Moving to Silicon Valley? Explore the Area by Data Science Methods

1. Introduction

1.1. Description and Discussion of the Background

Silicon Valley is a region in southern part of San Francisco Bay Area in Northern California that is considered as a global center for high technology and innovation. Most of Silicon Valley is in Santa Clara County, although it slightly extends to the neighboring counties such as San Mateo and Alameda Counties. The focus of this project is to explore Santa Clara County, the center point of Silicon Valley.

Many of the most well-known technology companies such as Google, Apple, Microsoft, Intel, IBM, Facebook, eBay, Cisco, Qualcomm, Texas Instruments, Netflix, and Oracle reside in Santa Clara County. Economic prosperity, world class universities, cultural diversity, a support system for entrepreneurs, and a rich engineering tradition attracts many tech talents to the region. The region's economy is heavily service based. Although technology, both hardware and software, dominates the service sector by value, Santa Clara County has its share of retail and office support workers.

In this project, I used GitHub repository as a database. I utilized geopy to convert addresses into latitude and longitude values. I used Foursquare API to explore cities in Santa Clara County specifically to retrieve the most common venue categories in each city and then used this feature to group cities into clusters. This task was completed using k-means clustering algorithm. I also used python Folium library to visualize cities and their respective clusters. Finally, I used open datasets to analyze demographics, economics, and other social aspects in Santa Clara County.

The main audience of this project are people who come to live and work in Santa Clara County. People consider many factors when thinking about a new destination to call home. Such factors include amenities such as popular venues in each city, demographics, education, housing, and school rating. This project uses a data science approach to explore these factors. In addition to new residents, home buyers, investors, policy makers or anybody interested in learning about Santa Clara County can benefit from this project.

1.2. Datasets and Toolkits

1.2.1. Prerequisite

This project requires a Foursquare developer account. To create an account, go to https://developer.foursquare.com/.

1.2.2. Datasets

Table 1 lists data sources that are used in this project along with their description.

Table 1 List of data sources that are used in this project.

Data Sources	Description
Spatial Data Repository of UC Berkley	JSON file of city boundaries
Foursquare API	location service API to explore venues
Santa Clara County Public Health Department	Dataset on demographics and education
Open Data Network	Dataset on population count and population density
Bayareamarketreports.com	Dataset on housing prices
Greatschools.org	Dataset on school ratings
RentCafe.com	Dataset on average rent

1.2.3. Python Libraries

Table 2 shows a list of Python libraries used in this project.

Table 2 List of python libraries used in this project.

Python Libraries	
numpy	Library to handle data in vectorized manner
pandas	Library for data analysis
json	Library to handle JSON files
geopy	Python client for geocoding web services
xlrd	Library for reading data and formatting information from Excel files
requests	Library to handle requests
matplotlib	Library for creating visualizations
scipy	Library for scientific computing
seaborn	Library for statistical data visualization
plotly	Library for scientific graphing
sklearn	Library for machine learning
folium	Library for map rendering

2. Methodology

2.1. Clustering Cities in Santa Clara County

Santa Clara County has a total of 15 cities. To segment the cities and explore them, we need a dataset that contains the name of the cities and their respective latitude and longitude coordinates. I used a GeoJSON file from [UC Berkeley Library GeoData]() to obtain the list of cities in Santa Clara County and I used python geopy library to get the latitude and longitude values for each city. The obtained data was then transformed into a panda data frame as shown in Table 3.

Table 3 Head of dataframe showing the coordinates of cities in Santa Clara County.

	City	Latitude	Longitude
0	MILPITAS	37.428272	-121.906624
1	GILROY	37.006508	-121.563172
2	MORGAN HILL	37.130408	-121.654497
3	MONTE SERENO	37.236333	-121.992458
4	SARATOGA	37.263832	-122.023015

Using python folium library, a map of Santa Clara County was created with cities superimposed on top (Figure 1).

