# **Coursera Capstone**

# **IBM Applied Data Science Capstone**

# **Capstone Project**

# **Finding optimal location for new restaurant in Yangon, Myanmar**

# **Author: Thandar Aung**

# **Date: 07 June 2020**

# **Introduction**

Yangon, Myanmar’s largest and former capital city, is by far the most exciting city in the country right now from former political exiles to current entertainment and artistic hub of the country. With the increase of foreign investment flocking in after the country has been reopened to the world, it is Yangon that reflects the most changes. High skyscraper condominium, hotels, and shopping malls are scattered in every parts of the city and certainly the increase in business of tourism sector.

# **Business Problem**

Along with the boom in tourism sector, this project is to identify the potential area to open restaurants to cater for business needs of increasing tourists. Being born in the country, I am keen to open a restaurant which will focus on the Burmese dishes (Myanmar local dishes) with the little fusion of European and Asian.

Taken into consideration of the price to open the restaurant, the purpose of this project is to find the optimal location in an area with ease of access for foreigners as well as local people. This project uses the unsupervised machine learning learnt from IBM data science courses combined with cluster validation methods to verify the output.

# **Target Audience of the project**

For this project, the targeted audience will be for those who are having similar idea to open the restaurant at the biggest cluster of tourist dense areas. In addition, this data will also be useful to check any top venues in the districts of Yangon, Myanmar.

# **Data**

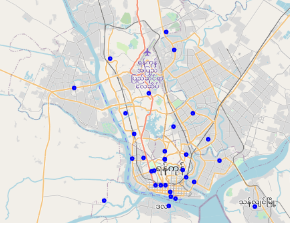
In order to proceed and process the analysis study, the dataset containing the districts and neighborhoods of Yangon are retrieved from Wikipedia. Then, with the help of geo-coder tool from the notebook, the geo-coordinates of the districts and neighborhoods are obtained. Thereafter, the top venues of the districts are queried using foursquare through an API.

# **Methodology**

The dataset was retrieved from the data source and tidied up to the format that needed for data analysis. After preparing the data format, the geo-coordinates of the districts and neighborhoods in Yangon are queried using the geo-code python library. In addition, the geocoder sometimes encounters error in retrieving the coordinates, therefore the looping of retrying the query until the coordinates have been fetched has been added. Thereafter, the new data frame was created to merge the districts, neighborhood together with the geo-coordinates data. After performing the merge, the following table in pandas data frame format is obtained.



Next, using the folium map, the visualization was made to see the scattered mark of districts and neighborhoods of Yangon on the map. For this case, the folium library was utilized.



*Folium Map showing all the neighborhoods in Yangon*

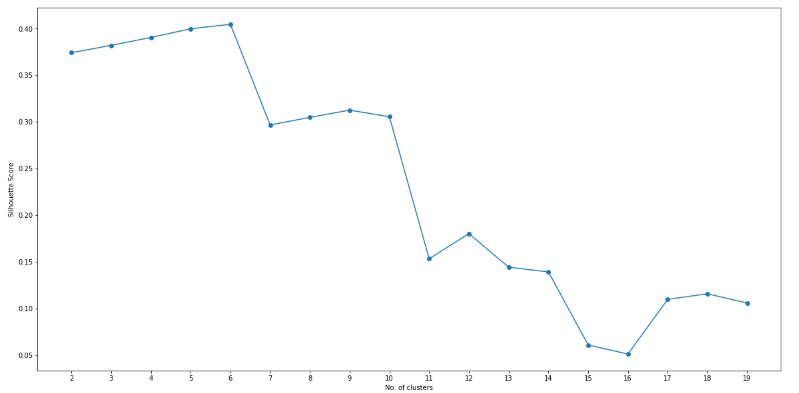
Foursquare API was applied to query the top venues of the neighborhoods using the retrieved latitude and longitude. Firstly, one of the particular neighborhoods was picked to check the effectiveness of the query by assessing the output.



Secondly, the general function to re-iterate for all the neighborhoods was built to retrieve of the top venues of individual neighborhood using the foursquare API. Thereafter, in order to analyze the district, one hot encoding method was applied for the venue category based on the data collected. Then, the new dataset is created by grouping the rows by neighborhood and by taking the mean of the frequency of occurrence of each category.

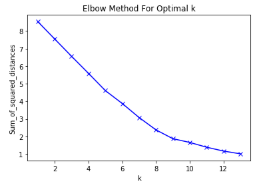
In this project, K-means clustering method was used to apply to the data set to cluster the neighborhoods. Beforehand, to determine the optimal cluster, two validation methods are applied to check the optimal cluster numbers.

The first method was to check the silhouette value, which is a measure of how similar an object to its own cluster compared to other clusters. The score value ranges from -1 to +1 where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The respective scores of the cluster ranging from 2 to 19 were collected by constructing a loop to run K-means clustering. From the diagram below, cluster 6 is the optimal number with the highest silhouette value.



*Diagram of Silhouette value output*

Another method was the Elbow method, which is a popular technique and the intent is to run k-means clustering for a defined range and for each value. Then the sum of squared distances from each point to its assigned center are calculated. The distortions are plotted and the plot looks like an arm then the “elbow” (the point of inflection on the curve) is the best value of k. Based on the output diagram, there is a point of inflection at cluster 6 , which identifies the same cluster number as previous method.

