# The Battle of Neighborhoods in San Francisco - Restaurants

Jessie Ying Lingshu

## 1. Introduction/Business Understanding

### 1.1 Discussion of the Background

San Francisco owes its world class status in part to its vibrant tourism and hospitality sectors. Combined with the city's cultural attractions, natural beauty, state-of-the-art convention facilities and strong business base, the tourism and hospitality sectors make San Francisco a top travel destination, playing a vital role in drawing visitors in record numbers. In 2019, San Francisco hosted 26.2 million visitors, contributing an estimated $10.2 billion to the local economy.

San Francisco is as famous for its restaurants and food trends as it is for its Golden Gate Bridge and cable cars. As a traveler, it is always difficult to make a choice from among many options since there is also too much information on the internet, furthermore,  as everybody's got their own take of where to go and it's all so fragmented that you have to assemble it yourself especially if you're interested in non-touristy recommendations.

By leveraging on Foursquare, the location data provider and data science methodology, this project aims to help visitors making decisions on where to find the appropriate neighborhoods for the food and beverage. Meanwhile it will also help visitors to have a good understanding on the most popular venues in San Francisco and where the according neighborhoods are.

### 1.2 Description of the Problem

The project is to help visitors making decisions on where to find the appropriate neighborhoods and having a good understanding on the most popular venues in San Francisco in general and where the according neighborhoods are.

Questions are proposed below:

1. What are the most popular venue categories in San Francisco?
2. What are the most popular restaurants in San Francisco?
3. Where are the neighborhoods to those popular categories?
4. What are the most visited neighborhoods in San Francisco?

## 2. Data Description

For this project we need following data:

1. San Francisco Neighborhoods as Zip Codes, this dataset consists of San Francisco Neighborhood names and zip codes.

   Datasource: http://www.healthysf.org/bdi/outcomes/zipmap.htm

2. United States Zip Codes, this dataset consists of US zip codes, longitude and latitude.

   Datasource: http://docs.gaslamp.media/wp-content/uploads/2013/08/zip_codes_states.csv

3. Foursquare API Data

   By using this Foursquare API we will get all the venues in each neighborhood. We can filter these venues to get such as restaurants, coffee shops or pizza places.

# 3. Methodology

## 3.1 Data Preparation

### 3.1.1 Scraping San Francisco Neighborhoods as ZIP Codes Table from Webpage

The table is scraped by using the BeautifulSoup package. And use pandas to drop the unnecessary column (Population) and format the data frame. It started as below:

```
response = requests.get("http://www.healthysf.org/bdi/outcomes/zipmap.htm")
soup = BeautifulSoup(response.text, "lxml")
table = soup.find_all("table")
df = pd.read_html(str(table))
df = pd.DataFrame(df[4])
```

```
df.head()
```

| | 0 | 1 | 2 |
|---|---|---|---|
| 0 | Zip Code | Neighborhood | Population (Census 2000) |
| 1 | 94102 | Hayes Valley/Tenderloin/North of Market | 28991 |
| 2 | 94103 | South of Market | 23016 |
| 3 | 94107 | Potrero Hill | 17368 |
| 4 | 94108 | Chinatown | 13716 |

After the formatting, it updated as below:

```
df.columns = df.iloc[0]
df = df.iloc[1:-1, :-1]
sf_data = df
sf_data.head()
```

| | Zip Code | Neighborhood |
|---|---|---|
| 1 | 94102 | Hayes Valley/Tenderloin/North of Market |
| 2 | 94103 | South of Market |
| 3 | 94107 | Potrero Hill |
| 4 | 94108 | Chinatown |
| 5 | 94109 | Polk/Russian Hill (Nob Hill) |

Convert the data type of Zip Code to prepare for the following data frame merge.

```
sf_data.dtypes
```

```
0
Zip Code        object
Neighborhood    object
dtype: object
```

```
sf_data[["Zip Code"]] = sf_data[["Zip Code"]].astype("int")
sf_data.dtypes
```

```
0
Zip Code         int64
Neighborhood    object
dtype: object
```

