### Housing Market in California during Pandemic & Predictions 2020

Zhendi Zhou

8/21/2020

1. **Introduction**

**1.1 Business Problem**

One might assume that in the middle of the coronavirus pandemic, when businesses are closed, one out of seven California workers are unemployed, and everyone is staying at home... one market that would be in complete decline would be housing.

But surprising enough, according to the real estate website Redfin, traffic is up about 40 percent, and the National Association of Realtors announced that pending home sales are up 15 percent from June to July, 2020.

What is the trend regarding area preference given that the way we work — or that some of us work — has fundamentally changed due to the pandemic?

This project will be taking a deep dive into the housing sales data collected by Realtor.com in combination with the venue data from Foursquare API in order to answer this question.

**1.2 Stakeholders (Target Audience)**

There's a widening gap about economy state between white-collar workers who can do their jobs from anywhere — and those who cannot. Understanding that mortgage interest rates are falling, and may drop further according to Forbes, California real estate investors and home-buyers will find it more affordable to buy a property during this time.

Choosing positive cash flow properties is the top priority for those who are looking for a home or considering investing into the real estate market in 2020. Real estate companies would also want to clientele using this information.

**1.3 Success Criteria**

The success criteria of the project will be a good recommendation of neighborhood in California according to the current trend due to pandemic.

1. **Data**

**2.1 Description of Data**

* Housing sales data contains areas/zip codes and hotness downloaded from Realtors.com (https://econdata.s3-us-west-2.amazonaws.com/Reports/Hotness/RDC\_Inventory\_Hotness\_Metrics\_Zip\_History.csv)

Realtor.com Market Hotness Index: scores and rankings based on days on market (supply index) and realtor.com views per property (demand index).

Monthly data updated on July 30th, 2020 with data through July 2020.

* zip code, latitude & longitude data downloaded from online open source Gaslamp.

* Venue data from Foursquare API

Notice that the data from Realtor.com is grouped by zip code, we will also retrieve the data by zip code & their corresponding longitude and latitude values from API.

**2.2 Methodology**

* Collect the Housing sales data from realtors.com data center.
* Select Top 20 zip codes based on Hotness score.
* Visualize the Ranking of zip codes by using Python Folium library.
* Using Foursquare API we will find all venues/amenities for each neighborhood(zip code).
* Create and investigate clusters (using k-means clustering) for neighborhoods to give recommendations to homebuyers.
* Compare 2020 & 2019 information to forecast trends caused by pandemic.

**2.3 Data Cleaning**

After downloading the dataset needed from Realtor.com & Gaslamp.com, I transformed them into dataframes so that they are easy to manipulate. I decided to use data between May 2020 and the end of July 2020 as "in-Pandemic" data since according to California Association of Realtors, housing sales rebounded starting from May 2020, after which point indicating home-buying decisions due to pandemic. And I used data between May 2019 to end of July 2019 as comparison.

I retrieved data for California, "month\_date\_yyyymm", "postal\_code", "zip\_name" and"hotness\_score", and combined with each zip code’s corresponding latitude & longitude, to be prepared for creating Choropleth maps and the following Foursquare API analysis.

For missing zip code values, I decided to drop them because they only account for about 1% of the data. It would not be worth the efforts searching for all the missing values and their location values.

**2.4 Data Analysis**

**2.4.1 Choropleth Map**

After data cleaning, there were 1,148 records for Pandemic year 2020, and 1,246 records for year 2019. The feature for those records is called Market Hotness Index. This is the score and ranking based on days on market (supply index) and realtor.com views per property (demand index). I decided to create Choropleth maps so that it is easy to visualize the hotness status of different areas just by looking at their colors on the map.

It is easy to easy to estimate that during Pandemic, housing market might be cooler than when the economic was booming. And this was supported by the data. The hot areas(red color) are shrinking.

Let’s zoom in, it is interesting to find that, during the Pandemic period, west LA areas such as Culver City, Bel Air, and Santa Monica, where in the same period of time last year were fairly hot, are losing interest this year. Going upper to the north, it is easy to visualize that the dark red zones near bay area are shrinking. Yet further areas such as Morgan Hill and Clovis are getting more popularities.

