Battle of the Neighborhoods in San Francisco, CA

Neeke Swart

April 14, 2021

(Photo: Paul Finnerty)

Capstone Project, ‘The Battle of the Neighborhoods’

The 'Applied Data Science Capstone' Coursera Course is part of IBM's Professional Certificate 'Data Scientist'

TABLE OF CONTENTS

1. Introduction

1.1 Background

The CBD, or cannabidiol, market is a fast-growing market which offers CBD in a broad range of products. Well-known product forms are oils/tinctures, lotions/balms, and gummies. Less known CBD infused products are bedsheets, bath bombs, and dog treats. According to Consumer Reports (2019) CBD products are primarily used to reduce stress, anxiety and joint pain (61%), and insomnia (10%). Recreational use comprises 10% of CBD use, and 33% of CBD-only users is 20-34 years old. The Brightfield Group (2017) expects the CBD market to grow to $25 billion by 2025, up from 4.67 million in 2017.

1.2 Client, goal, and problem

Currently the leading market in the US is California (Statista, 2021). Our client, a medium sized CBD seller with a broad range of products, has decided to open a store in San Francisco, CA. Before starting the search for a suitable location, the company asked us to make a data driven recommendation on an area that falls within the following guidelines (in order of importance):

1. The preferred area is a cluster of neighborhoods, although a single neighborhood is not rejected beforehand.
2. The dominant age group in the area is between 20 -34 years old.
3. The area has enough businesses to attract people that are interested health and/or are looking for fun. Examples are pharmacies with CBD, vitamins and supplements on their shelves, coffeeshops, restaurants and/or nightlife.

1.3 About this report

This report describes the subsequent steps followed in the research. The first section is the 'Data' section. Here the report shows how the data was gathered, cleaned and prepared. Next the 'Methodology and analysis' section describes which techniques are used to analyze the data and reports on the analysis. The outcome of the analysis is discussed in the 'Results and Discussion' section. Finally, some concluding remarks on this and on future research are made in the 'Conclusion' section.

2. Data acquisition and preparation

2.1 Data sources

Based on the guidelines above data was gathered regarding:

- Neighborhoods of San Francisco. Neighborhoods were defined by their zip code.
- Age distribution in the San Francisco neighborhoods
- Top 5 venues in the San Francisco neighborhoods

Following resources were used to get and process the information:

- The neighborhoods and their Zip Codes were scraped from the San Francisco Department of Public Health (2004).
- The latitudes and longitudes of the neighborhoods were added by geocoding using 'uszipcode' (zip code database in Python).
- The coordinates of San Francisco were obtained by using 'Geopy Library' (also Python).
- The age distribution per zip code were extracted from the American Community Survey (ACS), of the US Census Bureau (2019).
- The neighborhoods, the distribution of the 20-34 age group and the distribution of the venues were mapped using the Folium library.
- The top 5 venues per neighborhood, with their type and relative presence were obtained using Foursquare API.

2.2 Data preparation: cleaning and transforming

The data preparation involved four different datasets that each added features to the final set (see §2.3, table 1).

Zip codes and neighborhoods

The zip codes of San Francisco neighborhoods were, in csv format, scraped of the web (San Francisco Department of Public Health, 2004) and loaded into a dataframe. The dataframe counted 21 rows, which is in accordance with the number of neighborhoods.

Latitudes and Longitudes

Next, the latitudes and longitudes of the neighborhoods were added to the dataframe using ‘uszipcode’. The longitudes of ‘Outer Richmond’ and ‘Marina’ were incorrect. The label of Outer Richmond dropped into the sea, and the longitude value of Marina was the same as North Beach/Chinatown. Using the ‘Geopy Library’ the true values were determined and then corrected them in the dataframe. The Geopy Library was also used to find the latitude and longitude of San Francisco, these were used to make a map of San Francisco neighborhoods with the Folium library.

Age distribution per neighborhood

The age distribution per neighborhood was downloaded from the US Census Bureau (2019) in xlsx format. The first cleaning round was performed using Excel. The original data showed the absolute and proportional data distribution of each age group in each neighborhood. All the columns of age groups containing absolute numbers were dropped. Now, the table only contained proportional data on age groups. Next, several age groups were combined, which resulted in 5 age groups (see table 1). Lastly, the

table was transposed. The table was then read into a dataframe. Its shape was as expected: 21, 6 (21 zip codes, and 5 age groups). The first column was unnamed and contained zip codes. It was renamed 'Zip Code'.

Venues per neighborhood

With the use of Foursquare all the venues within a 500 meters radius of the neighborhood centers (according to Foursquare categorization) were collected and loaded into a new dataframe. The Foursquare information consisted of 223 unique categories. The data was normalized with the use of one-hot encoding. By grouping the venues together on venue type, the top 5 venues of each neighborhood were compiled. Finally, the total number of venues per neighborhood was computed and a column was added,

2.3 Data feature selection

The preparation of the four datasets described above resulted in 15 features (see table 1).

Table 1: Added features by each dataset during data preparation

Dataset	Features
Zip codes and neighborhoods	'Neighborhoods', 'Zip Codes'
Latitudes and Longitudes	'Latitudes', 'Longitudes'
Age distribution per neighborhood	'0-19 years', '20-34 years', '35-54 years', 55-64 years', '65+ years'
Venues per neighborhood	'Venues Total', '1st Most Common Venue', '2nd Most Common Venue', '3rd Most Common Venue', '4th Most Common Venue', '5th Most Common Venue'

3. Methodology and Analysis

3.1 Methodology

The collected and prepared data is analyzed in three steps:

1) Data exploration:

- San Francisco neighborhoods: this part sums up San Francisco neigborhoods and maps their location in San Francisco.
- Age distribution of 20-34 years per neighborhood: here the presence of the target age group, 20-34 years, in the neighborhoods is analyzed. Bar charts and a bubble map are used to visualize the neighborhoods where this group is dominant.
- Venue distribution per neighborhood: the distribution of venues is looked into with the use of a boxplot and a bubble map.

