1. Introduction / Business Problem

The following project is aimed to find out where in the city of San Francisco would it be best to open or build a new bar for it to achieve its maximum success. We will look at what area is safest and most popular for the bar to thrive. Anyone that is looking to a open a bar in San Francisco or someone looking for a bar to visit can benefit from this analysis.

2. Data

Public datasets from San Francisco will be used alongside Foursquare API.
We will be using 1) San Francisco Registered Business Data to help locate the number and type of business in each area, 2) San Francisco Crime Data to observe which neighborhood is safest to establish a new bar, and 3) Foursquare API Data for access to venue data.

a. San Francisco Registered Business Data

Using python, we can grab the data from this public data set and put it into a pandas data frame in order to show us the number of businesses located within each neighborhood of San Francisco.

```
In [12]: business = pd.read_csv('https://data.sfgov.org/api/views/g8m3-pdis/rows.csv?accessType=DOWNLOAD')
         print(business.shape)
         business.head()
```

```
(260309, 36)
```

Out[12]:

	Location Id	Business Account Number	Ownership Name	DBA Name	Street Address	City	State	Source Zipcode	Business Start Date	Business End Date	Locatio Start Da
0	1103593-08-161	1049564	Anjan Rajbhandari	Uber	28134 Harvey Ave	Hayward	CA	94544.0	03/24/2014	12/31/2017	03/24/2
1	1218784-04-191	1100756	Luisa Alberto	High Five Sf	467 14th St	San Francisco	CA	94103.0	04/15/2019	04/15/2019	04/15/2
2	1223199-05-191	1102424	Sunrun, Inc.	Sunrun, Inc	595 Market St	San Francisco	CA	94105.0	06/01/2008	06/01/2008	06/01/2
3	1220748-05-191	1101579	Felix Hernandez	Tru-Tec Electric	44 Mcaker Ct	San Mateo	CA	94403.0	05/06/2019	06/18/2019	05/06/2
4	1135452-02-171	1065102	Tirta Llc	Tirta	105 Hudson St 6s	New York City	NY	10013.0	06/09/2016	12/23/2018	06/09/2

The business registration data frame contains the name and number of businesses in the entire bay area and shows what neighborhood they are located

After cleaning up the data to see the number of businesses registered in San Francisco in the last 10 years group by neighborhood, this is the dataframe we get:

Out[14]:

	Neighborhood	Businesses
5	Financial District/South Beach	12196
18	Mission	6654
33	South of Market	5759
34	Sunset/Parkside	4396
0	Bayview Hunters Point	3680
25	Outer Richmond	2909
16	Marina	2750
2	Castro/Upper Market	2690
39	West of Twin Peaks	2520
9	Hayes Valley	2520

It looks like the Financial District has significantly more business registrations than everywhere else, but all the top 10 have a steady number of businesses.

b. San Francisco Crime Data

Next, we look at San Francisco's crime data to help us make sure we select a safe area for our bar.

In [17]:
```
crime = pd.read_csv('https://data.sfgov.org/api/views/wg3w-h783/rows.csv?accessType=DOWNL
OAD')
print(crime.shape)
crime.head()
```

```
(371095, 36)
```

Out[17]:

	Incident Datetime	Incident Date	Incident Time	Incident Year	Incident Day of Week	Report Datetime	Row ID	Incident ID	Incident Number	CAD Number
0	2020/05/12 05:45:00 PM	2020/05/12	17:45	2020	Tuesday	2020/05/13 09:46:00 AM	92897328150	928973	206082743	NaN
1	2020/05/19 09:00:00 PM	2020/05/19	21:00	2020	Tuesday	2020/05/20 05:43:00 PM	92899306244	928993	206082709	NaN
2	2020/05/16 06:00:00 PM	2020/05/16	18:00	2020	Saturday	2020/05/16 10:26:00 PM	92902428150	929024	206083296	NaN
3	2020/03/30 12:00:00 AM	2020/03/30	00:00	2020	Monday	2020/05/04 11:47:00 AM	92905305073	929053	206062193	NaN
4	2020/02/03 02:45:00 PM	2020/02/03	14:45	2020	Monday	2020/02/03 05:50:00 PM	89881675000	898816	200085557	20034287(

