Project Report

Opening a new café in Kuala Lumpur

IBM Applied Data Science Capstone

By: Ahmad Syahmi Adnan

7th August 2020

Introduction & Background Study

Coffee is the most popular beverage in the world, with more than 400 billion cups consumed each year. More than 450 million cups of coffee are consumed in the United States every day. It's the world's 2nd largest traded commodity. Coffee is consumed in great quantities, making it the most beloved beverage after water. It's worth is over $100 billion worldwide.

A café is a type of restaurant which typically serves coffee and tea, in addition to light refreshments such as baked goods or snacks. The term "café" comes from the French word meaning "coffee". A café setting is known as a casual social environment where you can find people reading newspapers and magazines, playing board games, studying or chatting with others about current events. It is known also regarded as a place where information can be exchanged. I am a coffee lover and have a plan to open my own café or coffee chop business in Kuala Lumpur. So the purpose of this project is to identify the best neighbourhood to open the coffee shop since there are tonnes of café in Kuala Lumpur.

Problem Statement

The objective of this capstone project is to analyse and select the best locations in the city of Kuala Lumpur, Malaysia to open a new café. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: Which area/neighbourhood in Kuala Lumpur is the best to open a new café?

The Location

Kuala Lumpur is the capital city of Malaysia. It is the largest city in Malaysia, covering an area of 243 km^2 (94 sq mi) with an estimated population of 1.96 million. Kuala Lumpur is the cultural, financial and economic centre of Malaysia and it is among the fastest growing metropolitan regions in Southeast Asia, in both population and economic development.

Data Collection & Data Source

Data that are needed for this project is:

- List of neighbourhood in Kuala Lumpur.
  - Web scrapping from (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur)
- Venue data in each neighbourhood to know what kind of venue by category in each area and to extract venue related to café and coffee house.
  - From Foursquare API
- Coordinate of neighbourhoods in Kuala Lumpur to get the venue data from Foursquare API and to plot it on the map.
  - Using Geocoder

Libraries

Libraries Which are used to Develop the Project:

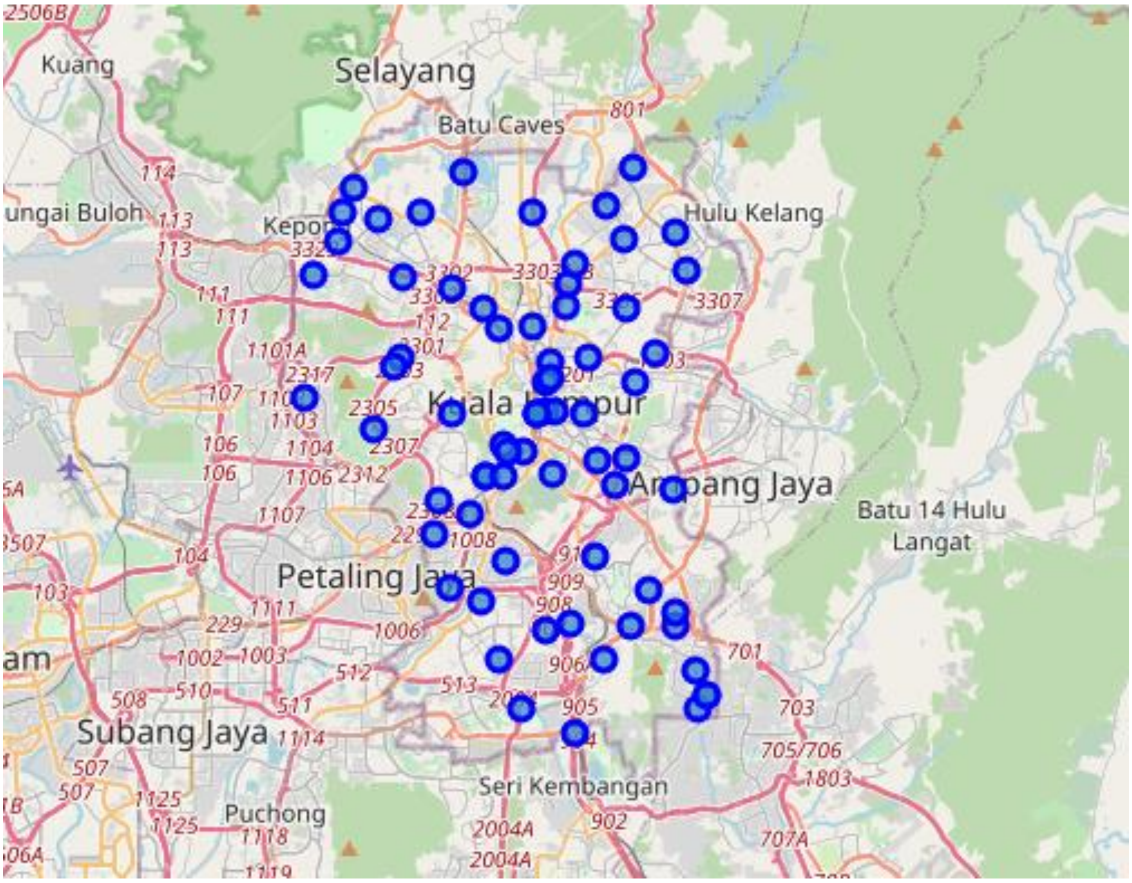
- Pandas: For creating and manipulating dataframes.
- Folium: Python visualization library would be used to visualize the neighbourhoods cluster distribution of using interactive leaflet map.
- Scikit Learn: For importing k-means clustering.
- JSON: Library to handle JSON files.
- XML: To separate data from presentation and XML stores data in plain text format.
- Geocoder: To retrieve Location Data.
- Beautiful Soup and Requests: To scrap and library to handle http requests.
- Matplotlib: Python Plotting Module.

