Capstone Project – The Battle of Neighborhoods

San Francisco Evictions – Data Analysis Report

Prepared By: Sujith Kollath

August, 2020

Table of Contents

1. Introduction

1.1 Overview and Analytical Approach

In a city like San Francisco with a population of 3.3 million[1] people and home to Silicon Valley, home evictions are getting very common. San Francisco has many areas, including the Bay Area, which are among the most densely populated areas in the United States. In these difficult times, through my project I wanted to explore and analyze the best neighborhoods for renting a house in SF, based on the probability of not getting evicted.

As mentioned, the population density in San Francisco makes is very challenging to rent a house in a good neighborhood. Understanding the details of a good neighborhood is key to having a peaceful shelter, especially when the occurrences of home evictions carried out by landlords are increasing. One of the primary reasons for increased evictions can be factored to easy availability of tenants, willing to pay high rents. In a real estate market where the buyer power is less than the seller power, selecting the right neighborhood, property and landlord will reduce the chances of becoming homeless.

Analyzing the above-discussed scenarios by creating density maps and information tables of the past evictions in San Francisco will provide good insights into the underlying neighborhoods. By segmenting and clustering the neighborhoods based data, one can select the best-fit neighborhood to rent a house/apartment in San Francisco.

2. Data Description

2.1 Data Requirements

For data requirements, I identified websites that can provide details of San Francisco in a specified format (CSV, Excel,.json etc.). The data should include details of the neighborhoods and boroughs in San Francisco, CA. Also, optional data regarding postal/zip code will be useful for analyzing.

2.2 Data Collection

In the data collection stage, I found the details of websites or articles that provide details of San Francisco's real estate. The data I am particularly interested in was the renter's marketplace.

I have used the following data sets to analyze the problem statement discussed in the above section:

[1] https://en.wikipedia.org/wiki/San_Francisco

- From data.world[2] > Make Over Monday data sets, I downloaded the CSV file for the San Francisco Eviction Notices. During the data cleansing, I grouped the neighborhoods to Supervisor Districts[3] and mapped the probability of an eviction in San Francisco using a choropleth map

- I used Foursquare API to analyze the most common venues around the San Francisco county (in San Francisco there is no concept of Borough, instead I used counties to segment)

- I did not have to use CSV file for Geospatial data since the data set included latitude and longitude coordinates corresponding to each data entry

- I used the JSON file **san-francisco.geojson** to create the choropleth map. This JSON file was used while analyzing the San Francisco crime-related data during the data visualization lab[4]

2.3 Data Understanding and Preparation

In the data understanding stage, I checked if the details provided in the data sets collected are representative of the San Francisco tenant market. I prepared data by cleansing and removing duplicates, addressing missing values, and ensuring it is properly formatted.

3. Methodology

3.1 Modeling and Evaluation

The data set – San Francisco house eviction notice (CSV file) used for the data analysis consists of 41043 rows and 29 columns.

3.1.1 Histogram

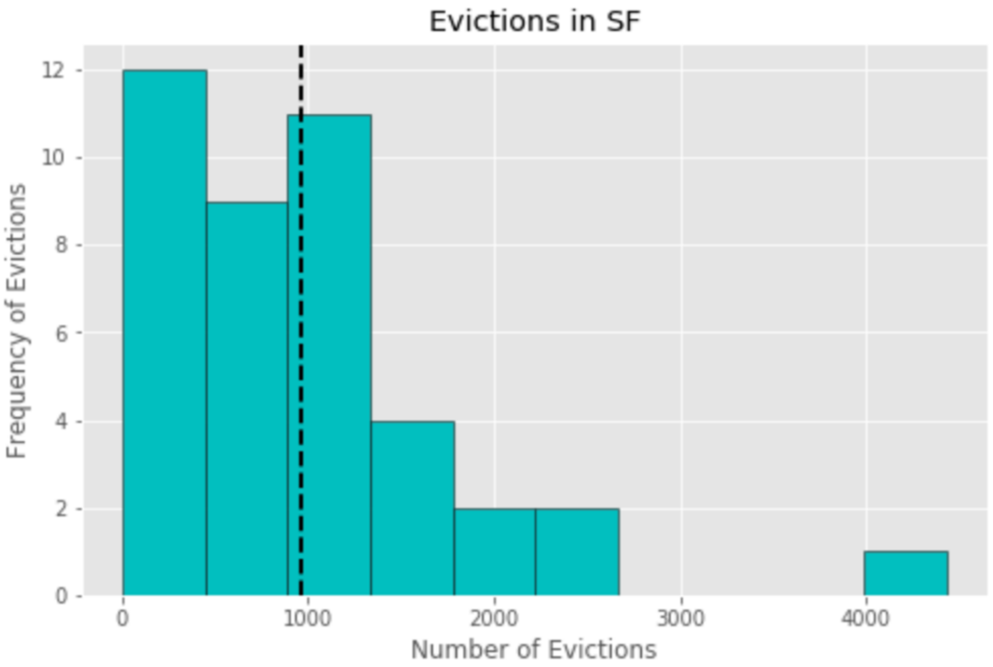
For my histogram analysis, I cleaned the data by removing duplicates and excluded rows and columns with missing data.

