Applied Data Science Capstone

House Hunting in San Francisco

Luis Macias-Navarro

Using unsupervised learning to cluster homes for sale in San Francisco based on nearby venues and crime statistics.

1. Introduction

1.1 Background

Purchasing a home is the largest financial transaction most people will conduct in their lifetimes. What's more, this purchase usually represents the lion's share of an individual or couple's net worth. As a result, such a large and important transaction requires a fair amount of time and due diligence to complete.

Luckily there are tools available online such as Redfin and Zillow that make searching for a home very easy and convenient. These real estate platforms provide data on a home's listing price, days on market, estimated sales price, property history, etc. and can even connect homebuyers with a real estate agent.

However, these are not the best tools when it comes to comparing many homes against one another across any number of features.

1.2 Problem

Homebuyers find it difficult to compare houses across different neighborhoods based on nearby venues and crime statistics.

Online real estate platforms such as Redfin provide generic scores based on walkability, transit options and cycling infrastructure. While the walkability score is useful it does not provide any granularity regarding what type of venues are in the vicinity (Figure 1.). Furthermore Redfin does not explicitly provide any information related to crime statistics for a given neighborhood.

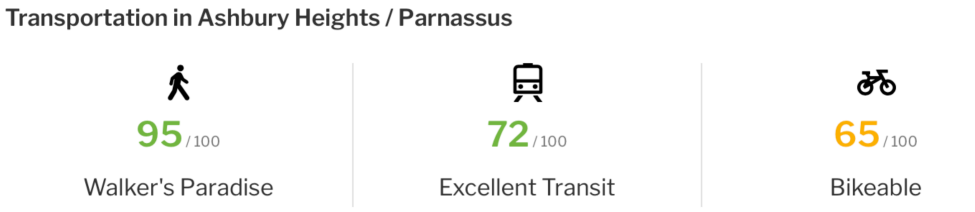


Figure 1. Redfin neighborhood information

The solution is to use unsupervised learning to cluster these homes on two dimensions:

1) Venue-based clustering - using venue data from Foursquare

2) Crime-based clustering - using crime data from SF Data

Having homes clustered based on venues allows buyers to understand whether homes in different parts of the city have access to similar type venues. Crime-based clustering tells them which homes have similar crime rates even if they're in different parts of the city.

1.3 Target Audience

Homebuyers that want to make an informed decision by taking into account nearby venues as well as crime statistics for each one of the homes they are considering.

2. Data

2.1 Data Gathering

The basis for this project was data generated from Redfin on 57 homes for sale in San Francisco that met the *Acquisition criteria* displayed in Table 1 below. Each one of those homes had a set of geo coordinates (latitude and longitude) that were then used as centers to which venue and crime data was assigned based on distance. The data was created by exporting from the Redfin website as a CSV file.

Venue data for each of the 57 homes was generated using the Foursquare API. Foursquare returned all the venues (limited to 100) within a 500-meter (0.31 miles) radius of each home. The distance of 500 meters was chosen because it is a reasonable, maximum distance to walk without needing to take public transportation or drive.

Crime data was acquired through the SF Data portal and downloaded into a CSV file; crime incidents reported had a unique set of geo coordinates (latitude and longitude). Similar to the venue-based data, crime data was generated for each home within a 500 meter radius. This was chosen to reflect how safe the immediate vicinity of each home is and what types of crime are perpetrated within walking distance. In addition, and due to the huge volume of incidents reported (yikes), only incidents reported during February of 2020 were included in this project.

Table 1 **Summary of** data sources

Source	Type of Data	Features used for modeling	Acquisition criteria	Acquisition Method
Redfin www.redfin.com	Real Estate	For each home: • Street address • Geographic coordinates	All homes in SF available for sale mid-Feb 2020, with at least 2 bedrooms and 2 bathrooms, selling price in the $1MM range.	CSV file downloaded from Redfin webpage
Foursquare developer.foursquare.com	Venues	For each venue: • Geographic coordinates • Venue category • Venue subcategory	All venues within 500m (0.31miles) of each home, limited to 100 venues	API call to Foursquare via IBM Watson
SF Data datasf.org/opendata/	Crime	For each reported crime • Geographic coordinates • Incident category • Incident subcategory	All crimes committed within 500m (0.31miles) of each home during Feb 2020	CSV file downloaded from SF Data webpage

2.2 Data cleaning

Data downloaded from Redfin required minimal cleaning because the export feature on the website allows for a CSV download with the selected criteria. In the CSV, each home was listed in a separate row and the geo coordinates were right alongside each address in a separate column for latitude and longitude. That was all the data needed from Redfin. Foursquare data required minimal wrangling, once the API was called, to get the data into a dataframe and matched up with the Redfin data.

