Undervalued ZIP Codes of Silicon Valley

A Capstone Project for the

IBM Professional Certificate in Data Science

-

Alissa McFarlin

June 2021

Contents

Introduction

What are the "undervalued" ZIP Codes of Silicon Valley which represent potential opportunities for investment in apartment buildings?

The core (sub)urban area of Santa Clara County, California, better known as Silicon Valley, is famously home to many well-known technology companies such as Apple and Google. It is also famous for its ever more expensive housing market, making real estate investment an attractive venture for those who have the funds to do so.

This analysis aims to identify possible “undervalued” ZIP Codes within Silicon Valley which have the mix of venue categories typical of ZIP Codes with higher rents yet are still less expensive. These “undervalued” ZIP Codes represent neighborhoods potentially on the rise – with the potential for rising rents as well. Taken together, these ZIP Codes form a short-list worthy of further research by individuals and developers interested in investing in apartment building construction, purchase, or renovation in Silicon Valley as apartments in these areas may be well-positioned to increase in value.

Data

Data from several sources is used in this analysis.

ZIP Code Data

ZIP Codes are the best way to define areas within Silicon Valley for analysis. The valley is a single, continuous (sub)urban sprawl spanning several cities of various sizes, and unofficially-named neighborhoods within cities are poorly defined. ZIP Codes conveniently divide the sprawl into smaller units, with the added advantage of being the unit of choice for government-collected rent data. (More on that below.)

A list of ZIP Codes in Santa Clara County, California and the polygon data necessary to map their boundaries can be found in a GeoJSON file provided by the City of San Jose’s Enterprise GIS department. San Jose is the county seat, and this list includes ZIP Codes covering the entire county. More information can be found at https://data.sanjoseca.gov/dataset/zip-code-boundary1/resource/cde0aa77-a995-4457-b6d1-6449c0274bfa.

The data in a GeoJSON file has a nested key-value structure. The data for ZIP Code 95124, one of San Jose’s ZIP Codes, looks like this:

{ "type": "Feature", "properties": { "OBJECTID": 2457, "ZIPCODE": "95124", "LASTUPDATE": "2021-03-09T23:46:55Z", "SANJOSELIMITS": "Yes", "SHAPE_Length": 102896.59358156379, "SHAPE_Area": 183088849.70612666 }, "geometry": { "type": "Polygon", "coordinates": [[[-121.913409700293727, 37.286604500046032], [-121.913410000055109, 37.286715999832488], [-121.913420446820155, 37.286820316448093] …

For the purposes of this analysis, a list of all the ZIP Codes in the county was extracted from this file, and those lying largely outside of Santa Clara County were removed. The resulting ZIP Code list defined the initial scope of the analysis and was used to extract venue category and rent data.

Using the ZIP Code list to extract venue category data required one additional piece of information: the “center” of each ZIP Code. Many Silicon Valley ZIP Codes span (sub)urban areas and adjoining rural areas, making a true geographic center a poor choice for extracting venues typical of the (sub)urban area of concern. Two geocoders were used to produce ZIP Code center coordinates with unsatisfactory

results. Further research revealed that the ZIP Code Tabulation Areas (ZCTAs) used by the US Census have better-defined centers. The non-profit GeoNames provides a convenient tab-delimitated text file containing ZIP Code coordinates derived from Census data. The zip file is available at https://download.geonames.org/export/zip/.

95124 looks like:

US	95124	San Jose	California	CA	Santa Clara	085	37.2563	-121.9229	4

The third and second to last items are the latitude and longitude, respectively. Once the ZIP Code center coordinates were set, venue category data could be obtained.

Venue Category Data

Foursquare's API offers several endpoints to access its rich venue data. The explore endpoint takes a latitude, longitude, radius (in meters) and returns a response containing up to 100 recommended venues with their names, locations, and categories. For more information, see: https://developer.foursquare.com/docs/places-api/endpoints/.

The results returned from the API are in JSON format which, like the GeoJSON data above, consist of nested key-value pairs. The first result returned for 95124's latitude and longitude is:


```
{'reasons': {'count': 0,
  'items': [{'summary': 'This spot is popular',
    'type': 'general',
    'reasonName': 'globalInteractionReason'}]},
 'venue': {'id': '4e278ec2b61c9d28cbec4c33',
  'name': 'Pink Beauty & Spa',
  'location': {'address': '2977 Union Ave',
   'crossStreet': 'Foxworthy',
   'lat': 37.26138658007297,
   'lng': -121.9257221536735,
   'labeledLatLngs': [{'label': 'display',
     'lat': 37.26138658007297,
     'lng': -121.9257221536735}],
   'distance': 618,
   'postalCode': '95124',
   'cc': 'US',
   'city': 'San Jose',
   'state': 'CA',
   'country': 'United States',
   'formattedAddress': ['2977 Union Ave (Foxworthy)',
    'San Jose, CA 95124',
    'United States']},
  'categories': [{'id': '4bf58dd8d48988d1ed941735',
    'name': 'Spa',
    'pluralName': 'Spas',
    'shortName': 'Spa',
    'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/shops/spa_',
     'suffix': '.png'},
    'primary': True}],
  'photos': {'count': 0, 'groups': []}},
 'referralId': 'e-0-4e278ec2b61c9d28cbec4c33-0'}
```

Foursquare assigns a category to each venue, in this case, “Spa”. The venue category data is by nature categorical and therefore is well-suited to clustering analysis to identify groups of ZIP Codes with similar mixes of venue categories. The categories for all venues were extracted and the proportion of a ZIP Code’s venues which belong to each category calculated. The resulting category frequency matrix became the feature set used by clustering analysis to identify groups of similar ZIP Codes.

Rent Data

As this analysis explores the relationship between ZIP Code venue category mix and rent, a measure of rent was required. Many apartment rental sites offer data based on their own listings, but these may not be consistent from site to site. Fortunately, the US Department of Housing and Urban Development (HUD) small area Fair Market Rent (FMR) data provides a reliable and consistent measure of rent.

The FMR is set at the 40th percentile of rent distribution and is used to determine values for Section 8 housing vouchers. Small area indicates that this data is defined at the level of ZIP Codes. For more information see https://www.huduser.gov/portal/datasets/fmr/smallarea/index.html.