We are going to use python again to clean up the data to see the number of incidents in the last 5 years, per neighborhood. This will give us a good idea of the safe and less safe areas. We get the resulting data frame:

	Neighborhood	Incidents
18	Mission	37755
35	Tenderloin	33523
5	Financial District/South Beach	30318
33	South of Market	28171
0	Bayview Hunters Point	19822
40	Western Addition	11003
22	North Beach	10119
2	Castro/Upper Market	10100
20	Nob Hill	9694
34	Sunset/Parkside	9398

3. Methodology: Data Visualization and Exploration

a. Narrowing Down Neighborhoods

By using graphs, we can examine our data sets and narrow down our options for our favorite neighborhoods for our new bar!

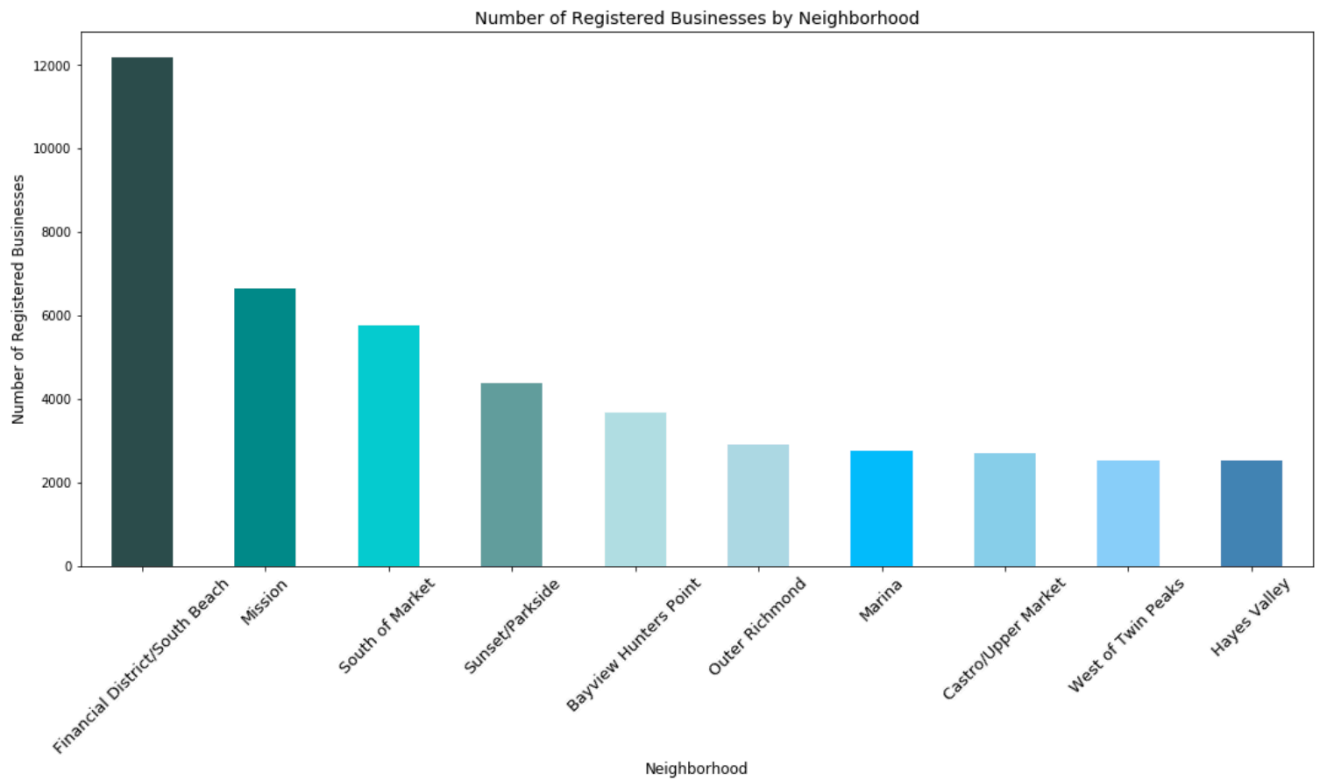

Figure 3: Count of business registered in the last 10 years for each neighborhood in San Francisco, sorted from most to least (showing top 10).