Crime data from SF Data was generated into a CSV. This data required the most wrangling because the distance between each home and each location where a crime had been reported had to be calculated. This was done in Excel using the Haversine formula:

$$d = 2r \arcsin\left(\sqrt{\operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)}\right)$$

Where:

- ϕ_1, ϕ_2: latitude of point 1 and latitude of point 2 (in radians),
- λ_1, λ_2: longitude of point 1 and longitude of point 2 (in radians).

After each distance was calculated, only distances <= 500 meters between a crime location and a home were taken into account. This resulted in a CSV file that matched up crime locations with each one of the 57 home addresses.

3. Methodology

This project was subdivided into two efforts:

1) venue-based clustering and 2) crime-based clustering. The methodology followed for both was similar in terms of exploratory data analysis as well as use of machine learning to cluster.

The first step was to import Redfin data and visualize all the homes for sale on a map of San Francisco, see Figure 2. Most of the homes for sale are on the Eastern side of the city[1]. However there are enough homes scattered throughout San Francisco to hopefully derive meaningful insight form clustering.

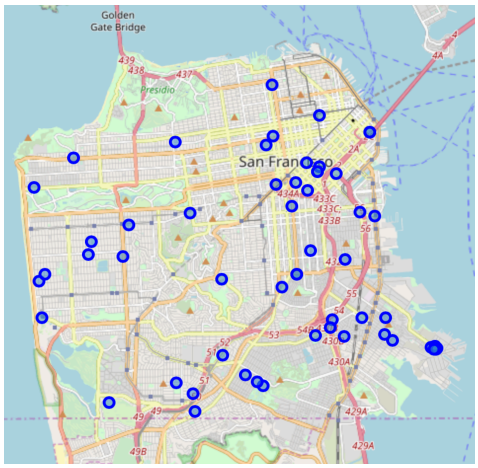


Figure 2 Homes for sale in San Francisco

3.1 Exploratory Data Analysis and Clustering

3.1.1 Venue-based analysis

The imported datasets from Redfin and Foursquare were merged into a single dataframe showing all of the venues within 500m of each address (Table 2).

Table 2 Abridged dataframe showing venues associated with each home

	ADDRESS	ADDRESS Latitude	ADDRESS Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	257 N Lake Merced Hls	37.712635	-122.481353	Holy Trinity Center	37.714081	-122.483060	Church
1	257 N Lake Merced Hls	37.712635	-122.481353	Lake Merced Run Loop	37.713463	-122.485606	Trail
2	257 N Lake Merced Hls	37.712635	-122.481353	San Francisco Golf Club	37.712459	-122.476884	Golf Course
3	257 N Lake Merced Hls	37.712635	-122.481353	Parkmerced Dog Run	37.715189	-122.479566	Dog Run
4	257 N Lake Merced Hls	37.712635	-122.481353	Lake Merced Fishing Pier And Nature Trail	37.713396	-122.486065	Trail

The total number of venues for all 57 homes totaled 2247 with 293 unique categories, an average of 39 venues per home.

The next step was to use *One Hot Encoding* to convert the categorical data into numeric data and then take the mean of the frequency of occurrence of each category. The result

[1] These homes were for sale in mid-Feb 2020. They are in the ~$1MM range with at least 2 bedrooms and 2 bathrooms

was a dataframe that listed each home alongside the most common occurring venues (Table 3).

Table 3 Abridged dataframe of most common venues for each home

ADDRESS	LATITUDE	LONGITUDE	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
257 N Lake Merced Hls	37.712635	-122.481353	Trail	Church	Dog Run	Golf Course	Dumpling Restaurant
1957 San Jose Ave	37.725763	-122.442057	Tennis Court	Coffee Shop	Light Rail Station	Chinese Restaurant	Dessert Shop
1328-1330 Silver Ave	37.731009	-122.409813	Chinese Restaurant	Grocery Store	Pizza Place	Playground	Recreation Center
1501 Greenwich St #202	37.800064	-122.424680	Spa	Coffee Shop	Hotel	Italian Restaurant	Park
69 Palm Ave #4	37.784415	-122.458306	Chinese Restaurant	Coffee Shop	Thai Restaurant	Italian Restaurant	Bank
878 47th Ave	37.771922	-122.507591	Bus Stop	Grocery Store	Chinese Restaurant	Garden	Board Shop
8 Octavia St #301	37.772549	-122.423361	Cocktail Bar	Wine Bar	Optical Shop	Sushi Restaurant	New American Restaurant
400 Beale St #1012	37.786905	-122.390920	Coffee Shop	Gym	Scenic Lookout	Food Truck	Lounge
222 Parnassus Ave Unit B	37.764716	-122.453196	Mexican Restaurant	Park	Middle Eastern Restaurant	Track Stadium	Café

