**Location for new Chain of coffee shops**

Tran Ngoc Tuyen

June 15, 2020

# 1. Introduction

## 1.1 Background

Vietnam is the second largest coffee producer in the world after Brazil. Coffee is only second to rice in value agricultural products exported in Vietnam. The coffee had become a culture in this country, from farming to consumption or enjoyment. Vietnamese would love to drink coffee. There are various of coffee shops in Vietnam and many companies, groups were established. It also has famous brands, widely known chain of coffee shops. Recently, when someone is going to create a new start-up, most people think about he will open a coffee shop. This is the truth because of its simplicity and fast growth, compared with other industries. The most important factor is very big market size and room for growth. Thus, there are many ideas for a new coffee shop even chains of coffee shop.

## 1.2 Problem

When we plan to open new shop selling coffee cups, we must choose the right locations. A lot of factors impact to the decision like population, center distance, around offices, spaces cost, … Data of venues locations might contribute to determining the worth area to place a new coffee shop. Assuming, there is a relationship between number of coffee shop and number of other venues. Example, in specific location has many buildings and coffee shops, another location has a few buildings and the coffee shop as well. We might predict number of shops in a particular area based on number of buildings.

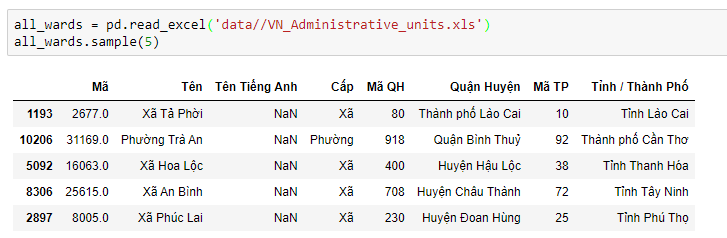
## 1.3 Interest

Business owners would be very interested in predicting number of coffee shops an area should have. This can help them to determine if that area is lacking or abundant of coffee shops. Then, they consider place their chain where is shortage of coffee shops.

# 2. Data preparation

## 2.1 Coordinates data

Firstly, get the geographic data of areas which have strongest consumption of coffee. In this project, I chose 3 big cities in Vietnam, they are Ha Noi, Da Nang and Ho Chi Minh city. It is very simple to downloading Vietnam administrative unit data from <https://www.gso.gov.vn/dmhc2015/>. The data is structured and stored in .xls format, so we can easy to read the file using Pandas library.



The location unit in project is ward. We have total 10619 wards in 63 provinces in Vietnam but we apply filter to get only wards in 3 big cities.



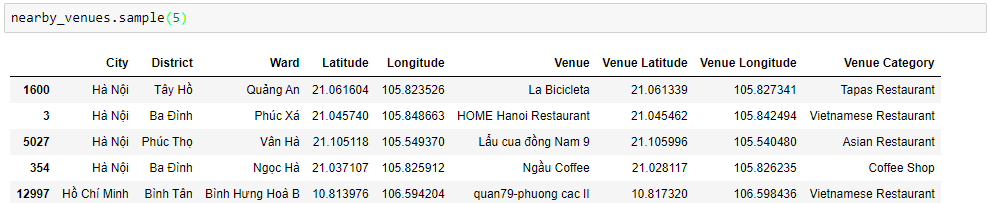
Next is getting coordinates of these wards. I used Geocoder library, after tried many providers I found that komoot is the fastest, almost fully locaction API, it's also free and unlimited. The data coordinates should look like



## 2.2 Venues data

Using Foursquare API, I got the venues information in a Ward like name, location, category, … For the more detail, please visit [Foursquare](https://foursquare.com/) webpage. The set of parameters for the get API is URI address, my personal ID and secret, version = 20200101, radius = 1000 meters, maximum venues = 50.

Because of my trail account and the unstable of API, I got 13611 venues in 754 wards. The number of data points is 754 as well.

