VENUES AND LIVING IN SAN JOSE

Capstone Project - The Battle of Neighborhoods

Samaa Gazzaz

Applied Data Science Capstone
IBM Data Science Professional Certificate

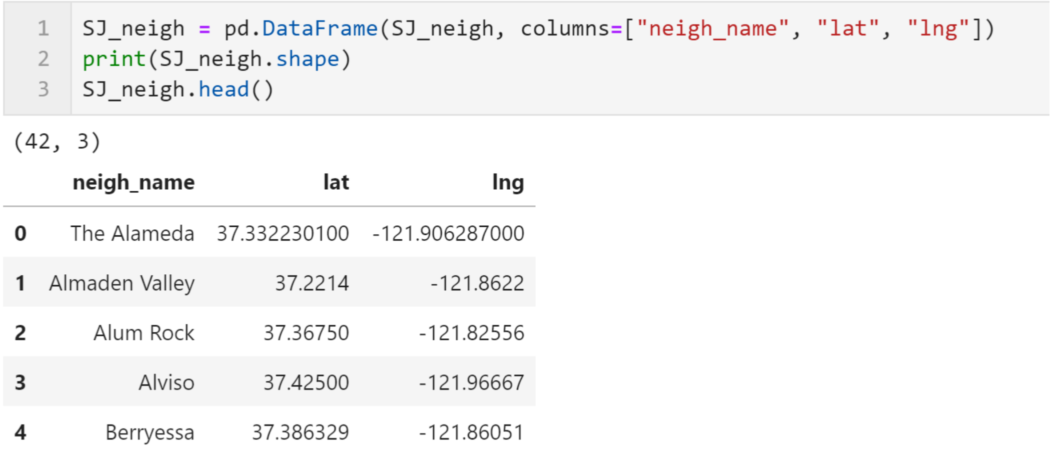
Introduction/Business Problem

Every year, more that 7 million Americans move to a different state within the US [1]. Along with it, comes the responsibility of finding a good location to lay roots and settle. One of the major indicators that people relay on when deciding where to move is the availability of venues that fit their needs and lifestyle. Having moved to San Jose lately myself, I discuss in this report the different venues in San Jose neighborhoods, their types, distribution and clusters that they fall into. By the end of the report, the knowledge produced shall be sufficient in guiding the moving decision based on venues nearby, and choosing the best fit neighborhood depending on one's needs.

Data

In this project, we focus on the city of San Jose and its neighborhood. Wikipedia provides a comprehensive list of San Jose's neighborhoods that we will be utilizing for this project [2]. As you can notice, the list is organized alphabetically, but other than that not much information is provided in this page other than the neighborhood names. Luckily, each neighborhood name links to a page dedicated to that neighborhoods information from where we shall acquire the latitude and longitude information of the center of that neighborhood.

There are multiple challenges with these data: First, some of the neighborhoods don't provide latitude and longitude information; thus, we eliminate such neighborhoods from our experiment. Next, the distances between the neighborhoods' centers are not normalized which causes duplicate venue entries when retrieving nearby venues. This challenge and its solution are discussed in the methodology section.



```
SJ_neigh = pd.DataFrame(SJ_neigh, columns=["neigh_name", "lat", "lng"])
print(SJ_neigh.shape)
SJ_neigh.head()
```

```
(42, 3)
```

	neigh_name	lat	lng
0	The Alameda	37.332230100	-121.906287000
1	Almaden Valley	37.2214	-121.8622
2	Alum Rock	37.36750	-121.82556
3	Alviso	37.42500	-121.96667
4	Berryessa	37.386329	-121.86051

Figure1: This shows a snippet of the code and the top of the dataframe holding all the neighborhood data of San Jose and their locations.

