Coursera Capstone Project

IBM Applied Data Science Final Capstone

New Indian Restaurant in Kuala Lumpur, Malaysia

Vishal Gawai
October 2019

Introduction:

Kuala Lumpur is the capital city of Malaysia. It is a great destination for tourists from all over the world due to vivid attractions the city provides. Also a large number of Indian tourists has started visiting Malaysia and Kuala Lumpur is one place that no one plans to miss. With this growing tourists, there arises a need to provide great quality and variety food for them. To satisfy the quest for food, Indian restaurants are a great opportunity not only for Indian tourists but also for tourists from all over the world who wants to try something new. Keeping this in mind, I will be focusing on identifying best locations to open new Indian restaurants in Kuala Lumpur, Malaysia.

Business Problem:

The goal of the Capstone project is to analyze neighborhoods in Kuala Lumpur, Malaysia and identify the best locations to open a new Indian restaurant. We will be making use of various Data Science strategies and Machine Learning tools that has been learned during the entire IBM Applied Data Science Course.

Problem: Which are the best neighborhoods in Kuala Lumpur to open new Indian Restaurants in Kuala Lumpur, Malaysia for new stakeholder who wants to open a new restaurant.

Data Gathering:

Data required

- Get the list of neighborhoods in Kuala Lumpur.
- Coordinates of neighborhoods in Kuala Lumpur. These are fetch the list of venues and plot them on map.
- Information on venue details will be required to perform clustering.

Data Source

- Wikipedia is used to fetch the list of neighborhoods in Kuala Lumpur, Malaysia. **Beautiful Soup** will be used to fetch the list as it is highly effective in web scrapping of Html data.
- **Geocode package** provided by python is used to fetch the coordinate details of the neighborhoods. The Latitude and Longitude information will be further useful to gather information about the various venues.
- **Foursquare API** will be used to gather information about various venues in the neighborhoods of the Kuala Lumpur. The API provides great and detail information about the venue such as location, category, details, tips, comments, photos etc.
- Machine Learning Algorithm **K-Means** will be used to perform clustering of neighborhoods. The venues are grouped by categories and take mean frequency of occurrences of venue categories.
- **Folium** package provided by Python is used to plot the venues on map to provide great visualization for better view of venues.

Methodology:

- Gather the list of Neighborhoods from Wikipedia using python package Beautiful Soup. This package is very handy and useful to perform web scrapping on HTML data and fetch information from it.

- Geocode package is helpful to get the coordinate details (Latitude and Longitude positions) of Neighborhoods of Kuala Lumpur, Malaysia.

- The Foursquare API is used to gather various details about the venues present in the different neighborhoods of Kuala Lumpur using the coordinates fetched previously.

- Then using the Pandas, Numpy and other python packages, we can perform different operations on the data to get the venue data in our required format to perform further analysis on it.

- Using Machine Learning algorithm k-Means we further cluster the neighborhood by grouping the venue categories on the basis of mean frequency of occurrence of each of them.

- Finally, we plot the result using Folium library which provides better visualization to analyze the data.

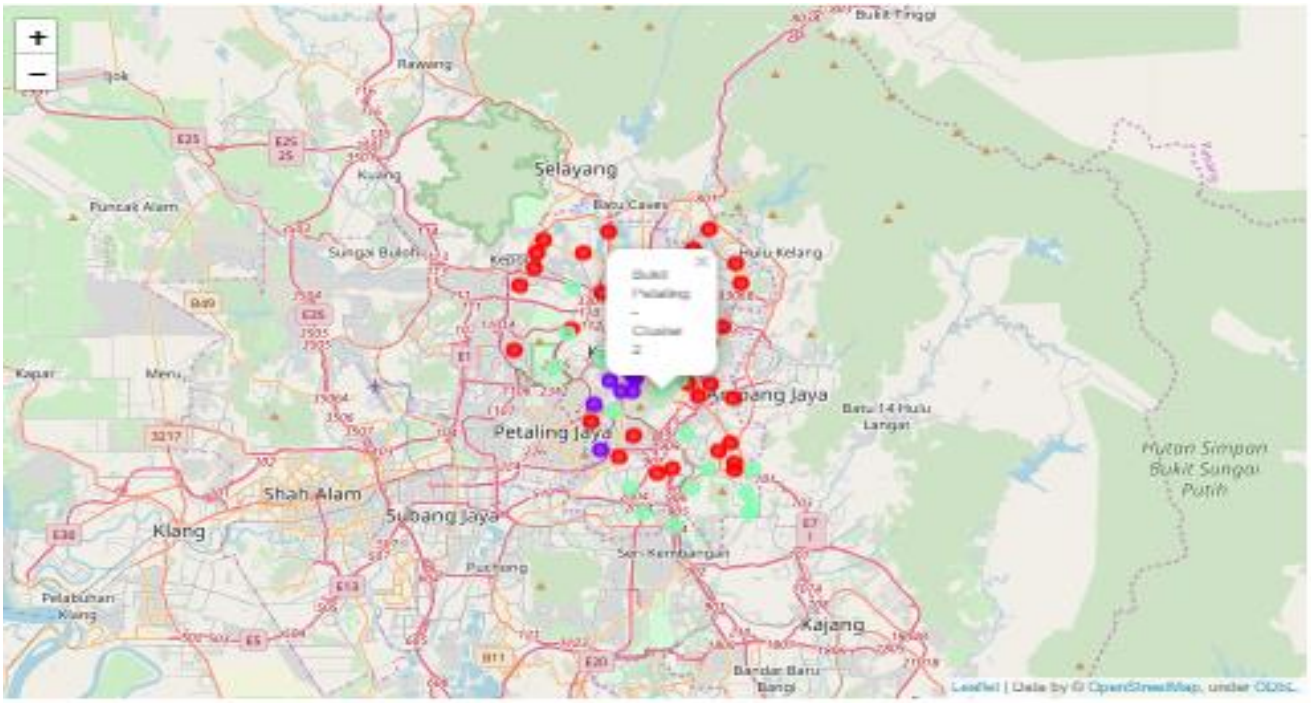
Results:

We have categorized the neighborhoods in 3 clusters based on the frequency of occurrence for **“Indian Restaurants”.** We have performed **k-means** clustering to derive these results:

- Cluster 0: Neighborhoods with least count of Indian Restaurants.
- Cluster 1: Neighborhoods with very high concentration of Indian Restaurants.
- Cluster 2: Neighborhoods with moderate number of Indian Restaurants.

The results are visualized below in the map below with

Cluster 0 – Red, Cluster 1 – Purple, Cluster 2 – Light Green colors.



Conclusion:

In the project, we have performed our analysis to identify the best locations to open an Indian restaurant in Kuala Lumpur. As per our findings, neighborhoods in cluster 0 is the least concentrated with Indian Restaurants. It thus provides the best opportunity to open a new Indian restaurant in neighborhoods of cluster 0.

Neighborhoods in Cluster 1 has moderate number of Indian restaurants and so provide adequate opportunity for the restaurant to flourish and grow.

Cluster 2 has the highest count of Indian Restaurants and should be the least choice to open Indian Restaurant for new stakeholders.

References:

- Wikipedia
- Foursquare
- Python Library documentation.