Coursera Capstone Week 5

1. BUSINESS PROBLEM

Buying a house is an extremely dauting task. You want to first consider where your house is located. This type of research considers different both economic and environmental factors, such as the proximity to hospitals, restaurants, and parks, to name a few. Once you have found a desirable location, you must find a house that matches your desired expectations. This process is lengthy and difficult. This capstone project aims to help customers narrow their search by providing a tool that allows them to compare and contrast four different locations, based on their surrounding environmental factors. This will aim to minimize the difficulties presented in the first step of buying a house in four locations in the Santa Clara County – San Jose, Sunnyvale, Mountain View, and Santa Clara.

2. DATA

a. Foursquare API data

By using the latitude and longitude coordinates of a particular location, the Foursquare Places API is able to provide a list of venues associated to the location. In order to obtain a list of venues within a specified area, we use the *explore* endpoint from the API via an HTTP request, which provides us with a JSON object. This JSON object gives details on the *location* (coordinate of each venue) and the different types of *categories*. This provides valuable information regarding various types of venues (e.g. entertainment, dining, etc.) surrounding a particular location.

b. Neighborhood Geographic

OpenCage Geocoder (opencagedata.com) provides an API to convert coordinates to and from places. This allows easy access to get the longitude and latitude details of the four locations.

c. Wikipedia

Wikipedia allows me to retrieve information regarding the population of several cities in the Santa Clara County:
https://en.wikipedia.org/wiki/Santa_Clara_County,_California

d. Real Estate History

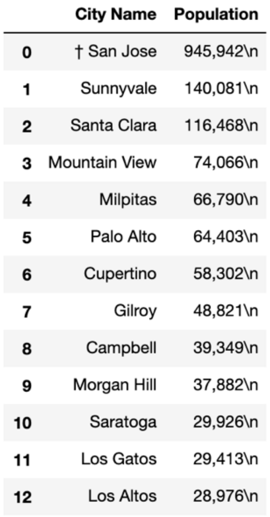
Real estate details, such as the average price of a house in a location and the housing market increase or decrease from year to year, can be collected from Zillow (Zillow.com) to make more informed decisions in creating this tool.

e. Zip Code Information

To retrieve further insights into each zip code, a special python database named uszipcode (pypi.org/project/uszipcode/) was used.

3. METHODOLOGY

Prior to beginning my exploratory analysis of the data, I imported the necessary python libraries and credentials that were necessary in collecting the data. More specifically, I listed the credentials from the various API websites I was using to gain access to them. Additionally, since it is very time consuming to analyze all the cities in the bay area, I chose a particular county, Santa Clara County, and created a short list of the most popular cities in this county by retrieving information from a Wikipedia page. After web scraping this data, I chose to further limit the scope of this project to the top four cities with the largest populations: San Jose, Sunnyvale, Santa Clara, and Mountain View, seen in Table 1. Finally, I retrieved the latitude and longitude points for each city, as these were the key to exploring FourSquare's information of venues located in and around each city.



	City Name	Population
0	† San Jose	945,942\n
1	Sunnyvale	140,081\n
2	Santa Clara	116,468\n
3	Mountain View	74,066\n
4	Milpitas	66,790\n
5	Palo Alto	64,403\n
6	Cupertino	58,302\n
7	Gilroy	48,821\n
8	Campbell	39,349\n
9	Morgan Hill	37,882\n
10	Saratoga	29,926\n
11	Los Gatos	29,413\n
12	Los Altos	28,976\n