Methodology

- Get the list of neighbourhoods in Kuala Lumpur by using Python requests and beautifulsoup packages to extract the list of neighbourhoods data from en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur.
- Use the Geocoder package to convert address into geographical coordinates in the form of latitude and longitude.
- Mapping each neighbourhoods on map to make sure all the locations are correct by using Folium package.
- Use Foursquare API to extract top 100 venues from each neighbourhoods in radius of 2 km.
- Analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category.
- Create new dataframe to determine the number of café in each neighbourhood.
- Run K-means to cluster the neighborhoods in Kuala Lumpur into 3 clusters.
- Analyse and visualise each cluster via Autoplotter and Geocoder packages.

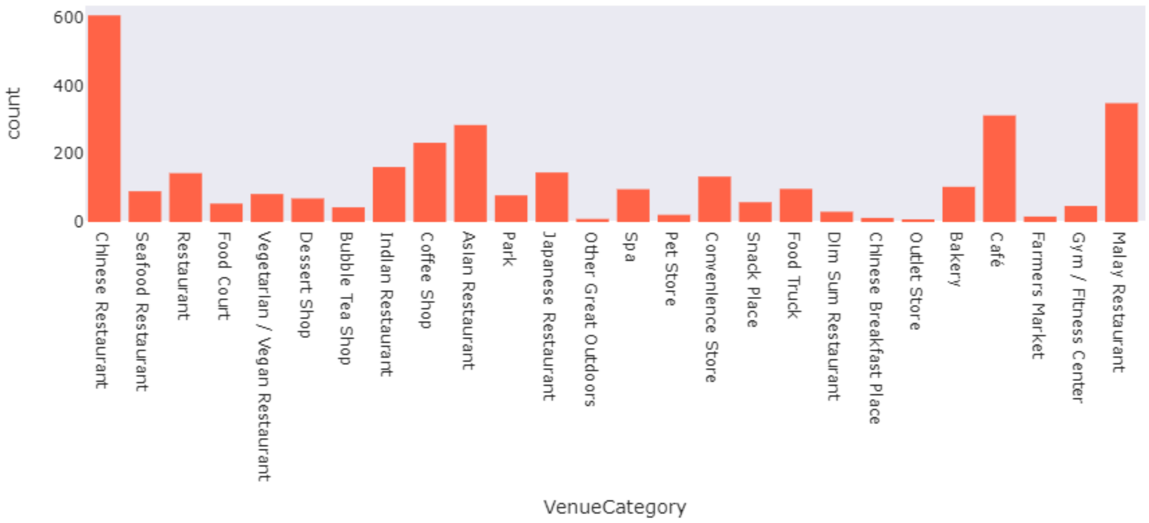
Results & Discussion

The web scrapping from en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur gives 71 neighbourhood in Kuala Lumpur. The Geocoder finds the coordinate based on the name of the neighbourhood. This is the result of the map of all neighbourhoods in Kuala Lumpur produced by Folium package:

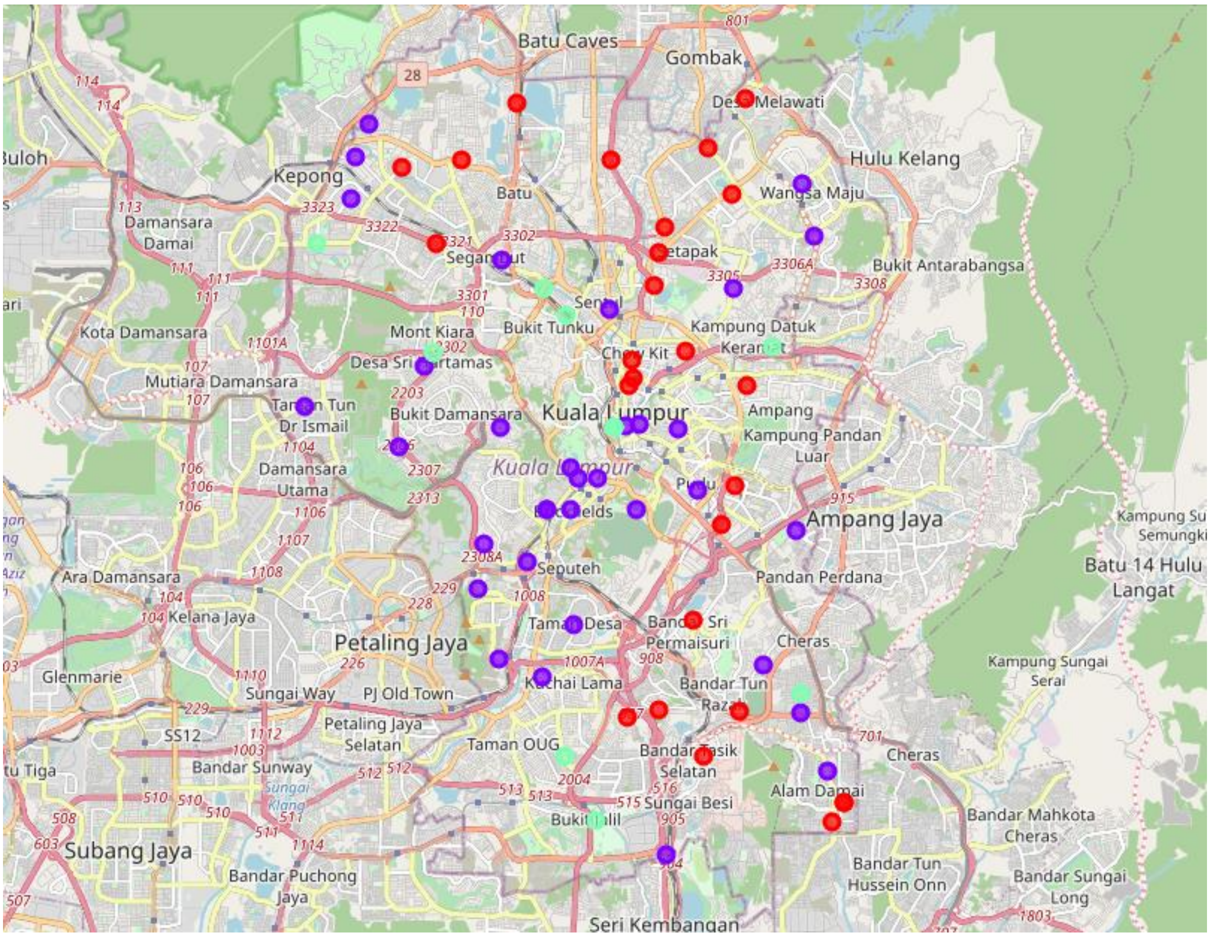


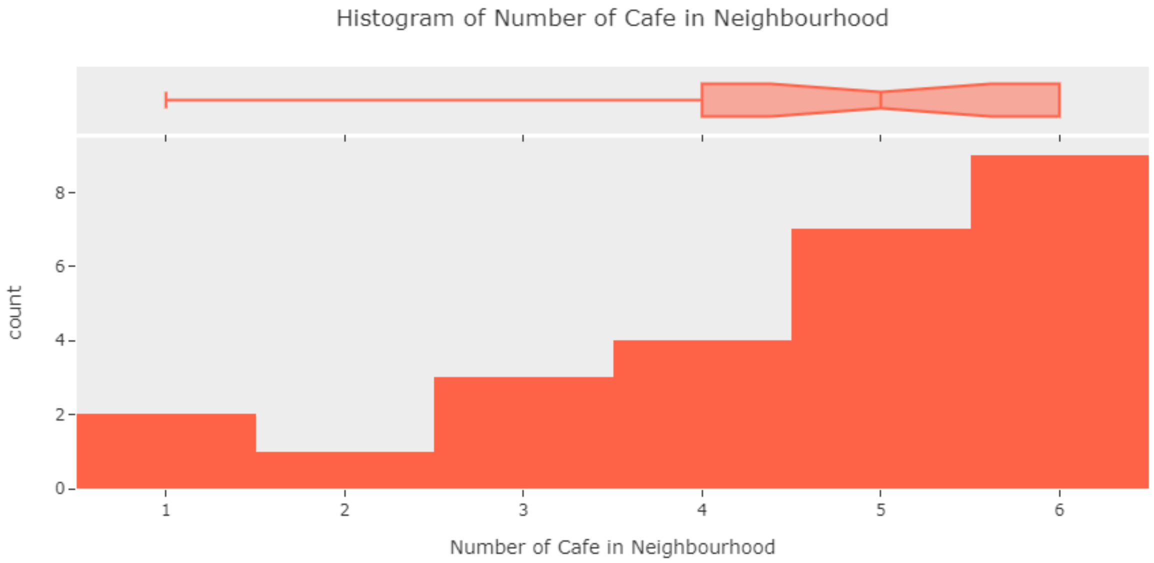
Based on the map above, its confirmed that all the neighbourhood are correctly plotted in the area of Kuala Lumpur.

After extracting top 100 venues from each neighbourhood, total of 309 categories are identified for all 7100 venues. Below shows some of the category with some The top 3 categories number of venues by category based on the graph below are Chinese Restaurant (611 venues), Malay Restaurant (354 venues) and Café (316 venues)

