

# The Battle of Neighborhoods

## IBM Applied Data Science Capstone - Final Report

Henry Hong – 2019-07-28

# Table of Content

- 1. Introduction Section**
- 2. Data Section**
- 3. Methodology Section**
  - 1) San Jose Map - Current residence and venues in San Jose**
  - 2) Import Schools Datasets**
  - 3) Import Parks dataset**
  - 4) Import Railroad Stations dataset**
  - 5) Import Public Libraries dataset**
  - 6) Import Crime Reports dataset**
  - 7) Download the house price trends data**
  - 8) Process the ZIP codes and Cities**
  - 9) Explore Cities in Santa Clara County**
  - 10) Import Recent Sold House Price Data**
- 4. Results Section**
- 5. Discussion Section**
- 6. Conclusion Section**

# 1.0 Introduction Section

# Business Problem

How to buy a dream house in Silicon Valley which complies with the requirements of price, features, safety, location and venue favors?

## The audience who also interested

This case is also applicable for anyone interested in exploring the ways of searching and analysis the location and real estate data for finding a suitable house to buy in Silicon Valley.

# Requirements

- The amenities in the selected neighborhood shall be similar to client's current residence apartment
- The **price** is under **1.5M**
- House must be at least **3** bedrooms, **2** bathrooms, **1** car garage, around **1800** to **2100** square footage of size
- Near the **park** (within **0.5** mile)
- Near the **library** (within **1** mile)
- Near the **school** (within **0.5** mile)
- The **schools** in the area should have high **rating** (Ranking greater than **7**)
- Not too close to the **railroad** (at least **3 miles** away)
- The location is near the **supermarket** (within **0.5** mile radius)
- The location is near the **shopping mall** (within **3 miles** radius)
- The location is close (within **1 mile**) to **venues** such as restaurants (Asian and Mexican foods ...etc.), parks and coffee shops
- The neighborhood/community should be safe and have low **crime rate**

## 2.0 Data Section

# Datasets and features to be extracted

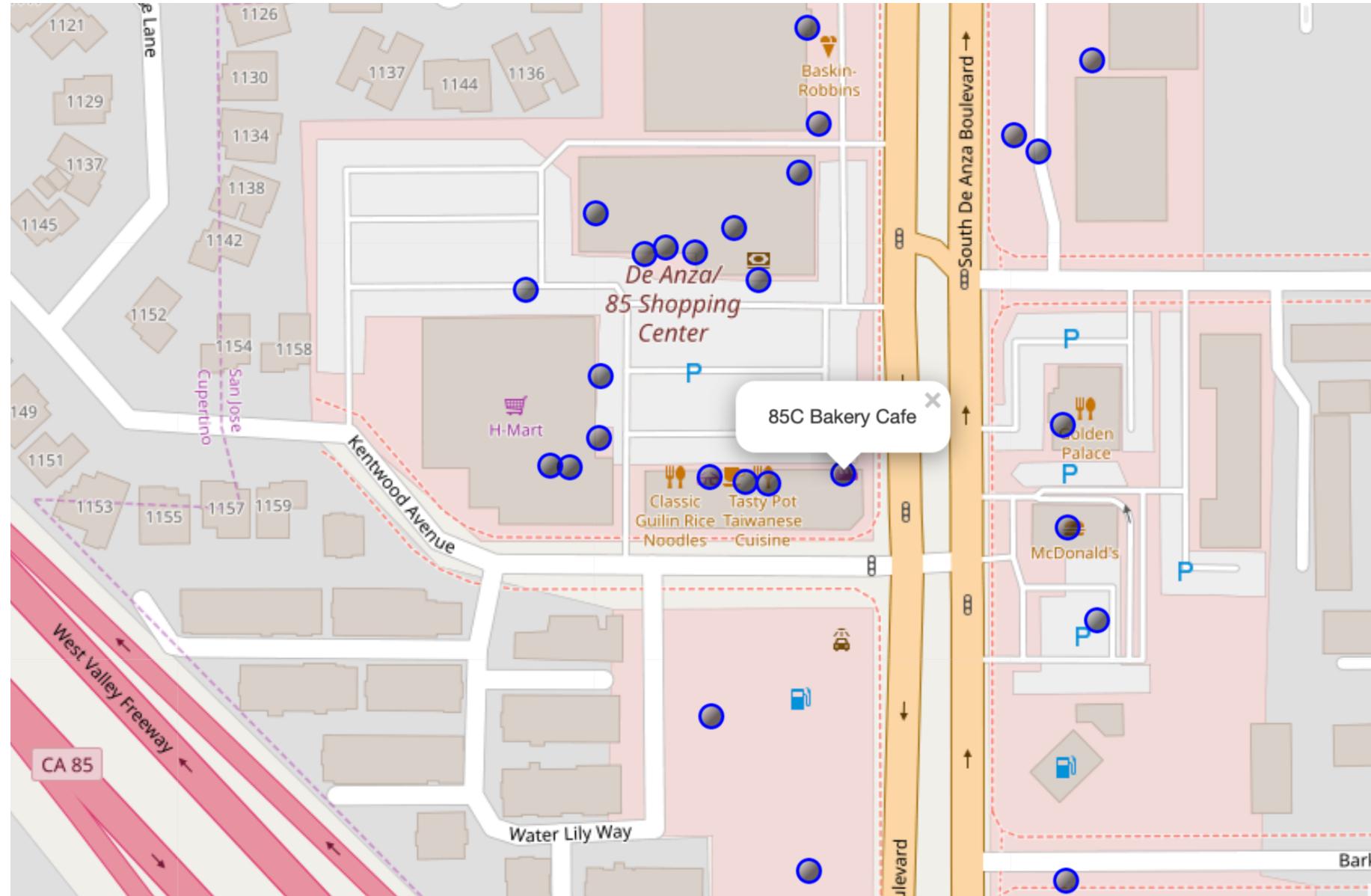
- Schools Dataset: ['ZIP', 'CITY', 'PLACENAME', 'ADDRESS', 'LATITUDE', 'LONGITUDE']
- School Rating Dataset(WebScraping by BeautifulSoup): ['SCHOOL', 'RANK']
- Parks dataset: ['PLACENAME', 'LATITUDE', 'LONGITUDE']
- Railroad stations dataset: ['PLACENAME', 'PLACETYPE', 'LATITUDE', 'LONGITUDE']
- Public Libraries dataset: ['ZIP', 'CITY', 'PLACENAME', 'ADDRESS', 'LATITUDE', 'LONGITUDE']
- Crime Reports dataset: ['city', 'incident\_type\_primary', 'parent\_incident\_type', 'latitude', 'longitude']
- House price trends data: ['City', 'Median Price', '2018 Change', '2017 Change', '2016 Change']
- Foursquare API: ['Neighborhood', 'Venue Category', 'venue', 'freq']
- Recent Sold House Price Data: ['ZIP', 'ADDRESS', 'ORGLD', 'ORIG LSPRC', 'LIST PRICE', 'SALE PRICE', 'SQFT', 'LOTSZ', 'COE', 'DOM']
- Zillow API Comparable Sales Analysis: ['zpid', 'zipcode', 'city', 'street', 'year\_built', 'lot\_size', 'finished\_size', 'bedrooms', 'bathrooms', 'zestimate', 'last\_sold\_price', 'last\_sold\_date', 'latitude', 'longitude', 'home\_details', 'similar\_sales']