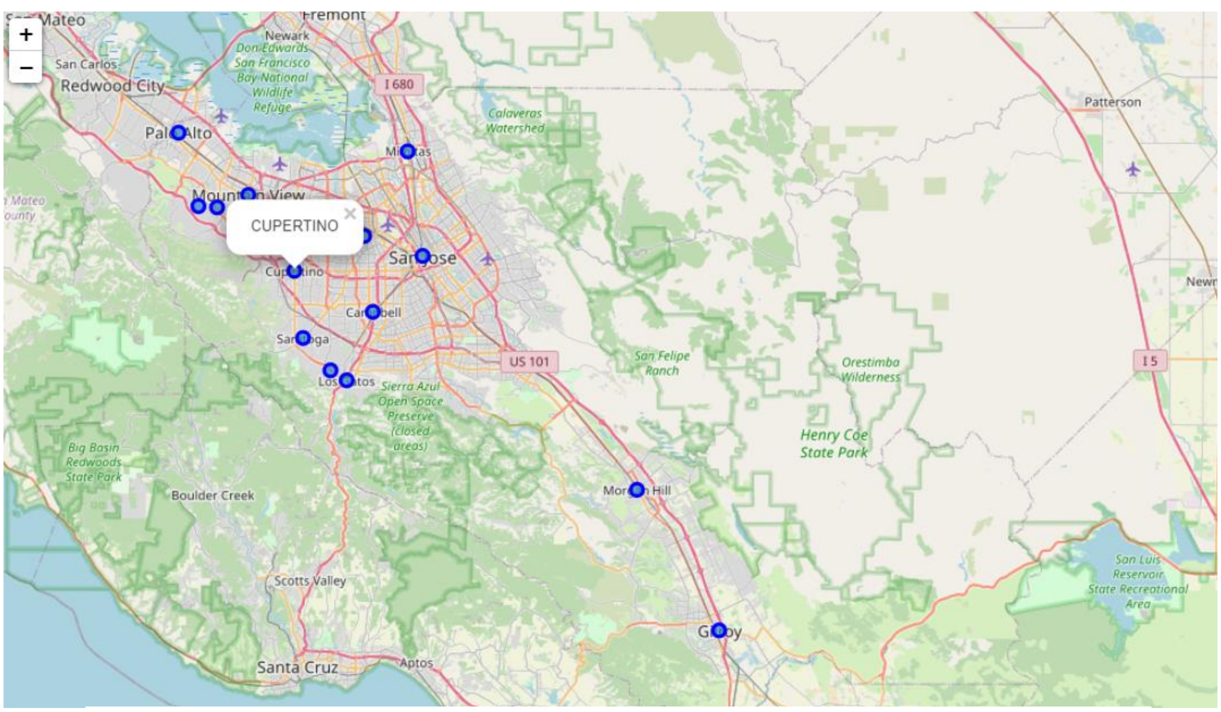

Figure 1 Map of Santa Clara County with location of cities superimposed on top.

Next, I used Foursquare API to explore the cities and segment them. I constructed a URL and sent requests to API to explore each city. I set the limit to 100 venues and the radius to 1600 meters (~ 1 mile). I cleaned the retrieved JSON file and structured it into a panda data frame as displayed in Table 4.

Table 4 Head of dataframe showing the returned venues by Foursquare API.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	MILPITAS	37.428272	-121.906624	Sea Link Cafe	37.427921	-121.906359	Café
1	MILPITAS	37.428272	-121.906624	Fosters Freeze	37.427821	-121.907548	Burger Joint
2	MILPITAS	37.428272	-121.906624	Com Tam Thien Huong	37.428375	-121.907346	Asian Restaurant
3	MILPITAS	37.428272	-121.906624	Anh Hong Saigon	37.428103	-121.911465	Vietnamese Restaurant
4	MILPITAS	37.428272	-121.906624	Black Bear Diner	37.428430	-121.909569	Diner

Figure 2 shows the total number of venues returned by Foursquare API for each city. As it can be observed, eight cities have reached the 100 limits. On the other hand, Foursquare has returned less than 50 venues for four cities including only 5 venues for Los Altos Hills. It should be noted that in this project the Foursquare inquiry was limited to a single latitude-longitude pair for each city. In other words, the inquiry investigated the venues within a 1-mile radius for each given latitude-longitude pair. A more thorough inquiry can include multiple latitude-longitude pairs for each city to obtain more venue information.

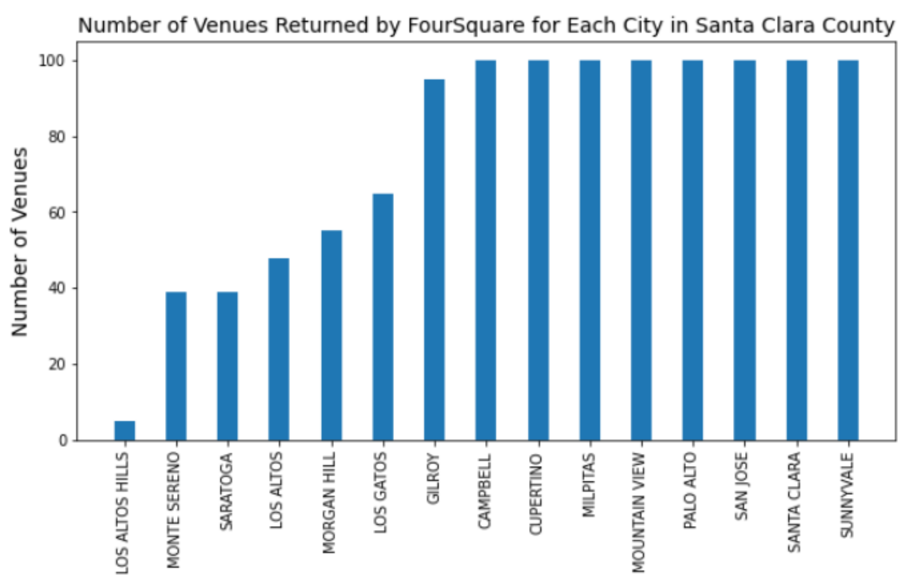

Figure 2 Total number of returned venues by Foursquare API for each city.

In total, 220 unique categories were returned by Foursquare API. The top 5 venue category for each city is displayed in the Table 5.

Table 5 Head of dataframe showing cities along with their top five common venues.

