# California City Segmentation for Real Estate Industry

Rui Luo

May 12 2019

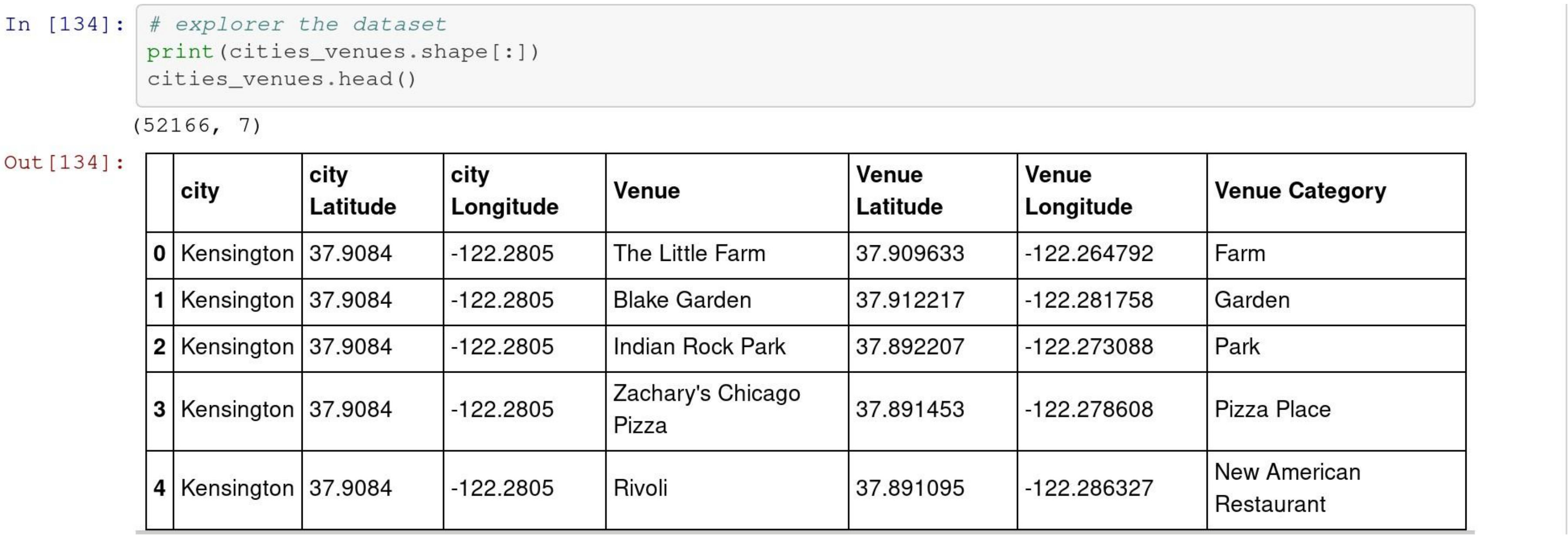
# City Segmentation is important for decision making

- This project aims to cluster a group of cities in California in order to split the cities into different groups so that the cities within a group are very similar type.
- Obtaining venues data for each city is the necessary process, and the cities will be grouped based on the most common ten venue types of each city.
- Real estate companies would be very interested in city segmentation since they could utilize this to find similar cities and build business strategy that can be applied to these similar cities.
- This could reduce the cost and potentially increase the profits.

# Data Acquisition and Cleaning

- California cities data can be obtained from simplemaps.com
- The venues data for each city can be obtained from Foursquare
- City name, latitude, longitude, and population density for each city
- Cities with 1000.0 or higher density
- Dataset will be sorted into a new dataset that has city name, 1st most common venue type, 2nd most common venue type, until 10th most common venue type.