We will also look at a visualization of neighborhoods that experience the most crime.

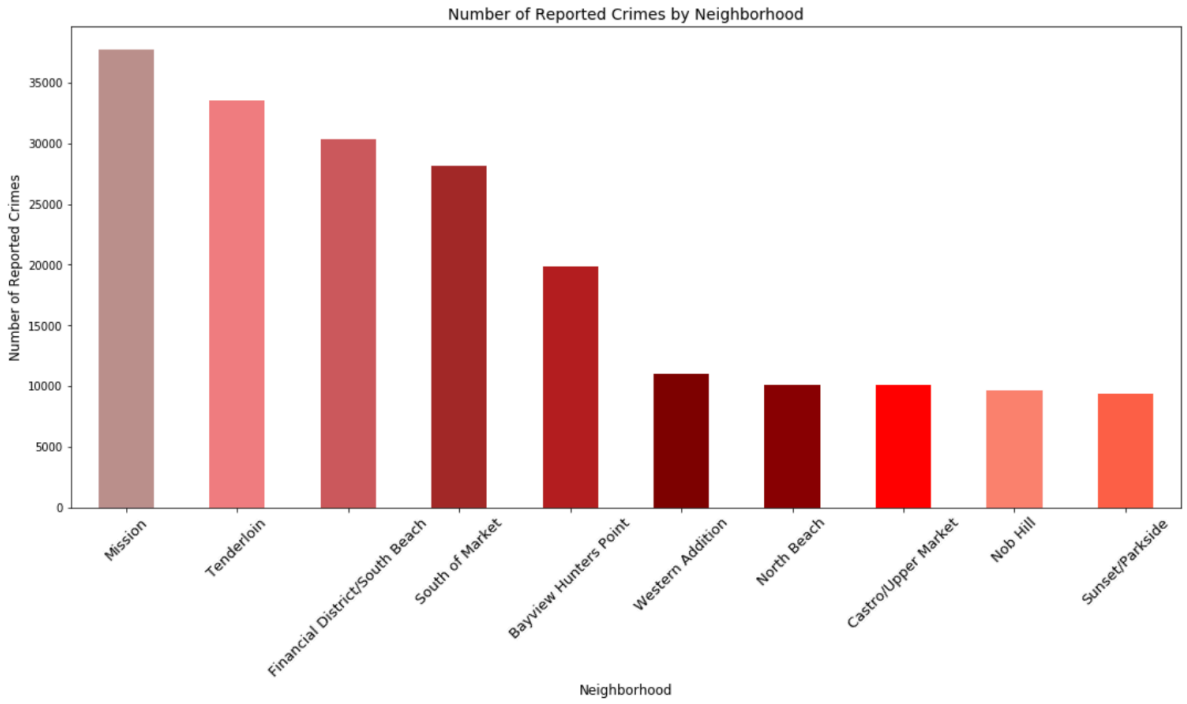

Figure 4: Number of crimes reported in the last five years in each neighborhood in San Francisco, sorted from most to least (showing top 10).

It looks like there is significantly more crime in the first 5 neighborhoods than the rest. We want to make sure we are located in a safe area so our clients feel safe. We will hence try and avoid these 5 dangerous neighborhoods.

Using Python, we merge our business and crime data frames into one to get a refined list of neighborhoods with a high number of businesses and low crime.

