Predicting location to open a restaurant

Predicting location to open a restaurant is both time – saving and money - saving

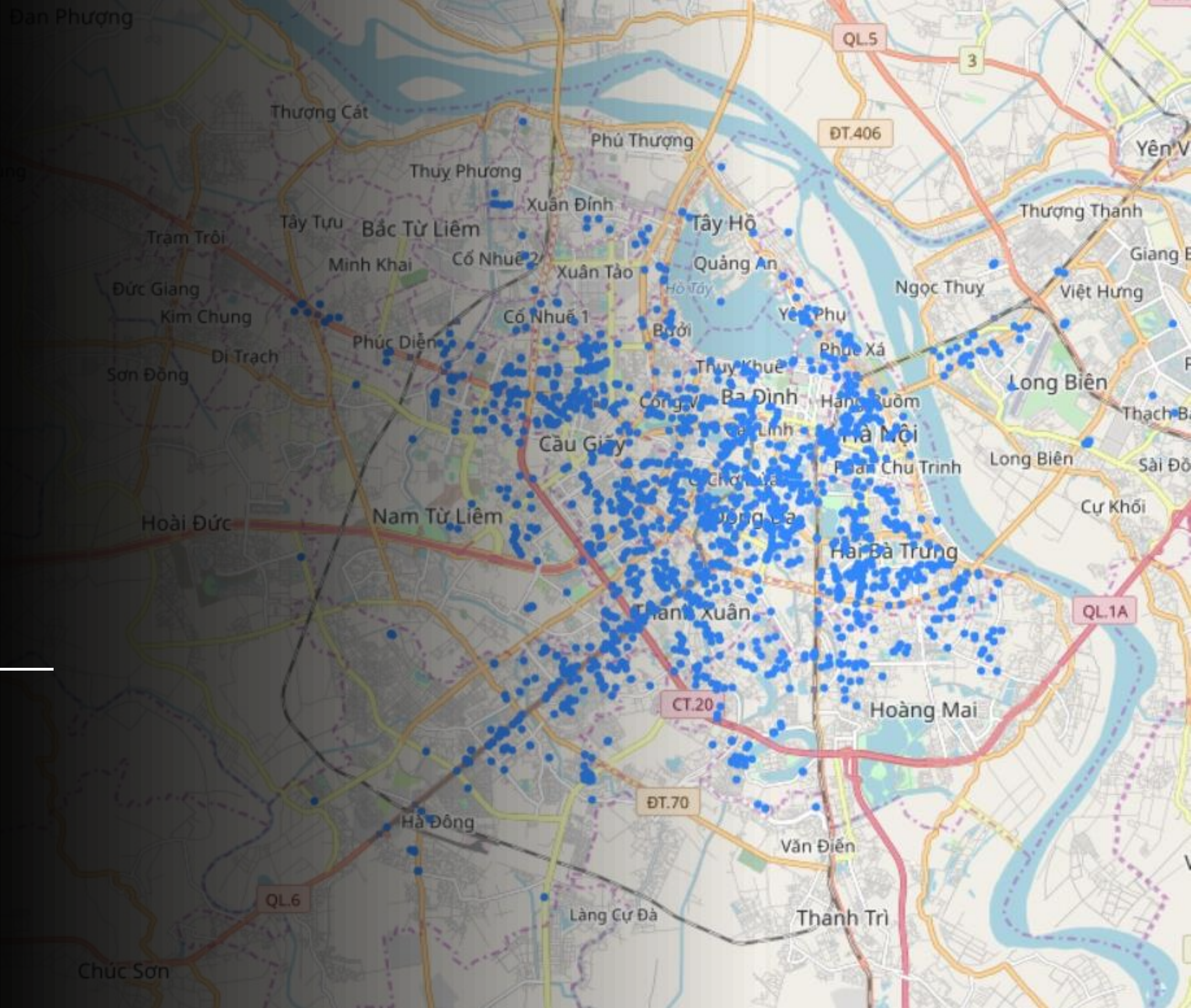
Know a good location to open a restaurant is often hard and time -demanding since it requires one ‘s deep understanding of many places. With the help of Data Science, one can overcome that job quickly and effortlessly.

Even knowing well several locations, one would have to try and test those places to see whether opening a restaurant in that place is good. A process that costs a lot can be reduced with the help of Data Science.