The K-means clustering algorithm was used to cluster the homes based on the frequency of associated venues. K of 6 was selected through trail and error (as opposed to elbow function) since it gave the most reasonable, intuitive and understandable clustering.

3.1.2 Crime-based analysis

The imported datasets from Redfin and SF Data (crime) were merged into a single dataframe showing all of the reported crime incidents within 500m of each address (Table 4).

Table 4 Abridged dataframe showing crimes associated with each home for the February 2020 timeframe

ADDRESS	ADDRESS Latitude	ADDRESS Longitude	CRIME	CRIME Latitude	CRIME Longitude	CRIME Category
257 N Lake Merced Hls	37.712635	-122.481353	Assault	37.716375	-122.479185	Simple Assault
257 N Lake Merced Hls	37.712635	-122.481353	Malicious Mischief	37.716375	-122.479185	Vandalism
257 N Lake Merced Hls	37.712635	-122.481353	Assault	37.716246	-122.483275	Simple Assault
1957 San Jose Ave	37.725763	-122.442057	Burglary	37.727861	-122.438085	Burglary - Hot Prowl
1957 San Jose Ave	37.725763	-122.442057	Drug Offense	37.723130	-122.435977	Drug Violation

The total number of crimes reported for all 57 homes totaled 3776 with 45 unique categories, an average of 66 incidents per home in the month of February 2020.

The next step was to use *One Hot Encoding* to convert the categorical data into numeric data and then take the mean of the frequency of occurrence of each category. The result was a dataframe that listed each home alongside the most common occurring crimes (Table 5).

Table 5 Abridged dataframe of most common crimes for each home

ADDRESS	1st Most Common Crime	2nd Most Common Crime	3rd Most Common Crime	4th Most Common Crime	5th Most Common Crime
1957 San Jose Ave	Larceny - From Vehicle	Motor Vehicle Theft	Vandalism	Missing Person	Weapons Offense
1501 Greenwich St #202	Larceny - From Vehicle	Vandalism	Larceny Theft - Other	Motor Vehicle Theft	Burglary - Other
69 Palm Ave #4	Larceny - From Vehicle	Larceny Theft - Other	Vandalism	Theft From Vehicle	Robbery - Commercial
8 Octavia St #301	Larceny - From Vehicle	Vandalism	Larceny Theft - Other	Simple Assault	Motor Vehicle Theft
2002 3rd #117	Larceny - From Vehicle	Larceny Theft - Other	Simple Assault	Larceny Theft - From Building	Warrant
1139 Judah St	Larceny - From Vehicle	Larceny Theft - Other	Vandalism	Burglary - Residential	Theft From Vehicle
1635 29th Ave	Larceny - From Vehicle	Larceny Theft - Other	Larceny Theft - From Building	Vandalism	Theft From Vehicle
5090 Diamond Heights Blvd Unit A	Larceny - From Vehicle	Fraud	Larceny Theft - Other	Burglary - Commercial	Vandalism

The final step was to cluster the homes based on the frequency of associated crimes using K-means clustering algorithm.

4. Results

4.1 Venue-based analysis

The k-means clustering algorithm resulted in five clusters based on venues. Each cluster is represented by a different color on the map in Figure 3.

