Predicting restaurant's location for opening in Ha Noi

The Vinh Nguyen

March 14, 2020

I. Table of contents

II. Introduction

1. Background: Opening a restaurant is the easiest way to make money. On the contrary, it is way more complicated. So how complicated? There are many factors one has to consider before opening a restaurant. One deciding factor is location. Why location? A luxury restaurant in a poor village, a beef-specialized restaurant in Hindus community are all examples of poor choice of location. How to choose a good location? Worst news is that there is no correct answer like in maths. We have to use try-and-error method to figure out which works well and which doesn't. Changing restaurant's location several time really hurts one's pocket. It is a costly process. A process which can be greatly reduced in cost if we utilize Data Science.

2. Problem
   Choosing location to open a restaurant is always an emotional decision because there are no specific guidelines to show one how to get the job done. Therefore, this project aims to predict a suitable restaurant's location by exploring criteria related to location vital to attractiveness of a restaurant.

3. Target Audience

Anyone looking for a place to open their restaurant.

III. Data

1. Data source:

   Data is scraped from now.vn, an online food – ordering service. The data has approximately 2000 rows and 11 columns which are:

   - RESTAURANT_NAME
   - ADDRESS
   - WEBSITE
   - RESTAURANT_KIND
   - NUMBER_OF_APPROACHED_PEOPLE
   - PRICE_RANGE
   - CATEGORY
   - FOOD_ITEM
   - FOOD_COST
   - LATITUDE
   - LONGITUDE

   The data is constantly updated so it can best reflect situation in Ha Noi. This is the part that took me most of the time.

2. Data cleaning:

   - Latitude, Longitude: there are some rows with missing latitude and longitude. To combat this, I decide to drop all those rows since they only have about 2,3 rows. It won't affect the overall quality of dataset. Afterward, I convert it to double datatype for easier readability for methods.
   - RESTAURANT_KIND: there are about nearly 200 rows with RESTAURANT_KIND unidentified. This is a huge number of row, I cannot simply drop it. At least, they can still be helpful in identifying density of restaurant in a zone. So, I assign filler word "NaN" to them.

- NUMBER_OF_APPROACHED_PEOPLE: the same as RESTAURANT_KIND, it is just too many of row missing it (10% of total row)
- FOOD_COST: originally, FOOD_COST is a list of cost for each kind of food a restaurant serves which I find it difficult to save it in .csv file. So I decide to save it in the form of <cost1>|<cost2>|<cost3> as a long string. If there is nan value, I will drop it since it only occupies a tiny portion.
- PRICE_RANGE: There are some restaurants with missing PRICE_RANGE's value. However, since I find that FOOD_COST is a much more detailed information about price than PRICE_RANGE, I simply not touch PRICE_RANGE.

3. Feature Selection

   The following features are needed for this project:
   - FOOD_COST
   - LATITUDE
   - LONGITUDE
   - RESTAURANT_KIND
   - NUMBER_OF_APPROACHED_PEOPLE

   The rest is dropped.