2) Age Venues Equation (AVE)

To express the importance of age over venues, the venues are weighted using the proportion of the 20-34 age group:

```
AVE = ('20-34 years' * 'Venues Total')/100
```

This equation results in a value between 0 and 1, the higher the value the more interesting this neighborhood will be for the client. The AVE distribution is depicted with a bubble map and a boxplot.

3) K-means Clustering

K-Means Clustering is used to find clusters of neighborhoods. The clusters will be mapped out with the Folium Library. Every cluster will be analyzed, looking at the AVE scores and the types of venues in the area.

3.2 Data exploration

3.1.1 San Francisco neighborhoods

The San Francisco Peninsula comprises 21 neighborhoods that are spread out over the upper half of the peninsula. The respective neighborhoods are summed up in table 2 and mapped out in figure 1.

Table 2: San Francisco neighborhoods and their Zip Codes

Zip Code	Neighborhood
94102	Hayes Valley/Tenderloin/North of Market
94103	South of Market
94107	Potrero Hill
94108	Chinatown
94109	Polk/Russian Hill (Nob Hill)
94110	Inner Mission/Bernal Heights
94112	Ingelside-Excelsior/Crocker-Amazon
94114	Castro/Noe Valley
94115	Western Addition/Japantown
94116	Parkside/Forest Hill
94117	Haight-Ashbury
94118	Inner Richmond
94121	Outer Richmond
94122	Sunset
94123	Marina
94124	Bayview-Hunters Point
94127	St. Francis Wood/Miraloma/West Portal
94131	Twin Peaks-Glen Park
94132	Lake Merced
94133	North Beach/Chinatown
94134	Visitacion Valley/Sunnydale

