**Introduction**

For this Capstone project, I am creating a hypothetical scenario whereby Chinese restaurateur wants to explore opening an authentic Chinese restaurant in Manhattan area. The idea behind this project is that there may not be enough Chinese restaurants that serve authentic food in Manhattan and it might present a great opportunity for this entrepreneur who is based in the United State. As Chinese food is very similar to other Asian cuisines, this entrepreneur is thinking of opening this restaurant in locations where Asian food is popular (aka many Asian restaurants in the neighborhood). With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I am designing this project to help him find the most suitable location.

**Business Problem**

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Chinese restaurant in Manhattan, United State. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Manhattan, if an entrepreneur wants to open a Chinese restaurant, where should they consider opening it?

**Target Audience**

The entrepreneur who wants to find the location to open authentic Chinese restaurant

**Data**

To solve this problem, I will need the following 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.

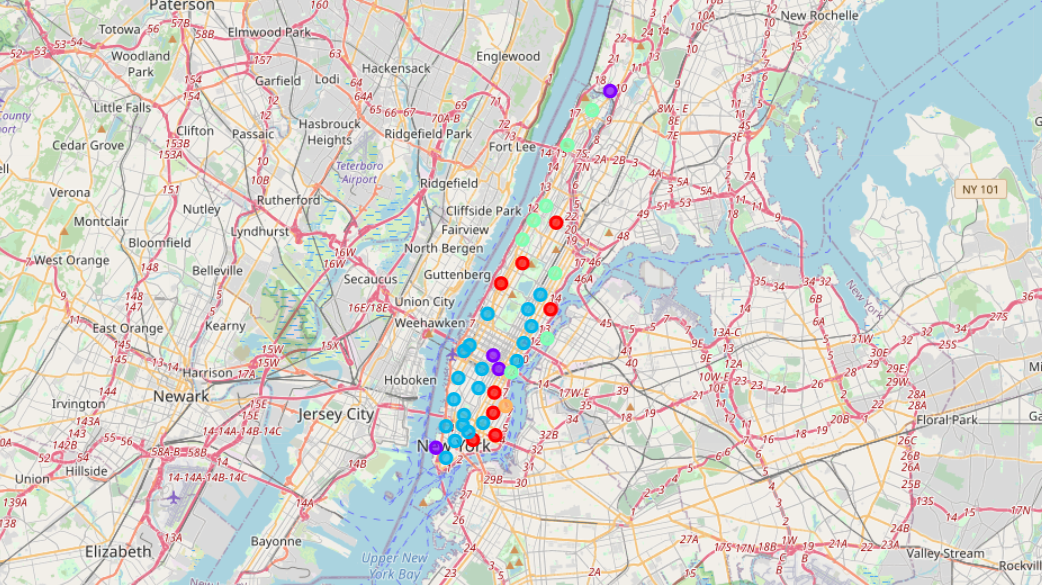
Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data,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

**Results**

Clusters



Cluster 1 (Red):



Cluster 2 (Purple):



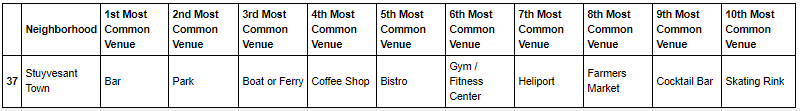
Cluster 3 (Blue):



Cluster 4 (Turquiose):



Cluster 5 (Brown):



The results from k-means clustering show that we can categorize Manhattan neighbourhoods into 5 clusters based on the top 10 most common neighbourhood:

* Cluster 1: Neighbourhoods with many restaurants especially Chinese restaurants
* Cluster 2: Neighbourhoods with few restaurants, no Chinese restaurants and few Asian restaurants
* Cluster 3: Neighbourhoods with many restaurants especially Western restaurants
* Cluster 4: Neighbourhoods with many restaurants and few Chinese restaurants
* Cluster 5: Neighbourhoods with no restaurants

The results are visualised on the map above with cluster 1 in red colour, cluster 2 in purple colour, cluster 3 in blue colour, cluster 4 in turquoise colour, and cluster 5 in brown colour.

**Recommendations**

Before proceeding with choosing a location for opening Chinese restaurant, it is better to have few selection criteria in mind which should include location having few or no Chinese restaurants, having not many restaurants, having at least few restaurants and the location should not have no restaurants at all. Examining cluster 1, we could immediately rule out the decision to build a Chinese restaurant in these neighbourhood because there are too many Chinese restaurants in the same area, leading to huge competition. It will be hard for an unknown Chinese restaurant to survive the first months of business before gaining enough popularity that ensures making of stable profit. Next, cluster 3 is not a suitable location to open Chinese restaurant because there are too many restaurants especially a western one which imply that the consumers in the neighbourhoods have higher tendency to opt for western food for their meals. Besides, cluster 5 should be ruled out as having only one neighbourhood in one cluster is not enough to generalise a cluster, not to mention that there are no restaurants over there, which could mean that there is no demand for a restaurant. Finally, there are only two cluster left to be chosen, which are cluster 3 that has few Chinese restaurants and cluster 2 that has few Asian restaurants. However, the market for restaurants in cluster 3 has saturated when there are too many restaurants and building more restaurants would most likely to cause bad return in investment. Therefore, it is highly recommended to open a restaurant in cluster 2 as there are only few restaurants. More importantly, the existence of Asian restaurants in these neighbourhood proves the potential market for Chinese restaurant as Chinese food is very similar to other Asian food and therefore should be able to build up popularity within a short time frame. There are few neighbourhoods to choose from, such as Marble Hill, Midtown, Murray Hill, and Battery Park City.

**Limitations and Suggestions for Future Research**

In this project, I only take into consideration of one factor: the top ten most common venues in each neighbourhood. There are many factors that can be taken into consideration. For instance, the location should be low in crime rate. High crime rates can make potential customers uncomfortable, and if they feel they’ll be mugged walking to their cars, it will only drive away business, no matter how legendary the restaurant is. Moreover, the rent of the place should be considered as well. Cost is always a bottom-line consideration for any business. If the rent or purchase of the space is more than you’ll bring in each month in profits, that location is not feasible at that time. Besides, parking spaces should be big enough to accommodate the number of customers based on the size of the restaurant. It is rare to have customers that are willing to spend hours looking for parking lots or even walk few hundred meters if they park their car somewhere far away from the restaurant. However, putting all these data into this project is not possible considering a short time frame for this capstone project. Future research can take into consideration of these factors.

**Conclusion**

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.