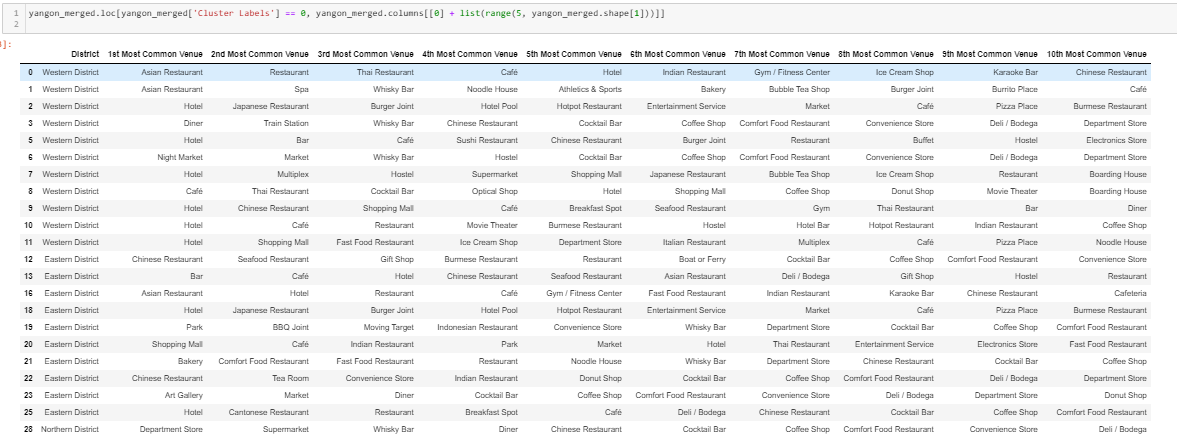


*Diagram of Elbow method’s output*

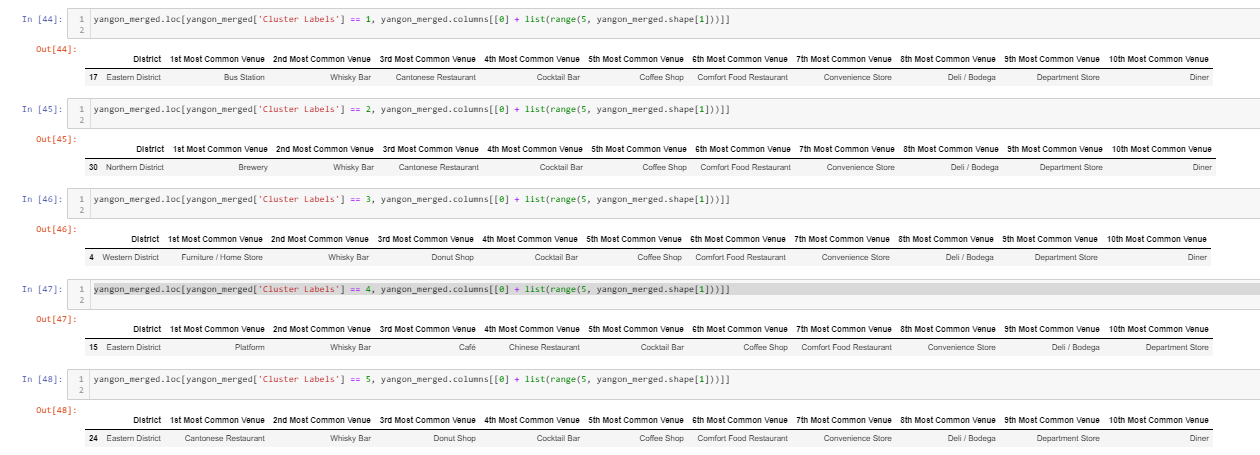
# **Results**

Based on the K-means clustering’s output, in order to decide which cluster is better for opening the restaurant, all the detailed data of the individual cluster were retrieved to validate.

1. Cluster 0



1. Cluster 1/2/3/4/5



# **Discussion and limitations**

Based on the output clusters, the result can be decided that cluster 0 is a potential place to consider for opening the restaurant. The cluster is basically near to the hotels and town center which satisfy the ease of access for tourists.

However, there are some limitations of this dataset. Firstly, the latitude and longitude that retrieved for districts of Yangon are limited and some of them need to be manually updated in order to be able to link with Foursquare search. Secondly, the collection of the venues using foursquare may have data limitations as foursquare app is not popularly used in the country.

# **Conclusion**

In summary, the business problem was analyzed thorough Python using popular libraries such as Pandas, Scikit and Folium. Then using the geocoder to retrieve geographical locations to be utilized in foursquare venue search. Machine learning comes into play after all the data preprocessing was done, in this case, K-means clustering method. With the help of clustering validation method, which are the silhouette value method and elbow method, the optimal cluster number is identified and then the data output is finalized. Finally, the output of the data analysis provides the potential location of a new restaurant with all the skills and tools learnt from the IBM Data Science Course.

# **Reference**

<https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_of_Yangon>