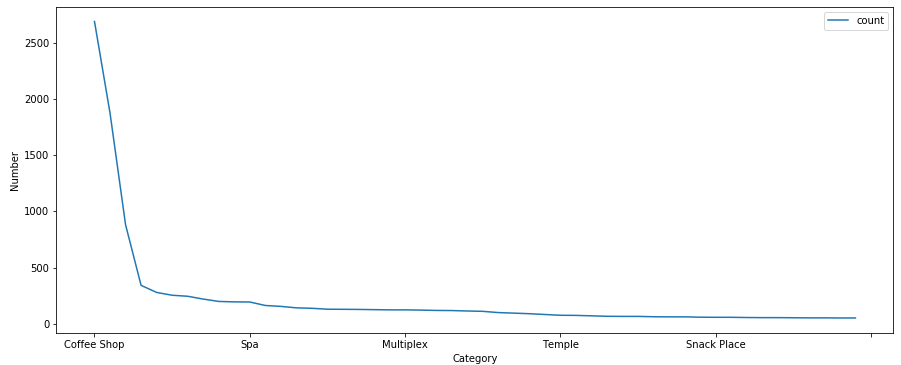


The detail step of cleaning data is in the notebook file.

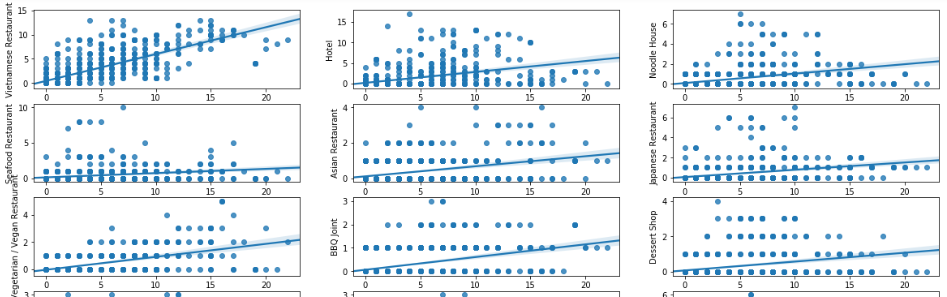
# 3. Data Exploratory

## 3.1 Category analysis

Now we have 325 venues categories. Count number of venues in each, the most category is Coffee shop with 2688 venues. That number is not surprise us, because this is country of coffee and it’s also the target.

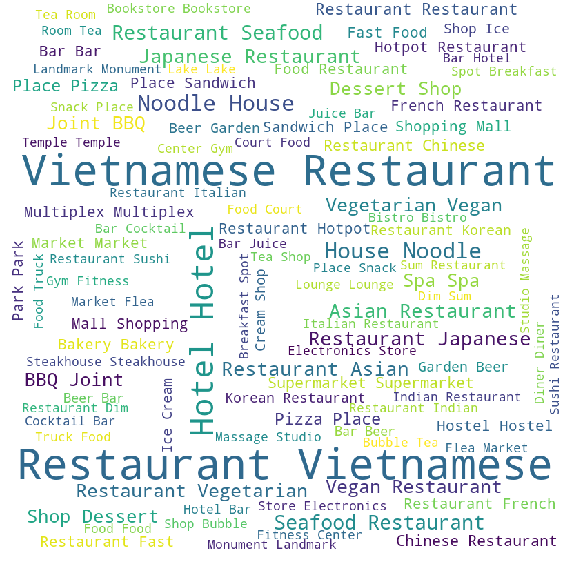


Regression relationship between each category to target.

