Exploring San Francisco Neighborhoods

by
Sree Lakshmi Gudreddi

Business Problem

Which neighborhood should I choose to live in San Francisco?

1. Problem Description and Background

San Francisco is the fourth populous city in the state of California located on the tip of a peninsula surrounded by the Pacific Ocean and San Francisco Bay. it is known for its iconic Golden Gate Bridge, cable cars, colorful Victorian houses and year round fog.

San Francisco is the city of cinematic, ethnic, and historic neighborhoods with plethora of galleries, boutiques, most alluring hiking trails and parks, local and stylish restaurants with bustling night life. Each neighborhood carries its own charm and attract young, urban professionals and family people who are ethnically diverse with Irish, Russian, Hispanic, Italian, and Chinese roots. Currently, San Francisco is a melting pot of diverse people which include families with children and dogs, young and urban professionals, tech workers, blue-collar workers, retired people, affluent people, artists, hipsters, surfers, students and homeless people.

How do people from various backgrounds choose a neighborhood to live in San Francisco? One answer is based on their life style i.e activities and interests. To achieve this people choose to live closer to their interested venues. For this reason, exploring San Francisco neighborhoods to find various venues in each neighborhood is necessary to help make better decision of choosing a neighborhood to live in.

In this project, we focus mostly on three groups of target audience: Families with children and dogs, young and urban professionals, and artists.

2. Data Description and Extraction

2.1 Data Description

For the San Francisco neighborhood data, a Wikipedia page exists that has all the neighborhood information in a tabular form. The tabular form includes four boroughs: Downtown, North of Downtown, Outside Lands, Western Additions, and Southern and 60 neighborhoods within them. This data is scraped from the Wikipedia page using Beautiful Soup. The page contents are wrangled, cleaned, and the borough and neighborhood names are read into a pandas data frame.

Next, for each neighborhood the geo spacial coordinates i.e. the latitude and longitude are obtained using geopy geo-coders library. For some reason, the geo-codes obtained were not available for few neighborhoods and I got few errors in the obtained values. For this reason, all the missing and error values are entered manually into a cvs file. The file is then read into a data frame and merged into the San Francisco neighborhood data frame. Now, we have a data frame with Borough, Neighborhood name, latitude, and longitude information.

2.2 Data Extraction

Using Foursquare API, a social location service, all the venues in the neighborhoods of San Francisco can be explored by making REST API calls. A request URL is created using the Foursquare credentials i.e the client id and client secret, the latitude and longitude of each neighborhood, the radius of area to explore in kilometers, and the limit value that sets the maximum number of venues returned. The limit is set to 100 for this study. The response of the GET request is a JSON file which has the list of venues Under each venue category, venues are listed with information including venue name, venue latitude and venue longitude. There are 246 unique categories with 1155 venues identified in San Francisco neighborhoods.

2.3 Feature Selection

The venue categories are selected from the unique category list that fit the life style of the targeted audience. The categories are chosen based on personal knowledge of the interests of the audience. The list is shown below in a tabular form.

Target Audience	Venue Categories
Families with Children and Dogs	Park', 'Garden', 'Trail', 'Library', 'Dance Studio', 'Music School','Skating Rink', 'Athletics & Sports', 'Soccer Field', 'Mini Golf','Bookstore','Church', 'Veterinarian','Pet Store','Shopping Mall','Supermarket'
Young and Urban Professionals	Light Rail Station', 'Bus Stop', 'Gym / Fitness Center', 'Gym', 'Pilates Studio', 'Yoga Studio', 'ATM' ,'Flower Shop', 'Boutique', 'Electronics Store' ,'Paper / Office Supplies Store', 'Comic Shop', 'Beer Bar', 'Cocktail Bar','Karaoke Bar' ,'Jazz Club', 'Nightclub'
Artists	History Museum', 'Art Museum', 'Art Gallery', 'Public Art', 'Sculpture Garden', 'Antique Shop' , 'Monument / Landmark', 'Outdoor Sculpture','Historic Site','Harbor / Marina', 'Beach', 'Mountain', 'Hobby Shop' ,'Arts & Crafts Store'

Category names with restaurants are purposefully eliminated from the feature set because there is no way of knowing the target audiences favorite restaurants. Note that these categories are not ignored when grouping the neighborhoods based on density of venues and based on proximity to similar venue categories using statistical methods while performing data analysis.

