**Silicon Valley Rent Segmentation**

**Introduction**

The Silicon Valley is one of the world's hub for technology and innovation. From long time companies like Intel and Hewlett-Packard, social media giants like Google, Facebook and Twitter to a plethora of startups - the Silicon Valley attracts high-skilled workers from all over the world.

When there is an influx of workforce into a region, the cost of living begins to increase. The Silicon Valley is currently going through a 'housing crisis' - with housing prices skyrocketing, affordability is a major issue. For someone moving to the Silicon Valley for a new job, buying a home might far from reach. The other option is finding a place to rent.

**Business Problem**

In this study, we will be looking at how median rent for a two-bedroom apartment varies within each zip code of the Santa Clara County. For those who are new and moving to the area, it is best to know where the house rents are relatively low but also closer to social venues like restaurants, theatres, malls, parks, etc. This analysis plans to address this issue.

**Data Requirement and Collection**

For this study I chose to divide Santa Clara county by zip code (since neighborhood were too small in size and cities with too big in size). Once I have the relevant zip codes for the respective regions in the county, I need to merge the data with the median rent for a two-bedroom apartment. Further using Geocoder, find the coordinates for each zip code region and form this into a data set. Finally, I will need data about the different venues in each region to help cluster similar regions. I will use Foursquare API to gather this information.

Once the data has been collected and processed, one should be able to view of a map with similar venue categories overlaid over a choropleth showing the median rent of that region.

For example, you can look for regions with Mexican restaurants and gyms in the neighborhood and compare what the median rent in those regions are (this is a hypothetical example prior to running the code).

**The following data must be collected:**

**1. A list of zip codes within the Santa Clara County**

*I was able to scrap data from Santa Clara County Public Health website:* [*https://www.sccgov.org/sites/phd/hi/hd/Pages/city-profiles.aspx*](https://www.sccgov.org/sites/phd/hi/hd/Pages/city-profiles.aspx)*. This provided a list of cities and neighborhoods within Santa Clara County along with zip codes.*

**2. Median rent for Santa Clara County by Zip code**

*This data was scrapped from a real estate search website:* [*https://www.zillow.com/santa-clara-county-ca/home-values/*](https://www.zillow.com/santa-clara-county-ca/home-values/)*. This website provides the most up to date data on housing prices and rents. I was able to scrap the data for 2019 median rent for each zip code in the county.*

**3. Geo Locator for geographical coordinates of the county**

*Importing Geocoder library to get the coordinates for Santa Clara County*

**4. List of venues in each neighborhood**

*A list of venues in each neighborhood will be pulled from Foursquare API*

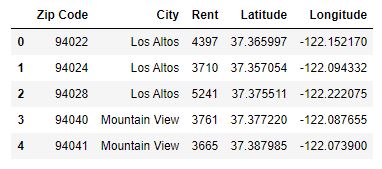
**5. Map generation**

*Using Folium to generate maps in Python. In order to plot a choropleth a json file of Santa Clara county was obtained from* [*https://geo.nyu.edu/catalog/stanford-qb712rn3742*](https://geo.nyu.edu/catalog/stanford-qb712rn3742)

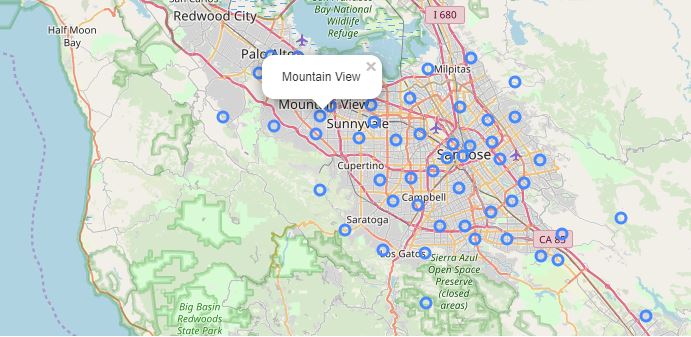
**Methodology**

First, I cleaned all the data obtained from the Santa Clara County Public Health as well as from Zillow.com in order to create a data frame with *Zip Code, City, Rent* data. In the cleaning process, I dropped the comma and dollar sign to convert the Rent value to an integer.

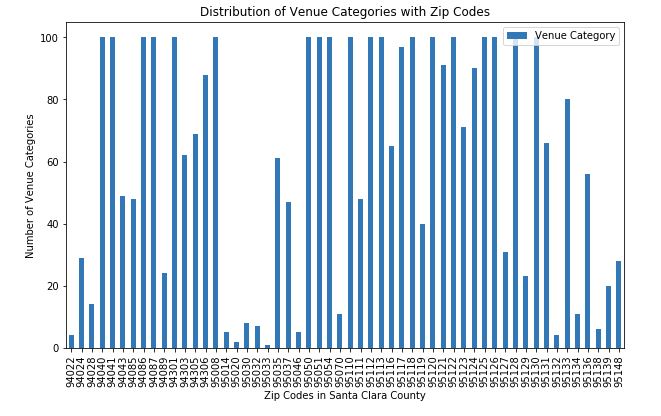
Then using the Geocoder Library, I obtained the geographical coordinates for each zip code and appended *Latitude and Longitude* data to the data frame as follows:



Using the Folium Library, I plotted each zip code, labelled by the city on the map of Santa Clara County. The data frame contained 55 unique zip codes as shown below:



Next, using the **Foursquare API in Python**, I ran a request for a maximum of 100 venues with a mile (1600 meters) of each zip code coordinate. I first ran a test to explore the first zip code (Los Altos, CA) and then expanding to obtain venues for all zip codes in the data frame. We see that areas like Los Gatos and Gilroy have much fewer venues than areas like Sunnyvale and San Jose.