# DATA processing - To answer following questions

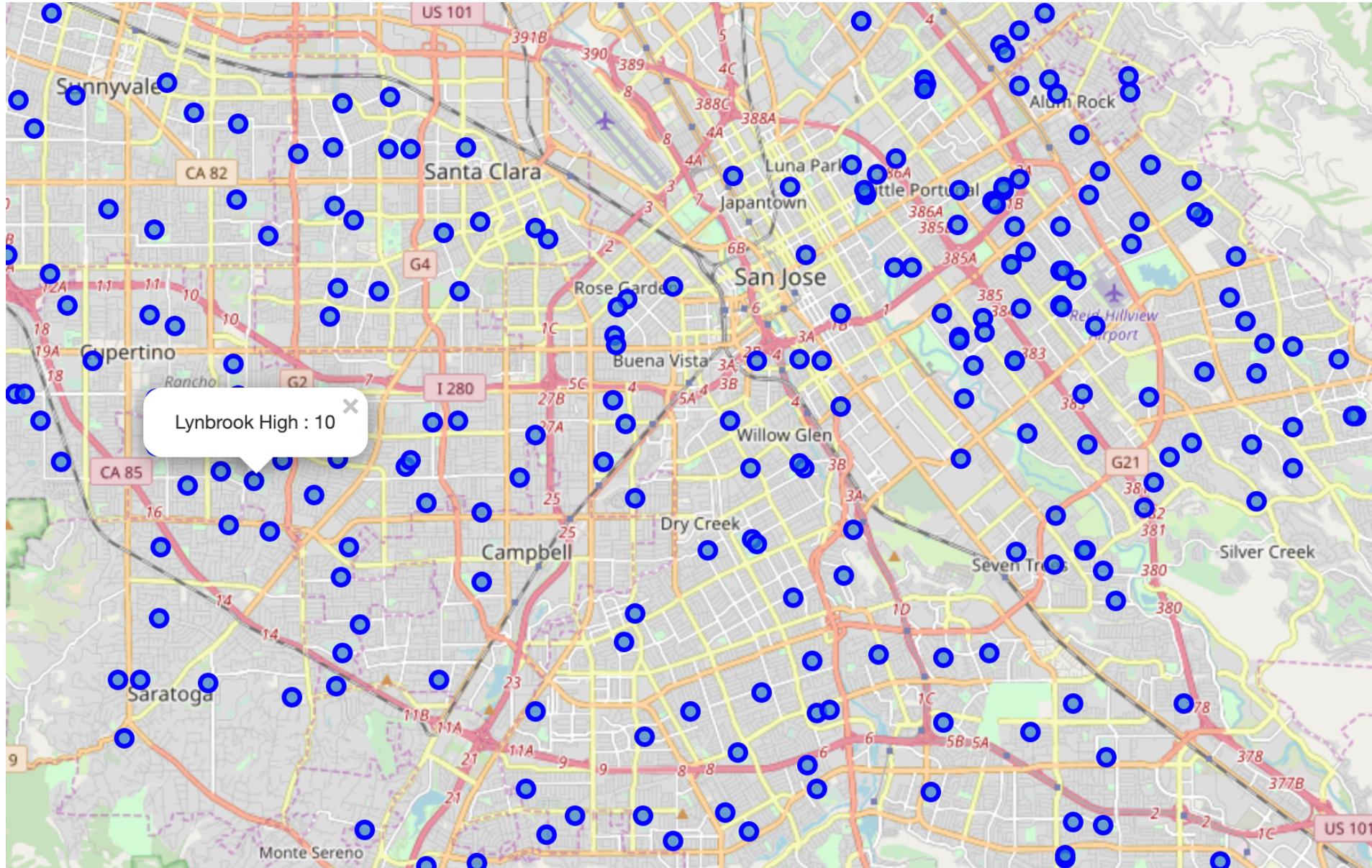
- Which city/zipcode is good to consider while considering the **price** (less than 1.5M)?
- Which city/zipcode is good to consider while considering the **school** (within 0.5 mile and ranking > 7)?
- Which city/zipcode is good to consider while considering the **park** (within 0.5 mile)?
- Which city/zipcode is good to consider while considering the **library** (within 1 mile)?
- Which city/zipcode is good to consider while considering the **railroad** (at least 3 miles away)?
- Which city/zipcode is good to consider while considering the **supermarket** (0.5 mile radius)?
- Which city/zipcode is good to consider while considering the **shopping mall** (within 3 miles radius)?
- Which city/zipcode is good to consider while considering the **foods, restaurants (Asian and Mexican foods ...etc) and coffee shops?**
- Which city/zipcode is good to consider while considering the **safety** (low crime rate)?
- Which city/zipcode has the **best housing pricing**?
- Are there **tradeoffs** between size, price, location and safety?
- What are the venues of the **two best house** to buy?
- Any other **interesting statistical data** findings from the **Foursquare** and **ZillowAPI** real estate data?

# 3.0 Methodology Section

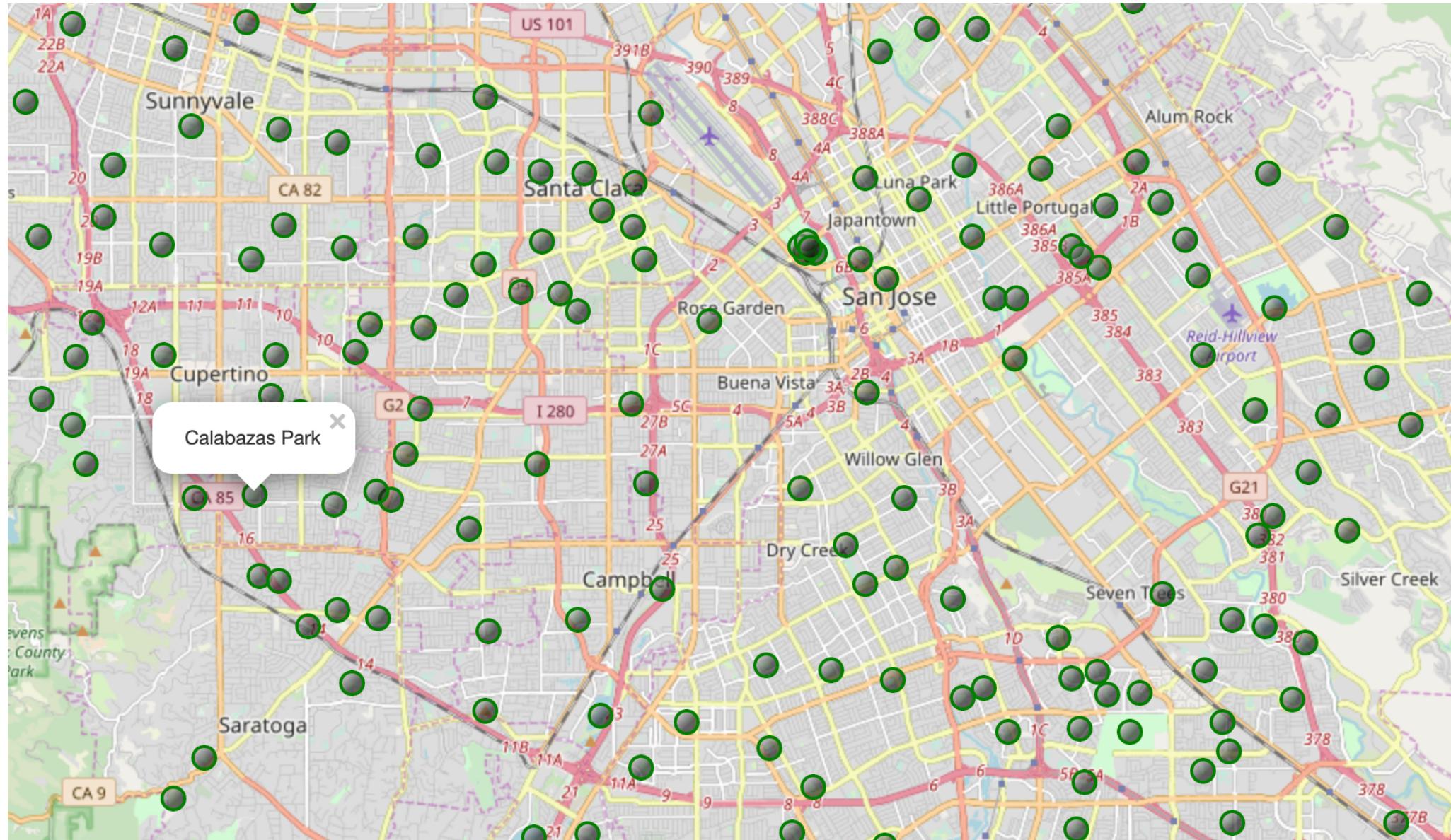
# San Jose Map - Current residence and venues in San Jose (95129)