The data is available as an Excel spreadsheet. The row pertaining to 95124 looks like this:

95124	METRO41940M41940	"San Jose-Sunnyvale-Santa Clara, CA HUD Metro FMR Area”									
$2,250	$2,025	$2,475	$2,580	$2,322	$2,838	$3,080	$2,772	$3,388	$4,020	$3,618	$4,422
$4,640	$4,176	$5,104									

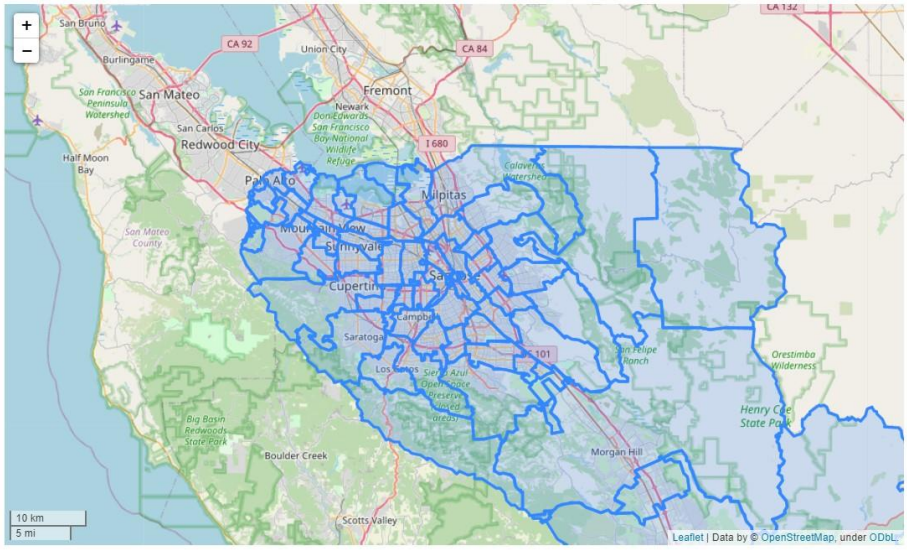
Note that FMRs are set separately for apartments of different sizes. This analysis used the one-bedroom apartment data, the fourth dollar amount listed. For 95124, it is $2,580.

If one wanted a realistic idea of rent in Silicon Valley, the 40th percentile may not be ideal. However, for the purpose of this analysis, a reliable measure of the *relative* cost of rent among ZIP Codes was needed so that “undervalued” ZIP Codes can be identified. The consistent data collection methodology behind the HUD FMR data made it an attractive choice for this purpose.

Methodology

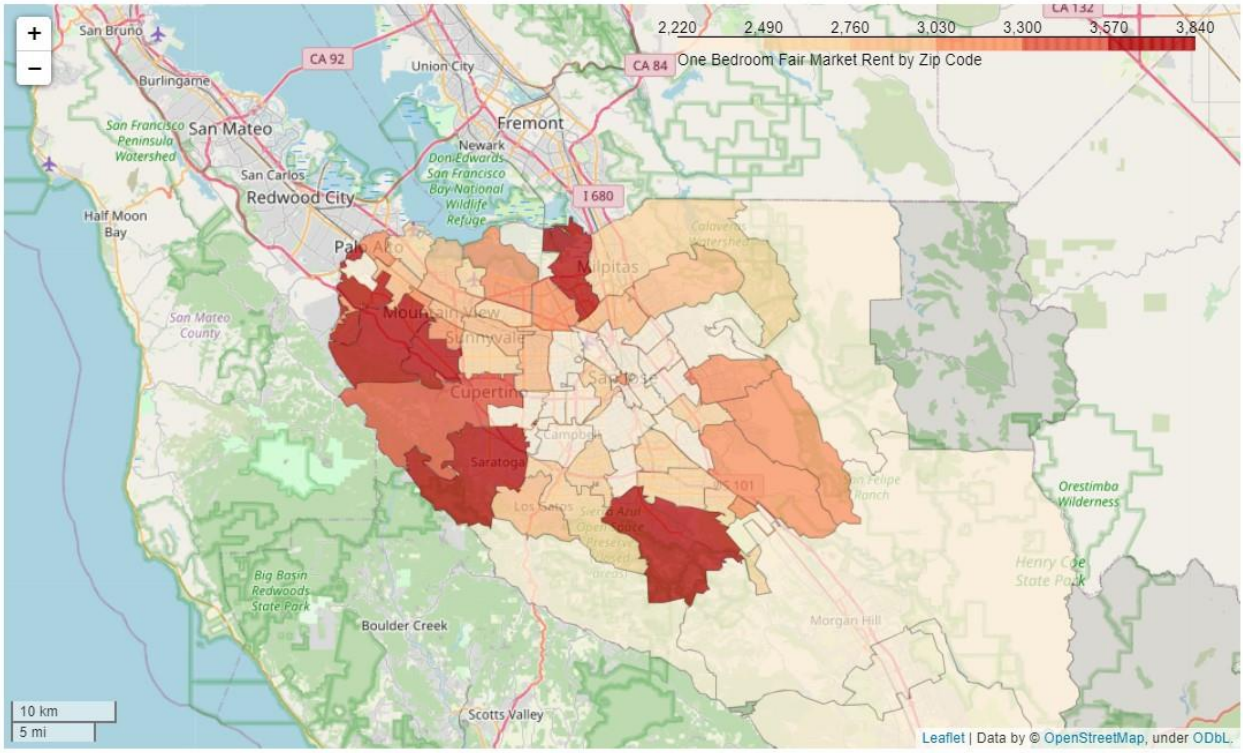
Setting the Scene

First, the ZIP Codes for Santa Clara valley were extracted from the GeoJSON file and mapped:



Three ZIP Codes were eliminated at this stage: 95076, which is mainly in Santa Cruz County; 94550 whose main city, Livermore, is in Alameda County; and 95023, whose main city, Hollister, is in San Benito County.