First step was to group the neighborhoods and count the number of evictions per neighborhood. The next step was to create bin ranges and using matplotlib library to create the below histogram.

[2] https://data.world/makeovermonday/2019w39

[3] https://en.wikipedia.org/wiki/San_Francisco_Board_of_Supervisors

[4] https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DV0101EN/DV0101EN-3-5-1-Generating-Maps-in-Python-py-v2.0.ipynb?lti=true



I categorized the bins for the above histogram as:

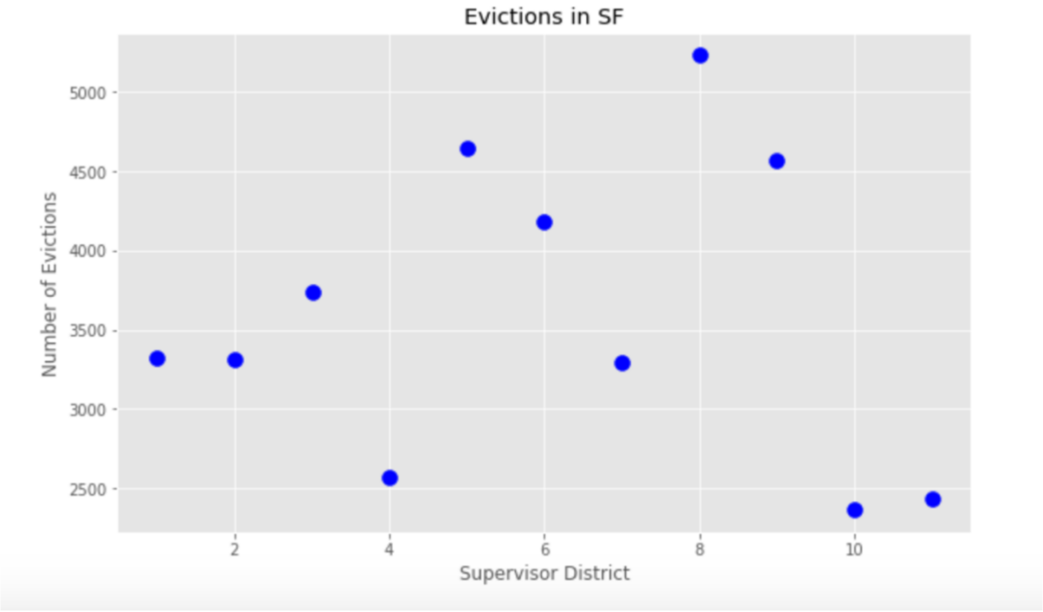
- Bin 1 (0–1000): Low number of evictions
- Bin 2 (1000–2000): Medium-I number of evictions
- Bin 3(2000–3000): Medium-II number of evictions
- Bin 4(3000–4000): High number of evictions
- Bin 5(> 4000): Very high number of evictions

On analyzing the above histogram, it is clear that the evictions range from 0 to ~4000 in different neighborhoods. With the average number of evictions (shown by dotted lines) approximately equal to 1000. More number of neighborhoods have evictions in the range of 0 to ~ 1500 and only one neighborhood has an eviction count of more than 4000.