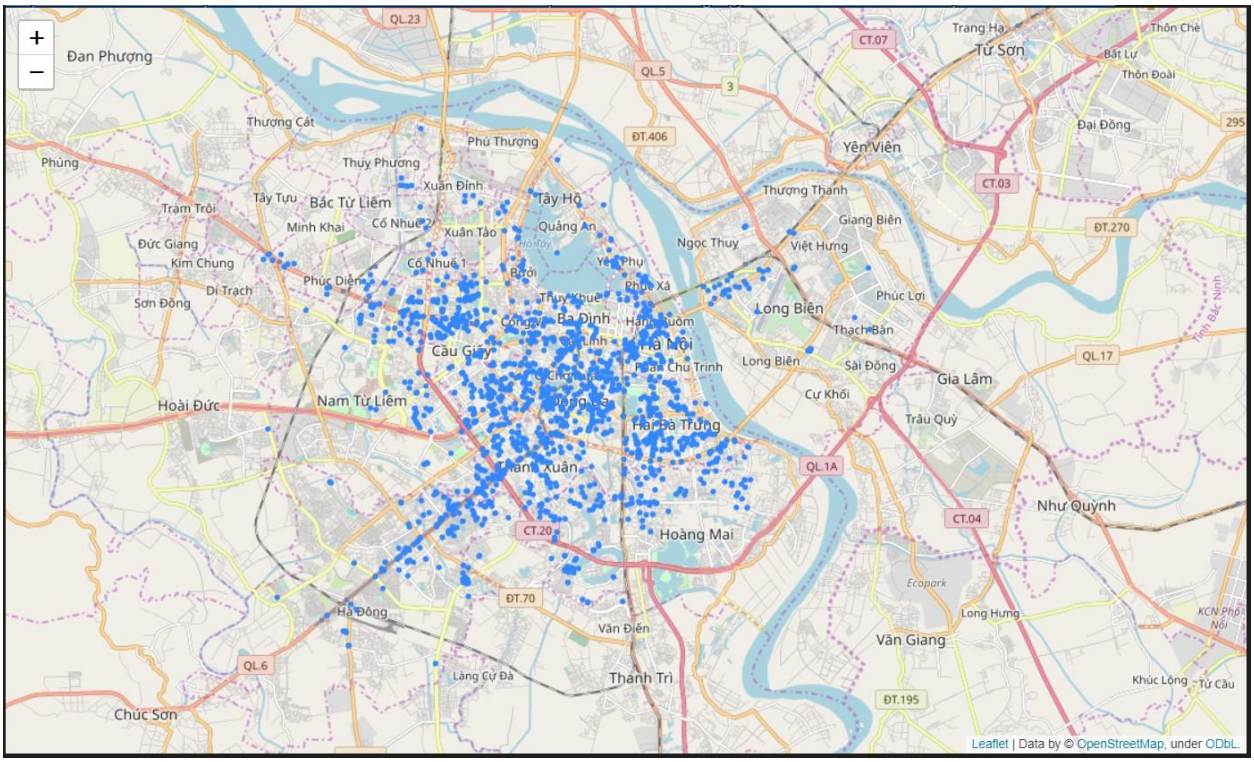
   Here's why:
   - LATITUDE, LONGITUDE: display restaurant location on a map for further analysis.
   - RESTAURANT_KIND: this information is needed to understand which zone is suitable to a kind of restaurant. At first, the data is rather chaotic with many different labels and missing values. I group it all down to 10 main labels: CAFÉ/DESSERT, QUÁN ĂN, SHOP ONLINE, ĂN VẶT/VỈA HÈ, NHÀ HÀNG, NaN, TIỆM BÁNH, ĂN CHAY, SHOP/CỬA HÀNG, QUÁN NHẬU with NaN as filler word for unknown restaurant kind.
   - NUMBER_OF_APPROACHED_PEOPLE: this information is needed to identify which zone has a high number of restaurants with high

number of approached people. Originally, this is the number of reviews on now.vn so it, maybe, can correctly reflect young people interest.

- FOOD_COST: this information is needed to identify which zone has a high number of restaurants with same price range. If it is high price range, the zone is more suitable to open a luxury restaurant. Identifying the mode price of all food items gives us a good estimation of a restaurant's price range for later comparison. It also makes PRICE_RANGE redundant.