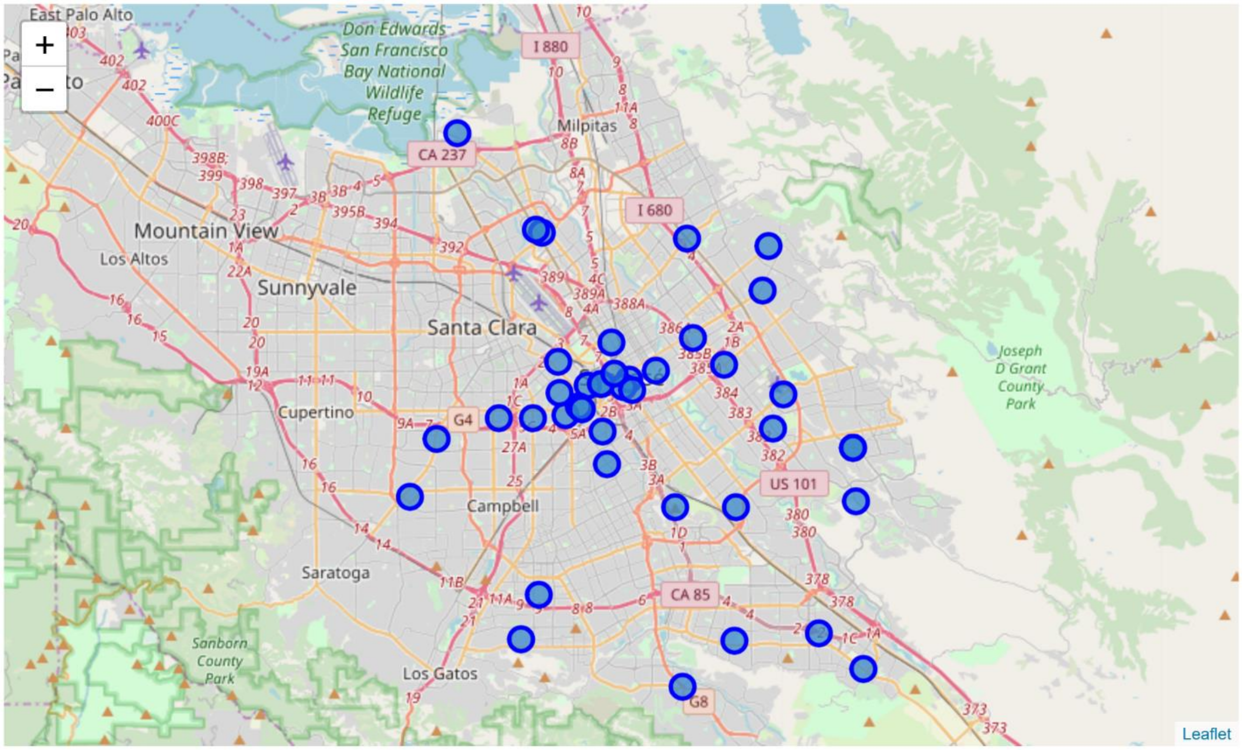


Figure 2: San Jose map showing locations of neighborhood centers.

Other data used in this project, is the data retrieved using the Foursquare API. Specifically, nearby venues to each neighborhood are recorded and utilized in understanding and qualifying the neighborhoods.

Methodology

In order to further explore these neighborhoods, we are in need of more information in regard to the nearby venues of each neighborhood. We are able to retrieve this information by requesting it from the Foursquare API. In order to understand the data, we create a dataframe where each row is a venue in San Jose and we list all its information. When we did so, we notice a very disturbing problem. There is a count of 1122 venues in the list, however, when the list is sorted, we notice huge redundancy in the venues. With further exploration, it is clear that the problem is that the neighborhood centers are too close as seen in figure 2 causing the venues within a specific radius to be returned with both calls from all nearby neighborhood centers.

In order to address this problem, we decided to remove centers that are too close where they almost represent the same location. The removed neighborhood names are ‘North San Jose Innovation District', 'West San Carlos', 'San Jose', 'Downtown Historic District (San Jose' and 'Downtown San Jose'. After the removal of these centers, we notice the decrease in the number of venues in San Jose when retrieved from Foursquare. This does not represent losing vital information, instead it is a sign of removing redundant venues. The number of venues decreased from 1122 venues to around 860.