3. Exploratory Data Analysis

The goal of this analysis is to group the neighborhoods based on three criteria: proximity to interested venues of each group of target audience, density of venues, and proximity to similar venue categories. Since we don't have trained data, unsupervised learning methods are used to extract meaning from the labelled neighborhood data. Clustering is one of the unsupervised methods to divide data into different groups based on similarity and proximity. In this analysis, Hierarchical clustering, Density based clustering, and K-Means clustering methods are used to group the neighborhoods of San Francisco.

3.1 Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering method is a bottom up clustering approach used to group neighborhoods based on interested venue categories of each group of target audience. This clustering method produces a dendrogram which shows the clusters of neighborhoods that have the interested venues specified in the data feature set of the target audience. In this method we need not specify the number of clusters to produce and it is easy to implement.

3.1.1 Methodology

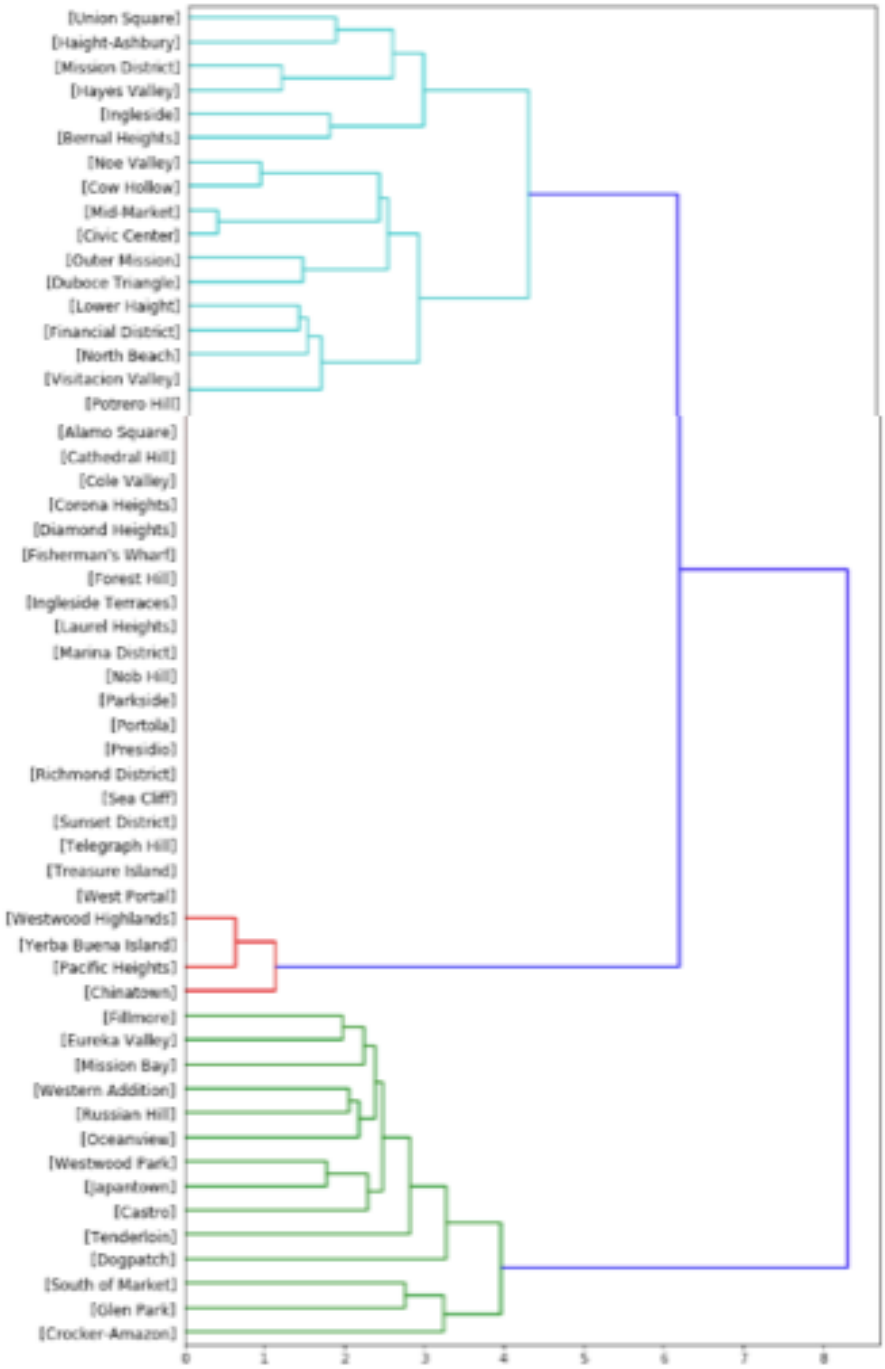
The data points are the neighborhood names and the dimensions are the venue categories specified in the feature set of each target audience. The data used has rows grouped by neighborhood with mean of frequency of occurrence of each category. The analysis is performed in the following steps.