3.1.2 Scatter Plot

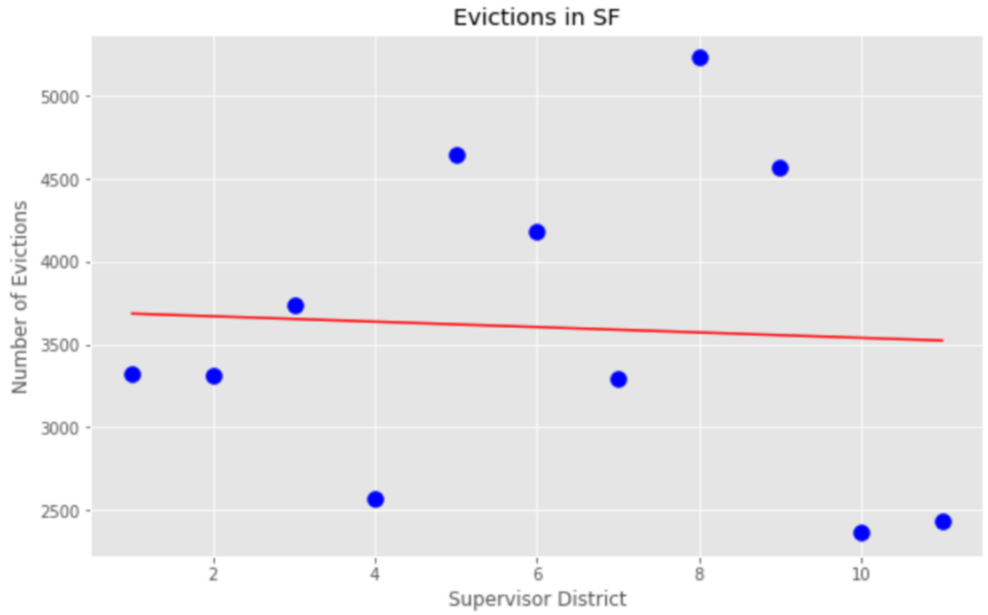
I used scatter plotting to analyze the relationship between the Supervisor district (i.e. Borough) and the number of evictions. To plot, I grouped the data by Supervisor district and counted the number of evictions in each Supervisor District/Borough.

By plotting the SF Supervisor districts in the x-axis and the number of evictions in the Y-axis, the below scatter plot is displayed.



On further analysis, we can observe that the relationship between X (Supervisor district/Borough) and Y (Number of Evictions) seems to be nonlinear.

To confirm my observation, I calculated the equation for the best fit line and plotted as shown below.



The best fit line has a negative slope.

The equation of the line is:
Number of Evictions = -16.26363636 * Supervisor District + 3702.58181818

The flat best fit line clearly depicts that there is no relationship between Supervisor Districts and the Number of evictions. Based on the above scatter plot, we can conclude that number of evictions does not depend on the Supervisor District. In other words, we cannot classify a Supervisor District to be bad or good for not getting evicted by landlords.

