Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Table of contents

1. Introduction: Business Problem

In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening an **Italian restaurant** in **Ho Chi Minh City**, Vietnam.

Since there are lots of restaurants in Ho Chi Minh City, we will try to detect **locations that are not already crowded with restaurants**. We are also particularly interested in **areas with no Italian restaurants in vicinity**. We would also prefer locations **as close to city center as possible**, assuming that first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

2. Data

Based on definition of our problem, factors that will influence our decision are:

* Number of existing restaurants in the neighborhood (any type of restaurant)

* Number of and distance to Italian restaurants in the neighborhood (if any)

* Distance of neighborhood from city center

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

* Centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using GeoPy **Nominatim API reverse geocoding**

* Number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**

* Coordinate of Ho Chi Minh City center will be obtained using GeoPy **Nominatim API geocoding**

3. Methodology

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

In first step, we will collect the required data: location and type (category) of every restaurant within 6km from city center. We will also have to identify Italian restaurants (according to Foursquare categorization).

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

In final step, we will focus on most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders. We will take into consideration locations with only one restaurant in radius of 250 meters, and we want locations without Italian restaurants in radius of 500 meters. We will present map of all such locations and 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

Geographical coordinates of Ho Chi Minh City center are [10.7758439, 106.7017555].

We create a grid of area candidates, equally spaced, within ~6km around city center. Our neighborhoods will be defined as circular areas with a radius of 300 meters, so our neighborhood centers will be 600 meters apart.

City center location and candidate neighborhood centers:
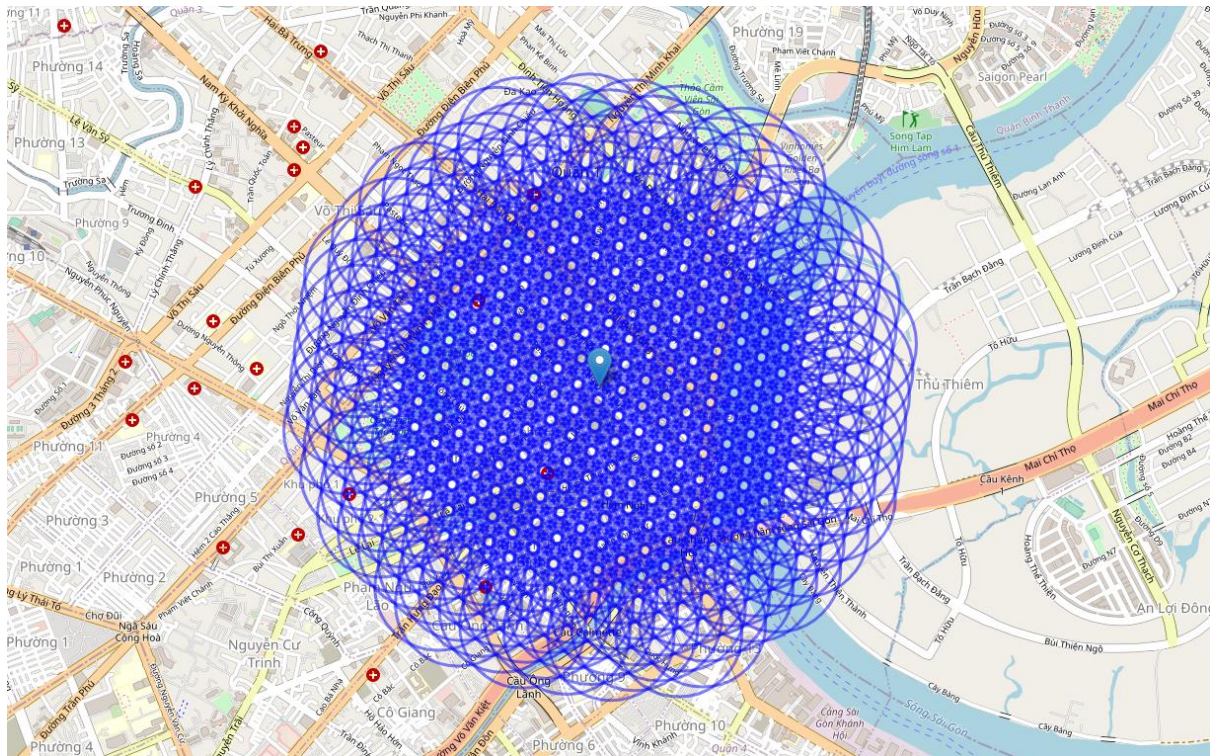
We use GeoPy Nominatim API reverse geocoding to obtain the approximate addresses of those centers of neighborhoods and place this information into a DataFrame:

