**1. Discussion and Background of the Business Problem:**

Problem Statement: Prospects of a Restaurant, Close to Office Areas in Shanghai, China.   
  
During the peak dinning time, office areas provide huge opportunities for restaurants. Usually the profit margin for a decent restaurant lie within 15−20% range but, it can even go high enough to 35%, as discussed here. Shanghai is made of 16 districts but, I will later concentrate on 8 most busiest business districts  —  Jing'an District(静安区), Huangpu District(黄浦区), Xuhui District(徐汇区), Changning District(长宁区), Pudong New Area(浦东新区), Hongkou District(虹口区), Putuo District(普陀区), Yangpu District(杨浦区)   
  
We will go through each step of this project and address them separately. I first outline the initial data preparation and describe future steps to start the battle of neighborhoods in Shanghai.   
  
**Target Audience**   
  
What type of clients or a group of people would be interested in this project?

* Business personnel who wants to invest or open a restaurant. This analysis will be a comprehensive guide to start or expand restaurants targeting the large pool of office workers in Tokyo during lunch hours.
* Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.
* New graduates, to find reasonable lunch/breakfast place close to office.
* Budding Data Scientists, who want to implement some of the most used Exploratory Data Analysis techniques to obtain necessary data, analyze it, and, finally be able to tell a story out of it.

# 2. Data Preparation:

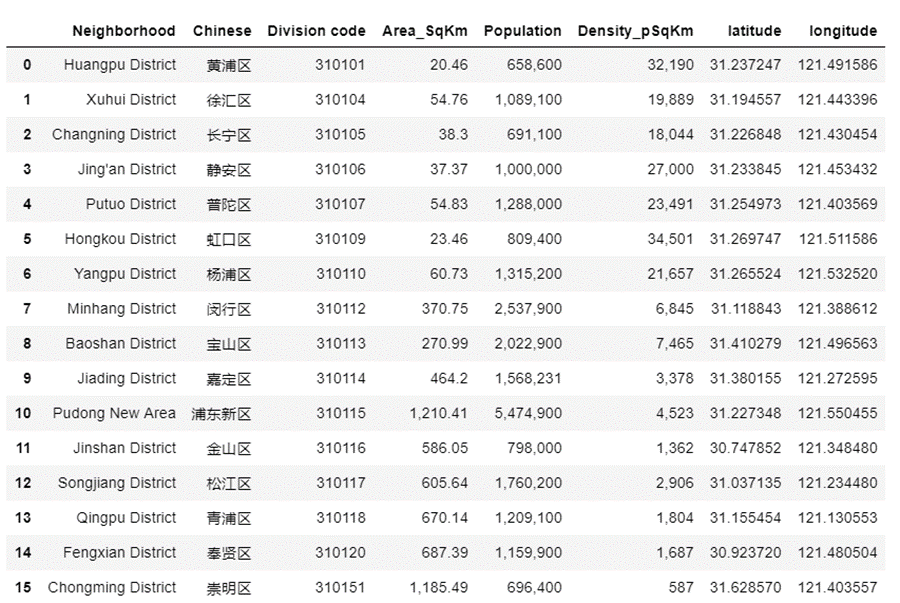
## 2.1. Extract data from Wikipedia

I used the list of the administrative districts in Shanghai as my original start, which come from Wikipedia (<https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai)>. The data was just extracted and saved. The data was shown as below.

## 

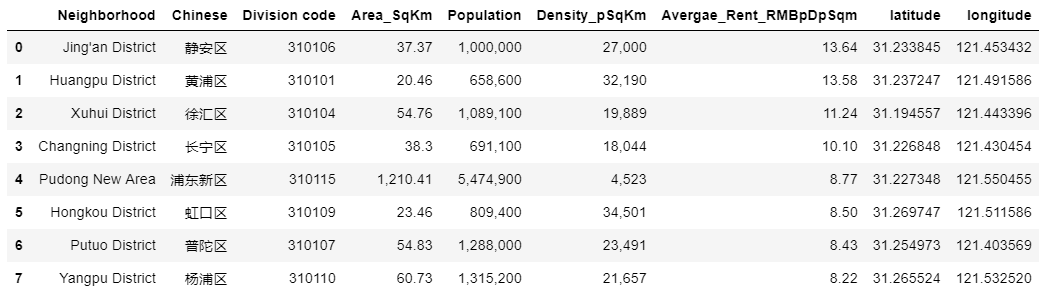
## 2.2 Getting Coordinates of Major Districts

Here it should be mentioned that the geopy package was first applied to get the geology information, however, after the data was visualized in the map, it indicates that the geology location was not accurate. In order to resovle this issue, I chose baidu map api to get the geology information which is a Chinese Map APP, and the results seem like more reasonable than previous API. The code was shown as below.