	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	CAMPBELL	Mexican Restaurant	Pizza Place	Coffee Shop	Italian Restaurant	Sandwich Place
1	CUPERTINO	Japanese Restaurant	Bakery	Coffee Shop	Chinese Restaurant	Park
2	GILROY	Mexican Restaurant	Furniture / Home Store	Fast Food Restaurant	Coffee Shop	Sandwich Place
3	LOS ALTOS	Pizza Place	American Restaurant	Park	Mexican Restaurant	Coffee Shop
4	LOS ALTOS HILLS	Park	Pool	Home Service	Soccer Field	Yoga Studio

The returned venues by Foursquare are more understandable if we consider demographics of Santa Clara County (demographics analysis is presented in section 2.2). For example, Asian is considered the majority ethnic group in Cupertino. Thereby, it is understandable that the most common venue category in Cupertino to be Asian Restaurants. On the other hand, Latino is the majority ethnic group in Gilroy. Expectedly, the first most common venue category for Gilroy is Mexican restaurants.

As it can be seen in Table 5, there are some common venue categories in the cities. To further examine this, I used unsupervised learning K-means algorithm to cluster the cities. As shown in Figure 3, analyzing K-Means with elbow method suggests an optimum k = 4 for the K-Means.

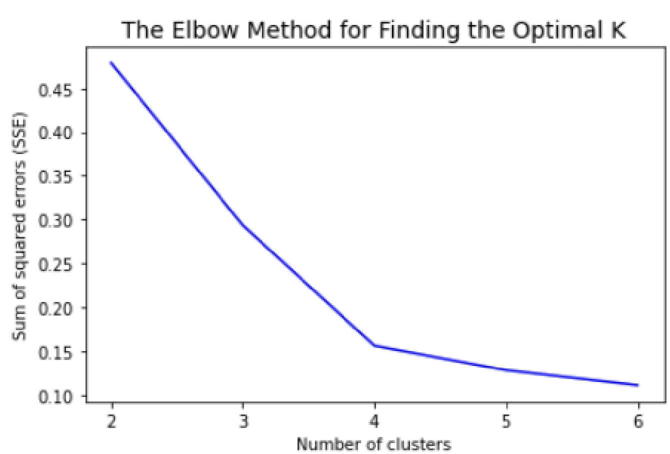

Figure 3 The elbow method for finding the optimal number of clusters.

I performed K-Means to cluster the cities into 4 clusters. The cities in each cluster are similar to each other in terms of the features included in the dataset. The merged data frame in Table 6 shows the cluster as well as the top 5 venues for each city.

Table 6 Head of merged dataframe showing the cluster as well as the top five common venues.

	City	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	MILPITAS	37.428272	-121.906624	0	Chinese Restaurant	Indian Restaurant	Korean Restaurant	Mexican Restaurant	Sandwich Place
1	GILROY	37.006508	-121.563172	0	Mexican Restaurant	Furniture / Home Store	Fast Food Restaurant	Coffee Shop	Sandwich Place
2	MORGAN HILL	37.130408	-121.654497	2	Italian Restaurant	Pizza Place	Brewery	Vietnamese Restaurant	Convenience Store
3	MONTE SERENO	37.236333	-121.992458	2	Pizza Place	Mexican Restaurant	Restaurant	Pet Store	Bakery
4	SARATOGA	37.263832	-122.023015	3	American Restaurant	Italian Restaurant\t	Coffee Shop	Burger Joint	Café

To label each cluster, a bar chart was created showing the number of **1st Most Common Venue** in each cluster (Figure 4).

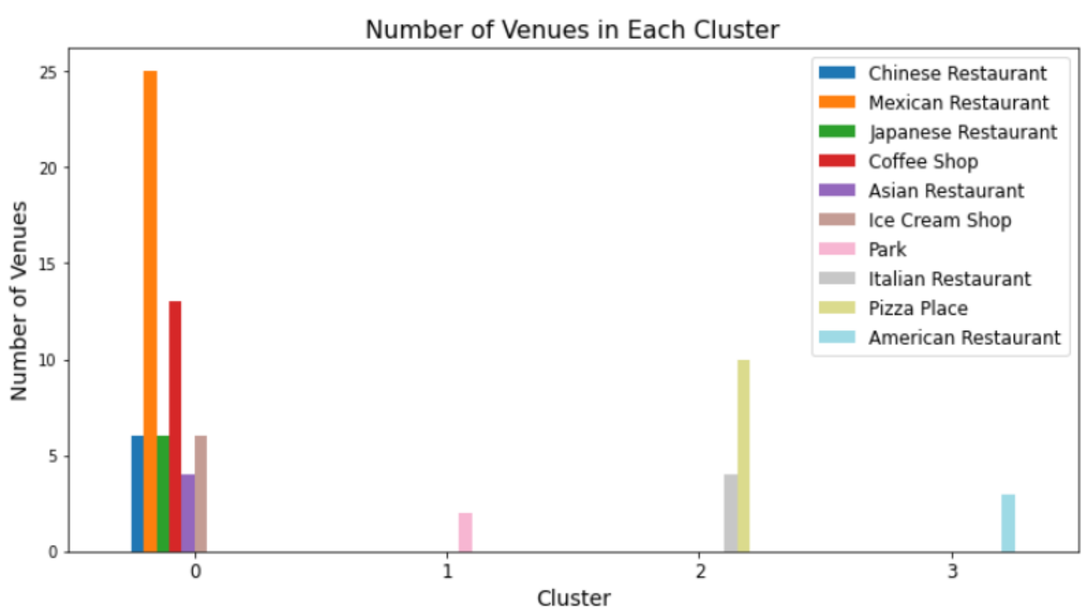

Figure 4 Number of 1st most common venues in each cluster.

By examining the graph, we can label clusters as shown in Table 7.