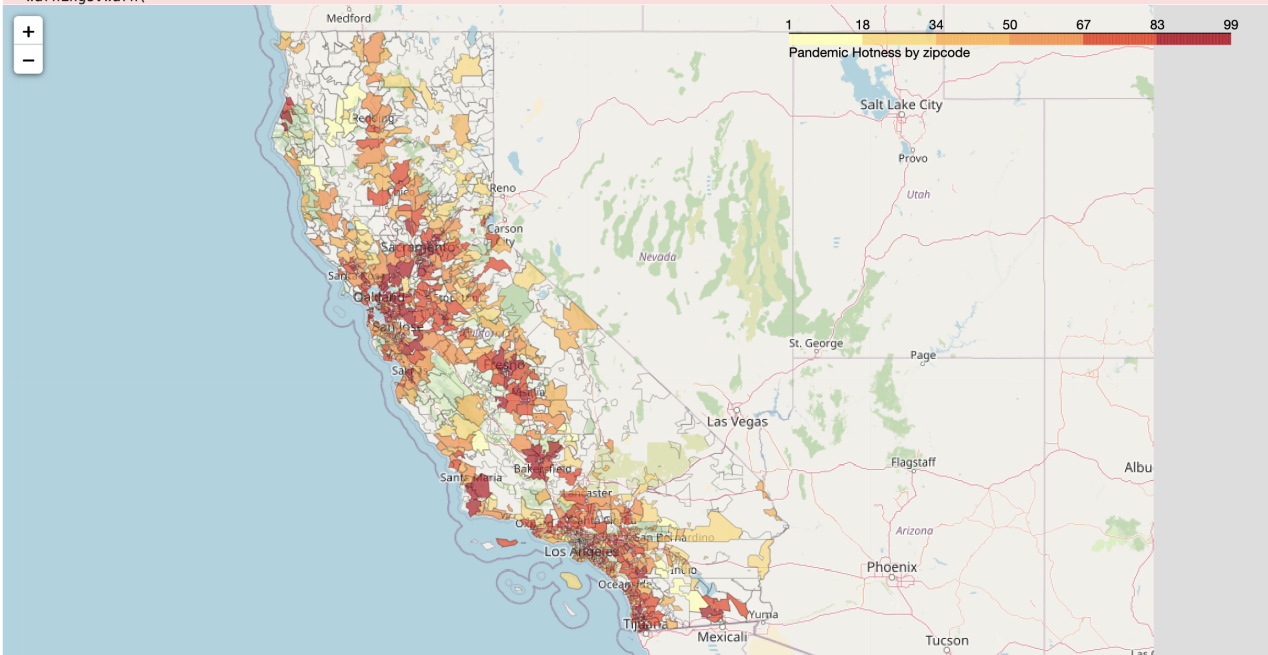


Figure 1. Hotness score by zip code in California between May, 2020 & July, 2020

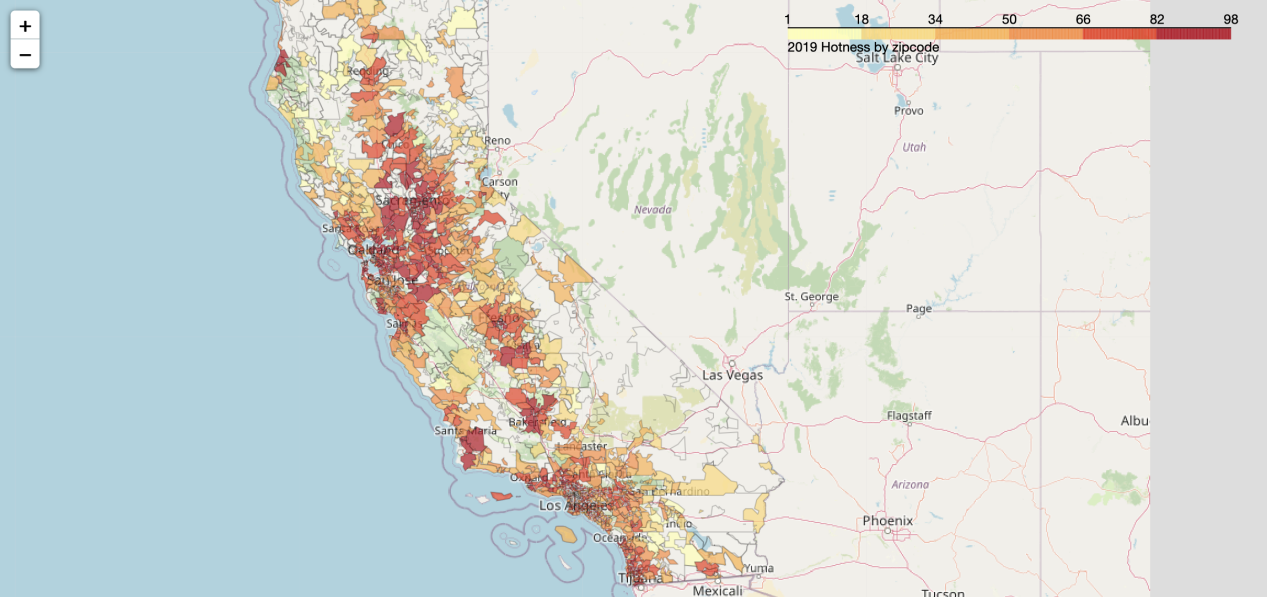


Figure 2. Hotness score by zip code in California between May, 2019 & July, 2019

**2.4.2 Utilizing Foursquare API to explore the neighborhoods**

After getting the 1,148 neighborhood(zip code) records for Pandemic year 2020, I borrowed the function from the Foursquare lab, so that I can get the top 100 venues that are in within a radius of 5000 meters of all these neighborhoods in California.

For each neighborhood, I decided to use one hot encoding in order to the top 5 most common venues for each zip code. Below shows some examples: for instance Burger Joints has a frequency of 0.12, which ranks the 1st among all the venue