**Data**

To solve this problem, I will need below data:

* List of neighbourhoods in Manhattan, United State.
* Latitude and Longitude of these neighbourhoods.
* Top 10 most common venues in each neighbourhood. This will help us find out the neighbourhood that is most suitable to open a Chinese restaurant.

**Extracting the data**

* Downloading geospatial data of neighbourhoods in New York from IBM developer skills network
* Getting Latitude and Longitude data of Manhattan city via Geocoder package
* Using Foursquare API to get venue data related to these neighbourhoods

**Methodology**

First, I need to get the list of neighbourhoods in New York, United State. This is possible by downloading the list of neighbourhoods from IBM developer skills network.

(“https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork\_data.json”)

Next, the json data is then transformed into a pandas data frame with columns including borough, neighbourhood, latitude, and longitude. The data frame is then filtered to retain data for neighbourhood located in Manhattan only. However, it does not contain the general geographical coordinate of Manhattan city which is needed to locate Manhattan city on the map. Therefore, geocoder package is employed to get the latitude and longitude of Manhattan city. With this geospatial data, folium is used to visualize the neighbourhood in Manhattan city.

Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues.

As for analyse each neighbourhood, one hot encoding is applied to the venue category so that each row of the data frame displays the existence of each venue category in a certain neighbourhood in binary form. This one-hot encoding is necessary to group the data by neighbourhood and calculate the mean of the frequency of each category. Then, the data frame is further reduced to contain only the top 10 most common venues for each neighbourhood.

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as far from each other as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighbourhoods in Manhattan into 5 clusters based on the top 10 most common venues for each neighbourhood. The clusters are then visualized on the Manhattan map. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.