Data acquisition and Cleaning

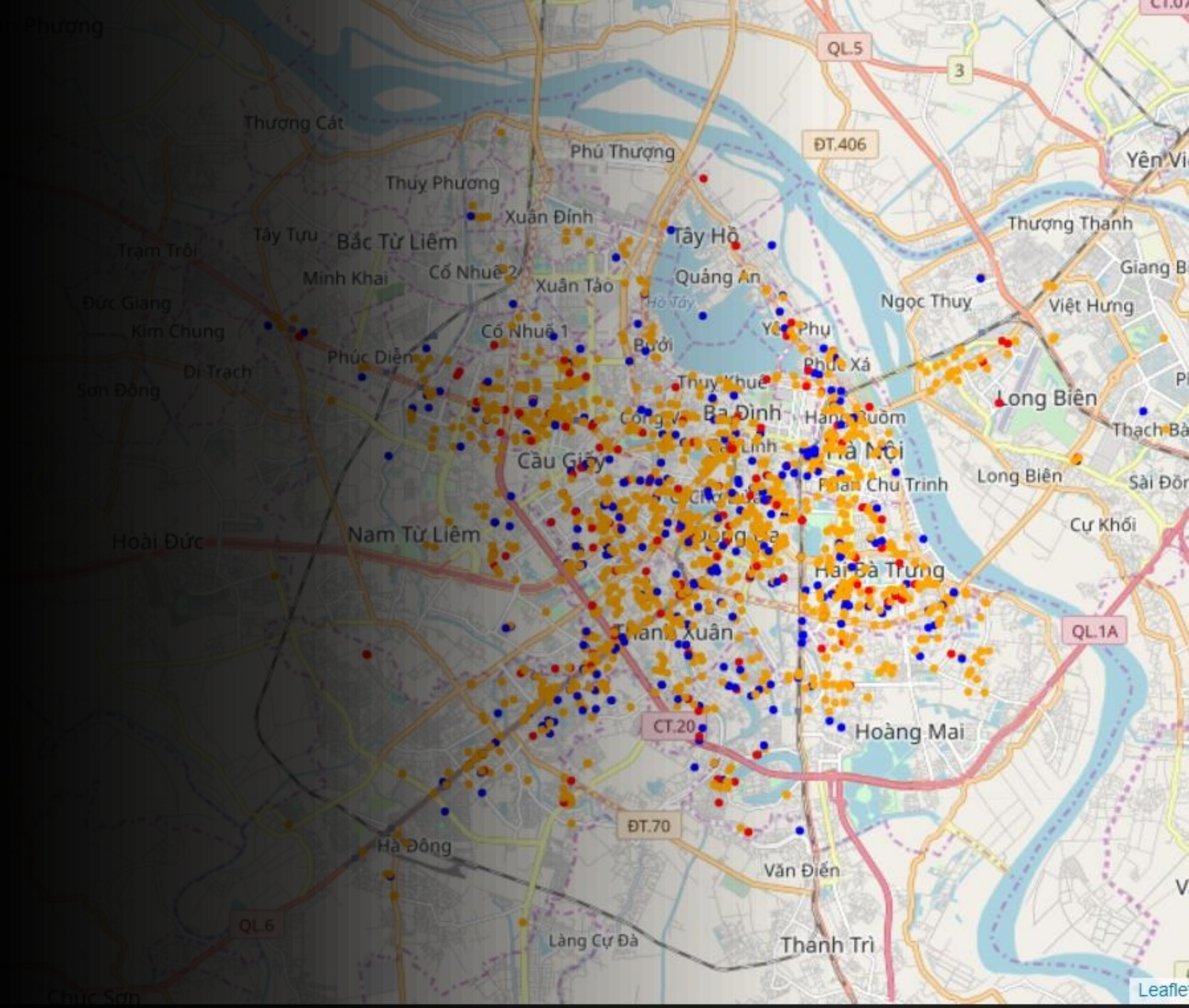
- Data is scrapped from Now.vn. The data has approximately 2000 restaurant's information including longitude, latitude, restaurant kind, food cost, and number of approached people.
- Missing value in small quantity is dropped
- Data in wrong format is changed to correct one
- Data with many different value is standardized to a specific number of value
- Cleaned data contains 5 features.

