# 1. Introduction

## 1.1. Business Problem

Can we predict a good place to start a bar in San Francisco city.

## 1.2. Target Audience

The target audience of this report would be anyone who wants to buy or build a bar in San Francisco, or anyone in San Francisco just looking for a nice area to grab a drink.

# 2. Data

In order to best make this decision, we're going to need some data. Fortunately, the city of San Francisco has hundreds of public data sets that describe various aspects of the city, and Foursquare API allows free access to some of its location and venue data.

Altogether, we are looking at three sets of data for our analysis:

1.San Francisco Registered Business Data

2.San Francisco Crime Data

3.Foursquare Data

## 2.1.San Francisco Registered Business Data

We can pull a list of every business registered in San Francisco from the last couple of decades from the data SF website. We can grab this data and place it into a pandas data frame using python. This is going to help us roughly gauge the foot traffic in each neighborhood of San Francisco (at least on the weekdays) by providing the number of businesses located within each neighborhood.

```
In [40]: business = pd.read_csv('https://data.sfgov.org/api/views/g8m3-pdis/rows.csv?accessType=DOWNLOAD')
         print(business.shape)
         business.head()
```

```
(250406, 26)
```

Out[40]:

| | Location Id | Business Account Number | Ownership Name | DBA Name | Street Address | City | State | Source Zipcode | Business Start Date | Business End Date | Location Start Date | Location End Date | Mail Address | Ma Cit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1103593-08-161 | 1049564 | Anjan Rajbhandari | Uber | 28134 Harvey Ave | Hayward | CA | 94544.0 | 03/24/2014 | 12/31/2017 | 03/24/2014 | 12/31/2017 | NaN | Na |
| 1 | 1218784-04-191 | 1100756 | Luisa Alberto | High Five Sf | 467 14th St | San Francisco | CA | 94103.0 | 04/15/2019 | 04/15/2019 | 04/15/2019 | 04/15/2019 | NaN | Na |
| 2 | 1223199-05-191 | 1102424 | Sunrun, Inc. | Sunrun, Inc | 595 Market St | San Francisco | CA | 94105.0 | 06/01/2008 | 06/01/2008 | 06/01/2008 | 06/01/2008 | NaN | Na |
| 3 | 1220748-05-191 | 1101579 | Felix Hernandez | Tru-Tec Electric | 44 Mcaker Ct | San Mateo | CA | 94403.0 | 05/06/2019 | 06/18/2019 | 05/06/2019 | 06/18/2019 | NaN | Na |

You can see there's a total of 250,406 entries in the business registration data frame (note: this is for the entire bay area and contains decades of information). it has the name of businesses, their neighborhood, and the date that they were registered

Let’s clean up our data. We want to see the number of businesses registered in San Francisco in the last 10 years, grouped by neighborhood. This will give us a rough indication of how much foot traffic each area of the city gets today. After coding in python to get the neighborhoods with the most business registrations, we get the resulting data frame:

| Neighborhood | Businesses |
|---|---|
| Financial District/South Beach | 11229 |
| Mission | 6180 |
| South of Market | 5275 |
| Sunset/Parkside | 4074 |
| Bayview Hunters Point | 3338 |
| Outer Richmond | 2729 |
| Marina | 2528 |
| Castro/Upper Market | 2478 |
| West of Twin Peaks | 2352 |
| Hayes Valley | 2349 |

It looks like the Financial District has significantly more business registrations than everywhere else, but the top 10 are all looking like they have a good amount.

## 2.2.San Francisco Crime Data

Next, we pull all of San Francisco’s crime data for the last couple decades from the [data SF website](). This is going to help us select one of the safest areas for our bar.

```
In [41]: crime = pd.read_csv('https://data.sfgov.org/api/views/wg3w-h783/rows.csv?accessType=DOWNLOAD')
         print(crime.shape)
         crime.head()
```

```
(294501, 35)
```

Out[41]:

| | Incident Datetime | Incident Date | Incident Time | Incident Year | Incident Day of Week | Report Datetime | Row ID | Incident ID | Incident Number | CAD Number | Report Type Code | Report Type Description | Filed Online |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2019/08/15 11:41:00 AM | 2019/08/15 | 11:41 | 2019 | Thursday | 2019/10/01 02:06:00 PM | 85424006374 | 854240 | 196208089 | NaN | II | Coplogic Initial | True |
| 1 | 2019/09/17 10:00:00 PM | 2019/09/17 | 22:00 | 2019 | Tuesday | 2019/10/02 10:01:00 PM | 85426606374 | 854266 | 196208205 | NaN | II | Coplogic Initial | True |
| 2 | 2019/10/04 02:25:00 PM | 2019/10/04 | 14:25 | 2019 | Friday | 2019/10/04 04:13:00 PM | 85442603474 | 854426 | 190746203 | 192772728.0 | II | Initial | NaN |
| 3 | 2019/10/03 07:30:00 PM | 2019/10/03 | 19:30 | 2019 | Thursday | 2019/10/03 11:25:00 PM | 85419706244 | 854197 | 190744514 | 192764437.0 | II | Initial | NaN |