3.1.3 Choropleth Map

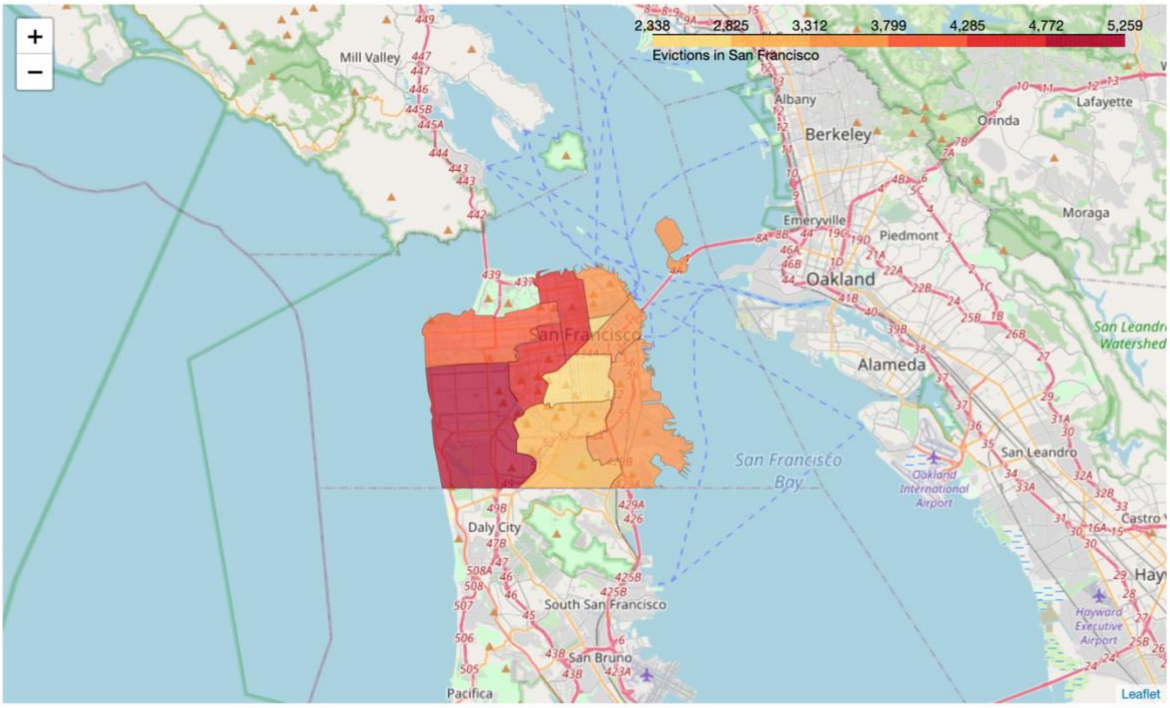
I utilized the choropleth map to visualize the evictions by Supervisor Districts/Boroughs.

The map displays the following information:

- Supervisor Districts/Boroughs in San Francisco
- Legend to identify the number of evictions by different colors
- Range of evictions

If higher the number of evictions in a Supervisor District, darker the color on the map. On further analysis, I observed that more areas have a medium number of evictions depicted by orange color.

Therefore, for supplementary evaluation, I selected a Supervisor District with an average number of evictions.



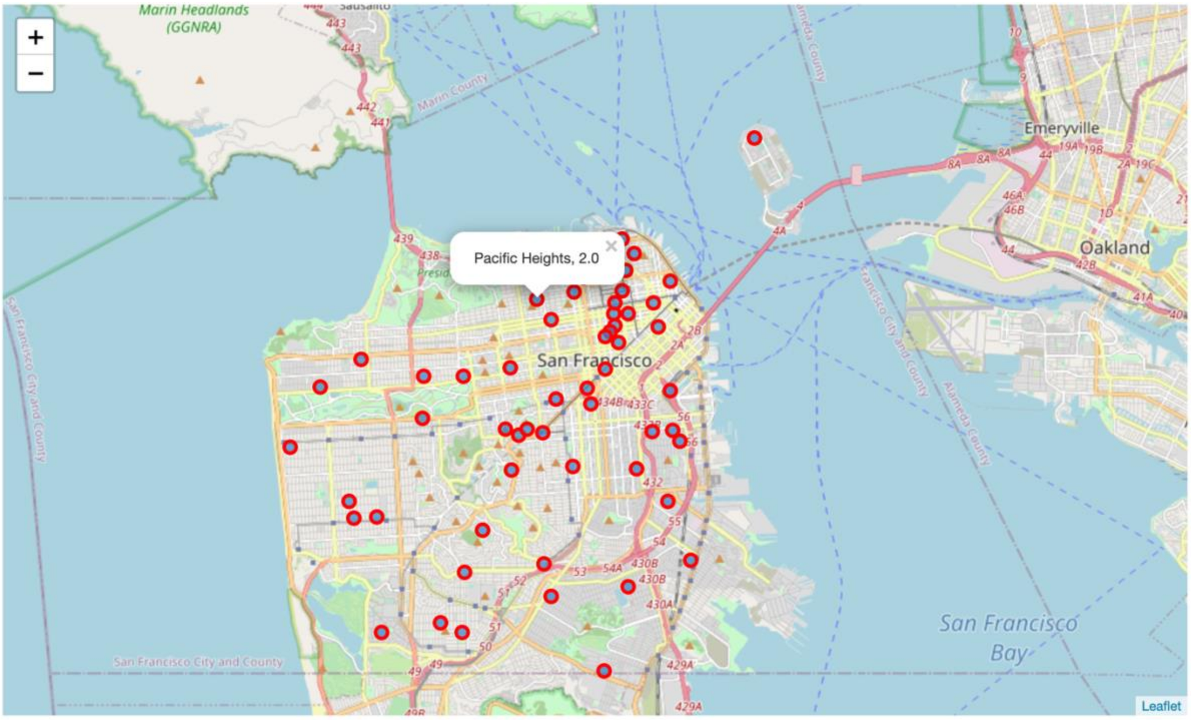
3.1.4 Foursquare API — Segmenting and Clustering