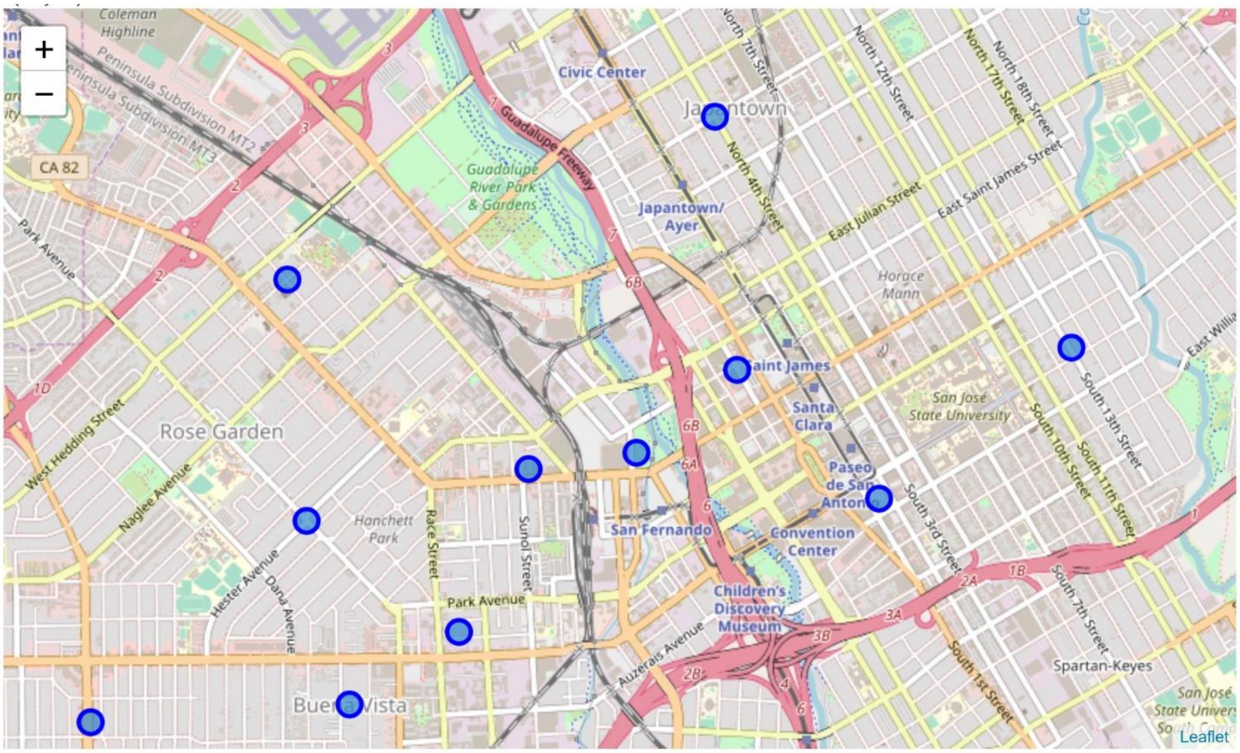


Figure 3: This map demonstrates how removing unnecessary neighborhood centers resulted in clear separation of neighborhoods.

When moving, we assume that the couple are looking for a lively neighborhood, which enlists a lot of nearby venues. This way, we can look for the neighborhoods with the greatest number of close by venues to start our search. We notice these neighborhoods in figure 4 as the ones with the higher count of nearby venues.