You can see a total of 294,501 incident reports are contained in this data frame. Let’s do some python again to clean up our data. We want to see the number of incident reports in the last 5 years grouped by neighborhood. This will give us a good indication of how safe or dangerous each area of the city is today. We get the resulting data frame:

| Neighborhood | Incidents |
|---|---|
| Mission | 30385 |
| Tenderloin | 26641 |
| Financial District/South Beach | 25127 |
| South of Market | 22567 |
| Bayview Hunters Point | 14959 |
| North Beach | 8629 |
| Western Addition | 8533 |
| Castro/Upper Market | 7965 |
| Nob Hill | 7499 |
| Sunset/Parkside | 7325 |

## 2.3.San Francisco Neighborhood Map

According to the data SF website, there are 41 neighborhoods in San Francisco. It will vary depending on who you ask, but the neighborhoods are approximately broken out into odd shapes across the peninsula, as seen below.

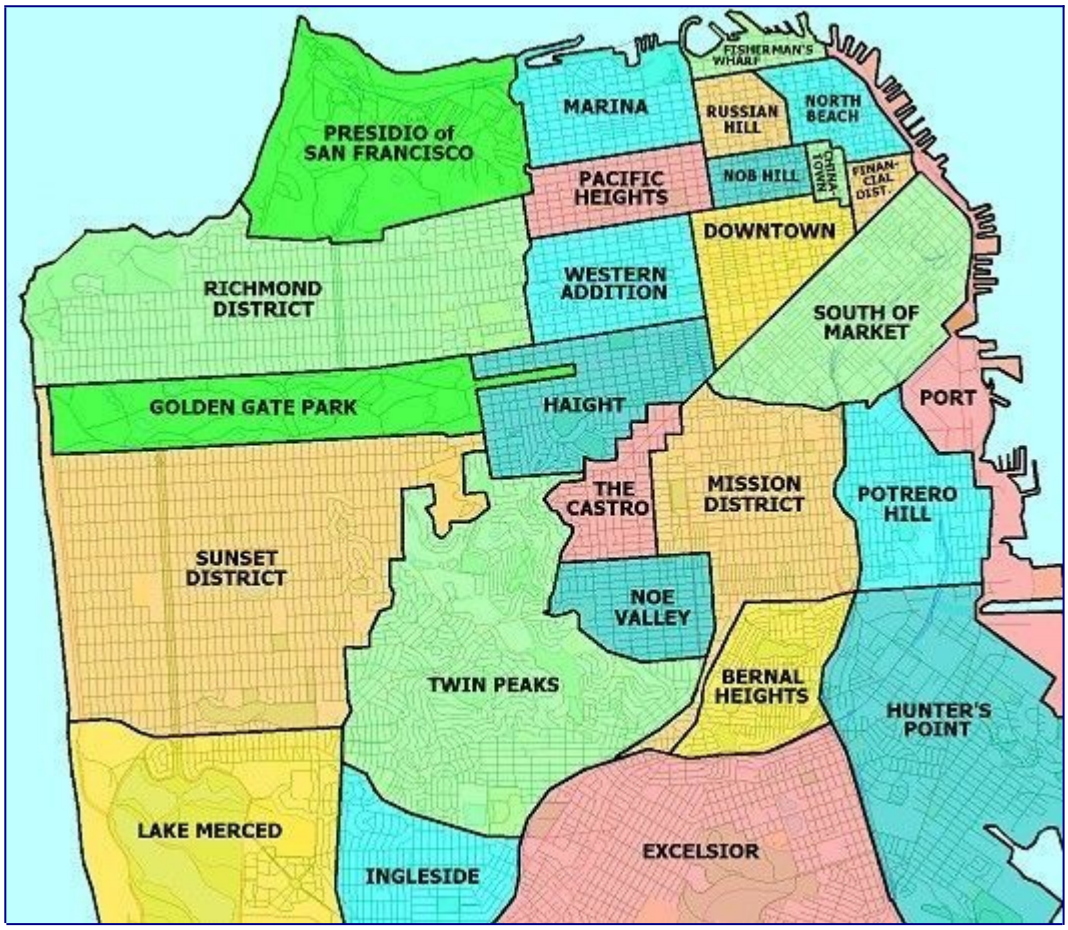


Figure 1: A map of San Francisco with each neighborhood outlined.

Money is also an important factor, so it's going to be useful for us to look at the median price for one month's rent in a one-bedroom apartment in each neighborhood. This will give us a rough idea of how expensive it will be to maintain a bar in each neighborhood. Fortunately, Zumper has this information posted online.