Restaurants are distributed mainly in the center and proximity area



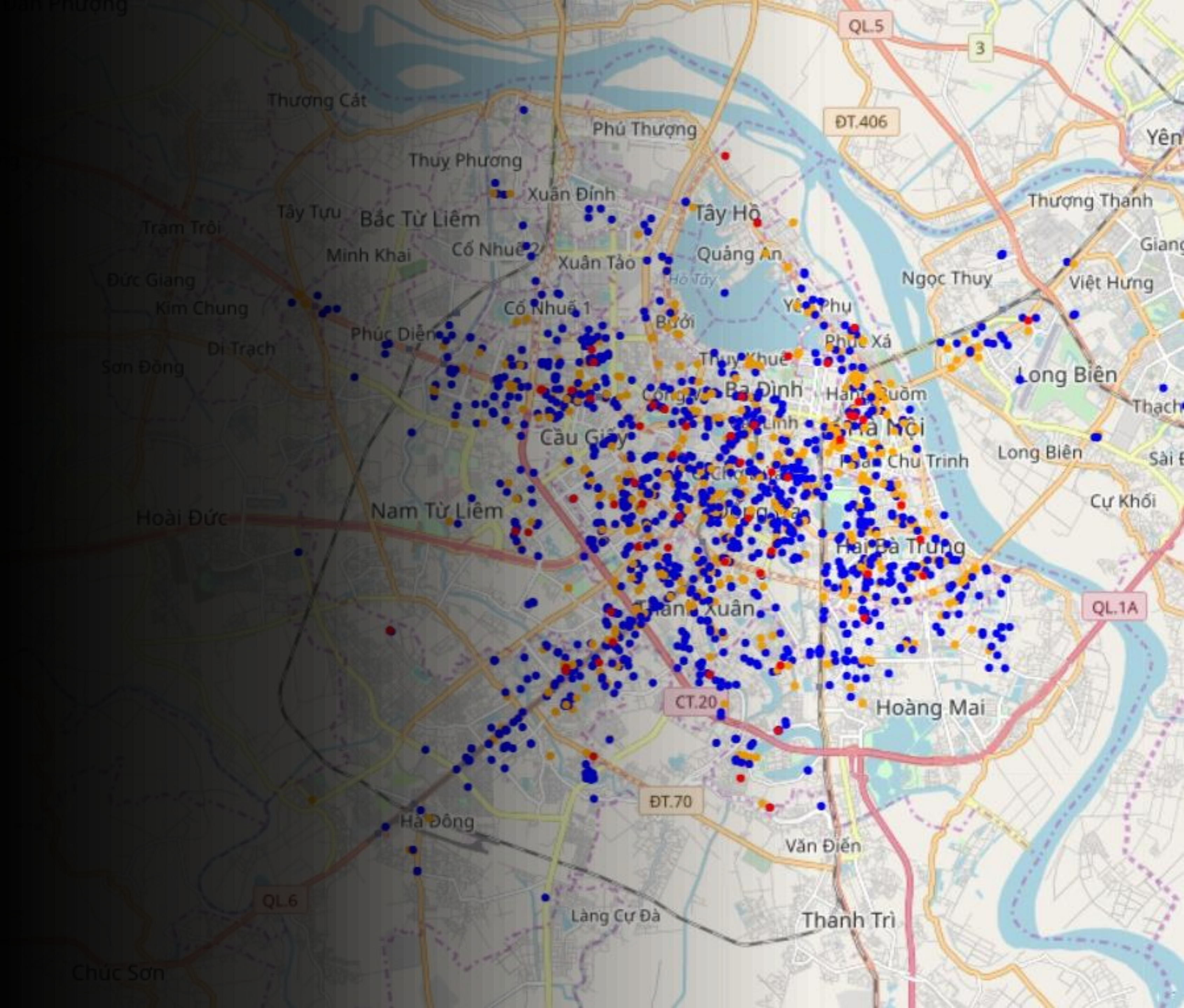
RESTAURANT_LIMIT_PER_AREA defines the number of restaurant an area can support

- Orange circle indicates restaurant with more than 100 reviews. Blue circle indicates restaurant with number of reviews below 50. Red points are unknown
- This picture shows that there are many blue circles are outcompeted by orange circle.