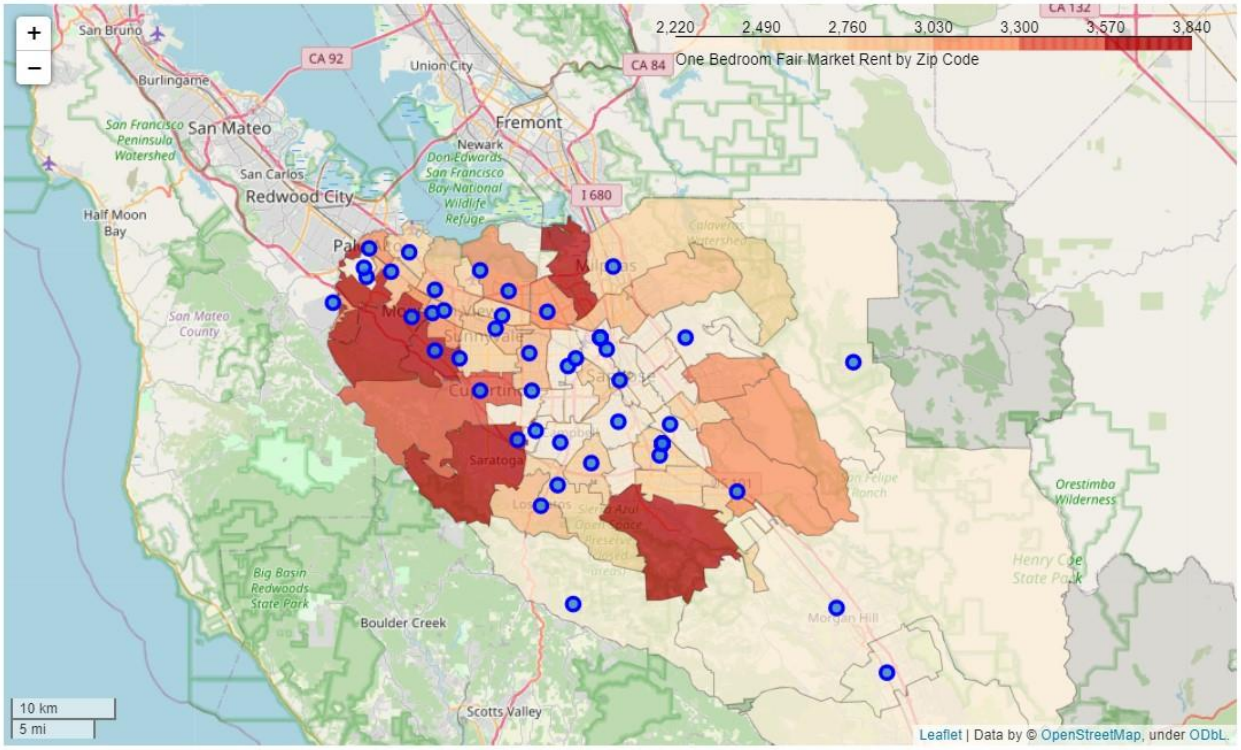
The next step was to merge in the FMR rent data. A choropleth map was produced to give an idea of the general pattern of ZIP Codes that might be expected. The gray ZIP Codes are those that were eliminated above.



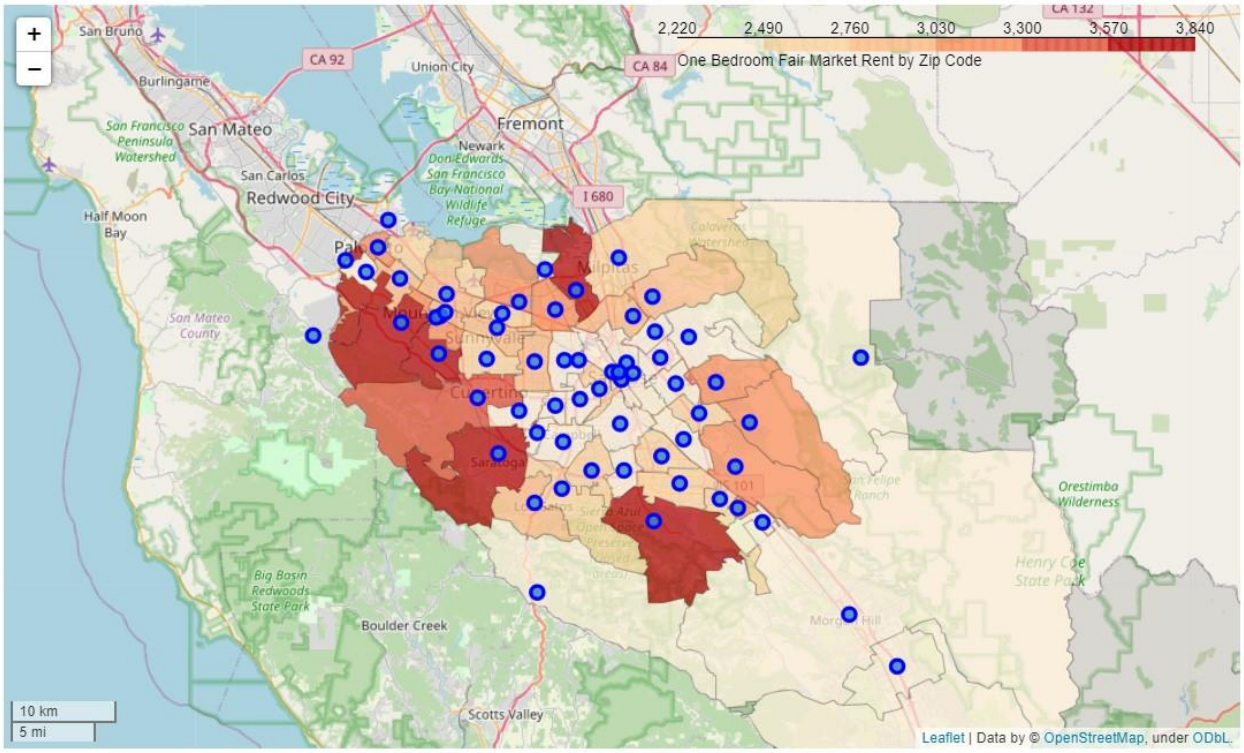
The most expensive areas are largely on the western side of the valley plus Almaden at the south. The core of San Jose and Campbell is less expensive, surrounded by a ring of posher suburbs, in turn bordered by less expensive rural areas to the south and east.

