**Introduction**

The business problem that I’m trying to solve in this project is to find the best location to open a coffee shop in New York City. The target audience are small business owners who struggle to utilize data analytics to make the decision on this subject. There are many factors should be taken into consideration while choosing the best location to open a coffee shop. Such as the demographics of the city in each neighborhood, foot traffic in each area, competitors and neighboring business in the same area, etc.

Particularly, to decide the target market/customers is the crucial step in terms of picking the right location. Since there are only five boroughs in New York City, and Manhattan is the most densely populated borough in New York City with approximately 1.6 million residents and 25 Business Improvement Districts (BIDs), I select Manhattan as the target borough to select the best location. The target customers are students or employees who need to get work done before or after hours.

Different types of data analytic techniques are performed in the project to support the decision-making process, and the conclusion carry out useful insights to help small business owners to have a better understanding about their prospective customers and competitors in the target market.

**Data**

To provide the stakeholders the necessary information I will explore neighborhoods in Manhattan, analyze each neighborhood and cluster neighborhoods by downloading data from [here](https://geo.nyu.edu/catalog/nyu_2451_34572), and making API calls to Foursquare. Based on the clusters, I would know how the competitors located across the neighborhoods in Manhattan. Nest, I will examine the rankings of each neighborhoods based on indicators defined by [NYU Furman Center](furmancenter.org/neighborhoods) to find the best location to open the coffee shop.

**Methodology**

I performed exploratory analysis on examining the New York City data, particularly, the neighborhoods of Manhattan, to find the most common venues in each neighborhood through making API calls to Foursquare. I applied K-means machine learning techniques to cluster the neighborhoods and gain a better understanding about where the competitors located across the neighborhoods.

**1) Data Pre-processing**

Since NYC data is readily available on the internet, I downloaded and opened the json data in the Notebook, and then transform it to Pandas DataFrame by extracting the ‘Borough’, ‘Neighborhood’, ‘Latitudes’ and ‘Longitude’ from the original dataset. In this step, the dataframe returns 5 boroughs and 306 neighborhoods.

As discussed early, Manhattan is our target borough to find the best location, therefore, I created a new dataframe of the Manhattan data in the same format as the original dataframe. Before making API calls to Foursquare, I obtained the geographical coordinates of Manhattan by using the *Nominatim* function from *geopy.geocoders.*

**2) Exploratory Analysis**

I explored the first neighborhood from Manhattan dataframe, Marble Hill, and got top 100 venues that are in Marble Hill within a radius of 500 meters through making API calls to Foursquare.

1. **Data Cleaning**

Firstly, I defined a function to extract the category of the venue and then clean the json data and structured it into a Pandas dataframe. As a result, 25 venues were returned by Foursquare.

1. **Explore all neighborhoods in Manhattan**

I created a function to repeat the same process to all the neighborhoods in Manhattan, and then wrote the code to run this function on each neighborhood and create a new dataframe called *manhattan\_venues.* All venues were grouped by the neighborhoods and I found out 331 unique categories can be curated from all the returned venues.

1. **Analyze each neighborhood**

One-hot encoding was applied and I grouped rows by neighborhood and by taking the mean of frequency of occurrence of each category. 332 unique categories were grouped into 40 neighborhoods in Manhattan. Next, I printed each neighborhood along with the top 5 common venues and then put it into Pandas dataframe. Next, I ranked the top 10 most common venues for each neighborhood in order to cluster the neighborhoods in the following step.

1. **K-means Clustering neighborhoods**

K-means machine learning technique was applied to group the neighborhoods into 5 clusters. Merged the dataframe of the top 10 most common venues with *mahattan\_data* that includes the ‘Latitude’ and ‘Longitude’ information. The last step step is to print out the dataframe of each cluster to further observe the results.

1. **Examine the Development in each neighborhood**

One of the important measurement to showcase the development of the community district is the Certificate of Occupancy issued by the Department of Buildings. This will has positive impacts on the increasing of total number of residents living in each neighborhood across a certain time period, as well as for the local economy and the consumer base. In this project I downloaded the data from [NYU Furman Center](file:///C:\Users\sz185\Desktop\furmancenter.org\neighborhoods), and ranked the total number of certificates issued in 2019 in all of community districts in Manhattan.

**Result**

By observing among the 5 clusters, cluster 3 has the most number of neighborhoods being grouped together with 8 neighborhoods have coffee shops as the 1st most common venue and 7 neighborhoods have it as the 2nd most common venue. The competition remains a high level in cluster 3. In the cluster 1, Park appears to be the first or second most common venue in these 3 neighborhoods, and it seems like there are not financial districts, shopping malls or school districts close to these 3 neighborhoods. Foot traffic from students and employees doesn’t seem to be high in this area. In cluster 2, the neighborhoods composes of diversified cultures, particularly Asian culture, coffee consumption doesn’t seem likely to generate a high sales volume in this area, as people might tend to favor tea/bubble tea over coffee.

**Discussion**

We observed that Upper West Side has the most new certificate of occupancy issued by the Department of Buildings in 2019, it is also a community district grouped in cluster 3, which has the most competitors (coffee shops) spread out and most of the neighborhoods being grouped together. Putting all these facts together, Upper West Side is under the development by the BIDs and more people will rush into this area as the new constructions completed and new jobs created. Also noticed that coffee shop is not even in the top 10 most common venues in Upper West Side, this indicated the level of competitiveness is low and it would be a good idea to open a coffee shop in this area.

**Conclusion**

This report conclude that to open a coffee shop in Upper West Side of Manhattan would be a great choice in terms of the low competition and positive changes that would be brought by community development. The development and new constructions would bring the increasing number of local residents, thus increase the consumer base to enable the business to generate a high sale volume.