In [23]:
```
'''start by merging the datasets and making a new dataset that incldues the neighborhood
s
which wereamong the top 10 for businesses AND are amond the top 5 for crime '''
Overlap = business6.merge(crime8, on=['Neighborhood'])
'''then take this joined dataframe and remove all common values from your list of top 10
neighborhoods for businesses'''
SF_Neighborhoods = business6[(~business6.Neighborhood.isin(Overlap.Neighborhood))]
'''and what you have is the top neighborhoods for businesses that are NOT the top
neighborhoods for crime'''
SF_Neighborhoods.head()
```

Out[23]:

	Neighborhood	Businesses
34	Sunset/Parkside	4396
25	Outer Richmond	2909
16	Marina	2750
2	Castro/Upper Market	2690
39	West of Twin Peaks	2520

Next, using our geopy library we get coordinates for each of our top neighborhoods.

```
from geopy.geocoders import Nominatim
geolocator = Nominatim(user_agent="SF_explorer")
SF_Neighborhoods['Coordinates'] = SF_Neighborhoods['Neighborhood'].apply(geolocat
or.geocode).apply(lambda x: (x.latitude, x.longitude))
SF_Neighborhoods
```

	Neighborhood	Businesses	Coordinates
34	Sunset/Parkside	4396	(37.751616, -122.490810)
25	Outer Richmond	2909	(37.780001, -122.490229)
16	Marina	2750	(37.801406, -122.439718)
2	Castro/Upper Market	2690	(37.762932, -122.435395)
39	West of Twin Peaks	2520	(37.739871, -122.460106)
9	Hayes Valley	2520	(37.776685, -122.422936)

We have narrowed down our options to 6 neighborhoods that have a high business registration count and a low crime rate:

1. Sunset
2. Outer Richmond
3. Marina
4. Castro
5. West of Twin Peaks
6. Hayes Valley

3.2 Foursquare Data Analysis

Foursquare API will help us retrieve information about the most popular venues in each neighborhood in San Francisco. This is insightful to know which type of venue is most popular in each neighborhood. Calling the Foursquare API returns a JSON file, which can be turned into a data frame for analysis in a python notebook.

We start by writing a function that will search for the most popular venues within a half mile radius of our neighborhoods

```
def getNearbyVenues(names, latitudes, longitudes, radius=800, LIMIT = 100):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_
secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in
venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

