Analysis of the Neighborhoods in San Jose, California for Starting a new Restaurant

Capstone Project - The Battle of Neighborhoods:

1. Introduction

San Jose, California, located in the center of Silicon Valley, is the largest city in Northern California by both population and area. With an estimated 2019 population of 1,021,795, it is the third-most populous city in California (after Los Angeles and San Diego) and the tenth-most populous in the United States. Located in the center of the Santa Clara Valley, on the southern shore of San Francisco Bay, San Jose covers an area of 179.97 square miles.

The large concentration of high-technology engineering, computer and microprocessor companies around San Jose has led the area to be known as Silicon Valley. And this also led the San Jose has a large variety of immigrations and different kind of restaurant across all the city. The diversity of the population of the city has also brought in a vast diversity in food. To start a new restaurant, it is very import to do a careful analysis of the neighborhoods in San Jose and find the best area, which is the main topic in this project and also the business problem we will solve.

2. Business Problem

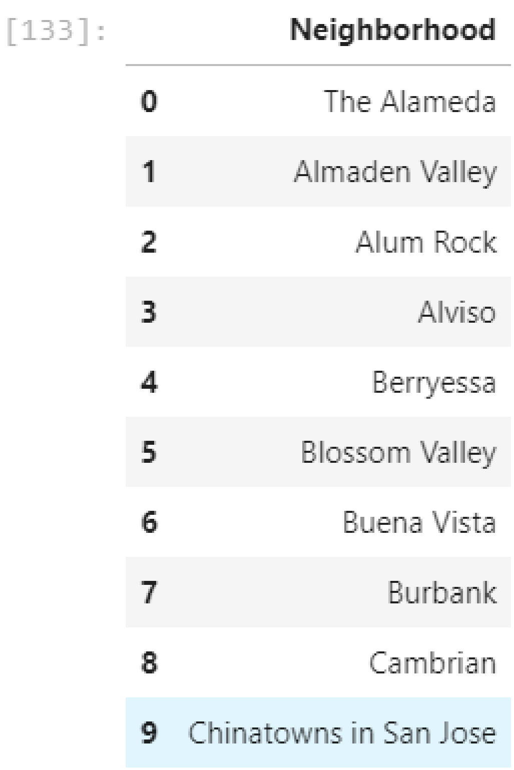
Our client is an investor and entrepreneur who is plan to start a new restaurant in San Jose. He approached us to study the neighborhoods in San Jose and suggest a location of area which would be in best interest of the business. Our goal is to extract and analysis the data of all the neighborhoods of San Jose, using machine learning techniques and provide a suggestion of locate to start a new restaurant.

3. Data

The following data is required for this project. We will divide it in to below sections.

3.1 Neighborhood Data

The neighborhood data includes all the neighborhood name. This information can be easily find in the wikipedia page.

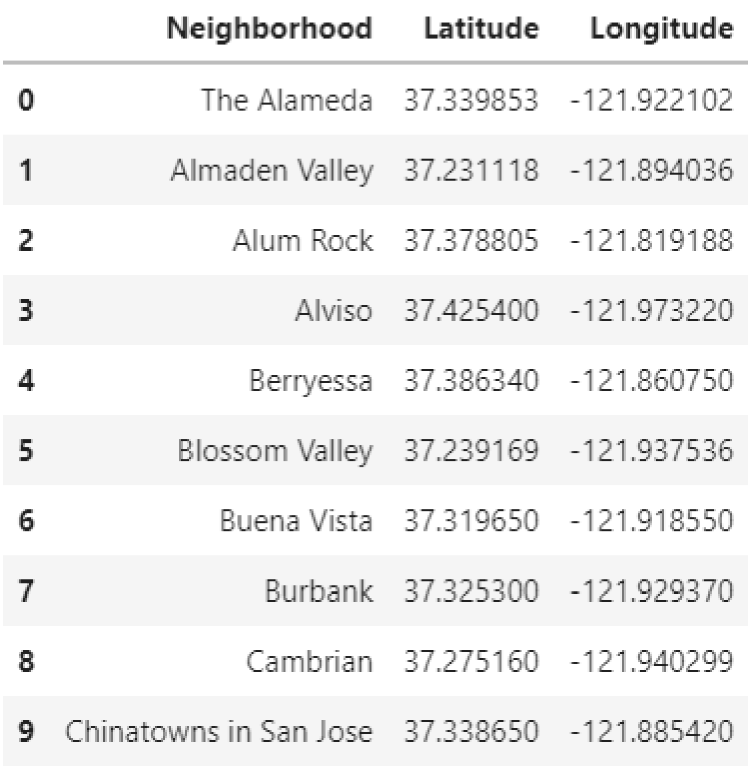


[133]:

	Neighborhood
0	The Alameda
1	Almaden Valley
2	Alum Rock
3	Alviso
4	Berryessa
5	Blossom Valley
6	Buena Vista
7	Burbank
8	Cambrian
9	Chinatowns in San Jose

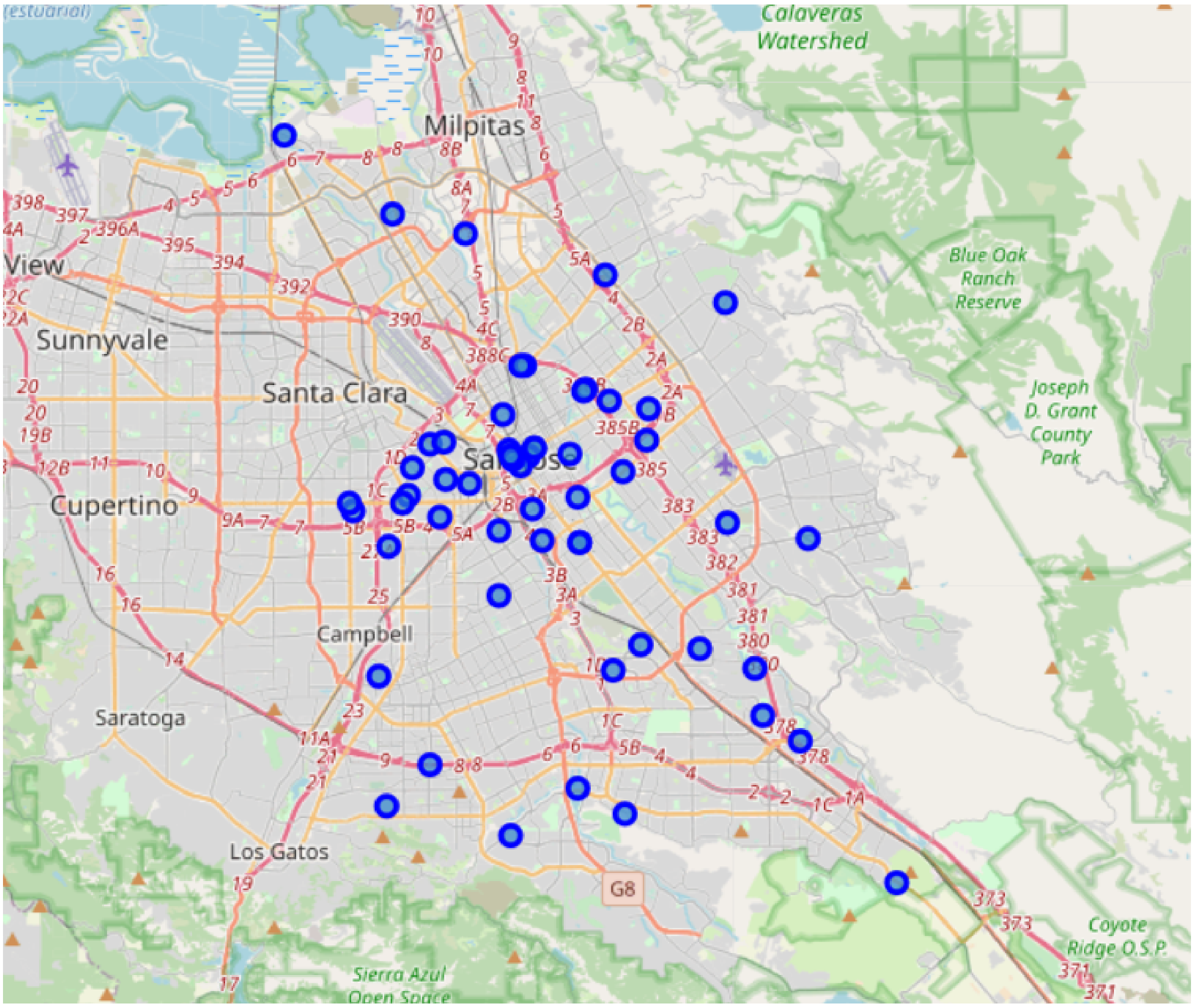
3.2 Geographical Coordinates

We will use the Geopy library in python to get all the geographical information. The geographical coordinates is very important for map plotting during the project. Here is the example data after pulling out all the geographical coordinates.