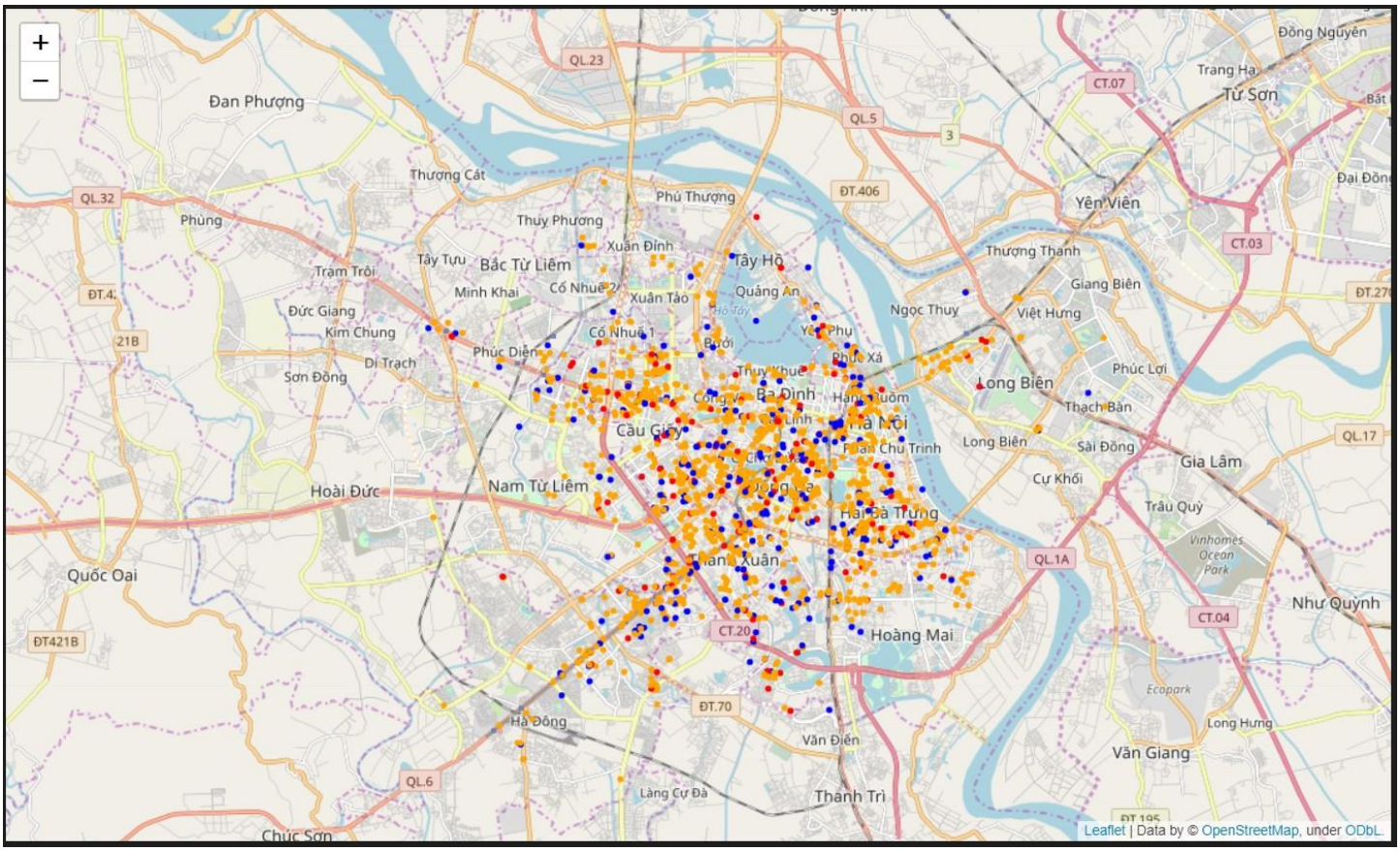
IV. Exploratory Data Analysis

- A first overview of all restaurant locations in the dataset.



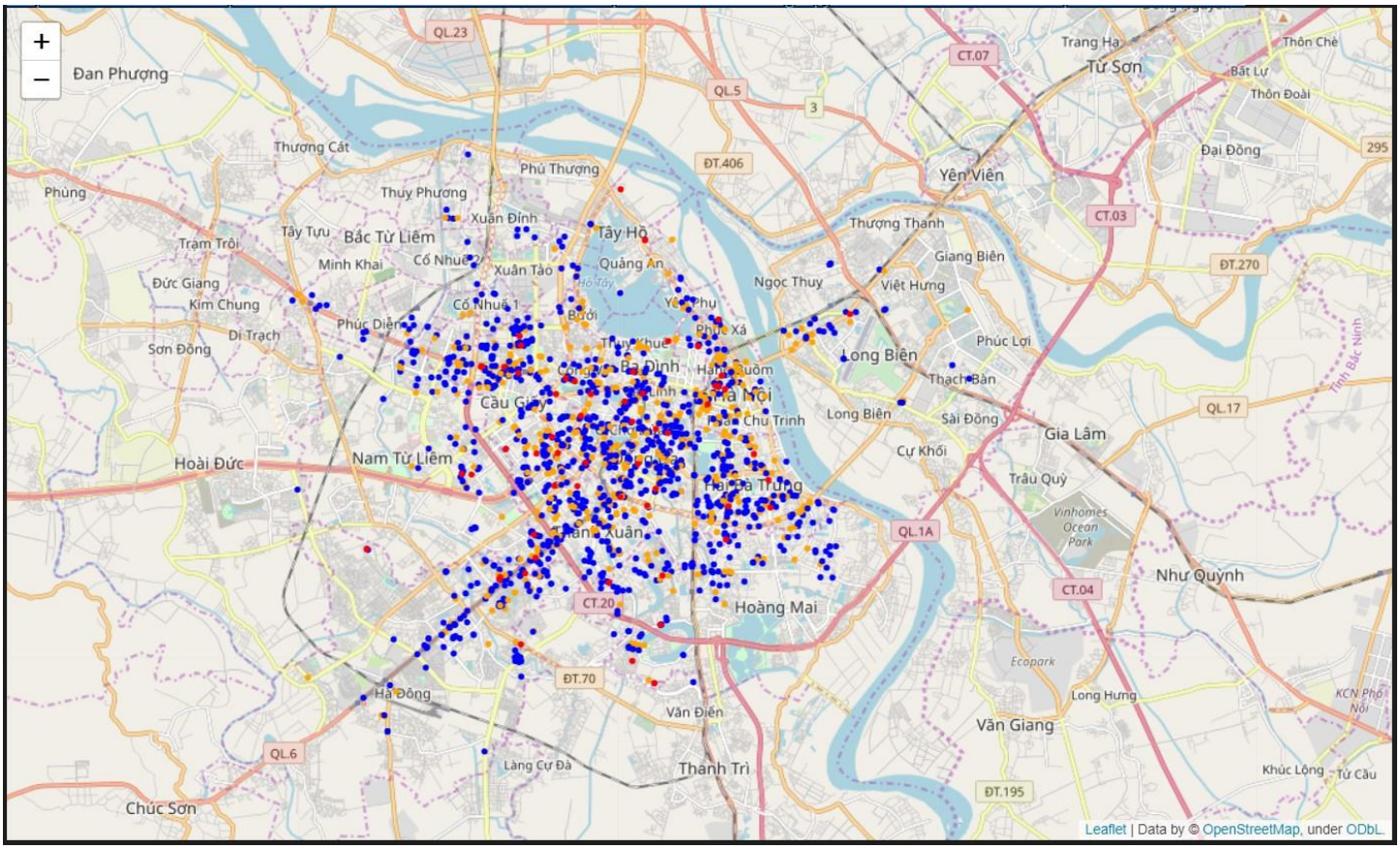
It can be clearly seen that there is a high number of restaurants in the center of Ha Noi. Further zoom into this picture, one can see that majority of restaurants face two-way roads and in dense neighborhood's capillaries (word I borrow from medical terms to depict really small roads in Viet Nam). The more density starts to diminish, the further away from the center of Ha Noi. The density diminishes along routes, which leads to the center of Hanoi.

