Assisting International Assignment Department in Identifying Suitable Neighborhoods for Employee

Chiew Seng Chun

May 27th, 2020

1. Introduction

1.1. Background

Many big corporations will have times when they need to relocate an employee from one country to another country as part of international assignment. Relocating the employee and their family can be a tricky task if it is not managed correctly. The family may not feel settled in the new area. There is a huge cost for the company as a mis-managed relocation can cause lost productivity of the employee and the employee not benefitting fully from the purpose of the international assignment which is to broaden the skills of the employee.

1.2. Problem

The international assignment department which handles this can provide recommendation of areas where employees can stay in the new location with a view that it may help alleviate some of the concerns. These recommendations are based on past experiences they have which may be a sound approach. However, such recommendation may not suit all employees. *How can the department provide more useful recommendations of where to live?* One way is to use the employee's current dwelling location and use that as a basis of comparison in recommending neighborhoods with similar characteristics. For this assignment, we will attempt to find out neighborhoods in another city which is similar to the location that an employee is currently living.

1.3. Audience

This report will be targeted to stakeholders within the international assignment department and the employee due to relocate.

1.4. Scenario

The following scenario has been developed for the purpose of this assignment.

“Imagine John Doe, a Microsoft employee who has worked in London, UK for a long time, has recently been reassigned to San Francisco, California in the United States. Due to the reassignment, John Doe and his family will be relocating to San Francisco. John Doe has a specific requirement to stay in an area in San Francisco that is similar to where their family is staying now in London. John Doe currently stays near the Richmond Underground tube station and the Microsoft Headquarters in San Francisco is at 555 California St 200, San Francisco, CA 94104, United States. John Doe would also like to stay somewhere close to the new office. The international assignment department now has to provide a list of recommended neighborhoods in San Francisco for John Doe.”

2. Data

2.1. Data Sources

Every individual has specific requirements when choosing an area to live in. This can be school rating, venues available around the area, housing prices, availability of houses to buy/rent, crime data etc.

For this assignment, we will be looking at venues around the area and will use the following data:

- Neighborhood data for San Francisco

This is obtained using the geojson data in the following the link: https://data.sfgov.org/Geographic-Locations-and-Boundaries/SF-Find-Neighborhoods/pty2-tcw4.The geojson data has the name of the neighborhood and the geometry of the neighborhood. We will use the geometry to find the centroid which will then be used to find nearby venues.

- Foursquare API Venues data

The Foursquare venues data will be used to identify venues within a certain area of a location. We will be interested in the venue categories for neighborhoods within San Francisco and near Richmond underground station. Venue categories is a good way of characterising a particular location. There may be different nomenclature of between Richmond, UK and San Francisco neighborhoods which we will have to make common to ensure appropriate clustering.

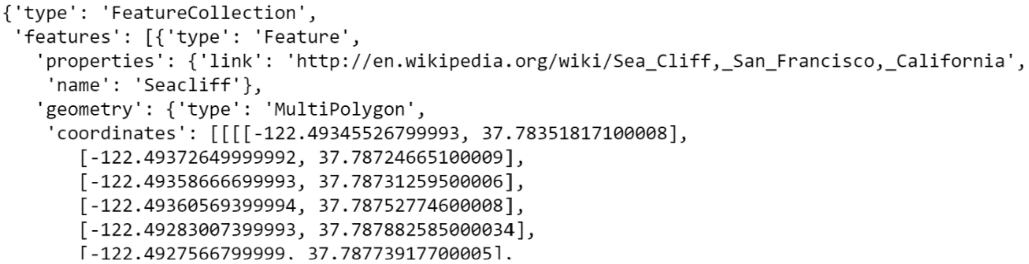
- Geodata

We will be using Geopy to identify the geodata for Richmond Underground Station, the new workplace and the distance between the neighborhoods and the new workplace.

