The Battle of the Neighborhoods

-Amala Mary Vincent

1. Introduction

1.1 Background

There are over a hundred restaurants in Chicago. Most of them are American restaurants and cafes. To open a new cafe, we need to find the best location in the city. The optimal location would be one less crowded with other restaurants especially coffee shops and other cafes. Also the site should not be far from the city center. Once the above two conditions are satisfied we will look for locations closer to the city center. This report is meant to help stakeholders who are interested in opening a cafe in Chicago. We will use data science methodologies to find the most promising sites and mention their advantages. Stakeholders can choose the best location according to their preferences.

1.2 Problem Statement

This project aims to find the best land site to start an cafe in Chicago. Based on definition of our problem, factors that will influence our decision are:

- number of existing restaurants in the neighborhood
- number of and distance to cafes in the neighborhood, if any
- distance of neighborhood from city center

2. Data

2.1 Data Sources

Data is collected from kaggle website. Chicago neighborhoods data is a json file that contains the longitude and latitudes of neighborhoods. The link to the data is :
https://www.kaggle.com/afernandezcan/chicago-neighborhoods-2012

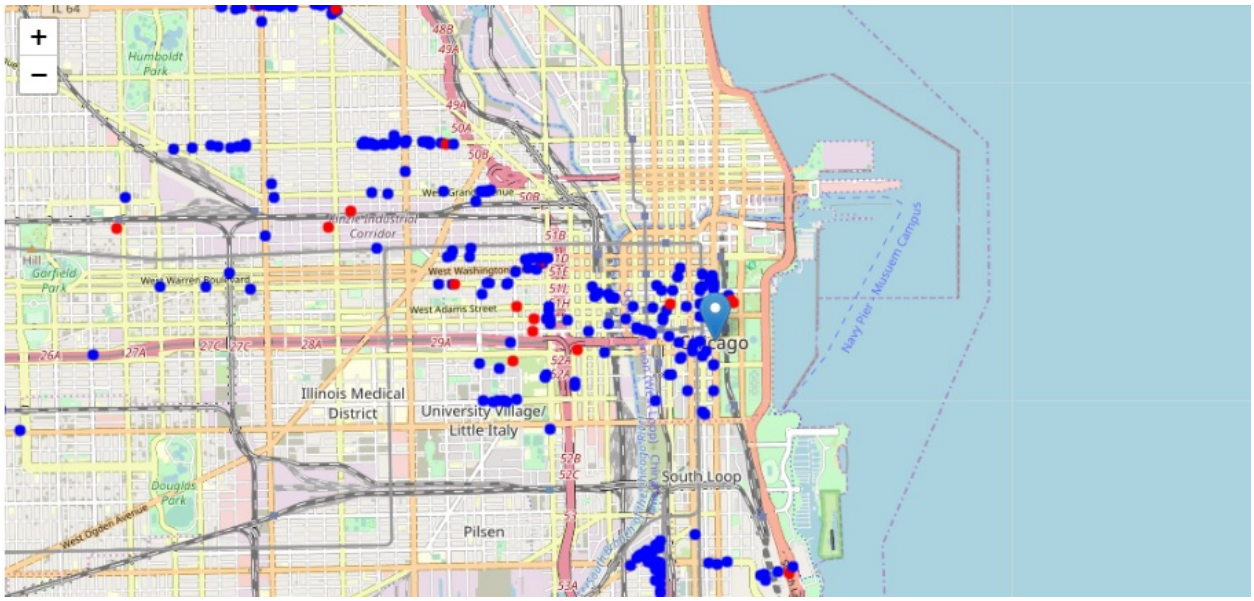
2.2 Load and explore the data

Data is downloaded as a zip file from kaggle site and the json file is extracted. After that, it is read into a pandas data frame named ‘neighborhoods’.