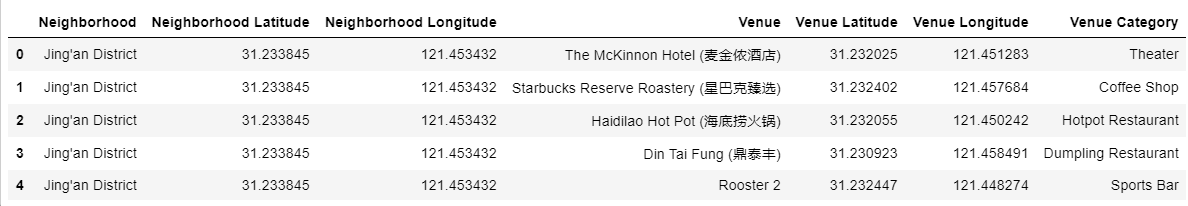




### 2.3 Average Rent Price in Districts of Shanghai

As mentioned in the beginning, the office area are always with high opptunities for restaruants,since the peak dinning hour. And this area always has a hihger rent price, which is a good factor that can guide us to decide which district would be beter to open a restaurant.   
  
Since there is a no website to provide compile data, I collect the information of every districts in the 'creprice' website, which has a latest rent price of different districts.(Ack. <http://www.creprice.cn/district/BS.html?city=sh&type=lease&proptype=22>) The rent price was compiled into the list of Shanghai districts. The 8 of the most higher price districts were selected which considered as the busiest business area, the data was shown as below.It should be noted that "Avergae\_Rent\_RMBpDpSqm" means Average rent per day per square in RMB.

### 2.4 Using Foursquare API to find the location data

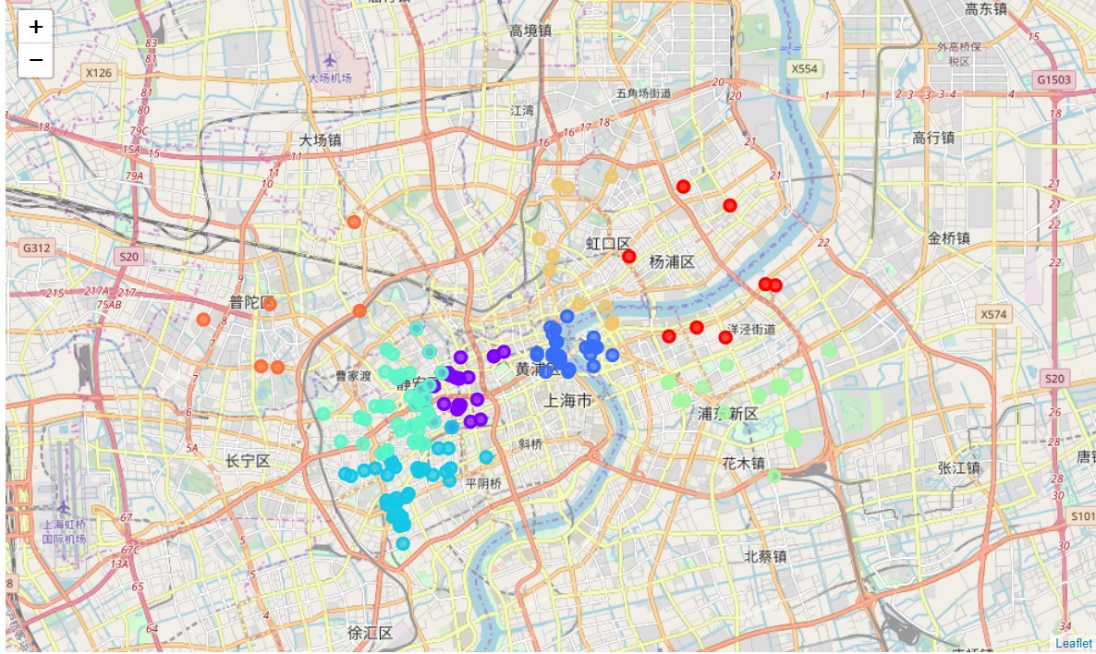
For this business problem I have used, as a part of the assignment, the Foursquare API to retrieve information about the popular spots around these 7 districts in Shanghai. The call returns a JSON file and we need to turn that into a dataframe. Here I’ve chosen 100 nearby spots for each major districts within a radius of 3 km. Below is the progress that I retrieve the information and dataframe obtained from the JSON file that was returned by Foursquare