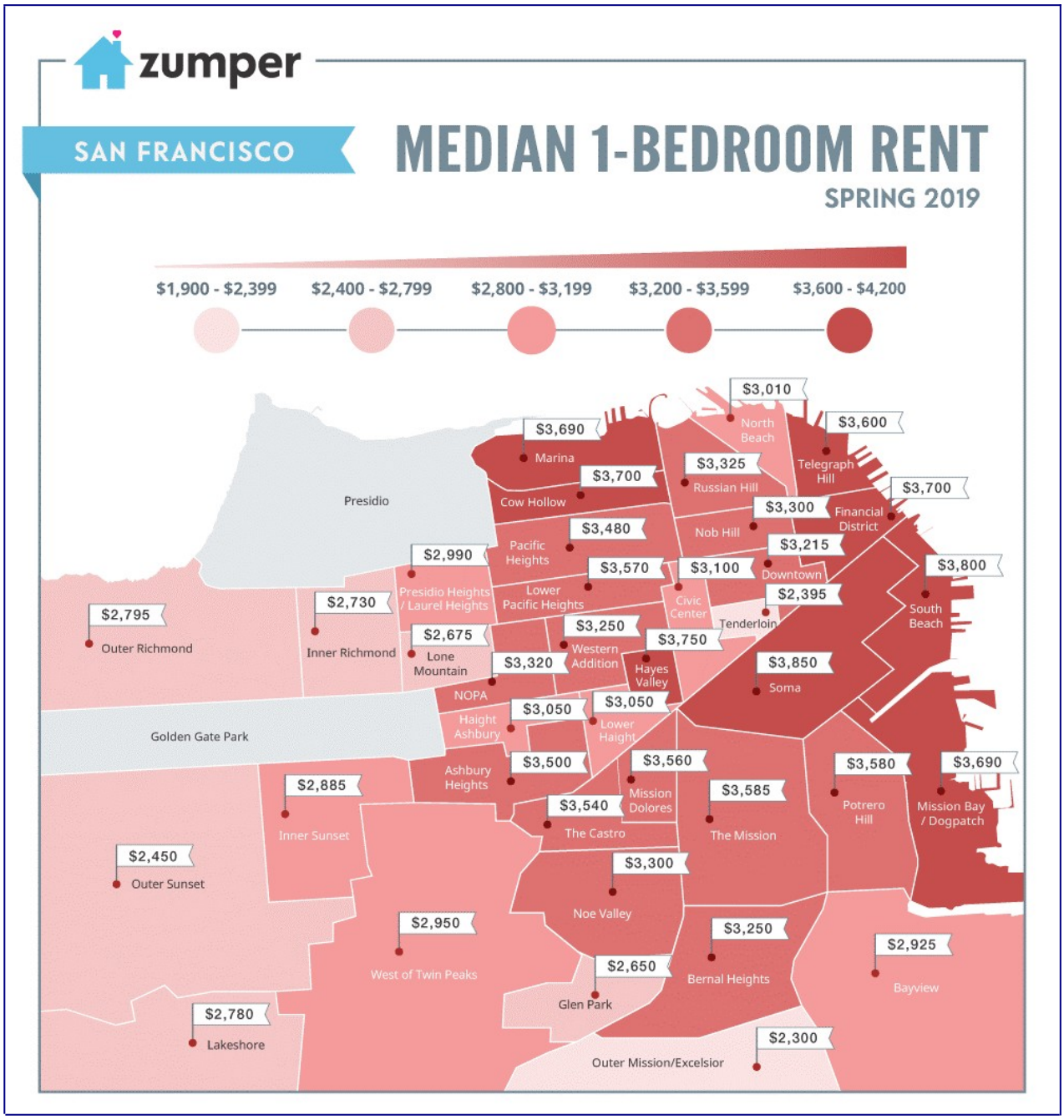


Figure 2: A map of San Francisco with the median rent for a one bedroom in each neighborhood.

# 3. Methodology: Data Visualization and Exploration

## 3.1.Narrowing Down Neighborhoods

We can use some simple visualizations to examine our data sets and narrow down our options for which neighborhood we'd like to host our bar. We want to first see the 10 neighborhoods with the most registered businesses.