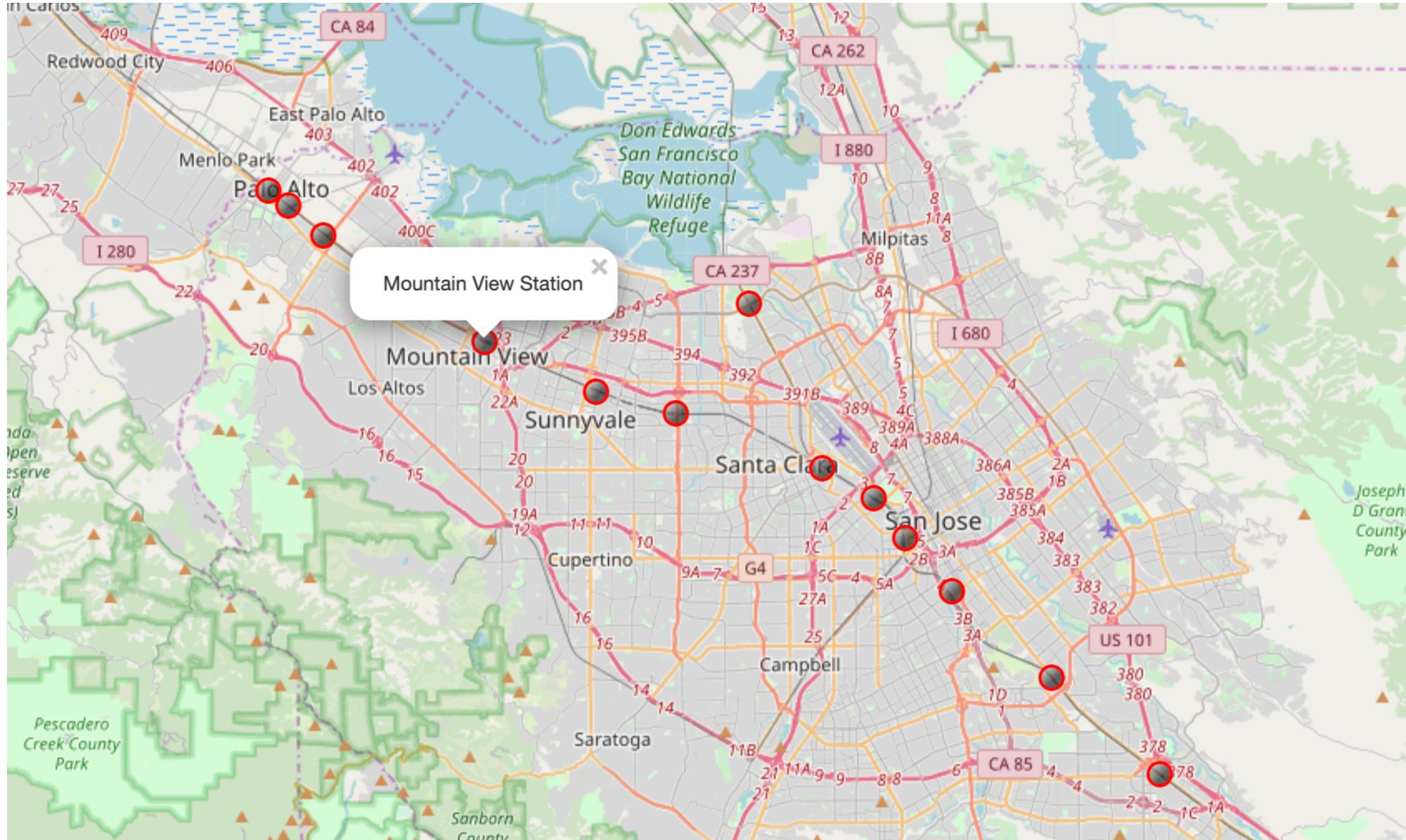
# Schools in Santa Clara County



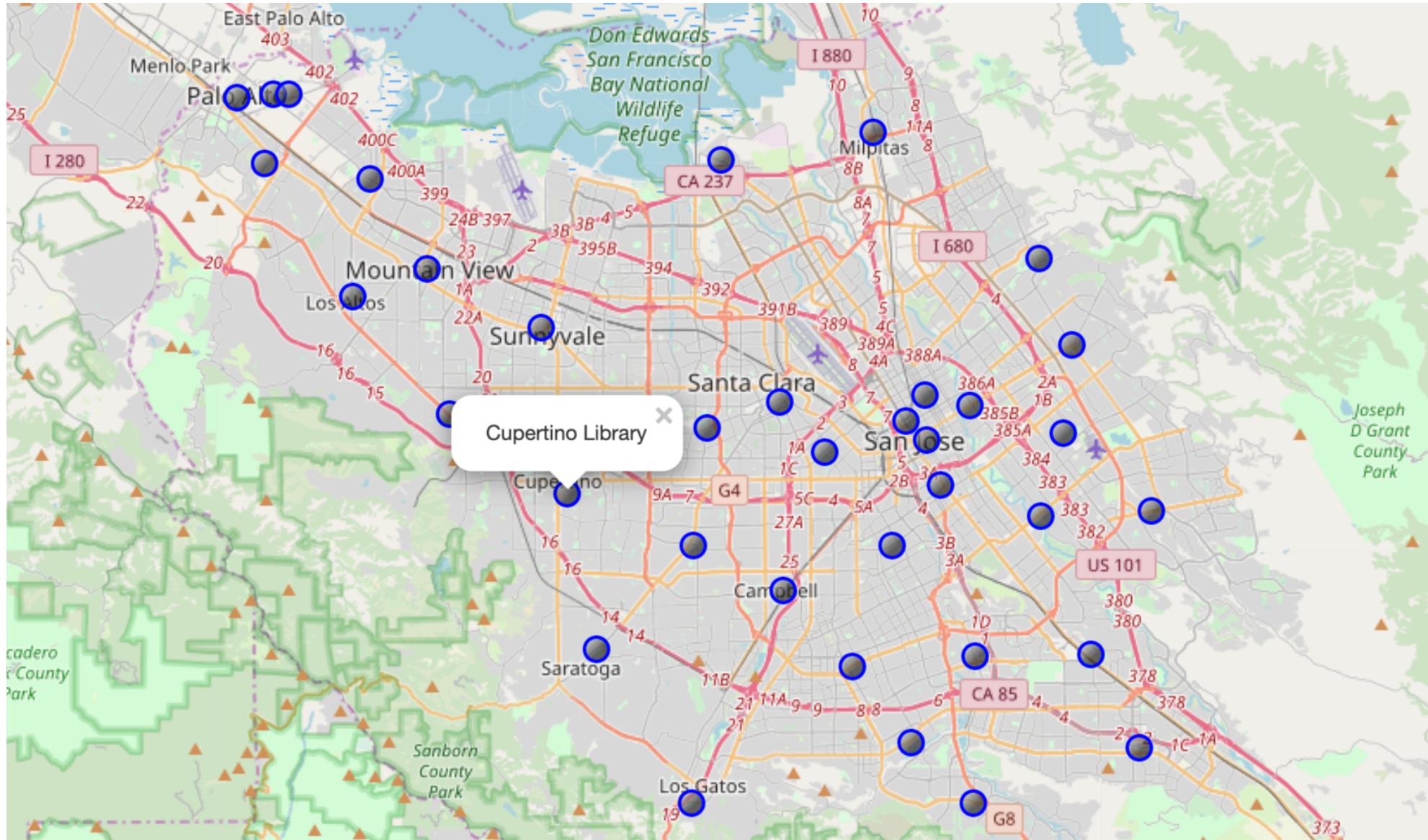