```
SF_venues = getNearbyVenues(names=SF['Neighborhood'],
                                  latitudes=SF['Latitude'],
                                  longitudes=SF['Longitude']
                                 )
```

We get a data frame of 472 entries in the 6 neighborhoods and 169 unique venue categories:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Sunset/Parkside	37.751616	-122.490810	TJ Brewed Tea and Real Fruit (TJ Cups)	37.753561	-122.490028	Bubble Tea Shop
1	Sunset/Parkside	37.751616	-122.490810	S&T Hong Kong Seafood	37.753702	-122.491278	Dim Sum Restaurant
2	Sunset/Parkside	37.751616	-122.490810	Donut Time	37.753651	-122.489439	Donut Shop
3	Sunset/Parkside	37.751616	-122.490810	Polly Ann Ice Cream	37.753454	-122.497765	Ice Cream Shop
4	Sunset/Parkside	37.751616	-122.490810	Sunset Recreation Center	37.757310	-122.487072	Playground

Here is a graphic representation of the most popular venue categories across all 6 neighborhoods:

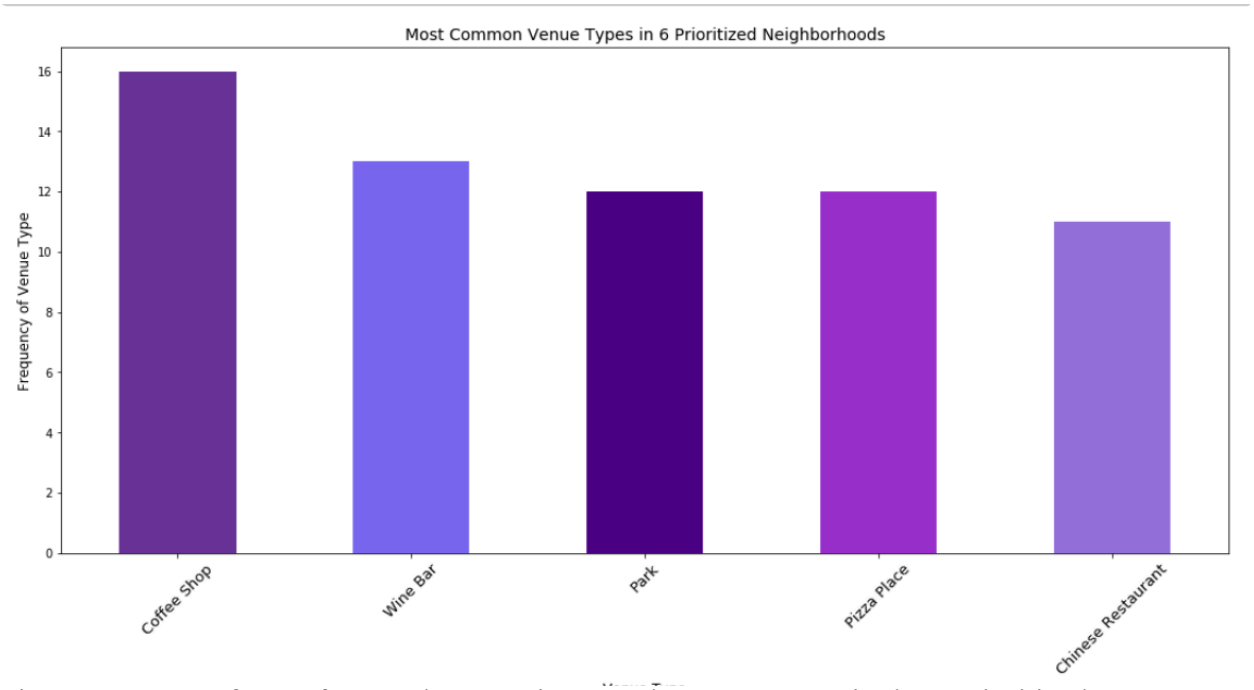

Figure 5: Count of most frequently occurring popular venue types in the 6 prioritized neighborhoods, sorted from most frequent to least (showing top 5).

Coffee shops are the most common popular venue type, followed by wine bars, parks, and various. This is good news as wine bars are definitely popular types of venues in our neighborhoods!

Let's dig further into each of the neighborhoods to see the most popular types of venues for each neighborhood. To do this, we will take the following steps:

1. Create a data frame of venue categories with pandas one hot encoding
2. Use pandas groupby to get the mean of the venue categories
3. Transpose data frame and arrange in descending order

```
# one hot encoding
SF_onehot = pd.get_dummies(SF_venues[['Venue Category']], prefix
="", prefix_sep="")

# add neighborhood column back to dataframe
SF_onehot['Neighborhood'] = SF_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [SF_onehot.columns[-1]] + list(SF_onehot.columns[:-
1])
SF_onehot = SF_onehot[fixed_columns]

SF_onehot.head()
```

```
#now group the data
SF_grouped = SF_onehot.groupby('Neighborhood').mean().reset_index()
print(SF_grouped.shape)
SF_grouped
```

```
(6, 170)
```

	Neighborhood	Accessories Store	Alternative Healer	American Restaurant	Antique Shop	Arcade	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Autom Shop
0	Castro/Upper Market	0.00	0.00	0.010000	0.000000	0.01	0.01	0.01	0.00	0.0000
1	Hayes Valley	0.01	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.0000
2	Marina	0.00	0.01	0.020000	0.000000	0.00	0.01	0.01	0.01	0.0000
3	Outer Richmond	0.00	0.00	0.011236	0.011236	0.00	0.00	0.00	0.00	0.0112
4	Sunset/Parkside	0.00	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.0000
5	West of Twin Peaks	0.00	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.0000