types in zip code 90001. Same way we are able to know zip code 90001 is an area where Burger Joints, Mexican Restaurants, and Fast Food Restaurants are most common.

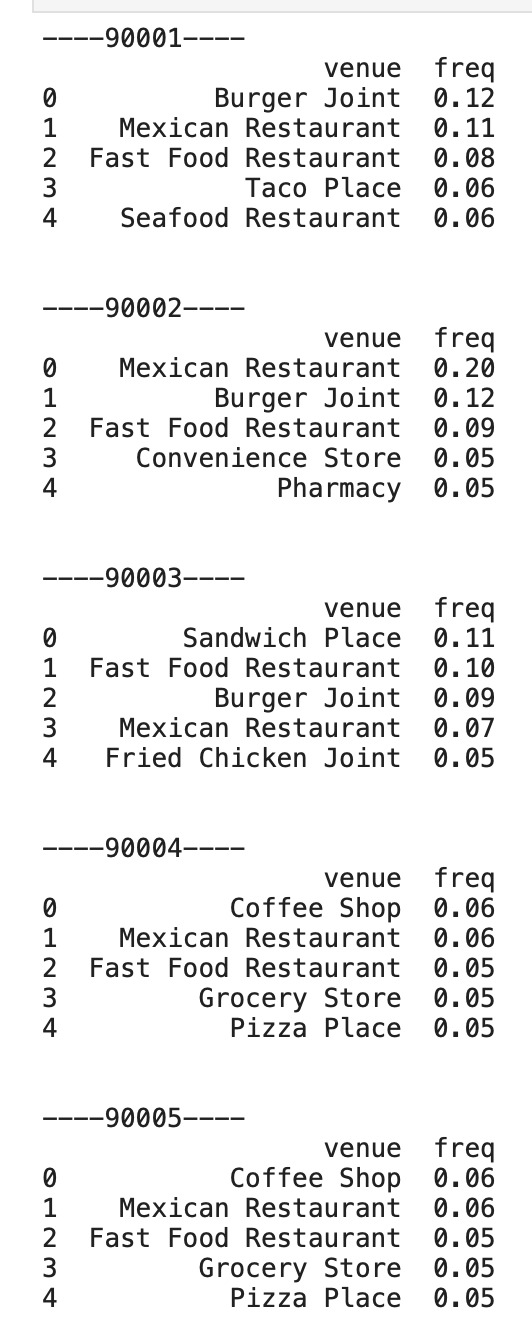


Figure 3. Top 5 most common venues for each zip code

To better visualize the top venues for each zip code, I borrowed the codes from Foursquare lab that create the new dataframe and display the top 10 venues for each neighborhood.



Figure 4. Top 10 venues for each zip code

**2.4.3 Modeling -- K-Means to cluster**

Then I used K-Means to cluster the records and then draw the clusters on the map, so that it is easy to visualize which areas are similar. To determine the optimal k(number of clusters), I used the elbow curve. By reviewing the below graph, we can determine the best k is 7.