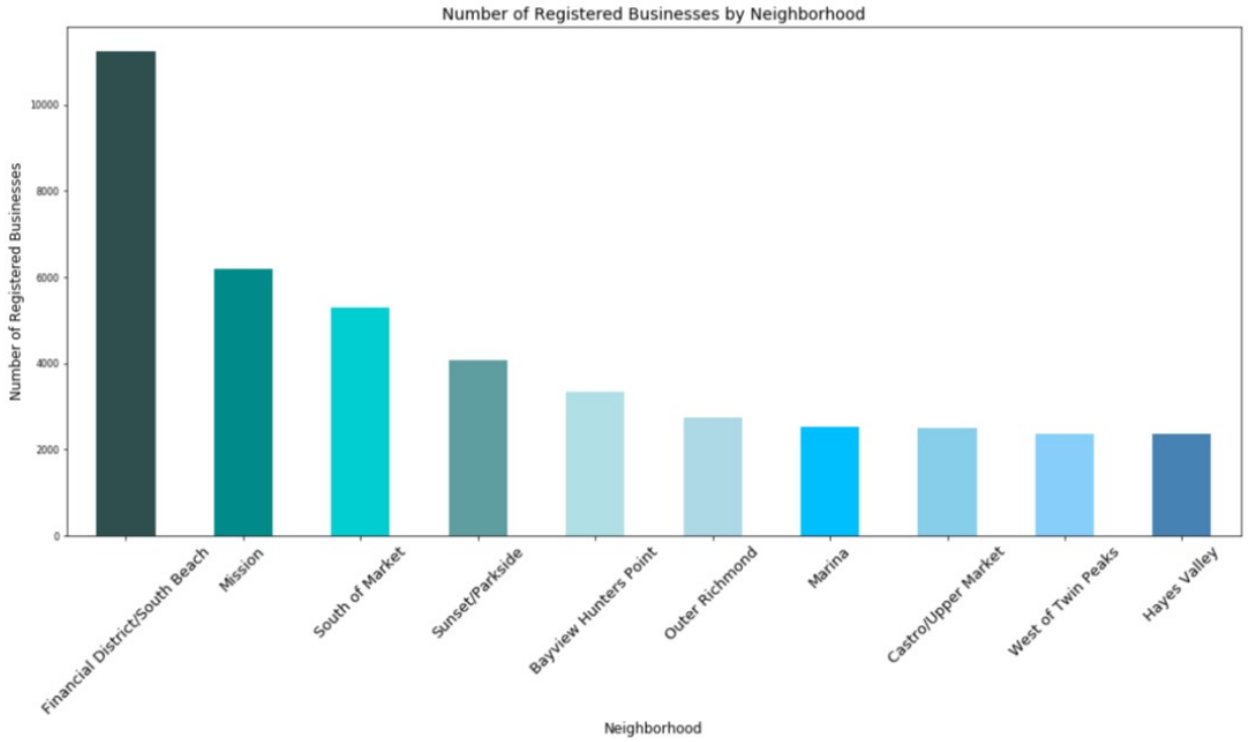


Figure 3: Count of business registered in the last 10 years for each neighborhood in San Francisco, sorted from most to least (showing top 10).

Let's also look at a visualization of neighborhoods that experience the most crime.

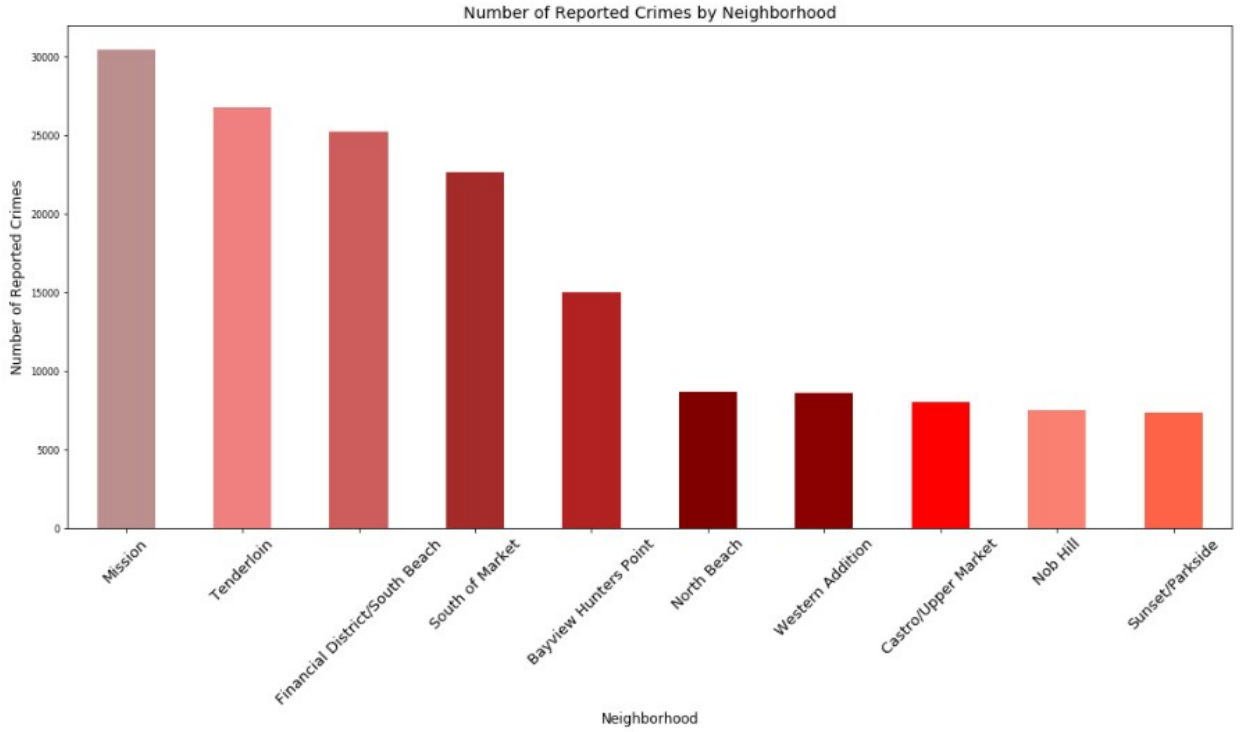


Figure 4: Number of crimes reported in the last five years in each neighborhood in San Francisco, sorted from most to least (showing top 10)

It looks like there is significantly more crime in the first 5 neighborhoods than the rest (Mission, Tenderloin, Financial District, South of Market, and Bayview). When we're opening a bar, we want our patrons to feel safe late at night, so let's try to avoid these 5 neighborhoods. We can do some python to merge our business and crime data frames and get a refined list of neighborhoods with high business registration and without high crime.