2.2. Data Cleaning

Looking at the geojson neighborhood data for San Francisco, there is a need to clean up the data to extract the geodata of each neighborhood. As can be seen from the figure below, the name of the neighborhood is under properties->name. This is followed by the geometry of the neighborhood with coordinates that matches the vertexes that form the boundary of the neighborhood on a map. Using any of the coordinates as the latitude and longitude of the neighborhood will be useless as it is at the perimeter of the neighborhood. The coordinates however can be used to find the centroid of the neighborhood.

To find the centroid, the shapely library will be used. Under the library, there is a Polygon class with a centroid method that can be used by passing a list of coordinates and it will return the

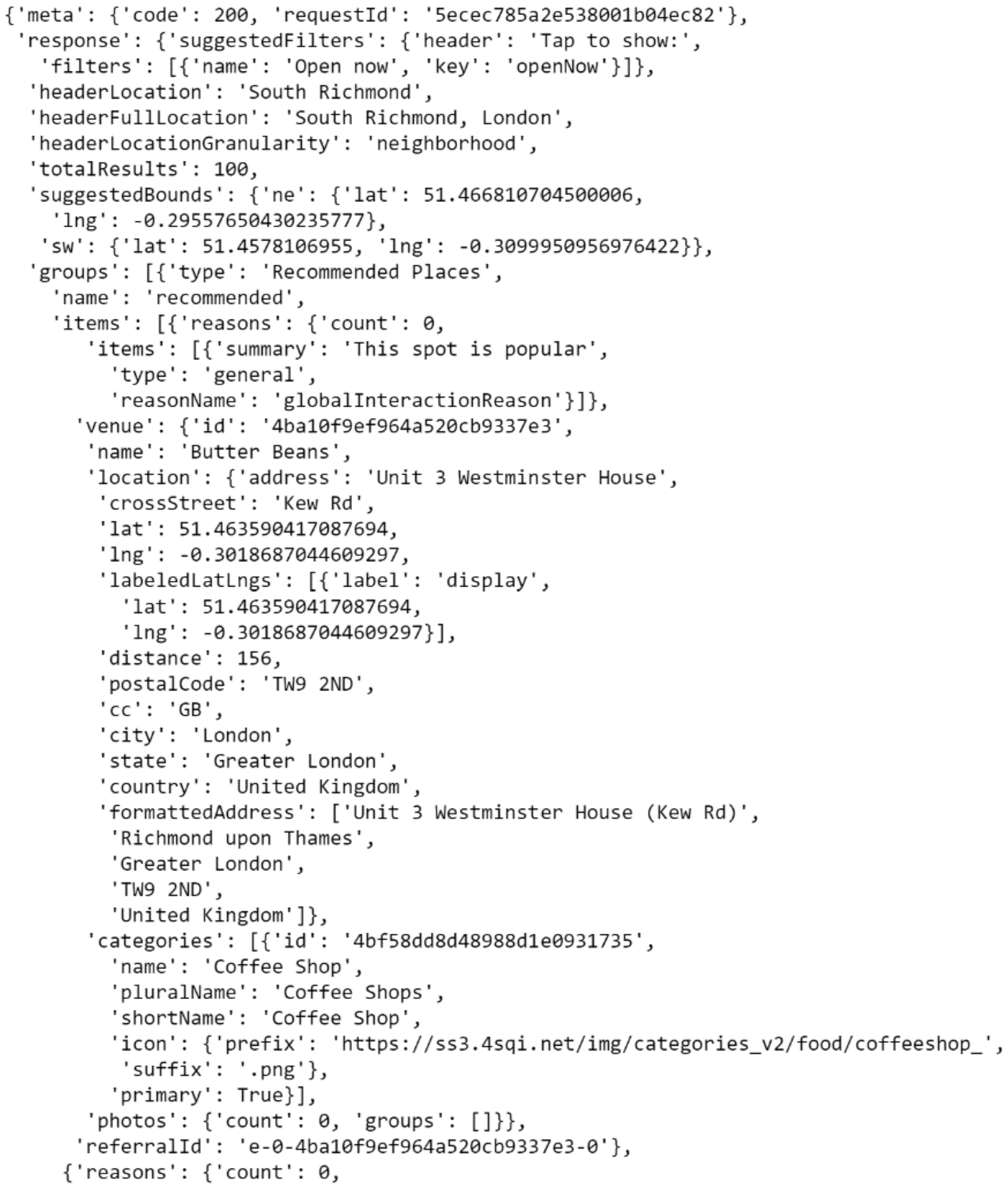


```
{'type': 'FeatureCollection',
 'features': [{'type': 'Feature',
   'properties': {'link': 'http://en.wikipedia.org/wiki/Sea_Cliff,_San_Francisco,_California',
    'name': 'Seacliff'},
   'geometry': {'type': 'MultiPolygon',
    'coordinates': [[[[-122.49345526799993, 37.78351817100008],
       [-122.49372649999992, 37.78724665100009],
       [-122.49358666699993, 37.78731259500006],
       [-122.49360569399994, 37.78752774600008],
       [-122.49283007399993, 37.787882585000034],
       [-122.4927566799999, 37.78773917700005],
```

