# The Battle of Neighborhoods - Final Report

## Applied Data Science Capstone by IBM/Coursera

### Introduction: Business Problem

In the western coast of the United States, there are four major cities: Los Angeles, San Francisco, Portland, and Seattle. These cities attract a lot of people to settle down and travel. For people who are planning to move into one of these cities, foods, shopping and entertainments might be the main factors to be considered. For travelers, they might care more about the hotels to live and places to visit, such as museum and scenic. In this project, we will explore venues around each city, categorize resulting venues, and find the best cities for both future residents and travelers.

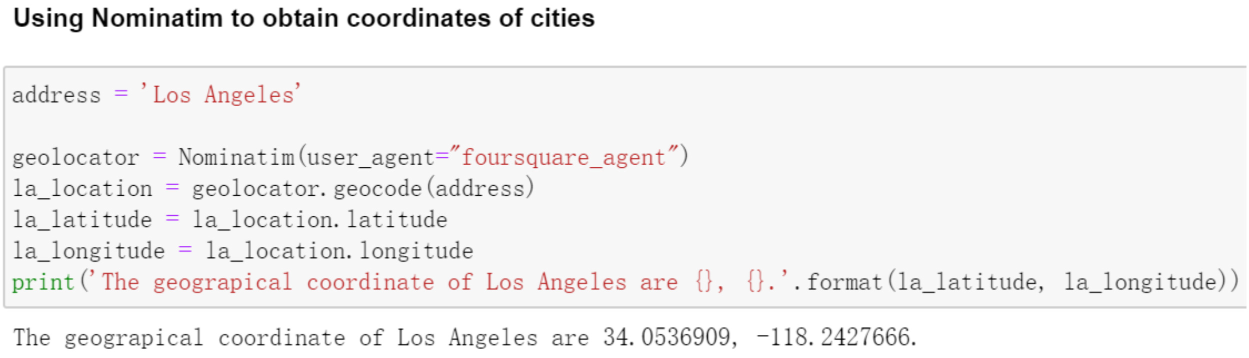
We selected stakeholders who are interested in moving to or travelling the western coast as our target audience. It is entirely possible that people who are selecting a city to live really care about the life convenience, which is part of human life quality and can be largely determined by close distribution of major living facilities. It is also necessary for travelers to think about the number of places worth visiting in the cities they are passionate about.

As is known to all, excluding other factors such as climate and job opportunities, all these four cities are ideal cities to live and travel in people's minds. When people are required to pick one of these, it might cause selecting difficulties. We believe this project will lower the selecting difficulties and provide constructive suggestions for our target audience.

Data Section

The choice of city to live or travel will basically depend on the number of each major category around each city. We selected shopping, food, entertainment, museum, scenic and hotel as six major venue categories in this project.

Before grouping into these six major categories, we will first use the Nominatim geolocation service to obtain the coordinates of Los Angeles, San Francisco, Portland and Seattle.



**Using Nominatim to obtain coordinates of cities**

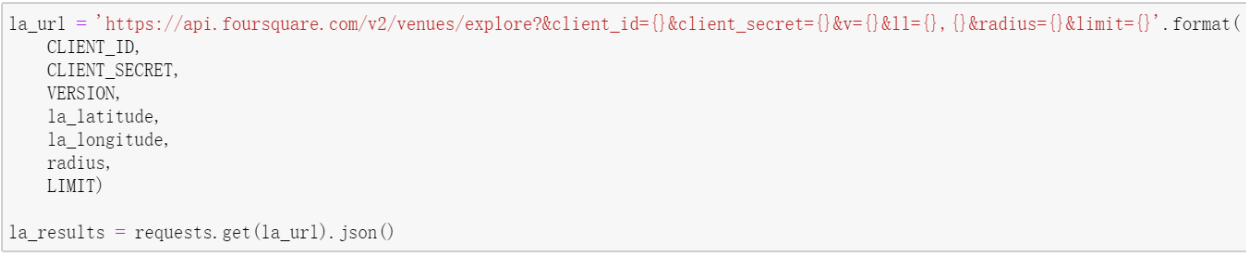
```
address = 'Los Angeles'

geolocator = Nominatim(user_agent="foursquare_agent")
la_location = geolocator.geocode(address)
la_latitude = la_location.latitude
la_longitude = la_location.longitude
print('The geograpical coordinate of Los Angeles are {}, {}.'.format(la_latitude, la_longitude))
```

```
The geograpical coordinate of Los Angeles are 34.0536909, -118.2427666.
```