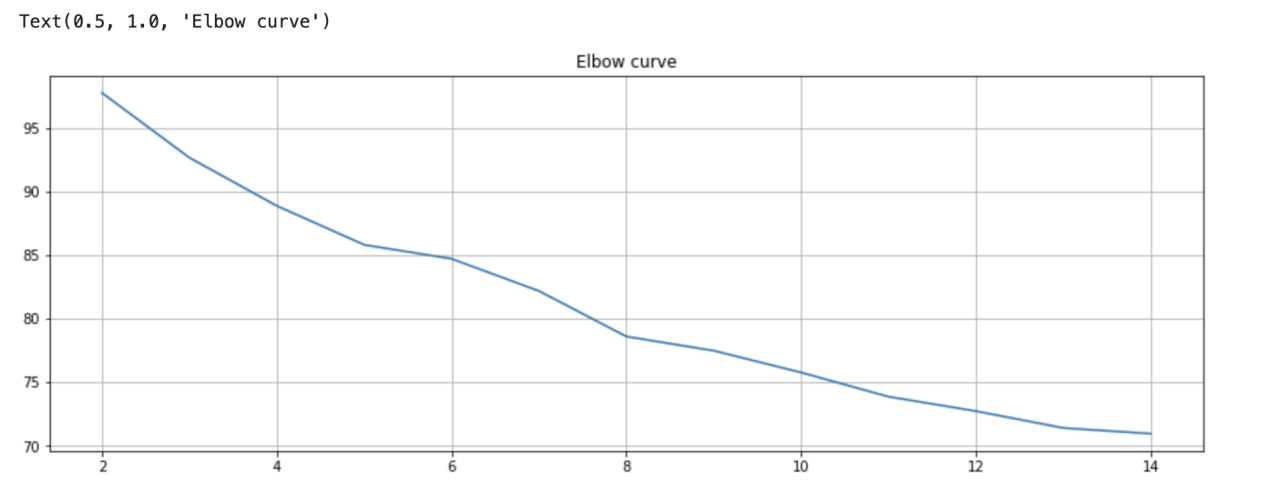


Figure 5. Elbow Curve of different k

Finally I run k-means to segment the neighborhood into 7 clusters. Let's visualize the resulting clusters.