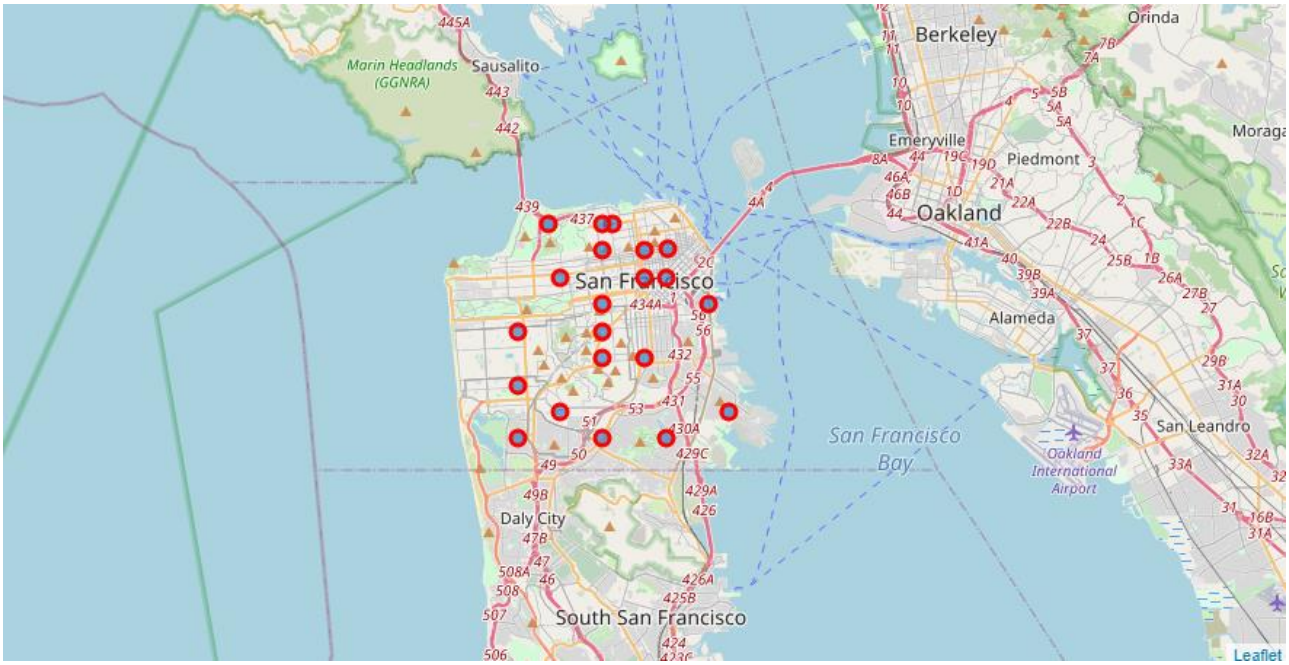


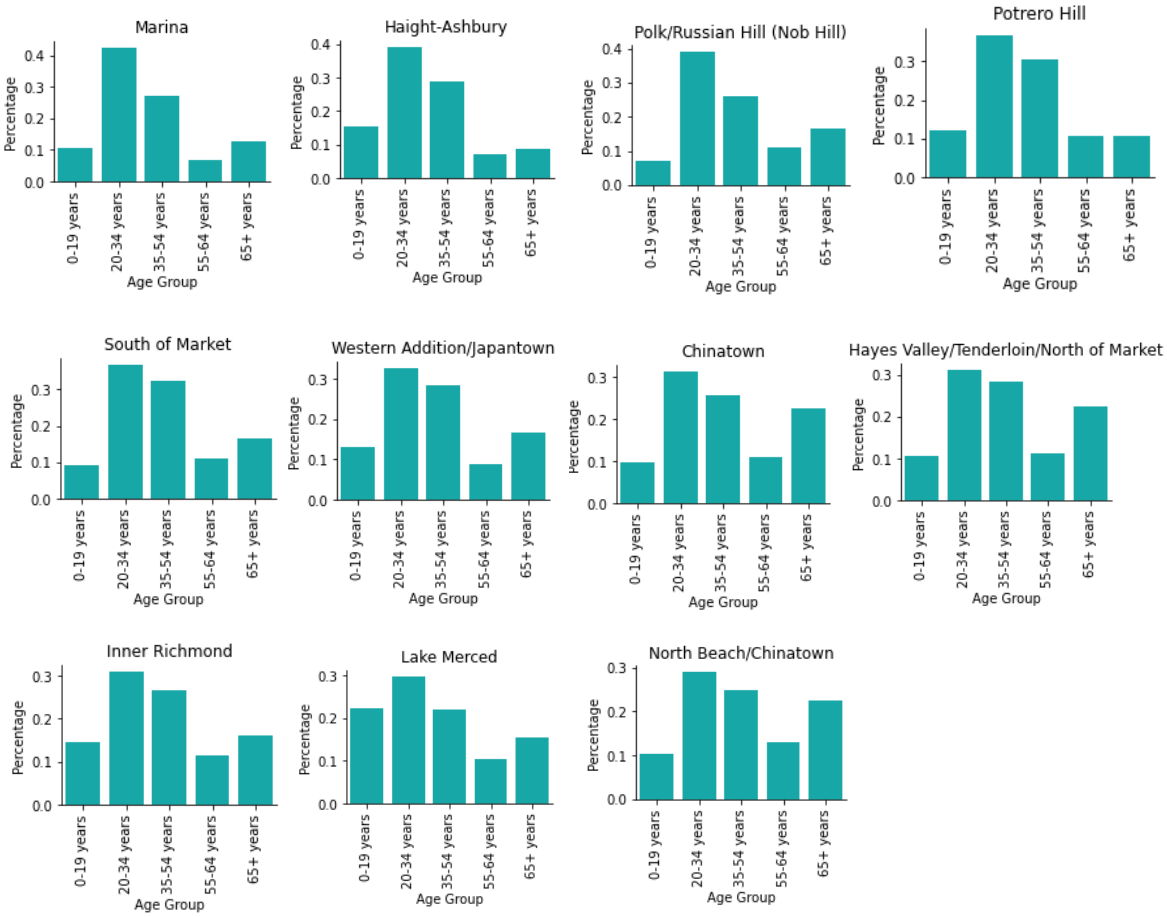
Figure 1: Map of San Francisco neighborhoods

3.1.2 Age distribution of 20-34 years old per neighborhood

The presence of the 20-34 age group in neighborhoods varies (see Appendix 1). Table 3 displays the bar charts of the 11 neighborhoods where the 20-34 years group is dominant. In six of those neighborhoods the target group makes up at least 33% of the population:

- Marina (0.43)
- Haight-Ashbury (0.39)
- Polk/Russian Hill (0.39)
- Portrero Hill (0.37)
- South of Market (0.37)
- Western Addition/Japantown (0.33)

Table 3: Neighborhoods where 20-34 years age group is dominant (ordered descending)



The bubble map below (figure 2) shows how the proportion of this age group is larger in the northeast neighborhoods than in other neighborhoods.

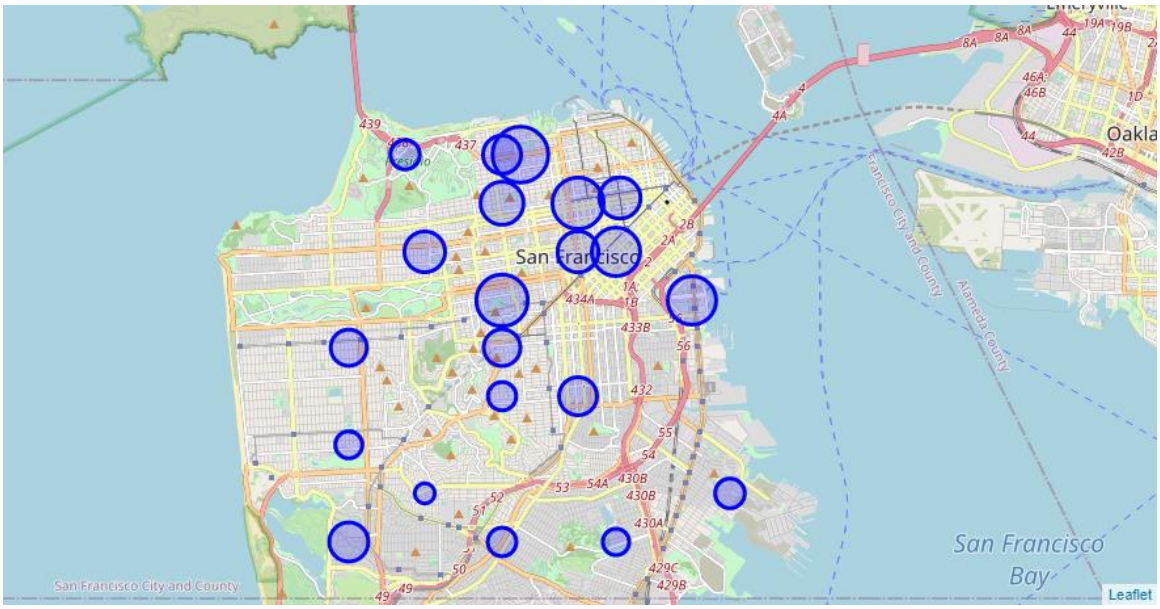


Figure 2. Bubble map of the relative presence of 20-34 age group in San Francisco neighborhoods

3.1.3 Venue distribution per neighborhood

The boxplot (figure 3) below shows the spread of the venue counts per neighborhood. The spread of the counts is wide; the range is 96 (min=3, max=99), the median is 56, and 75% of the venue counts is 71 or lower. The neighborhoods in the upper 25% are:

- Hayes Valley/Tenderloin/North of Market (99)
- Chinatown (98)
- Marina (98)
- Polk/Russian Hill (Nob Hill) (83)
- Portrero Hill (80)

The relative size of the venue counts per neighborhood are mapped out in figure 4. The biggest bubbles show up in the northeast of the Peninsula.