- Explore restaurant's approached people:
  Orange points have significantly more approached people than blue ones. Red points are unknown. One can see that the competition is rather dire. About 50% of all restaurants in dataset are orange points, while the rest is blue and red points. Further zoom can see that blue points are in very crowded zone or in too remote capillaries, area. This information has led me to deduce a variable known as
  RESTAURANT_LIMIT_PER_ZONE which indicates the number of restaurant allowed in a zone. This variable mainly concerns about quantity of pops. If higher than RESTAURANT_LIMIT_PER_ZONE, the competition will be fierce driving some out of business leading to rebalancing close to RESTAURANT_LIMIT_PER_ZONE. That's why it is important to identify zone and identify the variable RESTAURANT_LIMIT_PER_ZONE.



- Explore preferred price range:

Blue points indicate restaurants with cheap price. Orange points indicate restaurants with more expansive price. Red points are unknown. Orange points reside mainly in areas surrounding center zone and occupy a much smaller portion than blue points. This information has led me to variable AVG_PRICE_LIMIT_PER_ZONE. This variable mainly concerns about the quality of each pops. Tourists can afford much higher price than undergraduates. Still, it cannot give any significant information than the above picture. Henceforth, deeper analysis is required to understand RESTAURANT_LIMIT_PER_ZONE and AVG_PRICE_LIMIT_PER_ZONE.



V. Methodology

The goal is to understand which zone is most attractive to a certain types of restaurant and understand why that zone is attractive? To have a better insight into RESTAURANT_LIMIT_PER_ZONE and AVG_PRICE_LIMIT_PER_ZONE.

RESTAURANT_LIMIT_PER_ZONE has:

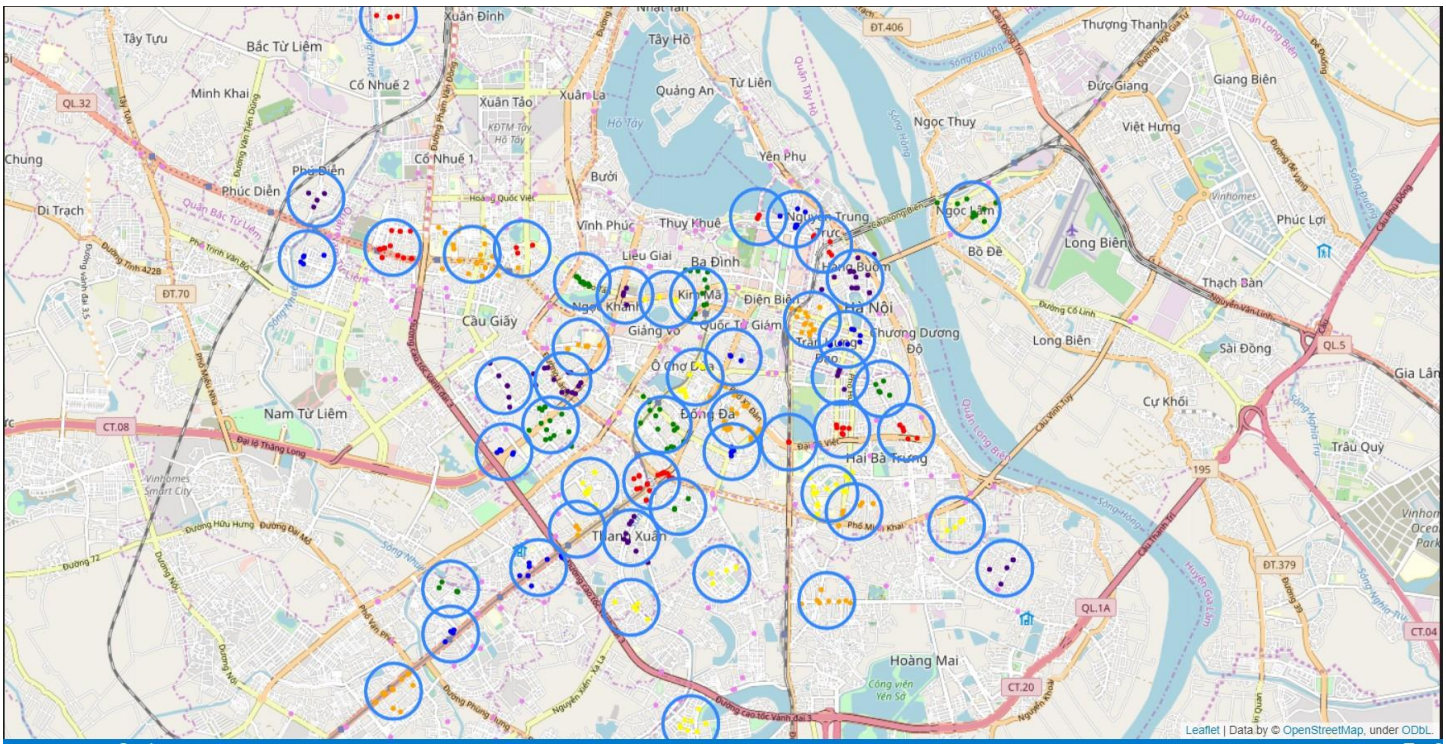
- ZONE_AREA: indicates how large a zone can be
- ZONE_FEATURES: indicates features attracting a large number of pops.