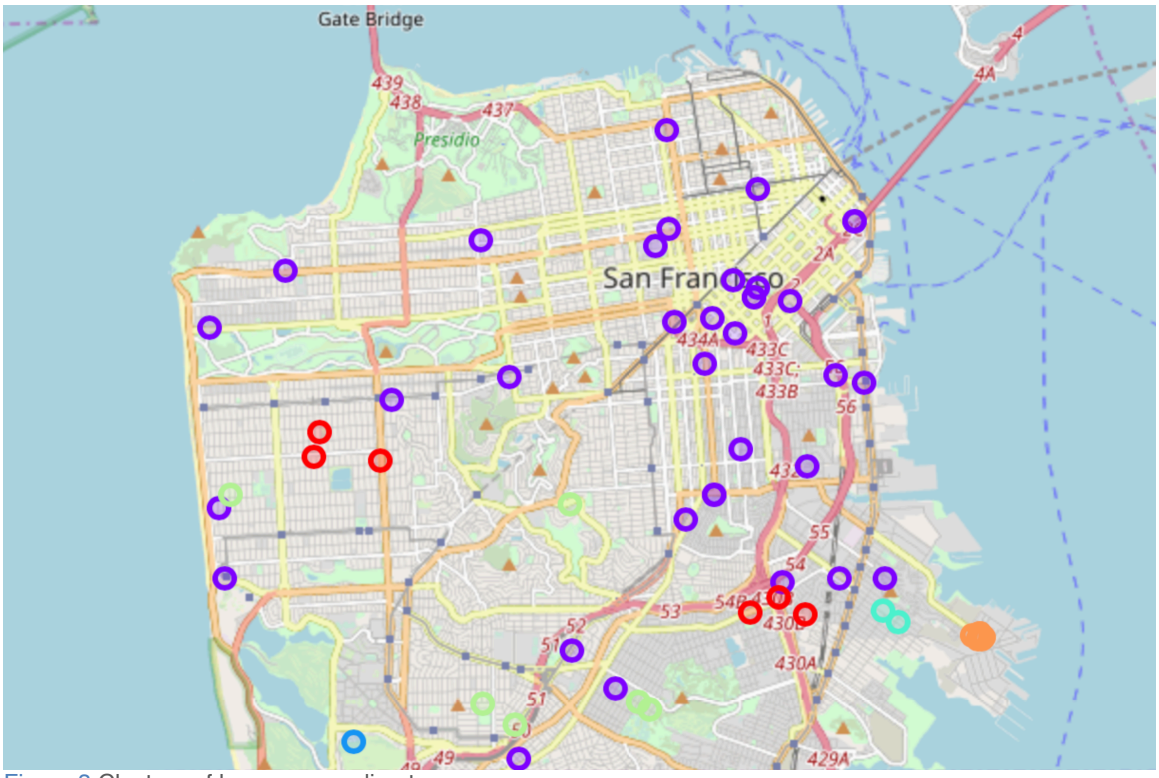
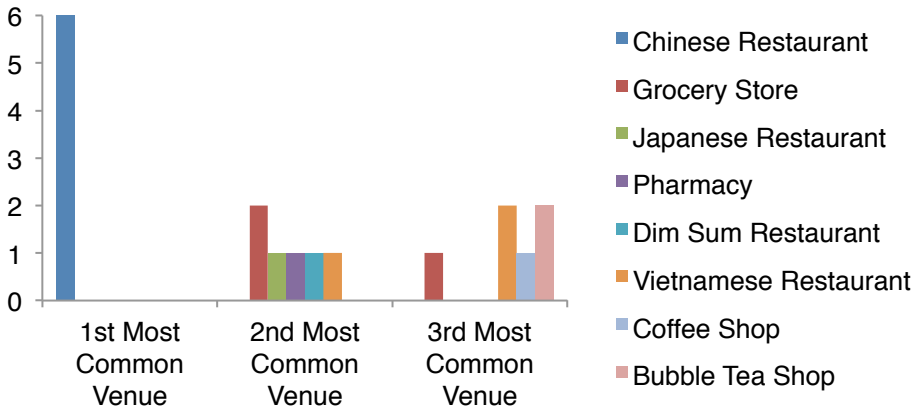


Figure 3 Clusters of homes according to venues

- Red Cluster – Chinese Restaurants

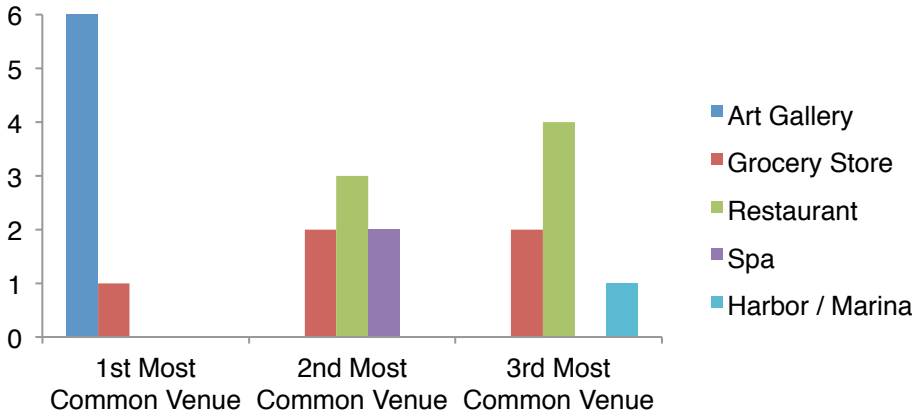
Chart 1 Most common venues in the Red Cluster



Homes in the Red Cluster primarily have easy access to Chinese restaurants and grocery stores. There are other Asian restaurants and pharmacies nearby as well.