### 3.1.2 Getting Coordinates of San Francisco Neighborhoods

The next objective is to get the coordinates of San Francisco Neighborhoods. The US zip code dataset shows as below:

```
df1.head()
```

| | zip_code | latitude | longitude | city | state | county |
|---|---|---|---|---|---|---|
| 0 | 501 | 40.922326 | -72.637078 | Holtsville | NY | Suffolk |
| 1 | 544 | 40.922326 | -72.637078 | Holtsville | NY | Suffolk |
| 2 | 601 | 18.165273 | -66.722583 | Adjuntas | PR | Adjuntas |
| 3 | 602 | 18.393103 | -67.180953 | Aguada | PR | Aguada |
| 4 | 603 | 18.455913 | -67.145780 | Aguadilla | PR | Aguadilla |

### 3.1.3 Merge the Above Two Data Frames

This is to combine two data frames by matching on the attribute zip code, firstly, let's align the zip code names:

**Sync the zip code attribute name and merge two dataframes**

```
sf_data.rename(columns={'Zip Code':'zipcode'},inplace=True)
sf_data.head()
```

| | zipcode | Neighborhood |
|---|---|---|
| 1 | 94102 | Hayes Valley/Tenderloin/North of Market |
| 2 | 94103 | South of Market |
| 3 | 94107 | Potrero Hill |
| 4 | 94108 | Chinatown |
| 5 | 94109 | Polk/Russian Hill (Nob Hill) |

```
df1.rename(columns={'zip_code':'zipcode'},inplace=True)
df1.head()
```

| | zipcode | latitude | longitude | city | state | county |
|---|---|---|---|---|---|---|
| 0 | 501 | 40.922326 | -72.637078 | Holtsville | NY | Suffolk |
| 1 | 544 | 40.922326 | -72.637078 | Holtsville | NY | Suffolk |
| 2 | 601 | 18.165273 | -66.722583 | Adjuntas | PR | Adjuntas |
| 3 | 602 | 18.393103 | -67.180953 | Aguada | PR | Aguada |
| 4 | 603 | 18.455913 | -67.145780 | Aguadilla | PR | Aguadilla |

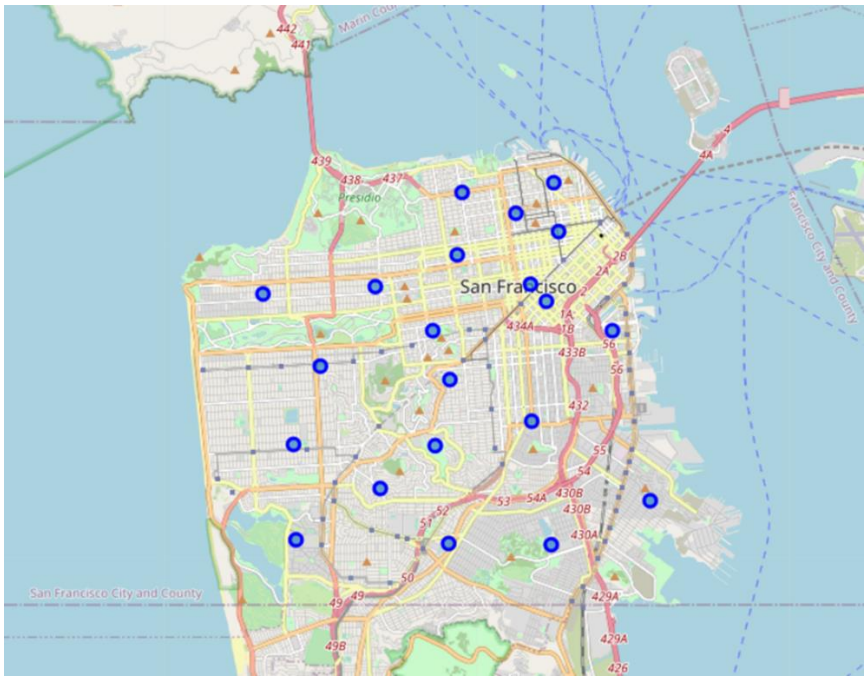
Then we can merge the two data frames successfully:

```
df2 = pd.merge(sf_data, df1, on='zipcode')
df2
```