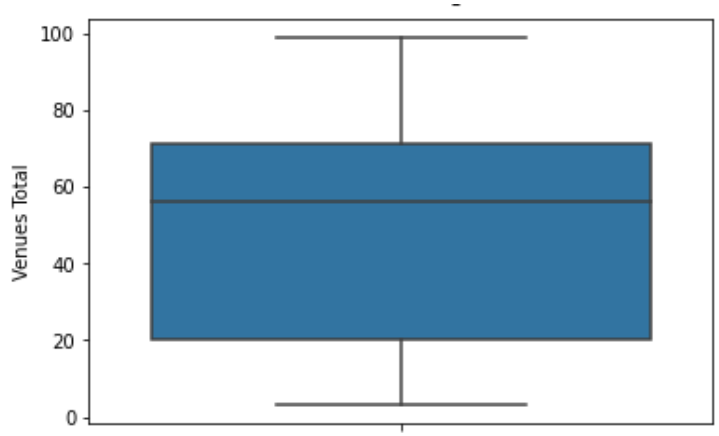


Figure 3: Boxplot of the venue counts per neighborhood

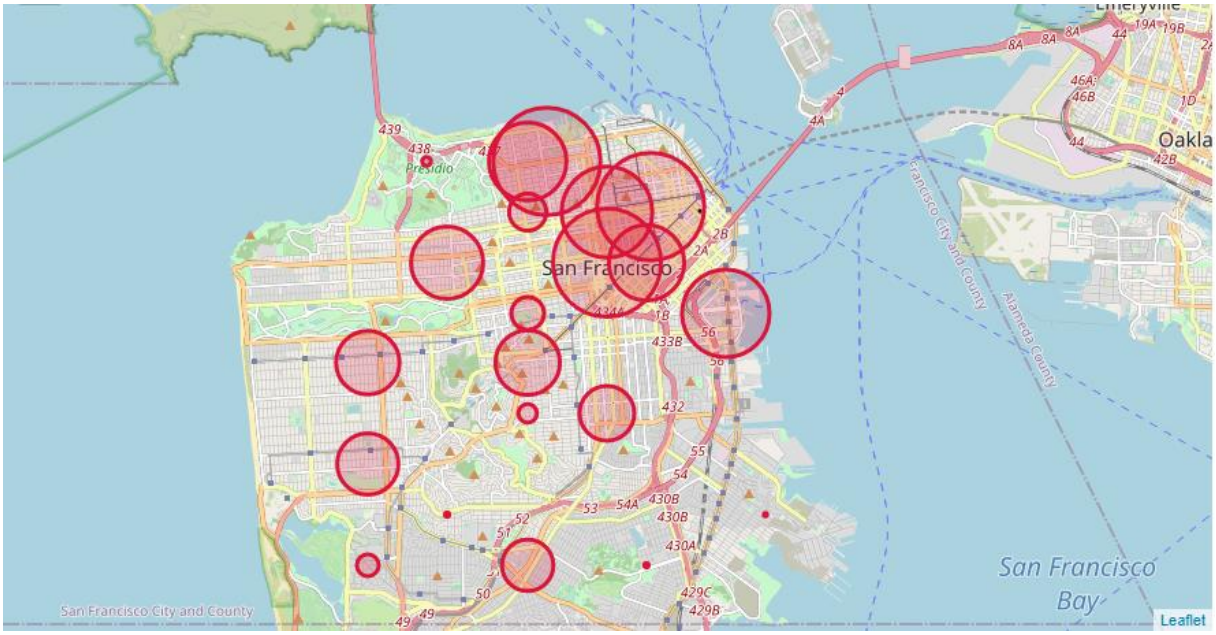


Figure 4. Bubble map of the relative venue counts in San Francisco neighborhoods

3.2 Age Venues Equation (AVE)

The range of the AVE scores is 0.411 and lies between 0.006 and 0.417 (figure 5). The median score is 0.145. Combining the boxplot with the overall scores (see Appendix 2) the neighborhoods in the fourth quartile are:

- Marina (0.417)
- Polk/Russian Hill (0.325)
- Hayes Valley/Tenderloin/north of Market (0.310)
- Chinatown (0.310)
- Portrero Hill (0.300)

As the bubble map in figure 6 shows, all are all situated in the northeast of the Peninsula.