1. First, the data is normalized using min-max scalar method so the values are with in the range 0-1.
2. Neighborhood proximity matrix is computed using Euclidian distance method and n x n matrix is computed where n is the number of neighborhoods.
3. In agglomerative clustering, at each iteration, the distance matrix is updated. The distance between the newly formed cluster and other clusters in the forest is specified in linkage variable to be 'Complete'. This means the maximum distance between two points in the cluster is measured.
4. The output is display in the form of a tree called a Dendrogram.

3.1.2 Analysis

The dendrograms produced show the clusters of neighborhood that have venue categories at close proximity for each target audience.

Dendrogram of target audience: Families with Children and Dogs



Dendrogram of target audience: Young and Urban Professionals

Dendrogram of target audience: Artists

3.2 DBSCAN

Density Based Spacial Clustering of Applications with Noise (DBSCAN) is the most popular density based clustering methods. This method used in the analysis because we have spatial data of the venues in each neighborhood of San Francisco. The DBSCAN method group neighborhoods based on the spacial location of the venues

using the venue latitude and venue longitude information. Clusters of arbitrary shape are formed with high density regions and the outliers are excluded from the clusters. Density is defined as the number of points, venues, within a specified radius. By analysis using this method all the venues closer to each other within that radius are grouped to form a cluster and also, there is no need to specify number of clusters. The clusters are visualized using basemap package, by transforming coordinates to map projections.

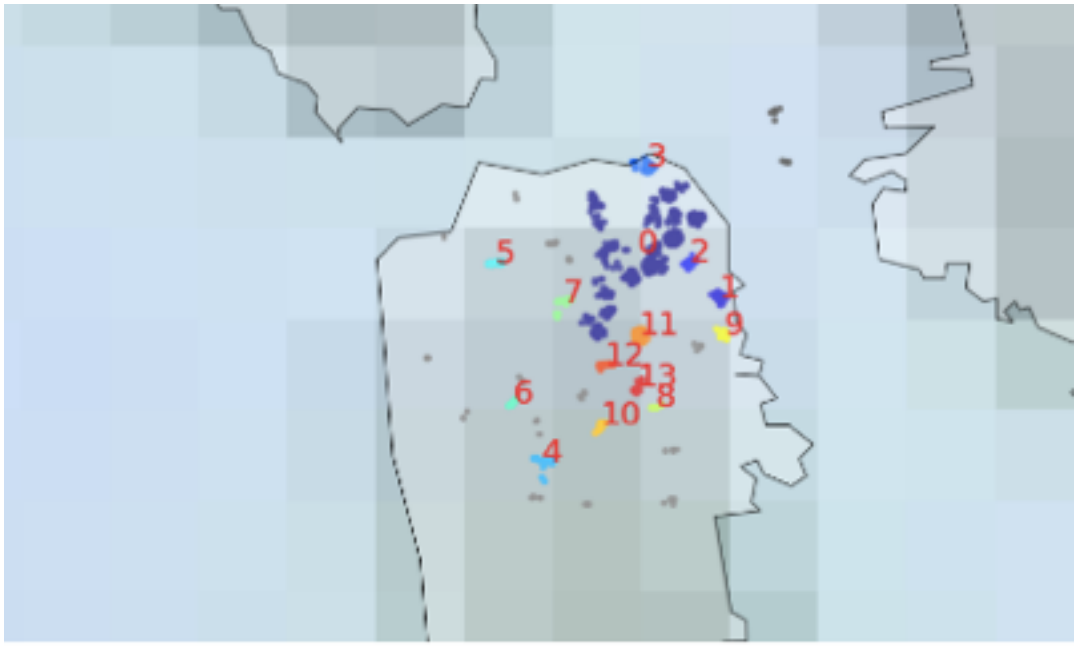
3.2.1 Methodology

The data set used contains the venue latitude and venue longitude.