2.3 Foursquare API

Once we have our location candidates, let's use Foursquare API to get info on restaurants in each neighborhood. We're interested in venues in 'food' category. We will include in out list venues that have 'restaurant', 'cafe', 'cafeteria', 'bakery', 'coffee shop' in category name, since they can all be considered competitors and we'll make sure to detect and include all the subcategories of

specific 'cafes' category, as we need info on cafes in the neighborhood. We will find the total number of restaurants, total number of cafes, percentage of cafes and the average number of restaurants in neighborhood from the data. Based on which we will plot the map of Chicago with red circles on cafes and blue circles on restaurants, using folium. Now we have all the restaurants in area within few kilometers from Chicago center, and we know which ones are cafes. We also know which restaurants exactly are in vicinity of every neighborhood candidate center. This concludes the data gathering phase - we're now ready to use this data for analysis to produce the report on optimal locations for a new café.



3. Methodology

In this project we will direct our efforts on detecting areas of Chicago that have low restaurant density, particularly those with low number of cafes. We will limit our analysis to area ~6km around city center.

In first step we have collected the required data: **location and type** (category) of every restaurant within 6km from Chicago center. We have also **identified cafes** (according to Foursquare categorization).

Second step in our analysis will be calculation and exploration of '**restaurant density**' across different areas of Chicago - we will use **heatmaps** to identify a few promising areas close to center with low number of restaurants in general (and no cafes in vicinity) and focus our attention on those areas.

In third and final step we will take into consideration locations with **no more than one restaurant in radius of 500 meters,** and we want locations **without cafes in radius of 400 meters**. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

4. Analysis

Let's perform some basic explanatory data analysis and derive some additional info from our raw data. First let's count the **number of restaurants** in every area candidate as shown in figure below.

[55]:
```
location_restaurants_count = [len(res) for res in location_restaurants]

neighborhoods['Restaurants in area'] = location_restaurants_count

print('Average number of restaurants in every area with radius=300m:', np.array(location_restaurants_count).mean())

neighborhoods.head(10)
```

```
Average number of restaurants in every area with radius=300m: 22.8
```

[55]:

	Neighborhood	Latitude	Longitude	Distance	X	Y	Restaurants in area	Distance to cafe
0	Grand Boulevard	41.816814	-87.606708	2.211630e+07	-5.389478e+06	1.152288e+07	8	88.692431
1	Printers Row	41.874371	-87.627607	2.211316e+07	-5.379867e+06	1.152233e+07	59	909.465834
2	United Center	41.888852	-87.667069	2.211578e+07	-5.376142e+06	1.152609e+07	51	938.323448
3	Sheffield & DePaul	41.921661	-87.658335	2.211193e+07	-5.371461e+06	1.152342e+07	55	103.370888
4	Humboldt Park	41.887823	-87.740596	2.212319e+07	-5.373476e+06	1.153454e+07	4	2909.053530
5	Garfield Park	41.888185	-87.695400	2.211866e+07	-5.375158e+06	1.152937e+07	17	1432.966189
6	North Lawndale	41.869869	-87.720239	2.212280e+07	-5.377003e+06	1.153314e+07	7	907.699344
7	Little Village	41.834801	-87.687399	2.212272e+07	-5.383628e+06	1.153119e+07	6	5000.000000
8	Armour Square	41.847127	-87.629201	2.211580e+07	-5.383975e+06	1.152390e+07	43	1967.812453
9	Avalon Park	41.751502	-87.585655	2.212014e+07	-5.400296e+06	1.152381e+07	5	5000.000000

Now let's calculate the **distance to nearest cafe from every area candidate center** (not only those within 300m - we want distance to closest one, regardless of how distant it is) as in figure below.