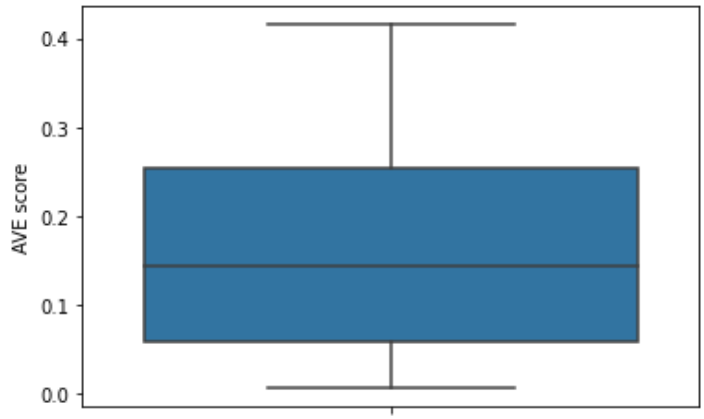


Figure 5: Boxplot of the AVE scores per neighborhood

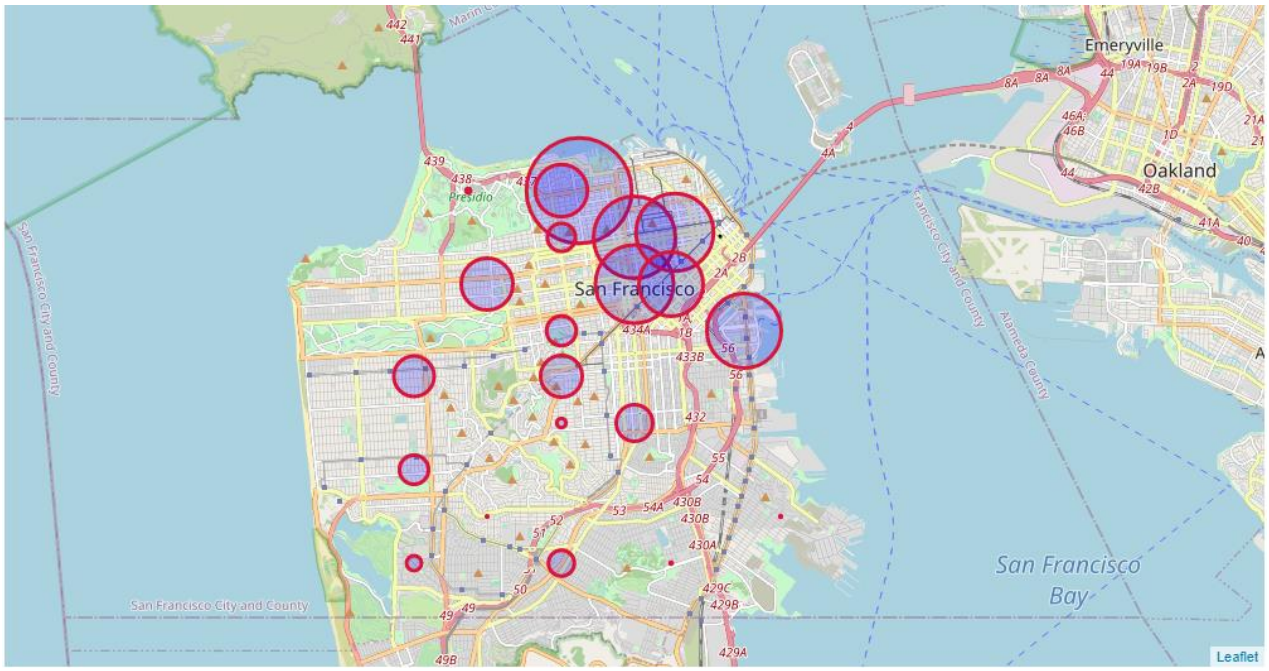


Figure 6. Bubble map of the AVE scores in San Francisco neighborhoods

3.3 K-Means Clustering

3.3.1 Determining optimal K

The venues were converted to weighted values which makes it possible to use K-Means Clustering. The neighborhoods, of course, stayed categorical. This means that the optimal K cannot be determined using the elbow method and/or the silhouette method. In its place we used visual inspection of the bubble maps on the relative presence of 20-34 age group (figure 2), the venue distribution (figure 4) and the AVE scores (figure 6). It was expected that the optimal K would return the perceived cluster in the northeast corner of the San Francisco Peninsula, and that the cluster would show certain amount of homogeneity in age distribution, and in the number and types of venues.

Next, neighborhoods were clustered on k=2, k=3, k=4, k=5 and k=6. The optimal K was found to be k=3. Here, the expected northeast cluster appeared and there seemed to be a relation between the neighborhood's age distributions, and the number and kind of venues in the neighborhoods within the cluster. The other ks presented clusters that where fuzzy and did not show a clear uniformity on the parameters age and venues.

3.3.2 Clustering with k=3

Three clusters were formed using the K-Means algorithm. An overview of the neighborhoods and venues in each cluster can be found in Appendixes 3, 4 and 5. As figure 7 shows, cluster 2 (purple) is in the northeast corner of San Francisco.

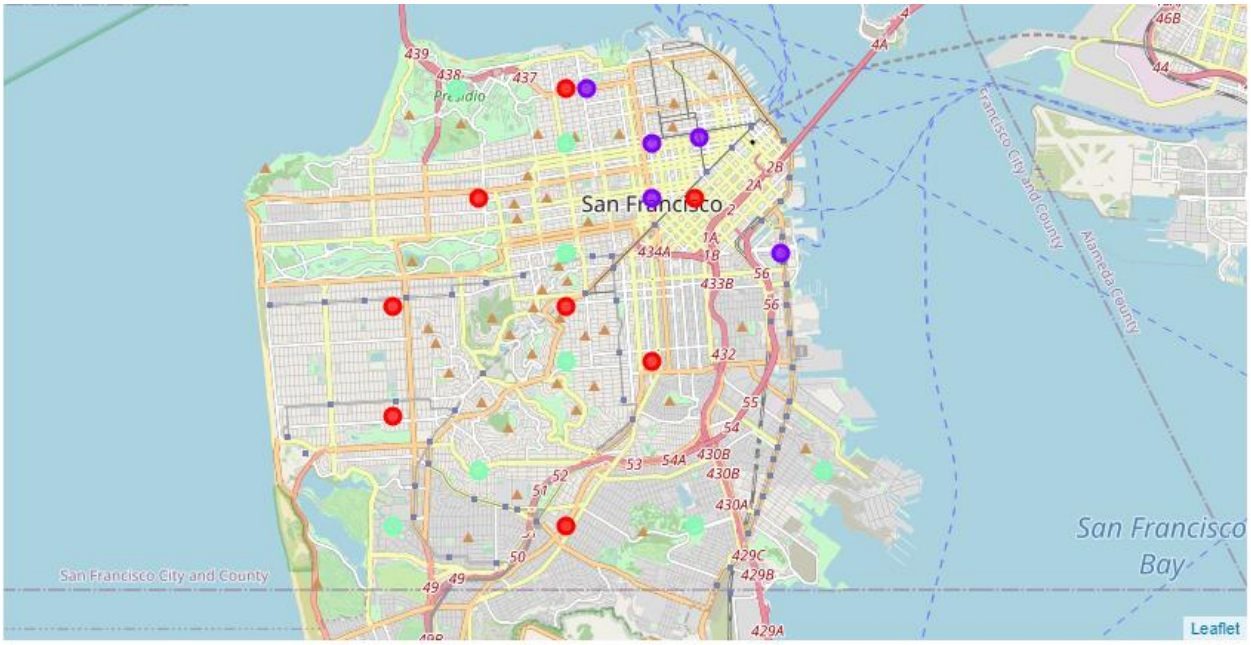


Figure 7. Bubble map of three clusters of San Francisco neighborhoods based on venue types

3.3.3 Analyzing clusters

Cluster 1: the 'Mixed-Use' cluster

This cluster consists of eight neighborhoods that are spread out over the Peninsula (see figure 7). The neighborhoods and their top 5 venues can be found Appendix 3.