To start segmenting and clustering, I installed the geopy library to get the latitude and longitude coordinates.

However, instead of Geospatial data, I used the location coordinates in my data set, split, and saved it into latitude and longitude for each data entry.

Firstly, the initial consideration was to explore the neighborhoods of San Francisco city. To create the below map, I used folium library and passed the details of the coordinates of each neighborhood.

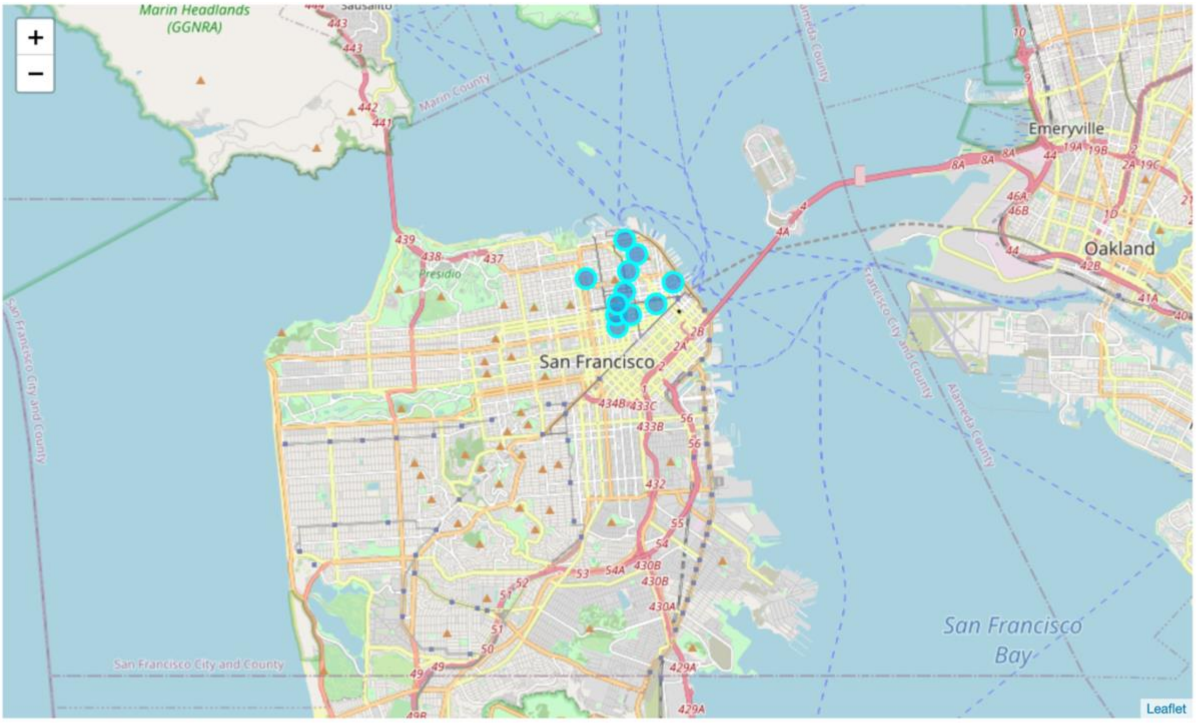
After dropping the duplicates and cleaning the data, the 59 neighborhoods with evictions are superimposed as shown below.