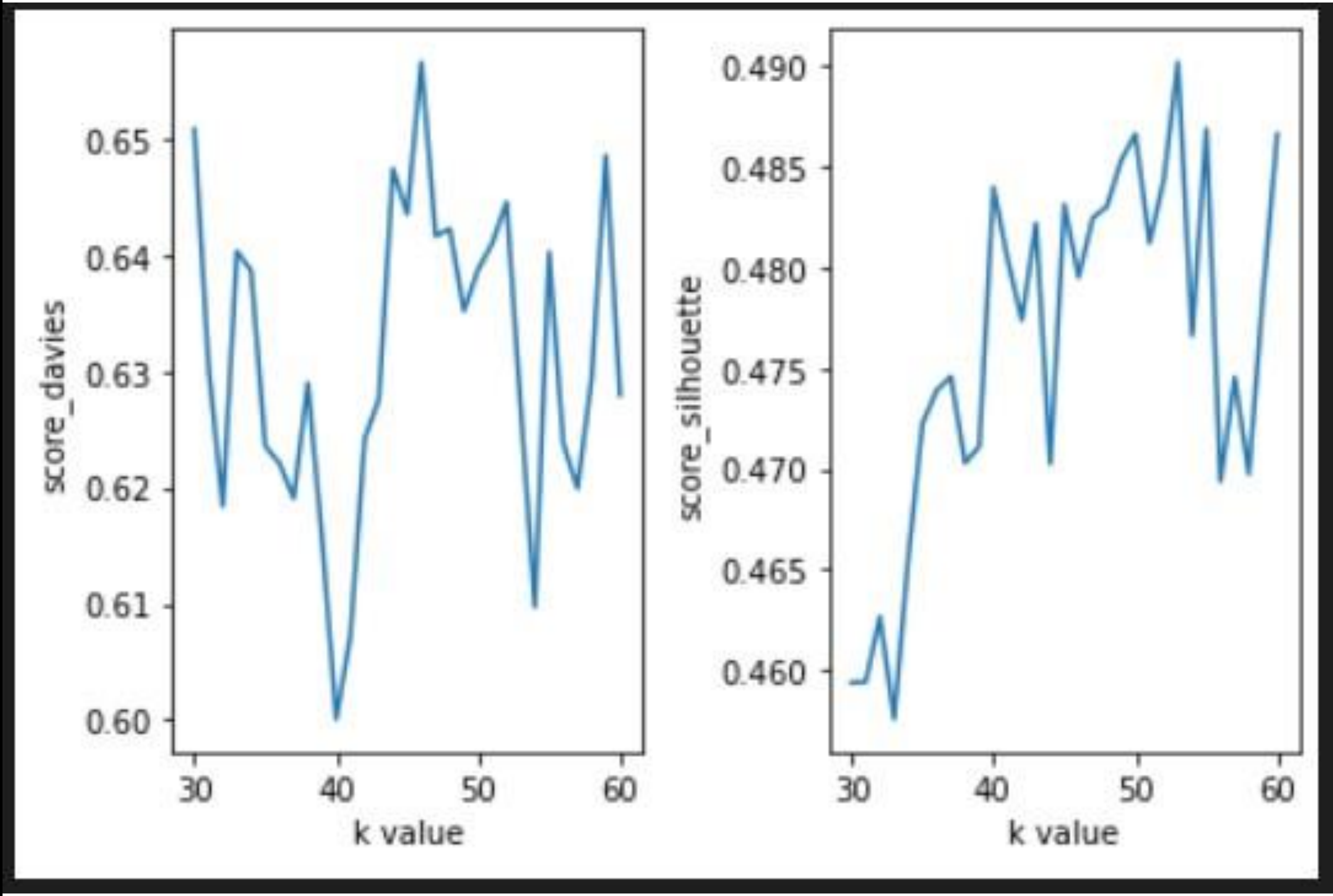
AVG_PRICE_LIMIT_PER_ZONE indicates the price range of a zone a restaurant should follow

- Blue points indicate restaurants with cheap price. Orange points indicate restaurants with more expansive price. Red points are unknown.
- This picture shows a high concentration of orange points in some areas while lower in the others.



