venues. These results should now help the municipality of California in zoning areas around the state and classifying new areas.