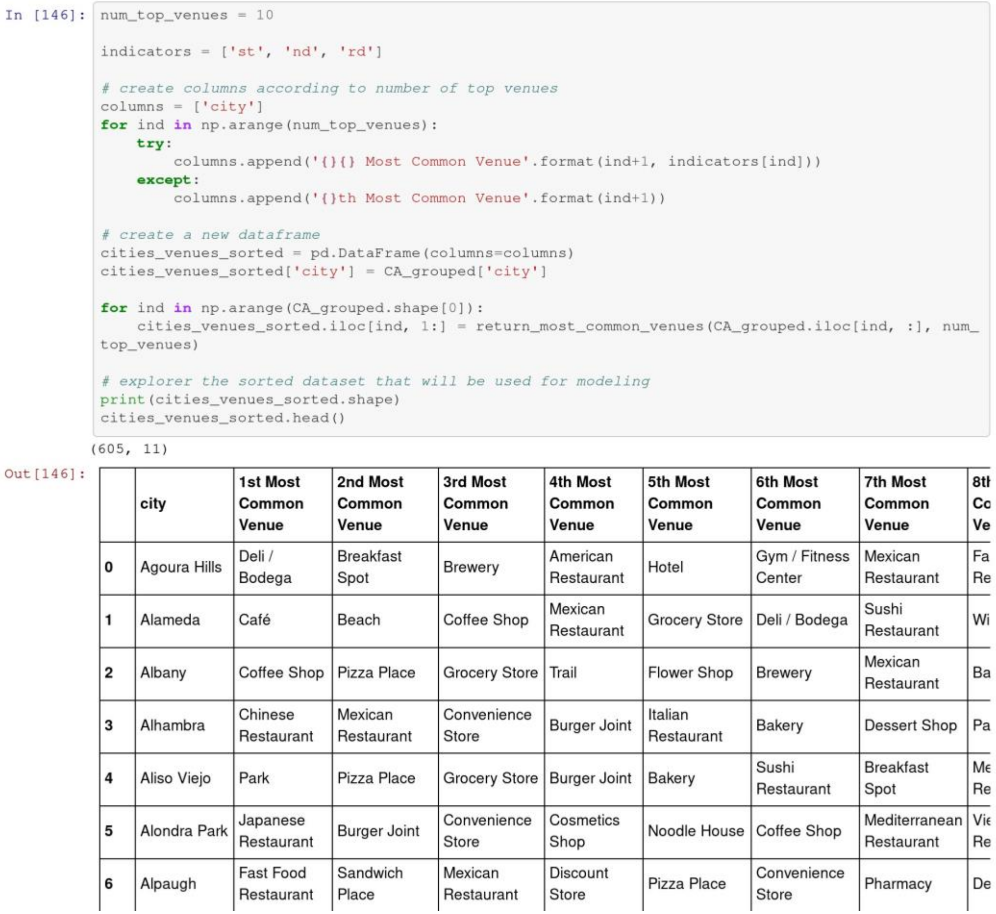
# Gathering venues for each city from Foursquare API



```
In [134]: # explorer the dataset
          print(cities_venues.shape[:])
          cities_venues.head()

          (52166, 7)

Out[134]:
```

| | city | city Latitude | city Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **0** | Kensington | 37.9084 | -122.2805 | The Little Farm | 37.909633 | -122.264792 | Farm |
| **1** | Kensington | 37.9084 | -122.2805 | Blake Garden | 37.912217 | -122.281758 | Garden |
| **2** | Kensington | 37.9084 | -122.2805 | Indian Rock Park | 37.892207 | -122.273088 | Park |
| **3** | Kensington | 37.9084 | -122.2805 | Zachary's Chicago Pizza | 37.891453 | -122.278608 | Pizza Place |
| **4** | Kensington | 37.9084 | -122.2805 | Rivoli | 37.891095 | -122.286327 | New American Restaurant |

# Integrated into a dataset that will be used for k-means



In [146]:

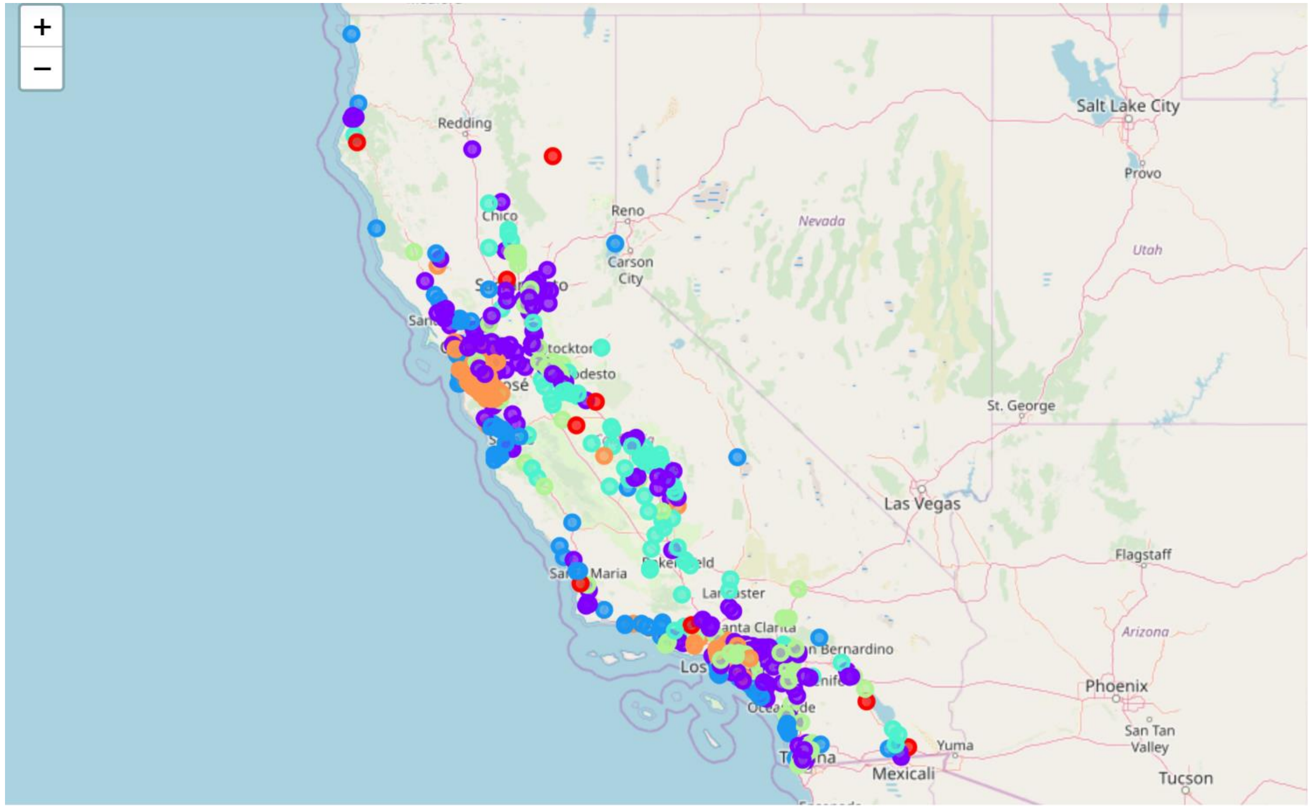
```python
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['city']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
cities_venues_sorted = pd.DataFrame(columns=columns)
cities_venues_sorted['city'] = CA_grouped['city']

for ind in np.arange(CA_grouped.shape[0]):
    cities_venues_sorted.iloc[ind, 1:] = return_most_common_venues(CA_grouped.iloc[ind, :], num_
top_venues)

# explorer the sorted dataset that will be used for modeling
print(cities_venues_sorted.shape)
cities_venues_sorted.head()
```

```
(605, 11)
```

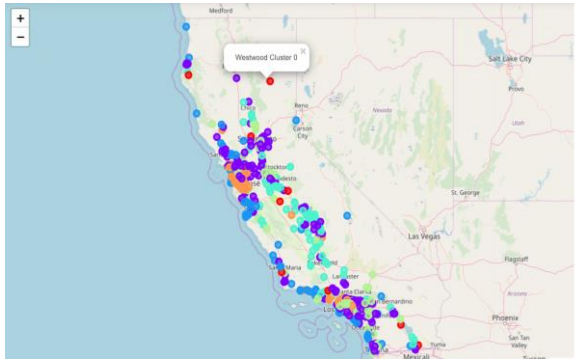
Out[146]:

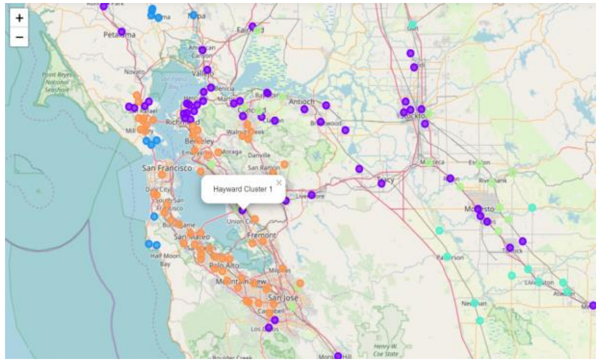
| | city | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8tl Co Ve |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Agoura Hills | Deli / Bodega | Breakfast Spot | Brewery | American Restaurant | Hotel | Gym / Fitness Center | Mexican Restaurant | Fa Re |
| **1** | Alameda | Café | Beach | Coffee Shop | Mexican Restaurant | Grocery Store | Deli / Bodega | Sushi Restaurant | Wi |
| **2** | Albany | Coffee Shop | Pizza Place | Grocery Store | Trail | Flower Shop | Brewery | Mexican Restaurant | Ba |
| **3** | Alhambra | Chinese Restaurant | Mexican Restaurant | Convenience Store | Burger Joint | Italian Restaurant | Bakery | Dessert Shop | Pa |
| **4** | Aliso Viejo | Park | Pizza Place | Grocery Store | Burger Joint | Bakery | Sushi Restaurant | Breakfast Spot | Me Re |
| **5** | Alondra Park | Japanese Restaurant | Burger Joint | Convenience Store | Cosmetics Shop | Noodle House | Coffee Shop | Mediterranean Restaurant | Vie Re |
| **6** | Alpaugh | Fast Food Restaurant | Sandwich Place | Mexican Restaurant | Discount Store | Pizza Place | Convenience Store | Pharmacy | De |

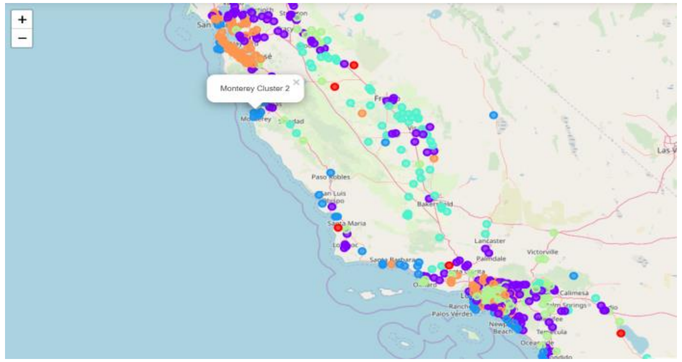
# Clustering results – 6 groups

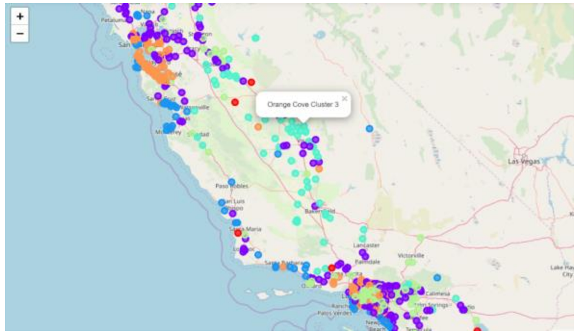


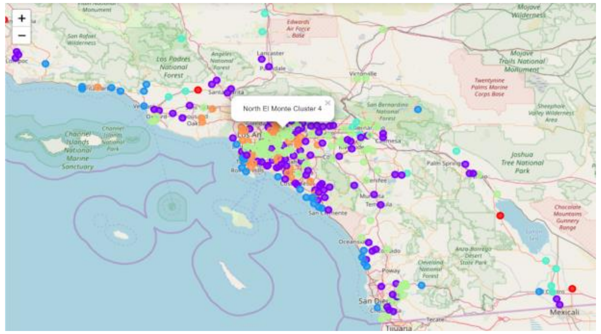
# Groups overview

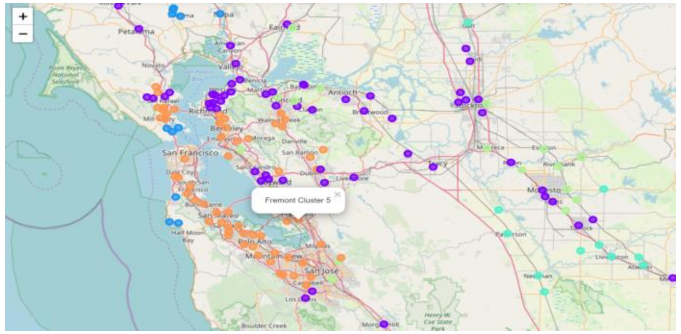












# Conclusion

- Acquired California city data from Simplemaps
- Extracted the cities with 1000.00 or higher population density
- Retrieved venues of each city from FourSquare API and integrated them with city data
- K-means machine learning technique and 6 clusters of cities
- Real Estate company could use the results to apply strategy to the similar cities