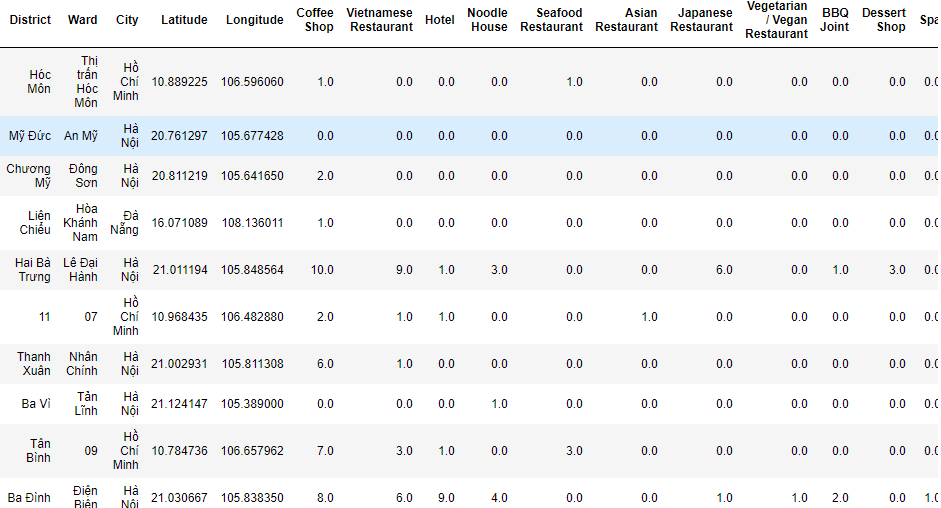


## 3.2 Feature engineering

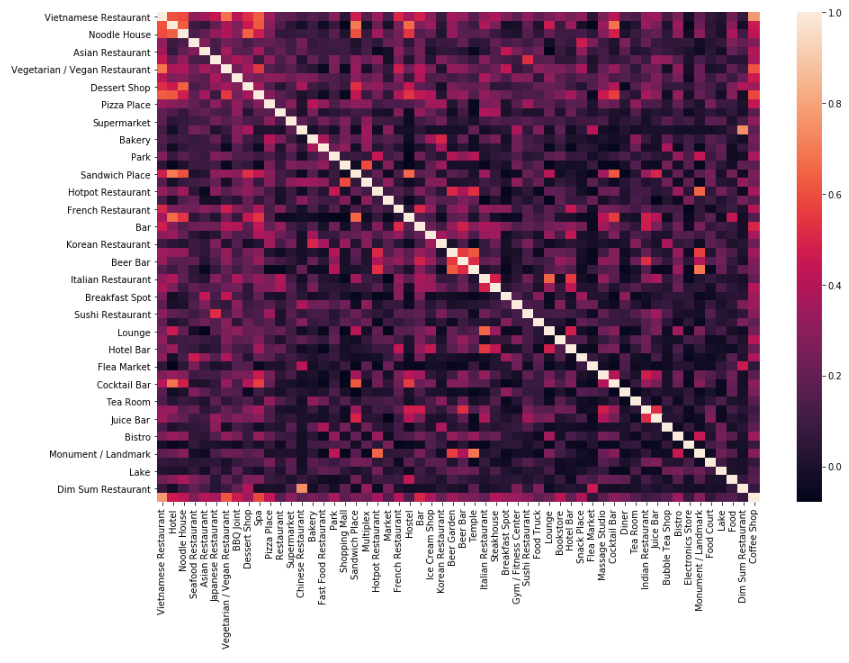
I assumed that these 325 categories are independent variables. But they’re likely too much and unreliable. I chose only categories which have more than 50 venues and the number of features is 53.



Next step is transforming the value of categories. Each Ward has 53 features / major categories, the value of these features is simply number of its venues. Group by Ward, I created matrix with 754 vectors and 53 dimensions:



Let’s check the correlation between features, Pearson standard coefficient is used



We can consider removing features with high correlation (greater than 0.7). In project, I keep all of features.

|  |  |
| --- | --- |
| The top features which high correlation with target are ‘Vietnamese Restaurant’, ‘Vegetarian’, ‘Spa’, ‘Bar’, ‘Hotel’.  These features may be strong variables in predictive model. |  |

# 4. Predictive modeling

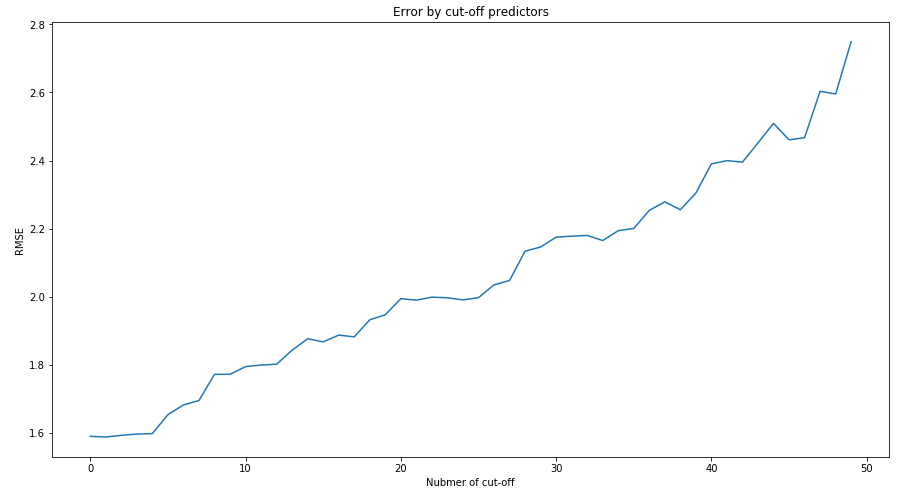
Because we predict the number of coffee shops, type of model should be regression.