## 3. Visualization and Data Exploration:

### 3.1. Folium Library and Leaflet Map:

Folium is a python library that can create interactive leaflet map using coordinate data. Before we further explore the restaurant information,I want to show restaurants location in these 8 districts and the spots related to the restaurants were selected out based on the 'Venue Category'. Since some spots are restaurants but without a 'restaurant' in 'Venue Category' text, I maually selected out the relevant categories and shown as below.

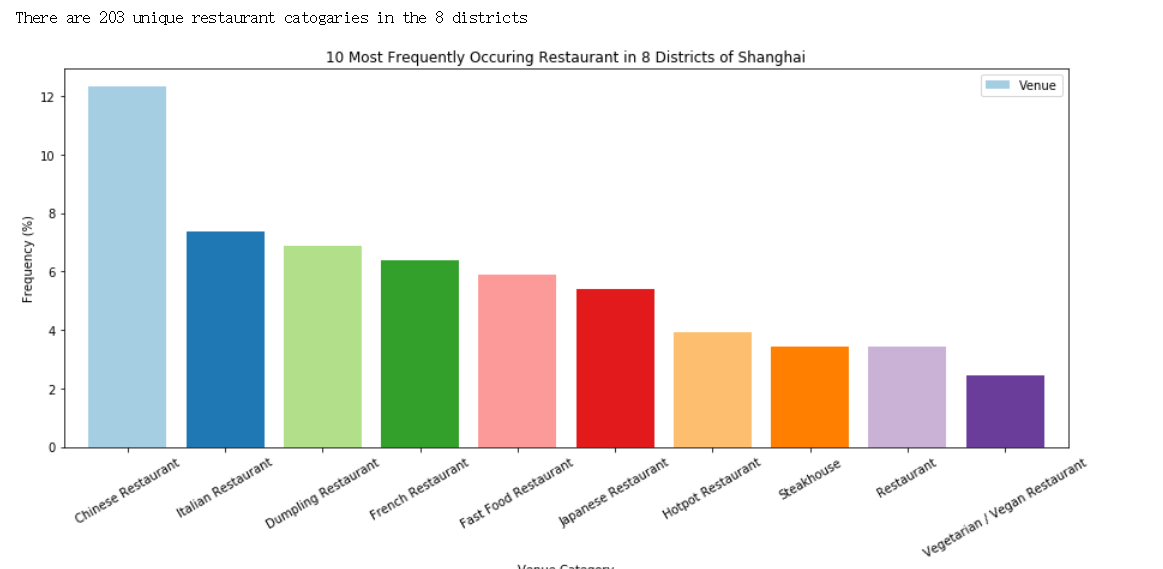
Based on the dataframe, a leaflet map with Folium is created to see the distribution of the 8 districts. Circular marks represent the relevant restaurants in the 7 Districts (Putuo - Orange,Changning - Cyan,Xuhui - Light Blue, Jing'an - Purple, Huangpu - Blue, Hongkou - Light Yellow,Yangpu - Red, Pudong New Area - Green ) districts of Shanghai, according to Foursquare data.   
  
It is easy to see that the density of restaurants in Changning Xuhui, Jing'an and Huangpu districts is much higher than the rest districts.



