Opening an New Restaurant in Kuala Lumpur, Malaysia

DATA SCIENCE CAPSTONE PROJECT

BY: VISHAL ANAND RAO

Introduction

- Malaysia is a hub of tourist places with an annual footfall of about 26,100,800 in 2019.
- Whereas its capital city Kuala Lumpur welcomes around 11.4 million international tourists annually according to Mastercard.
- The arrival of large number of tourists in the capital city booms the local economy and promotes the development of local businesses like restaurants, cafe's, shopping malls etc in the city and its neighborhood places.
- The businessmen and investors from around the world are showing their interest in setting up new businesses in the city to get the handsome returns in future. The businessmen are searching for some best locations around the city to setup new restaurants to cater the ever growing demand of tourists and locals. They are searching for locations with minimal competition to avoid the risk of failure of their business.
- As squares already crowded with number of restaurants offers less scope for success of new restaurant.

Business Problem

- The objective of this Capstone Project is to analyze the data using Data Science methodology and Machine Learning techniques like clustering to provide solutions to answer the business problem: If a businessmen or investor is looking for some best locations with less competition to open a new restaurant in city Kuala Lumpur or its neighborhood, which places you will be recommending?

Target Audience

- Businessmen
- Investors
- Analysts

Data Acquitition

To solve the problem, we will need the following data:

- List of neighborhoods in Kuala Lumpur. This defines the scope of this project which is confined to the city of Kuala Lumpur, the capital city of the country of Malaysia in South East Asia.

- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

- Venue data, particularly data related to restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of Data

- Wikipedia Page - - (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur)
- Foursquare API - One of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the ‘Restaurant’ category in order to help us to solve the business problem put forward.

Methodology

- Firstly, get the list of neighborhoods in the city of Kuala Lumpur. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur)
- Do web scraping using Python ‘requests’ and ‘beautifulsoup’ packages to extract the list of neighborhood names data.
- To get the latitude and longitude coordinates, use the wonderful ‘Geocoder’ package that will allow us to convert address into geographical coordinates in the form of latitude and longitude.

Contd….

- Now populate the data into a ‘pandas’ DataFrame and then visualize the neighborhoods in a map using ‘Folium’ package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by ‘Geocoder’ are correctly plotted in the city of Kuala Lumpur.
- Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters by passing make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop.

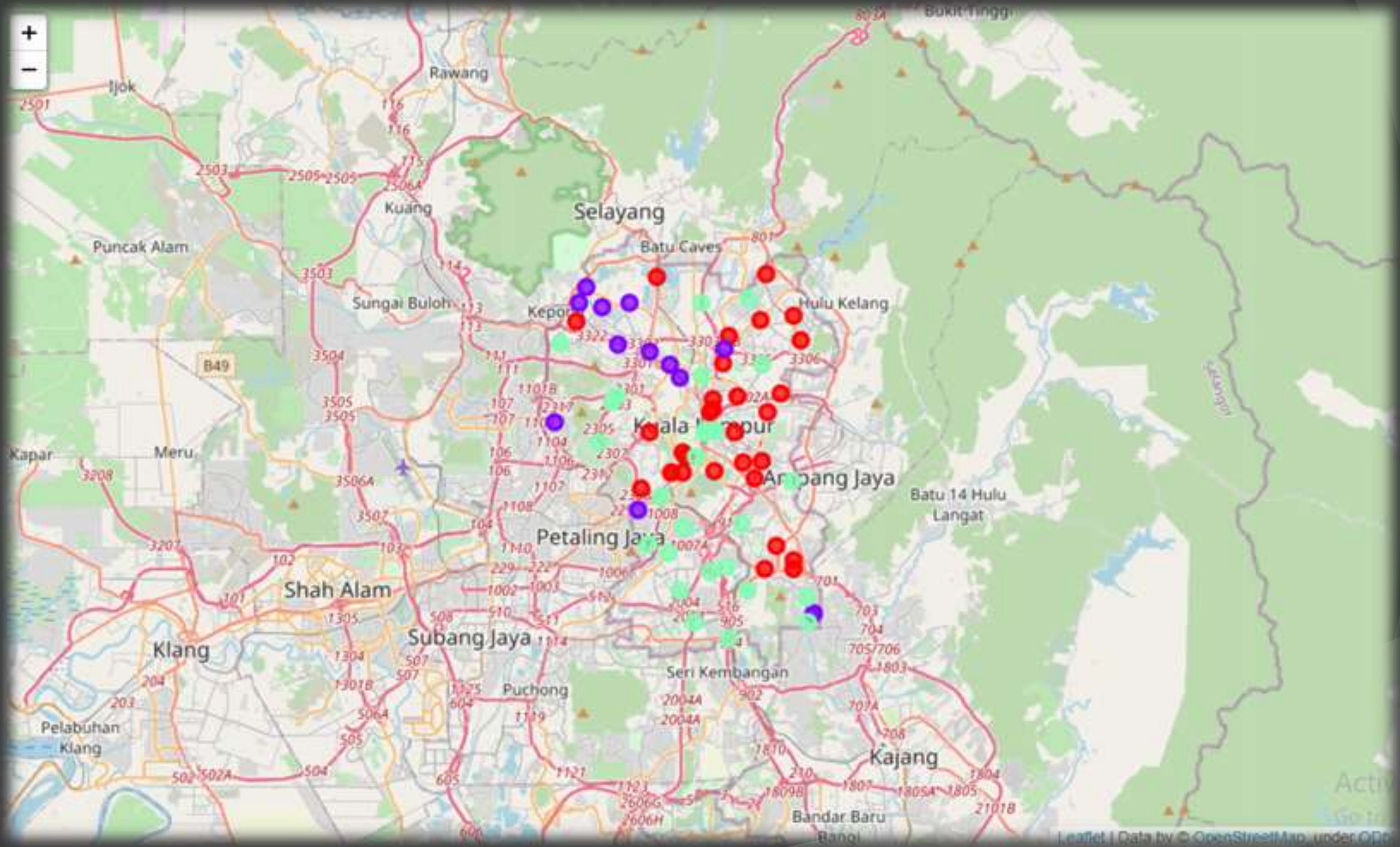
- Extract the venue name, venue category, venue latitude and longitude from the JSON file returned by Foursquare.
- With the data examine the unique categories that can be curated from all the returned venues. Then analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.
- Now filter the “Restaurants” as venue category for the neighborhoods.
- Perform clustering on the data by using k-means clustering

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Restaurant”:

- Cluster 0: Neighborhoods with less number of restaurants
- Cluster 1: Neighborhoods with high concentration of restaurants
- Cluster 2: Neighborhoods with moderate of restaurants
- The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

- Cluster 0 : Represents a great opportunity and high potential areas to open new restaurants as there is very little to no competition from existing restaurants
- Cluster 1 : Restaurants are likely suffering from intense competition due to oversupply and high concentration of restaurants.
- Cluster 2 : Restaurants with unique selling strategy can also opt for moderate competition

Thank you