| | zipcode | Neighborhood | latitude | longitude | city | state | county |
|---|---|---|---|---|---|---|---|
| 0 | 94102 | Hayes Valley/Tenderloin/North of Market | 37.779278 | -122.416582 | San Francisco | CA | San Francisco |
| 1 | 94103 | South of Market | 37.775678 | -122.412131 | San Francisco | CA | San Francisco |
| 2 | 94107 | Potrero Hill | 37.769029 | -122.393681 | San Francisco | CA | San Francisco |
| 3 | 94108 | Chinatown | 37.791028 | -122.408782 | San Francisco | CA | San Francisco |
| 4 | 94109 | Polk/Russian Hill (Nob Hill) | 37.795219 | -122.420782 | San Francisco | CA | San Francisco |

## 3.2 San Francisco Geographic Details Visualization

Python Folium Library is implemented to visualize the San Francisco Geographic Details with provided longitude and latitude of neighborhoods.




## 3.3 Exploratory Data Analysis:

Exploratory data analysis is used here to uncover hidden properties of data and provide useful insights to the San Francisco visitors.

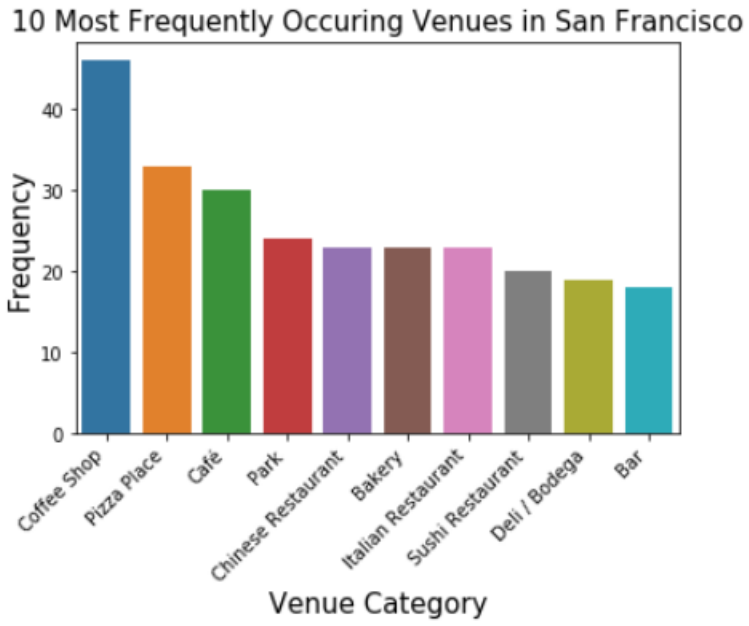
### 3.3.1 Using Foursquare Location Data

Let's make use of Foursquare API to explore San Francisco neighborhoods with the merged data frame.

```
sf_venues.head()
```

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Hayes Valley/Tenderloin/North of Market | 37.779278 | -122.416582 | Asian Art Museum | 37.780178 | -122.416505 | Art Museum |
| 1 | Hayes Valley/Tenderloin/North of Market | 37.779278 | -122.416582 | Orpheum Theatre | 37.779315 | -122.414790 | Theater |
| 2 | Hayes Valley/Tenderloin/North of Market | 37.779278 | -122.416582 | Philz Coffee | 37.781266 | -122.416901 | Coffee Shop |
| 3 | Hayes Valley/Tenderloin/North of Market | 37.779278 | -122.416582 | Ananda Fuara | 37.777693 | -122.416353 | Vegetarian / Vegan Restaurant |
| 4 | Hayes Valley/Tenderloin/North of Market | 37.779278 | -122.416582 | The Strand | 37.779888 | -122.413138 | Theater |

By exploring further, the top 10 most visited venue categories plot shows below:




Let's analyze each neighborhood to know about the top 5 venues of each one. So, we proceed as follows:

1. Create a data-frame with pandas one hot encoding for the venue categories. And use pandas groupby on neighborhood column and calculate the mean of the frequency of occurrence of each venue category.