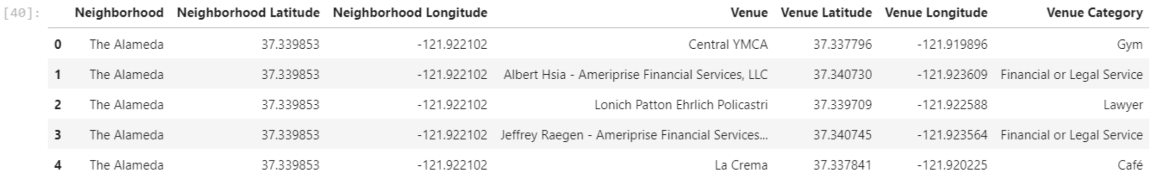
	Neighborhood	Latitude	Longitude
0	The Alameda	37.339853	-121.922102
1	Almaden Valley	37.231118	-121.894036
2	Alum Rock	37.378805	-121.819188
3	Alviso	37.425400	-121.973220
4	Berryessa	37.386340	-121.860750
5	Blossom Valley	37.239169	-121.937536
6	Buena Vista	37.319650	-121.918550
7	Burbank	37.325300	-121.929370
8	Cambrian	37.275160	-121.940299
9	Chinatowns in San Jose	37.338650	-121.885420

We can also plot the neighborhood on the map:



3.3 Venue Data

The venue data will be pulled from the FourSquare API. The main query type we will use is explore. And the result information will be used for data clustering. Here is an example of the Venue data for one neighborhood.



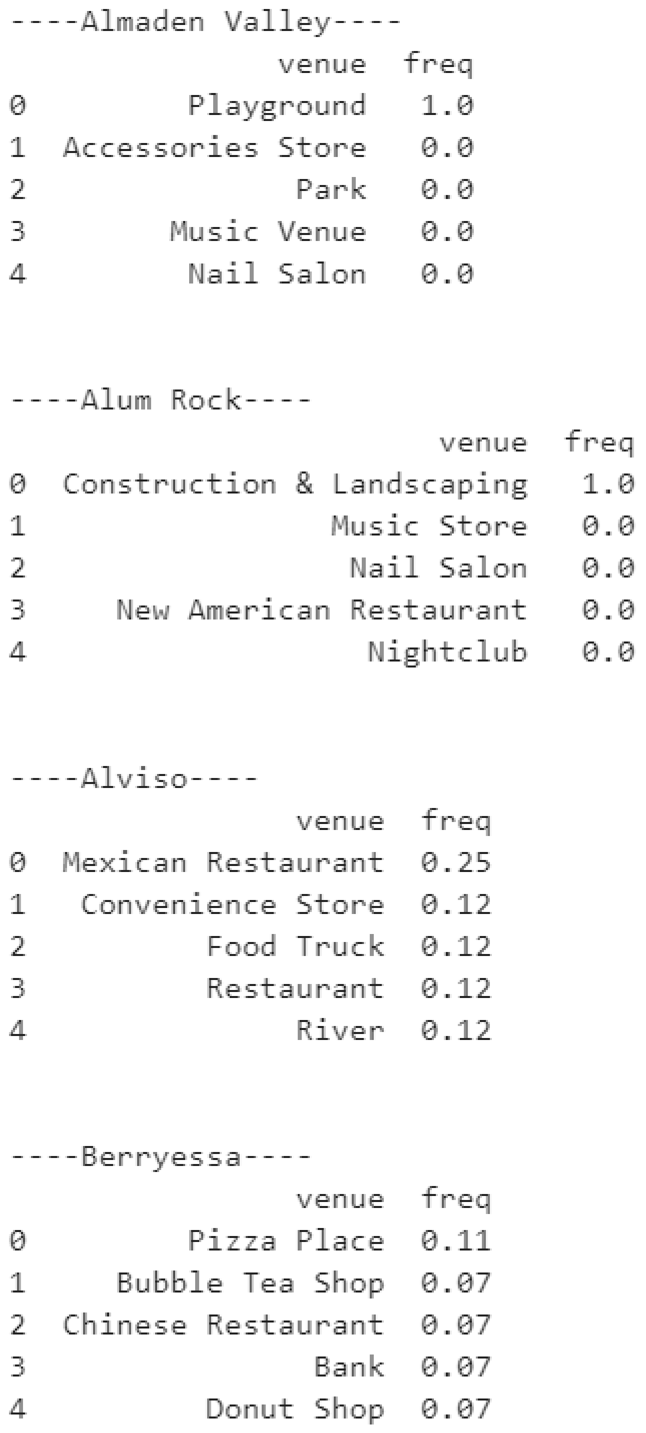
[40]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	The Alameda	37.339853	-121.922102	Central YMCA	37.337796	-121.919896	Gym
1	The Alameda	37.339853	-121.922102	Albert Hsia - Ameriprise Financial Services, LLC	37.340730	-121.923609	Financial or Legal Service
2	The Alameda	37.339853	-121.922102	Lonich Patton Ehrlich Policastri	37.339709	-121.922588	Lawyer
3	The Alameda	37.339853	-121.922102	Jeffrey Raegen - Ameriprise Financial Services...	37.340745	-121.923564	Financial or Legal Service
4	The Alameda	37.339853	-121.922102	La Crema	37.337841	-121.920225	Café