AVG_PRICE_LIMIT_PER_ZONE has:

- ZONE_AREA: indicates how large a zone can be
- ZONE_FEATURES: indicates features attracting pops of higher quality.

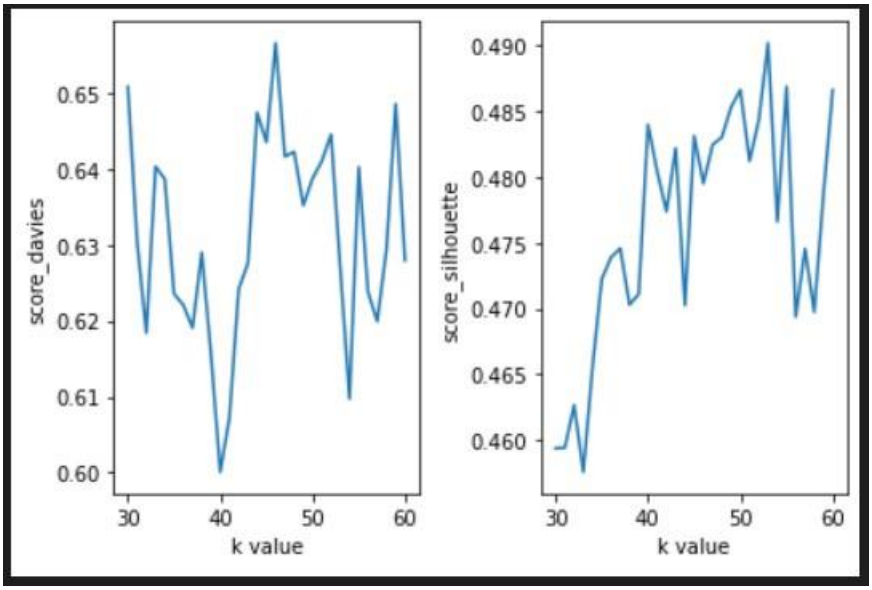
For the sake of simplicity, I choose a kind of restaurant which is 'CAFÉ/DESSERT' to analyze.

- I will create a heat map to focus on an area to analyze what criteria makes it a good place.
- After having identified some deciding factors, I will proceed to identify area suitable for opening 'CAFÉ/DESSERT'.
- I will cluster all ‘CAFÉ/DESSERT’ into zones using DBSCAN clustering. Here is the map of clusters: it produces 49 clusters, 169 noises