In [13]:
```
'''start by merging the datasets and making a new dataset that incldues the neighborhoods
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[13]:

| | Neighborhood | Businesses |
|---|---|---|
| 34 | Sunset/Parkside | 4074 |
| 25 | Outer Richmond | 2729 |
| 16 | Marina | 2528 |
| 2 | Castro/Upper Market | 2478 |
| 39 | West of Twin Peaks | 2352 |

Next, we can run some code to call on our geopy library and pull coordinates for each of these neighborhoods.

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

Not all the coordinates will come out accurately, so we'll have to check with google and pull in coordinates to get it right. Then, we can split the Coordinates column into Latitude and Longitude and add in our Crime data to get a complete picture of our neighborhoods.

| | Neighborhood | Businesses | Crimes | Coordinates | Latitude | Longitude |
|---|---|---|---|---|---|---|
| 0 | Sunset/Parkside | 4074 | 7325 | (37.751616, -122.490810) | 37.751616 | -122.490810 |
| 1 | Outer Richmond | 2729 | 5623 | (37.780001, -122.490229) | 37.780001 | -122.490229 |
| 2 | Marina | 2528 | 6157 | (37.801406, -122.439718) | 37.801406 | -122.439718 |
| 3 | Castro/Upper Market | 2478 | 7965 | (37.762932, -122.435395) | 37.762932 | -122.435395 |
| 4 | West of Twin Peaks | 2352 | 5082 | (37.739871, -122.460106) | 37.739871 | -122.460106 |
| 5 | Hayes Valley | 2349 | 7118 | (37.776685, -122.422936) | 37.776685 | -122.422936 |

It looks like we've 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 data is robust and provides location data for Apple and Uber. Foursquare API will allow us to retrieve information about the most popular spots in each neighborhood in San Francisco. This will be an insightful indication of foot traffic for different venue types. Calling the Foursquare API returns a JSON file, which can be turned into a data frame for analysis in a python notebook.

We can 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):

    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 180 entries, having 30 venues for each of our 6 neighborhoods and 97 unique venue categories:

```
print(SF_venues.shape)
SF_venues.head()
```

```
(180, 7)
```

3]:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Sunset/Parkside | 37.751616 | -122.490810 | S&T Hong Kong Seafood | 37.753702 | -122.491278 | Dim Sum Restaurant |
| 1 | Sunset/Parkside | 37.751616 | -122.490810 | Quan Ngon Vietnamese Noodle House | 37.753624 | -122.490549 | Vietnamese Restaurant |
| 2 | Sunset/Parkside | 37.751616 | -122.490810 | TJ Brewed Tea and Real Fruit (TJ Cups) | 37.753561 | -122.490028 | Bubble Tea Shop |
| 3 | Sunset/Parkside | 37.751616 | -122.490810 | Sunset Recreation Center | 37.757310 | -122.487072 | Playground |
| 4 | Sunset/Parkside | 37.751616 | -122.490810 | Polly Ann Ice Cream | 37.753454 | -122.497765 | Ice Cream Shop |

We can see a graphic representation of the most popular venue categories across all 6 neighborhoods:

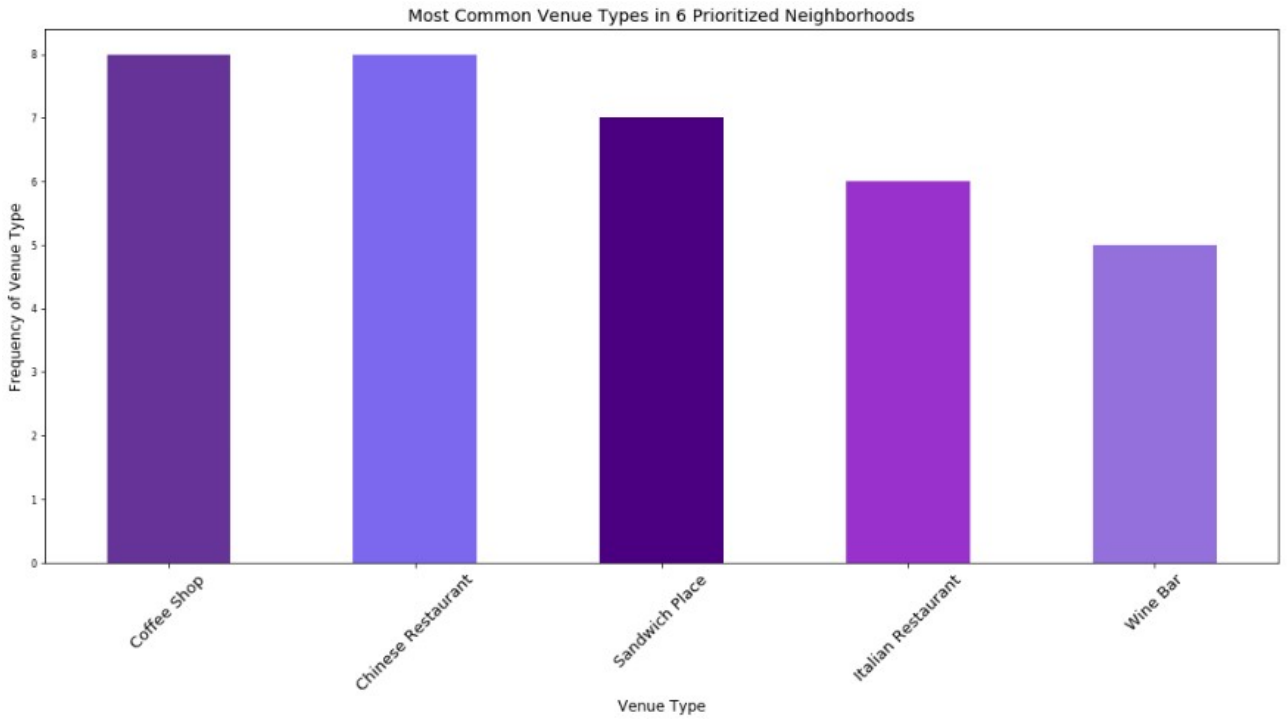


Figure 5: Count of most frequently occurring popular venue types in the 6 prioritized neighborhoods, sorted from most frequent to least (showing top 5).

It looks like coffee shops are the most common popular venue type, followed by various restaurants and then wine bars. So clearly bars are not the most popular type of venue in our neighborhoods overall, but maybe they are more popular in some neighborhoods than others.

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 one-hot encoded venue categories
3. Transpose the 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
```