Table 7 List of clusters and their labels.

Cluster	Cluster Label
Cluster 0	"Various Social Venues Including Intensive Mexican & Asian Restaurants"
Cluster 1	"Parks"
Cluster 2	"Pizza/Italian Restaurants"
Cluster 3	"American Restaurants"

Finally, the resulting clusters were visualized on an interactive map as shown in Figure 5.

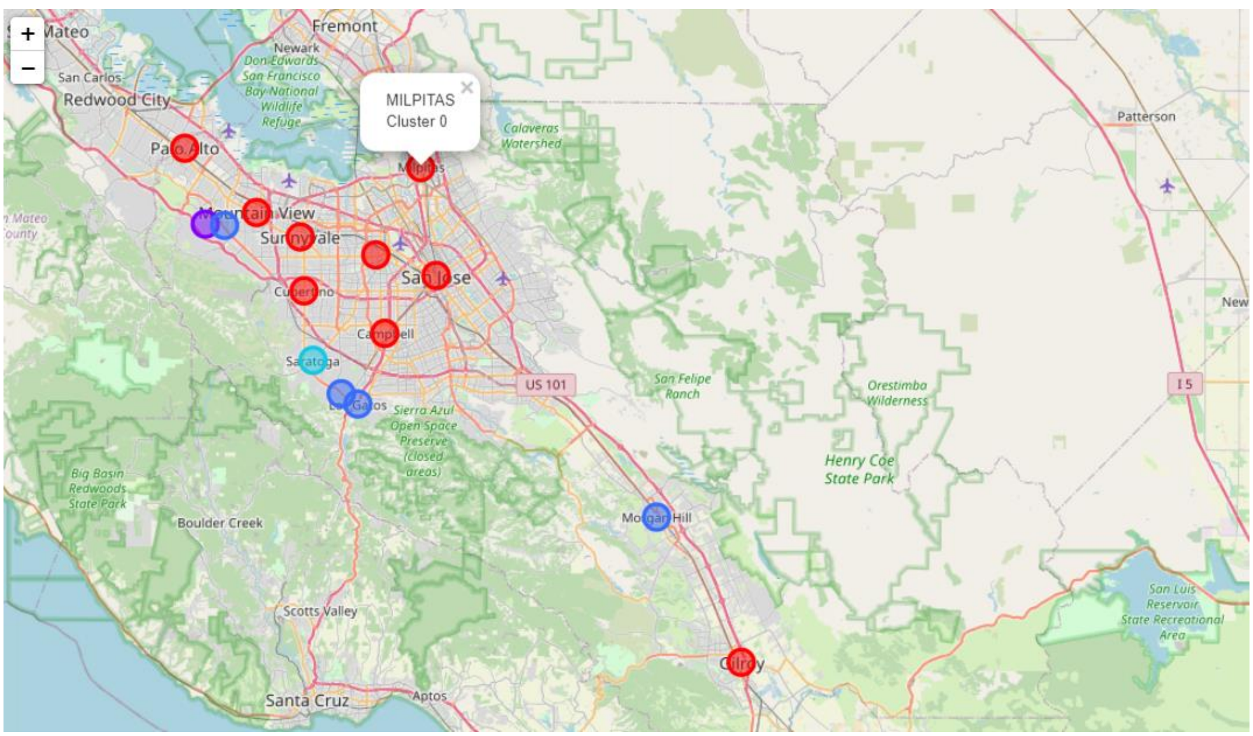

Figure 5 Visualizing the resulting clusters on Folium map.

2.2. Demographics

To investigate demographics of Santa Clara County, data on Open Data Network was scraped and processed into a panda data frame. The data frame was then visualized using python matplotlib library. Figure 6 displays population count and population density (per square mile) of cities in Santa Clara County.

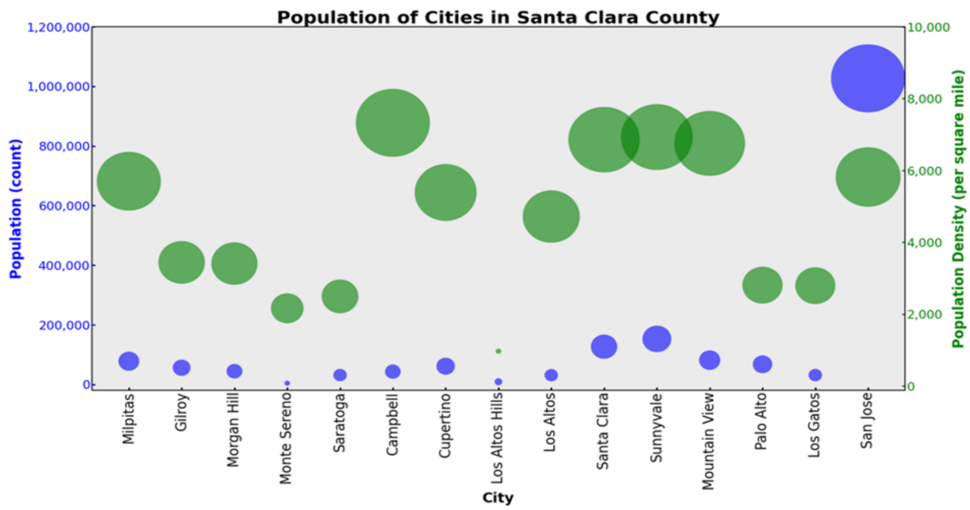

Figure 6 Population size and population density in cities of Santa Clara County.