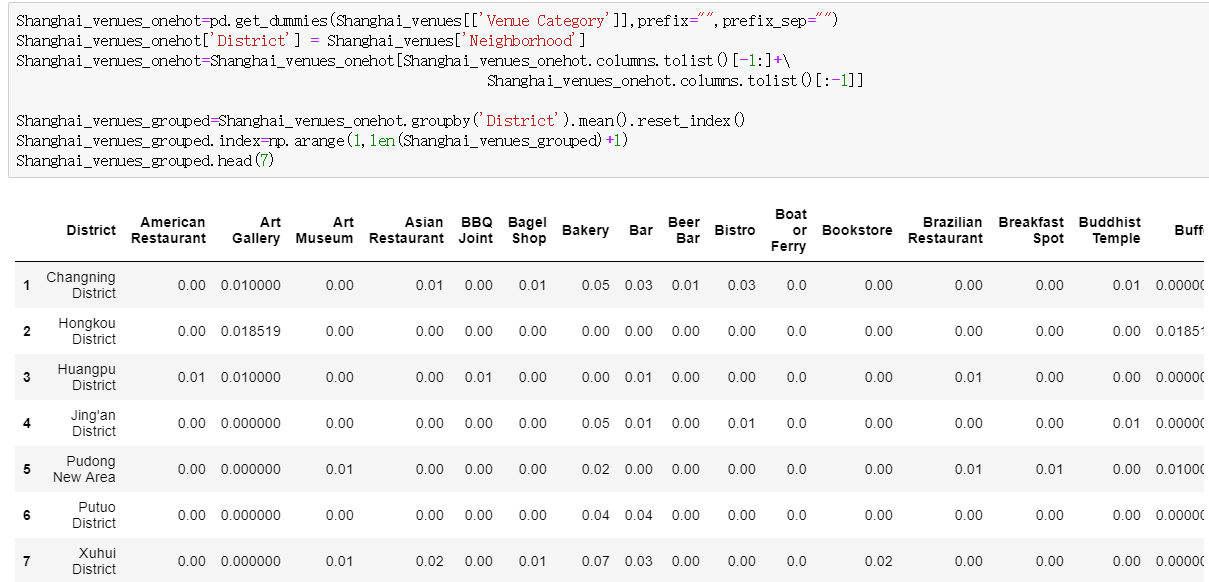
### 3.2. Exploratory Data Analysis:

Let's explore more details of these restaruants, after the analysis we can conclude:

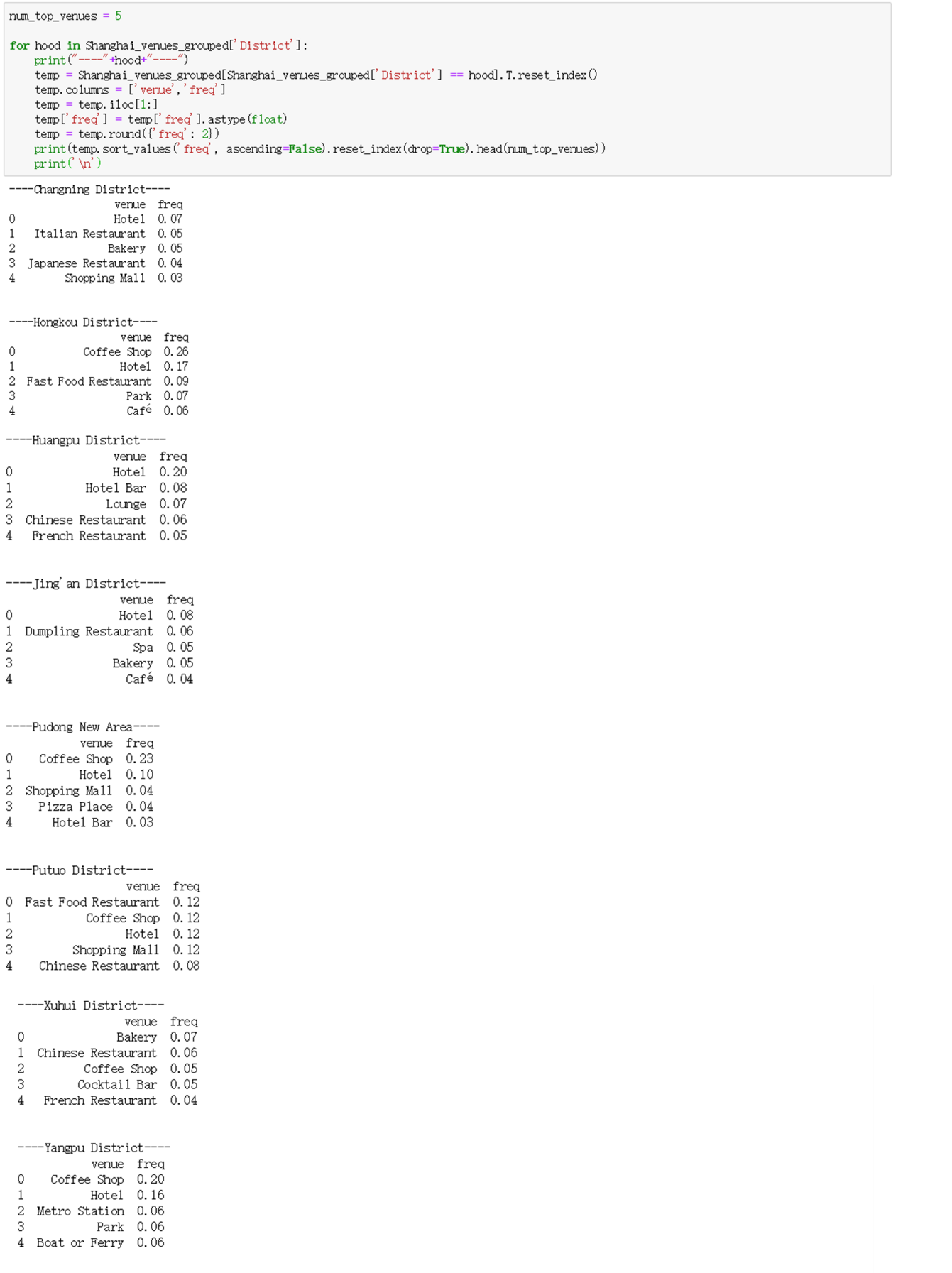
* There are 53 unique restaurant categories in the 8 districts
* In these unique restaurant categories, Chinese Restaurants are the most popular categories.
* Except Chinese Restaurants, some Chinese traditional restaurants such as dumpling and hotpot restaurants are also in the list