| | Neighborhood | American Restaurant | Antique Shop | Arts & Crafts Store | Bagel Shop | Bakery | Bank | Bar | Beer Garden | Bookstore |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Castro/Upper Market | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.033333 | 0.000000 | 0.000000 | 0.000000 | 0.033333 |
| 1 | Hayes Valley | 0.000000 | 0.000000 | 0.000000 | 0.033333 | 0.000000 | 0.000000 | 0.000000 | 0.033333 | 0.000000 |
| 2 | Marina | 0.033333 | 0.000000 | 0.033333 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 3 | Outer Richmond | 0.000000 | 0.033333 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | Sunset/Parkside | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.033333 | 0.033333 | 0.000000 | 0.000000 |
| 5 | West of Twin Peaks | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.033333 | 0.000000 | 0.033333 | 0.000000 | 0.033333 |

```
#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_venues))
    print('\n')
```

The code provides us with the top 5 venues for each neighborhood:

```
----Castro/Upper Market----
                   venue  freq
0             Coffee Shop  0.10
1  New American Restaurant  0.07
2             Yoga Studio  0.03
3             Pizza Place  0.03
4          Clothing Store  0.03


----Hayes Valley----
               venue  freq
0  French Restaurant  0.10
1           Wine Bar  0.07
2       Optical Shop  0.07
3        Coffee Shop  0.07
4       Cocktail Bar  0.07


----Marina----
                 venue  freq
0     French Restaurant  0.10
1         Deli / Bodega  0.07
2  Gym / Fitness Center  0.07
3   American Restaurant  0.03
4        Clothing Store  0.03


----Outer Richmond----
                venue  freq
0               Café  0.10
1   Chinese Restaurant  0.07
2       Sandwich Place  0.07
3   Seafood Restaurant  0.07
4  Japanese Restaurant  0.03


----Sunset/Parkside----
                venue  freq
0   Chinese Restaurant  0.13
1           Playground  0.07
2   Dim Sum Restaurant  0.07
3  Japanese Restaurant  0.07
4                Lake  0.03


----West of Twin Peaks----
                venue  freq
0      Sandwich Place  0.07
1  Italian Restaurant  0.07
2                Park  0.07
3       Movie Theater  0.03
4                 Pub  0.03
```

This data is important because it is giving us an idea of the atmosphere of each of these neighborhoods. As someone trying to open a bar, I might want to know whether my location is already a hot spot for other bars and restaurants. So, I'm going to create several data frames from this information and use them to examine the atmosphere of our potential neighborhoods.