[56]:
```
distances_to_cafe = []

for ax, ay in zip(neighborhoods.X, neighborhoods.Y):
    min_distance = 5000
    for res in cafe_places.values():
        cx = res[7]
        cy = res[8]
        d = calc_xy_distance(ax, ay, cx, cy)
        if d<min_distance:
            min_distance = d
    distances_to_cafe.append(min_distance)

neighborhoods['Distance to cafe'] = distances_to_cafe
```

[57]:
```
neighborhoods.head(10)
```

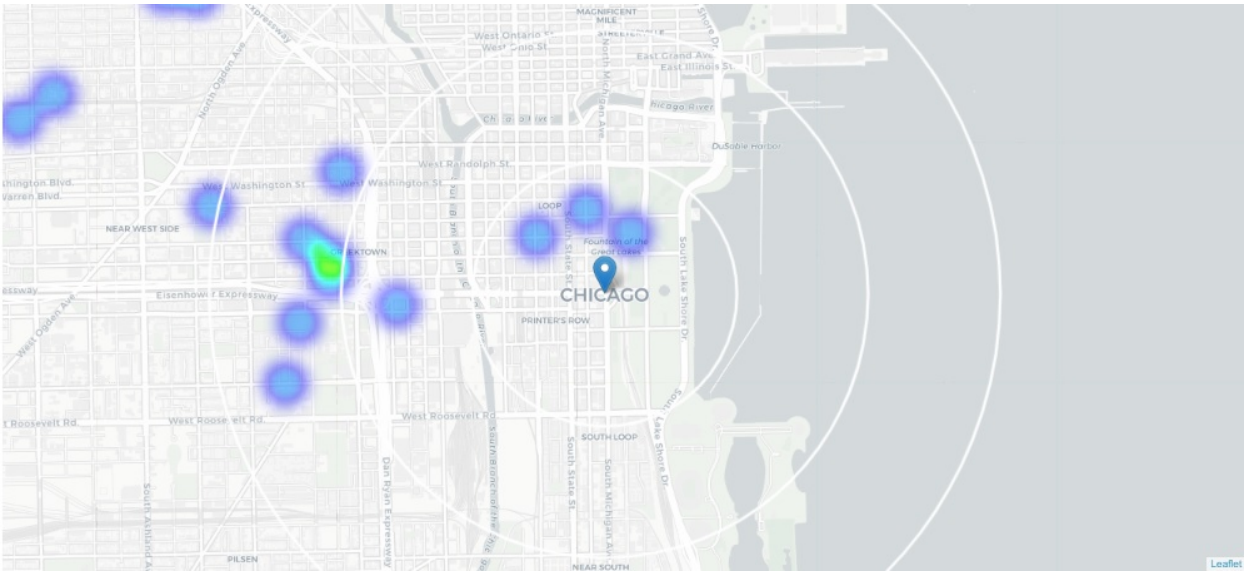
[57]:

	Neighborhood	Latitude	Longitude	Distance	X	Y	Restaurants in area	Distance to cafe
0	Grand Boulevard	41.816814	-87.606708	2.211630e+07	-5.389478e+06	1.152288e+07	8	88.692431
1	Printers Row	41.874371	-87.627607	2.211316e+07	-5.379867e+06	1.152233e+07	59	909.465834
2	United Center	41.888852	-87.667069	2.211578e+07	-5.376142e+06	1.152609e+07	51	938.323448
3	Sheffield & DePaul	41.921661	-87.658335	2.211193e+07	-5.371461e+06	1.152342e+07	55	103.370888
4	Humboldt Park	41.887823	-87.740596	2.212319e+07	-5.373476e+06	1.153454e+07	4	2909.053530
5	Garfield Park	41.888185	-87.695400	2.211866e+07	-5.375158e+06	1.152937e+07	17	1432.966189

Find the average distance to closest cafe from each area center. We got it as 1888 meters. So on average cafe can be found within ~2km from every area center candidate. We need to filter our areas carefully.