## 4.1 Linear Regression

Split random data into train, test set with test\_size = 25%. Evaluate metric is root-mean-square-error. The first model had RMSE = 1.59

|  |  |
| --- | --- |
|  |  |

I tried to cut-off number of features by remove weak features from model. The weak feature is low correlation with target.



As we can see, the more features we cut-off, the increase in error. The acceptable cut-off thresh hold is 4, because it did not increase model error too much.

The final linear regression model has 49 predictors, RMSE = 1.6

|  |  |
| --- | --- |
|  |  |

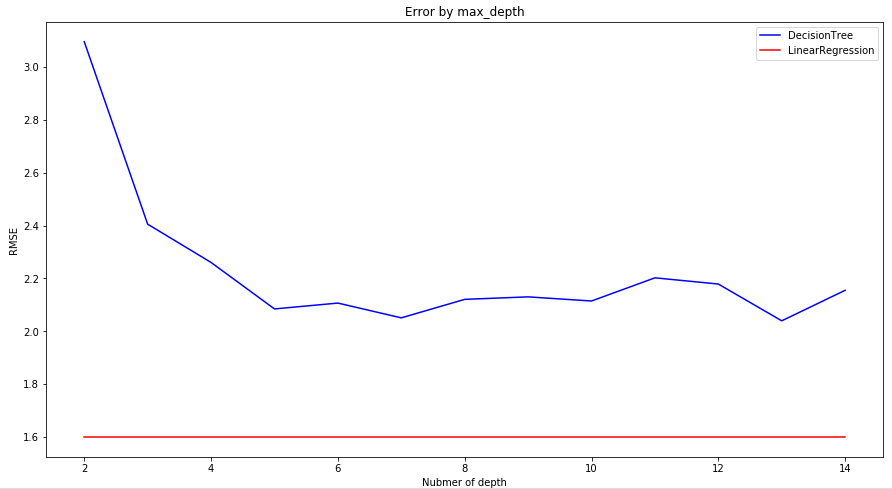
## 4.2 Decision Tree

Another algorithm is decision tree, it can work well with regression type.

Draft tree has max\_depth = 5, criterion by mse. The result has RMSE = 2.09

|  |  |
| --- | --- |
|  |  |

It looks like does not better than Linear Regression. I was thinking optimize the decision tree model by change the max\_depth parameter.



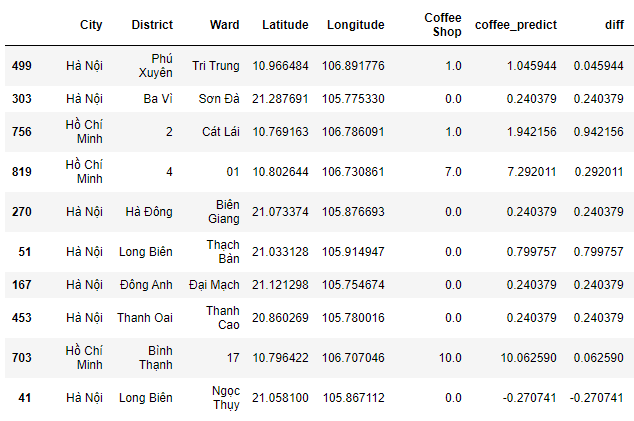
The error on Decision tree model is always higher than Linear Regression model’s.

# 5. Location for new coffee shop

In this part, by comparison number of coffee shop should be in prediction with the current, we can choose the area for new coffee shops.