Next, latitude and longitude coordinates for each ZIP Code were obtained. The first attempt used two free geocoders available through the geopy library, Nominatim and ArcGIS. Both successfully returned coordinates, but when plotted on the map, they did not well represent the population centers of many ZIP Codes. In some cases, coordinates actually landed in neighboring ZIP Codes or were missing altogether. Using these coordinates with Foursquare API calls would not give accurate results!

Below are the Nominatim coordinates to give an idea of the problem. The ArcGIS data was similar.



Research into the nature of ZIP Code definitions revealed that what was needed were the center coordinates of the US Census ZIP Code Tabulation Areas (ZCTAs). The ZCTA-based coordinates from GeoNames better represent the population centers of the ZIP Codes under consideration, as seen below:



Note that the centers are not at the geographical centers but in the functional centers. Take, for example, Cupertino, the terra cotta colored blob on the western side of the valley between the two dark red ones. Its ZIP Code extends well into the Santa Cruz Mountains to the west, but the coordinates place the "center" closer to the where most of the population and venues are. A couple of the ZIP Codes on the northwest edge seem a bit off, but this is due to the fact that parts of those ZIP Codes extend outside of Santa Clara County. This should not affect the venue data much as the (sub)urban sprawls spills over the county lines in this direction. In other words, the venue mix does not change.

As the final step in setting the scene, the information so far was collected together for ease of analysis. Here are the first five ZIP Codes:

	zipcode	fmr1bdrm	lat	long
0	94022	3620	37.3814	-122.1258
1	94024	3650	37.3547	-122.0862
2	94028	3230	37.3702	-122.2182
3	94035	2560	37.3861	-122.0839
4	94040	3140	37.3855	-122.0880

From Foursquare to Feature Set

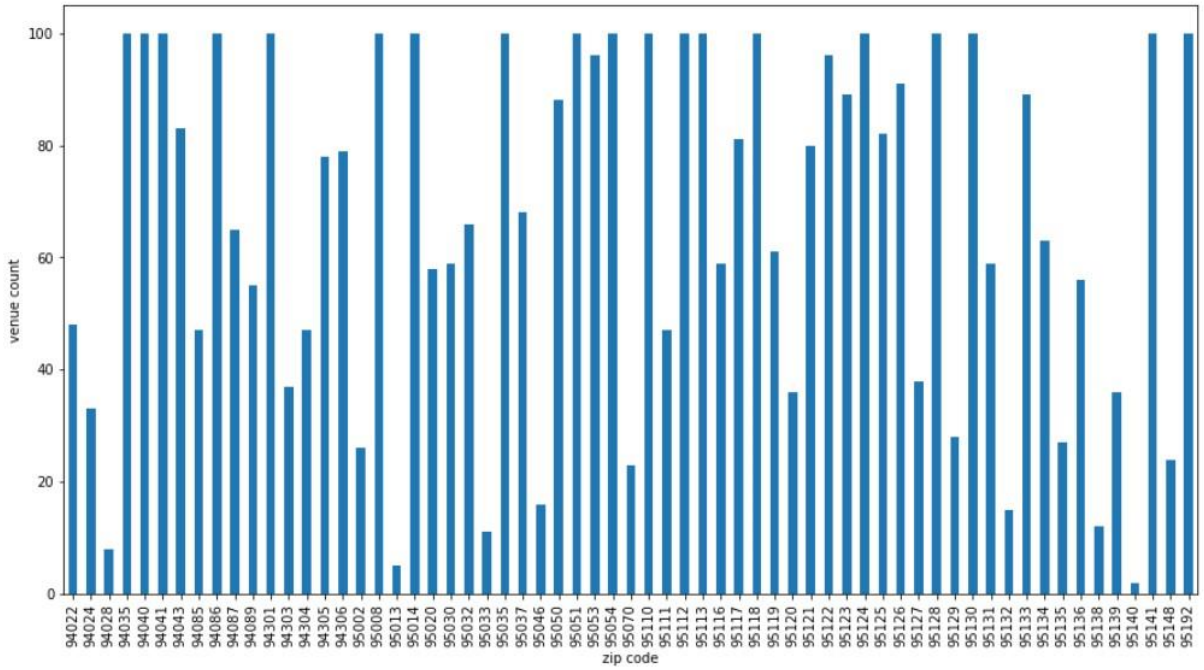
Once ZIP Code center coordinates were found, explore requests to the Foursquare API were made to retrieve nearby venue information. Requests include a latitude, longitude, and search radius (in meters). The response contains the names, locations, and categories of up to 100 recommended venues.

Setting the search radius is an art. Too small, and too few venues are returned. Too big, and the search area for neighboring ZIP Codes overlaps significantly. Both conditions would result in sub-optimal clustering. Most of the valley is suburban sprawl, so the best radius will be bigger than for truly urban areas with higher venue density and smaller ZIP Codes. An initial radius of 1000 meters returned only 49 venues for a medium-sized ZIP Code, 95124. A radius of 1250 meters returned 81, and a radius of 1500 meters returned the full 100. Looking at the ZIP Code map above, it is clear that going any larger than 1500 would cause too much overlap. As such, 1500 meters was chosen as the search radius.

Next, API calls were made to all of the ZIP Codes using a radius of 1500 meters. The venue name, location, and category information were extracted and tabulated. The first five venues for the first ZIP Code look like this:

	zipcode	zip_lat	zip_long	venue	venue_lat	venue_long	venue_category
0	94022	37.3814	-122.1258	Tin Pot Creamery	37.378643	-122.118322	Ice Cream Shop
1	94022	37.3814	-122.1258	Linden Tree	37.379216	-122.116963	Bookstore
2	94022	37.3814	-122.1258	Manresa Bread	37.379234	-122.117080	Bakery
3	94022	37.3814	-122.1258	Asa	37.379490	-122.116821	Spanish Restaurant
4	94022	37.3814	-122.1258	Satura Cakes	37.378835	-122.115860	Bakery

The number of venues returned for each ZIP Code was visualized with a bar chart:



Even with a 1500 meter radius, few venues were returned for the more rural ZIP Codes. Since this analysis is focused on the (sub)urban core of the county, the low-data, outlying ZIP Codes were removed at this stage. They are 94028, 95013, 95033, 95037, 95046, and 95140. Note that 95037 (Morgan Hill) did return a fairly high number of venues, but was eliminated because it is in the far south of the county, separated from Silicon Valley by rural ZIP Codes. At this point, 57 ZIP Codes remained.