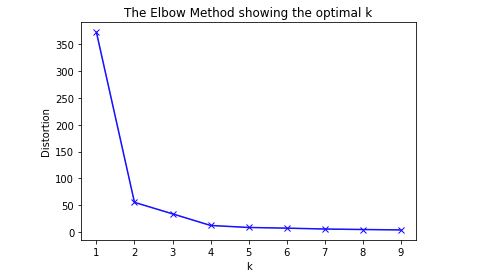
I was able to obtain 302 unique venue categories for this data set.

Note: The number of venues is tied to the specific coordinates for each zip code. Adjusting the coordinates might results in a different number for each region.

Further, I analyzed the data to check the occurrence of each category for every zip code region. This was done by creating a data set with each category as columns and rows show the frequency of occurrence for every zip code. This enabled us to visualize the top 10 most common venue category in each zip code.



I then used **K-Means Clustering** to group the zip code based on similar venue categories. In order to obtain the optimal value for k (number of clusters) I used the **Elbow Method**. This indicated that the optimal value for **k is 2**, however, I needed that the data not be binary and so I picked the second optimal value for **k as 4**.

****

Up on analyzing the cluster, I divided them into 4 Segment:

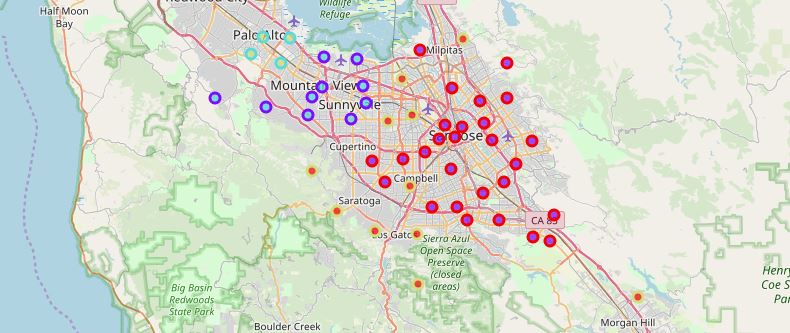
* *Cluster 0 - Restaurants, Coffee Shops and Stores*
* *Cluster 1 - Parks, Grocery Stores, Eateries*
* *Cluster 2 - Hotels and Social Venues*
* *Cluster 4 - Parks, Hills, Restaurants*

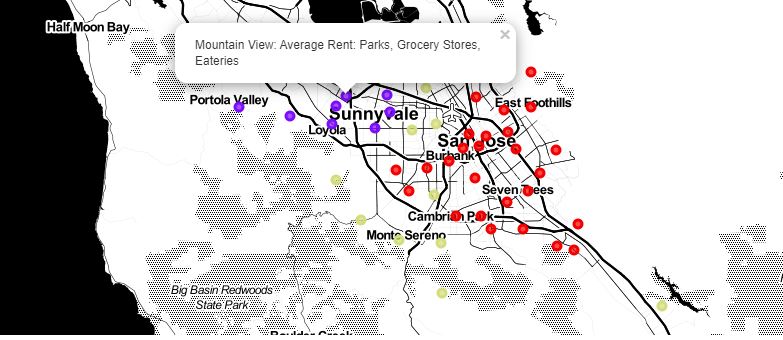
I also divided the Rent Price into three segments:

* *Low - $2,556 - $3,451*
* *Average - $3,452 - $4,346*
* *High - $4,346 - $5,241*

**Results**

The result from clustering was plotted on the map using Folium. Each label would provide information about the City, the Rent Level and the Venue Cluster



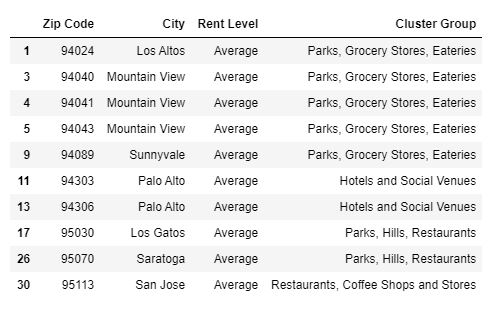
****

The clusters were also mapped along with the Rent Levels to visualize what Cluster each Zip Code and Rent Level fall into:

* *Rent Level: High*



* *Rent Level: Average*



* *Rent Level: Low*



**Discussion**

Although the median rent in the Santa Clara County is much higher than the National Median, we see that most zip codes cluster within the ‘Low’ rent level. Only 4 zip codes fall in the ‘High’ rent level and these are in very affluent neighborhood that are surrounded by Parks and Social Venues.

Additional work can be done on this using neighborhood instead of zip codes to see the granularity of the clustering. Additionally, one other factor that could affect rents is neighborhood school rating, and this could be used to analyze the data.

**Conclusion**

When people arrive in the Silicon Valley, they are much focused on the job that brought them there. Resources, through data science, can be made available to those looking for the best place to live in their price point. This would be helpful in creating a more socio-economically diverse neighborhood.