*Figure 1 Get geographical coordinate of Los Angeles*

Next, we will implement Foursquare API with credentials to get JSON data of venues around each city as our main data source. We will also need to catch relevant attributes and data from raw JSON data and save them into a pandas data frame.



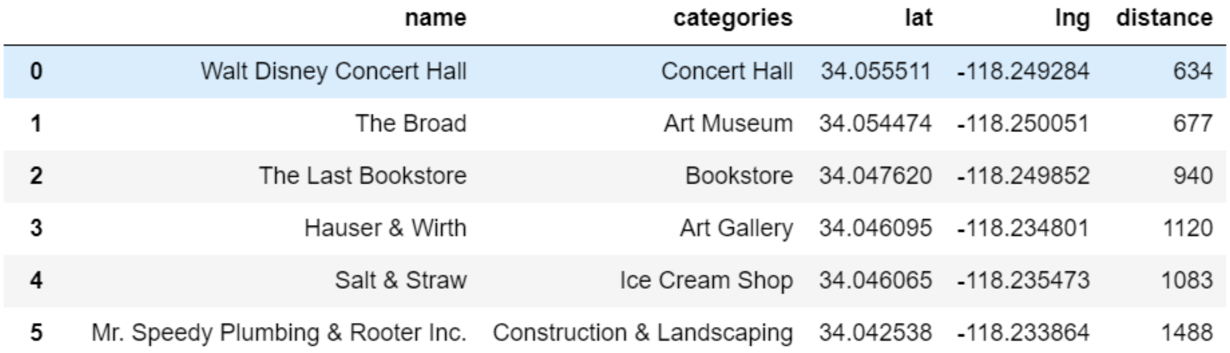
```
la_url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    la_latitude,
    la_longitude,
    radius,
    LIMIT)

la_results = requests.get(la_url).json()
```

*Figure 2 Request raw JSON data of venues using Foursquare API*

Methodology Section

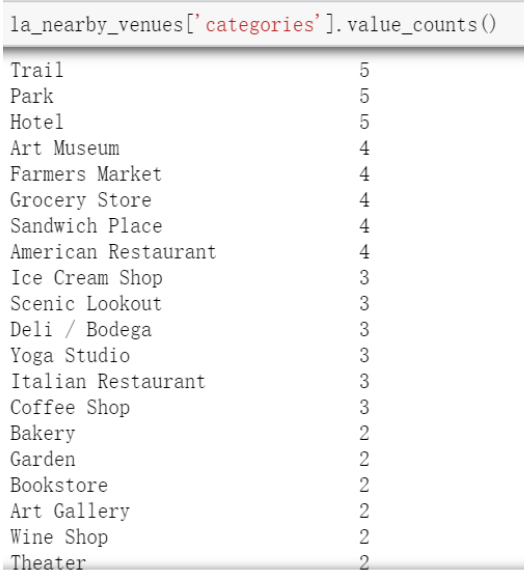
First of all, we requested raw JSON data of up to 250 venues by searching a radius of 100 kilometers (62 miles) of each city. Cleaning and extracting only needed data from raw JSON data that we requested from Foursquare API is definitely the first step of our data preprocessing. We implemented feature engineering to extract important features from raw JSON data (i.e. Venue name, category, latitude, longitude, and distance). After the process of feature engineering we eventually saved these data into a pandas data frame, which helps to present the data more clearly and succinctly. The data frame shown below is the venue information around the Los Angeles area.