The final step before clustering was to transform the raw venue data into a useable feature set for clustering. First, the venue categories were encoded using one hot encoding. There are 331 unique venue categories in this data set, and each with its own column. The first five venues look like this:

	zipcode	ATM	Accessories Store	Acupuncturist	Afghan Restaurant	Airport Terminal	American Restaurant	Andhra Restaurant	Arcade	Art Gallery	...	Vietnamese Restaurant	Volleyball Court	Warehouse Store
0	94022	0	0	0	0	0	0	0	0	0	...	0	0	0
1	94022	0	0	0	0	0	0	0	0	0	...	0	0	0
2	94022	0	0	0	0	0	0	0	0	0	...	0	0	0
3	94022	0	0	0	0	0	0	0	0	0	...	0	0	0
4	94022	0	0	0	0	0	0	0	0	0	...	0	0	0

The data was then grouped by ZIP Code, and the mean of each category was calculated for each ZIP Code. These values are the percentage of each ZIP Code's venues in each category and therefore will be between 0 and 1. The resulting frequency matrix for the first five ZIP Codes looks like this:

	zipcode	ATM	Accessories Store	Acupuncturist	Afghan Restaurant	Airport Terminal	American Restaurant	Andhra Restaurant	Arcade	Art Gallery	...	Vietnamese Restaurant	Volleyball Court	Warehouse Store
0	94022	0.020833	0.0	0.0	0.0	0.0	0.062500	0.0	0.020833	0.0	...	0.000000	0.0	0.0
1	94024	0.030303	0.0	0.0	0.0	0.0	0.030303	0.0	0.000000	0.0	...	0.030303	0.0	0.0
2	94035	0.010000	0.0	0.0	0.0	0.0	0.010000	0.0	0.000000	0.0	...	0.000000	0.0	0.0
3	94040	0.000000	0.0	0.0	0.0	0.0	0.010000	0.0	0.000000	0.0	...	0.000000	0.0	0.0
4	94041	0.010000	0.0	0.0	0.0	0.0	0.010000	0.0	0.000000	0.0	...	0.010000	0.0	0.0

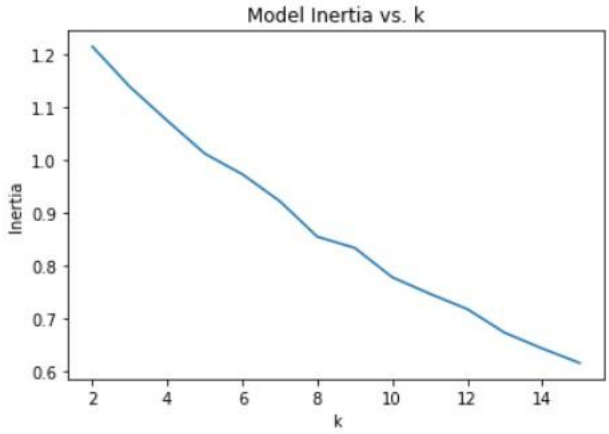
This is the feature set that was used cluster ZIP Codes in the next stage of analysis.

K-Means Clustering of ZIP Codes on Venue Category Frequency

K-Means clustering is a simple yet often effective unsupervised machine learning algorithm which is used for many applications. In this case, ZIP Codes in the core (sub)urban part of Santa Clara County were clustered based on the frequency of different venue categories within them. Clustering identified clusters of ZIP Codes with similar mixes of venue categories and, ideally, a similar "feel" to them.

The algorithm starts with k number of cluster centroids, chosen at random. It then assigns data points to the closest centroid, relocates the centroids to the mean of the data points assigned to them, and repeats the process until the centroids stop moving. The goal is to minimize the intra-cluster distance between points as measured by inertia, the sum of the squared distance between each data point and its centroid.

K-Means requires the number of clusters (k) be set in advance. One way to determine the optimal value for k is to use the "elbow method". Models are created for a range of k values. The inertia will always go down when the number of clusters increased. However, on a plot of inertia versus k, a point or "elbow" can be identified where the decrease in inertia gained by increasing k starts to flatten out. For this analysis, models were created for values of k from 2 to 15, with the following results:



The elbow is at k = 8. A final K-Means clustering model was created with k=8, and the resulting cluster labels (0 through 7) were added to the ZIP Code data:

	zipcode	fmr1bdrm	lat	long	label
0	94022	3620	37.3814	-122.1258	1
1	94024	3650	37.3547	-122.0862	1
2	94035	2560	37.3861	-122.0839	1
3	94040	3140	37.3855	-122.0880	1
4	94041	2900	37.3893	-122.0783	1

Clusters were visualized through mapping and through tabulating the top ten venue categories in each.

Clusters vs. Rent

Rent data, in the form of ZIP Code Fair Market Rents (FMRs), were grouped by cluster. The mean FMR and FMR distribution for each cluster was examined visually through box and violin plots. These plots allow quick determination of the relationship between cluster membership and rent.

To determine if the difference in mean FMR between the two largest clusters is significant, the Kruskal-Wallis H statistic and associated p value were calculated. The Kruskal-Wallis H test is similar to the ANOVA (F test) test for equal means. However, unlike ANOVA, Kruskal-Wallis does not assume a normal distribution within samples nor equal variances for all samples. It is appropriate here because the FMR distribution within clusters is not normal and the cluster FMR variances are not equal. The Kruskal-Wallis test works best for samples with more than five points. As such, the two largest clusters were compared.

As a final step, the ZIP Codes in the cluster identified as having a significantly higher mean FMR were sorted by FMR. The five ZIP Codes in this cluster with the lowest FMR were investigated further by the simple means of googling them and seeing what is there.