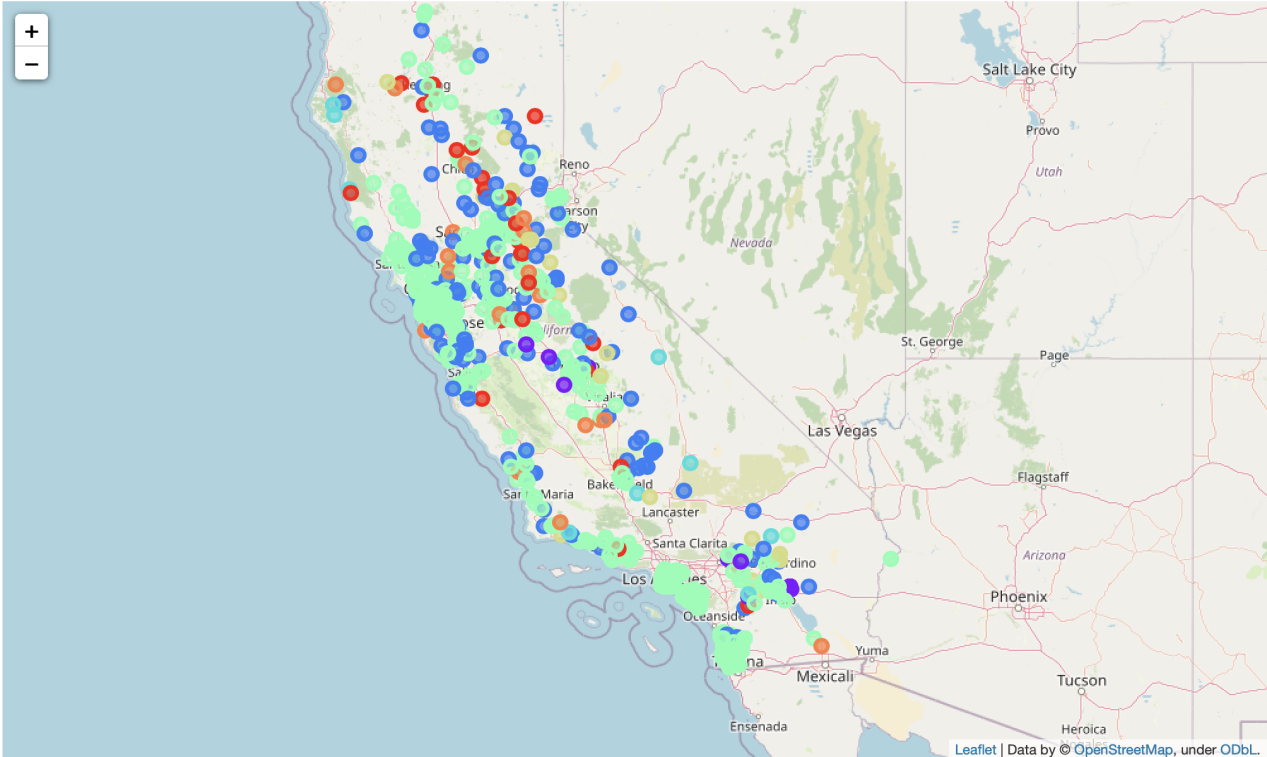


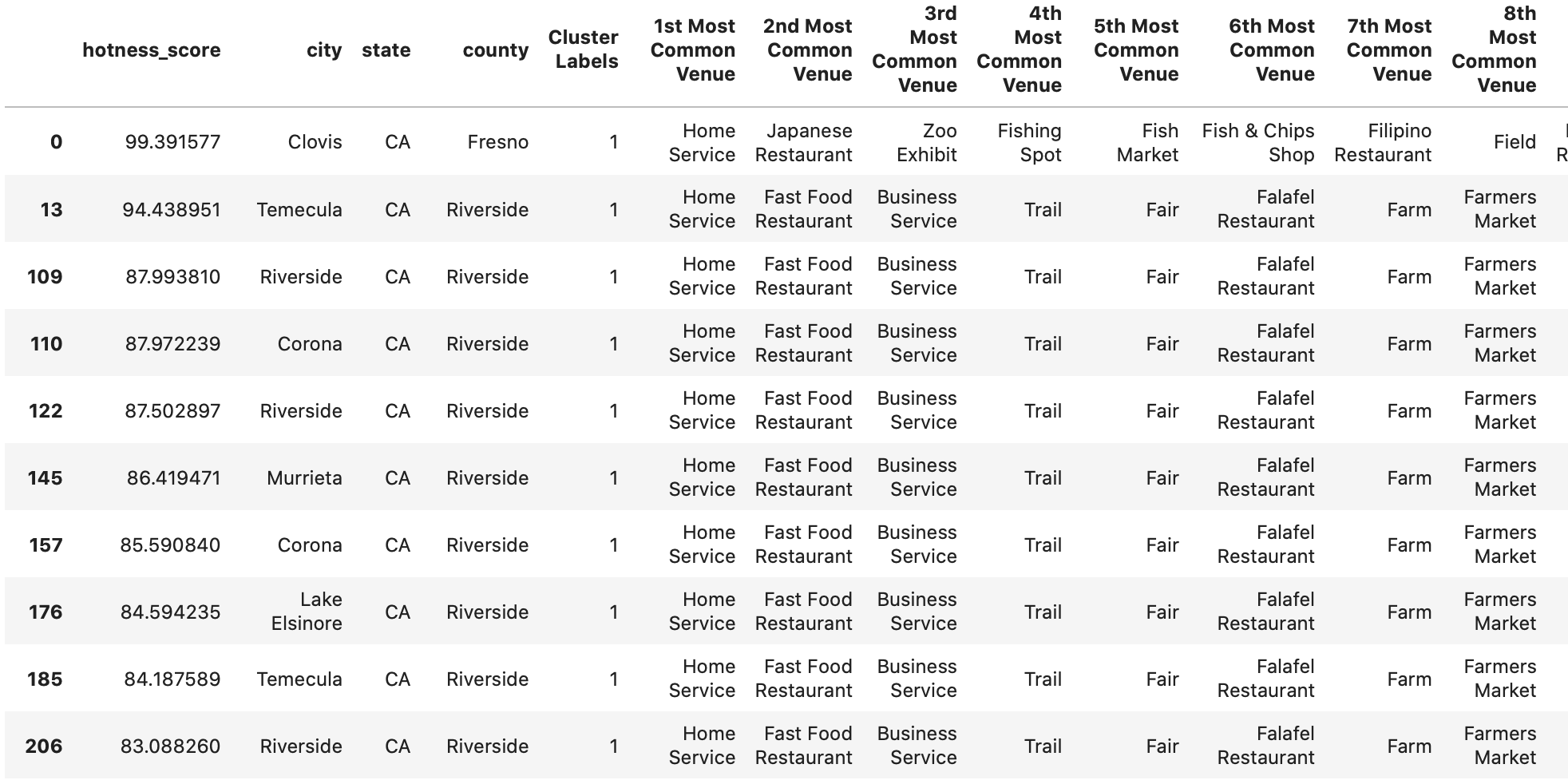
Figure 6. Clusters of neighborhoods

I examined each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a name to each cluster.