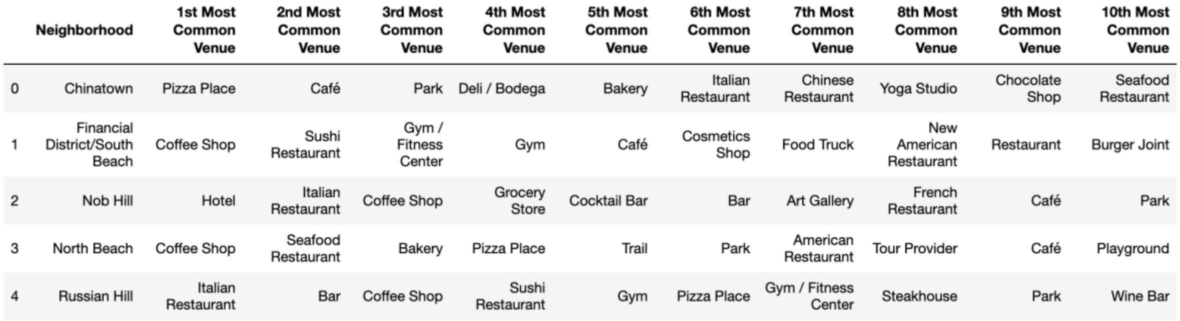
Secondly, based on my previous results, I selected a Supervisor District/Borough (i.e. County in SF) with an average rate of evictions. During my research, I instituted that San Francisco County is the county corresponding to Supervisor District 3[5]

On superimposing the 11 neighborhoods with house evictions for Supervisor District 3 (San Francisco County), I received the below folium map.

[5] https://en.wikipedia.org/wiki/San_Francisco

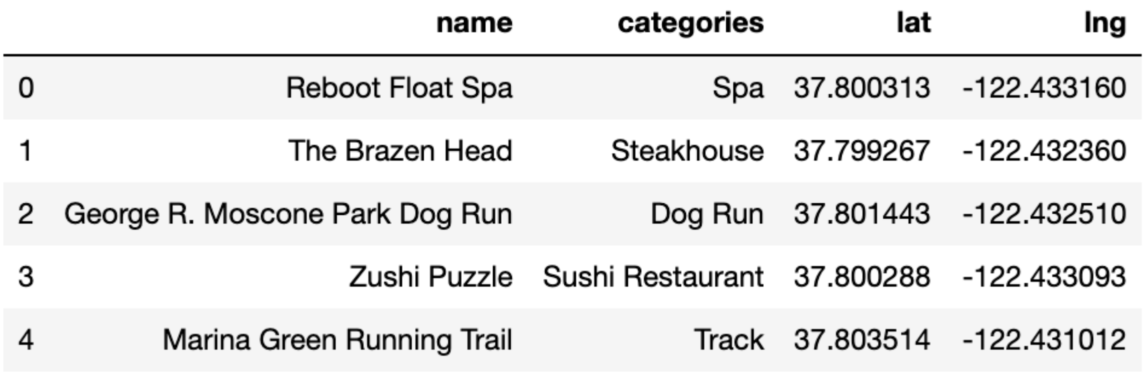


Finally, Foursquare API was used to find the most common nearby venues of the neighborhoods. The top 10 venues in each neighborhood in San Francisco county are shown below.