- Orange Cluster – Art Galleries and Spas

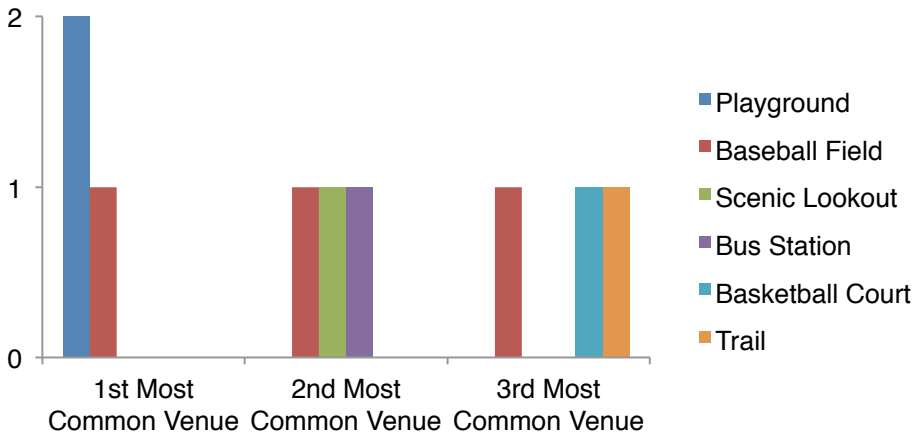
Chart 2 Most common venues in the Orange Cluster



The Orange Cluster represents a group of homes that are geographically very close and somewhat isolated from the rest (on the southeast corner of the map in Figure 3). As such they stand out as a unique cluster due to the unique mix of venues available nearby such as art galleries, spas and the marina.

- Green Cluster – Baseball and Playgrounds

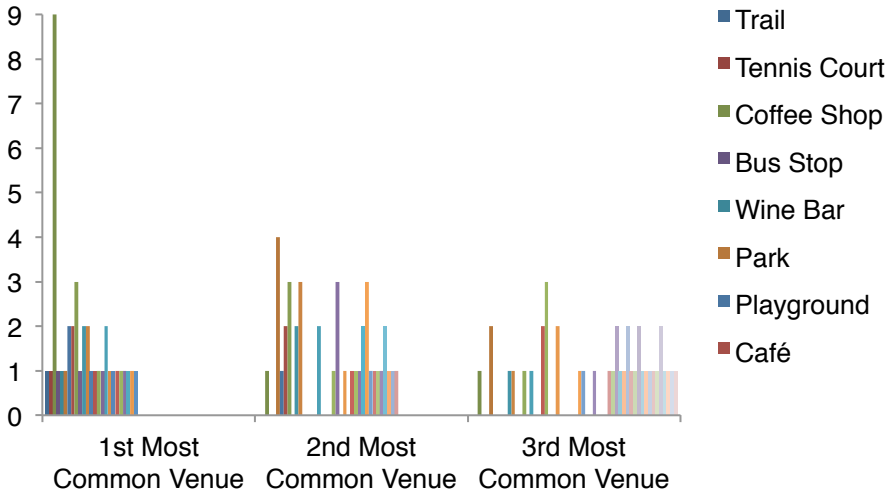
Chart 3 Most common venues in the Green Cluster



The Green Cluster represents homes that have access to Baseball fields and playgrounds as their most common venues.

- Purple Cluster – Coffee shops and everything else

Chart 4 Most common venues in the Purple Cluster



The purple cluster represents homes that have such a high density of venues nearby that it's difficult to describe. The most salient feature of this cluster is the abundance of coffee shops followed by many other types of venues.

- Aquamarine and Blue Clusters

These clusters are very small and boast unique venues such as motorcycle and electronic shops.

4.2 Crime-based analysis

The k-means clustering algorithm resulted in five clusters based on reported crimes. Each cluster is represented by a different color on the map in Figure 4.