| | Neighborhood | African Restaurant | American Restaurant | Asian Restaurant | Burmese Restaurant | Cantonese Restaurant | Caribbean Restaurant | Chinese Restaurant | Dim Sum Restaurant | Dumpling Restaurant |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Castro/Noe Valley | 0.000000 | 0.166667 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 1 | Chinatown | 0.000000 | 0.142857 | 0.000000 | 0.000000 | 0.095238 | 0.000000 | 0.047619 | 0.047619 | 0.000000 |
| 2 | Haight-Ashbury | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 3 | Hayes Valley/Tenderloin/North of Market | 0.000000 | 0.153846 | 0.076923 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 4 | Ingelside-Excelsior/Crocker-Amazon | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.142857 | 0.000000 | 0.000000 |

2. Output each neighborhood along with the top 5 most common venues.

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue |
|---|---|---|---|---|---|---|
| 0 | Castro/Noe Valley | Thai Restaurant | Seafood Restaurant | Szechuan Restaurant | Japanese Restaurant | American Restaurant |
| 1 | Chinatown | American Restaurant | Italian Restaurant | Sushi Restaurant | Cantonese Restaurant | Vietnamese Restaurant |
| 2 | Haight-Ashbury | Vegetarian / Vegan Restaurant | Vietnamese Restaurant | Filipino Restaurant | Jewish Restaurant | Japanese Restaurant |
| 3 | Hayes Valley/Tenderloin/North of Market | American Restaurant | Vegetarian / Vegan Restaurant | Vietnamese Restaurant | Southern / Soul Food Restaurant | Malay Restaurant |
| 4 | Ingelside-Excelsior/Crocker-Amazon | Vietnamese Restaurant | Mexican Restaurant | Chinese Restaurant | Latin American Restaurant | Thai Restaurant |
| 5 | Inner Mission/Bernal Heights | Mexican Restaurant | New American Restaurant | American Restaurant | Indian Restaurant | Chinese Restaurant |
| 6 | Inner Richmond | Chinese Restaurant | Sushi Restaurant | Korean Restaurant | Japanese Restaurant | Fast Food Restaurant |
| 7 | Lake Merced | Mexican Restaurant | Vietnamese Restaurant | Filipino Restaurant | Japanese Restaurant | Japanese Curry Restaurant |
| 8 | Marina | Italian Restaurant | French Restaurant | Mexican Restaurant | American Restaurant | Vegetarian / Vegan Restaurant |
| 9 | North Beach/Chinatown | Italian Restaurant | Chinese Restaurant | Sushi Restaurant | Mexican Restaurant | Latin American Restaurant |
| 10 | Outer Richmond | Japanese Restaurant | Chinese Restaurant | New American Restaurant | Ramen Restaurant | Vietnamese Restaurant |

I will use prescriptive analytics (Kmeans Clustering) to help visitors to decide a location to go for a restaurant.

Finally, we try to cluster these 20 districts based on the venue categories and use K-Means clustering. So, our expectation would be based on the similarities of venue categories, these districts will be clustered. I have used the code below:

### Cluster Neighborhoods

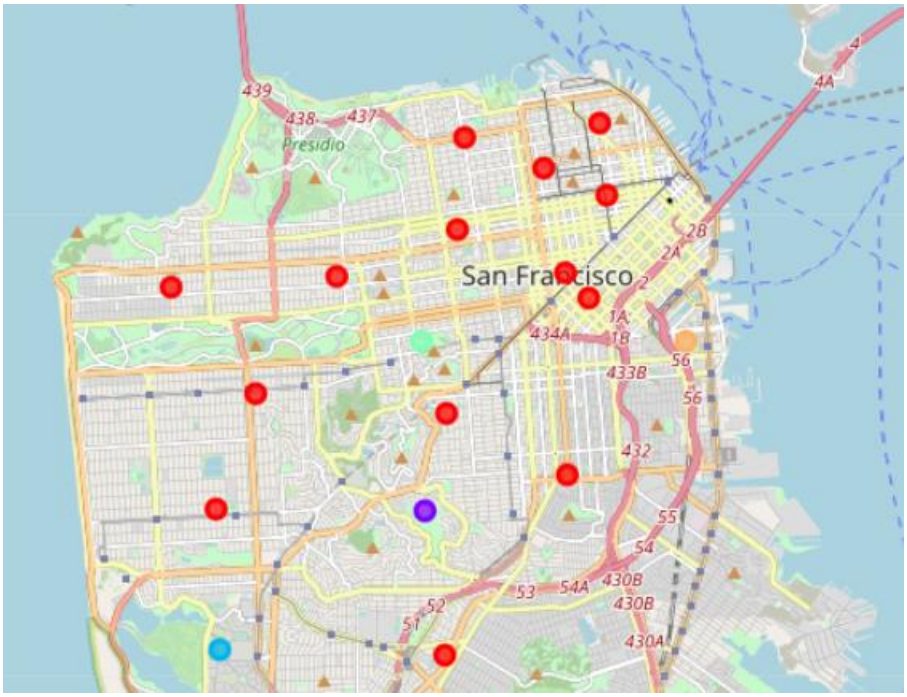
```
# set number of clusters
kclusters = 5

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]
```

```
array([0, 0, 3, 0, 0, 0, 0, 2, 0, 0], dtype=int32)
```

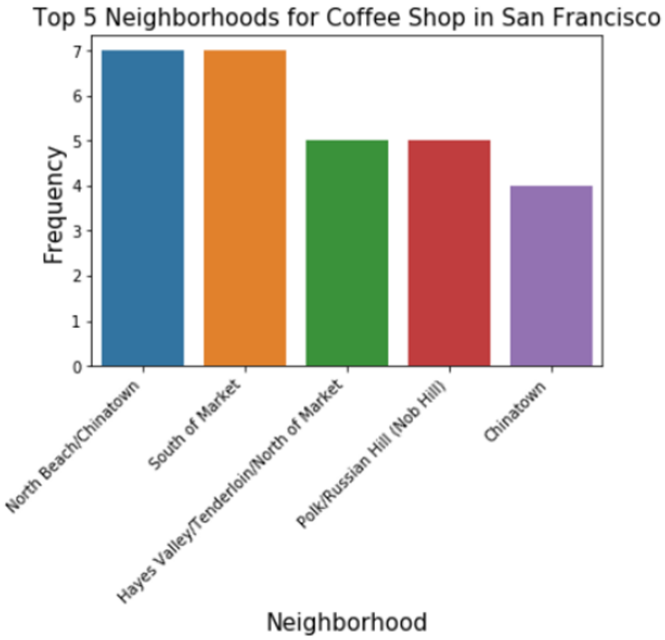
```
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
sf_merged = df2
sf_merged = sf_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')
sf_merged
```