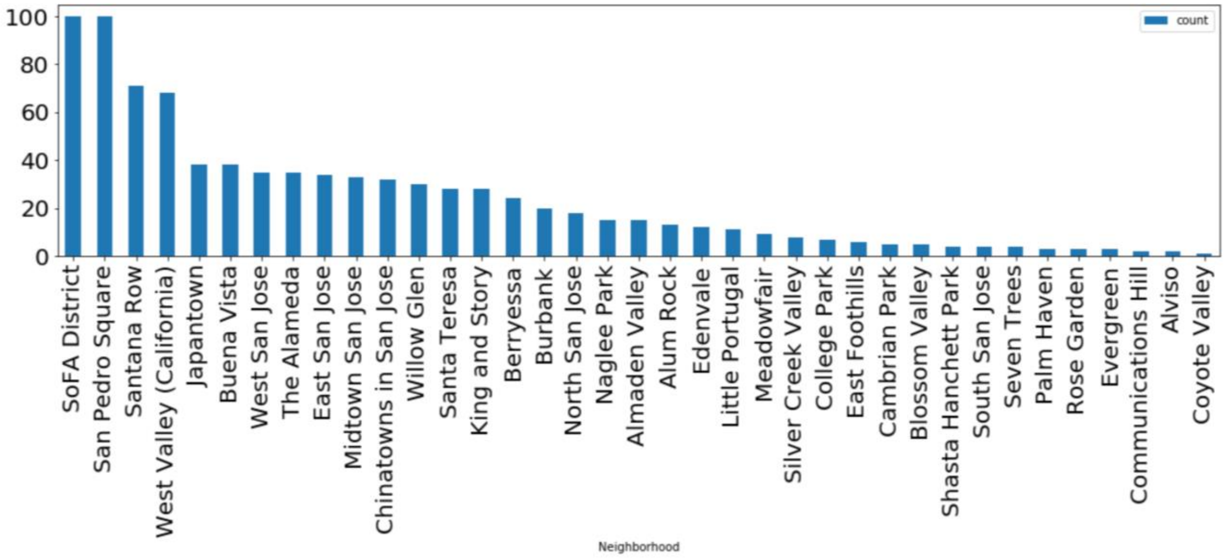


Figure 4: Neighborhoods sorted by the highest count of nearby venues

From the chart, we notice that SoFA (South of First Area) and San Pedro Square are the neighborhoods with the greatest number of venues. This is very predictable as these two areas are very close to the San Jose Downtown area.

Another very important exploration is the exploration of the most frequent venue category appearing in San Jose. This is done by grouping the list of all venues by their categories. We can view the results as the bar chart shown in figure 5.