Figure 1: Geojson data

centroid coordinates for the neighborhood. This centroid coordinate can be verified by putting it in Google Maps.

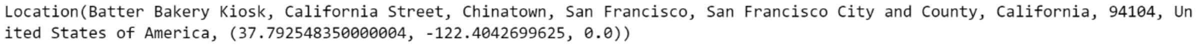
For the Foursquare API data, the data that we want is located under venue. As per the figure below, the data of interest are the venue name, venue latitude, venue longitude and venue categories name. The parameter for the API call will limit the results return to within 500m of the location provided. The maximum information returned by Foursquare API per venue call is 100 venues of interest in JSON format. The results will be parsed through a json method to allow population of a dataframe for later analysis



```
{'meta': {'code': 200, 'requestId': '5ecec785a2e538001b04ec82'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'South Richmond',
  'headerFullLocation': 'South Richmond, London',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 100,
  'suggestedBounds': {'ne': {'lat': 51.466810704500006,
    'lng': -0.29557650430235777},
   'sw': {'lat': 51.4578106955, 'lng': -0.3099950956976422}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4ba10f9ef964a520cb9337e3',
       'name': 'Butter Beans',
       'location': {'address': 'Unit 3 Westminster House',
        'crossStreet': 'Kew Rd',
        'lat': 51.463590417087694,
        'lng': -0.3018687044609297,
        'labeledLatLngs': [{'label': 'display',
          'lat': 51.463590417087694,
          'lng': -0.3018687044609297}],
        'distance': 156,
        'postalCode': 'TW9 2ND',
        'cc': 'GB',
        'city': 'London',
        'state': 'Greater London',
        'country': 'United Kingdom',
        'formattedAddress': ['Unit 3 Westminster House (Kew Rd)',
         'Richmond upon Thames',
         'Greater London',
         'TW9 2ND',
         'United Kingdom']},
       'categories': [{'id': '4bf58dd8d48988d1e0931735',
         'name': 'Coffee Shop',
         'pluralName': 'Coffee Shops',
         'shortName': 'Coffee Shop',
         'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/coffeeshop_',
          'suffix': '.png'},
         'primary': True}],
       'photos': {'count': 0, 'groups': []}},
      'referralId': 'e-0-4ba10f9ef964a520cb9337e3-0'},
     {'reasons': {'count': 0,
```

Figure 2: Foursquare API call

The geodata from geopy will be used to obtain specific geodata of a particular location. For this assignment, it is known that the address of Richmond Underground Station is The Quadrant, Richmond TW9 1EZ. This is found using Google Search. The address for the Microsoft Headquarters in San Francisco is 555 California St 200, San Francisco, CA 94104, United States, also found using Google Search. The geodata could have been obtained using Google Maps but for this assignment, the geopy library is used. Specifically, the Nominatim geocoder will be used as it is free. Passing the above addresses to the geocoder will return the location found, the address, the latitude and the longitude of the location as shown in the figure below. It is the latitude and longitude which is of interest.



```
Location(Batter Bakery Kiosk, California Street, Chinatown, San Francisco, San Francisco City and County, California, 94104, Un
ited States of America, (37.792548350000004, -122.4042699625, 0.0))
```