```
#print each neighborhood with the top 5 most common venues
num_top_venues = 5

for hood in SF_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = SF_grouped[SF_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_v
enues))
    print('\n')
```

We can see here the top venues for each neighborhood:

	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
0	Castro/Upper Market	Gay Bar	Coffee Shop	Park	New American Restaurant	Thai Restaurant	Indian Restaurant	Japanese Restaurant
1	Hayes Valley	Wine Bar	Pizza Place	New American Restaurant	Bakery	Dessert Shop	Coffee Shop	Cocktail Bar
2	Marina	Gym / Fitness Center	Cosmetics Shop	Wine Bar	Italian Restaurant	French Restaurant	Park	Sandwich Place
3	Outer Richmond	Chinese Restaurant	Café	Sushi Restaurant	Playground	Vietnamese Restaurant	Korean Restaurant	Seafood Restaurant
4	Sunset/Parkside	Chinese Restaurant	Japanese Restaurant	Elementary School	Dim Sum Restaurant	Playground	Pharmacy	Pizza Place
5	West of Twin Peaks	Burger Joint	Pizza Place	Park	Sandwich Place	Italian Restaurant	Mexican Restaurant	Coffee Shop

This data is important because it is giving us an idea of the atmosphere of each of these neighborhoods. The first two neighborhoods, Castro and Hayes Valley both have a type of bar as their more common venue. We can now narrow our search down to these two as we know people go there for the bar scene.

Next we look at a dataframe that merges the crime and business data to see which of the neighborhoods, Castro or Hayes Valley, has a lower crime rate.

Let’s look at each neighborhood and determine what percentage of their top 30 popular venues are bars or restaurants:

	Neighborhood	Businesses	Crimes	Coordinates	Latitude	Longit
0	Sunset/Parkside	4396	9398	(37.751616, -122.490810)	37.751616	-122.4
1	Outer Richmond	2909	7351	(37.780001, -122.490229)	37.780001	-122.4
2	Marina	2750	8193	(37.801406, -122.439718)	37.801406	-122.4
3	Castro/Upper Market	2690	10100	(37.762932, -122.435395)	37.762932	-122.4
4	West of Twin Peaks	2520	6594	(37.739871, -122.460106)	37.739871	-122.4
5	Hayes Valley	2520	9281	(37.776685, -122.422936)	37.776685	-122.4

We can see in the above figure that Hayes Valley has the third least crimes and is hence the better choice against Castro, the neighborhood with most crimes between the 6 neighborhoods.

b. Neighborhood Clustering

Finally, we can cluster our 6 neighborhoods based on their popular venue categories. This will help us get a feel for which neighborhoods are like each other based on the venues people like to visit in each one. We use K-Means clustering, detailed in the code below, to group our neighborhoods into 3 clusters.

```
# set number of clusters
kclusters = 3

SF_grouped_clustering = SF_grouped.drop('Neighborhood',
1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit
(SF_grouped_clustering)

# check cluster labels generated for each row in the data
frame
kmeans.labels_[0:10]
```

```
array([1, 1, 1, 2, 0, 1], dtype=int32)
```

```
SF_merged = SF
```

```
# merge SF_grouped with SF_data to add latitude/longitude
for each neighborhood
SF_merged = SF_merged.join(neighborhoods_venues_sorted.se
t_index('Neighborhood'), on='Neighborhood')
```

```
SF_merged['Latitude'] = SF_merged['Latitude'].astype(floa
t)
SF_merged['Longitude'] = SF_merged['Longitude'].astype(fl
oat)
SF_merged['Cluster Labels'] = SF_merged['Cluster Label
s'].astype(int)