**Cluster 0** -- Fish Market/Construction & Landscaping



**Cluster 1** -- Home Service/Japanese Restaurant/Trail

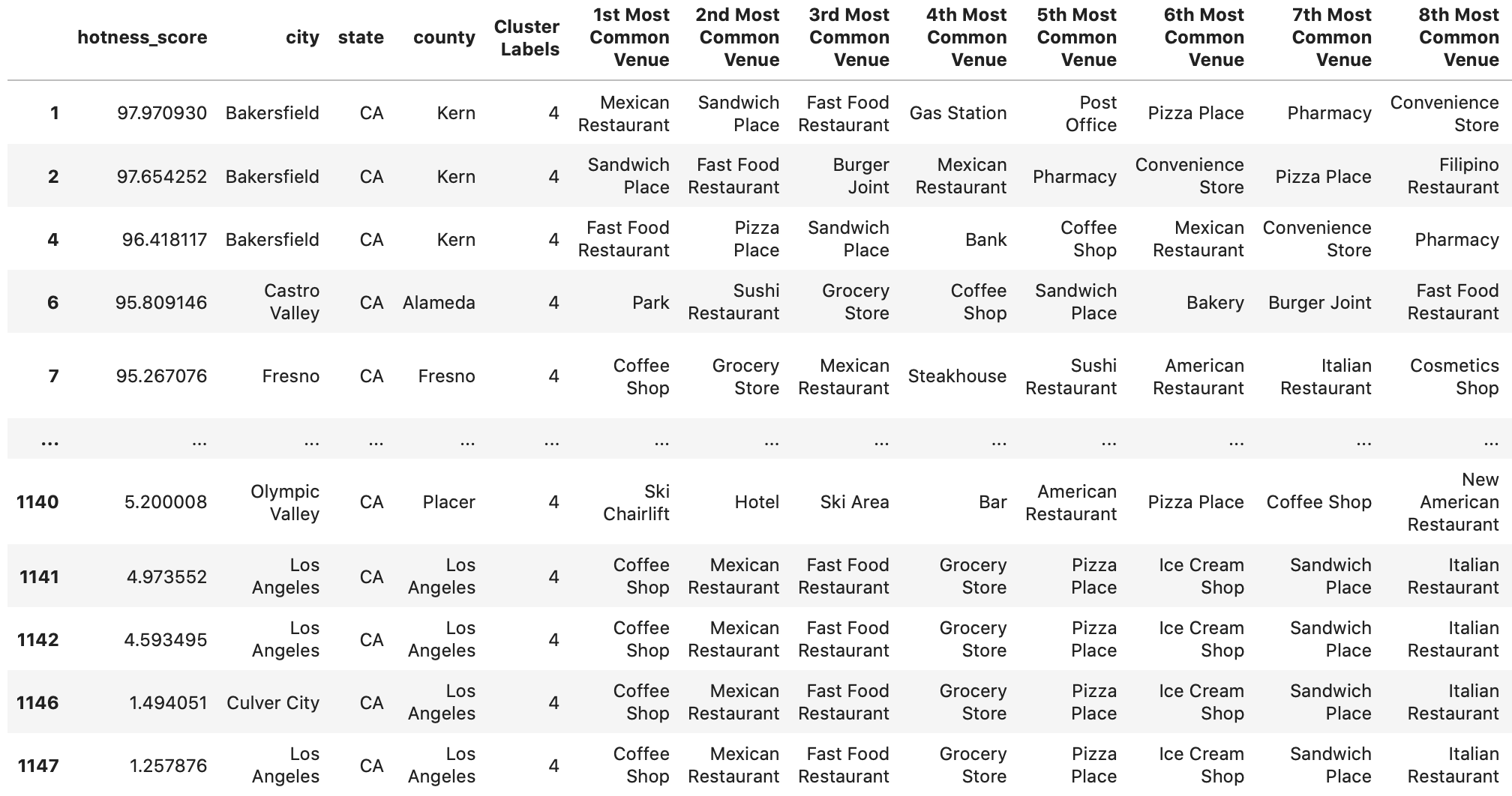


**Cluster 2** -- Winery/Farm/Park/Trail

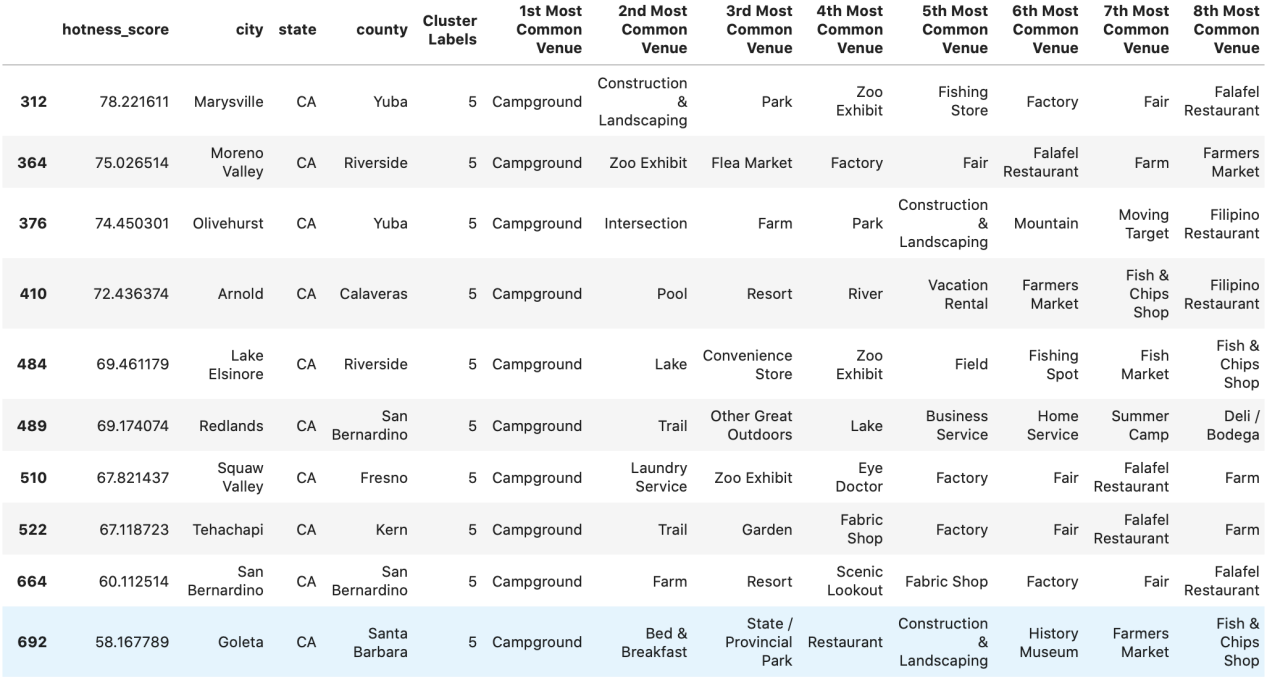


**Cluster 3** -- Trail/Garden/Farm 

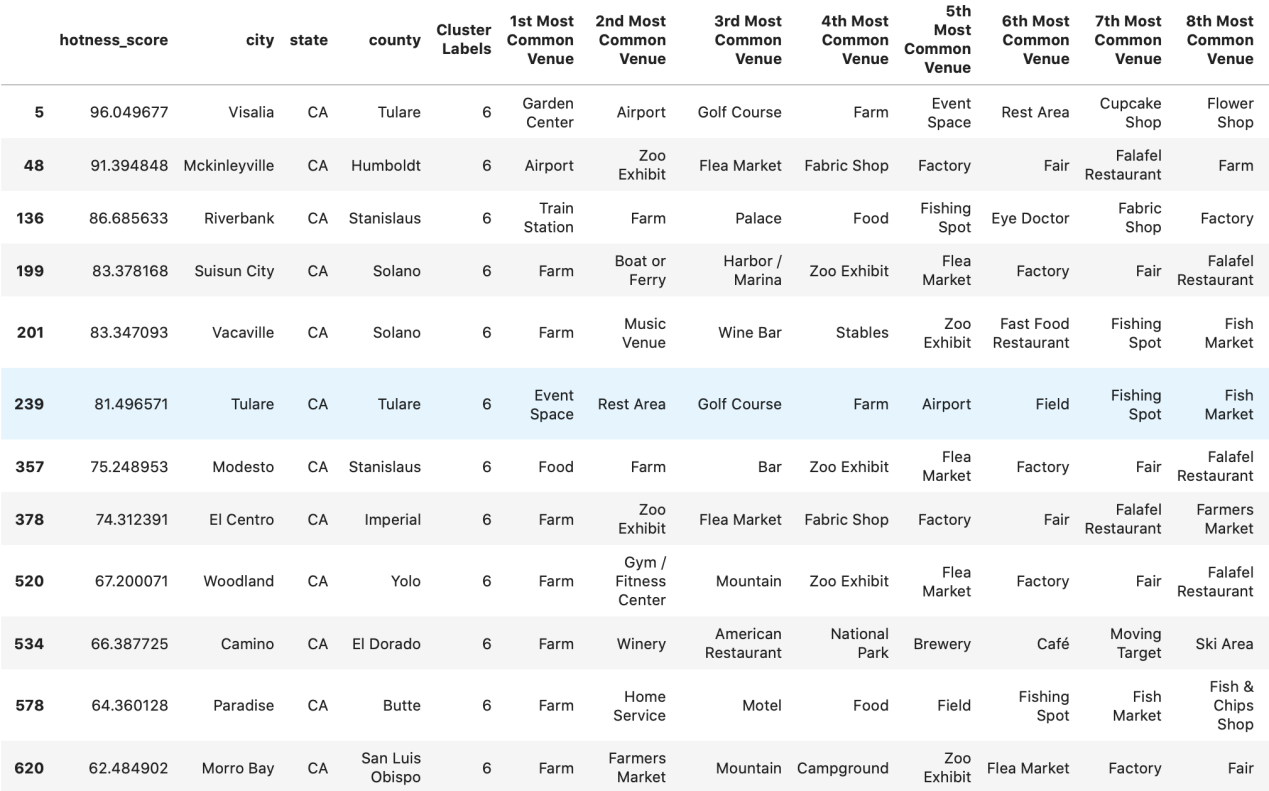
**Cluster 4** -- Mexican Restaurant/Sandwich Place/Fast Food Restaurant



**Cluster 5** --Campground



**Cluster 6** -- Farm



1. **Conclusion**

We discussed our results under two main perceptions:

First, by visualizing Choropleth map, it is interesting to find that, during the Pandemic period, west LA areas such as Culver City, Bel Air, and Santa Monica, where in the same period of time last year were fairly hot, are losing interest this year. And areas such as Burbank, Buena Park, Norwark and La Miranda are picking up. Going upper to the north, it is easy to visual that the dark red zones near bay area are shrinking. Yet further areas such as Morgan Hill and Clovis are getting more popularities.

Second, we can examine our results by the seven clusters we created. On average, Cluster 4 has the highest ranking of hotness scores: it almost covers all of the highest 50 hotness areas(such as Bakersfield and Fresno). And Cluster 2 also includes hot areas such as Lompoc and Elk Grove. We can conclude two main pattern from these clusters. Clusters 4 will attract home buyers who enjoy convenient communities with easy access to fast food restaurants, pharmacy stores, grocery stores and parks. Cluster 2 is for people who love easy access to wineries, parks and trails.

Home buyers can pay more attention to the above areas mentioned that pick up popularity during the pandemic.

1. **Future Directions**

Factors such as prices, the type of the house (ie: how many bedrooms, yard size, single family house or apartment etc) could also contribute to decision-making process during the pandemic. For example, people might prefer single family house than apartment because of the living spaces which is extremely important during the stay-at-home period. I expect to gather more data in order to make more accurate forecast on the house purchasing trend due to pandemic.