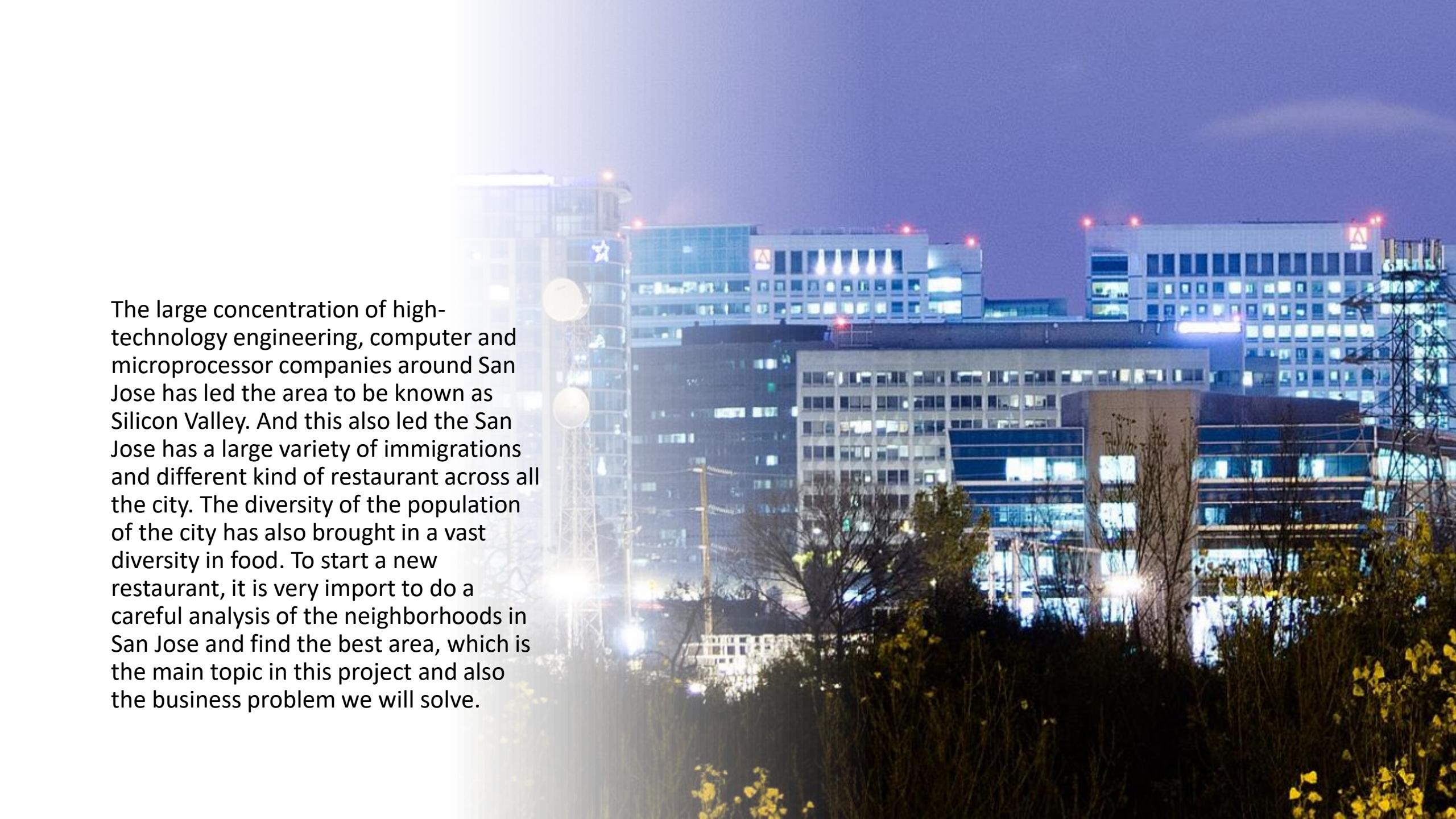


Analysis of the Neighborhoods in San Jose, California for Starting a new Restaurant

Capstone Project - The Battle of Neighborhoods

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.

A nighttime photograph of a city skyline, likely San Jose, California. The image shows several large, modern office buildings with many windows lit up, indicating they are still in use. A prominent communication tower with two large satellite dishes is visible on the left side. The foreground is dark, with some foliage and a few yellow flowers visible on the right. The sky is a deep blue, suggesting twilight or early night.

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.

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.