After have a roughly understanding of the venues, let’s get back to exploring the data a little more. To know about the top 5 venues of each district we proceed as follows

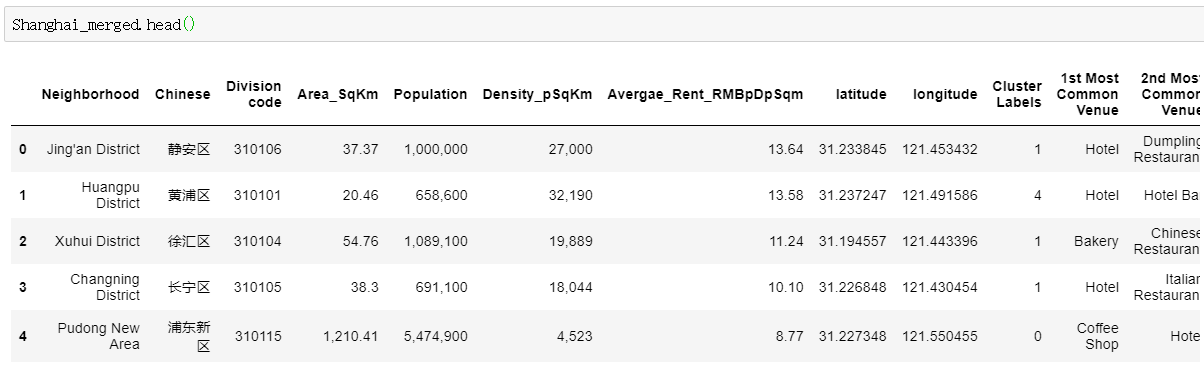
* Create a dataframe with pandas one hot encoding for the venue categories.
* Use pandas groupby on the District column and obtain the mean of the one-hot encoded venue categories.
* Transpose the data-frame at step 2 and arrange in descending order. Let’s see the code below