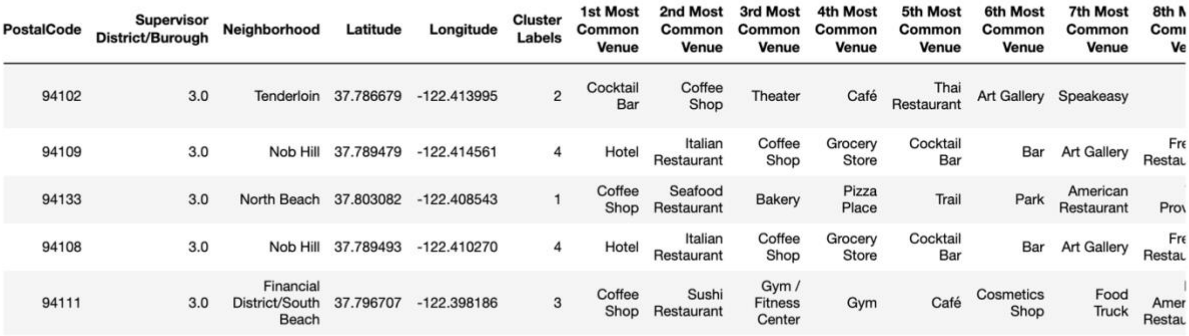
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Chinatown	Pizza Place	Café	Park	Deli / Bodega	Bakery	Italian Restaurant	Chinese Restaurant	Yoga Studio	Chocolate Shop	Seafood Restaurant
1	Financial District/South Beach	Coffee Shop	Sushi Restaurant	Gym / Fitness Center	Gym	Café	Cosmetics Shop	Food Truck	New American Restaurant	Restaurant	Burger Joint
2	Nob Hill	Hotel	Italian Restaurant	Coffee Shop	Grocery Store	Cocktail Bar	Bar	Art Gallery	French Restaurant	Café	Park
3	North Beach	Coffee Shop	Seafood Restaurant	Bakery	Pizza Place	Trail	Park	American Restaurant	Tour Provider	Café	Playground
4	Russian Hill	Italian Restaurant	Bar	Coffee Shop	Sushi Restaurant	Gym	Pizza Place	Gym / Fitness Center	Steakhouse	Park	Wine Bar

On exploring the top 100 venues in a radius of 500m of San Francisco county, Foursquare returned 74 venues under 144 unique categories.



	name	categories	lat	lng
0	Reboot Float Spa	Spa	37.800313	-122.433160
1	The Brazen Head	Steakhouse	37.799267	-122.432360
2	George R. Moscone Park Dog Run	Dog Run	37.801443	-122.432510
3	Zushi Puzzle	Sushi Restaurant	37.800288	-122.433093
4	Marina Green Running Trail	Track	37.803514	-122.431012

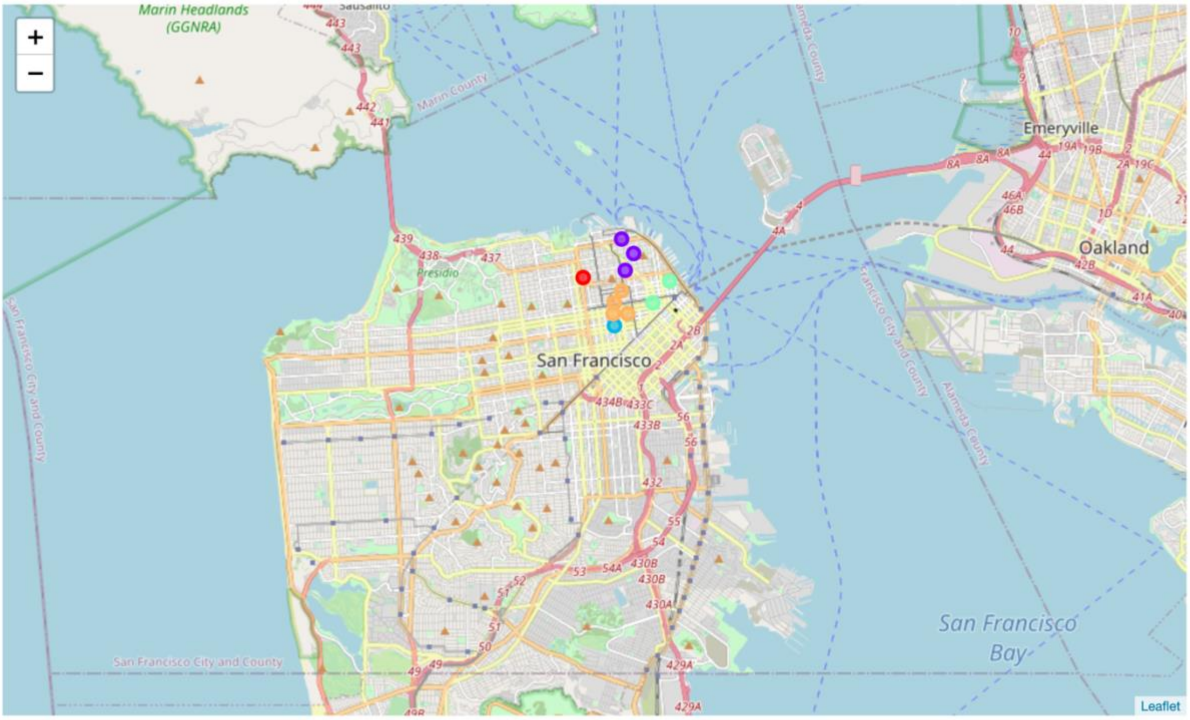
K-Means algorithm, which is an unsupervised learning method was used to cluster the neighborhoods. I chose to use K nearest neighbors equal to 5 for clustering the 11 identified neighborhoods.