Data



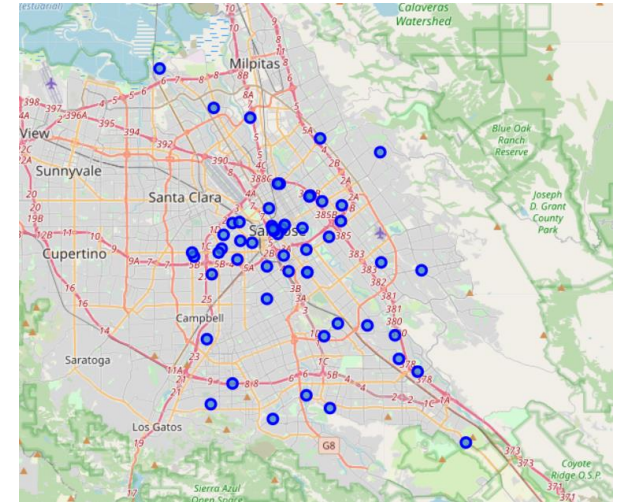
Neighborhood Data

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

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.



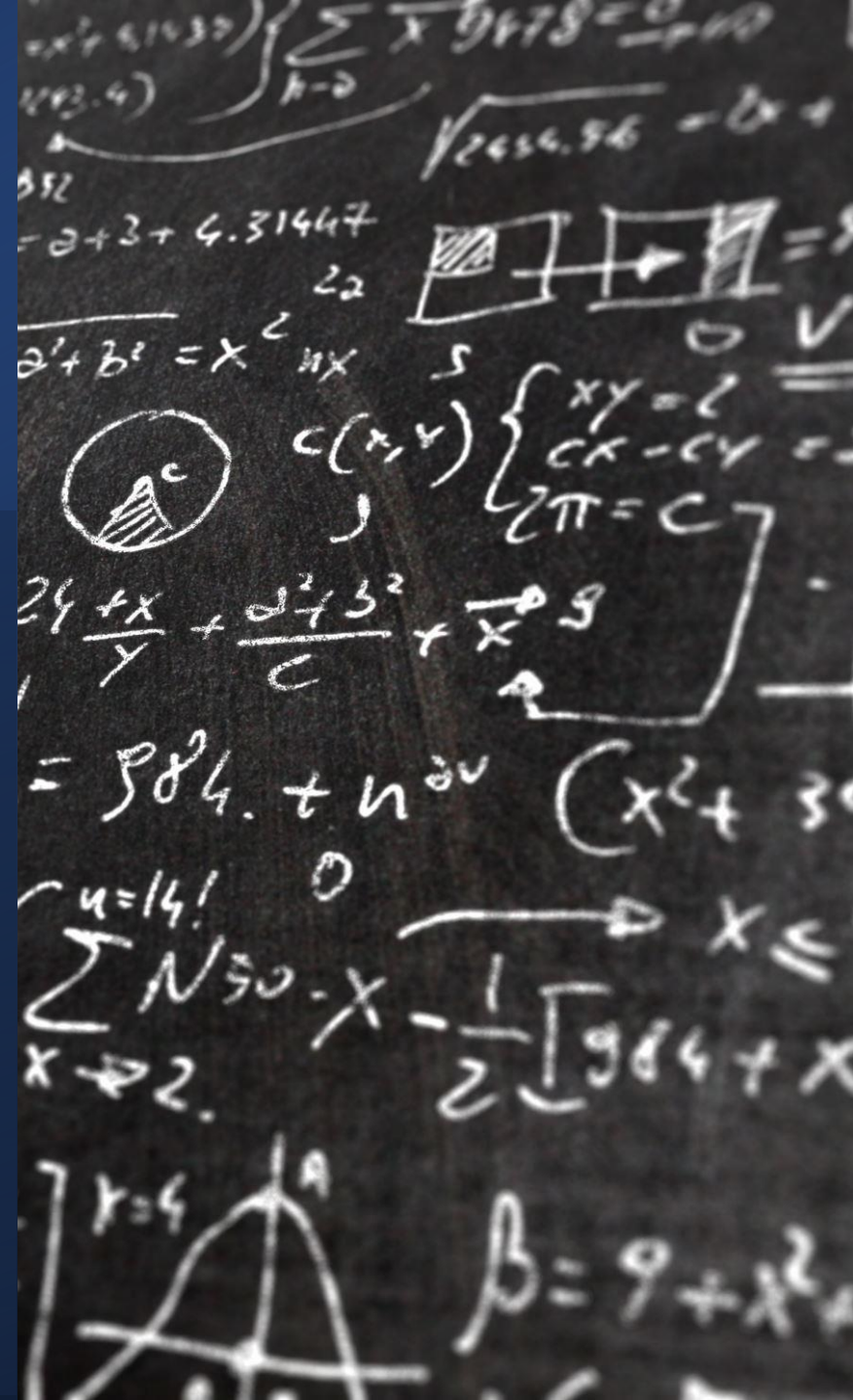
	Neighborhood	Latitude	Longitude
0	The Alameda	37.339853	-121.922102
1	Almaden Valley	37.231118	-121.894036
2	Alum Rock	37.378805	-121.819188
3	Alviso	37.425400	-121.973220
4	Berryessa	37.386340	-121.860750
5	Blossom Valley	37.239169	-121.937536
6	Buena Vista	37.319650	-121.918550
7	Burbank	37.325300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