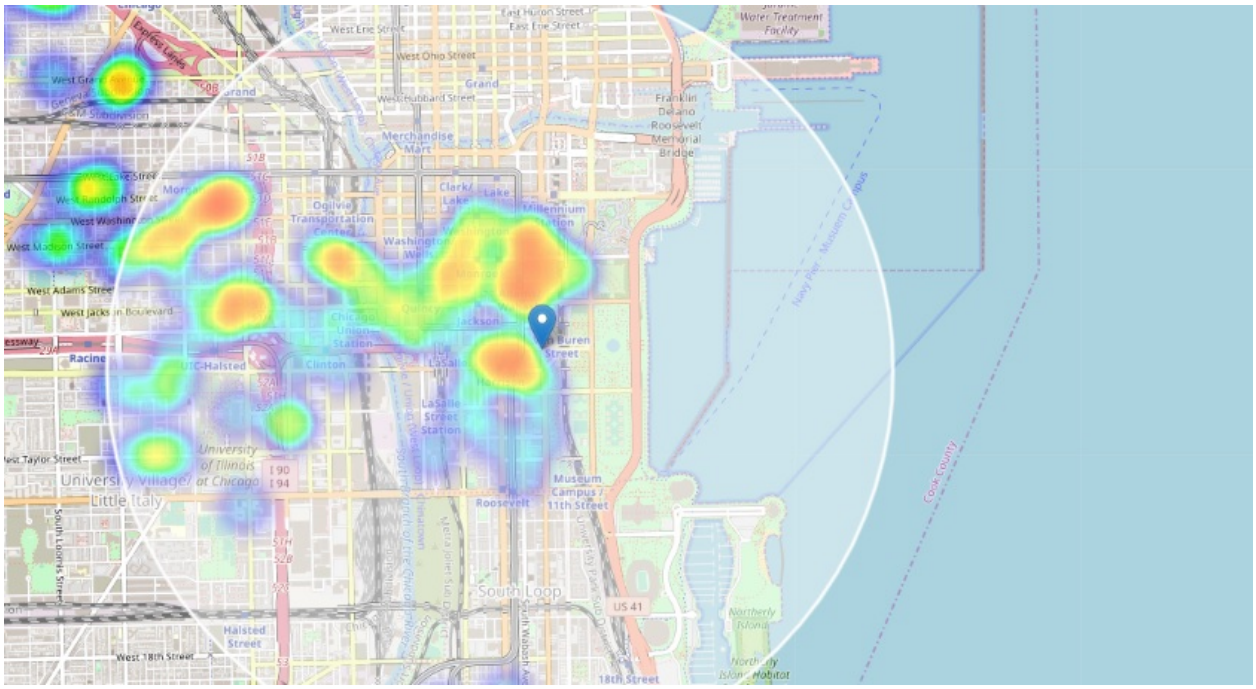
Let's create a map showing heatmap / density of restaurants and try to extract some meaningful insights from that. Also, let's show few circles indicating distance of 1km, 2km and 3km from Chicago center.

Now we create another heatmap map showing **heatmap/density of cafes** only.



This map indicates higher density of existing cafes directly north and west from Chicago center, with closest pockets of **low cafe density positioned north, south-west and south from city center**.

Based on this we will now focus our analysis on areas *north, south-west and south from Chicago center* - we will move the center of our area of interest and reduce its size to have a radius of **2.5km**.



This nicely covers all the pockets of low restaurant density closest to Chicago center.

Let's also create new, more dense grid of location candidates restricted to our new region of interest (let's make our location candidates 500m apart).

```
k = math.sqrt(3) / 2 # Vertical offset for hexagonal grid cells
x_step = 500
y_step = 500 * k
roi_y_min = roi_center_y - 2500

roi_latitudes = []
roi_longitudes = []
roi_xs = []
roi_ys = []
for i in range(0, int(51/k)):
    y = roi_y_min + i * y_step
    x_offset = 50 if i%2==0 else 0
    for j in range(0, 51):
        x = roi_x_min + j * x_step + x_offset
        d = calc_xy_distance(roi_center_x, roi_center_y, x, y)
        if (d <= 2501):
            lon, lat = xy_to_lonlat(x, y)
            roi_latitudes.append(lat)
            roi_longitudes.append(lon)
            roi_xs.append(x)
            roi_ys.append(y)

print(len(roi_latitudes), 'candidate neighborhood centers generated.')
```

```
50 candidate neighborhood centers generated.
```