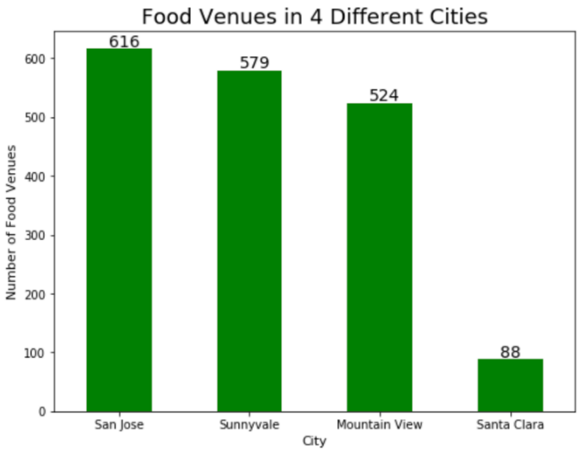
Table 1

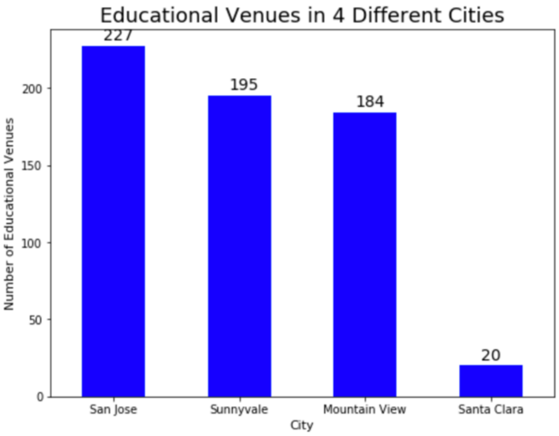
a. Exploratory Analysis

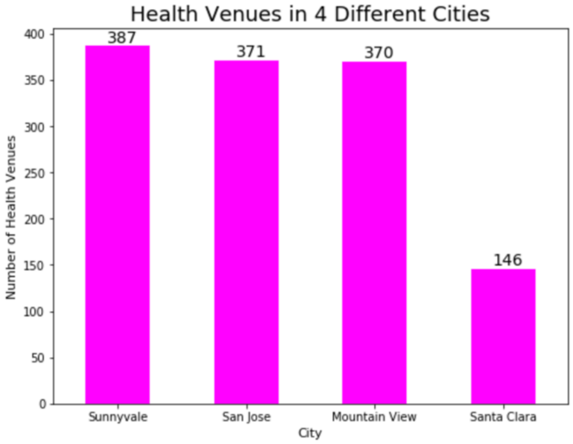
In order to gain a better understanding of the venues located in each city, I conducted an exploratory analysis for each city. This is a huge factor for those shopping for a house and I based this analysis on the following five main factors:

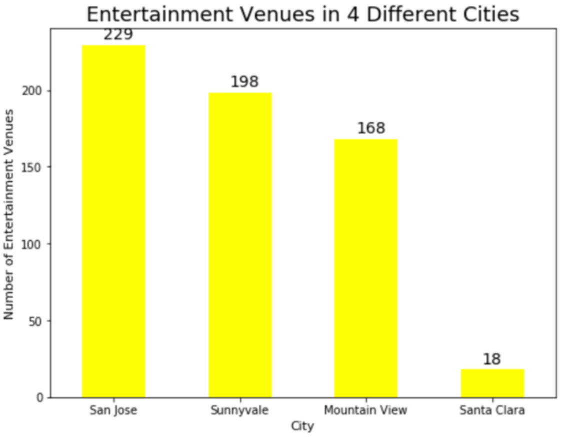
- Food (e.g. restaurants, grocery stores, etc.)
- Education (e.g. elementary schools, high schools, etc.)
- Health (e.g. hospitals, pharmacies, etc.)
- Entertainment (e.g. bars, theaters, etc.)
- Average prices of a house

These factors were established based on the factors I would consider when buying a house and they are listed in order of importance for me. For each category of venue, I sent a get request to the FourSquare API to retrieve the total number of venues that exist for that particular category. For example, the total number of venues for the food category consisted of the number of restaurants, grocery stores, and coffee shops in that city. Then, I plotted an individual bar graph for each factor listing the cities from highest to lowest number of venues. This is shown below.