SF_merged
```

We can then display these clusters on a leaflet map using the Folium library:

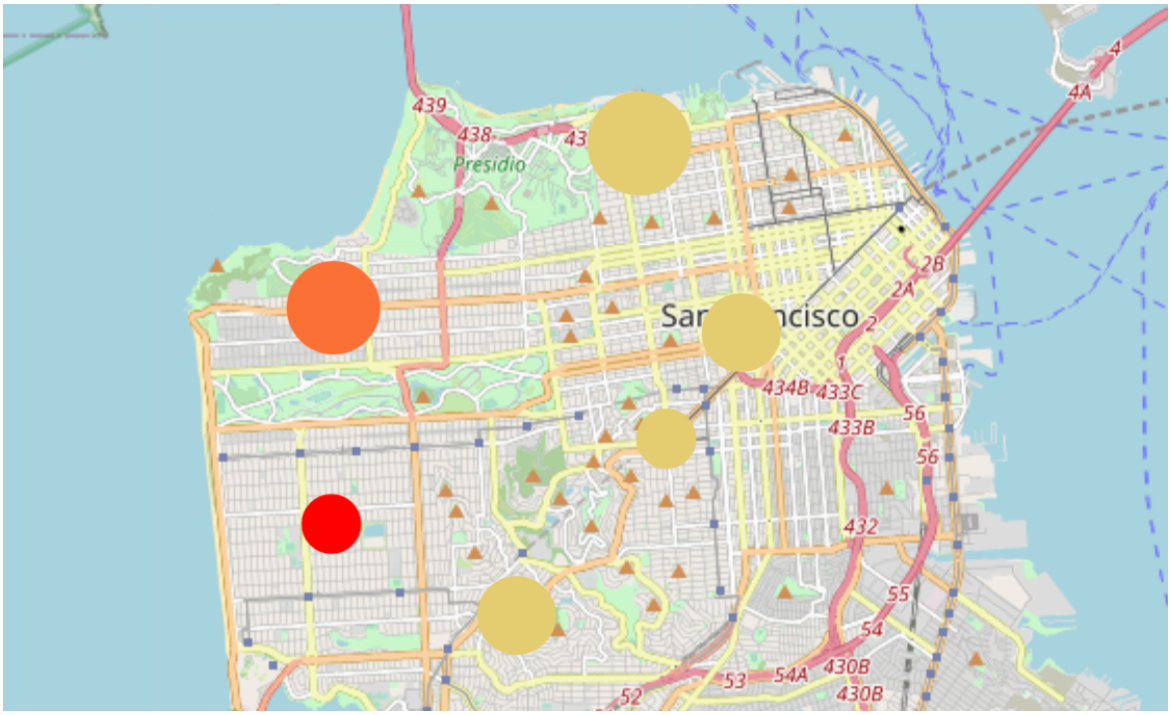

Figure 7: A map of San Francisco with each of our 6 preferred neighborhoods clustered into 3 groups (can differentiate them by color) based on the types of popular venues in each neighborhood. The size of each dot represents the number of bars and restaurants listed as popular venues.

West of Twin Peaks alongside with Castro, Hayes Valley, and Marina are all clustered together. Meanwhile, outer Richmond was placed in its own cluster as was Sunset.

3. Results and Discussion

By using information from datasets of business registrations and crime rates, as well as Foursquare, we have narrowed down neighborhood options to open up a bar in one of six areas in San Francisco

The most common venues in our areas of interest were discovered to be coffee shops, wine bars, parks, pizza places, and Chinese restaurants.

Clustering neighborhoods based on their most popular venues grouped West of Twin Peaks alongside with Castro, Hayes Valley, and Marine. Meanwhile, outer Richmond was placed in its own cluster as was Sunset.

Castro and Hayes Valley are the two neighborhoods where bars are the most popular venues.

Because Castro is the neighborhood that has the most crime rates out of the six areas we looked at, I have come to the conclusion that Hayes Valley is the neighborhood where a new bar should be opened because it is the 3rd out of 6 in crime rates and has much of a bar scene.

4. Conclusion

This data science project used python libraries to manipulate and transform datasets and Foursquare API to explore the neighborhoods of San Francisco. Folium map was used for clustering and segmenting the neighborhoods. These analytical tools allow for in depth analysis and problem solving as seen in this case. With public data of San Francisco alongside with the libraries I was able to make an educated choice on where I believe is the optimal neighborhood to open a bar in San Francisco.