| | name | categories | lat | lng | distance |
|---|---|---|---|---|---|
| 0 | Walt Disney Concert Hall | Concert Hall | 34.055511 | -118.249284 | 634 |
| 1 | The Broad | Art Museum | 34.054474 | -118.250051 | 677 |
| 2 | The Last Bookstore | Bookstore | 34.047620 | -118.249852 | 940 |
| 3 | Hauser & Wirth | Art Gallery | 34.046095 | -118.234801 | 1120 |
| 4 | Salt & Straw | Ice Cream Shop | 34.046065 | -118.235473 | 1083 |
| 5 | Mr. Speedy Plumbing & Rooter Inc. | Construction & Landscaping | 34.042538 | -118.233864 | 1488 |

*Figure 3 Save nearby venue data of Los Angeles into a pandas data frame*

In order to group the venues into six major categories, we will first need to see what venue categories a city might contain and how many venues each category have by using value_counts().

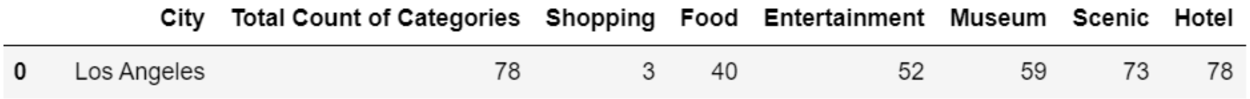


```
la_nearby_venues['categories'].value_counts()
```

```
Trail                              5
Park                               5
Hotel                              5
Art Museum                         4
Farmers Market                     4
Grocery Store                      4
Sandwich Place                     4
American Restaurant                4
Ice Cream Shop                     3
Scenic Lookout                     3
Deli / Bodega                      3
Yoga Studio                        3
Italian Restaurant                 3
Coffee Shop                        3
Bakery                             2
Garden                             2
Bookstore                          2
Art Gallery                        2
Wine Shop                          2
Theater                            2
```

*Figure 4 Use value_counts() to group resulting venue categories of Los Angeles*

Since we only have a single column (aka. attribute) about venue categories, we will not be able to use computer power to do the clustering process. In other words, we will need to define six clusters by ourselves. We started with manually created six lists showing six major categories (i.e. Shopping, Food, Entertainment, Museum, Scenic, and Hotel) and adding unique venue categories obtaining from the Foursquare API (see Figure 4) to the corresponding lists based on our knowledge. After we define each cluster, we then counted how many venues of each city of four belong to each major category of six. We put the counting results in a pandas date frame as shown below.



| | City | Total Count of Categories | Shopping | Food | Entertainment | Museum | Scenic | Hotel |
|---|---|---|---|---|---|---|---|---|
| 0 | Los Angeles | 78 | 3 | 40 | 52 | 59 | 73 | 78 |

*Figure 5 Clustering results of venues in Los Angeles*

Lastly, we used six different colors representing six major venue categories and implemented folium to visualize venues on the map of each city using same mapping scale (i.e. zoon_start = 11). The mapping results will show us the distributions of these venue and how convenient people can reach these venues.

