## Introduction

An investor wants to invest in top cities of California. However, he is not sure of the kind of business and city in California where he should put his money to get the best returns.

This project will help this investor, identify city and business category, he should be putting his money on**.**

## Objective/Business Problem

Aim of this project is to segment cities in California based on the deposits held in Banks by the residents of the city.

Once we have identified the top 30 branches with maximum deposits, we can gather information about different categories of venues available in these cities by leveraging the data from four square, define clusters of these cities based on business venues, and leverage this clustered data to find the best business and place to put money on.

## Data

Following Data will be used for this analysis:

1. FDIC publishes deposits at branch level at frequent intervals. The latest report available on their website is extracted on Dec 31st 2018. This data is has the branch details and deposits reported for this branch. Sorting this data for the city of California on the basis of deposits will give us the top 30 banks having the most deposits. (FDIC URL: <https://www5.fdic.gov/sod/ShowFileWithStats1.asp?strFileName=ALL_2018.zip>)
2. Clean the data after sorting into a pandas DataFrame for further analysis
3. Data published by FDIC also has the Latitude and longitude of the bank’s branch. This information will be used to get different types of venues available in this vicinity along with the coordinates of the venue
4. Statistical analysis of data obtained from above 2 sources will help us in getting the answers to our problem statment

## Methodology

### Key Assumptions

1. Higher are the deposits in bank branch, higher is the income of the residents living closer to the branch.
2. People with good bank balance are more likely to spend money on things they like the most.
3. This analysis is currently confined to California state only, however same methodology can be applied to any state in united states by changing the state in our filter condition
4. K Means Clustering methodology is used to cluster cities in California
5. Foursquare API is used to get details on all possible venues. As user account with four squarer is required to execute this API

### Methodology - Steps

#### Step 1: Identify top 30 Bank branches in California having most deposits:

#### Download Bank deposits data from FDIC website 🡪 <https://www5.fdic.gov/sod/ShowFileWithStats1.asp?strFileName=ALL_2018.zip>

#### Filter data for “California” state by applying the filter State=”CA”. State information is stored in the column (“STALPBR”)

#### Sort the above data based on the deposits value. This value is stored in the column “DEPSUMBR” of the data extract.

#### Remove redundant columns and create a dataframe with column information that we will require for our analysis

#### Plot these branches on the map of California. The output will be like this:

#### Call foursquare API with appropriate credentials to get top 100 venues around these branches

#### Step 2: Explore neighborhoods in the cities with most deposits:

#### Get venues within top 100 venues within a radius of ½ a mile from these top 30 cities.

#### Get the count of these top venues for each of the cities

#### Based on the above information San Francisco, Los Angeles, San Diego, Beverly Hills and Walnut Creek are most likely the areas where we should be investing in. Let’s do further analysis to strengthen our hypothesis.

#### Step 3: Analyze each city to get more information points

#### Apply One hot encoding to the venue data to get statistical information in numeric format for further sorting and analysis. This will store the top venues as columns and corresponding analytical values as numbers

#### Group rows based on the city and use mean values to identify the frequency of occurrence of each of these venues. Check how many venues and unique categories were returned for these cities.

#### Identify top 10 venues for each of the shortlisted cities

#### Step 4: Create clusters of these venues by performing k-means analysis

#### Merge the sorted venue data with the original dataset created in Step1. Plot the cluster distribution of venues on the map of California.

#### 

#### Step 5: Examine these clusters

#### With an assumption of 5 clusters, use K-Cluster algorithm to come up with 5 different clusters in California with similar set of Venues.