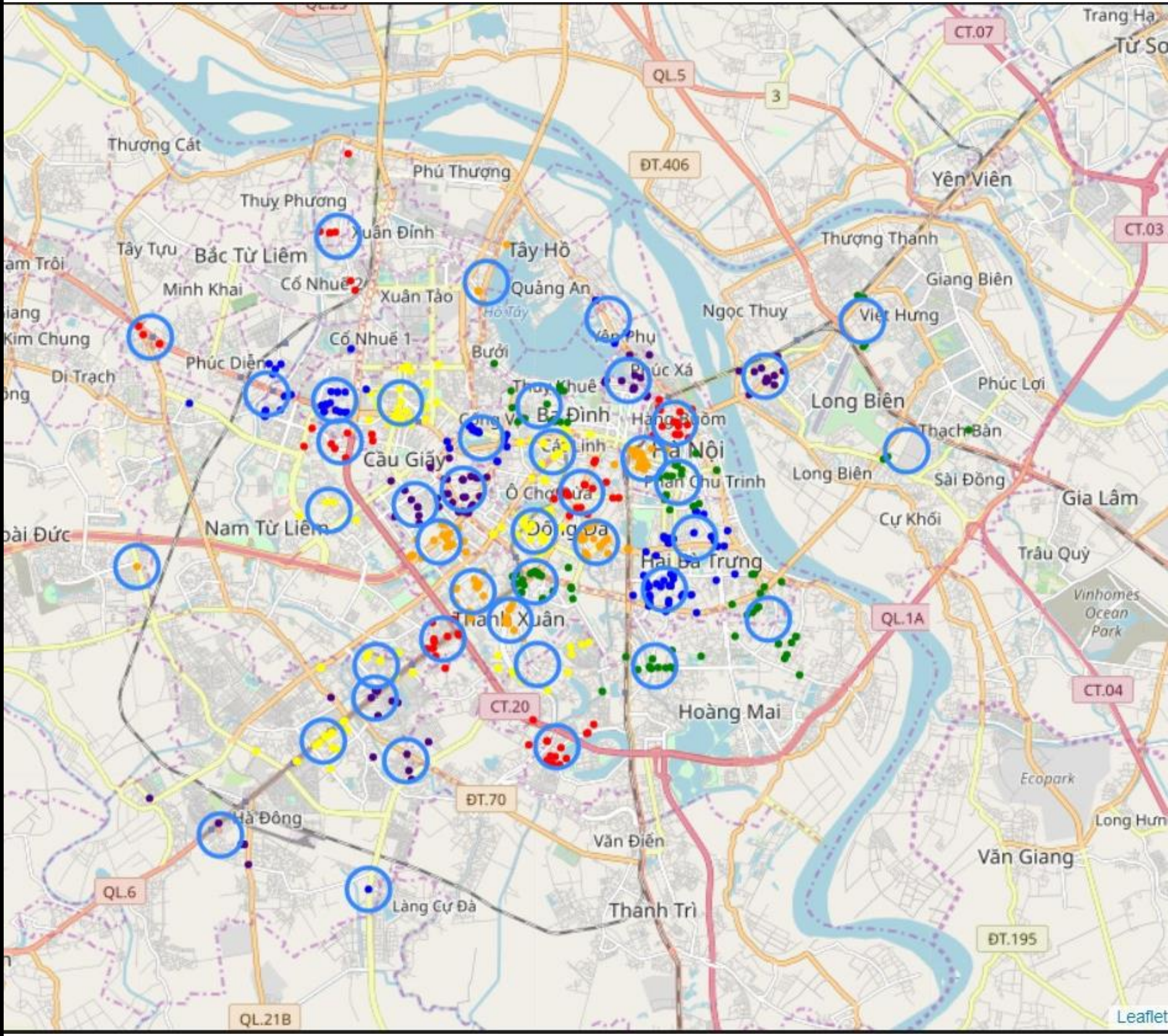
Clustering model: k - means

- The value of different k results in Davies – Bouldin Index and Silhouette coefficient is poor.
- Lowest point of Davies – Bouldin Index is 0.6. Highest point of Silhouette coefficient is 0.5.