The AVE scores in this cluster lie between 0.103 and 0.255, which is in the 25-75% range of the total number of AVE scores (median = 0.145). The number of venues in the neighborhoods range from 47 to 71. Which is, compared to other neighborhoods on the peninsula, quite high (total venues median= 56).

The venues can be characterized by three themes:

- **Day-to-day living**, with bakeries, groceries, deli's, pet stores, coffee shops and restaurants.
- **Public spaces**, like parks and light rail stations.
- **Night life**, with theatres, cafes, and bars.

This cluster can be characterized as a mixed-use cluster. On the one hand there are residents, with shops, venues, and recreational spaces for their specific use, and on the other hand there are night life venues.

Cluster 2: the 'Hospitality, Health and Luxury' cluster

In contrast to the other clusters, this cluster is formed by five adjacent neighborhoods (see figure 7): Marina, Polk/Russian Hill, Hayes Valley/Tenderloin/north of Market, Chinatown, and Portrero Hill. All five are characterized by a large population, 30-42%, of 20-34 years old (see also §3.1.2) and a high number of businesses, 80-99 (total venues median = 56, max=99). This results in AVE scores that add up to the top 25% of the San Francisco scores (0.295-0.417).

Looking at the top 5 venues in each neighborhood (Appendix 4) two themes can be distinguished:

- **Hospitality**, with hotels, coffee shops, wine bars and restaurants
- **Health and Luxury**, with cosmetic shops, fitness centers, massage studios, and boutiques

The presence of many businesses and the relative absence of common venues, like grocery stores, hairdressers or pizza restaurants, makes this cluster less residential and more business oriented with an emphasis on Hospitality, Health and Luxury.

Cluster 3: the 'Residential' cluster

The eight neighborhoods in this cluster are, like Cluster 1, spread out over the Peninsula (see figure 7). The members have low AVE scores, 0.006-0.118, compared to the total AVE score in San Francisco (range= 0.006-0.417, median = 0.145). Note that two of the neighborhoods, Haight-Ashbury and Western Addition/Japantown are mentioned in §3.1.2, because they have high scores on the target population (>= 33%).

As shown in Appendix 5 the number of venues in the neighborhoods range from 3 to 34. The business presence in this cluster is low; 3 out of the 8 neighborhoods list only three or four venues. The cluster can be characterized by two motifs:

- **Day-to-day living**, with a bakery, a grocery, a salon/barbershop, a juice bar, a few coffee shops and restaurants, a furniture/home store, a gym, a yoga studio, and a spa.
- **Public and/or recreational spaces**, this cluster has, compared to the other clusters, a lot of public and/or recreational spaces like parks, a garden, a fountain, a dog run, and a lookout point, an art gallery, a tennis court, and a baseball field.

The spaces in this cluster are primarily used for living with a high presence of public and/or recreational spaces and an overall low presence of businesses.

4. Results and Discussion

This report describes how the age and venue distribution in San Francisco neighborhoods were analyzed in order to make a data driven recommendation on the most suitable area to establish a CBD selling point. The recommendation needed to fall within three parameters (in order of importance): 1) the preferred area is a cluster of neighborhoods, 2) the dominant age group in the area is 20 -34 years (the target age group), and 3) the area has enough businesses to attract people that are interested health and/or are looking for fun.

Looking at age we determined 11 candidate neighborhoods where the target group is dominant over other age groups. In six of those the target age group constitutes at least 33% of the neighborhood population. Since the age distribution has precedence over venues, we then calculated the weighted venue score per neighborhood, the AVE. Next, using the AVE scores and mapping the scores in a bubble map five candidates situated in the northeast of the peninsula emerged: Marina, Polk/Russian Hill, Hayes Valley/Tenderloin/north of Market, Chinatown, and Portrero Hill.

Using K-Means clustering the neighborhoods were then clustered on the type of venues within their borders. This resulted in three clusters. Two of the clusters fall within the parameters: the Mixed-Use cluster (MU cluster), and the Hospitality, Health and Luxury cluster (HHL cluster). The MU cluster comprises eight neighborhoods that are spread out over the peninsula. Their AVE scores and the number of venues are medium high. There are retail and recreational spaces for day-to-day living, and there is night life, like theatres, cafes, and bars. The HHL cluster comprises five neighborhoods with a relative high number of the 20-34 age group and the highest AVE scores: Marina, Polk/Russian Hill, Hayes Valley/Tenderloin/North of Market, Chinatown, and Portrero Hill. All are next to one another and situated in the northeast of the Peninsula. The cluster is primarily business oriented with an emphasis on Hospitality (Hotels, Bars, Cafes), Health (Spa, Fitness, Cosmetics) and Luxury (Boutiques), which falls into the stated guidelines.

Our analysis points towards the HHL cluster as the preferred starting point for the location search. The layout of this cluster, the five adjacent neighborhoods, strengthens this recommendation. When going into the next stage of the decision-making process the client should keep an open mind to the other neighborhoods that emerged in this research. Especially South of Market, North Beach/Chinatown and Western Addition/Japantown, which are within/next to the HHL cluster and have a 20-34 years old population of around 33% of the total neighborhood population. Looking for an actual location in the northeast area of the Peninsula will also require the collection of new data, like real estate availability and prices, the spread of existing CBD selling points in the area vs the size of the local CBD market, and the movements of (potential) customers to, through and from the cluster.

Finally, we would like to discuss two considerations related to the data. The first is that the Foursquare data might be skewed by 1) the lack of users in a certain area, and/or 2) by imitation behavior of users. Certain types of venues might get more than average attention, while others might get less. Also, certain types of venues might never appear in the app. Follow-up research should take this into account and use multiple data sources to get a more valid dataset. The second consideration is the fluctuating nature of the Foursquare data. Even when analyzing the venues, minor shifts in the top 5s were observed. This makes reproduction of the research difficult and the reliability less robust.