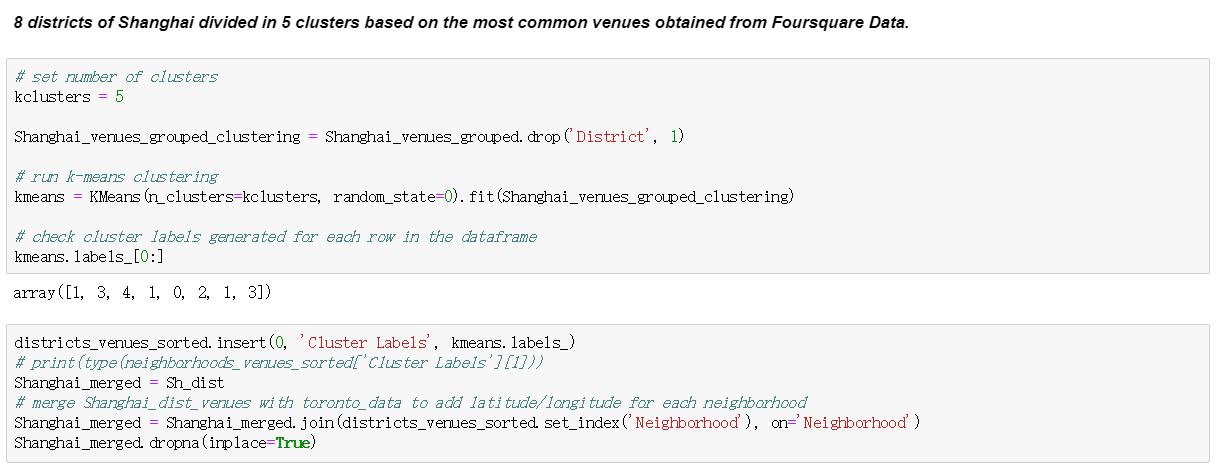
And the final results are shown as below



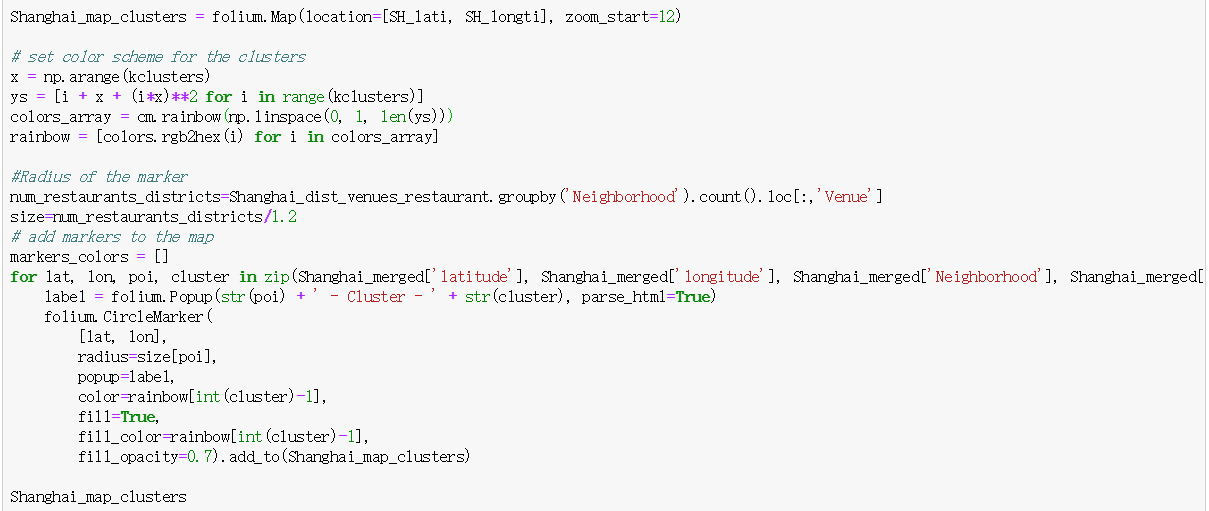
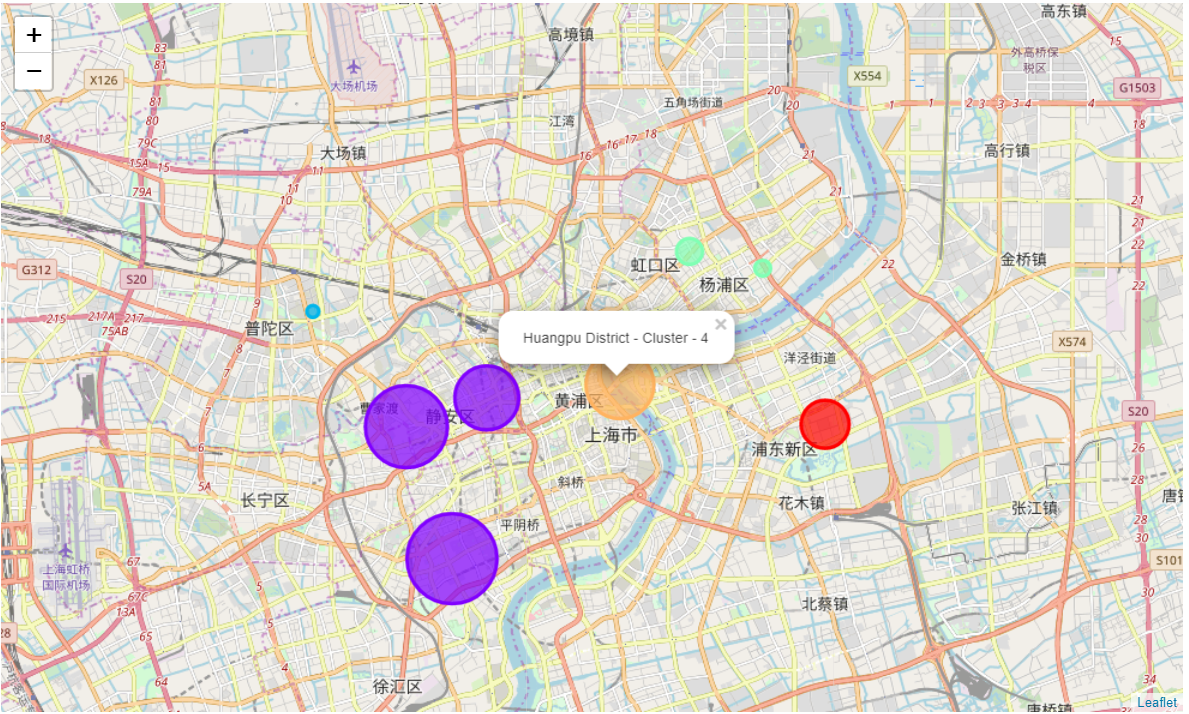
## 4. Clustering the Districts