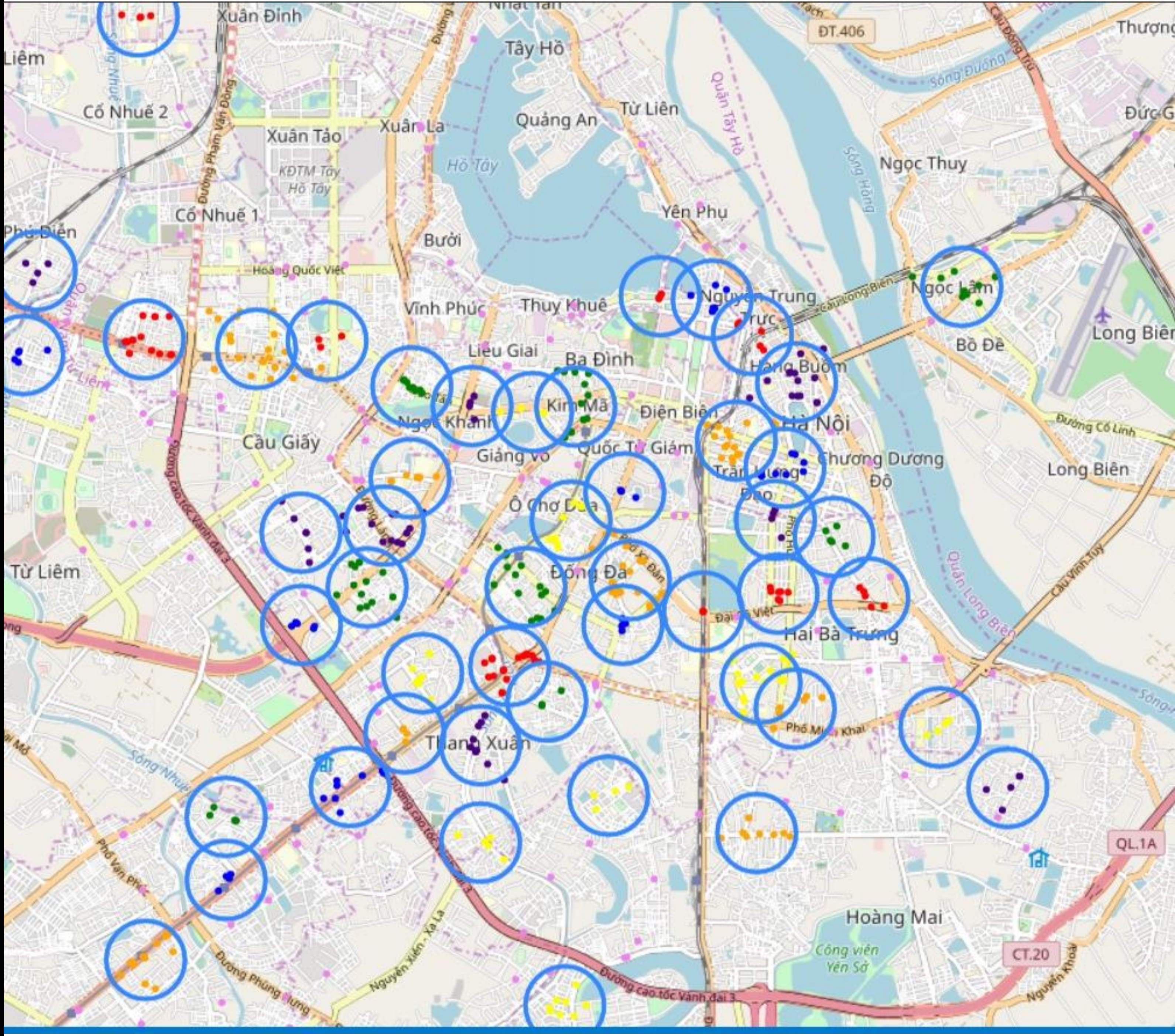
K – means too susceptible to noises in the data

- With the most prominent k value of 42 according to the chart, the result is 42 clusters and zero noise. Making clusters useless.