# Parks in Santa Clara County



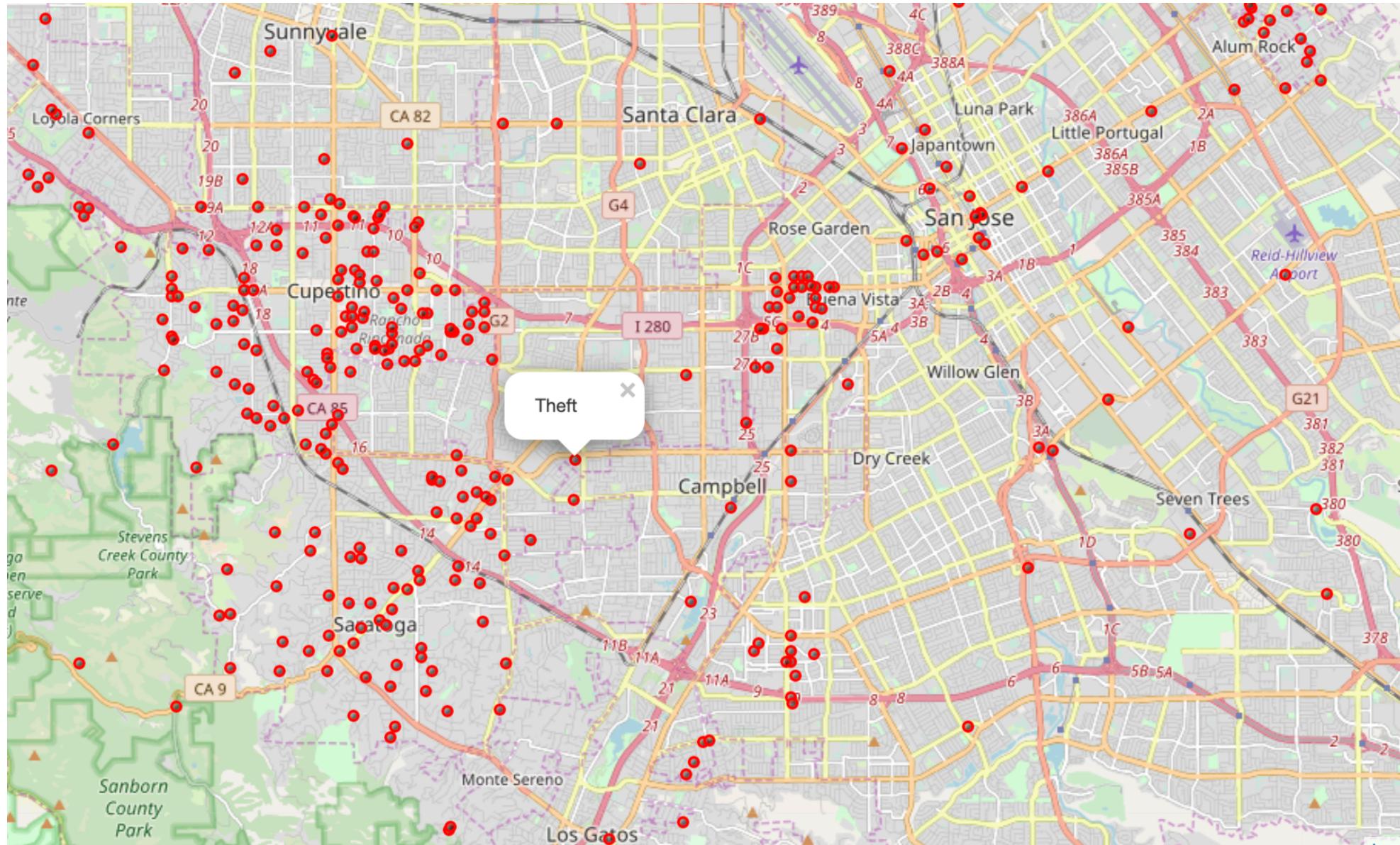
# Railroad Stations in Santa Clara County