It can be observed that cities such as Campbell and Mountain View have a relatively small population size with respect to city of San Jose. However, the population density of both cities is slightly higher than population density of San Jose. On the other hand, cities of Monte Sereno and Los Altos Hills both have a small population count and a small population density. This is important as a higher population density is usually linked with a higher probability of services such as hospitals.

Race/ethnicity is an important factor for understanding the composition of the county and comparing the diversity of cities. Understanding the racial/ethnic composition is especially important for city managers and policy makers who target programs to meet the needs of the residents. It is also important for people who want to live in an area more suited to their needs in terms of ethnic food or local events. To investigate race/ethnicity, data from Santa Clara County Public Health Department was imported and structured into a panda dataframe. Figure 7 presents the percentage of population in five racial/ethnic groups (African American, Asian/Pacific Islander, Latino, White, and Other) for each city in Santa Clara County.

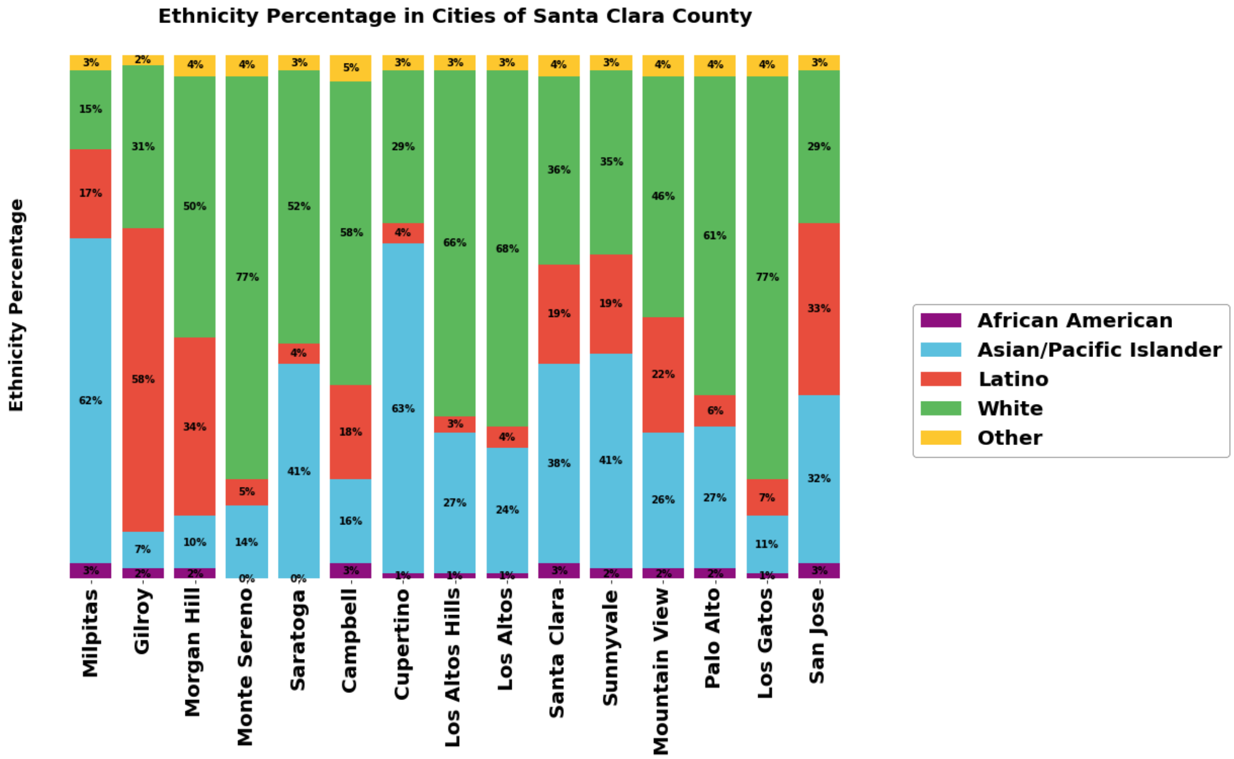

Figure 7 Ethnic/racial percentage in cities of Santa Clara County.