Figure 3: Geopy result

3. Methodology

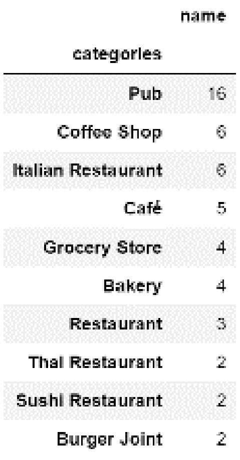
3.1. Tackling the problem

Based on the scenario in 1.4 above, the following is the plan of action that will be reflected in the code notebook.

- Obtain location of Richmond Underground Station and explore the area within 500m of the station
- Obtain location of neighborhoods for San Francisco.
- Explore each neighborhoods and add the explored data from Richmond.
- Cluster the neighborhoods and highlight which San Francisco cluster is similar to Richmond.
- Within that cluster, sort neighborhoods by distance to Microsoft Headquarters.

3.2. Analysis

On obtaining the latitude and longitude of Richmond Underground Station, the json result was parsed and put into a dataframe. Within 500m of that location, 100 venues of interest were found of which there are 53 unique venue categories. Out of the 53 unique venue categories, pubs were the highest with 16 count, followed by coffee shop at 6 count and Italian restaurant at 6 count. Below is a figure showing the top ten venue categories found in Richmond.



	name
categories	
Pub	16
Coffee Shop	6
Italian Restaurant	6
Café	5
Grocery Store	4
Bakery	4
Restaurant	3
Thai Restaurant	2
Sushi Restaurant	2
Burger Joint	2

Figure 4: Top ten venue categories near Richmond Underground Station

For this assignment one would expect that a similar neighborhood would have the same characteristics.

The geojson data for San Francisco neighborhoods were parsed and the centroid for the each neighborhood was established and put into a dataframe. There are 117 neighborhoods identified in San Francisco. A geographical representation was created using Folium to visualise the 117 neighborhoods in San Francisco as per the figure below.