# Public Libraries in Santa Clara County

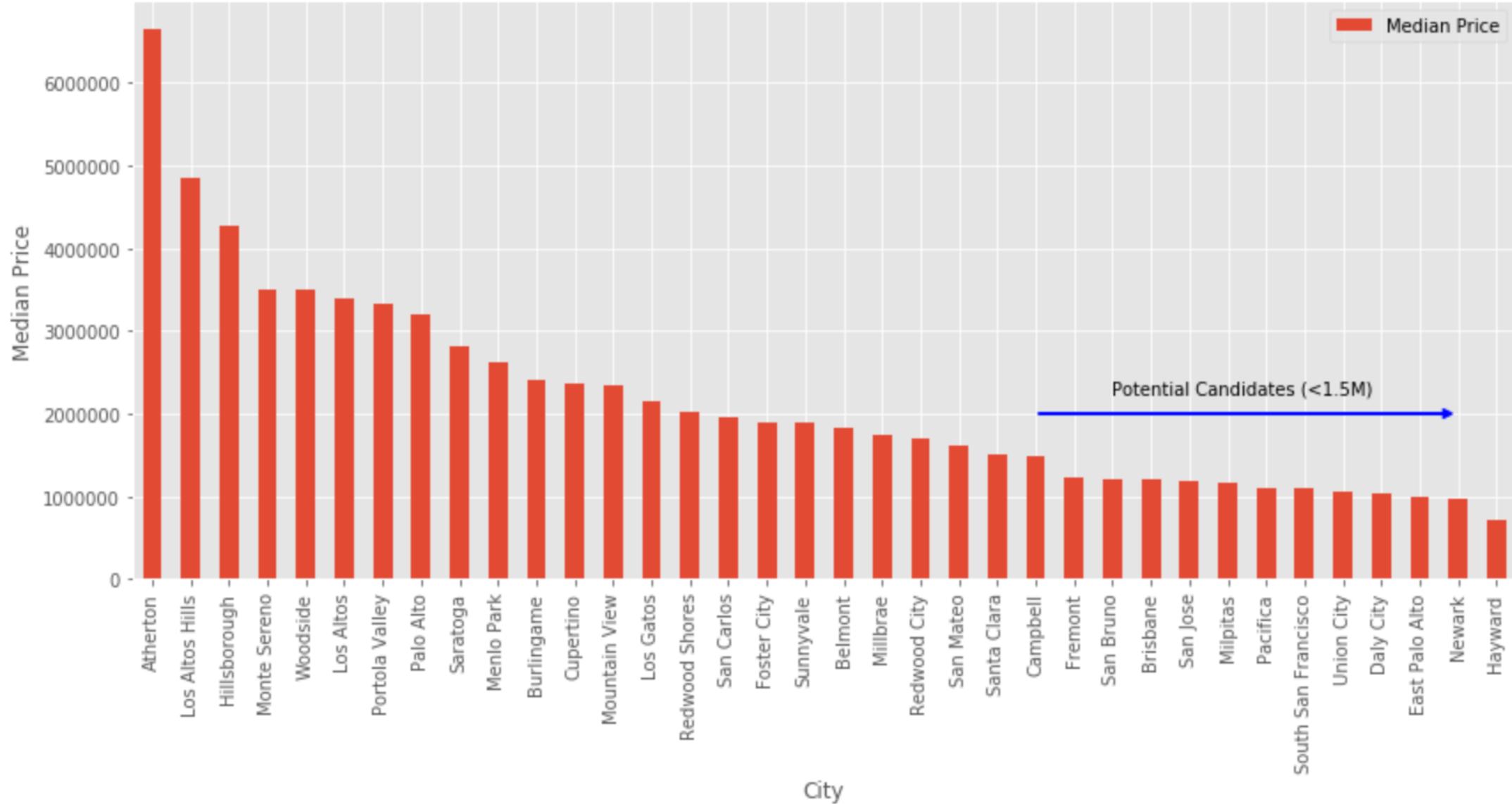


# Crime Reports in Santa Clara County

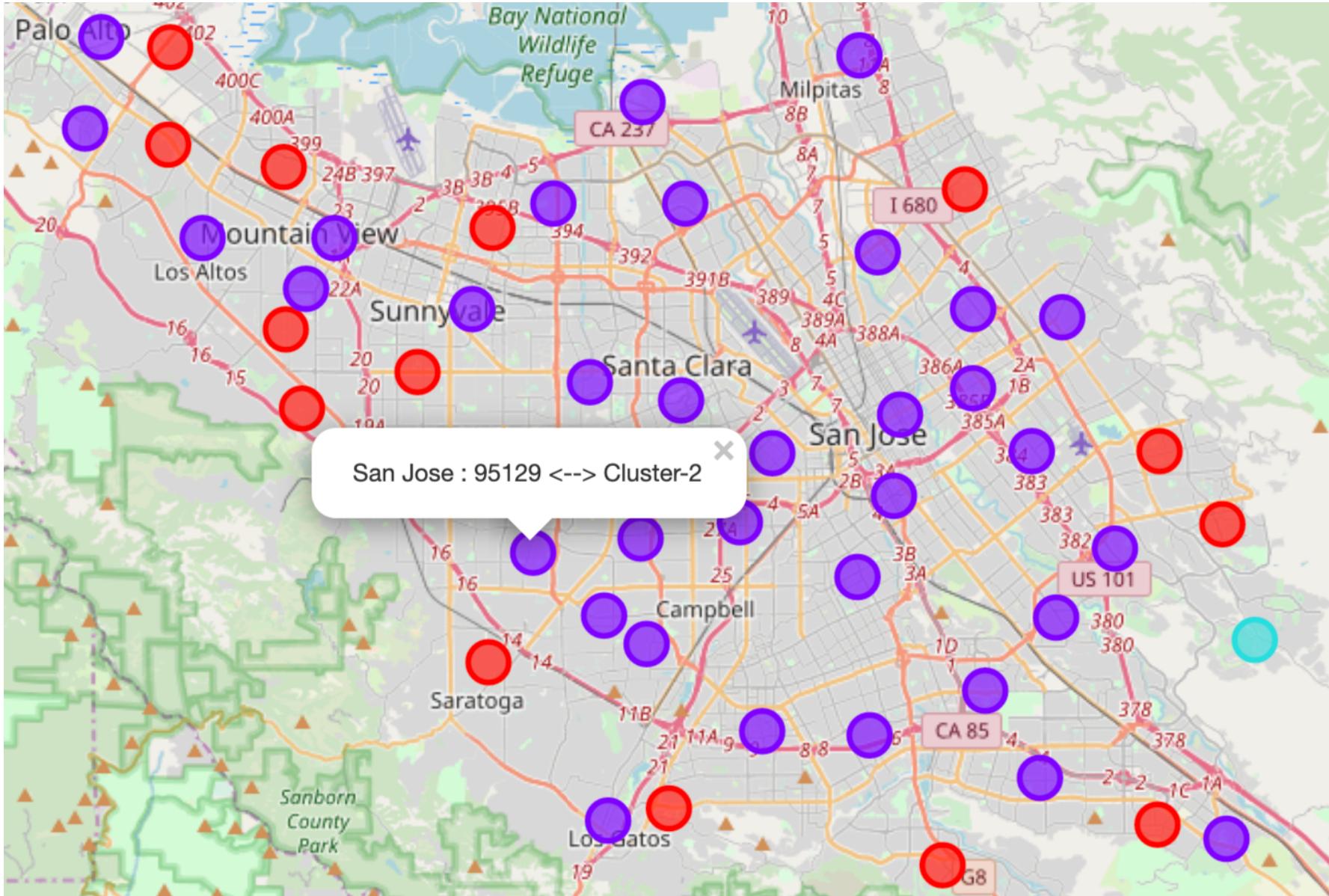


# Median Prices For Silicon Valley Cities