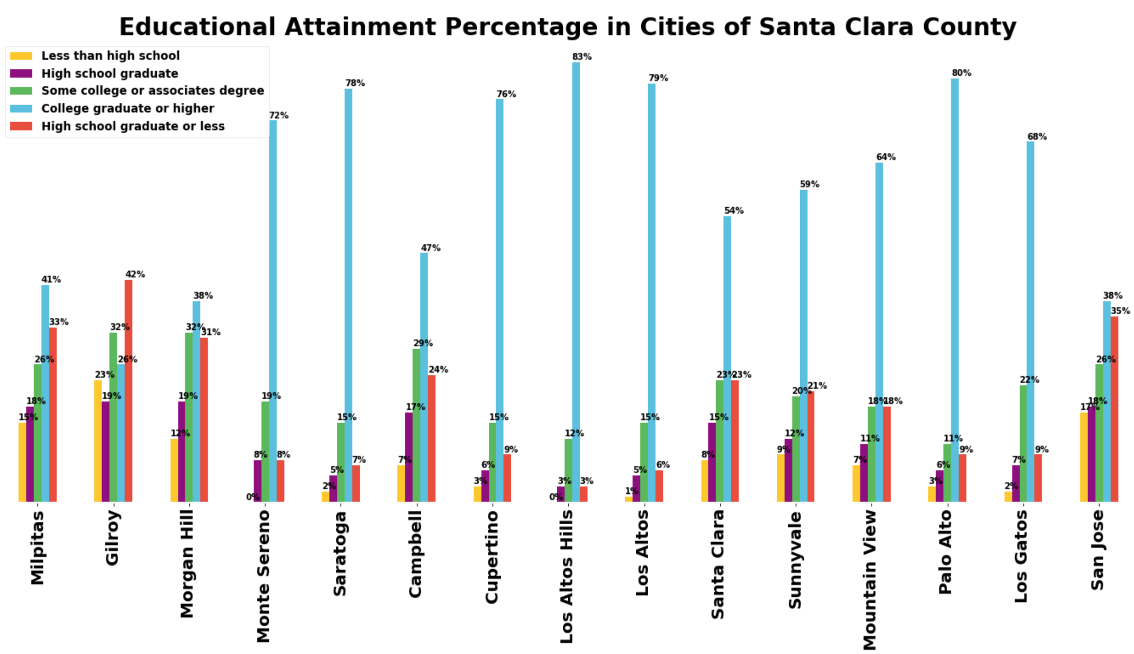

Figure 8 Educational attainment percentage in cities of Santa Clara County.

2.3. Education and Economics

2.3.1. Education

Education is linked to higher employment rate, income levels and overall quality of life. To investigate education attainment, data from Santa Clara County Public Health Department was processed and analyzed. Figure 8 shows percentage of population ages 25 or older with educational attainment in Santa Clara County. As shown, Los Altos Hills and Gilroy have the highest and lowest percentage of college graduates, respectively.

2.3.2. Economics

Housing is one of the most important factors in choosing where to live once people decide on a destination. Households that face high housing costs may have reduced financial resources for their other needs. High housing cost may also force people to move frequently or reside in areas with poorer quality housing and higher crime rates. To investigate housing costs, data regarding the average rent in Santa Clara County in year 2020 was scraped from RENTCafe and analyzed. Figure 9 shows a choropleth map with the average rent information for Santa Clara County cities. As it can be observed, the most affordable cities in Santa Clara County are Gilroy and Morgan Hill. On the other hand, the most expensive cities are Mountain View, Palo Alto, and Cupertino.