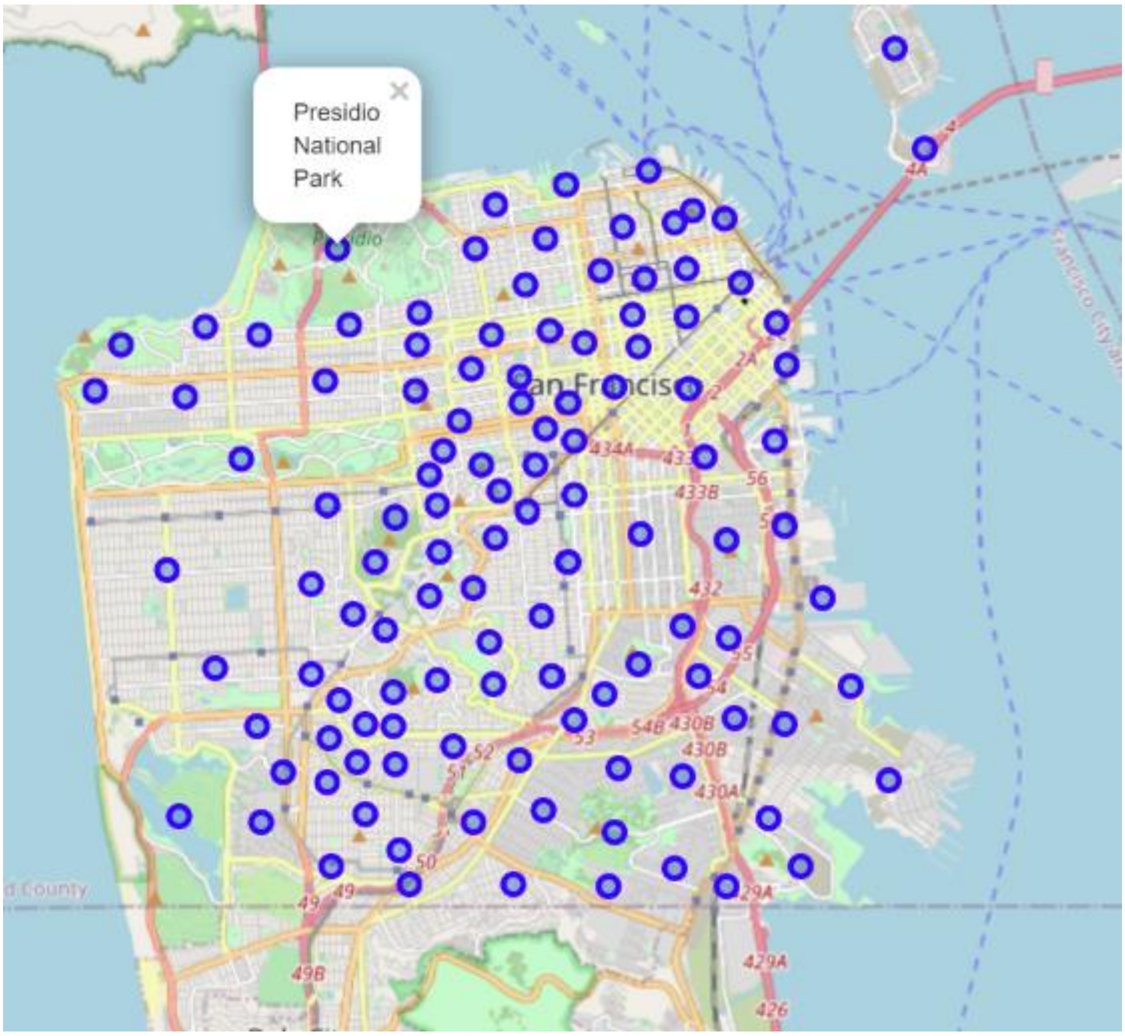


Figure 5: Location of neighborhoods in San Francisco

As can be seen, there is no overlap of neighborhoods. There are some neighborhoods which will be close. The radius limit of 500m set for the Foursquare API call should ensure there is minimal duplication of venues. To confirm that the centroid is approximately the center of the neighborhood, Google Maps have been used as verification. In the above figure, a point has been picked: Presidio National Park. The figure below shows a highlighted area of the Presidio of San Francisco. Comparing both, it can be seen that the centroid computed using the Shapely library matches approximately to the center of Presidio of San Francisco in Google Maps.

Once the centroid has been verified, the latitude and longitude of the centroid is pass through as a parameter for the Foursquare API call. This is an iterative process to get results from all 117 neighborhoods. Once the results are obtained, it is presented in a dataframe. To ensure that no further calls are required, the results were stored locally as a csv file for future data reload.