Now let's calculate two most important things for each location candidate: **number of restaurants in vicinity** (we'll use radius of **250 meters**) and **distance to closest cafe**. Then we will **filter** those locations: we're interested only in **locations with no more than one restaurant in radius of 250 meters**, and **no cafes in radius of 400 meters**.

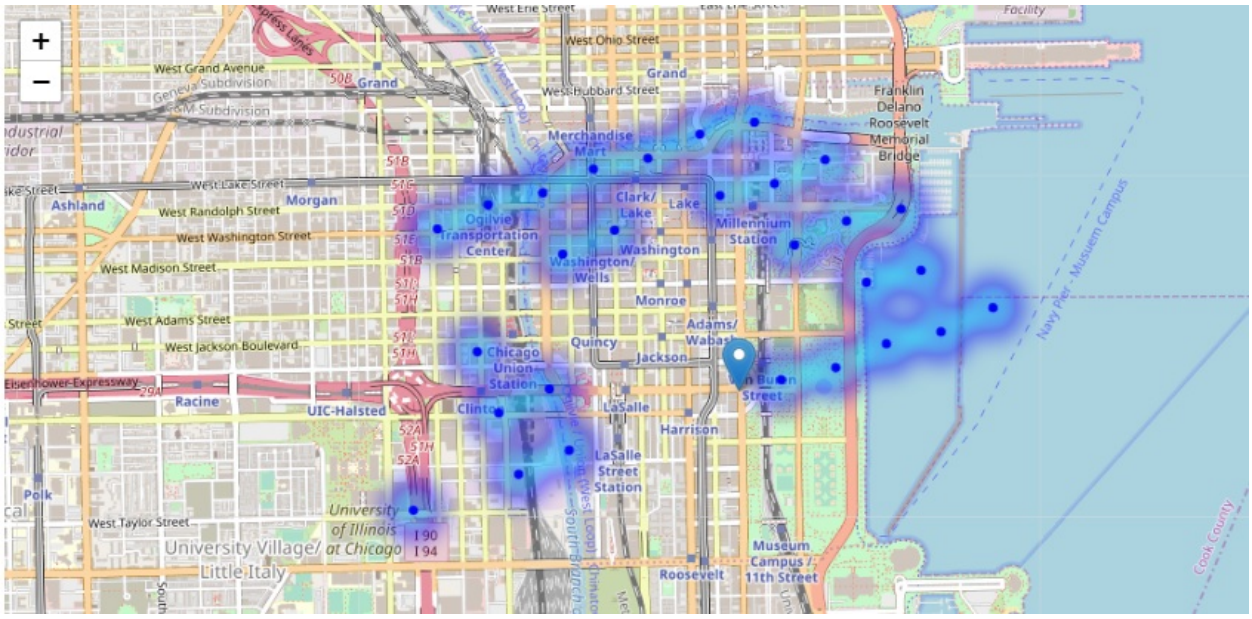
```
good_res_count = np.array((df_roi_locations['restaurants nearby']<=1))
print('Locations with no restaurants nearby:', good_res_count.sum())

good_caf_distance = np.array(df_roi_locations['Distance to cafe']>=400)
print('Locations with no cafes within 400m:', good_caf_distance.sum())

good_locations = np.logical_and(good_res_count, good_caf_distance)
print('Locations with both conditions met:', good_locations.sum())

df_good_locations = df_roi_locations[good_locations]
```

```
Locations with no restaurants nearby: 32
Locations with no cafes within 400m: 41
Locations with both conditions met: 28
```

In the map, this appears as follows:

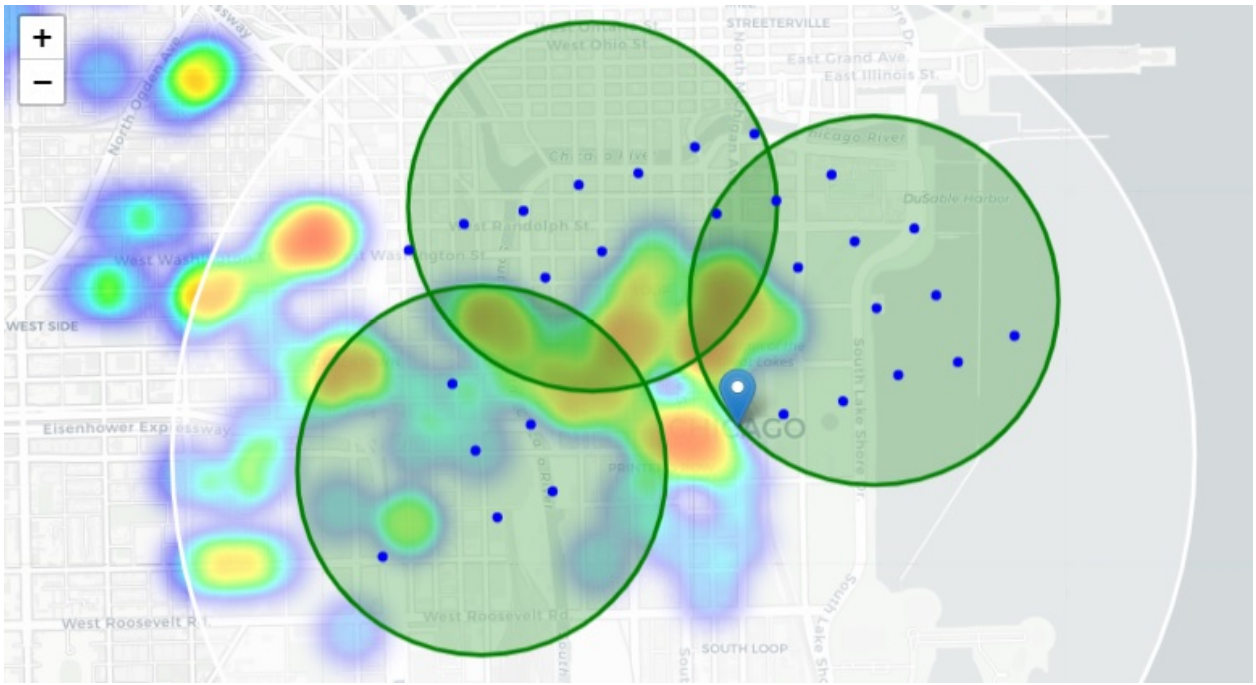
We now have a bunch of locations fairly close to Chicago center and we know that each of those locations has no more than one restaurant in radius of 250m, and no cafes closer than 400m. Any of those locations is a potential candidate for a new cafe, at least based on nearby competition.

Let's now show those good locations in a form of heatmap:



What we have now is a clear indication of zones with low number of restaurants in vicinity, and *no* cafess at all nearby.