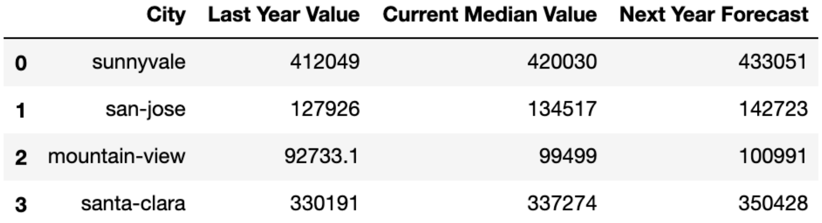








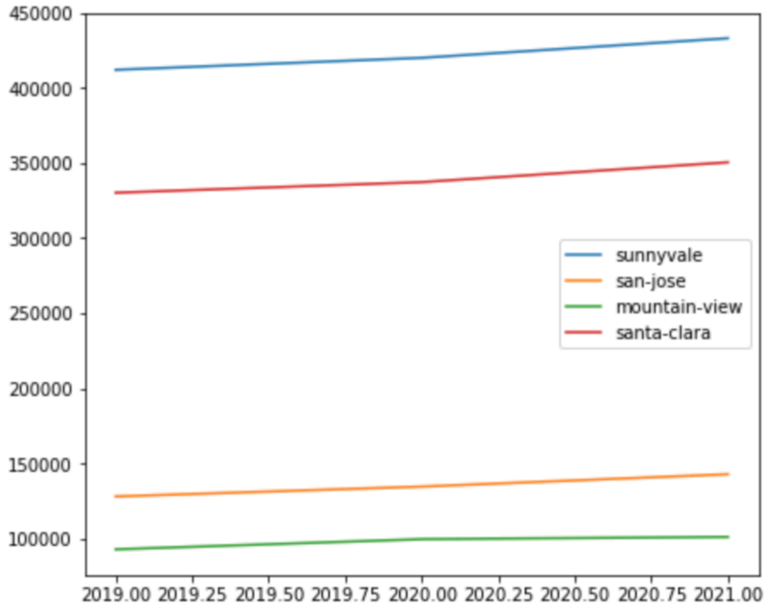
After discovering analyzing the category venues for each city, it is in the best interest to see the monetary value of housing prices in each city. In order to do this, I sought the average price of the house in the past year (2019), the current average price of the house (2020), and the predicted price of the house in the upcoming year (2021). I gained this information by using Zillow and I created a table displaying these values:



	City	Last Year Value	Current Median Value	Next Year Forecast
0	sunnyvale	412049	420030	433051
1	san-jose	127926	134517	142723
2	mountain-view	92733.1	99499	100991
3	santa-clara	330191	337274	350428

Table 2

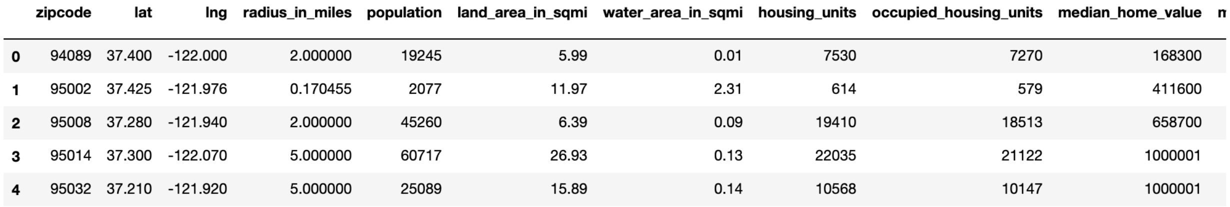
Then, I plotted the different values of the median prices of the houses for each city on using a line graph:



Based on the above analysis of the five different factors, San Jose has the most to offer in terms of venues. Thus, I focus the remaining exploration on the city of San Jose.

b. Statistical Analysis and Machine Learnings Used

I'm going to do further statistical analysis to investigate the different neighborhoods in San Jose to decide where is the best location to look for a house in San Jose. In order to do this, I first retrieve all the zip codes in the San Jose area by using the 'city-data' website. Then, I get additional information regarding each zip code by using the 'uszipcode' library provided by python. For example, the new information collected includes details like the housing unit, the median home value, and the median house price. This gathered data is held in a data frame so that it can be easily accessed (seen in Table 3).



	zipcode	lat	lng	radius_in_miles	population	land_area_in_sqmi	water_area_in_sqmi	housing_units	occupied_housing_units	median_home_value	n
0	94089	37.400	-122.000	2.000000	19245	5.99	0.01	7530	7270	168300	
1	95002	37.425	-121.976	0.170455	2077	11.97	2.31	614	579	411600	
2	95008	37.280	-121.940	2.000000	45260	6.39	0.09	19410	18513	658700	
3	95014	37.300	-122.070	5.000000	60717	26.93	0.13	22035	21122	1000001	
4	95032	37.210	-121.920	5.000000	25089	15.89	0.14	10568	10147	1000001	