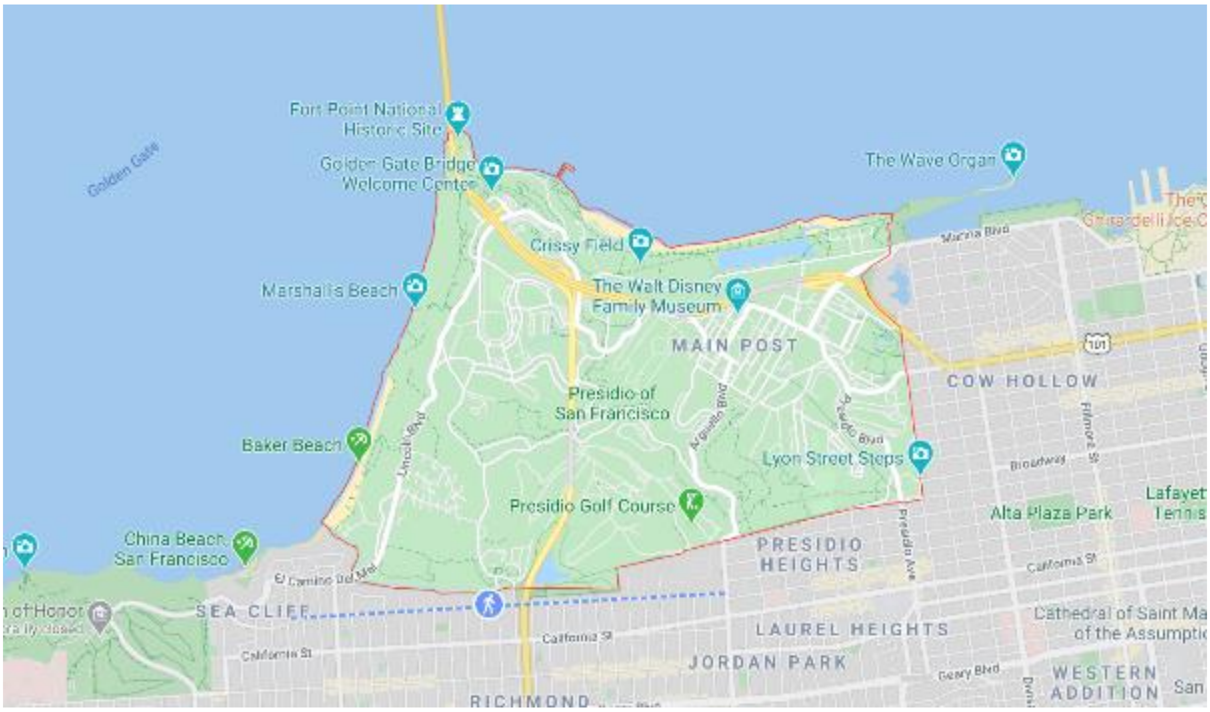
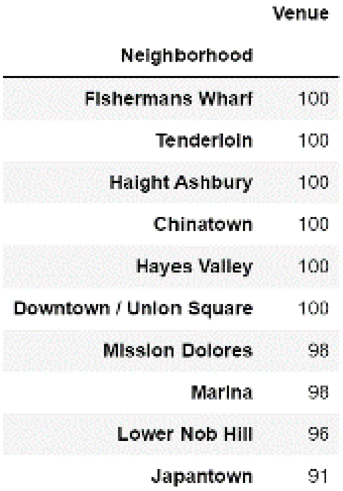


Figure 6: Google Map search of Presidio of San Francisco

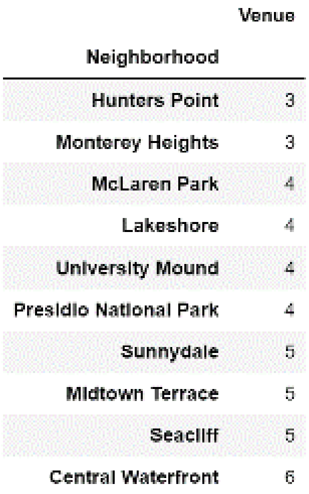
The result contained 4632 venues spread across 117 neighborhoods each with a 500m radius search. Out of the 117 neighborhoods, Fishermans Wharf, Tenderloin, Haight Ashbury, Chinatown, Hayes Valley and Downtown/Union Square all returned 100 venues. Below is a figure showing the top ten neighborhood with the highest venue count.



