The Battle of the Neighborhoods: San Francisco, CA

I. Introduction

San Francisco, California is consistently ranked as one of the best cities for finding a new job. However, it is also one of the most expensive cities in the United States. Many young professionals moving there will not have the savings necessary to purchase real estate. Instead, they will likely rent. Additionally, even those further along into their career may prefer to rent rather than make the commitment of buying a home, especially if they do not plan to stay in San Francisco in the long term.

In this project, I will explore the neighborhoods of San Francisco. I want to find which neighborhoods have the cheapest rent, and which of those have attractive venues nearby so newcomers will not have to sacrifice entertainment and convenience for cheaper rent. I will use the Foursquare API's "explore" function to get the recommended venues around each neighborhood. I plan to examine not only the amount of nearby venues, but also the varieties of venues for each neighborhood to ascertain that neighborhood's desirability.

In short, my goal is to find the most desirable neighborhood for renters in San Francisco based on both average rent and popular surrounding venues.

II. Data

A. RENTCafé's San Francisco, CA Rental Market Trends

RENTCafé tracks trends in the rental market such as overall average rent, year-over-year change, and average apartment size. In particular, I am interested in the table titled "Average Rent in San Francisco, CA By Neighborhood," which I will

webscrape for my project to determine the cheapest neighborhoods in San Francisco.

B. Foursquare API

After webscraping to find the average rent in each neighborhood, I will use the Foursquare API to get the recommended surrounding venues for each of the cheaper neighborhoods. From this data, I can also pull the categories to see the types of venues in each neighborhood. I am looking not only for a neighborhood with many venues, but with diverse types of venues (for example, a neighborhood that contains multiple grocery stores may not be as desirable as one that contains fewer grocery stores, multiple restaurants, and a museum). Based on this, I will make a suggestion regarding the best neighborhood for renters.

III. Methodology

For the first part of my project, I needed to find out the average rent of San Francisco neighborhoods. RENTCafé publishes data on the rental market trends for many U.S. cities, and so I was able to web scrape the table of average monthly rent by neighborhood using the BeautifulSoup package.

After arranging the data into a Pandas dataframe, I wrote a for loop using the GeoPy library's geolocator feature to get the coordinates for each neighborhood. To accomplish this, I had to clean up the data a bit, as GeoPy did not recognize some of the neighborhood names RENTCafé used. Then, using the Folium library, I visualized the neighborhoods on a map of San Francisco to get a better idea of their whereabouts.

Once I had my map, I was ready to utilize the FourSquare API. First, I explored the venues around the first city in my dataframe, Treasure Island. I created a GET request URL to return the data for Treasure Island, and then defined a function to extract the category type from each venue in the Foursquare dataset. I cleaned up the JSON data and structured it into a Pandas dataframe, and printed the amount of venues Foursquare returned for Treasure Island. Then, I iterated the process throughout each of the neighborhoods in the dataframe. Using the Pandas groupby function, I got both the number of venues in each neighborhood (capped at 100 venues), and the number of unique venue categories in each neighborhood to see which had most venue types.

Next, I examined the top most common venues in each neighborhood. Using the one-hot encoding method, I converted the categorical variables to numerical ones. Again using the Pandas groupby function, I took the average frequency of occurrence for each category in each neighborhood. With this new dataframe, I was able to run a for loop to get the top 5 most common venue categories for each neighborhood. I then ran a function to sort the venues from most common to least, and put the 10 most common venue types for each neighborhood in a dataframe.

Lastly, I wanted to compare the venue types of each neighborhood, and find which neighborhoods were similar. To do this, I used *k*-means clustering and broke the data up into four clusters. Once again, I used the Folium library to visualize a map of the neighborhoods, this time color-coded based on their clusters.

IV. Results

Through my analysis, I found that the ten cheapest neighborhoods were, respectively, Treasure Island, Van Ness - Civic Center, Tenderloin, Downtown District 8 - Northeast, Marina, Hayes Valley, Russian Hill, Nob Hill, Western Addition, and Bernal Heights. The table containing each neighborhood's average rent is displayed here:

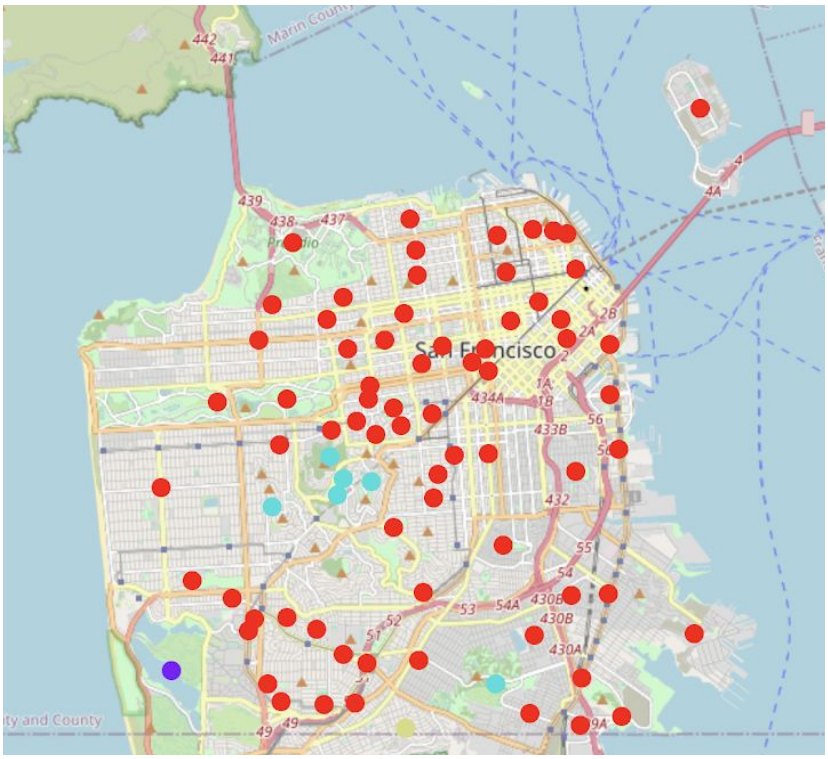
	Neighborhood	Average Rent
0	Treasure Island	$2,616
1	Van Ness - Civic Center	$2,944
2	Tenderloin	$2,944
3	Downtown District 8 - North East	$2,956
4	Marina	$2,974
5	Hayes Valley	$2,983
6	Russian Hill	$2,983
7	Nob Hill	$2,996
8	Western Addition	$3,030
9	Bernal Heights	$3,061

The neighborhoods with the most venues overall were Yerba Buena, Cow Hollow, Downtown District 8 - Northeast, Tenderloin, Sunnyside, Hayes Valley, South Beach, Inner Mission, and Marina. Each of these neighborhoods hit the 100 venue limit.

Sunnyside and Inner Mission were the neighborhoods with the most unique venues at 68 unique categories each. They are followed closely by Cow Hollow and Marina at 66 unique categories and South of Market at 65 unique categories.

Once the neighborhoods were broken up into four clusters, I found that the first, made up of the red dots on the map, was the largest. It contains the neighborhoods with

diverse category types. The second cluster, marked by the purple dot on the map, contained only Lake Shore. This is likely because the only venue type within Lake Shore is the lake itself. Likewise, the fourth cluster, marked by the chartreuse dot which almost blends into the bottom of the map, contains only Crocker Amazon, a neighborhood whose only venues are liquor store and dog run. The third cluster, turquoise on the map, is made up of neighborhoods whose primary venues are trails and parks. The map is shown below:



Something to be noted about the clusters is that, when displayed as a dataframe, the remaining top venue columns for neighborhoods with fewer than 10 venues each appear to have a filler category type, and can therefore be misleading about the amenities in those neighborhoods.

V. Discussion

Based on my analysis, I would recommend Marina as the best neighborhood for renters in San Francisco. On average, the rent is only $358 more than the cheapest neighborhood, making it the 5th cheapest neighborhood in San Francisco. Furthermore, it is both one of the neighborhoods with the most venues overall as well as one of the neighborhoods with the largest amount of unique venue types.

Marina is the neighborhood that would allow renters both to rent a cheaper apartment and experience the same wider variety of venues that exist in more expensive neighborhoods.

VI. Conclusion

At only 46.27 mi^2, San Francisco is a small city compared to other famous ones, such as New York, London, Tokyo, etc. Because of this, it is not surprising that so many neighborhoods ended up in the second cluster. Nearby neighborhoods likely share some venues, making them more homogeneous in the eyes of the clustering algorithm. The benefit of this is that many cheaper neighborhoods have similar amenities to their more expensive counterparts, so long as you know where to look.