#### Map all the defined clusters in the City of California.

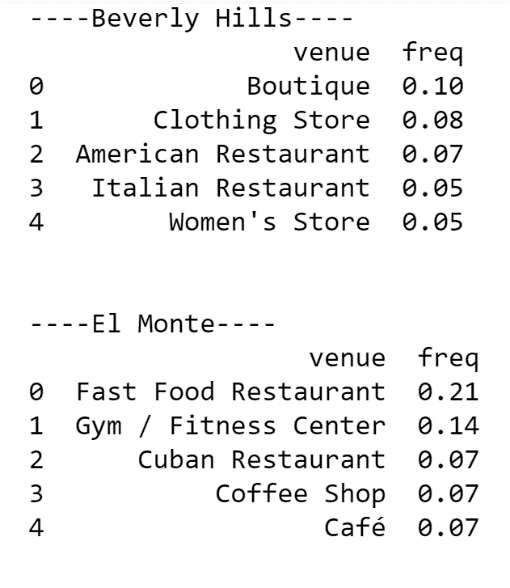
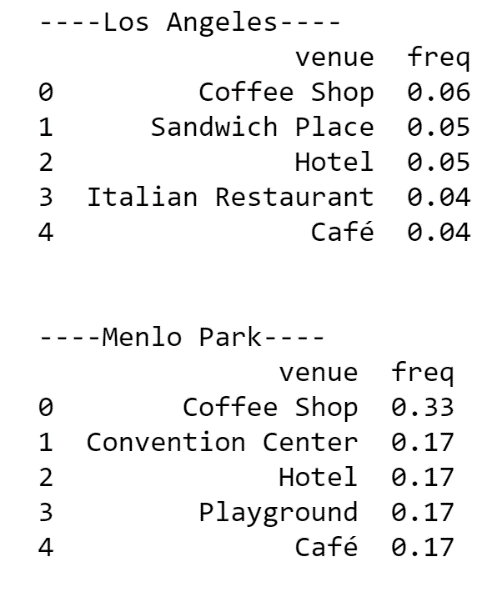
#### Examine clusters and predict:

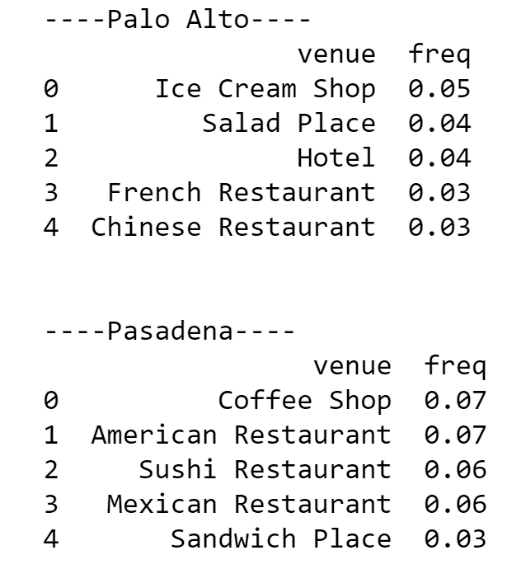
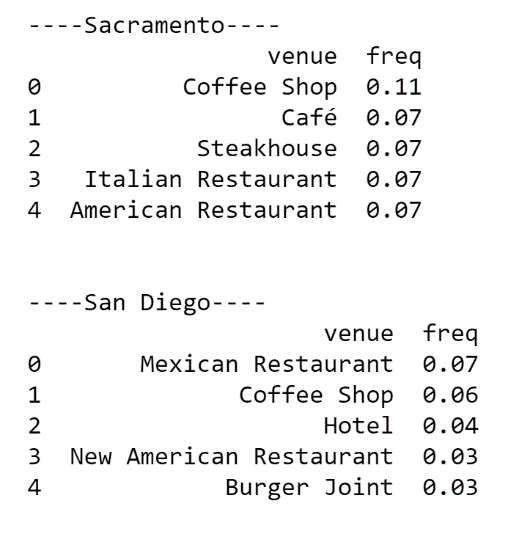
#### Examine each cluster and determine the discriminating venue categories that distinguish each cluster.

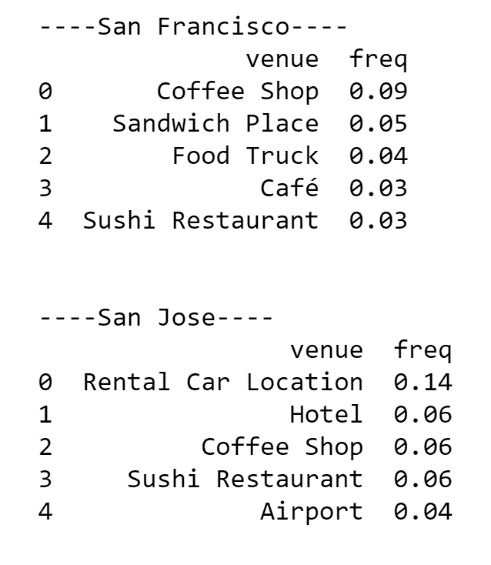
#### Identify the clusters & Cities and identify business entities having maximum frequency