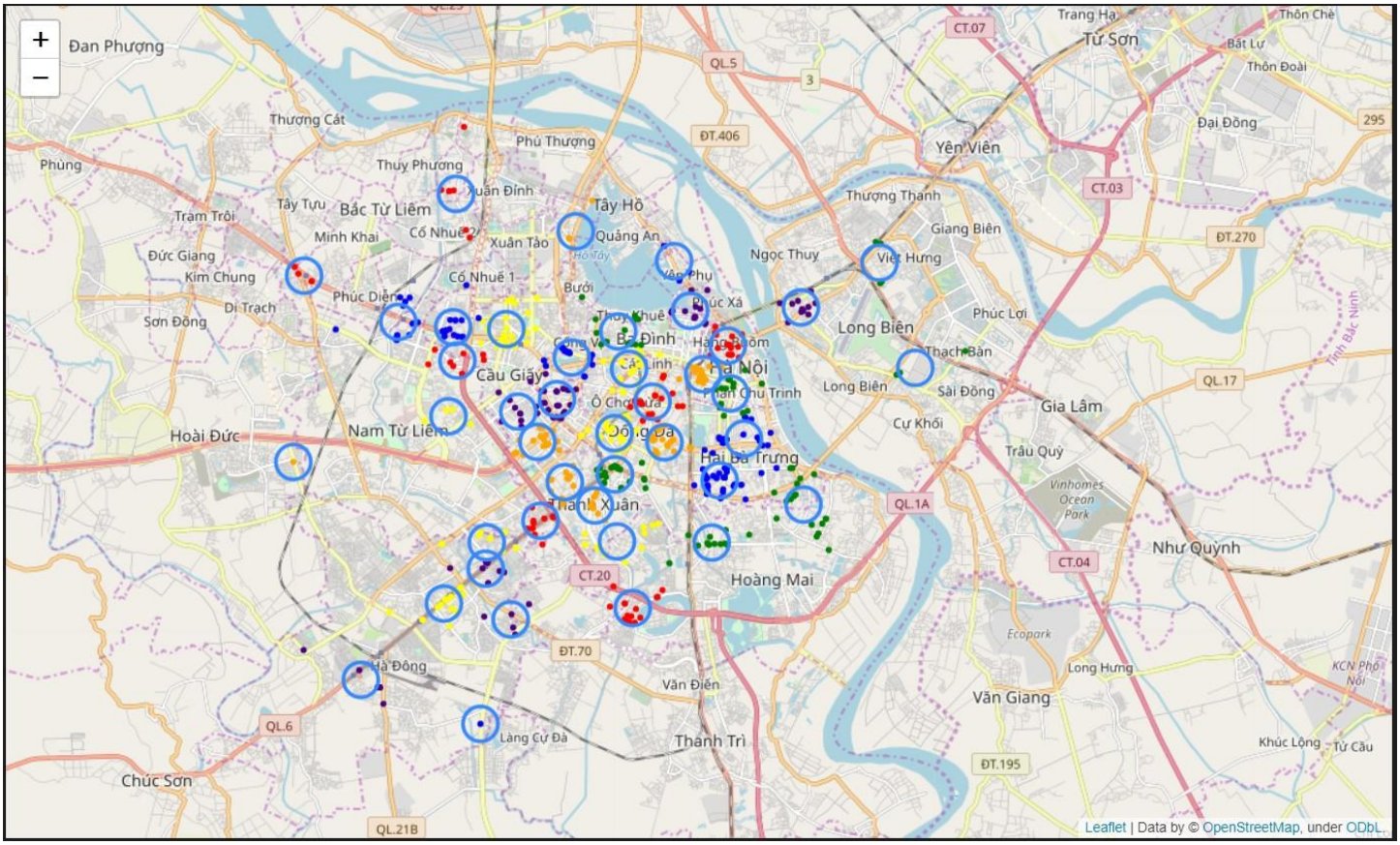


- I will also cluster all ‘CAFÉ/DESSERT’ into zones using K – means clustering for comparison.

To find k value, I try value from 30 to 60 using Davies – Bouldin Index and Silhouette coefficient to calculate how good model is. K value of



about 40 seems pretty good, it has low Davies – Bouldin Index of 0.6 (which is bad) and a high Silhouette coefficient of 0.48 (which is also bad). The model has 42 clusters and zero noise. Here is the picture after plotting:



Overall, it is bad, it is too susceptible to noises that doesn't correctly reflect some sparse areas. So, DBSCAN clustering is the one.

- Find center points of those zones and proceed to create a circular region to identify ZONE_AREA.
- Proceed to predict zones with identified factors excluding clustered zones.

VI. Conclusion

Heat map of RESTAURANT_LIMIT_PER_ZONE:

It disperses evenly in developed area of Ha Noi, around the center and diminishes further away from it.

Many cafés concentrate in tourist attraction zones which have both quantity and quality. However, these places will easily be noticeable by everyone so I don't consider it in here.

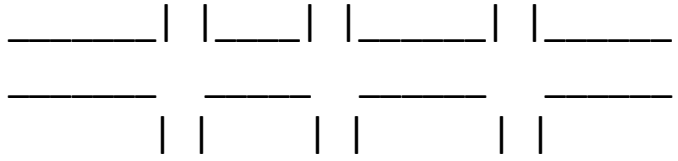
Lakes are also where many cafés reside but only lakes surrounded by large neighborhoods that cafés thrive. These places are also easily noticeable.

Apart from those traditionally famous place that everyone knows, I also spot some places that are out of the ordinary. These places don't have two ZONE_FEATURES listed above.

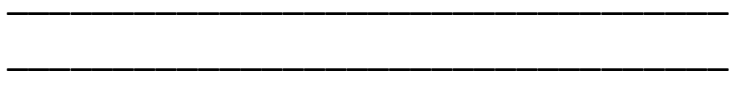
On the map, I identify a dense area which is Đào Tấn street and a sparse area near it. I then go there for further investigation.