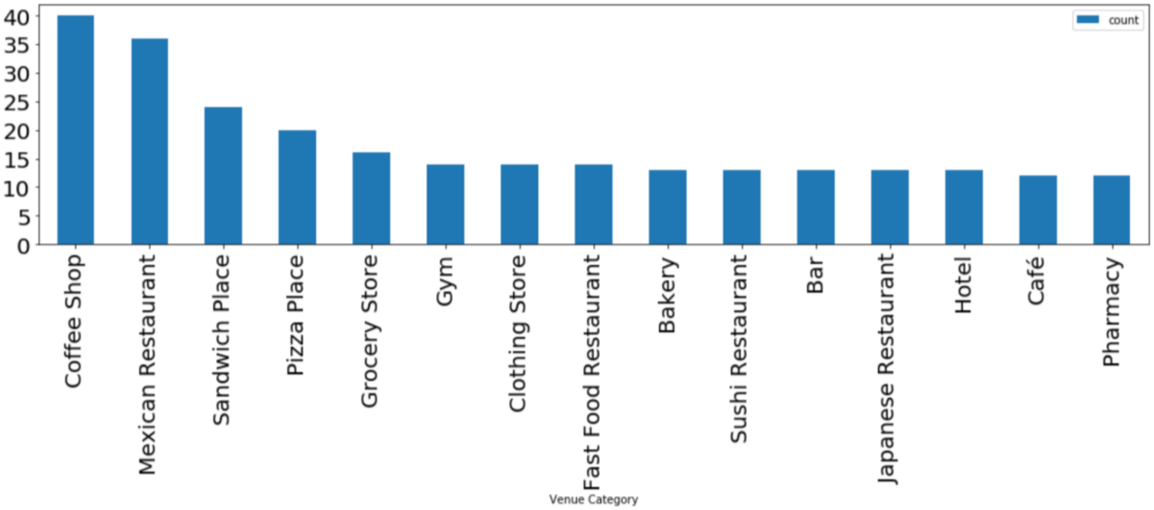


Figure 5: Listing the most frequent venue categories in San Jose

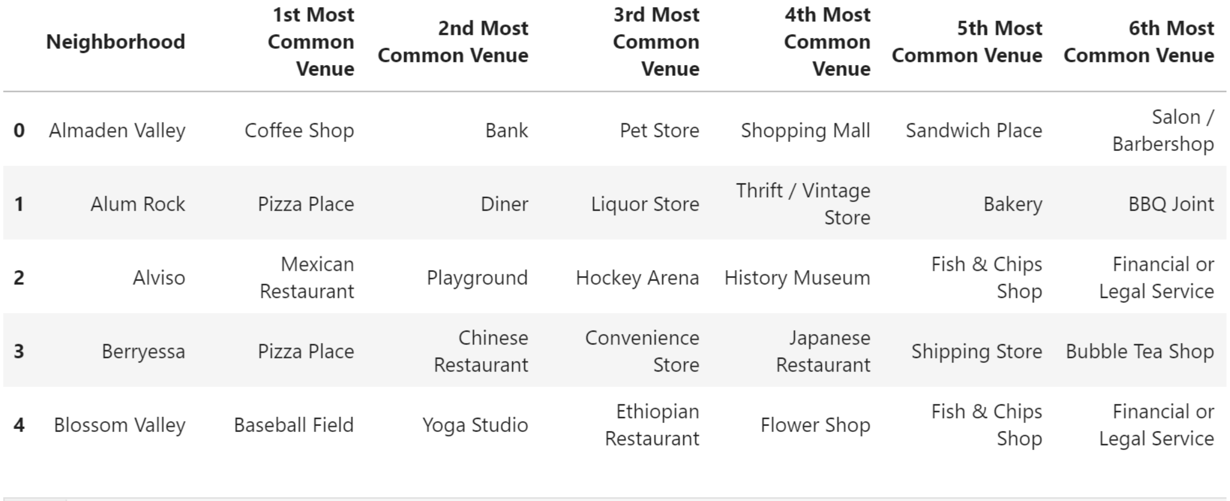
The most frequent venue is Coffee Shops. This is great news as our hypothetical couple moving to San Jose are very interested in finding a location that features a lot of coffee shops and sandwich joints. We notice from this chart that sandwich places are the third most frequent venue in San Jose.

Moreover, we are very interested in knowing the most frequent venue categories in each of the neighborhoods. We can use one hot technique to rule out the venues we are most interested in. We can notice from the results that there are 3 different neighborhoods that feature these two kinds of venues as the most frequent venue in the hood. These neighborhoods are: The Alameda neighborhood, San Pedro Square and Naglee Park.

Results

From the previous exploratory knowledge, we can determine that San Jose is a good match for our couple. There are a great number of coffee shops and sandwich venues. In addition to neighborhoods being close to each other’s leading to notice the ease of visiting venues in other neighborhoods.

Moreover, it is a great time to explore how the neighborhoods of San Jose connect to each other and how similar these areas are. In order to discover these relationships, we employ the K-means clustering algorithms. We use this algorithm because we do not have set labels that we could use to predict relations. Also, we are interested in finding out the clusters of neighborhoods and labels to these clusters are not very important. We set the number of clusters we would like to categorize our neighborhoods in to 4. The following figure views the results.



	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
0	Almaden Valley	Coffee Shop	Bank	Pet Store	Shopping Mall	Sandwich Place	Salon / Barbershop
1	Alum Rock	Pizza Place	Diner	Liquor Store	Thrift / Vintage Store	Bakery	BBQ Joint
2	Alviso	Mexican Restaurant	Playground	Hockey Arena	History Museum	Fish & Chips Shop	Financial or Legal Service
3	Berryessa	Pizza Place	Chinese Restaurant	Convenience Store	Japanese Restaurant	Shipping Store	Bubble Tea Shop
4	Blossom Valley	Baseball Field	Yoga Studio	Ethiopian Restaurant	Flower Shop	Fish & Chips Shop	Financial or Legal Service

Figure 6: The dataframe that summarizes the 6 most common venues in each neighborhood (only the first 5 neighborhoods are shown)