5. Conclusion

This research analyzed age -, venue distribution, and neighborhood clustering in San Francisco to make a data driven recommendation on the most suitable area to start the search for a CBD shop location. After analysis two strong candidates emerged. With the stronger representation of the target age group, with better AVE scores, and with relevant businesses like cafes, bars, spas, and fitness centers the Hospitality, Health and Luxury cluster (HHL cluster) is the most likely starting point. This position is strengthened by the adjacent position of the neighborhoods within the HHL cluster and the existence of three interesting neighborhoods around/within the HHL cluster.

Finally, looking at the results of this research the question arises if this report might also be interesting for other businesses that are targeting 20-34 years old with an interest in health and a healthy lifestyle. Of course, it cannot be used as a final answer but, just like it is intended here, as a starting point to find a suitable location.

References

Brightfield Group(2017). Understanding cannabidiol. Retrieved from https://daks2k3a4ib2z.cloudfront.net/595e80a3d32ef41bfa200178/59946dd86c6b200001c5b9cb_CBD_-_HelloMD_Brightfield_Study_-_Expert_Report_-_FINAL.pdf

Consumer Reports (2019). CBD Goes Mainstream. Retrieved from https://www.consumerreports.org/cbd/cbd-goes-mainstream/

Leung, V. (2019). Data Clustering in San Francisco Neighborhoods. Retrieved from https://towardsdatascience.com/kickstart-your-first-clustering-project-in-san-francisco-neighborhoods-e258e659440c

San Francisco Department of Public Health (2004). San Francisco Burden of Disease & Injury Study: Determinants of Health. Retrieved from http://www.healthysf.org/bdi/outcomes/zipmap.htm

Statista (2012). Estimated dollar sales of the CBD market in the United States in 2019, by state (in million U.S. dollars). Retrieved from https://www.statista.com/statistics/1065838/dollar-sales-of-us-cbd-market-by-state/

US Census Bureau (2019). American Community Survey. Retrieved from https://data.census.gov/cedsci/table?q=United%20States&t=Age%20and%20Sex&g=8600000US94102,94103,94107,94108,94109,94110,94112,94114,94115,94116,94117,94118,94121,94122,94123,94124,94127,94131,94132,94133,94134&y=2019&tid=ACSDP5Y2019.DP05&moe=false&tp=true&hidePreview=true

All sources were accessed in April, 2021

Appendix 1. Age Groups per Neighborhood (in %) ordered by ’20-34 years’

	Zip Code	Neighborhood	0-19 years	20-34 years	35-54 years	55-64 years	65+ years
14	94123	Marina	0.107	0.426	0.270	0.068	0.127
10	94117	Haight-Ashbury	0.154	0.393	0.290	0.073	0.089
4	94109	Polk/Russian Hill (Nob Hill)	0.070	0.391	0.260	0.111	0.167
2	94107	Potrero Hill	0.123	0.369	0.307	0.106	0.106
1	94103	South of Market	0.092	0.369	0.324	0.111	0.167
8	94115	Western Addition/Japantown	0.132	0.328	0.284	0.090	0.166
3	94108	Chinatown	0.096	0.314	0.256	0.111	0.225
0	94102	Hayes Valley/Tenderloin/North of Market	0.106	0.313	0.285	0.111	0.225
11	94118	Inner Richmond	0.147	0.311	0.267	0.116	0.160
18	94132	Lake Merced	0.222	0.298	0.219	0.105	0.156
19	94133	North Beach/Chinatown	0.104	0.291	0.249	0.130	0.225
5	94110	Inner Mission/Bernal Heights	0.156	0.290	0.342	0.106	0.106
7	94114	Castro/Noe Valley	0.118	0.280	0.352	0.114	0.134
13	94122	Sunset	0.151	0.276	0.291	0.117	0.164
15	94124	Bayview-Hunters Point	0.236	0.227	0.265	0.130	0.142
12	94121	Outer Richmond	0.160	0.225	0.284	0.140	0.191
6	94112	Ingelside-Excelsior/Crocker-Amazon	0.178	0.219	0.286	0.144	0.172
17	94131	Twin Peaks-Glen Park	0.169	0.217	0.320	0.132	0.163
9	94116	Parkside/Forest Hill	0.172	0.207	0.277	0.132	0.212
20	94134	Visitacion Valley/Sunnydale	0.200	0.202	0.283	0.149	0.164
16	94127	St. Francis Wood/Miraloma/West Portal	0.207	0.154	0.287	0.146	0.208

Appendix 2. AVE Scores per Neighborhood ordered by ‘AVE’

	Zip Code	Neighborhood	Latitude	Longitude	20-34 years	Venues Total	AVE
14	94123	Marina	37.800	-122.435205	0.426	98	0.41748
4	94109	Polk/Russian Hill (Nob Hill)	37.790	-122.420000	0.391	83	0.32453
0	94102	Hayes Valley/Tenderloin/North of Market	37.780	-122.420000	0.313	99	0.30987
3	94108	Chinatown	37.791	-122.409000	0.314	98	0.30772
2	94107	Potrero Hill	37.770	-122.390000	0.369	80	0.29520
1	94103	South of Market	37.780	-122.410000	0.369	69	0.25461
19	94133	North Beach/Chinatown	37.800	-122.440000	0.291	71	0.20661
11	94118	Inner Richmond	37.780	-122.460000	0.311	66	0.20526
7	94114	Castro/Noe Valley	37.760	-122.440000	0.280	59	0.16520
13	94122	Sunset	37.760	-122.480000	0.276	58	0.16008
5	94110	Inner Mission/Bernal Heights	37.750	-122.420000	0.290	50	0.14500
10	94117	Haight-Ashbury	37.770	-122.440000	0.393	30	0.11790
9	94116	Parkside/Forest Hill	37.740	-122.480000	0.207	56	0.11592
8	94115	Western Addition/Japantown	37.790	-122.440000	0.328	34	0.11152
6	94112	Ingelside-Excelsior/Crocker-Amazon	37.720	-122.440000	0.219	47	0.10293
18	94132	Lake Merced	37.720	-122.480000	0.298	20	0.05960
17	94131	Twin Peaks-Glen Park	37.750	-122.440000	0.217	17	0.03689
12	94121	Outer Richmond	37.800	-122.465453	0.225	8	0.01800
20	94134	Visitacion Valley/Sunnydale	37.720	-122.410000	0.202	4	0.00808
15	94124	Bayview-Hunters Point	37.730	-122.380000	0.227	3	0.00681
16	94127	St. Francis Wood/Miraloma/West Portal	37.730	-122.460000	0.154	4	0.00616