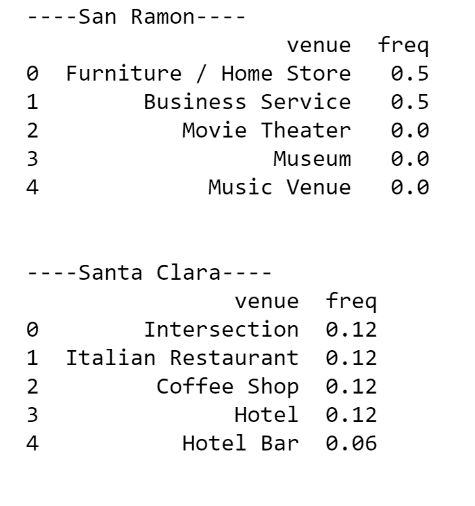
## Results

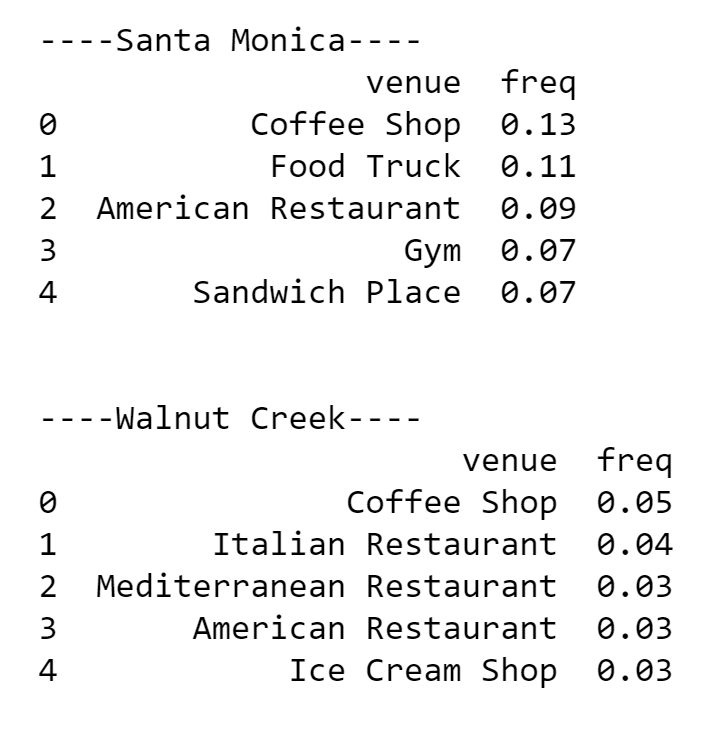
1. There are 2123 locations among cities with maximum deposits and there are 228 categories of business opportunities in these locations
2. These businesses are spread across 14 cities in the state
3. Top 5 venues of each of these cities are:









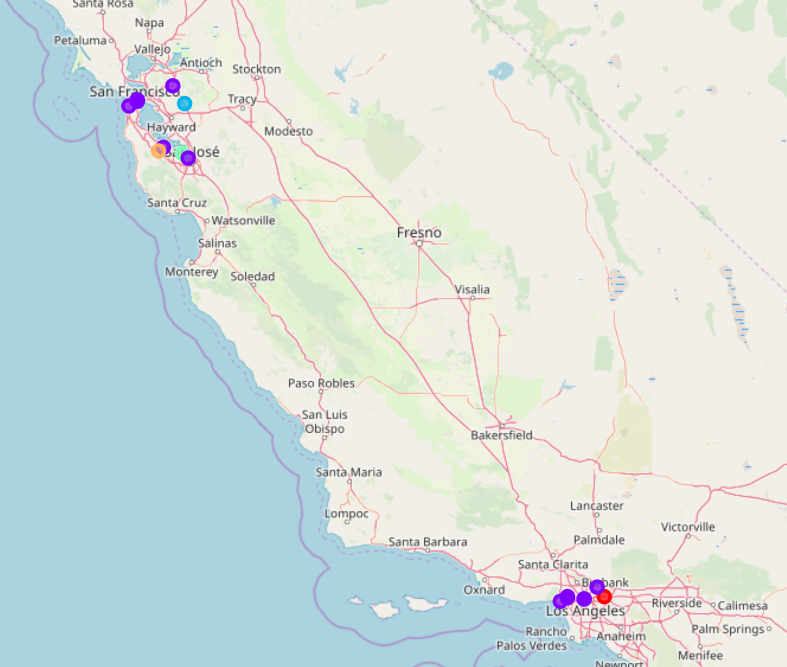


**Cluster Cities**

1. The clustering algorithm resulted the following 5 clusters using K-Cluster algorithm

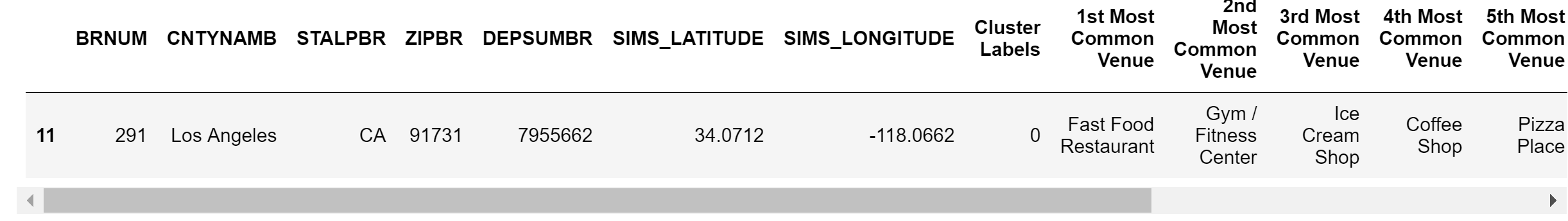


1. And Once all the defined clusters are mapped on the City of California it looks as follows:

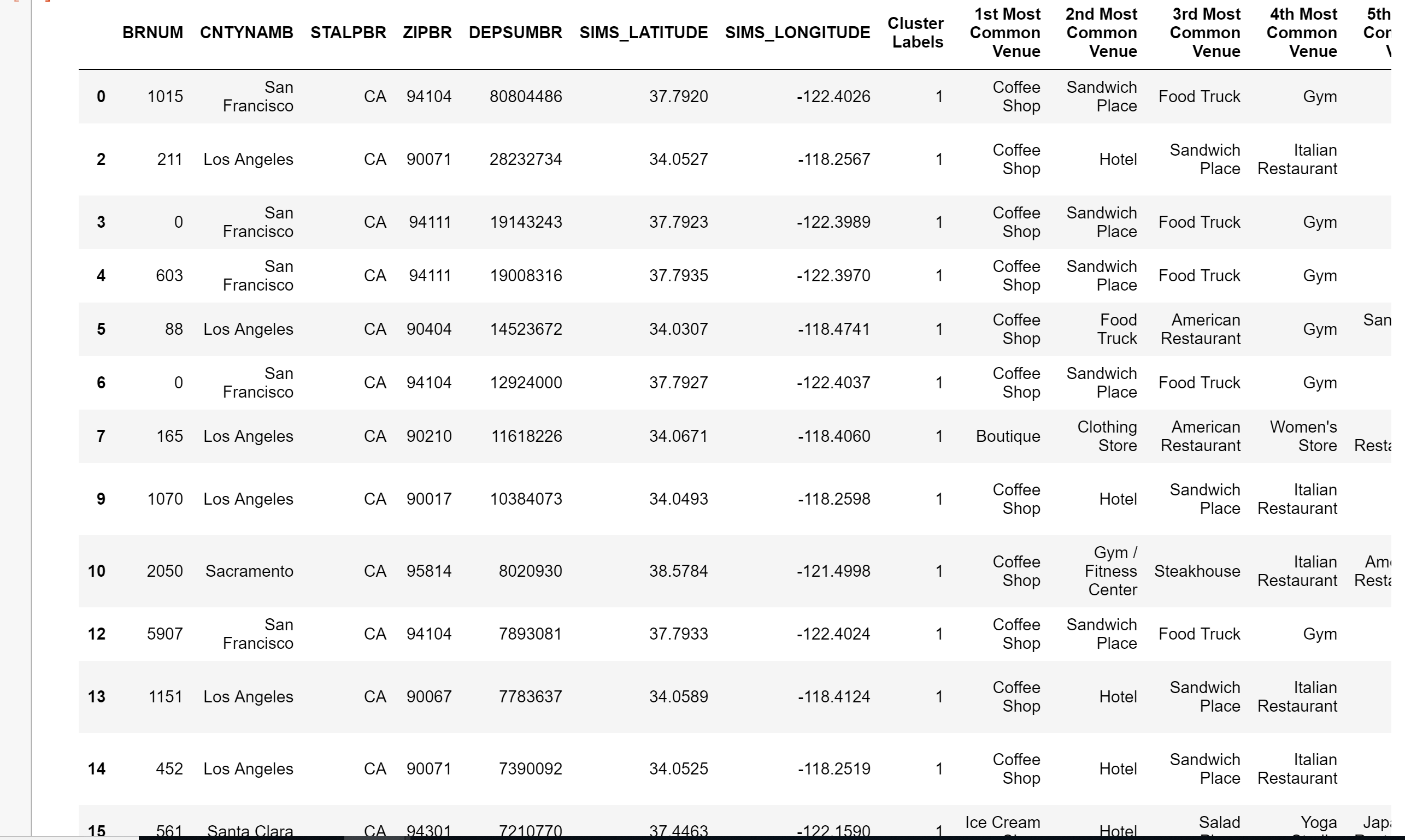


#### Examine clusters and predict

**a) The clusters defined looks as follows:**

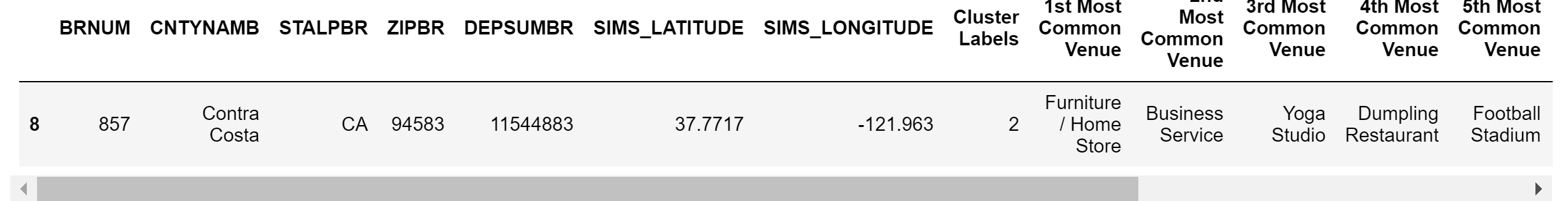
**Cluster-1**

**Cluster-2**

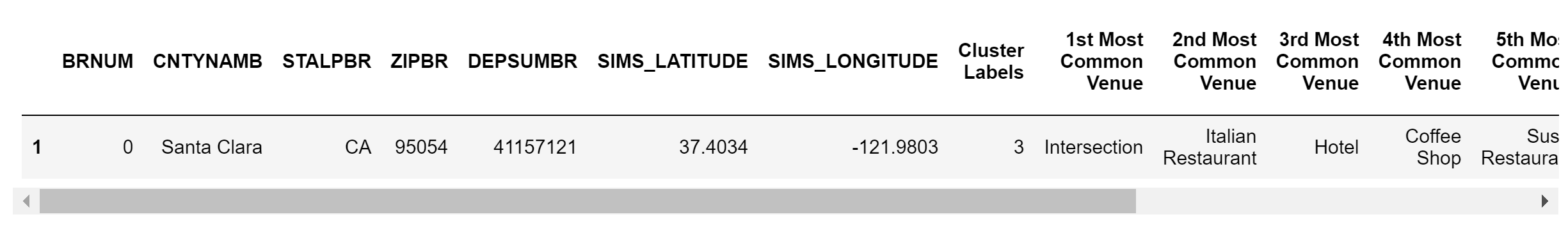




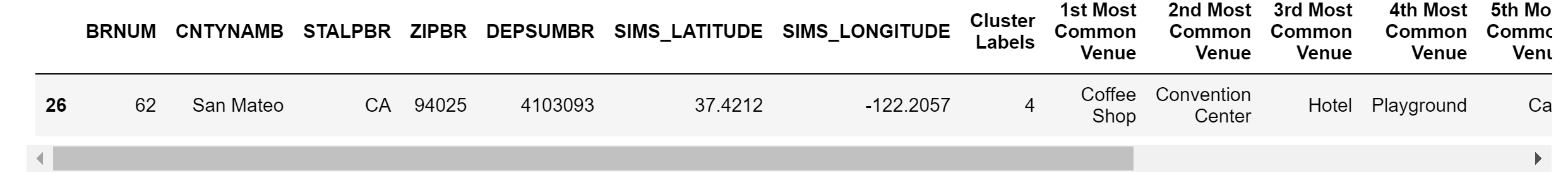
**Cluster-3**



**Cluster-4**



**Cluster-5**



**b) Based on the cluster data investing in a coffee shop in any of the locations listed in cluster 2 will be profitable. Second best investment opportunity could be a sandwich shop and the third one is a hotel**

## Discussion

**Based on our initial assumption of the cluster with most common business venue would be a good investment area. On this basis Cluster 2 and a coffee shop in this area would be the right business opportunity. Top 5 areas for this are the ones having zip codes:**

1. 94104
2. 90071
3. 94111
4. 90404
5. 90210

**Conclusion**

**Clustering as an algorithm gave a good split among the clusters and was able to identify uniqueness well. With the current data used, this looks like one of the best possible solution for the customer to identify location and business category is in Cluster 2.**

**If he needs to choose an area he can safely bet on area having zip code 94104 which is SanFrancisco California, and possibly one of the top five group of neighborhoods as explained in the Discussion section.**

**Note: It is quite possible to have a different set of assumptions or calculation logic used to identify the right City and this might significantly change the result of this model.**