- The data set contains the neighborhood name, venue latitude, venue longitude
- The data is standardized using StandardScalar so for each column the mean = 0 and standard deviation is 1.
- The epsilon and min_samples values are set to 0.25 and 10.
- The DBSCAN is applied to the standardized data set.
- The clusters are shown in color coded map.

3.2.2 Analysis

The color coded clusters are visualized on geographical map of San Francisco with clusters numbered from 0 to 13.



DBSCAN Clustering of Venue locations

3.3 K-Means Clustering

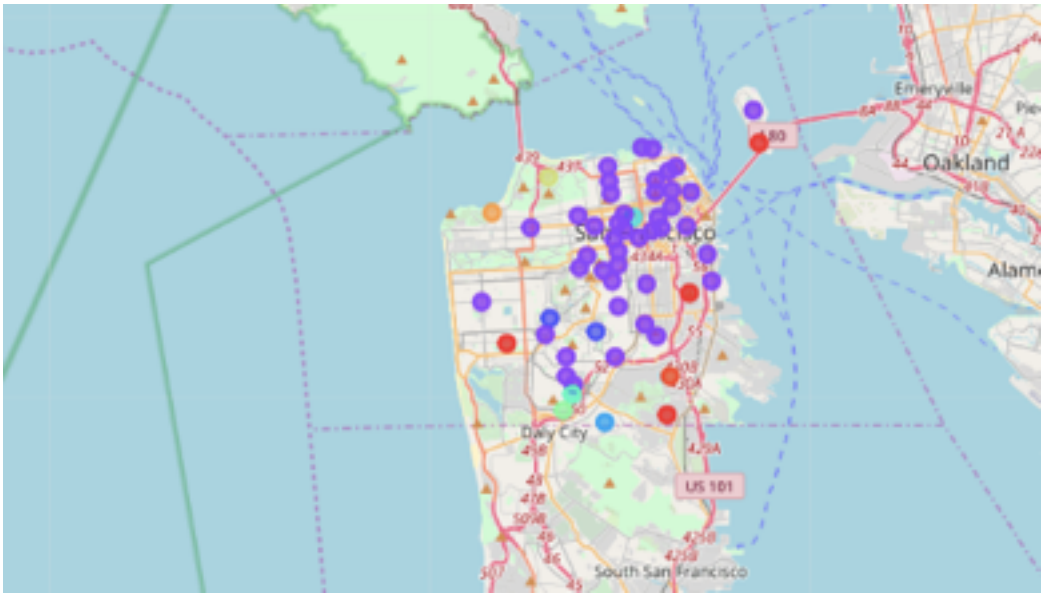
K-Means clustering method divides the data into K groups with similar records. Here, the neighborhoods are divided into clusters based on similar venue categories located in them. Here, we specify K, the number of clusters by obtaining it from the elbow methof. Since k-means algorithm does not work on categorical variables, the neighborhood rows are grouped and mean of the frequency of occurrence of each category is computed. The cluster labels are obtained after fitting the model.

3.3.1 Methodology

The data used has rows grouped by neighborhood with mean of frequency of occurrence of each category.

- First, the number of clusters K using the elbow method and set the value.
- K-means clustering is applied on the data set and the k-means labels are obtained.
- A sorted data set is created with top 10 common venues in each neighborhood.
- The cluster labels are merged with the sorted set.
- Each cluster is examined to determine the common venue in them.
- The clusters are visualized in a color coded map.