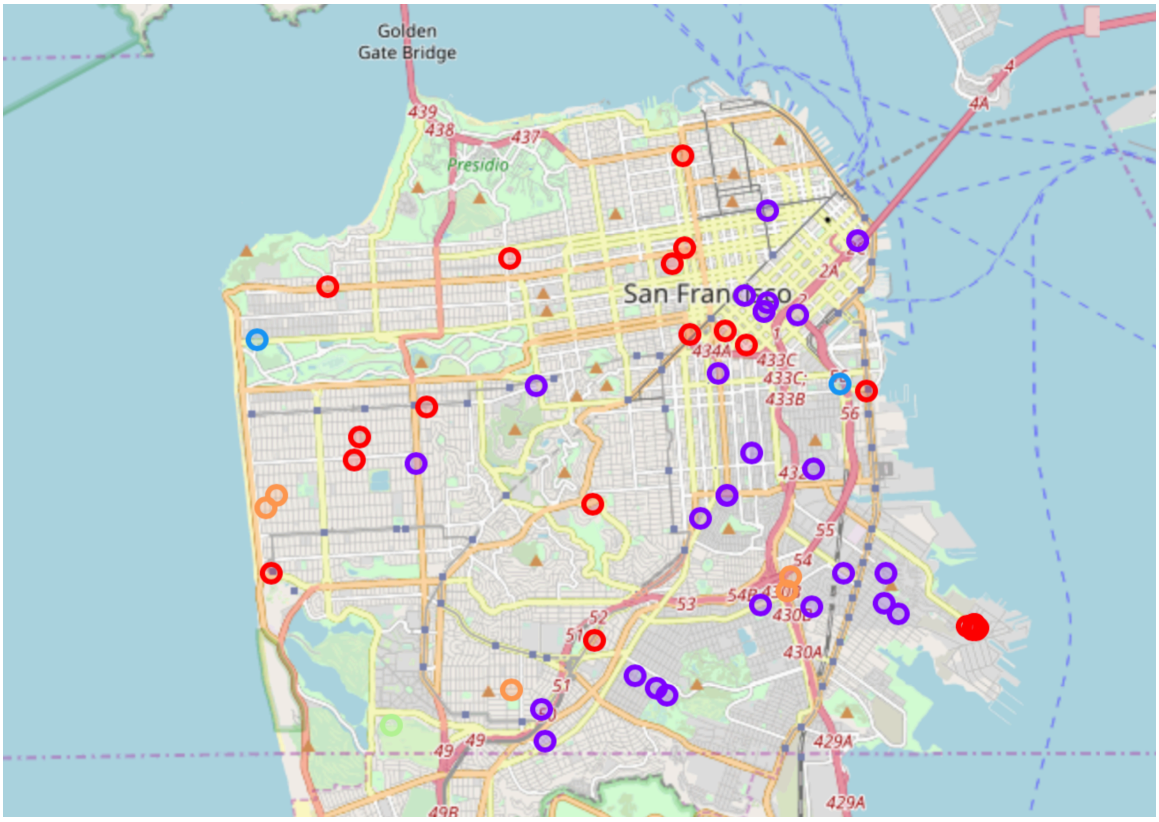
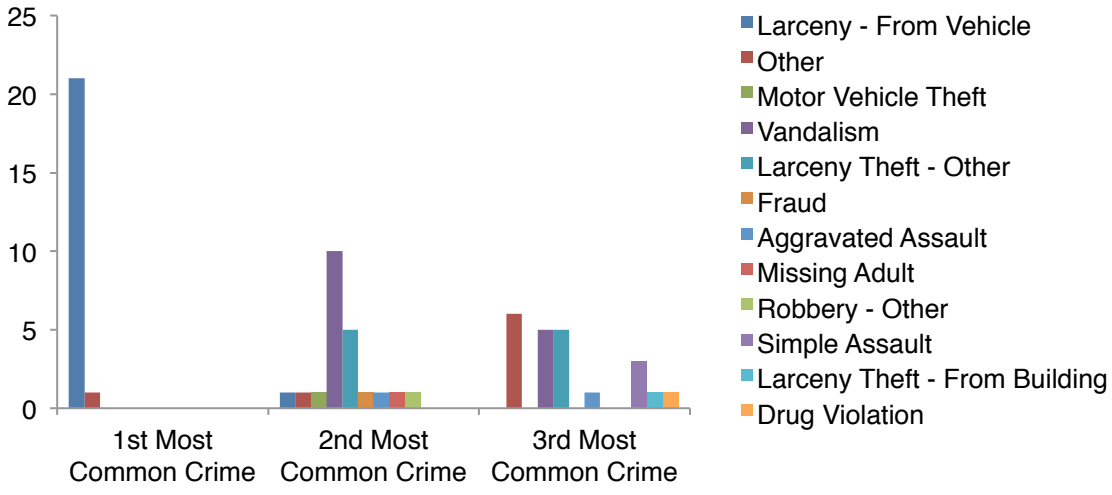