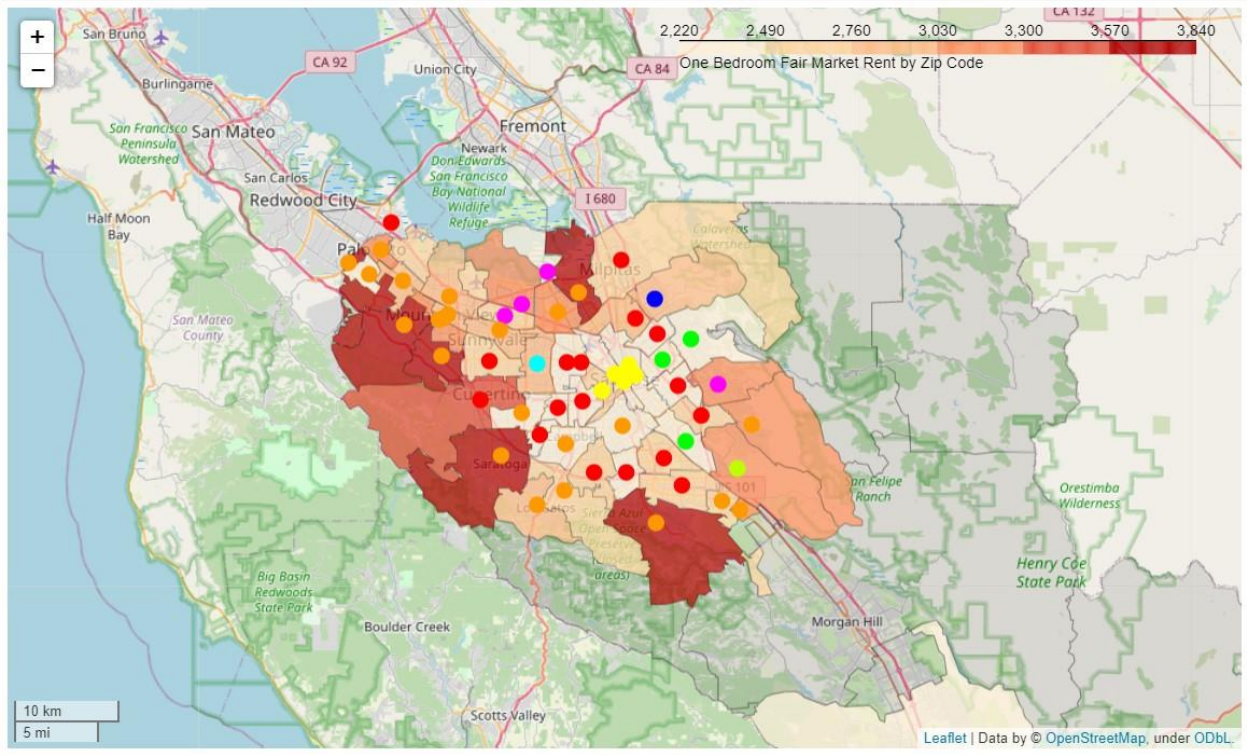
Results

The Clusters

K-Means clustering of ZIP Codes on venue category frequency data with k=8 resulted in eight clusters, numbered 0 through 7, with the following number of ZIP Codes in each one:

label	0	1	2	3	4	5	6	7
# of zip codes	18	23	6	1	3	1	1	4

The clusters can be visualized on a map:



Although the ZIP Codes were clustered solely on venue category data, the resulting clusters form clear geographical patterns. Cluster 3 (yellow) is downtown San Jose, the county's largest city. Cluster 0 (red) forms an inner suburban ring around the downtown area. Cluster 1 (orange) predominates in the western and southern suburbs of the valley while cluster 4 (green) is limited to the eastern suburbs. Cluster 7 (pink) is mostly on the northern edge by San Francisco Bay. The remaining clusters appear to be outliers. A cursory glance suggests that cluster 1 (orange) is associated with ZIP Codes with higher rents. More on that below.

Another way to understand the clusters is to extract the top ten venue categories for each:

	label	color	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	red	Coffee Shop	Sandwich Place	Vietnamese Restaurant	Pizza Place	Grocery Store	Fast Food Restaurant	Mexican Restaurant	Convenience Store	Bubble Tea Shop	Chinese Restaurant
1	1	orange	Coffee Shop	Park	Pizza Place	Mexican Restaurant	Grocery Store	Sandwich Place	Chinese Restaurant	Fast Food Restaurant	Gym / Fitness Center	Italian Restaurant
2	2	yellow	Mexican Restaurant	Coffee Shop	Bar	Cocktail Bar	Sandwich Place	Sushi Restaurant	Pub	Pizza Place	Ice Cream Shop	Grocery Store
3	3	lime	Yoga Studio	Hotel	Playground	Seafood Restaurant	Music Venue	Sushi Restaurant	Motel	Gym / Fitness Center	Mexican Restaurant	Baseball Field
4	4	green	Mexican Restaurant	Fast Food Restaurant	Convenience Store	Intersection	Pizza Place	Sandwich Place	Vietnamese Restaurant	Bakery	Video Store	Chinese Restaurant
5	5	light blue	Korean Restaurant	Indian Restaurant	Chinese Restaurant	Coffee Shop	Sushi Restaurant	BBQ Joint	Sandwich Place	Park	Spa	Rental Car Location
6	6	blue	Park	Light Rail Station	Tennis Court	Fried Chicken Joint	Playground	Food & Drink Shop	Bubble Tea Shop	Volleyball Court	Burrito Place	Donut Shop
7	7	pink	Park	Hotel	Food Truck	Mexican Restaurant	Convenience Store	Gym	Indian Restaurant	Vietnamese Restaurant	Video Store	Gym / Fitness Center

Cluster 2 (yellow) is the downtown cluster. Venue categories unique to its top ten are bars, cocktail bars, and pubs, indicating this is a key evening entertainment destination.

Cluster 0 (red) is the inner suburban ring. No categories are unique to its top ten. The difference may be in its relative lack of venues in categories more common farther from downtown such as parks.

Cluster 1 (orange) is the mostly western cluster. It has a similar mix as the inner suburban ring, with a higher portion of its venues being parks, gyms, and Italian restaurants.