3.3.2 Analysis



K-Means Clustering of Venue Categories

4. Discussion

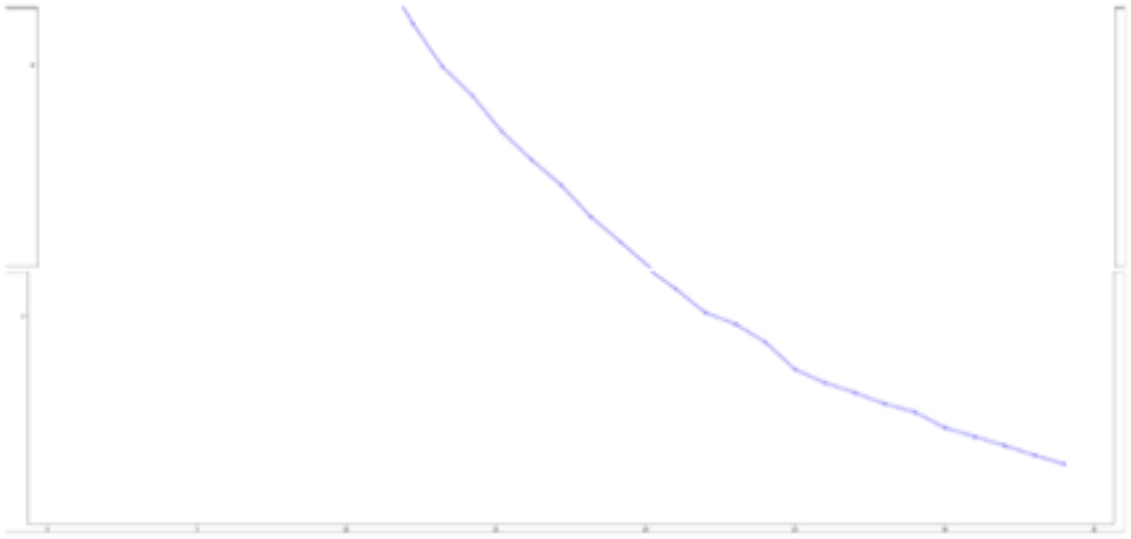
In this section, we discuss the results obtained from the statistical analysis of the San Francisco neighborhood data. The dendrograms obtained from the hierarchical clustering clearly shows the color coded clusters of neighborhoods suitable for the respective target audience. The leaves of the trees show the neighborhoods and the length of the branches show the degree of dissimilarity between the corresponding clusters. We can observe that the neighborhoods are grouped into three clusters for all the audience groups. The clusters are not of the same size but living in any neighborhood in the cluster will have access to the features venues in the other neighborhoods of the cluster.

Target Audience	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Families with Children and Dogs	Mission District Hayes Valley Haight-Ashbury Outer Mission Dogpatch Ingleside Bernal Heights Noe Valley Cow Valley Castro FIllmore Eureka Valley Russioan Hill Mission Bay Oceanview Crocker-Amazon Tenderloin Union Square South of Market Japantown Glen Park	North Beach Financial District Visitacion Valley Potrero Hill Lower Haight Duboce Triangle Mid-Market Civic Center	Pacific Heights Chinatown Alamo Square	Western Addition Hayes Valley

Target Audience	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Young and Urban professionals	Union Square Haight-Ashbury Mission District Hayes Valley Ingleside Bernal Heights Noe Valley Cow Hollow Mid-Market Civic Center Outer Mission Duboce Triangle Lower Haight Financial District North Beach Visitacion Valley	Westwood Highlands Yerba Buena Island Pacific Heights Chinatown	Fillmore Eureka Valley Mission Bay Western Addition Russioan Hill Oceanview Westwood Park Japantown Castro Tenderloin Dogpatch South of Market Glen Park Crocker-Amazon	
Artists	Ingleside Terrace Civic Center Nob Hill Sea Cliff Telegraph Hill Marina District Visitacion Valley Mid-Market Corona Heights Richmond District North Beach Mission District South of Market China Town Eureka Valley	Castro Alamo Square Duboce Triangle Noe Valley Cow Hollow	Pacific Heights Dogpatch Haight-Ashbury Fisherman's Wharf	

K-Means clustering method is used to group similar venue categories in each neighborhoods. Number of clusters, k can also be specified in K Means clustering. While finding the value of K using elbow method, it can be noticed that the line is linear. Which implies that there is one single large cluster with other small clusters. This proves that the venues are mostly concentrated in few neighborhoods with very limited venues in other neighborhoods. For example, the number of neighborhoods in each cluster are shown for different values of k. It can be observed that both have similar cluster sizes.

If k = 10, the largest cluster 1 has 43 neighborhoods, cluster 0 has 4 neighborhoods, cluster 2 has 2 neighborhoods and the rest have 1.
if k = 30, the largest cluster 29 has 18 neighborhoods, cluster 28 has 3 neighborhoods, cluster 4 has 6 neighborhoods, cluster 1 has 3 neighborhoods, and the rest of clusters have 1 neighborhood.