Figure 4 Clusters of homes according to crime statistics

- Red Cluster – Larceny theft from vehicle

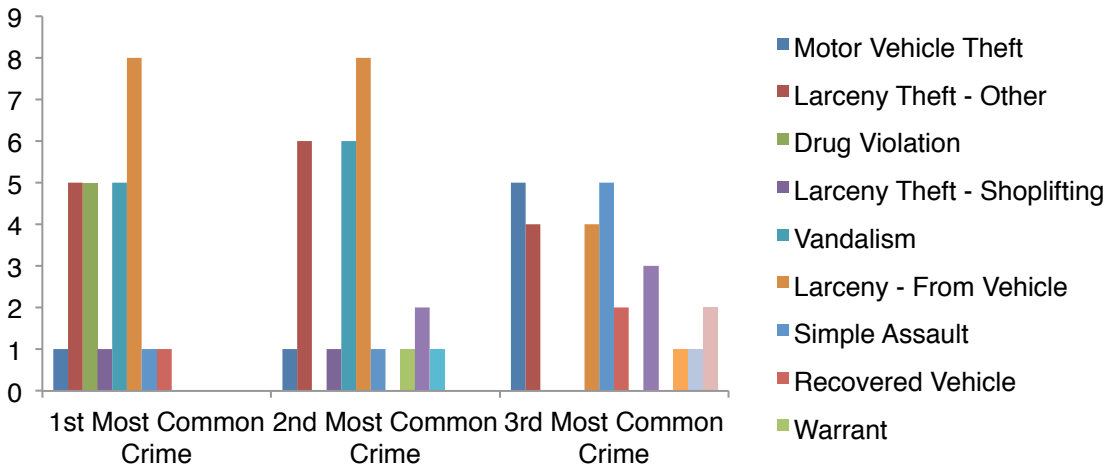
Chart 5 Most common crimes in the Red Cluster



Larceny theft from vehicle occurs overwhelmingly near the homes that belong to the Red Cluster; it's the most common crime in those areas by far. The second most common crime is Vandalism with a variety of other crimes occurring to a lesser extent.

- Purple cluster - Larceny theft from vehicle mixed in with a variety of others

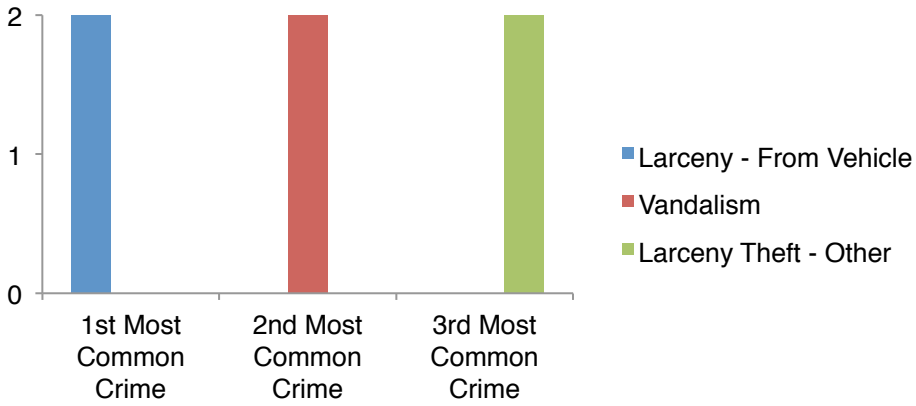
Chart 6 Most common crimes in the Purple Cluster



The Purple cluster is characterized by a large amount of Larceny theft from vehicle; it's the most and second most common crime. However a number of other types of crimes are also committed frequently such as Vandalism, Other Larceny as well as Drug Violations.