Finally, we try to cluster these 8 districts based on the venue categories and use K-Means clustering. So our expectation would be based on the similarities of venue categories, these districts will be clustered. I have used the code below —





We can represent these 3 clusters in a leaflet map using Folium library as below.   
8 major districts of Shanghai segmented into 5 clusters based on the most common venues. The size of the circles represents number of restaurants as most common venues for each district, which is highest at Xuhui and lowest at Putuo as shown in map.



## 5. Results and Discussion:

I have used data from web resources like Wikipedia, python libraries like Baidu API, and Foursquare API, to set up a very realistic data-analysis scenario. We have found out that —

* The category of Chinese restaurants is the top of all restaurant categories in the 8 districts.
* Xuhui District is dominated by Venues related to food and drink such as Bakery, restaurants and bar as the the most common venue whereas the common venues of other districts are always come with hotel and restaurants .
* Xuhui District has maximum number of restaurants as the most common venue whereas has Putuo district has the least.
* Since the clustering was based only on the most common venues of each district, Xuhui, Jing'an and Changning are in the same cluster, this may be caused by the similar constitution of the venues with many food and drink venues. Huangpu district and Pudong New Area are separated into different clusters.Since the Huangpu District has similar number of resturants with Xuhui, Jing'an and Changning, the reason for it fall into a different cluster could be the more hotels in this area.
* As for the Pudong New Area district, I believe it is a better place to start a new restaurant, which will provide least competition for an upcoming restaurant since restaurants is not the most common venue in this area. In addition, Considering the land price and coffe store frequency, which is the top of the common venue, may indicate Pudong New Area could be a fast-pace or busist area.The frequency of restaurants as common venue are lower compared to the remaining busist districts.  
    
  Hence, this region could potentially be a target for starting quality restaurants, especially for fast food and Chinese restaurant, which may be really popular for the feature of the people there. Some drawbacks of this analysis are — the clustering is completely based on the most common venues obtained from Foursquare data. Since land price, distance of the venues from closest stations, number of potential customers, benefits and drawbacks of Shinagawa are not carefully taken here, could all play a major role and thus, this analysis is definitely far from being conclusory. However, it certainly gives us some very important preliminary information on possibilities of opening restaurants around the major districts of Shanghai.