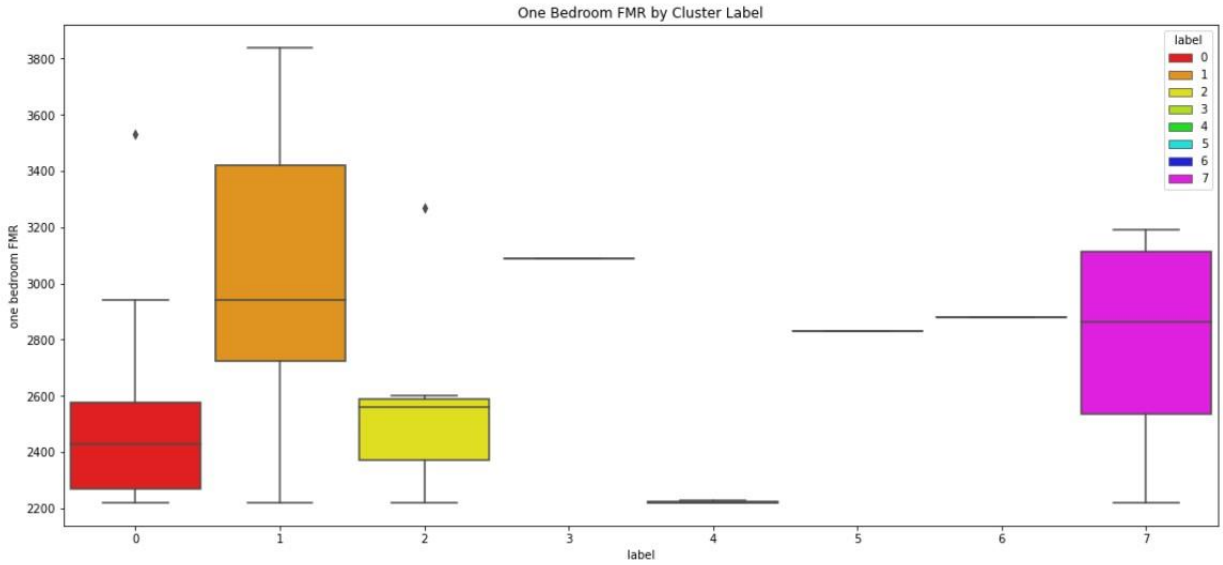
Cluster 4 (green) is the eastern cluster. Compared to cluster 1 (western), Cluster 4 has a higher portion of its venues being convenience stores, video stores, and bakeries and a lower portion being coffee shops and parks.

Cluster 7 (pink) is the mainly northern cluster. This cluster and the single-ZIP Code cluster 5 (light blue, just south of the pink cluster) are the only two with Indian restaurants in the top ten categories.

The remaining clusters only have one ZIP Code in them. Cluster 5 (light blue) is 95051 is the only cluster with Korean restaurants in its top ten. Not surprisingly, this area is known as Koreatown. Cluster 6 (blue) is 95132. Only 15 venues were returned for it, and the small sample size may have made this an outlier. Cluster 3 (lime green) is 95138. Only 12 venues were returned for it, and, again, the small sample size may have made this an outlier.

Mean Rent and Rent Distribution within Clusters

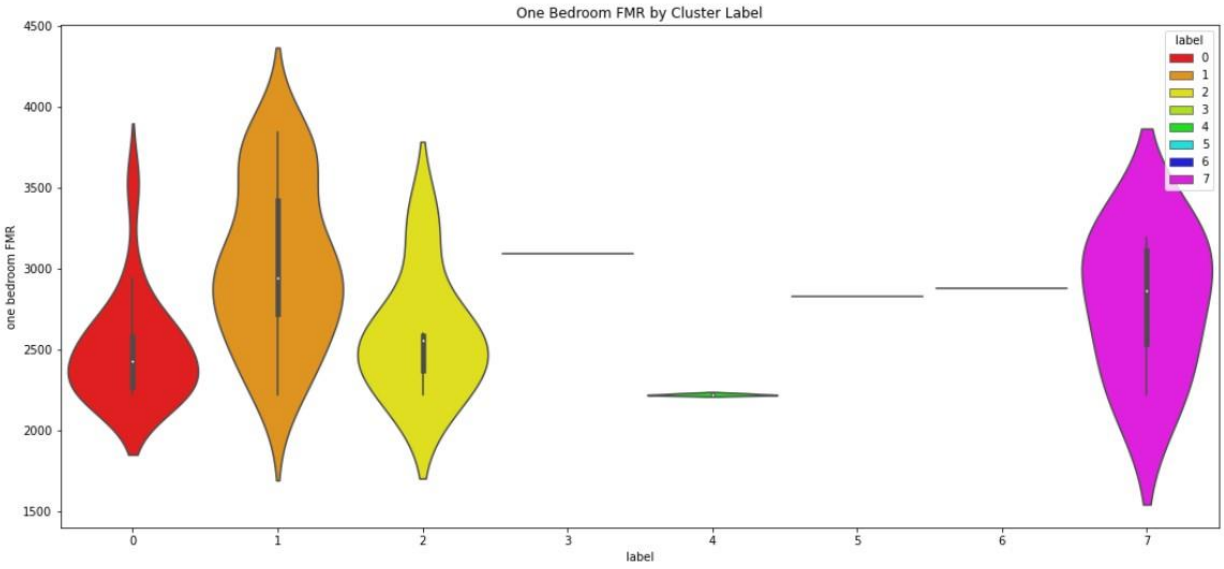
The distribution of ZIP Code Fair Market Rents (FMRs) within clusters and the difference in mean FMR between clusters can be seen in a box plot. The colors are the same as used in the map above.



Recall that clusters 3, 5, and 6 only have one ZIP Code and therefore only one FMR. Cluster 4 (east) has three ZIP Codes, all of which have lower FMRs. On the other hand, cluster 7 (north) has four ZIP Codes, which are spread out over a much wider range.

Among the three clusters with the most ZIP Codes, 0 (inner suburban ring), 1 (west), 2 (downtown), cluster 1 has the highest mean FMR.

A cousin of the box plot, the violin plot shows the distribution of FMRs within the clusters more clearly:



It can be seen that although cluster 1 (west) has a higher mean FMR, it has a very wide distribution of FMRs. To determine if the mean FMR of cluster 1 is statistically significant from mean FMR of cluster 0

(the adjoining suburban inner ring), a Kruskal-Wallis H test for equal means was run. The results are an H statistic of 13.26484 and a p value of 0.00027. Therefore, with a 95% confidence level, the null hypothesis that the cluster means are the same can be rejected as the p value is less than the significance level (0.05).