Plot of k and Sum of squared distance

Cluster Id	Count of Neighborhoods	Venue Category
0	4	Island/Park
1	43	Coffee Shop/Restaurant
2	2	Trail
3	1	Light Rail Station
4	1	Cupcake shop
5	1	Playground
6	1	Park
7	1	stables
8	1	Beach
9	1	Bus Station

To prevent the problem of uneven clustering, DBSCAN clustering method is used to group the neighborhoods where the radius of the cluster and the minimum number of data points within the cluster can be specified. Radius of 0.25 and min sample was set to 10. The number of clusters obtained in this method are 15 in total with one outlier. In this method, individual audience groups are not considered. Entire venue data is used in obtaining the clusters contrary to venue categories in hierarchical clustering. The

purpose of this analysis to show the grouping of neighborhoods based on the density of venues, venues that are closer to each other. This analysis is useful for all the target audience in deciding to pick a neighborhood closer to cluster of venues that interest them. The clusters are shown in the map in analysis section but a sample cluster with venues is shown here.

For example, cluster 4 includes the following venues:

487 Cluster 4, Venue: Whole Foods Market, Category...
488 Cluster 4, Venue: Philz Coffee, Category: Coff...
489 Cluster 4, Venue: Beep's Burgers, Category: Bu...
490 Cluster 4, Venue: Pakwan Restaurant, Category:...
491 Cluster 4, Venue: Pokihub, Category: Poke Place
492 Cluster 4, Venue: City College: Cafeteria, Cat...
493 Cluster 4, Venue: Sno-Crave Tea House, Categor...
494 Cluster 4, Venue: 21 Taste House, Category: As...
495 Cluster 4, Venue: Randy's Place, Category: Div...
496 Cluster 4, Venue: Quan Pho Viet, Category: Vie...
497 Cluster 4, Venue: Pizza Joint, Category: Pizza...
498 Cluster 4, Venue: K-Line, 29 Bus, & 91 Owl Sto...
499 Cluster 4, Venue: La Parilla, Category: Mexica...
500 Cluster 4, Venue: MUNI Bus Stop - Phelan Ave /...
501 Cluster 4, Venue: Wiley's No Limit Liquor & Fo...
502 Cluster 4, Venue: Orchids Cafe, Category: Cha ...
503 Cluster 4, Venue: Tea Me, Category: Café
504 Cluster 4, Venue: Wiley's Liquor, Category: Li...
505 Cluster 4, Venue: Ocean Ave. Mural, Category: ...
506 Cluster 4, Venue: Holloway Market, Category: L...
507 Cluster 4, Venue: The Gym @ Avalon Ocean Avenu...
508 Cluster 4, Venue: Snookies House, Category: Pl...
509 Cluster 4, Venue: merced heights eastern summi...
571 Cluster 4, Venue: Ocean Ale House, Category: G...
572 Cluster 4, Venue: Super Cue Cafe, Category: Bu...
573 Cluster 4, Venue: The Mayflower Restaurant, Ca...
574 Cluster 4, Venue: Taqueria El Jalapeño, Catego...
575 Cluster 4, Venue: Viking Giant Subs, Category:...
576 Cluster 4, Venue: Chase Luck Bakery, Category:...
577 Cluster 4, Venue: Relax Feet Spa, Category: Spa
578 Cluster 4, Venue: Sakesan Sushi | Robata, Cate...
579 Cluster 4, Venue: Hawaiian Sandwich Shop, Cate...
580 Cluster 4, Venue: Jolie Nail Spa, Category: Spa
581 Cluster 4, Venue: Go Go 7, Category: Korean Re...
582 Cluster 4, Venue: Copy Edge, Category: Paper /…

5. Conclusion

In this study, the neighborhoods of San Francisco are explored to provide insight for target audience to choose suitable neighborhood to live in. The suitable location is identified based on the proximity to their interested venue categories obtained from the Foursquare API. Various unsupervised statistical clustering methods are applied to the categorical data to group the neighborhoods based on proximity to interested venues, similarity of venue categories located, density of the venues. The output clusters are depicted using dendrograms and maps that show various neighborhoods with groups of color coded clusters for easy visualization. This analysis will definitely provide insight into various San Francisco neighborhoods that will help the target audience in finding the right neighborhood to live.

Though this study provides insight by exploring the neighborhood data there are other factors like the home price, population, schools etc. which will play a major role in choosing the right neighborhood to live in San Francisco.