	Address	Latitude	Longitude	X	Y	Distance from center
0	Lẩu Kichi Kichi, Lê Quý Đôn, Võ Thị Sáu, Phườn...	10.780574	106.691692	1.522905e+07	1.090728e+07	5992.495307
1	Võ Thị Sáu, Phường Võ Thị Sáu, Quận 3, 7000000	10.779478	106.691541	1.522965e+07	1.090728e+07	5840.376700
2	Phòng Khám Sản Nhi Diamond, 9, Trần Quốc Thảo,...	10.778382	106.691389	1.523025e+07	1.090728e+07	5747.173218
3	Maritime, Võ Văn Tần, Võ Thị Sáu, Phường Võ Th...	10.777287	106.691238	1.523085e+07	1.090728e+07	5715.767665
4	Tip top Karaoke, Trương Định, Võ Thị Sáu, Phườ...	10.776191	106.691087	1.523145e+07	1.090728e+07	5747.173218
5	Jollibee, 204, Nguyễn Thị Minh Khai, Khu phố 1...	10.775096	106.690936	1.523205e+07	1.090728e+07	5840.376700
	...					

Now that we have our location candidates, we will use Foursquare API to get information on restaurants in each neighborhood.

We are interested in venues in 'food' category, but only those that are proper restaurants. We will include in our list only venues that have 'restaurant' in category name. We will make sure to detect

and include all the subcategories of specific 'Italian restaurant' category, as we need information on Italian restaurants in the neighborhood.

From Foursquare, we have:

- Total number of restaurants: 395
- Total number of Italian restaurants: 19
- Percentage of Italian restaurants: 4.81%
- Average number of restaurants in every area with radius=300m: 13.5

Restaurants in our area of interest are in blue, Italian restaurants are in red:
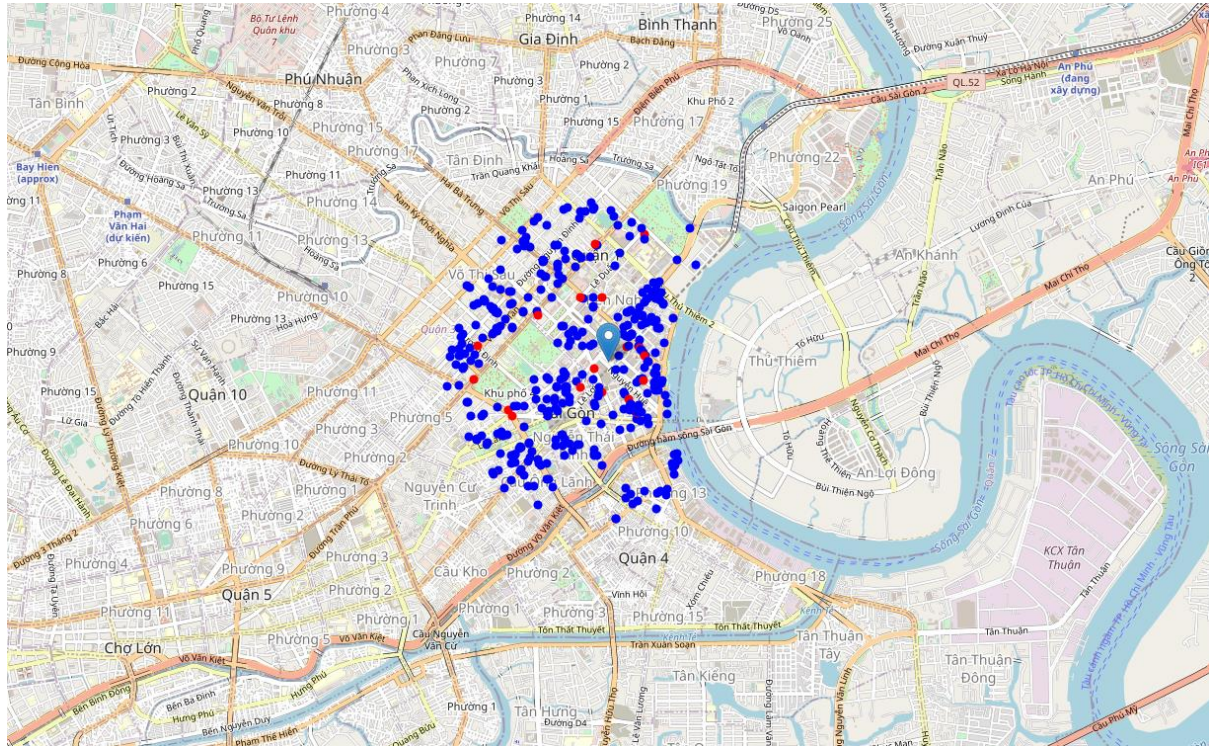