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

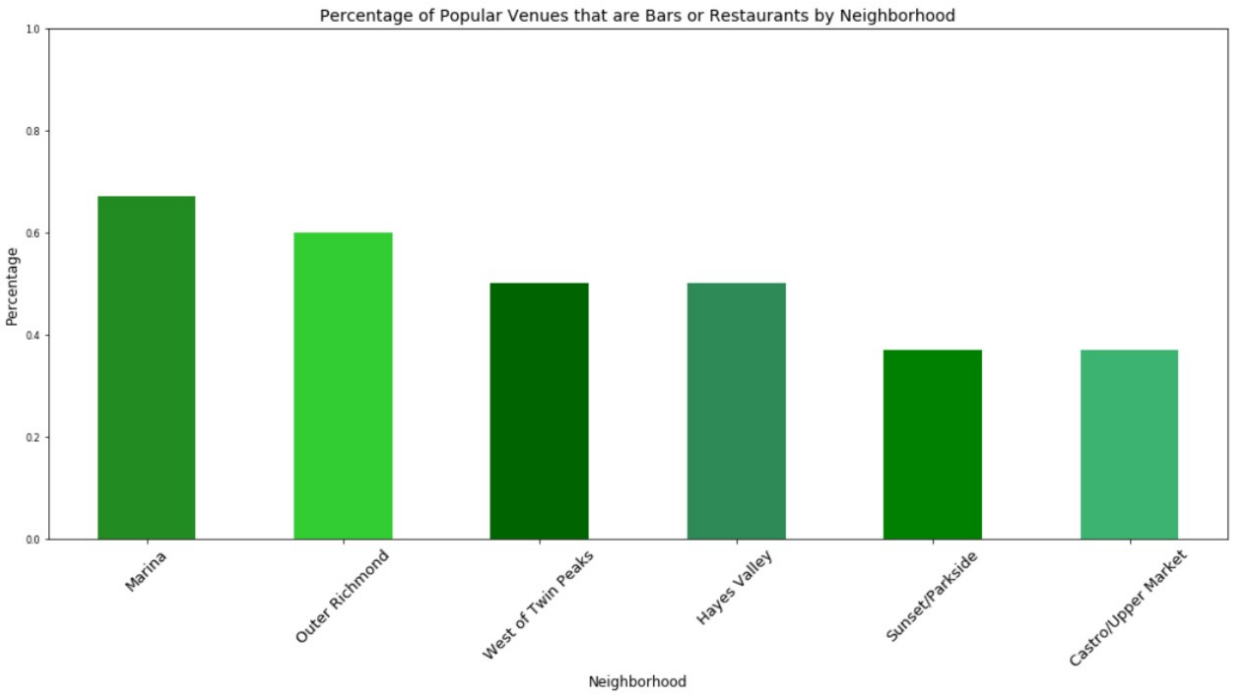


Figure 6: Percentage of top 30 venues in each neighborhood that are either a bar or a restaurant.

This figure tells us which neighborhoods are popular because of the bars and restaurants they have. Neighborhoods like Marina and Outer Richmond have a higher percentage of bars and restaurants as their popular venues, which is an indication that the bar and restaurant scene for these neighborhoods is a major factor for people choosing to visit. Sunset and Castro scored lower, meaning the popular venues in these neighborhoods are majorly other types of facilities like parks or grocery stores.

## 3.3 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 dataframe
kmeans.labels_[0:10]
```

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

```
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

SF_merged = SF
```

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

```
SF_merged['Latitude'] = SF_merged['Latitude'].astype(float)
SF_merged['Longitude'] = SF_merged['Longitude'].astype(float)
SF_merged['Cluster Labels'] = SF_merged['Cluster Labels'].astype(int)