4. Methodology

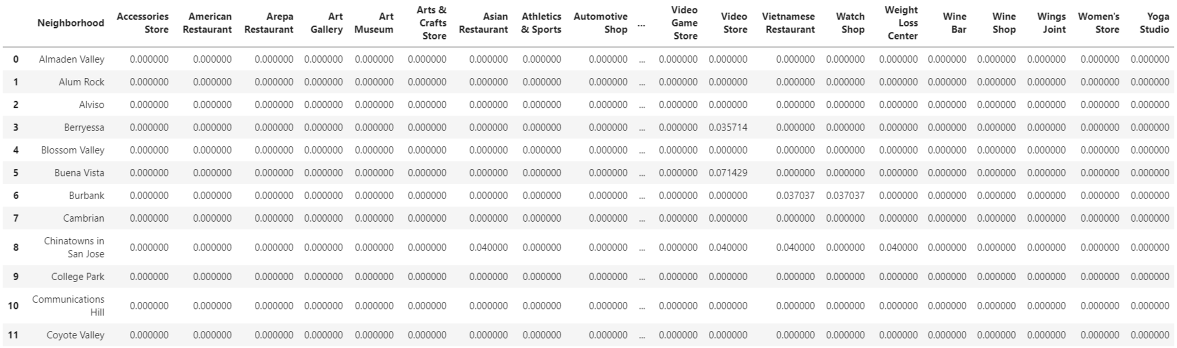
4.1 Feature Extraction

With the help our Foursquare API, we can extract the feature the feature of each neighborhood. First, we create a function and looped it to all the neighborhood to get the explore result. The Foursquare API will feedback us a list of interesting venues.



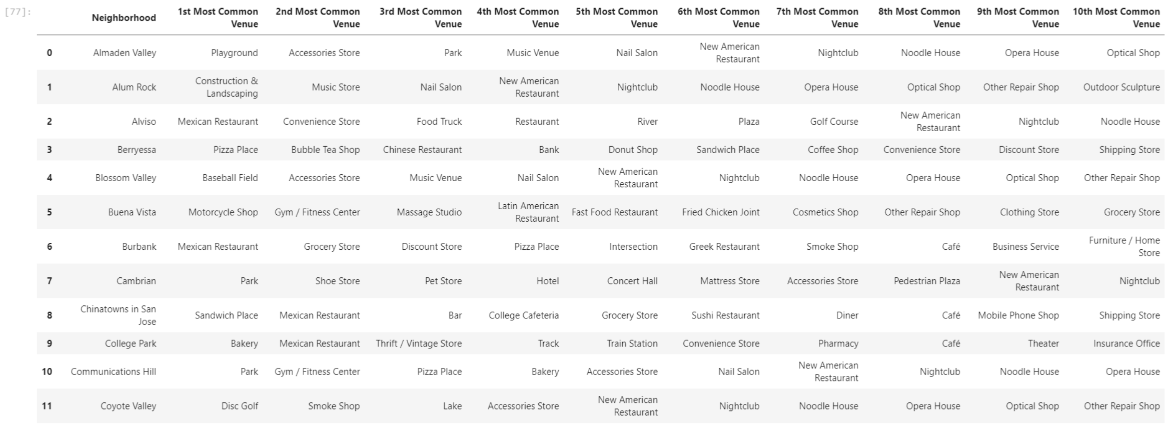
```
----Almaden Valley----
               venue  freq
0         Playground   1.0
1  Accessories Store   0.0
2               Park   0.0
3        Music Venue   0.0
4         Nail Salon   0.0


----Alum Rock----
                        venue  freq
0  Construction & Landscaping   1.0
1                 Music Store   0.0
2                  Nail Salon   0.0
3     New American Restaurant   0.0
4                   Nightclub   0.0


----Alviso----
                venue  freq
0  Mexican Restaurant  0.25
1   Convenience Store  0.12
2          Food Truck  0.12
3          Restaurant  0.12
4               River  0.12


----Berryessa----
                venue  freq
0         Pizza Place  0.11
1      Bubble Tea Shop  0.07
2  Chinese Restaurant  0.07
3                Bank  0.07
4          Donut Shop  0.07
```