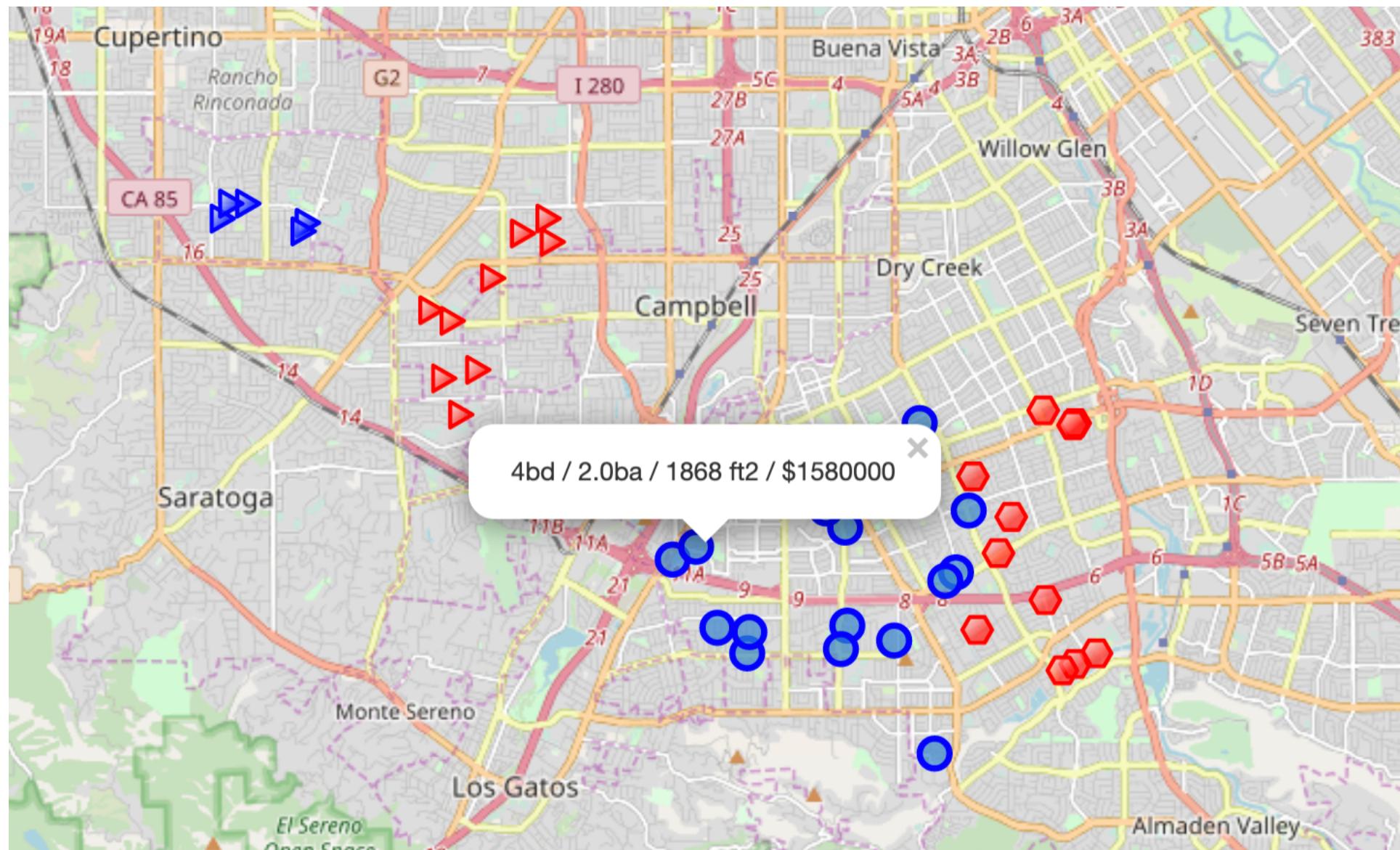
2019 Silicon Valley Cities Single House Median Price



# Clustering Cities/Zipcodes



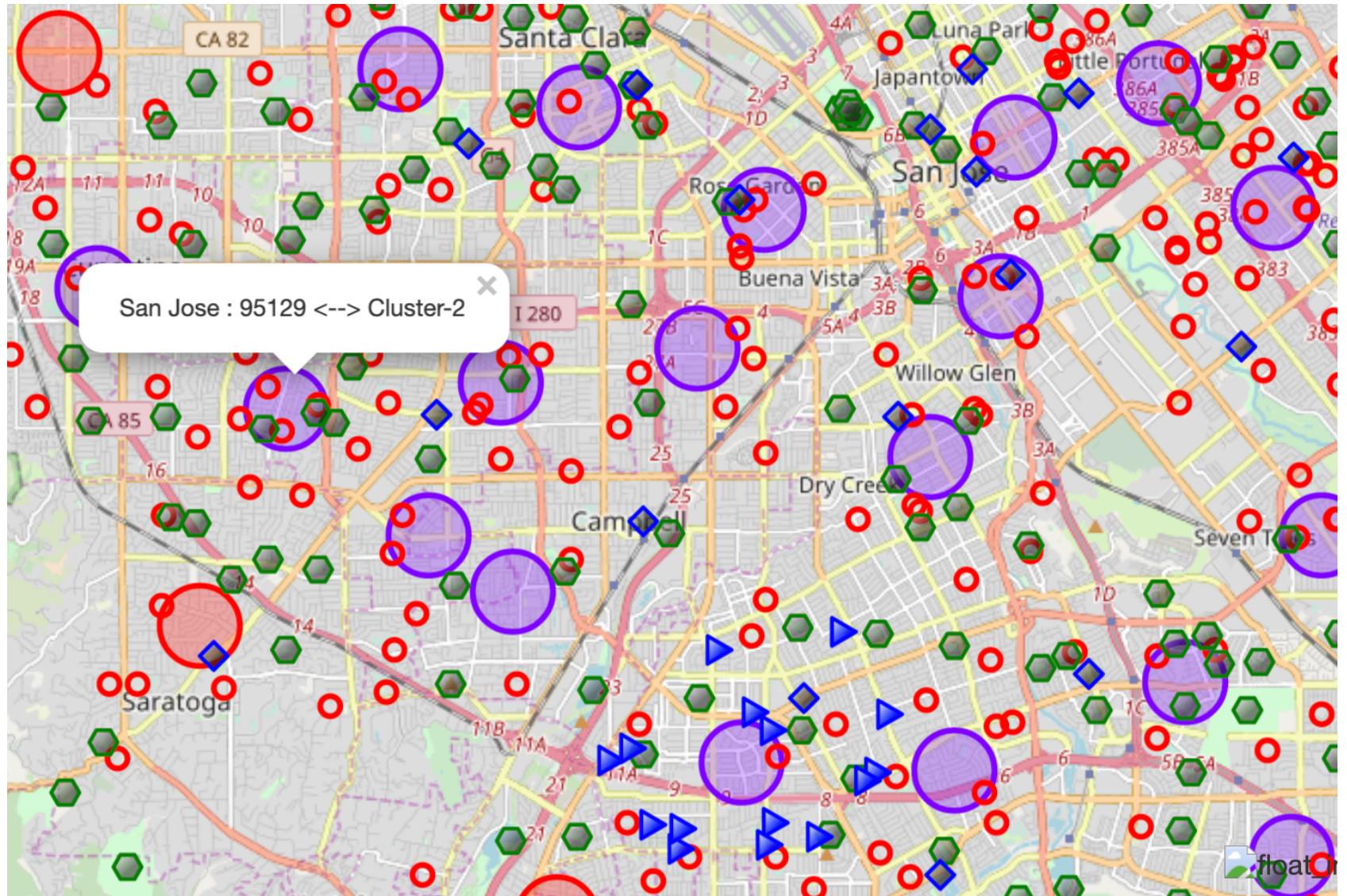
# Recent Sold Houses in 95129/95130/95118/95124