	Venue
Neighborhood	
Fishermans Wharf	100
Tenderloin	100
Haight Ashbury	100
Chinatown	100
Hayes Valley	100
Downtown / Union Square	100
Mission Dolores	98
Marina	98
Lower Nob Hill	96
Japantown	91

Figure 7: Top ten neighborhoods with highest venue count

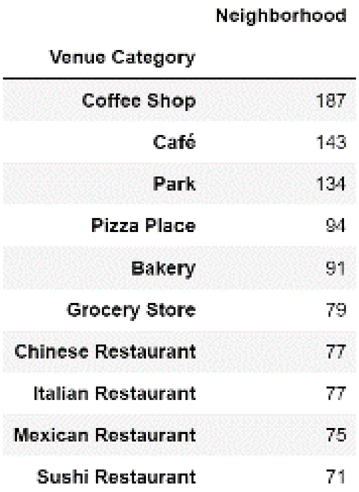
The following neighborhoods returned the least venues.



	Venue
Neighborhood	
Hunters Point	3
Monterey Heights	3
McLaren Park	4
Lakeshore	4
University Mound	4
Presidio National Park	4
Sunnydale	5
Midtown Terrace	5
Seacliff	5
Central Waterfront	6

Figure 8: Bottom ten neighborhoods with lowest venue count

In all San Francisco neighborhoods, 368 unique venue categories were returned. How can we compare against Richmond? Fishermans Wharf returned 100 venues so it can be used as a comparison. Filtering the dataframe to just Fishermans Wharf revealed that there are 55 unique categories. The same was done on Tenderloin neighborhood and it returned 54 unique categories. This is comparable to Richmond's 53 unique categories. Hence, for the whole of San Francisco neighborhoods having 368 unique venue categories may not seem to far-fetched. Analysis of all 368 unique venue categories revealed that 187 were coffee shops, 143 were cafes, 134 were parks and 94 were pizza places. The top 10 venue category across all neighborhoods are shown in the figure below.

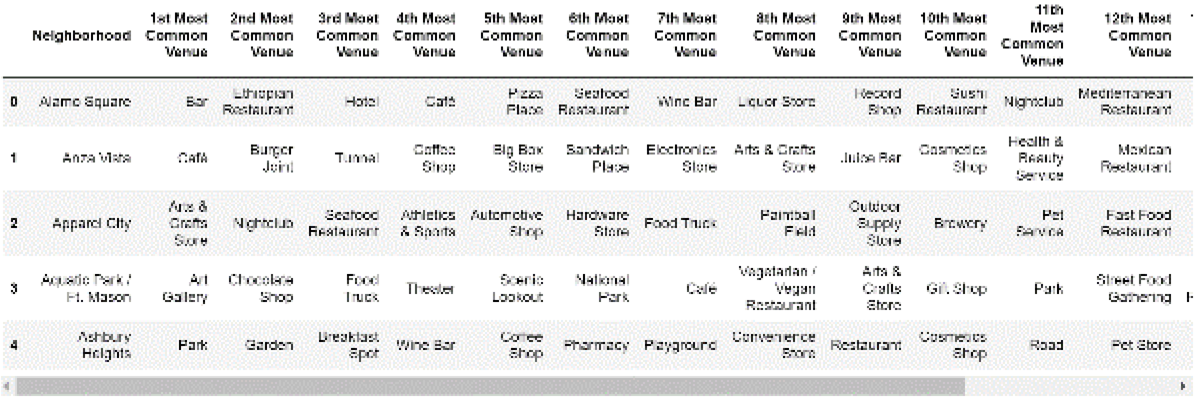


	Neighborhood
Venue Category	
Coffee Shop	187
Café	143
Park	134
Pizza Place	94
Bakery	91
Grocery Store	79
Chinese Restaurant	77
Italian Restaurant	77
Mexican Restaurant	75
Sushi Restaurant	71

Figure 9: Top ten venue categories across all San Francisco neighborhoods

Are all 53 Richmond's unique venue categories in San Francisco's 368 unique venue categories? Analysis reveals that 49 venue categories out of the 53 are common across Richmond and San Francisco's neighborhoods. This is good as it means Richmond can be clustered somewhere within San Francisco. If it was a lot less say half of 53, then it can be difficult to ensure suitable clustering. Further analysis reveals that 4 San Francisco neighborhood do not have any of the 49 venue categories that is common between Richmond and San Francisco. The neighborhoods are Hunters Point, Midtown Terrace, Monterey Heights and Seacliff. This is reflected by the fact that they are within the bottom ten neighborhoods with the lowest venue count. As such, it is expected that Richmond will not be clustered with them.

