**Restaurant Reccomendation in New York**

*Dang Nguyen Hoang*

*December 2020*

1. **Introduction:**
2. Background

New York is one of the most populous city in the United States, and also the [most densely populated](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population_density" \o "List of United States cities by population density) major city in the United States with an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2).

With this population and dense, it seems that opening a Restaurant in New York is both a chance and challenge.

1. Problem

The most challenging is to locate the for the Restaurant under some conditions:

* High Density of living
* Low competitors

1. Solution

The problem of finding High Density should be solved by NY\_Airbnb database, using ‘DBScan’ and ‘Kmeans’ Technology while checking for Competitors can be done by Foursquare API

1. **Data Acquisition and Cleaning**
2. Data Sources

The data for New York Airbnb is acquired from [this link](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data) of Kaggle.

The API that used for checking the Restaurants around the locations is [FourSquare](https://foursquare.com/)

1. Data Cleaning

Data of Airbnb that downloaded from Kaggle that only need longitude, latitude, neighbour and neighbour\_group which were enough and no need to clean.

Data that get from FourSquare need to be extracted, especially in Categories column

1. Feature Selection

As mentioned above, the Features were selected by 2 datasets are:

**New York Airbnb:**

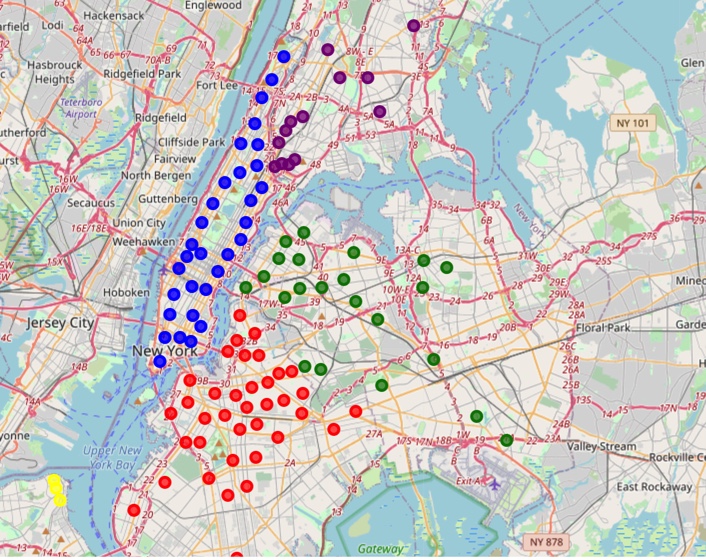
* + Neighbourhood\_group
  + Neighbourhood
  + Longitude
  + Latitude

**FourSquare:**

* + Longitude
  + Latitude
  + Categories
  + Name

1. **Exploratory Data Analysis**

3.1 There are 48895 locations of Airbnb found from the dataset. According to my experience, the Airbnb are dense in the center of the districts. However, there are some located far from the center of the District/Neighbourhood. If we use coordiantes of those Airbnb, then we can not find the center of the District. For this reason, DB-Scan was used in first hand to remove the ‘outliers’ in coordinates. After that, K-means was used to find Centers of each Neighbors



After cleaning and applying DB-scan, K-means, we can estimate the Center of Dense Zone in each neighbor as above