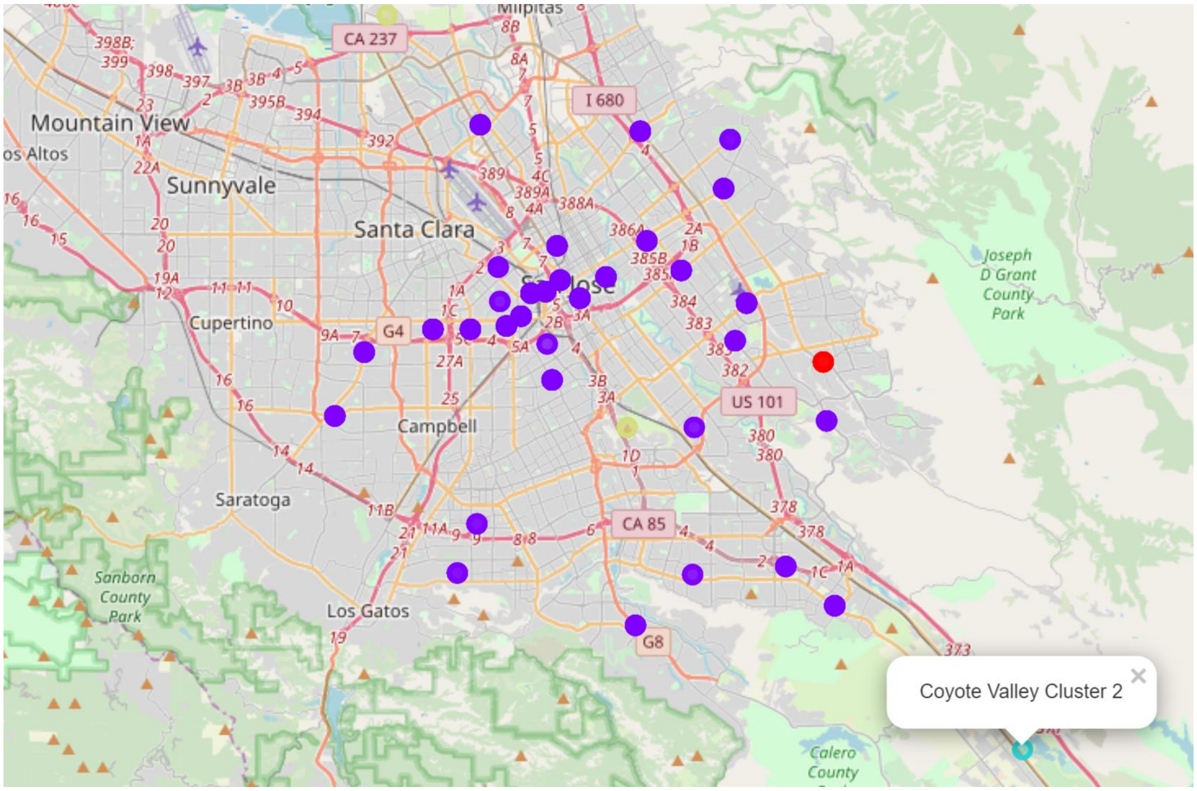


Figure 7: Clustered neighborhoods of San Jose

Discussion

From the last map and the inspection of each cluster that we did in the implementation, we notice huge gain in knowledge. Most of San Jose's areas/neighborhoods are very similar when it comes to nearby venues. The two unique areas detected were the Coyote Valley on the very south end of San Jose and the Rose Garden area in the east. These areas can be labeled as rural or could be considered suburbs of San Jose.

Going back to our client couple who like a lot of lively venues, we can definitely encourage moving in to San Jose. There is a great number of venue varaiety in the area. Moreover, the most frequent venue in San Jose is coffee shops which is something that the couple are very interested in. Since they are also interested in Sandwich locations, we found that the frequency of coffee shops and sandwich places in The Alameda, Naglee Park and San Pedro Square are very high. However, according to our earliest exploration, we notice that San Pedro Square is one of the neighborhoods with the greatest number of venues. This leads us to recommend San Pedro Square as the best match for our couple to move into.

Conclusion

The problem at hand was revolving around recommending the best location for living to a new couple moving to San Jose. The preferences were a lively area where lots of coffee shops and sandwich venues can be found. Data detailing the neighborhoods of San Jose was scrapped off of Wikipedia along with its latitude and longitude information. In addition, Foursquare API was utilized in acquiring close by venues in regards to each neighborhood. We employed multiple techniques in our approach to clean and prepare the data. Furthermore, we took advantage of clustering algorithms in gaining knowledge about San Jose neighborhoods. Finally, we were able to draw great conclusions and recommend the best fit neighborhoods based on the preferences given and the knowledge gained from the process.

References

[1] New State Residents Statistics, Demographic Data. Retrieved April 4, 2019, from https://www.governing.com/gov-data/residents-moving-to-new-state-demographics-population-statistics.html

[2] Category:Neighborhoods in San Jose, California. (2017, November 15). Retrieved April 4, 2019, from https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Jose,_California