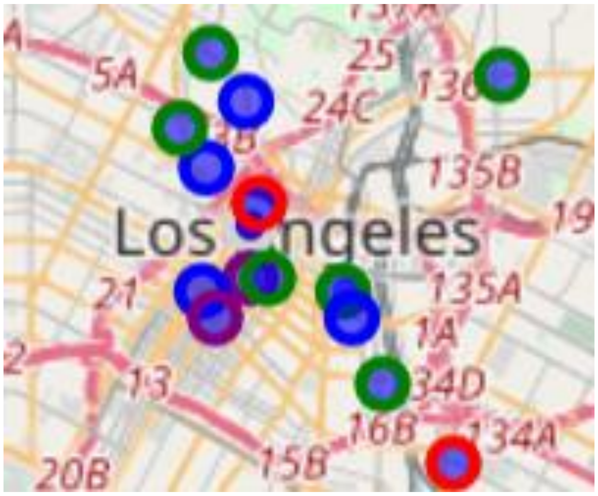
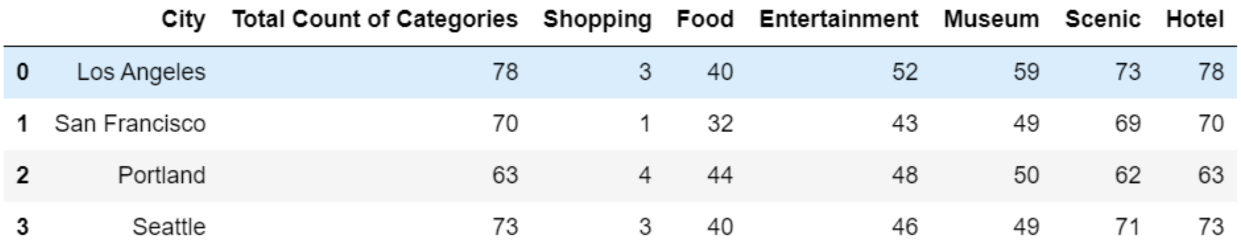


*Figure 6 Mapping venues of Los Angeles downtown area*

Result Section

The first result from our data analytical process is the pandas data frame showing the numbers of six major venue categories in each city of four.



| | City | Total Count of Categories | Shopping | Food | Entertainment | Museum | Scenic | Hotel |
|---|---|---|---|---|---|---|---|---|
| 0 | Los Angeles | 78 | 3 | 40 | 52 | 59 | 73 | 78 |
| 1 | San Francisco | 70 | 1 | 32 | 43 | 49 | 69 | 70 |
| 2 | Portland | 63 | 4 | 44 | 48 | 50 | 62 | 63 |
| 3 | Seattle | 73 | 3 | 40 | 46 | 49 | 71 | 73 |

*Figure 7 Clustering results of venues in Los Angeles, San Francisco, Portland and Seattle*

Next, we will present the mapping results of clusters (aka. major venue categories) in each city. Six colors will be used for different clusters on the maps. Each major category and its corresponding color would be:

- Shopping: Red
- Food: Green
- Entertainment: Purple
- Museum: Lightgray
- Scenic: Blue
- Hotel: Yellow