SF_merged
```

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

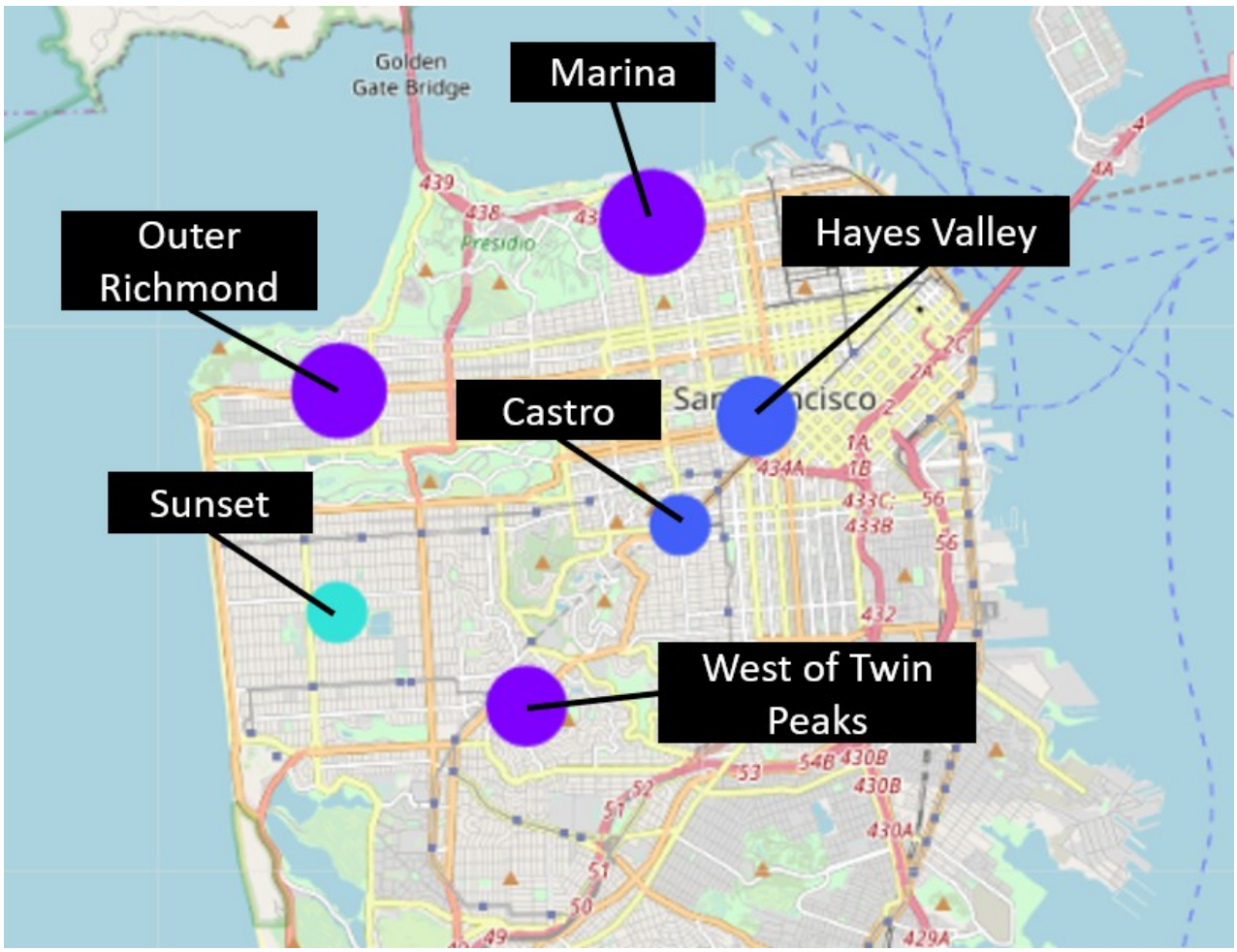


Figure 7: A map of San Francisco with each of our 6 preferred neighborhoods clustered into 3 groups based on the types of popular venues in each neighborhood. The size of each dot represents the number of bars and restaurants listed as popular venues.

It looks like Marina, Outer Sunset, and West of Twin Peaks fit to one cluster, Castro and Hayes Valley fit to a second cluster, and Sunset stands alone as a third.

# 3. Results and Discussion

We have pulled data on crime rates and business registrations for every neighborhood in San Francisco and used this information to narrow down our neighborhood options to 6 neighborhoods. Our analysis has informed us that:

- Coffee shops, Chinese Restaurants, Sandwich Shops, French restaurants, and Wine Bars are the most common venues in our 6 preferred neighborhoods.
- Clustering neighborhoods based on their most popular venues grouped Hayes Valley with Castro into a cluster, West of Twin Peaks, Outer Richmond and Marina into another cluster, and Sunset as its own independent cluster.
- Marina and Outer Richmond have majority bars and restaurants as popular venues, whereas most of the popular venues in Sunset and Castro are not bars or restaurants, but locations like parks and yoga studios.
- From Zumper’s report in Figure 2, we know Marina, Castro, and Hayes Valley are more expensive places to live, with the median rent for a one-bedroom of at least $3,500.

- Sunset, Outer Richmond, and West of Twin Peaks are more affordable and have a median rent for a one-bedroom of less than $3,000.

Based on this analysis, Outer Richmond seems to offer a good balance between foot traffic, popularity for restaurants and bars, and rent prices. Marina seems to be a hot spot for restaurants and bars, but also comes with the high cost of rent. Castro and Hayes Valley come with the same expense as Marina, but neither is as hot of a destination for restaurants and dining. West of Twin Peaks seems to have a similar feel to Marina and Outer Richmond based on clustering but is the least busy neighborhood of the 6. Sunset also has cheaper rent, but the popular spots tend to be more recreational in nature (e.g. lakes and playgrounds).

Ultimately, the optimal bar spot depends on what type of bar you would like to open. An upscale and trendy bar might fare better against competition in an expensive and bustling area like Marina, whereas a dive bar may be the go-to spot in an area like West of Twin Peaks, which likely receives most of its foot traffic exclusively from its residents.

A major drawback of this analysis is that the clustering was completely based on Foursquare’ s data for popular venues. There are plenty other ways to assess popularity of neighborhoods and the spots inside them, venue popularity is just one of them. It may also be helpful to look exclusively at bars in an area, how many there are, and how popular they are on weekdays and weekends.

# 4. Conclusion

Finally, we have executed an end-to-end data science project using common python libraries to manipulate data sets, Foursquare API to explore the neighborhoods of San Francisco, and Folium leaflet map to cluster and segment neighborhoods.