# 4.0 Results Section

# Consolidated Map of Santa Clara county with schools, parks, libraries and cluster of cities

- School --- Red Circle
- Park --- Green Hexagon
- Library --- Blue Square
- Cluster of city --- Large Circle  
(same cluster = same color)
- Comparable Houses in 95124 ---  
- Blue Triangle



## 4.1 Problem Resolution:

The amenities in the selected neighborhood shall be similar to his current residence apartment

### Solution:

My client lives in 95129 which is cluster-2 (label starts from 0) on the map. Therefore, the candidate cities/zipcodes should come from the same cluster. The number of candidate zipcodes is going down from 53 to 37 (Number will be different based on clustering result).

## 4.2 Problem Resolution: The price of house for sales is under 1.5M

### Solution:

There are only two cities remained (Campbell and San Jose) if we consider the affordable prices and the location (Gilroy and Morgan Hill are too far away from my client's current resident 95129). The number of candidate zipcodes goes down to 21 (Number will be different based on clustering result)

## 4.3 Problem Resolution: Not too close to the railroad (at least 3 miles away)

### Solution:

When we consider at least 3 miles away from railroad, the candidate zipcodes will be reduced to 9 (Number will be different based on clustering result). That is a lot easier when making the final decision

4.4 Problem Resolution: 1) Near the school (within 0.5 mile), 2)  
The schools in the area should have high rating (Ranking  
greater than 7)

**Solution:**

My client can use the average school rating to select the location of his potential home.

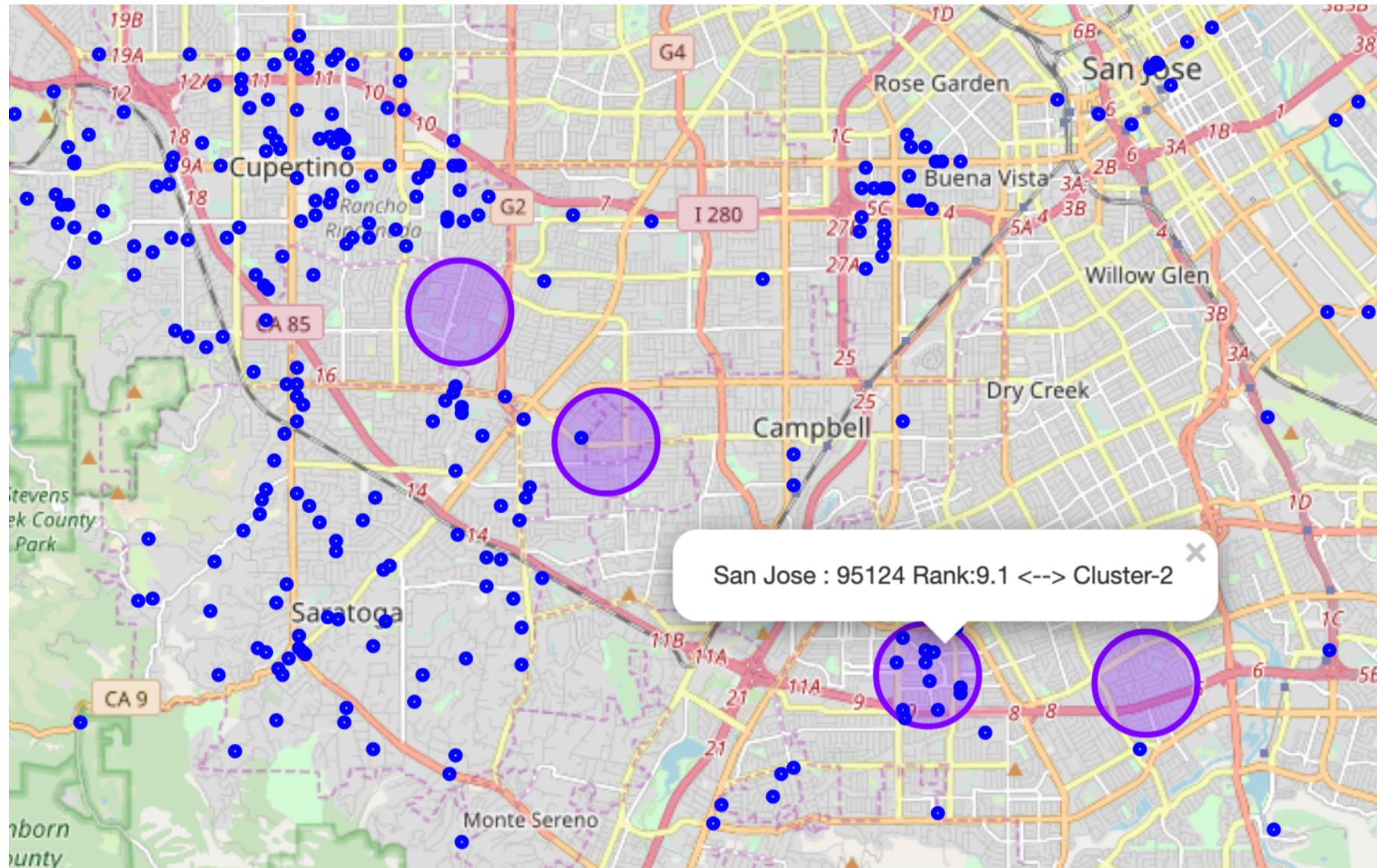
```
The Average School Rating of 95008: 6.0
The Average School Rating of 95117: 6.0
The Average School Rating of 95118: 7.6
The Average School Rating of 95124: 9.1
The Average School Rating of 95127: 4.1
The Average School Rating of 95129: 9.7
The Average School Rating of 95130: 9.8
The Average School Rating of 95131: 7.0
The Average School Rating of 95133: 5.7
```

4.5 Problem Resolution: The neighborhood/community should be safe and have low crime rate

**Solution:**

We display the crime locations and the final candidate zipcodes on the same map

# Crime/Theft locations and final candidate zipcodes (95129, 95130, 95124, 95118)



## 4.6 Problem Resolution:

- 1) Near the park (within 0.5 mile)
- 2) Near the library (within 1 mile))

### Solution:

My client can use the location data (latitude and longitude) of the library when making buying decision. The parks do not have zip code associated with them. However, we can use their latitude and longitude to know the distances from the candidate houses.

## 4.7 Problem Resolution:

House must be at least 3 bedrooms, 2 bathrooms, 1 car garage, around 1800 to 2100 square footage of size

### Solution:

Follow section 3.10 Recent Sold House Price Data, we can obtain the recent sold houses information. This information can then be used when negotiating the deal with the seller.