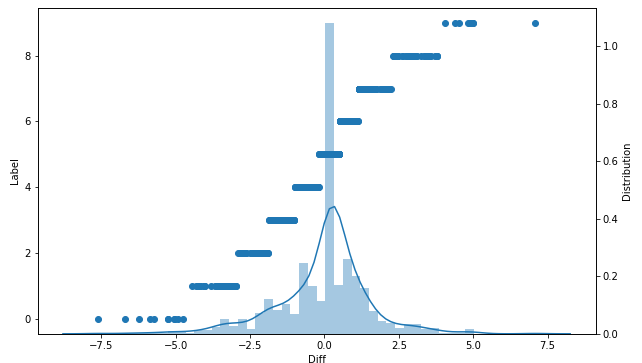
## 5.1 Prediction

Using the final model in part 4 to predict number of coffee shop of all Ward in 3 big cities. Then calculate the difference between prediction and reality.



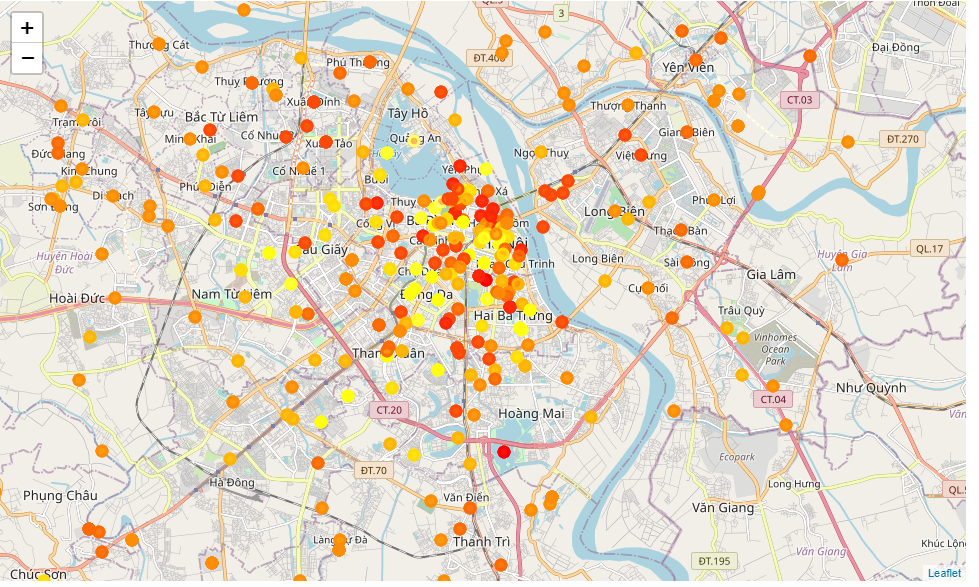
## 5.2 Shortage of coffee shop

I created visualize map and mark all wards by folium library. It is necessary to divide Diff value into bins, so I used k-means algorithm to create 10 clusters for Diff levels.



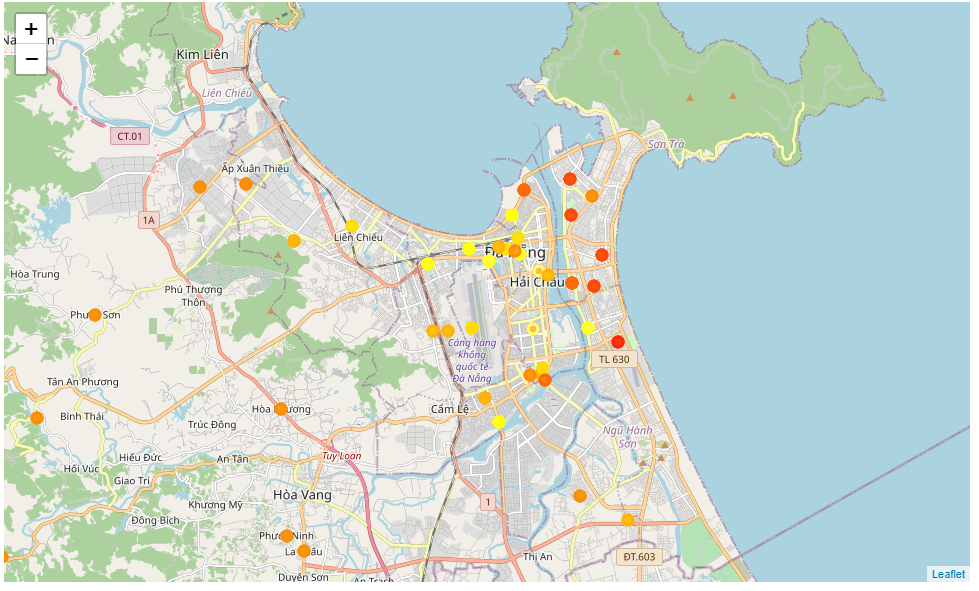
The Wards have label greater than 5 or positive Diff, is lack of coffee shops. In the map, these Wards denoted by lighter yellow circles. In the other hand, the red circles are Wards which have much of coffee shops.