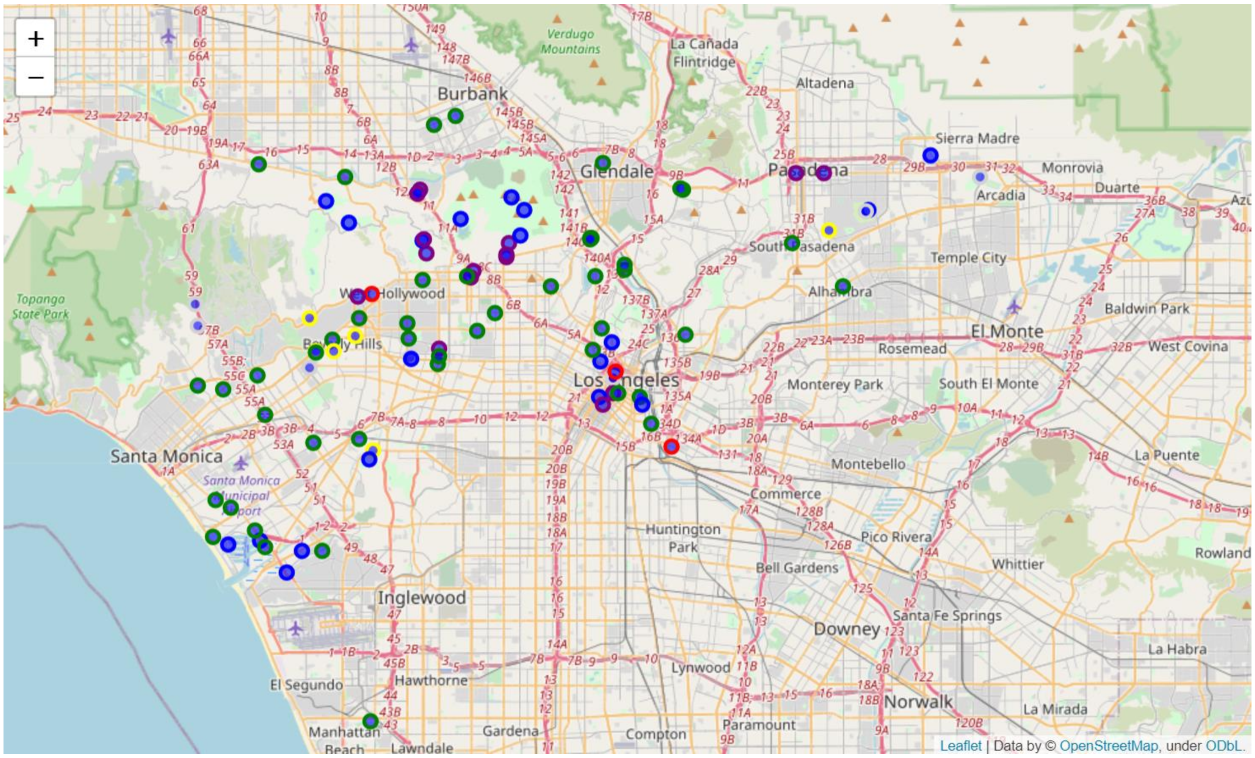


*Figure 8 Mapping venues of Los Angeles, CA*

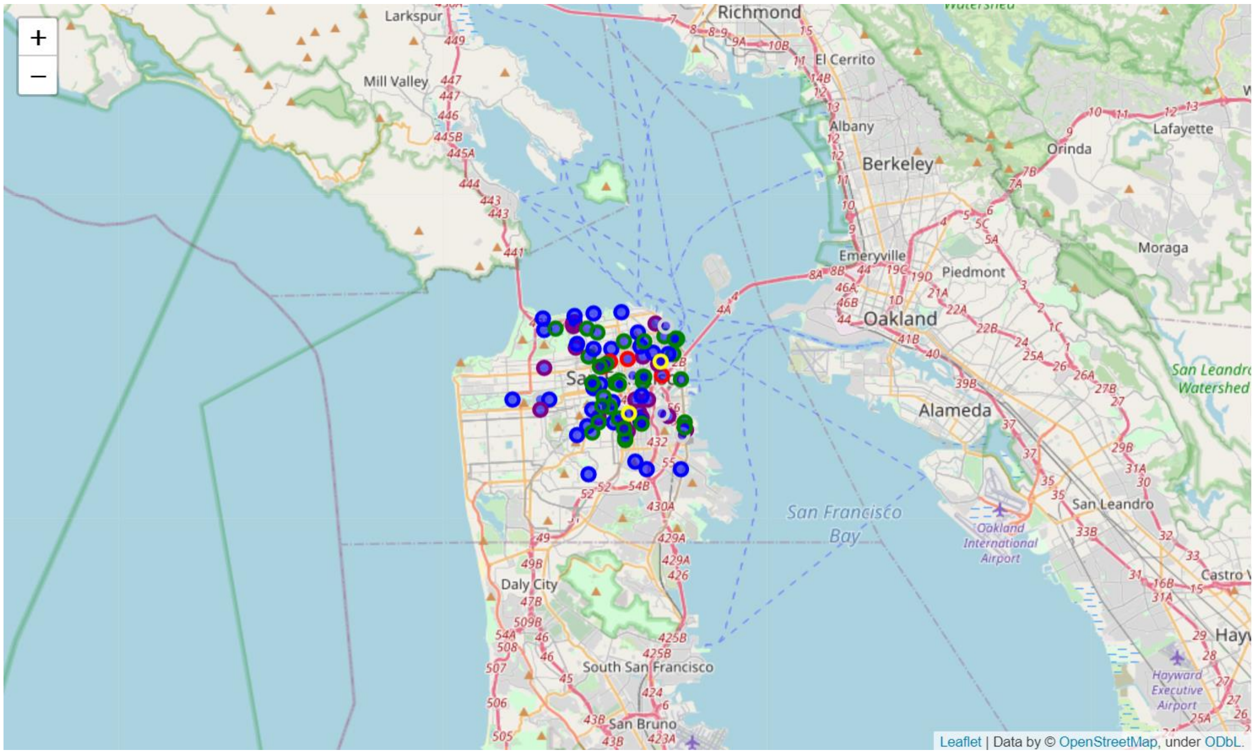


*Figure 9 Mapping venues of San Francisco, CA*

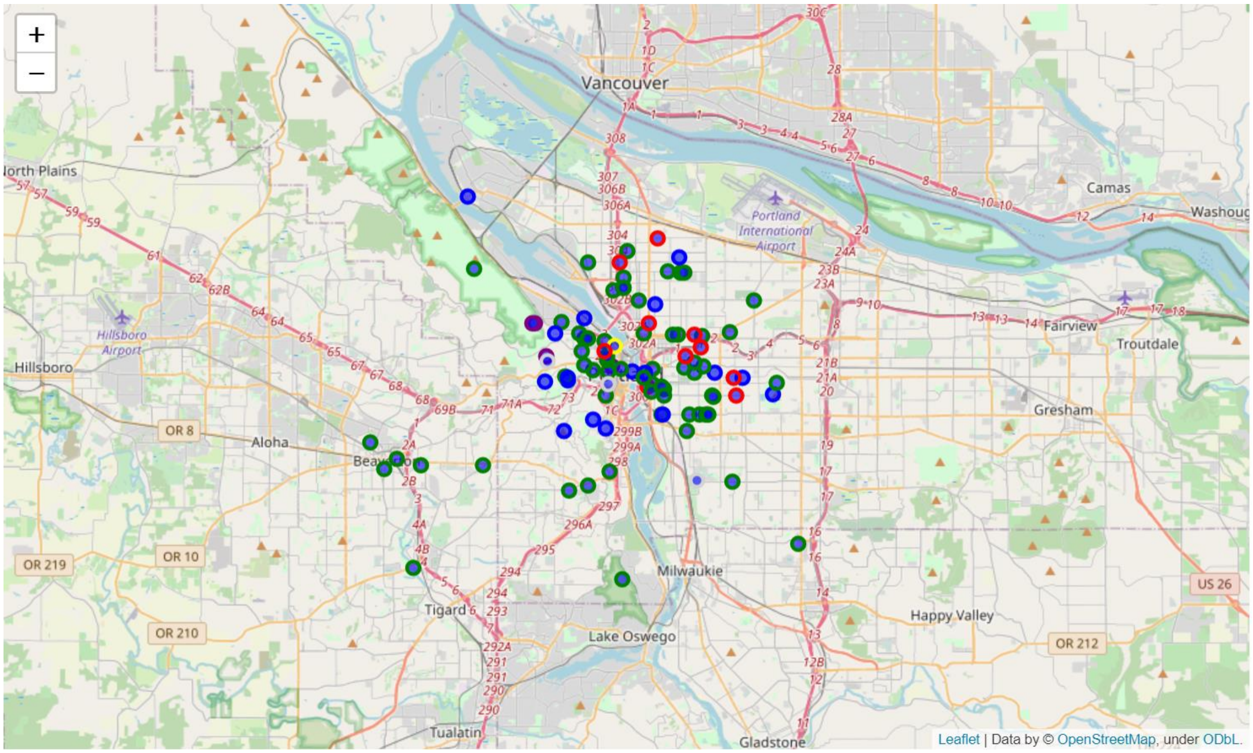


*Figure 10 Mapping venues of Portland, OR*

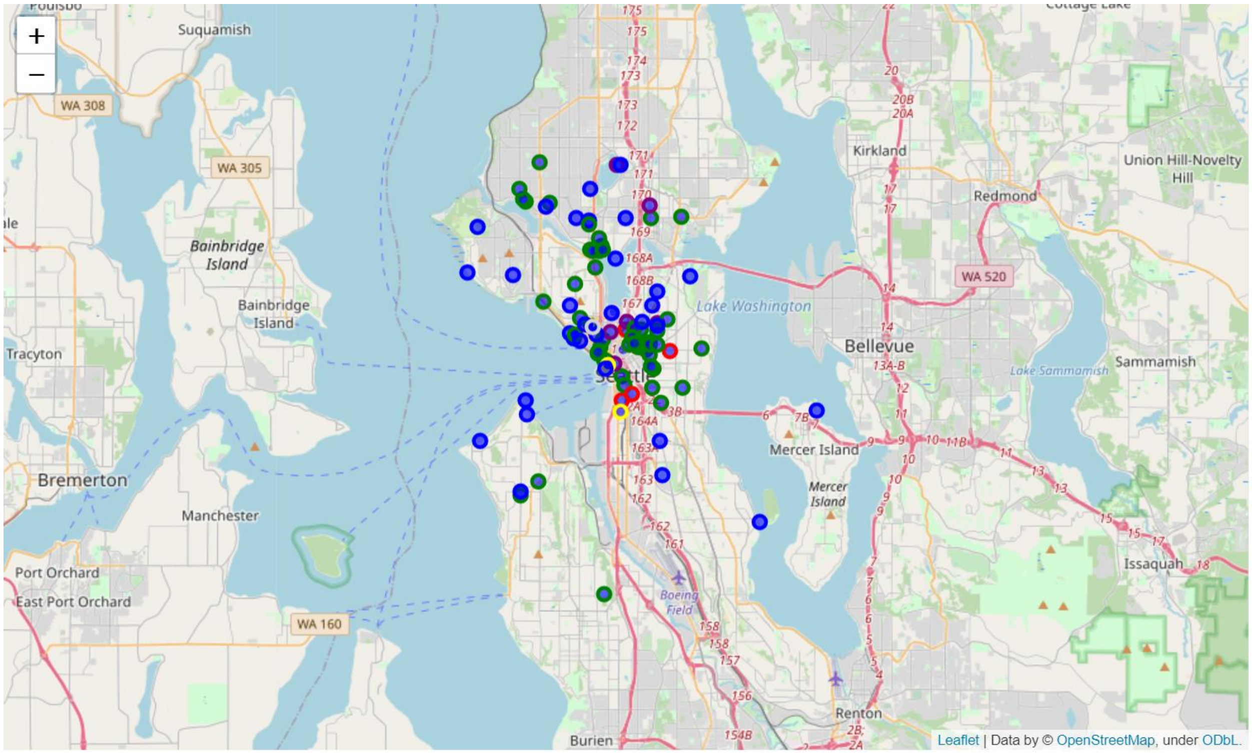


*Figure 11 Mapping venues of Seattle, WA*

## Discussion Section

From the data and visualization results, we have some findings and suggestions to our target audience, who are future residents and travelers.

Los Angeles has the maximum number of venues lying in the categories of Entertainment, Museum, and Scenic, but the distribution these venues are less dense. Residents might take more time to food spots or entertainment spots in a different area of Los Angeles. However, the greatest number of scenic spots and widest choice of hotels prove Los Angeles the worthiest city to visit. Hence, based on the data we have, I will recommend travelers to have Los Angeles as their first choice.

San Francisco has considerable numbers of food spots and entertainments spots that are densely distributed on the map. People who are living in San Francisco have the most convenient lives since they have a lot of food and entertainment choices in vicinity. Therefore, San Francisco could be the best city for future residents to settle down.

## Conclusion Section

From the data analysis and visualization results of nearby venues of Los Angeles, San Francisco, Portland and Seattle, we conclude Los Angeles as the best city to travel due to a great number of scenic spots and flexible choice of hotel. We believe tourists will have a good time travelling Los Angeles. We also consider San Francisco as the most convenient city to live because of its closely distributed living facilities, such as entertaining places and restaurants, along with a large number of scenic views.