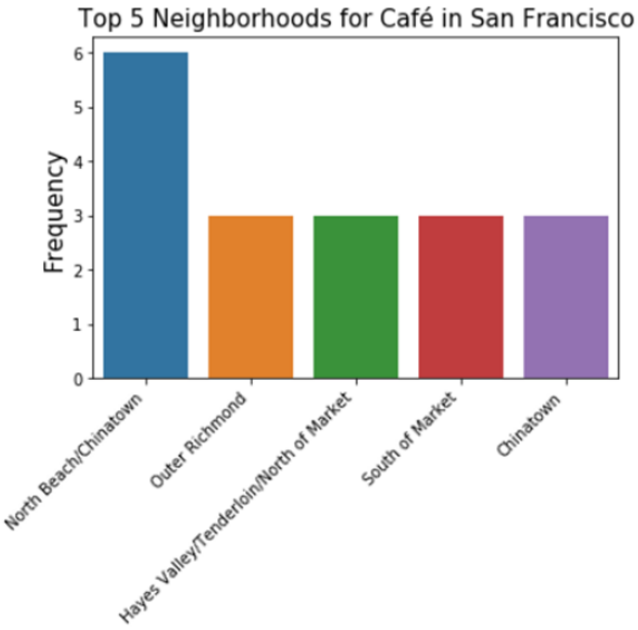
## 4. Results

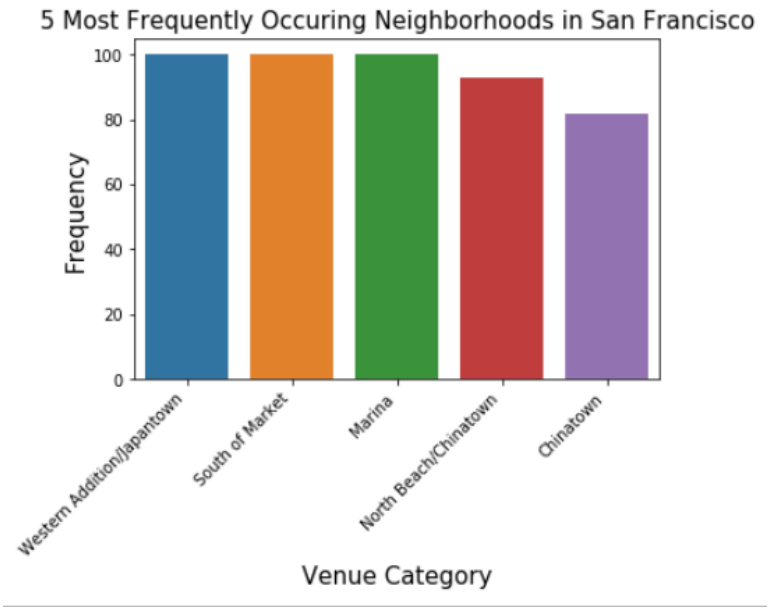
We can represent these 5 clusters in a leaflet map using Folium library as below:




Let’s explore more on the insights based Foursquare San Francisco Neighborhood Data:




## 5. Discussion

A glimpse of the neighborhoods in San Francisco has shown and there are some interesting insights uncovered which might be useful to visitors. Let's summarize our findings:

- The top visited neighborhoods are:
  1. Western Addition/Japantown
  2. South of Market
  3. Marina
  4. North Beach/Chinatown
  5. Chinatown

- It is very interesting to observe that the most popular restaurants in Chinatown are Italian Restaurants, followed by Mexican restaurants.
- Chinese Restaurants, Italian Restaurants and Sushi Restaurants are the most popular restaurants in San Francisco.
- The most popular venue categories are coffee shop, pizza place and café.
- If you want to explore popular coffee shops, you can go to the below neighborhoods:

  1. North Beach/Chinatown
  2. South of Market
  3. Hayes Valley/Tenderloin/North of Market
  4. Polk/Russian Hill (Nob Hill)
  5. Chinatown

- If you want to explore popular pizza places, you can go to the below neighborhoods:
    1. North Beach/Chinatown
    2. Inner Richmond
    3. Western Addition/Japantown
    4. Inner Mission/Bernal Heights
    5. Outer Richmond

- If you want to explore popular cafes, you can go to the below neighborhoods:

    1. North Beach/Chinatown
    2. Outer Richmond
    3. Hayes Valley/Tenderloin/North of Market
    4. South of Market
    5. Chinatown

- North Beach/Chinatown is the right neighborhood for you to explore coffee shops, pizza places and cafes.

## 6. Conclusion

In this analysis, other factors like range of prices of restaurants, restaurants review are ignored, since some data is not available and it would be difficult to farm it for a small exploratory study like this. Hence, this analysis only provide visitors references on the overview of restaurants and other popular places of San Francisco.

In a fast-moving world, there are many real-life problems or scenarios where data can be used to find solutions to those problems. Like seen in the example above, data was used to cluster neighborhoods in San Francisco based on the most common food venues (restaurants, venue categories). The results can help visitors to decide about the neighborhood that fit their needs.

I have made use of some frequently used python libraries to scrap web-data, use Foursquare API to explore the neighborhood of San Francisco and saw the results of segmentation of districts using Folium leaflet map.

This project has shown me a practical example to solve a real-life situation by using Data Science tools and methodologies. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.