Once the San Francisco venues are loaded, Richmond venues are appended to the dataframe to prepare for clustering. One hot encoding is used to pivot all unique venue categories into columns and then grouped by neighborhoods to get an idea of the average occurrences of each venue category in each neighborhoods. The dataframe is confirmed to be 118 rows (San Francisco neighborhoods + Richmond) and 372 columns (368 unique venue categories in San Francisco + 4 unique venue categories in Richmond). The data is then transformed to identify the top 15 venue categories for each neighborhood as shown in the figure below.



	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue
0	Alamo Square	Bar	Ethiopian Restaurant	Hotel	Café	Pizza Place	Seafood Restaurant	Wine Bar	Liquor Store	Record Shop	Sushi Restaurant	Nightclub	Mediterranean Restaurant
1	Anza Vista	Café	Burger Joint	Tunnel	Coffee Shop	Big Box Store	Sandwich Place	Electronics Store	Arts & Crafts Store	Juice Bar	Cosmetics Shop	Health & Beauty Service	Mexican Restaurant
2	Apparel City	Arts & Crafts Store	Nightclub	Seafood Restaurant	Athletics & Sports	Automotive Shop	Hardware Store	Food Truck	Paintball Field	Outdoor Supply Store	Brewery	Pet Service	Fast Food Restaurant
3	Aquatic Park / Ft. Mason	Art Gallery	Chocolate Shop	Food Truck	Theater	Scenic Lookout	National Park	Café	Vegetarian / Vegan Restaurant	Arts & Crafts Store	Gift Shop	Park	Street Food Gathering
4	Ashbury Heights	Park	Garden	Breakfast Spot	Wine Bar	Coffee Shop	Pharmacy	Playground	Convenience Store	Restaurant	Cosmetics Shop	Road	Pet Store

Figure 10: Example showing neighborhood and top venue categories

Now this is ready for clustering.

3.3. K-Means Clustering

For the purpose of this assignment, K-Means clustering has been chosen to attempt clustering. However, what k value should be chosen? For this we will count the square error cost of each k value from 1-11. The resulting plot shows an elbow at k=8.

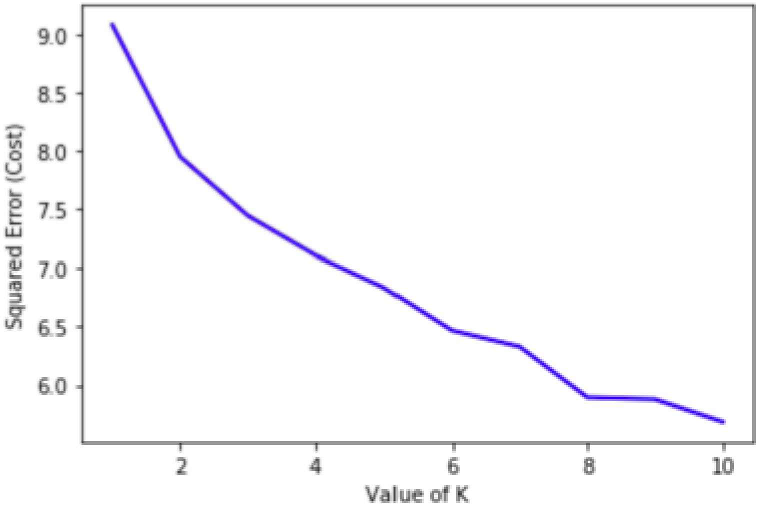


Figure 11: Plot of square error against K-value