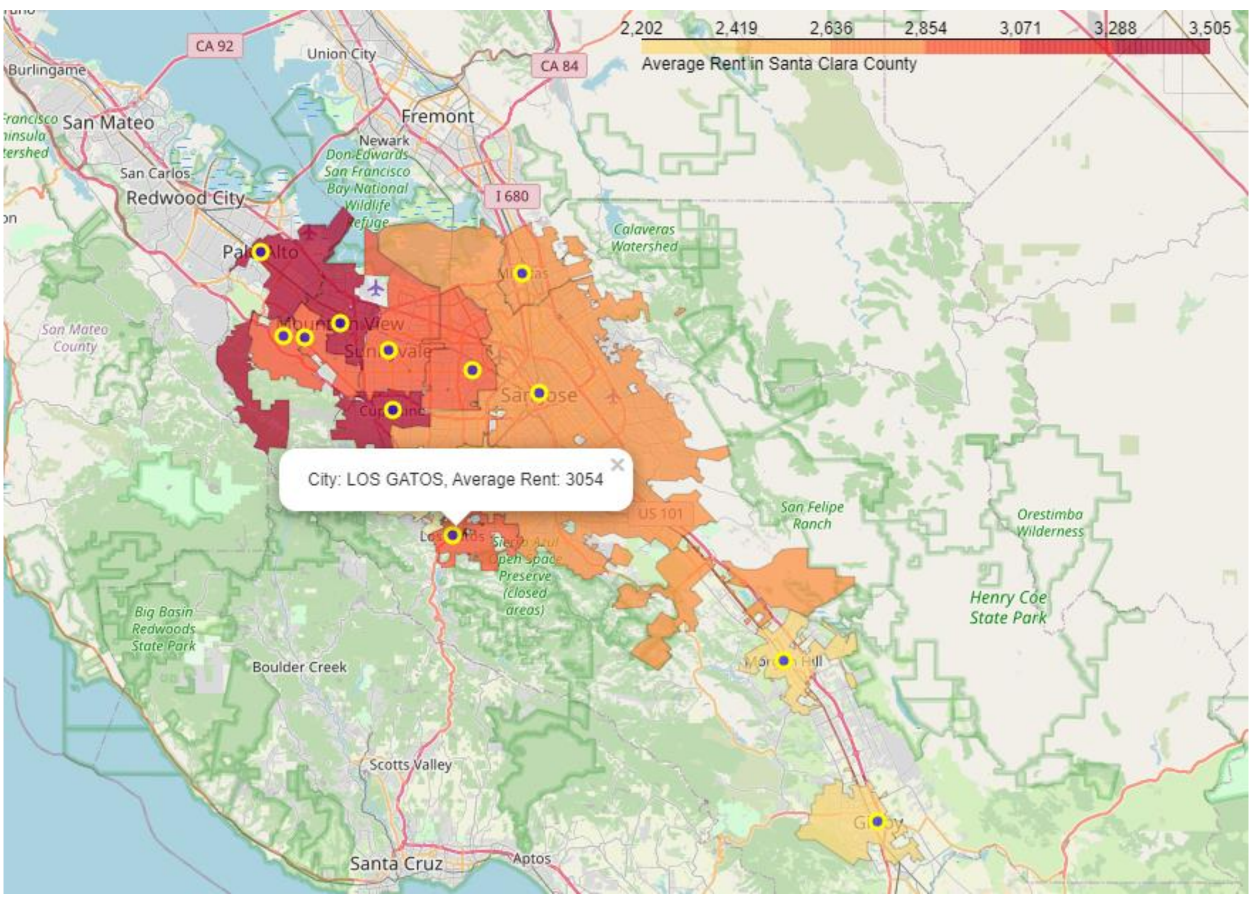

Figure 9 Choropleth map depicting average rent in cities of Santa Clara County.

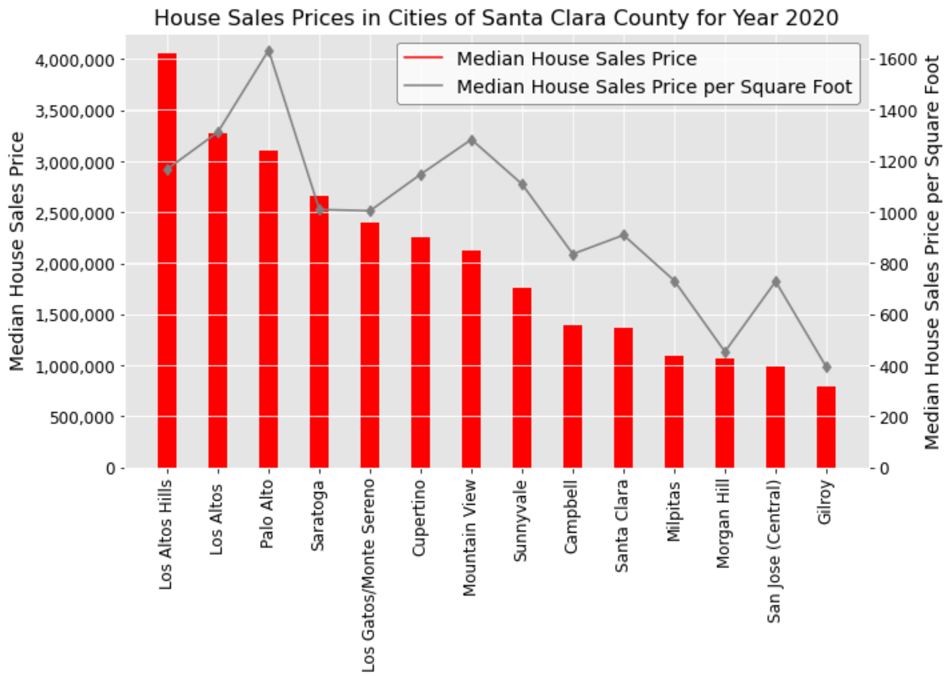

Figure 10 Median house sales price and median house sales price per square foot in year 2020..

To further investigate the housing cost, I obtained the dataset on median house sales prices in Santa Clara County from Bay Area Market Reports and analyzed it. Figure 10 depicts the median house sales prices in cities of Santa Clara County in 2020. As it can be observed, there is a wide range in house prices in the county where Los Altos Hills and Gilroy have the highest and lowest median house sales prices, respectively.

2.4. School Rating

School choices is a top criterion for people who have school-aged children or planning on starting a family in the future. To investigate the schools in the area, the data on greatschools.org was scraped and processed. The choropleth map shown in Figure 11 compares the average school rating across different cities in Santa Clara County.

Figure 12 shows a scatter plot of school rating and median house sales price in Santa Clara County. As it can be seen, there is a positive direct correlation between the two variables with p-value << 0.001, a strong evidence that correlation is significant. This shows that school rating is a good predictor of house price in Santa Clara County.