Clustering model: DBSCAN is much better

- With the result of 49 clusters, 169 noises, DBSCAN can cluster much better than K - means



Conclusion of deciding factors of RESTAURANT_LIMIT_PER_ZONE

- 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:

```
_______|  |____|  |______|  |______

_______   _____   ______    ______

       |  |      |  |        |  |
```

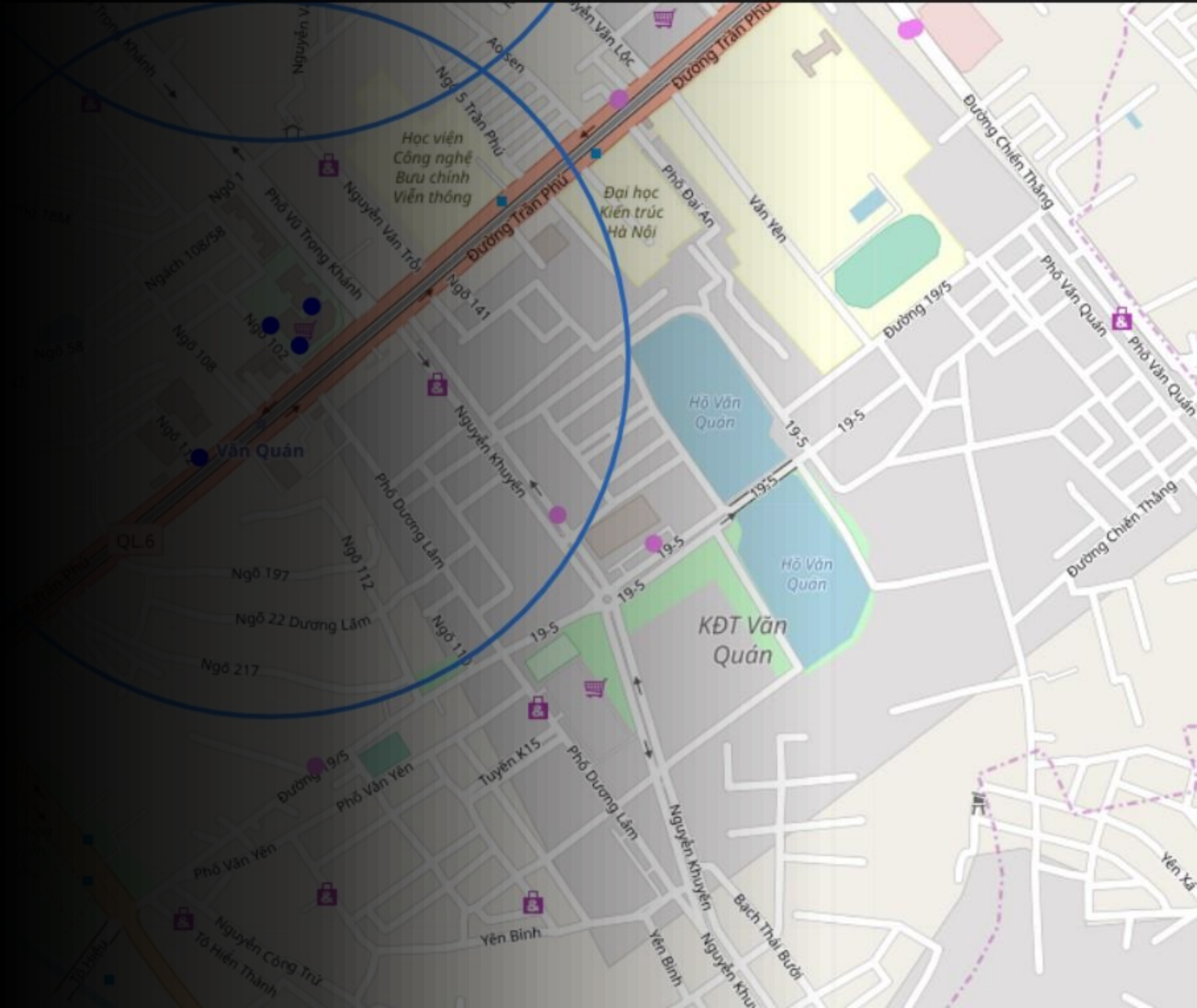
This type of road is much better than

```
___________________________________

___________________________________
```