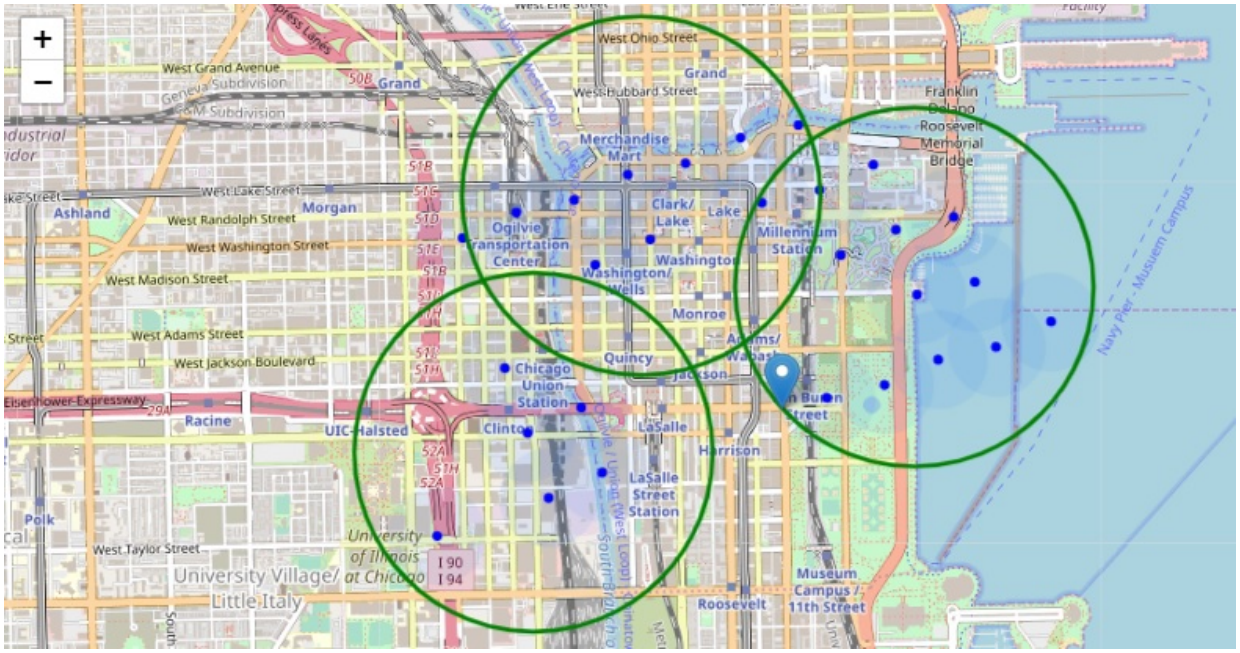
Let us now **cluster** those locations to create **centers of zones containing good locations**. Those zones, their centers and addresses will be the final result of our analysis.



Our clusters represent groupings of most of the candidate locations and cluster centers are placed nicely in the middle of the zones 'rich' with location candidates.

Addresses of those cluster centers will be a good starting point for exploring the neighborhoods to find the best possible location based on neighborhood specifics.

Let's see those zones on a city map without heatmap, using shaded areas to indicate our clusters:

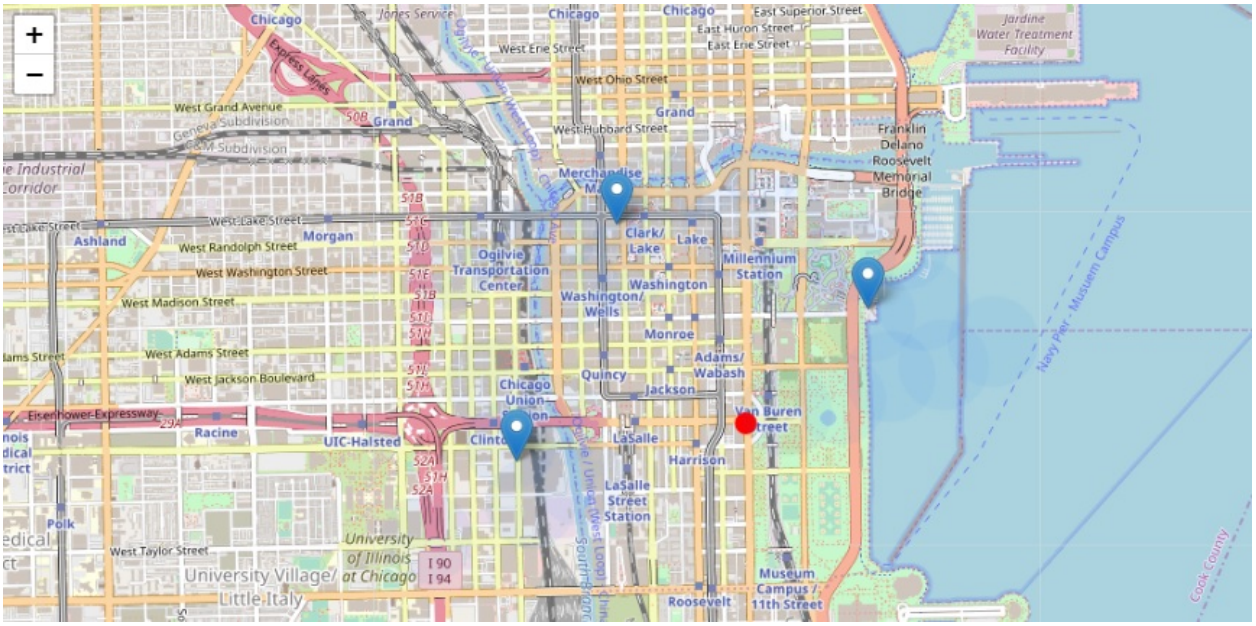


Now, we will list the latitudes and longitudes of centers of this cluster.

```
candidate_area_addresses = []
print('================================================================')
print('Lat-Long of centers of areas recommended for further analysis')
print('================================================================\n')
for lon, lat in cluster_centers:
    geolocator = Nominatim(user_agent="ch_explorer")
    location = geolocator.reverse(lon, lat)
    candidate_area_addresses.append(location.address)
    x, y = lonlat_to_xy(lon, lat)
    d = calc_xy_distance(x, y, X, Y)
    print('{},{} => {:f}m from Chicago center'.format(lat, lon, d/10000))
```

```
================================================================
Lat-Long of centers of areas recommended for further analysis
================================================================

41.87367302564292,-87.63954943427389 => 2211.441701m from Chicago center
41.88114797943391,-87.61638420659332 => 2211.142952m from Chicago center
41.885278274755876,-87.63300173893865 => 2211.271132m from Chicago center
```

This concludes our analysis. We have created 3 addresses representing centers of zones containing locations with low number of restaurants and no cafes nearby, all zones being fairly close to city center (all less than 2km from Chicago center). Although zones are shown on map with a radius of ~500 meters (green circles), their shape is actually very irregular and their centers/addresses should be considered only as a starting point for exploring area neighborhoods in search for potential restaurant locations.



Map shows the final three locations.

5. Results and Discussion

Our analysis shows that although there are a great number of restaurants in Chicago, there are pockets of low restaurant density fairly close to city center. Highest concentration of restaurants was detected west from city center, so we focused our attention to areas north, east, south-west and south.

After directing our attention to the narrow area of interest we first created a dense grid of location candidates (spaced 500m apart); those locations were then filtered so that those with more than one restaurant in radius of 250m and those with an Italian restaurant closer than 400m were removed.

Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Latitude and longitudes of centers of those zones were identified, which can be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is 3 zones containing largest number of potential new cafe locations based on number of and distance to existing venues - both restaurants in general and cafes particularly. This, of course, does not imply that those zones are actually optimal locations for a new cafe! Purpose of this analysis was to only provide info on areas close to Chicago center but not crowded with existing restaurants (particularly cafes) - it is entirely possible that there is a very good reason for small number of restaurants in any of those areas, reasons which would make them unsuitable for a new

restaurant regardless of lack of competition in the area. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in location which has not only no nearby competition but also other factors taken into account and all other relevant conditions met.

6. Conclusion

Purpose of this project was to identify areas close to Chicago center with low number of restaurants (particularly cafes) in order to aid stakeholders in narrowing down the search for optimal location for a new cafe. By calculating restaurant density distribution from Foursquare data we have first identified general neighborhood locations that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.