Second, we can group the venues by neighborhood. This can be done by using one-hot encoding. That is, if a venue belongs to a neighborhood, we will assign the frequency to the corresponding row and venue's column. We do this for all the neighborhood and we can get a matrix like below:



	Neighborhood	Accessories Store	American Restaurant	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Automotive Shop	...	Video Game Store	Video Store	Vietnamese Restaurant	Watch Shop	Weight Loss Center	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	Almaden Valley	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Alum Rock	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Alviso	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Berryessa	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.035714	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Blossom Valley	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	Buena Vista	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.071429	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	Burbank	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.037037	0.037037	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	Cambrian	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	Chinatowns in San Jose	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	0.000000	...	0.000000	0.040000	0.040000	0.000000	0.040000	0.000000	0.000000	0.000000	0.000000	0.000000
9	College Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
10	Communications Hill	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
11	Coyote Valley	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Third, with the data above, we can create another data frame that will represent the most common venue in each neighborhood. The matrix is like below:



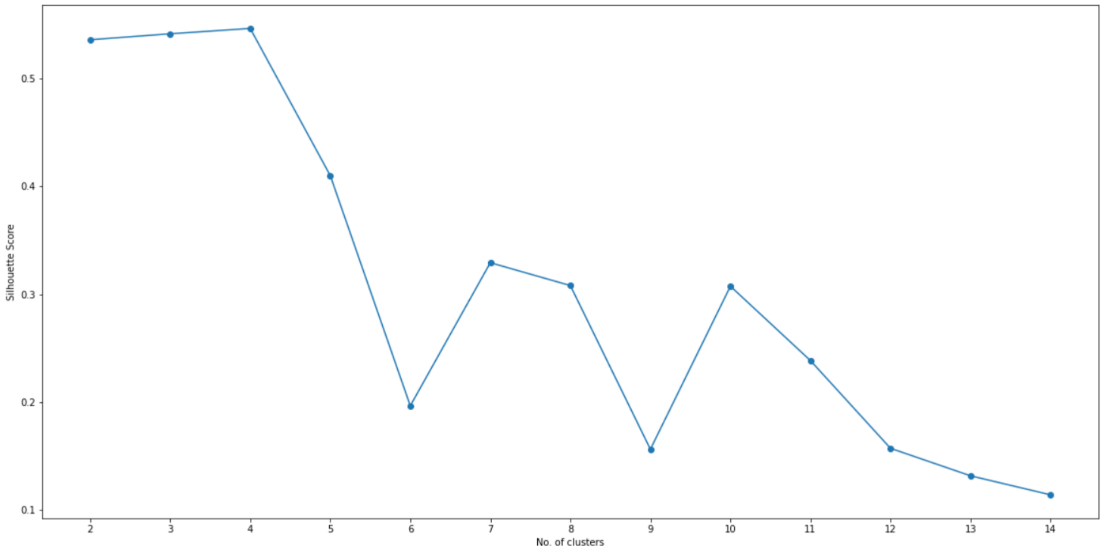
[77]:

	Neighborhood	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Almaden Valley	Playground	Accessories Store	Park	Music Venue	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop
1	Alum Rock	Construction & Landscaping	Music Store	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop	Outdoor Sculpture
2	Alviso	Mexican Restaurant	Convenience Store	Food Truck	Restaurant	River	Plaza	Golf Course	New American Restaurant	Nightclub	Noodle House
3	Berryessa	Pizza Place	Bubble Tea Shop	Chinese Restaurant	Bank	Donut Shop	Sandwich Place	Coffee Shop	Convenience Store	Discount Store	Shipping Store
4	Blossom Valley	Baseball Field	Accessories Store	Music Venue	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop
5	Buena Vista	Motorcycle Shop	Gym / Fitness Center	Massage Studio	Latin American Restaurant	Fast Food Restaurant	Fried Chicken Joint	Cosmetics Shop	Other Repair Shop	Clothing Store	Grocery Store
6	Burbank	Mexican Restaurant	Grocery Store	Discount Store	Pizza Place	Intersection	Greek Restaurant	Smoke Shop	Café	Business Service	Furniture / Home Store
7	Cambrian	Park	Shoe Store	Pet Store	Hotel	Concert Hall	Mattress Store	Accessories Store	Pedestrian Plaza	New American Restaurant	Nightclub
8	Chinatowns in San Jose	Sandwich Place	Mexican Restaurant	Bar	College Cafeteria	Grocery Store	Sushi Restaurant	Diner	Café	Mobile Phone Shop	Shipping Store
9	College Park	Bakery	Mexican Restaurant	Thrift / Vintage Store	Track	Train Station	Convenience Store	Pharmacy	Café	Theater	Insurance Office
10	Communications Hill	Park	Gym / Fitness Center	Pizza Place	Bakery	Accessories Store	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House
11	Coyote Valley	Disc Golf	Smoke Shop	Lake	Accessories Store	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop

When we reach here, we have complete all the data preparation for the data clustering.