### 5.2.1 Check the locations in Ha Noi city



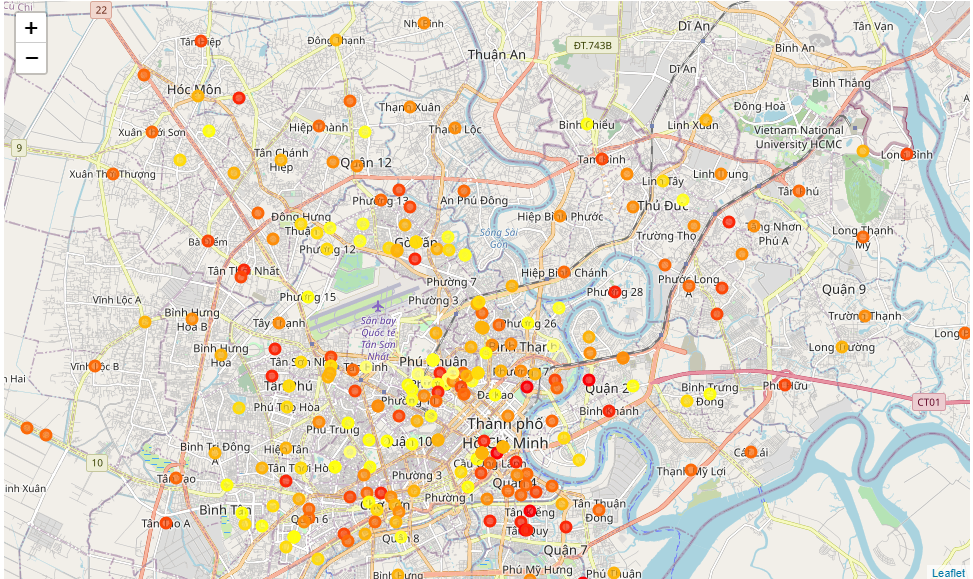
There are many places, even in center of city which now lack of coffee shops.

### 5.2.2 Check the locations in Da Nang city



Same as Ha Noi city, we may find at least 5 Wards to open new coffee shops in center of Da Nang city.

### 5.2.3 Check the locations in Ho Chi Minh city

Unlike above cities, HCMC has less Wards to open new coffee shop.

# 6. Conclusions

In this project, I analyzed the relationship between number of coffee shops and number of other venues. Some category venues have strong impact to appearance of coffee shops or vice versa. They are Vietnamese Restaurant, Bar or Hotel venues. I built models and I optimized them to predict number of coffee shops a Ward should have. From this model, I estimated which area now is shortage of coffee shops and plot them in visualize maps. This information can be useful for the starter, who want to open new coffee shops or for the chain of coffee owners to expand /collapse their business.

# 7. Future directions

The scope of the study is only 3 cities in Vietnam, I will get more data from many cites over world. In part 2.2, I got venues of 754 / 946 wards, it does not mean that in left 192 wards do not have any venues, I think the problem come from wrong coordinates or unstable Foursquare API. I should find other venues provider to improve this. In the project, I used only 2 algorithm to create model, I think it can be better if I try others like SVM, Random Forest, Gradients Boots ,…

This may be an omission that I predict the shortage of coffee shops only on number of venues. I ignored other factors like the cuisine, the needs, or the living standards of the area. I the future, I will add these in the study and the predictive model would be significantly improved.