Here is the conclusion:

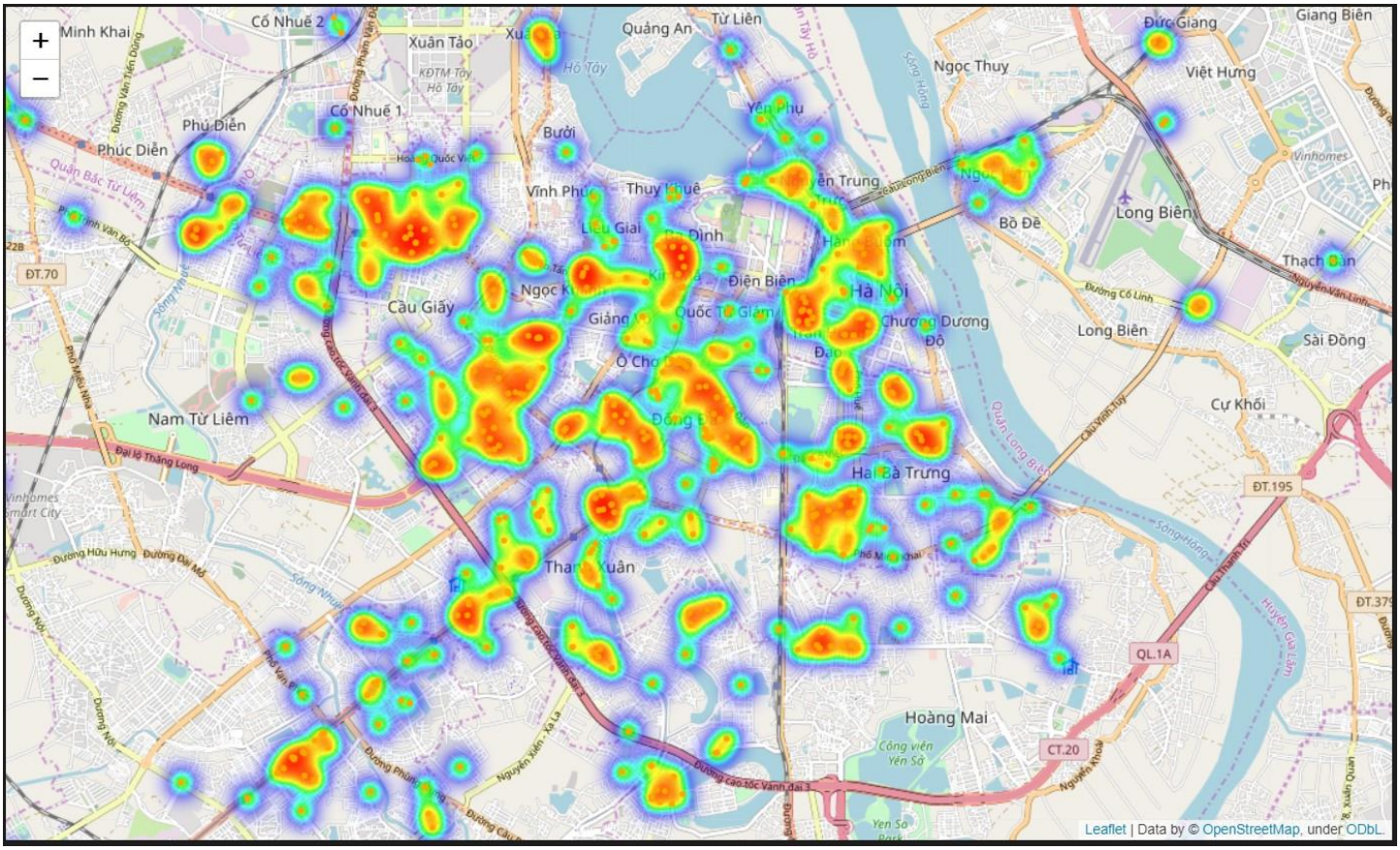
- Zones that attract a huge number of people like Japanese embassy, office districts, universities, factories, …
- Shape of road that can support easy travel to a restaurant