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

Methodology



Feature Extraction

With the help of our Foursquare API, we can extract 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----

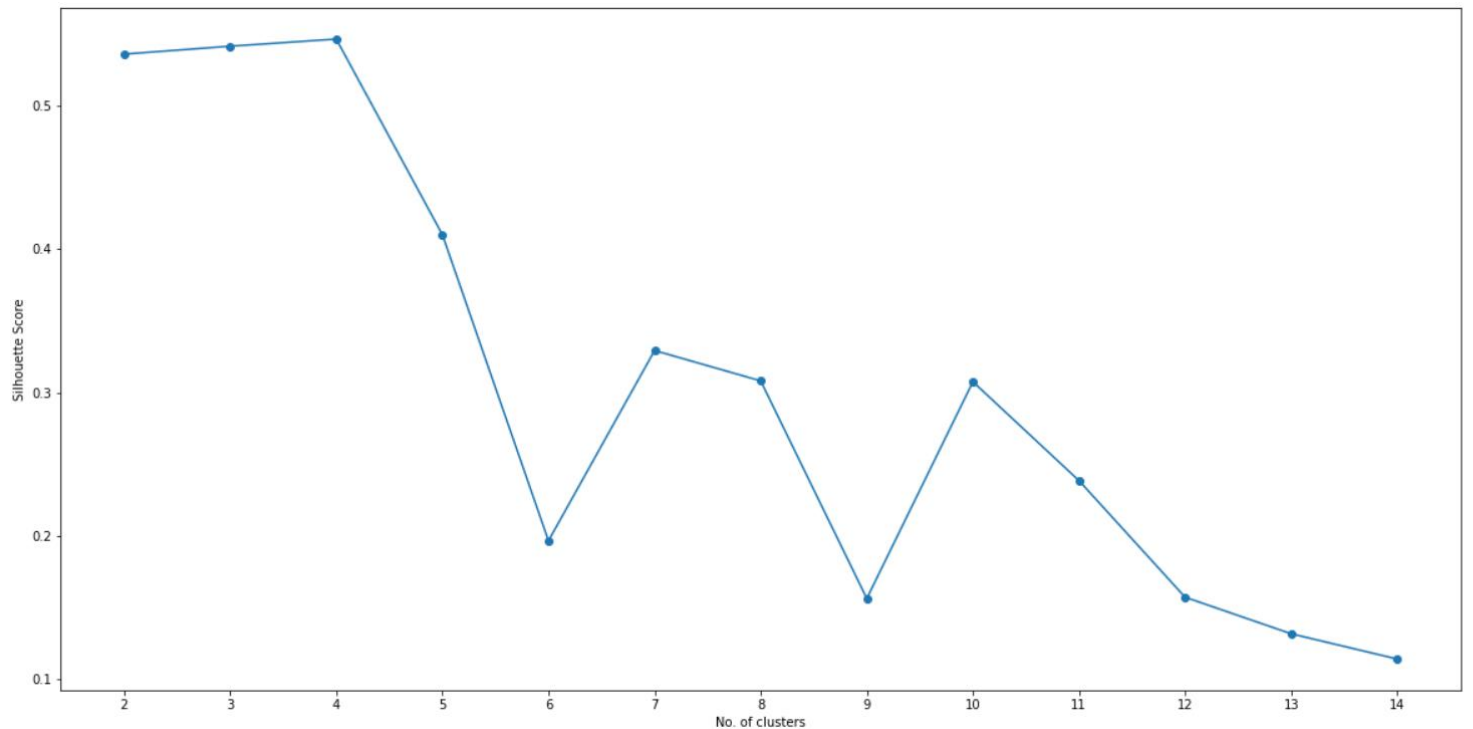
	venue	freq
0	Mexican Restaurant	0.25
1	Convenience Store	0.12
2	Food Truck	0.12
3	Restaurant	0.12
4	River	0.12