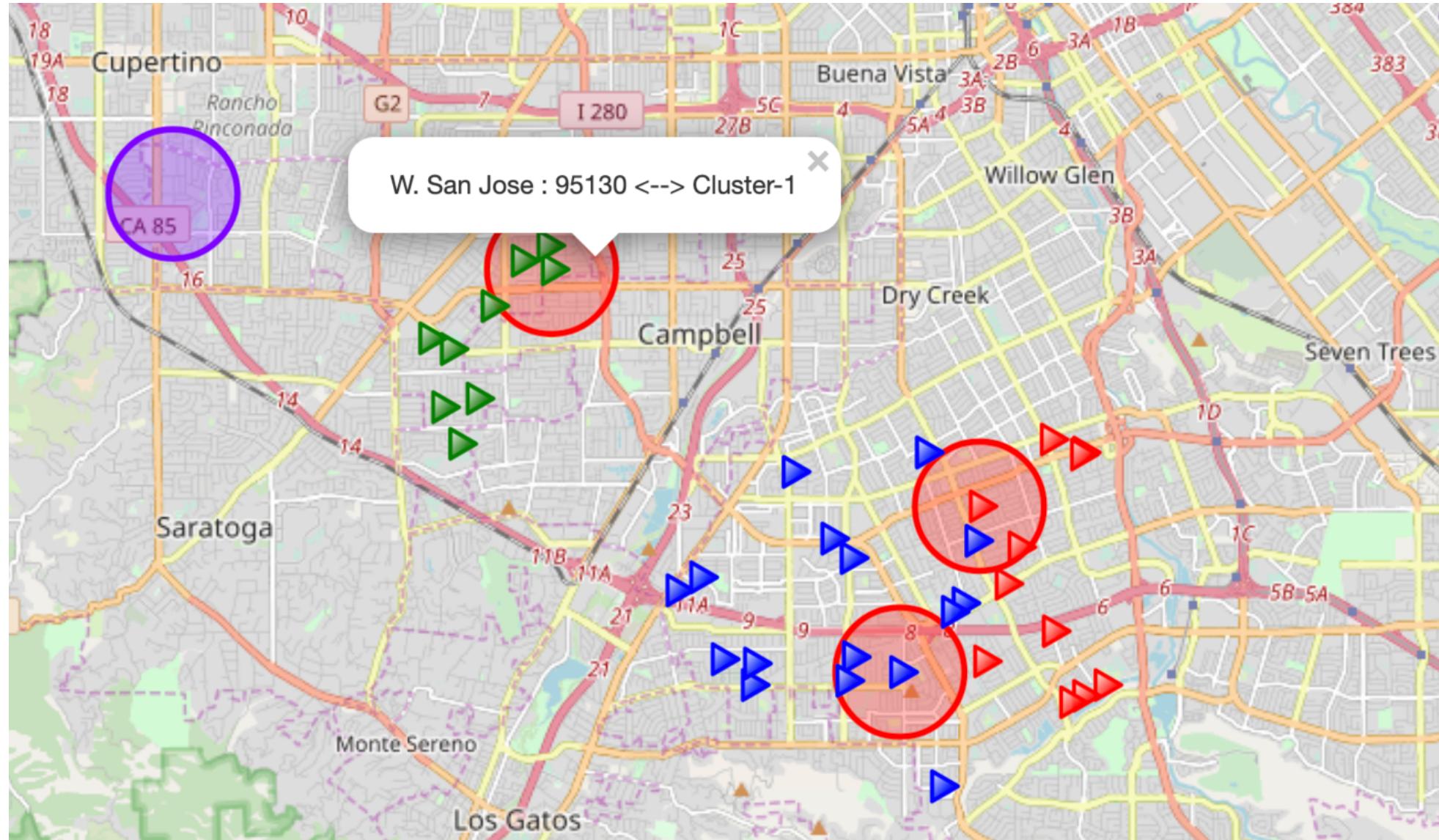
## 4.8 Problem Resolution:

- 1) The amenities in the selected neighborhood shall be similar to his current residence apartment
- 2) The location is near the supermarket (within 0.5 mile radius)
- 3) The location is near the shopping mall (within 3 miles radius),
- 4) The location is close (within 1 mile) to venues such as restaurants (Asian and Mexican foods ...etc.), parks and coffee shops ...

## Solution:

Refer to section 3.9 Explore Cities in Santa Clara County, we can run the Foursquare on three final dream home candidates (simulated by picking from recent sold houses; one for each zipcode; 95130, 95118, 95124) and current living apartment in 95129. The result will be able to address above requirements.

# Recent Sold Houses in 95130 (Green), 95118 (Red), 95124 (Blue)



## 4.9 House Selection:

The visualization maps and data analysis above will enable my clients to explore more options for his dream home hunting journey.

After examining all the alternatives, my client decides to buy a house at **95124** zip code area. Following is the summary ratings of each zip code:

Zip Code	Price	Crime Rate	School Rank	Park	Library	Railroad	Venues Favors
95129	1.9M	Low	9.7	Good	Good	Good	Good
95130	1.7M	Low	9.8	Good	Good	Good	Good
95118	1.4M	Low	7.6	Good	Good	Good	Fair
95124	1.5M	Low	9.1	Good	Good	Good	Good

So, what should my client do next? He can either go to hire a Real Estate broker or surfing homes for sale websites as below links:

- \* <https://www.zillow.com/>
- \* <http://julianalee.com/real-estate/#homes-for-sale>

## 5.0 Discussion Section

- The visualization maps and data analysis in section 3 and 4 will enable my clients to explore more options on his dream home hunting journey. I know it is hard and somehow frustrated when decide to buy a house especially in the Silicon Valley. However, if you have no time and not really know what exactly you should be looking for, then this capstone project might help.
- By the way, I encounter the mod\_security issue when I used BeautifulSoup webscraping tools. Well, the real estate website hosting company is blocking some kind of requests to their servers (where your site is located). The whole mod\_security Apache module is disabled for my site. There are no other ways around it and I have to manually handle some house sales data.
- The Zillow API is quite useful. However, the most recent sales information may not be accessed via API. I can understand that Zillow may want to protect their valuable data. But, I may not use that API again if someone provide me more open one.
- The kclusters (number of clusters) selection is tricky when using KMeans for the clustering analysis. The search radius (default is 500m) is also essential if you don't have a rough idea about the city. Sometimes, the result looks weird so you have to be patient and trying different combinations.
- The Foursquare API has a limit of 950 Regular API Calls per day and 50 Premium API Calls per day for Sandbox Tier Accounts.
- In general, I am quite impressed with the overall organization, content and lab works presented during the Coursera IBM Certification Course. I feel this Capstone project presented me a great opportunity to practice and apply the Data Science tools and methodologies learned. I will present the result of my study to my client. Of course, I would like to see this can help him to find his dream home in the coming months.

## 6.0 Conclusion Section

- This project has shown me a practical application to resolve a real situation that gives me the intention to use Data Science tools.
- The Folium is a powerful tool for data visualization which makes the analysis and decision making a lot easier. The FourSquare API is by all means a super tool to obtain all kinds of data based on the location. The ZillowAPI is good as long as you are not accessing to their protected data. The BeautifulSoup is terrific since webscraping is somehow exciting and it helps you to find your own treasure. I enjoy using all those tools.