This type of road is much better than

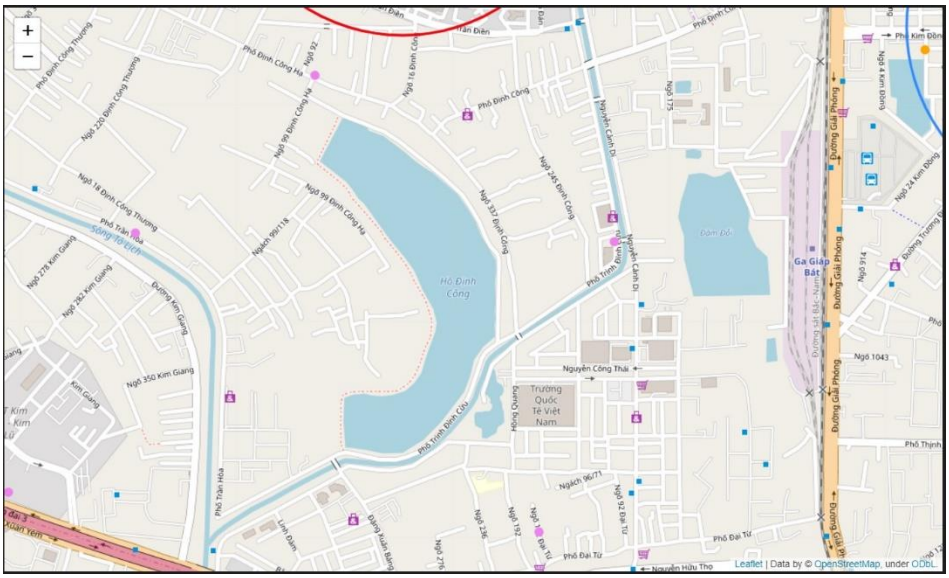


That’s why in newly developed zone, the second type of road is dominant and therefore have less concentrated area than in old imperial city zone, which the first type of road is dominant.

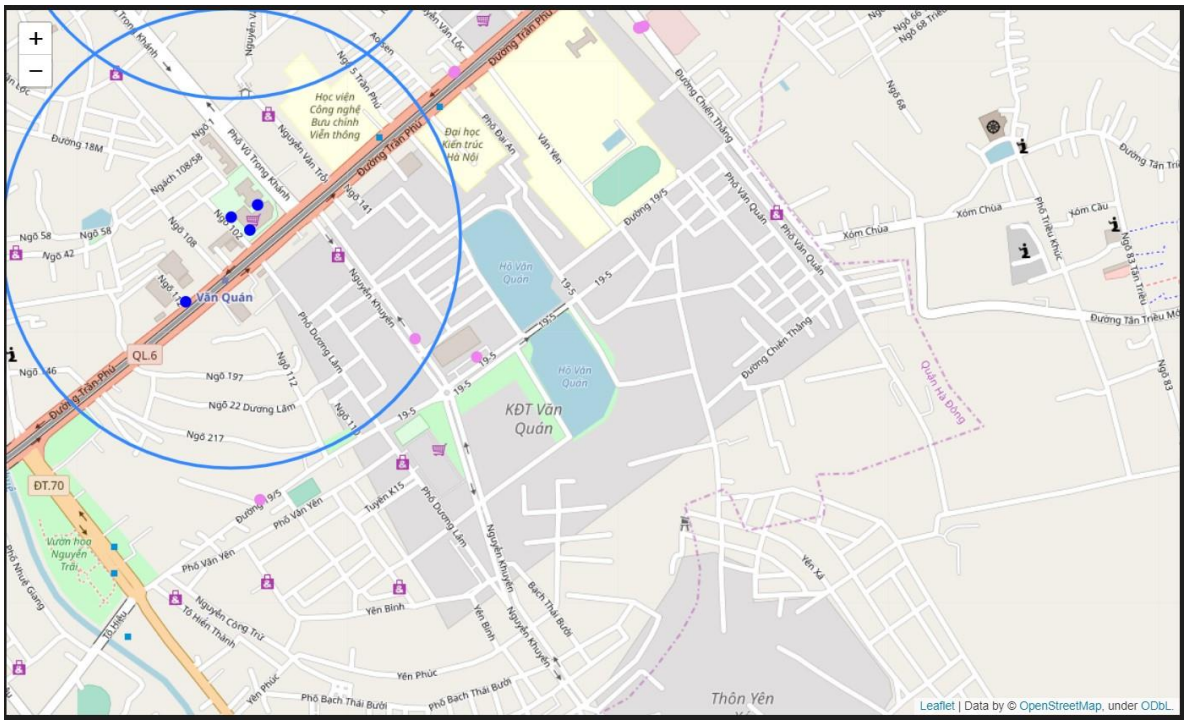


VII. Future Prediction

Due to lack of ZONE_FEATURES data needed to teach model to make prediction, I will have to rely on the map to make prediction.



As in the second and third picture, we have learned that the center region is packed with cafés and the competition is high. Based on picture 4, I look for zone with low density of Café displayed on the map. In the above picture, it has an advantage which is a lake. However, Its disadvantages are too great. The road around the lake is so sparse, and the road surrounding it is the second type of road which discourages easy travel. It doesn't have a huge population zone around it. This place has a small RESTAURANT_LIMIT_PER_ZONE. I would not open a café in this area. Here is another picture of a good potential place:



It's Văn Quán lake, surround it are two universities, a huge population zone. It also has the second type of road which allows easy travel. This is a place that I would like to open a café.