PostalCode	Supervisor District/Burough	Neighborhood	Latitude	Longitude	Cluster Labels	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 M Com Ve
94102	3.0	Tenderloin	37.786679	-122.413995	2	Cocktail Bar	Coffee Shop	Theater	Café	Thai Restaurant	Art Gallery	Speakeasy	
94109	3.0	Nob Hill	37.789479	-122.414561	4	Hotel	Italian Restaurant	Coffee Shop	Grocery Store	Cocktail Bar	Bar	Art Gallery	Fre Restau
94133	3.0	North Beach	37.803082	-122.408543	1	Coffee Shop	Seafood Restaurant	Bakery	Pizza Place	Trail	Park	American Restaurant	Prov
94108	3.0	Nob Hill	37.789493	-122.410270	4	Hotel	Italian Restaurant	Coffee Shop	Grocery Store	Cocktail Bar	Bar	Art Gallery	Fre Restau
94111	3.0	Financial District/South Beach	37.796707	-122.398186	3	Coffee Shop	Sushi Restaurant	Gym / Fitness Center	Gym	Café	Cosmetics Shop	Food Truck	Amer Restau

I divided and named the clusters depending on the most common venues as:

- Cluster 0: Restaurant cluster
- Cluster 1: Fast food cluster
- Cluster 2: Recreational cluster
- Cluster 3: Lifestyle cluster
- Cluster 4: Hotel cluster



4. Results

1. The histogram plotting empowered me to get better insights into the number of evictions in each neighborhood
2. The scatter plot and best fit line (with a slight negative slope) showed a lack of relationship between Supervisor Districts/Boroughs and evictions
3. I was able to visualize the color-coded Supervisor Districts and the corresponding evictions using the Choropleth maps
4. Analysis of the Foursquare API common venues gave me better insights into the neighborhoods of San Francisco county (i.e. Supervisor District = 3.0)

5. Discussion

In the study, I analyzed the neighborhoods in San Francisco to understand the implications of increased house evictions in the city. I identified that the evictions are not directly related to neighborhoods or Supervisor Districts/Boroughs. However, in my analysis, I was able to classify the neighborhoods with a high rate of evictions from others. These findings will be useful for a newcomer to the city when selecting a neighborhood to rent a house.

By segmenting and clustering San Francisco county (i.e. Borough) with an average rate of evictions, it was evident that most of its neighborhoods have the necessary amenities for a new settler. The clustering distinguished the area to restaurants, stores, health - lifestyle, and hotel, which are the basic needs people will look for when they choose a neighborhood to rent or buy a house.

6. Conclusion

Through this project, I evaluated the chances of evictions in each neighborhood of San Francisco. The insights provided by the analysis will enable a new renter to carefully evaluate his options before renting an apartment or house in San Francisco.

To conclude, with the data analysis completed in this project, a stakeholder new to the San Francisco city can easily choose the best neighborhood to rent a house with minimal chances of evictions.