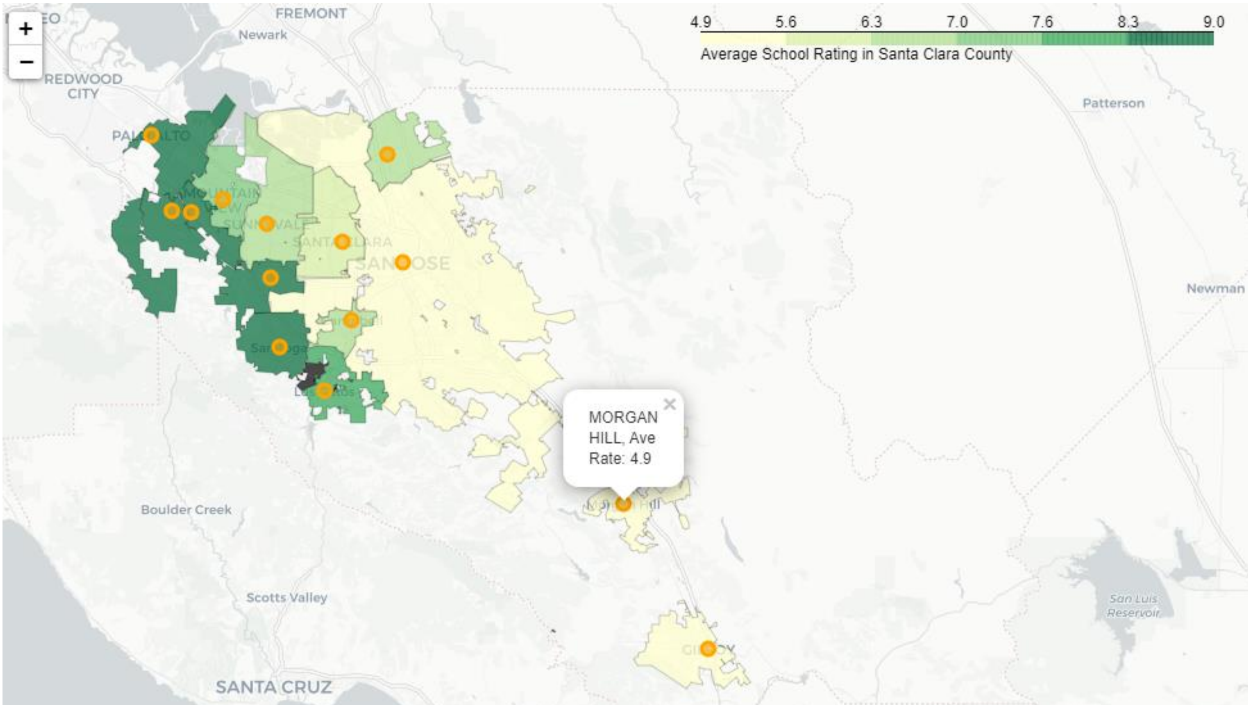

Figure 11 Choropleth map showing average school rating in cities of Santa Clara County.

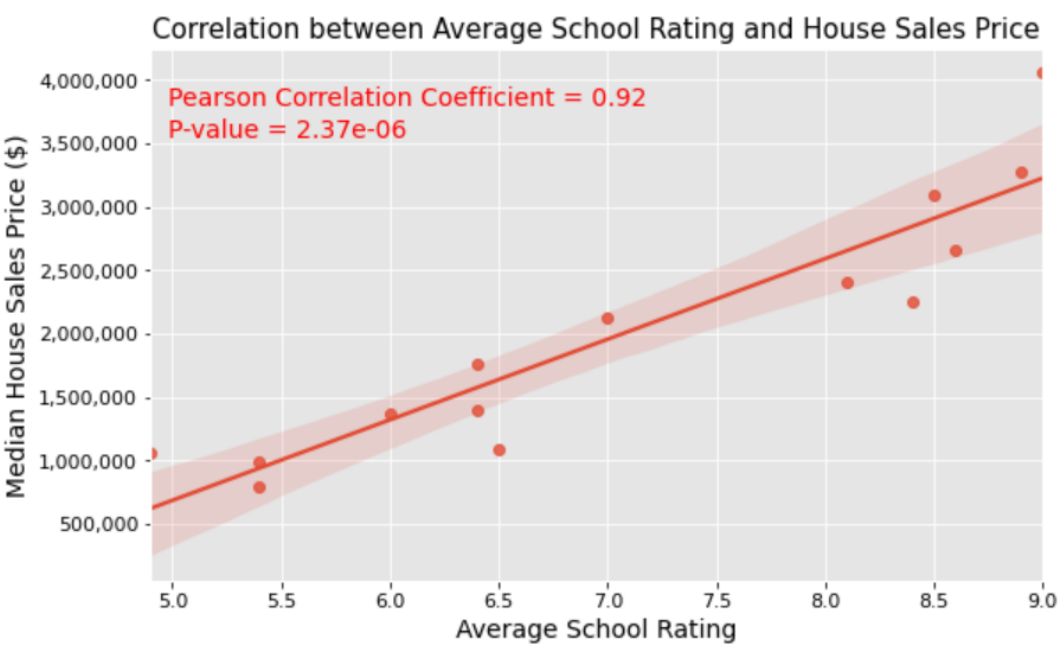

Figure 12 Correlation between average school rating and median house sales price in Santa Clara County.

3. Discussion

Foursquare API was used to retrieve the most common venues in each city of Santa Clara County. Then, using K-Means algorithm, cities were clustered into four groups. The clustering results were in agreement with the analyzed demographics of each city. It should be noted that this project explored one pair of latitude-longitude coordinates for each city. To increase the clustering accuracy, the coordinate dataset can be expanded to explore multiple coordinates for each city.

Population size and population density of each city were investigated. The diversity of cities in terms of ethnic groups were also analyzed. Educational attainment percentage and cost of living in terms of average rent price and median house sales price were investigated. These are all important information for new residents and other stakeholders to consider when choosing an optimal location based on their interests and needs. It should be noted that the economic analysis presented here was performed using static data. For future studies, these data can be accessed dynamically from specific platforms or packages.

4. Conclusion

Home to Silicon Valley, Santa Clara County attracts many people each year to the region. Newcomers can achieve better outcomes if they can have access to platforms introducing them with different aspects of living in Silicon Valley. Such platforms which can be obtained through data science methods are valuable not only for the new residents, but also for future investors and city officials.

5. References

- Wikipedia __ Silicon Valley
- Foursquare API
- UC Berkeley Geodata
- Open Data Network
- Santa Clara County Public Health Department
- RENTCafe
- Bay Area Market Reports
- Greatschools.org