- Blue cluster - Larceny theft from vehicle and Vandalism

Chart 7 Most common crimes in the Blue Cluster



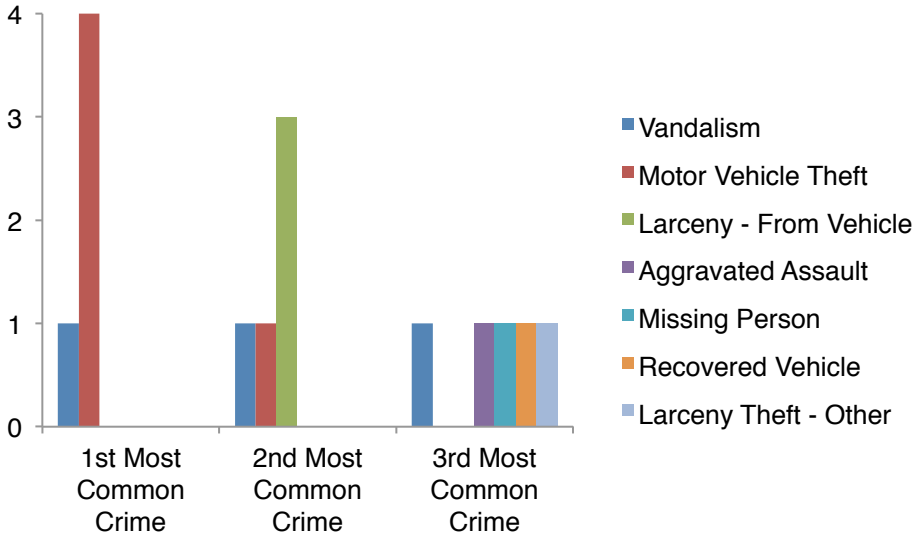
The Blue Cluster contains only two homes (at opposite ends of the city) and they have the exact same crime pattern for most common crimes. This could be a fluke since only February data is taken into account here and the common crimes exhibited are very similar to those which occur in the Red and Purple clusters.

- Green Cluster – Simple Assault

The Green cluster contains only one home but the crime pattern is different enough from the rest that the algorithm gave it its own cluster. The only three reported crimes here were Simple Assault, Vandalism and Weapons Offense (no chart needed to show this). This home has a very low crime rate (only three instances in February) so more data would be needed to know if violent crimes occur often.

- Orange Cluster - Motor Vehicle Theft

Chart 8 Most common crimes in the Orange Cluster



The most common crime in the Orange Cluster is Motor Vehicle theft followed by Larceny from vehicle. This does not mean, however, that motor vehicle theft does not occur frequently in other areas. In other areas other types of thefts are even more frequent so Motor Vehicle Thefts get diluted and do not make it into the top three crimes.

5. Discussion

Two large clusters emerged from both venue and crime based clustering (red and purple for both). These clusters are heavily weighted towards the Eastern side of the city and both have tremendous diversity of venues, and crime. This is especially true for the purple cluster in both cases. The purple venue cluster was described as “Coffee and everything else” while the purple crime cluster was described as “Larceny theft from vehicle mixed in with a variety of others”. Homes in the purple cluster have a variety of venues to enjoy, however this high level of activity brings with it a high level and variety of criminal activity. The logical assumption is that the purple cluster’s proximity to notorious neighborhoods like the Tenderloin accounts for this criminal activity.

Homes that belong to the red cluster for crime ("Larceny theft from vehicle") appear to be farther away from the central axis of criminal activity. In this case one's car is more likely to be broken into than other types of crime that happen in the purple cluster with more frequency.

On the subject of cars, the orange crime cluster ("Motor vehicle theft") makes sense because these homes are very far away from where the epicenter of criminal activity lies so the cars are stolen rather than broken into (the thieves need a way to get back home). Turning to venues, the homes in the orange crime cluster overlap somewhat with the green venue cluster ("baseball and playgrounds"). These homes would be good for families that enjoy playing outside but do not have a car, or do not mind having it stolen!

Even though the crime data was only for February of 2020; it is interesting to see relatively well-delineated patterns across clusters. Further study would be required with larger data sets to see if those patterns hold. My hunch is that the patterns will largely hold.

6. Conclusion

Fifty seven homes that were for sale in San Francisco were clustered based on their proximity to venues and crime incidents reported.

This project showed that unsupervised learning can be used as a complementary tool for homebuyers to be able to make an informed, data driven decision.

This tool can be refined and developed so that the user can choose any home for sale and see what type of cluster it belongs to based on the venues that are nearby as well as crime statistics. This would make comparing homes very easy and convenient.

Such a tool would be particularly useful in a city like San Francisco where nearby venues are important since it's such a walkable city. Unfortunately, any homebuyer must also take care to understand what types of crimes he or she would have to potentially deal with on the way to those venues or at home.