Appendix 3. Cluster 1 Neighborhoods and their Top 5 Venues (normalized)

Total venues in South of Market : [69]

	venue	freq
0	Coffee Shop	0.10
1	Theater	0.06
2	Bakery	0.04
3	Vietnamese Restaurant	0.04
4	Sandwich Place	0.04

Total venues in North Beach/Chinatown : [71]

	venue	freq
0	Italian Restaurant	0.08
1	Gym / Fitness Center	0.04
2	Wine Bar	0.04
3	French Restaurant	0.04
4	Pizza Place	0.03

Total venues in Inner Richmond : [66]

	venue	freq
0	Sushi Restaurant	0.06
1	Pizza Place	0.05
2	Pet Store	0.05
3	Bar	0.03
4	Bakery	0.03

Total venues in Castro/Noe Valley : [59]

	venue	freq
0	Gay Bar	0.07
1	Coffee Shop	0.05
2	Park	0.05
3	Wine Bar	0.03
4	Grocery Store	0.03

Total venues in Sunset : [58]

	venue	freq
0	Bubble Tea Shop	0.09
1	Vietnamese Restaurant	0.09
2	Bakery	0.09
3	Chinese Restaurant	0.05
4	Deli / Bodega	0.05

Total venues in Inner Richmond : [66]

	venue	freq
0	Sushi Restaurant	0.06
1	Pizza Place	0.05
2	Pet Store	0.05
3	Bar	0.03
4	Bakery	0.03

Total venues in Parkside/Forest Hill : [56]

	venue	freq
0	Chinese Restaurant	0.18
1	Park	0.07
2	Light Rail Station	0.05
3	Pizza Place	0.04
4	Café	0.04

Total venues in Ingelside-Excelsior/Crocker-Amazon : [47]

	venue	freq
0	Pizza Place	0.09
1	Mexican Restaurant	0.09
2	Café	0.06
3	Chinese Restaurant	0.06
4	Bar	0.06

Appendix 4. Cluster 2 Neighborhoods and their Top 5 Venues (normalized)

Total venues in Marina : [98]

	venue	freq
0	Cosmetics Shop	0.06
1	Italian Restaurant	0.04
2	Gym / Fitness Center	0.04
3	French Restaurant	0.03
4	Sandwich Place	0.03

Total venues in Polk/Russian Hill (Nob Hill) : [83]

	venue	freq
0	Grocery Store	0.05
1	Sushi Restaurant	0.05
2	Massage Studio	0.04
3	Thai Restaurant	0.04
4	Pet Store	0.04

Total venues in Hayes Valley/Tenderloin/North of Market : [99]

	venue	freq
0	Café	0.05
1	Hotel	0.04
2	Sandwich Place	0.04
3	Vietnamese Restaurant	0.04
4	Coffee Shop	0.04

Total venues in Chinatown : [98]

	venue	freq
0	Hotel	0.07
1	Coffee Shop	0.07
2	Boutique	0.05
3	Bubble Tea Shop	0.03
4	Sushi Restaurant	0.03

Total venues in Potrero Hill : [74]

	venue	freq
0	Food Truck	0.20
1	Coffee Shop	0.07
2	Gym	0.05
3	Pharmacy	0.04
4	Café	0.04

Appendix 5. Cluster 3 Neighborhoods and their Top 5 Venues (normalized)

Total venues in Haight-Ashbury : [30]

	venue	freq
0	Coffee Shop	0.10
1	Grocery Store	0.10
2	Bakery	0.06
3	Yoga Studio	0.03
4	Recreation Center	0.03

Total venues in Western Addition/Japantown : [34]

	venue	freq
0	Park	0.09
1	Chinese Restaurant	0.09
2	Sushi Restaurant	0.06
3	Furniture / Home Store	0.06
4	Spa	0.06

Total venues in Lake Merced : [20]

	venue	freq
0	Gym	0.20
1	Park	0.10
2	Café	0.10
3	Juice Bar	0.05
4	Mexican Restaurant	0.05

Total venues in Twin Peaks-Glen Park : [17]

	venue	freq
0	Park	0.17
1	Yoga Studio	0.06
2	Dog Run	0.06
3	North Indian Restaurant	0.06
4	Gift Shop	0.06

Total venues in Outer Richmond : [8]

	venue	freq
0	Park	0.12
1	Trail	0.12
2	Gymnastics Gym	0.12
3	Art Gallery	0.12
4	National Park	0.12

Total venues in Visitacion Valley/Sunnydale : [4]

	venue	freq
0	Music Venue	0.25
1	Park	0.25
2	Garden	0.25
3	Baseball Field	0.25
4	Nail Salon	0.00

Total venues in Bayview-Hunters Point : [3]

	venue	freq
0	Motorcycle Shop	0.33
1	Lighting Store	0.33
2	Coffee Shop	0.33
3	ATM	0.00
4	Park	0.00

Total venues in St. Francis Wood/Miraloma/West Portal : [4]

	venue	freq
0	Fountain	0.25
1	Bus Line	0.25
2	Park	0.25
3	Scenic Lookout	0.25
4	Paper / Office Supplies Store	0.00