----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

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:

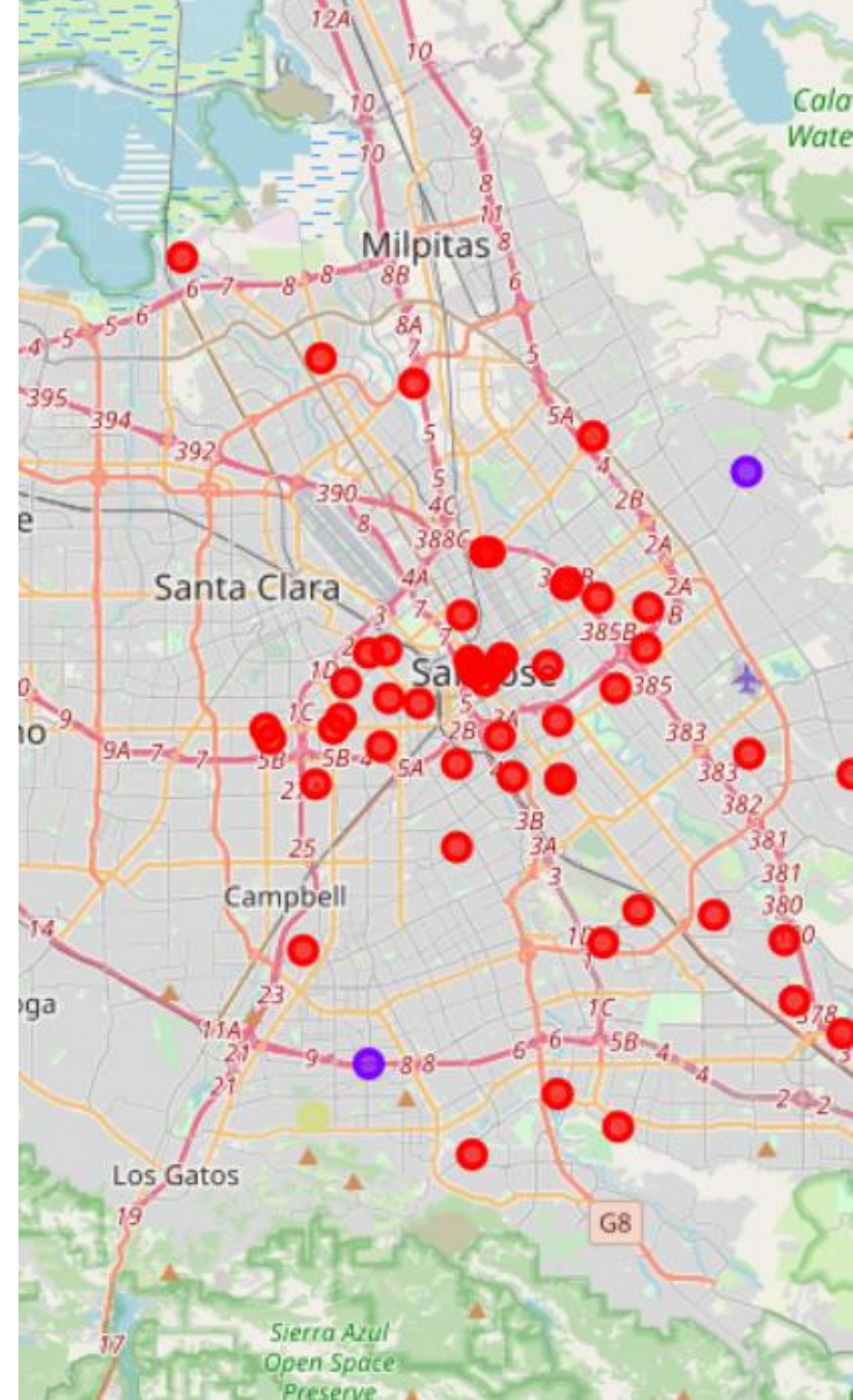


Results



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





Discussion

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 Historic 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.

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.

Thank You!