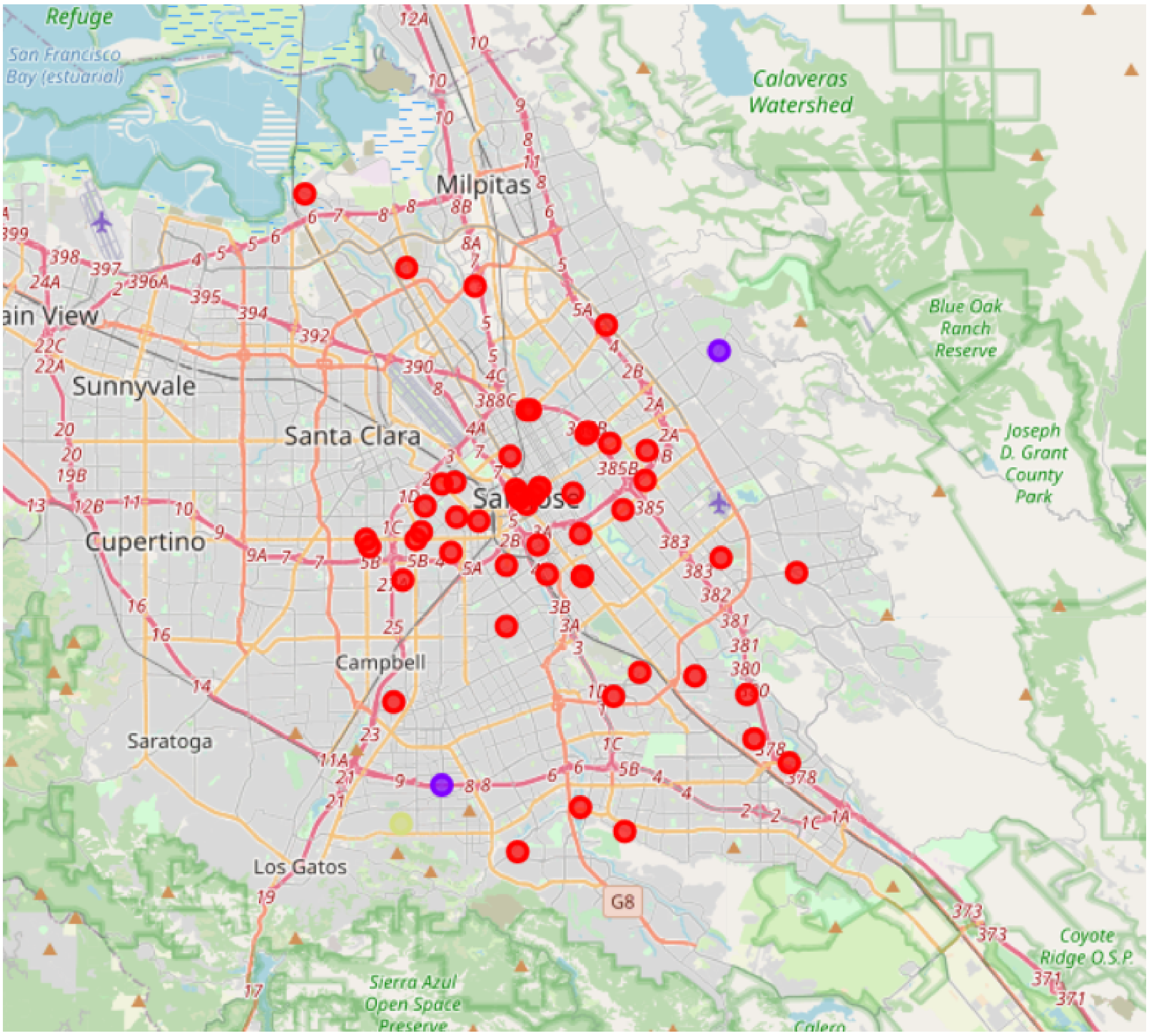
4.2 Unsupervised Machine Learning

We will use K-means to do the unsupervised machine learning to cluster the neighborhood to different group, which have some special features. For the number K, we will use the K-mean model to loop through different k and get the number of k vs Silhouette score plot. Then we can decide which K to use with the highest Silhouette score. The result is in below:

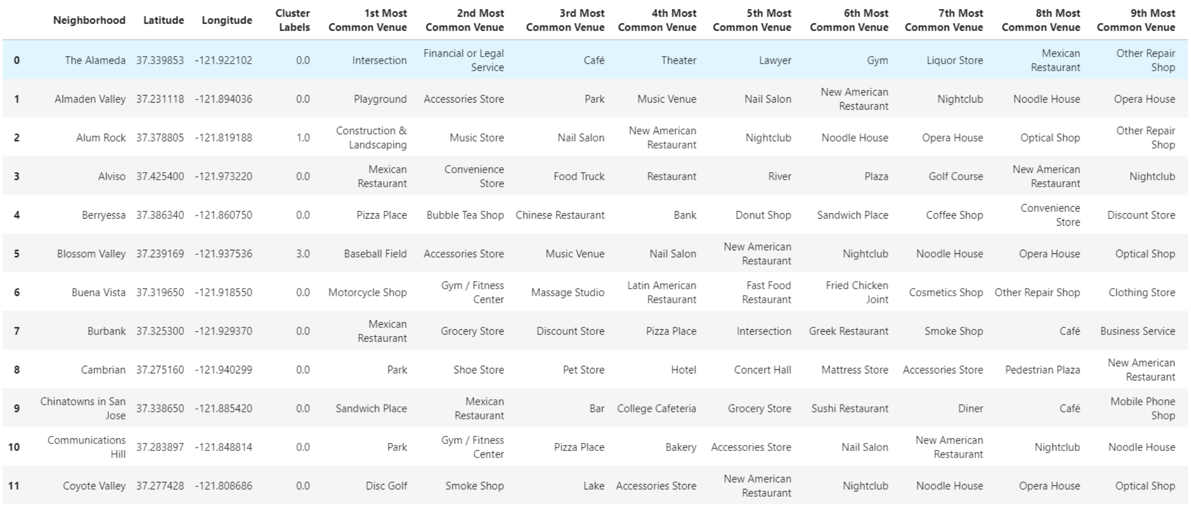


5. Results

The clustering model clusters the neighborhoods in San Jose and also provides a label. We can merge the label result into the most common venue matrix in chapter 4.1. Then we can see the features of each cluster. The below is the map plot of the clustering

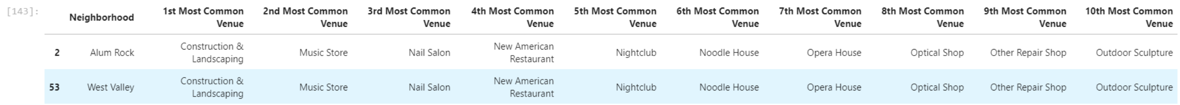


And for each cluster, the first cluster:



	Neighborhood	Latitude	Longitude	Cluster Labels	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	8th Most Common Venue	9th Most Common Venue
0	The Alameda	37.339853	-121.922102	0.0	Intersection	Financial or Legal Service	Café	Theater	Lawyer	Gym	Liquor Store	Mexican Restaurant	Other Repair Shop
1	Almaden Valley	37.231118	-121.894036	0.0	Playground	Accessories Store	Park	Music Venue	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House
2	Alum Rock	37.378805	-121.819188	1.0	Construction & Landscaping	Music Store	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop
3	Alviso	37.425400	-121.973220	0.0	Mexican Restaurant	Convenience Store	Food Truck	Restaurant	River	Plaza	Golf Course	New American Restaurant	Nightclub
4	Berryessa	37.386340	-121.860750	0.0	Pizza Place	Bubble Tea Shop	Chinese Restaurant	Bank	Donut Shop	Sandwich Place	Coffee Shop	Convenience Store	Discount Store
5	Blossom Valley	37.239169	-121.937536	3.0	Baseball Field	Accessories Store	Music Venue	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop
6	Buena Vista	37.319650	-121.918550	0.0	Motorcycle Shop	Gym / Fitness Center	Massage Studio	Latin American Restaurant	Fast Food Restaurant	Fried Chicken Joint	Cosmetics Shop	Other Repair Shop	Clothing Store
7	Burbank	37.325300	-121.929370	0.0	Mexican Restaurant	Grocery Store	Discount Store	Pizza Place	Intersection	Greek Restaurant	Smoke Shop	Café	Business Service
8	Cambrian	37.275160	-121.940299	0.0	Park	Shoe Store	Pet Store	Hotel	Concert Hall	Mattress Store	Accessories Store	Pedestrian Plaza	New American Restaurant
9	Chinatowns in San Jose	37.338650	-121.885420	0.0	Sandwich Place	Mexican Restaurant	Bar	College Cafeteria	Grocery Store	Sushi Restaurant	Diner	Café	Mobile Phone Shop
10	Communications Hill	37.283897	-121.848814	0.0	Park	Gym / Fitness Center	Pizza Place	Bakery	Accessories Store	Nail Salon	New American Restaurant	Nightclub	Noodle House
11	Coyote Valley	37.277428	-121.808686	0.0	Disc Golf	Smoke Shop	Lake	Accessories Store	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop

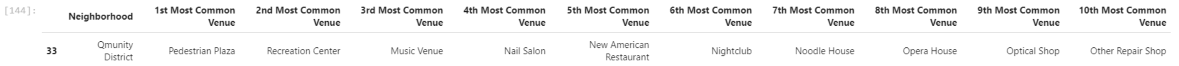
Second cluster:



[143]:

	Neighborhood	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Alum Rock	Construction & Landscaping	Music Store	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop	Outdoor Sculpture
53	West Valley	Construction & Landscaping	Music Store	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop	Outdoor Sculpture

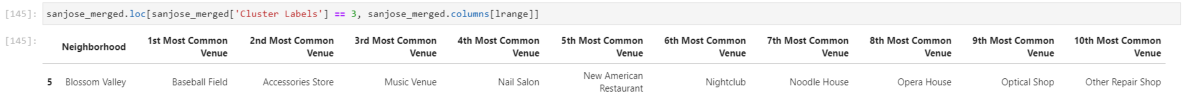
Third cluster:



[144]:

	Neighborhood	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
33	Qmunity District	Pedestrian Plaza	Recreation Center	Music Venue	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop

Forth cluster:



```
[145]: sanjose_merged.loc[sanjose_merged['Cluster Labels'] == 3, sanjose_merged.columns[lrange]]
```

[145]:

	Neighborhood	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Blossom Valley	Baseball Field	Accessories Store	Music Venue	Nail Salon	New American Restaurant	Nightclub	Noodle House	Opera House	Optical Shop	Other Repair Shop

6. Discussion

The cluster result shows San Jose city is a kind of 'average' city. The first cluster contains most of the neighborhoods. And all the neighborhoods in the first cluster has at least one kind of restaurant in its most 3 common venue. And the rest venues with high degree is coffee shop, grocery store, mall and so on. It indicates that the shopping trend is very high in this kind of neighborhood. The overall conclusion is that cluster one is the most suitable one for starting a restaurant. Within a close look at the neighborhoods, we recommend Alviso, Berryessa, Burbank, Downtown Histroic District, Downtown San Jose, East San Jose, Evergreen and South San Jose.

The second cluster is actually not very suitable for the restaurant since most interesting venues in this area are construction and landscaping company, music store and nail salon.

The third cluster is also not very suitable for the restaurant since the interesting venues are recreation center and nail salon. It has some restaurant but the most common venues are not showing the neighborhood has large trend on shopping and dining.

The fourth cluster also not suitable for restaurant since the most common venues are baseball field, accessories store, music venue and so on. All of the venues has no relation ship with restaurant at all.

7. Conclusion

In this project, the neighborhoods of San Jose, California has been successfully analyzed and we use the result to determine the best place to start a new restaurant. Based on the analysis result, the neighborhoods in cluster one are recommend. However, since the cluster one contains a large number of neighborhoods, a more detailed analysis can also conducted based on the interest of stake holders and investors.

Notes:

The repo for this project is: https://github.com/ArizonaTea/Coursera_Capstone