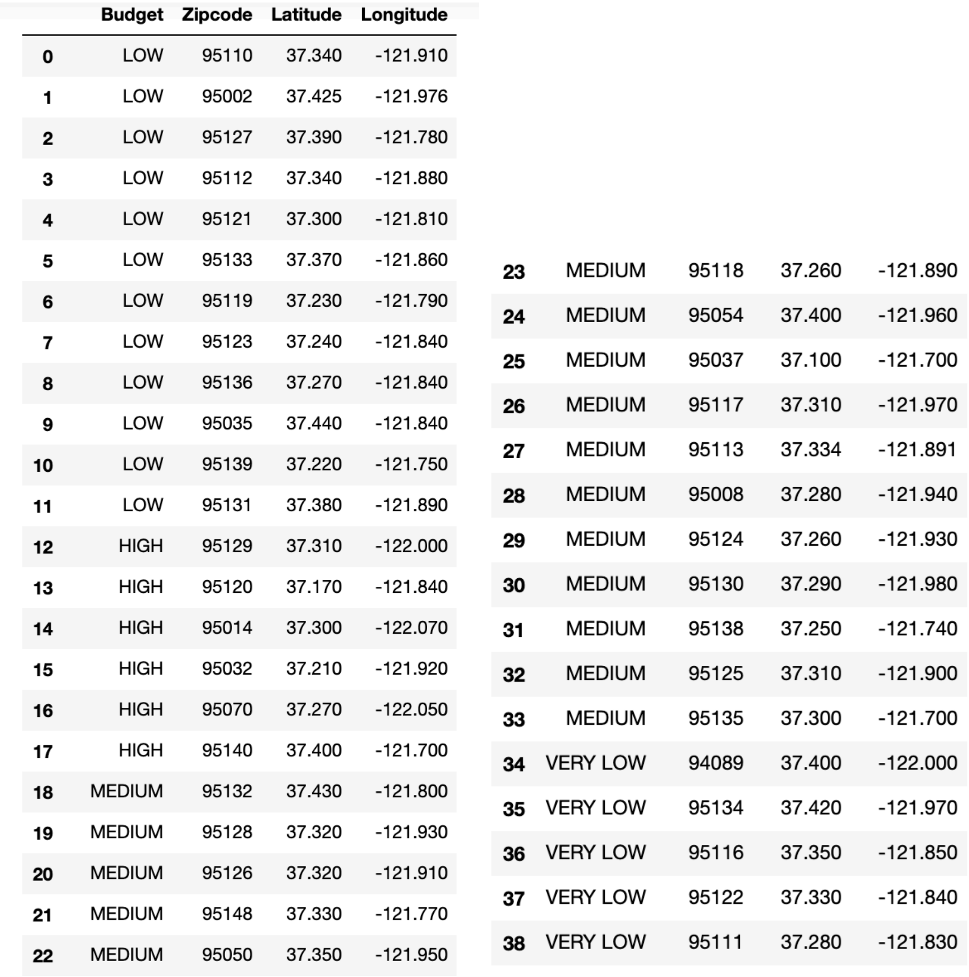
Table 3

Now that the information is in a data frame, I am able to apply a commonly used machine learning algorithm: Clustering. This algorithm allows me to see how different clusters split up and helps gain better insight about each cluster. After setting the cluster size to 4 and running this on the data frame, I get the San Jose data frame split into four separate clusters. Now, I can analyze each cluster individually, paying special attention to the median price of a house in each cluster:

- **Cluster 1**: 386,200 – 518,200 - Low

- **Cluster 2**: 862,100 – 1,000,001 – High
- **Cluster 3**: 568,000 – 758,300 - Medium
- **Cluster 4**: 168,300 – 334,300 - Very Low

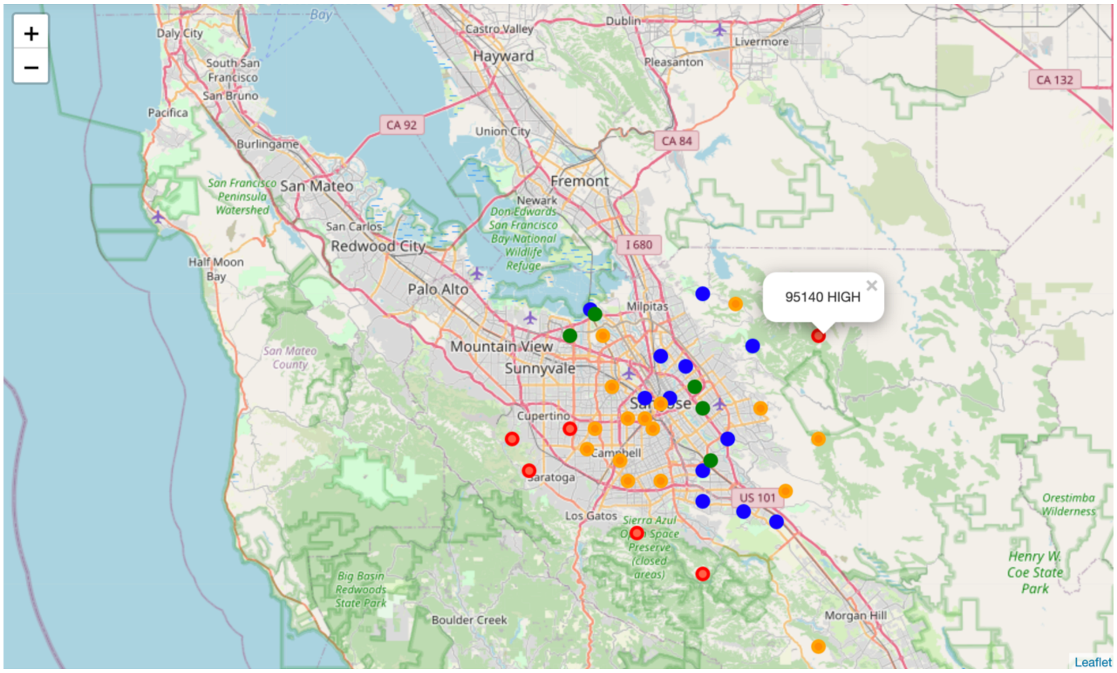
As seen above, I can clearly see that each cluster gave us a bucket range of the median home value. each cluster shows a minimum range and a maximum range for the median price of a house by zip code. To make this data more usable, I added a budget column that showed each zipcode in the San Jose area and its relative budget category based on the budget categories I provided above. This would further help house buyers to analyze what houses they want to look at based on their price range:



	Budget	Zipcode	Latitude	Longitude
0	LOW	95110	37.340	-121.910
1	LOW	95002	37.425	-121.976
2	LOW	95127	37.390	-121.780
3	LOW	95112	37.340	-121.880
4	LOW	95121	37.300	-121.810
5	LOW	95133	37.370	-121.860
6	LOW	95119	37.230	-121.790
7	LOW	95123	37.240	-121.840
8	LOW	95136	37.270	-121.840
9	LOW	95035	37.440	-121.840
10	LOW	95139	37.220	-121.750
11	LOW	95131	37.380	-121.890
12	HIGH	95129	37.310	-122.000
13	HIGH	95120	37.170	-121.840
14	HIGH	95014	37.300	-122.070
15	HIGH	95032	37.210	-121.920
16	HIGH	95070	37.270	-122.050
17	HIGH	95140	37.400	-121.700
18	MEDIUM	95132	37.430	-121.800
19	MEDIUM	95128	37.320	-121.930
20	MEDIUM	95126	37.320	-121.910
21	MEDIUM	95148	37.330	-121.770
22	MEDIUM	95050	37.350	-121.950
23	MEDIUM	95118	37.260	-121.890
24	MEDIUM	95054	37.400	-121.960
25	MEDIUM	95037	37.100	-121.700
26	MEDIUM	95117	37.310	-121.970
27	MEDIUM	95113	37.334	-121.891
28	MEDIUM	95008	37.280	-121.940
29	MEDIUM	95124	37.260	-121.930
30	MEDIUM	95130	37.290	-121.980
31	MEDIUM	95138	37.250	-121.740
32	MEDIUM	95125	37.310	-121.900
33	MEDIUM	95135	37.300	-121.700
34	VERY LOW	94089	37.400	-122.000
35	VERY LOW	95134	37.420	-121.970
36	VERY LOW	95116	37.350	-121.850
37	VERY LOW	95122	37.330	-121.840
38	VERY LOW	95111	37.280	-121.830

Table 4

Finally, a map of San Jose is plotted which displays the different budget areas.



4. RESULTS & DISCUSSION

Through my analysis of the four cities in the Santa Clara County, I found that San Jose has the largest population size and is the most versatile city in terms of offering. Across the first four category of venues, it had the highest number of venues for the food venues (616), educational venues (227), and entertainment venues (229). Sunnyvale had a marginally bigger number of health venues, 387 venues, to San Jose’s 371 health venues. By looking at the housing market, I also saw that San Jose was reasonably affordable compared to the other cities I was looking at, which makes this city a good fit for a first house.

By looking at the clusters, I saw that San Jose offers four different categories of housing prices. I denoted these as ‘VERY LOW’, ‘LOW’, ‘MEDIUM’, and ‘HIGH’. By looking at Table 4, a house buyer can make better decisions on where to look for a home in San Jose by using their budget costs. The map also provides an easy way to look at this clustering information.

5. CONCLUSION

In order to make buying houses easier, I went through an exploratory and statistical analysis of a small test case using the popular cities in Santa Clara County: San Jose, Sunnyvale, Mountain View, and Santa Clara. This research showed me that San Jose had the most to offer when evaluated across the five factors pre-determined. By using the clustering method, I was able to provide even more insight on the different median prices of the houses found in San Jose and provided helpful information for those who have a particular budget in mind.

37.336191	-121.890583