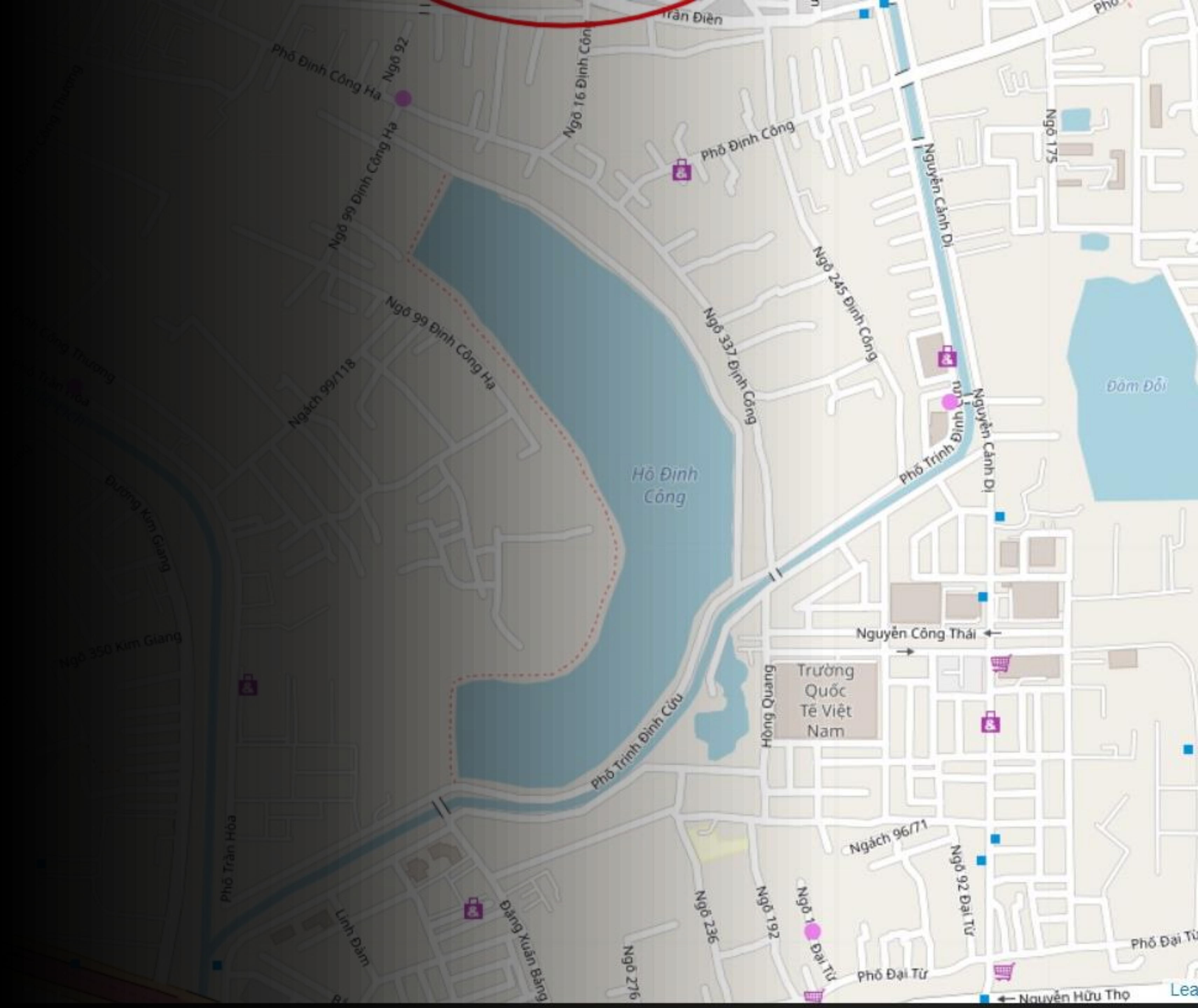
Prediction: a good location

- 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.



Prediction: a bad location

- In the 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.


Trần Điền
Phố Định Công Hạ
Ngõ 92
Ngõ 16 Định Công
Phố Định Công
Ngõ 99 Định Công Hạ
Ngõ 175
Nguyễn Cảnh Dị
Ngõ 245 Định Công
Ngõ 337 Định Công
Ngách 99/118
Đầm Đồi
Hồ Định Công
Phố Trịnh Đình Cửu
Nguyễn Công Thái
Trường Quốc Tế Việt Nam
Hoàng Quang
Ngõ 350 Kim Giang
Đường Kim Giang
Phố Trần Hòa
Ngách 96/71
Ngõ 92 Đại Từ
Phố Đại Từ
Ngõ 236
Ngõ 192
Ngõ 276
Đặng Xuân Bảng
Linh Đàm
Nguyễn Hữu Thọ