In other words, while cluster membership is not strongly predicative of rent (the FMRs of cluster 1 span the entire FMR range), the difference in mean FMR between clusters 0 and 1 is statistically meaningful. Cluster 1 is the "posh" cluster.

Potentially Undervalued ZIP Codes

To identify possibly "undervalued" ZIP Codes within cluster 1, its ZIP Codes were sorted by FMR, with the ten cheapest ZIP Codes as follows:

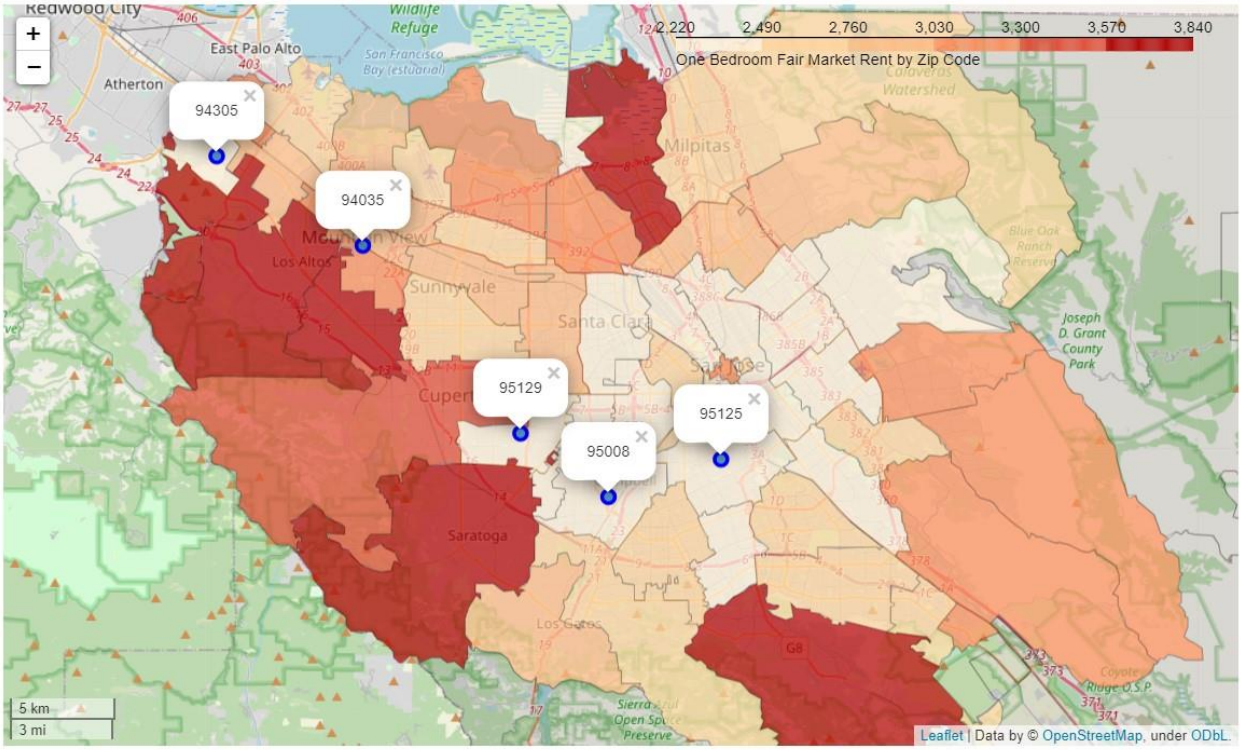
	zipcode	fmr1bdrm	label
0	94305	2220	1
1	95125	2320	1
2	95129	2410	1
3	95008	2440	1
4	94035	2560	1
5	95032	2700	1
6	95119	2750	1
7	94086	2760	1
8	95030	2770	1
9	94043	2890	1

The five ZIP Codes with the lowest FMR are 94305 (Stanford University), 95125 (Willow Glen), 95129 (West San Jose), 95008 (Campbell), and 94035 (Moffett Field).

Discussion

Cluster 1, a collection of ZIP Codes mainly in the western and southern suburbs of Silicon Valley, is distinguished by a higher portion of parks, gyms, and Italian restaurants. Additionally, its mean one-bedroom Fair Market Rent (FMR) is significantly higher than that of cluster 0, the inner suburbs. This suggests that there may be good opportunities for investment in apartment buildings in those ZIP Codes of cluster 1 which are not yet expensive but have already attracted the sorts of venues that are associated with higher rent areas.

Of particular interest are the five ZIP Codes in cluster 1 with the lowest FMRs: 94305, 95125, 95129, 95008, and 94035.



94305 is the Stanford University campus. This is a nice place to live, but there are no apartment investment opportunities as the land is owned by the University. The rent is likely lower because of the subsidized apartments the University has built for its graduate students and faculty.

95125 is Willow Glen. This San Jose neighborhood is known for its cute core shopping/dining area. It is worth exploring for rental investment opportunities as a possible up and coming neighborhood.

95129 is on the western side of San Jose. Sandwiched between the wealthier areas of Saratoga (95070, the most expensive ZIP Code in the entire dataset) and Cupertino (95014), this area is also worthy of further research.

95008 is Campbell. Campbell has revitalizing its core shopping/dining area and has the "feel" of a neighborhood on the rise. Again, worth a look.

94035 is limited to Moffett Field, a joint civil-military airport. The lower rents in this ZIP Code are likely the product of government housing for military families. Like Stanford, this is not a viable area for investment in apartment buildings.

Conclusion

This analysis of ZIP Codes in the core (sub)urban area of Santa Clara County, California has identified three ZIP Codes which have the venue category mix of areas with more expensive rents but yet themselves have lower rents: 95125 (Willow Glen), 95129 (West San Jose), and 95008 (Campbell). These ZIP Codes offer potential opportunities for investment in apartment buildings as the rents may be positioned to go up.

Before investing, however, further research is needed. There may be factors not considered in this preliminary analysis which bring the rent down in these areas, such as school ratings, crime rates, or traffic noise. There may also be limited investment opportunities due to zoning laws. Nevertheless, having a short list of candidate ZIP Codes to investigate will decrease the effort and cost of this further research.