We count the **number of restaurants in every area candidate** and calculate the **distance to nearest Italian restaurant from every area candidate center**:

	Address	Latitude	Longitude	Distance from center	Restaurants in area	Distance to Italian restaurant
0	Lẩu Kichi Kichi, Lê Quý Đôn, Võ Thị Sáu, Phườn...	10.780574	106.691692	5992.495307	32	1921.587500
1	Võ Thị Sáu, Phường Võ Thị Sáu, Quận 3, 7000000	10.779478	106.691541	5840.376700	29	1339.665224
2	Phòng Khám Sản Nhi Diamond, 9, Trần Quốc Thảo,...	10.778382	106.691389	5747.173218	30	785.434468
3	Maritime, Võ Văn Tần, Võ Thị Sáu, Phường Võ Th...	10.777287	106.691238	5715.767665	23	398.888074
	...					

The map showing **density of restaurants**:
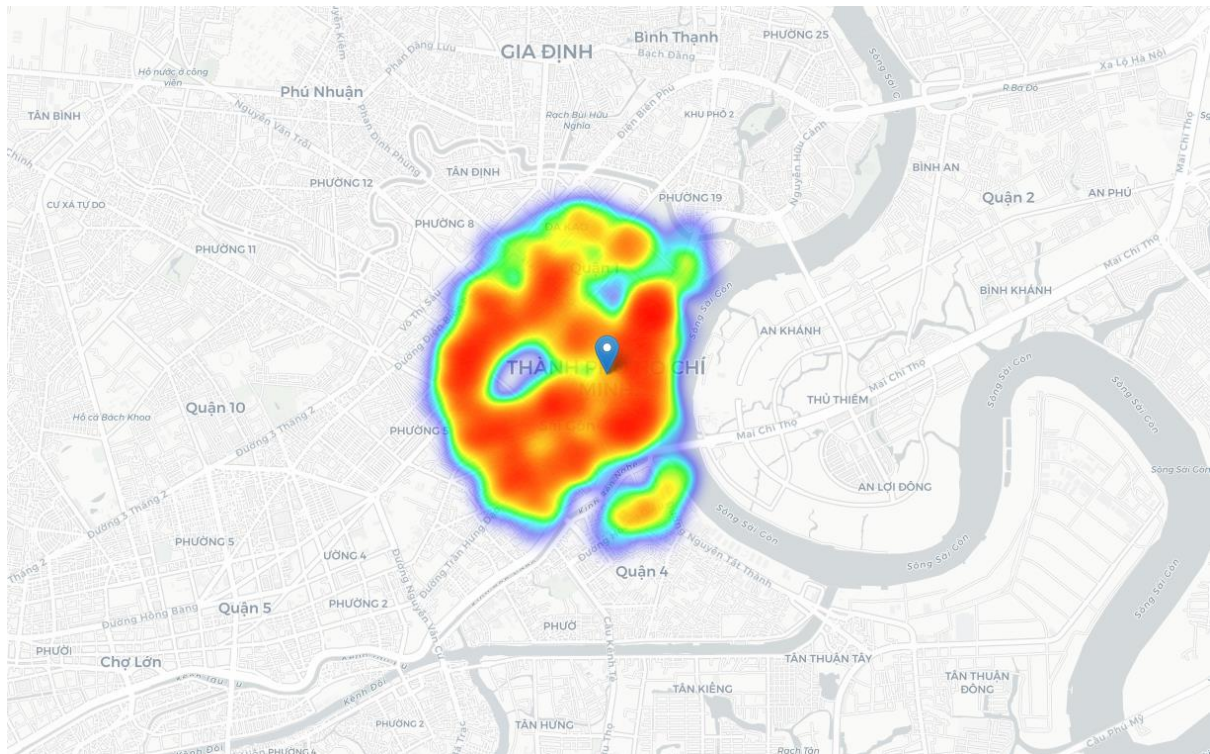
The map showing **density of Italian restaurants** only:
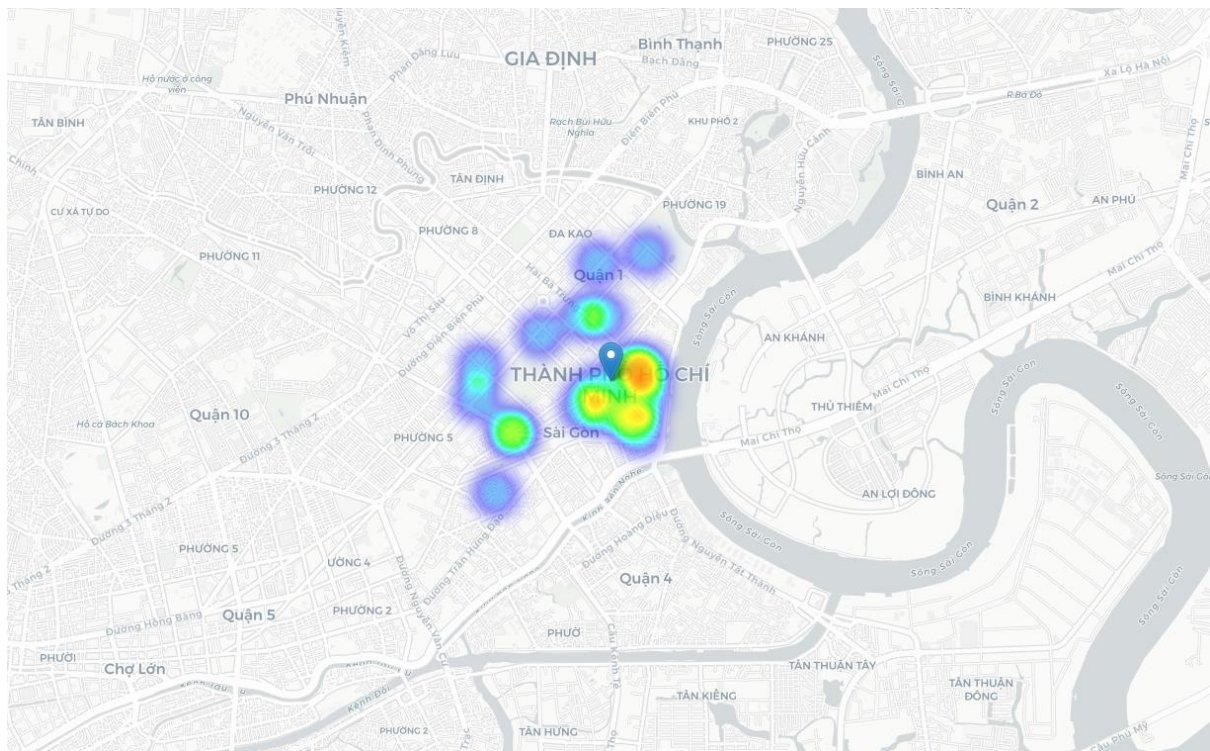
This map is not so 'hot' as Italian restaurants represent only a subset of ~4.8% of all restaurants in city center.

We are interested only in **locations with one restaurant in radius of 250 meters**, and **no Italian restaurants in radius of 500 meters****. So after filtering:

- Locations with only one restaurant nearby: 449
- Locations with no Italian restaurants within 500m: 358
- Locations with both conditions met: 324
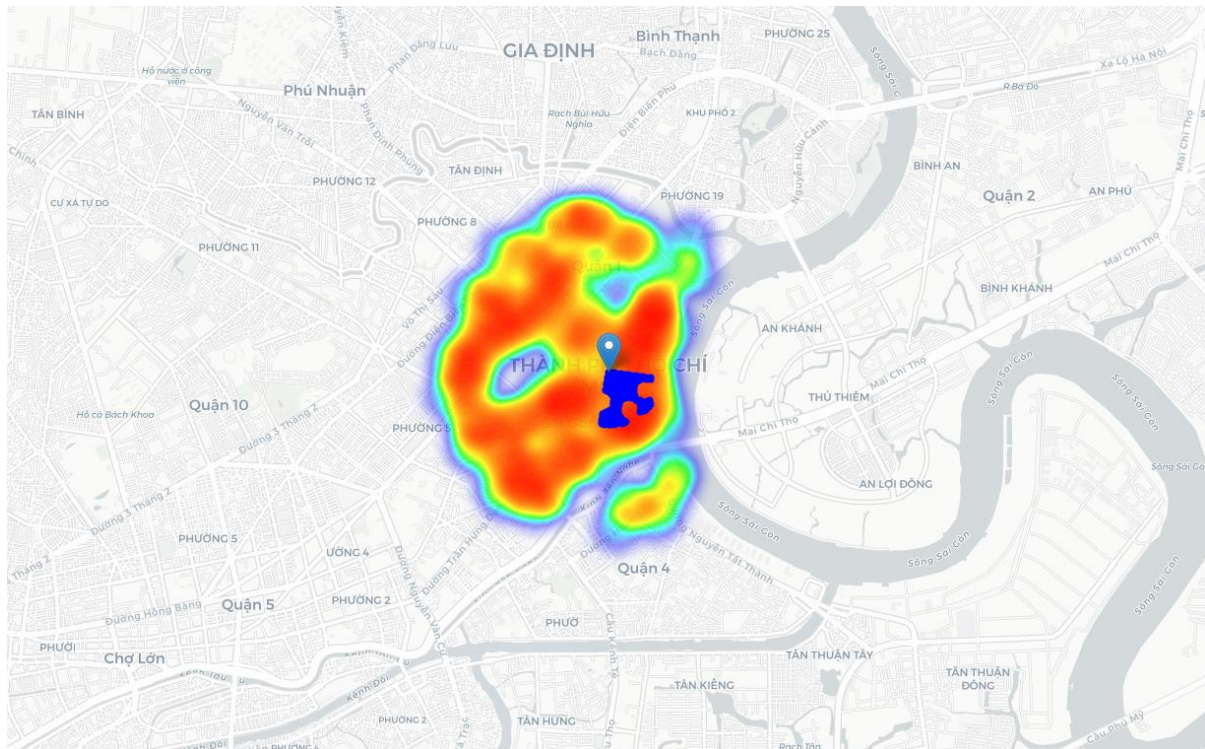
We cluster locations with both conditions met to create **centers of zones containing good locations**.
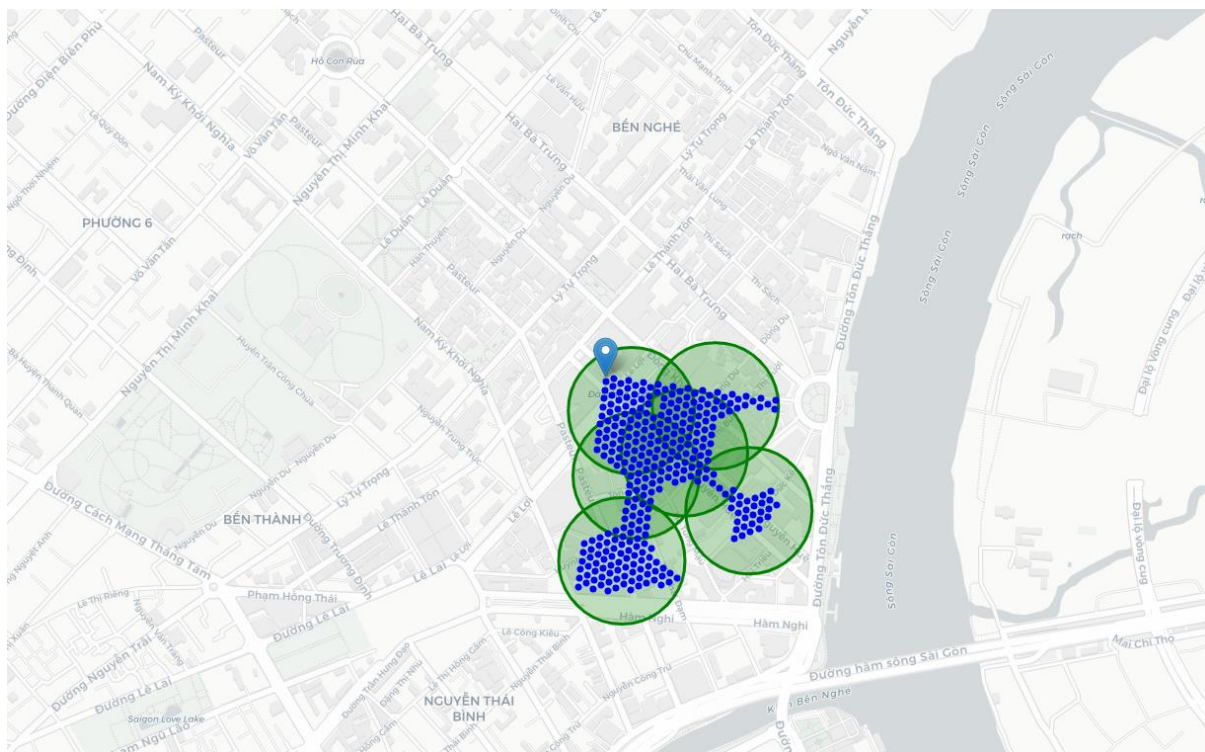
Those zones, their centers and addresses will be the result of our analysis. Here are the addresses of centers of areas recommended for further analysis:

1. Chung cư 42 Tôn Thất Thiệp, 42, Tôn Thất Thiệp, Phường Bến Nghé, Quận 1, 00084
   => 1.2km from Ho Chi Minh City center
2. Phố đi bộ Nguyễn Huệ, Nguyễn Huệ, Phường Bến Nghé, Quận 1, 71006
   => 0.5km from Ho Chi Minh City center
3. Phố đi bộ Nguyễn Huệ, Nguyễn Huệ, Phường Bến Nghé, Quận 1, 71006
   => 1.3km from Ho Chi Minh City center
4. Phố đi bộ Nguyễn Huệ, Nguyễn Huệ, Phường Bến Nghé, Quận 1, 71006
   => 2.3km from Ho Chi Minh City center
5. 86, Pasteur, Phường Bến Nghé, Quận 1, 00084
   => 2.1km from Ho Chi Minh City center
6. Trung Nguyen Coffee, Đông Du, Phường Bến Nghé, Quận 1, 00084
   => 1.3km from Ho Chi Minh City center
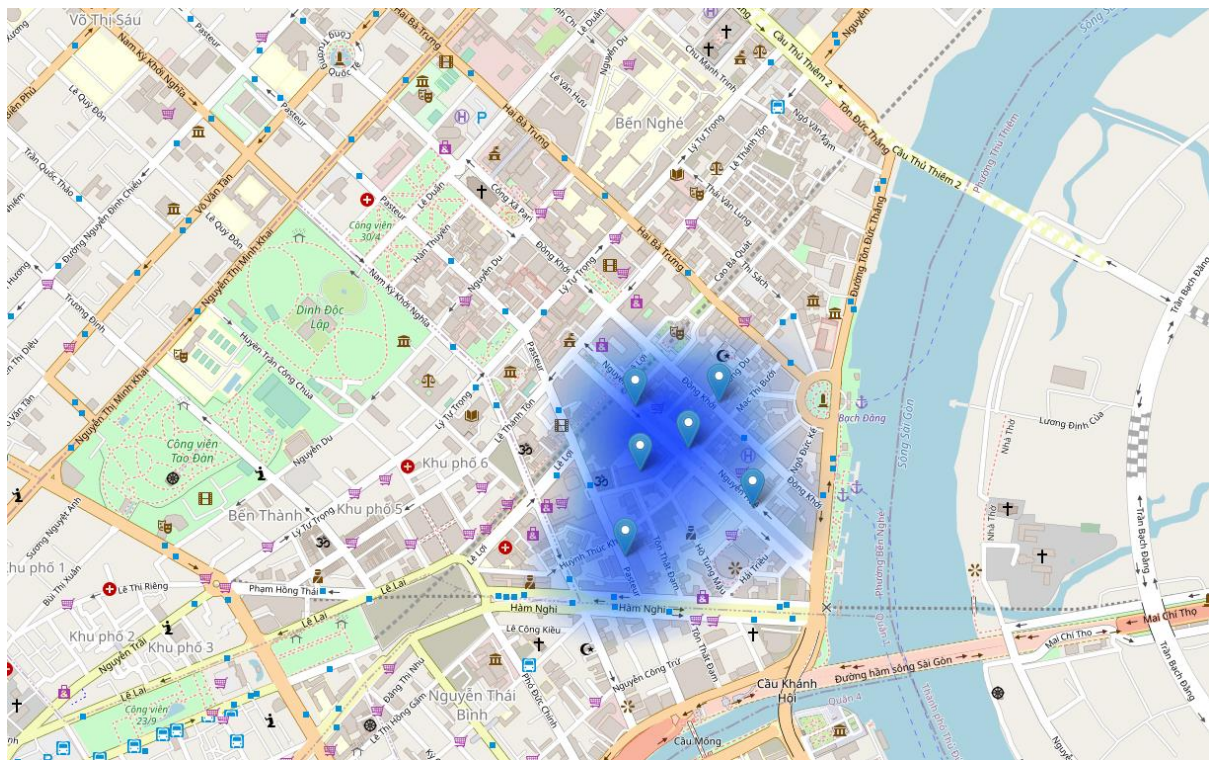
We have created 6 addresses representing centers of zones containing locations with low number of restaurants and no Italian restaurants nearby, all zones being close to city center. The centers/ addresses should be considered only as a starting point for exploring area neighborhoods in search for potential restaurant locations.

5. Results and Discussion

Our analysis shows that although there are nearly 400 restaurants in Ho Chi Minh City center, there are pockets of low restaurant density. We first created a dense grid of location candidates (spaced

100m 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 500m were removed.

Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all this is 6 zones containing largest number of potential new restaurant locations based on number of and distance to existing venues - both restaurants in general and Italian restaurants particularly. This, of course, does not imply that those zones are actually optimal locations for a new restaurant! Purpose of this analysis was to only provide info on areas close to city center but not crowded with existing restaurants (particularly Italian) - 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 Ho Chi Minh City center with low number of restaurants (particularly Italian restaurants) in order to aid stakeholders in narrowing down the search for optimal location for a new Italian restaurant. By calculating restaurant density distribution from Foursquare data we generated 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, levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.