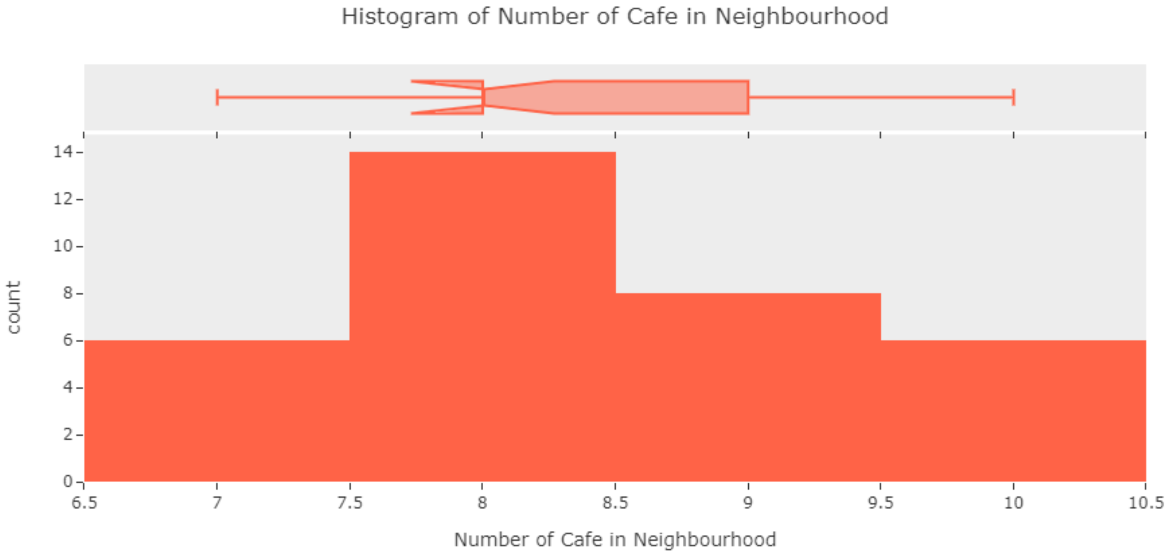


The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Café": The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour. Explain each cluster like in notebook with graph of each cluster

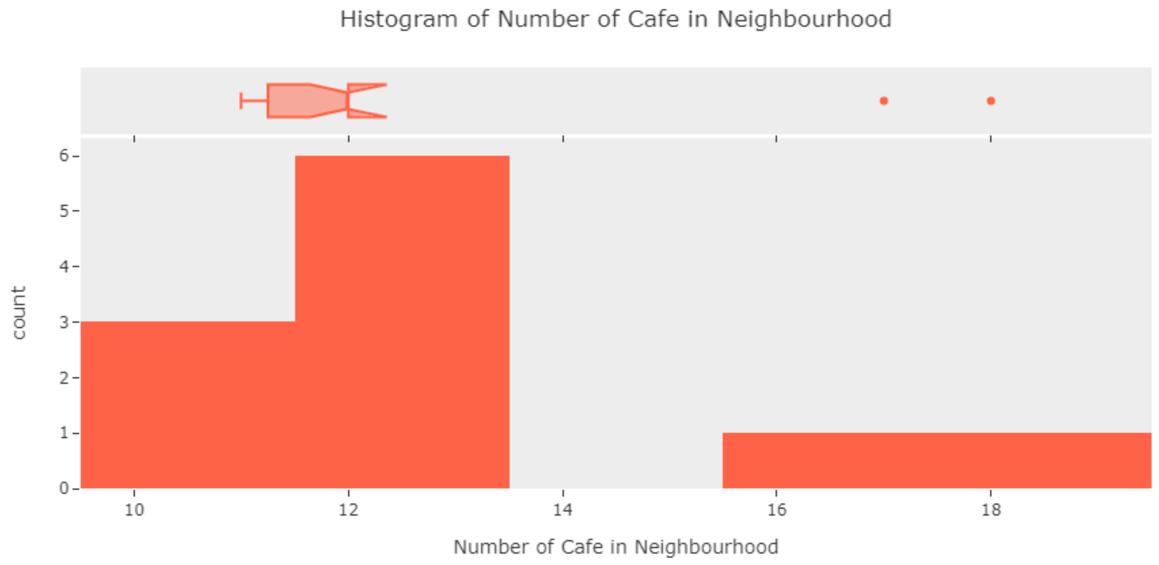




Cluster 0 shows the low density of coffee shops which are *from 1 to 6 cafés* in each neighbourhood. This cluster consists of 22 neighbourhoods out of 70 neighbourhoods in Kuala Lumpur and most of the the neighbourhoods in this cluster located at northern and southeaster part of Kuala Lumpur. This area has high potential for opening of a new cafe since there is less competitor compared to neighbourhoods in cluster 1 and 2.



Cluster 1 shows the second highest density of cafés in the neighbourhoods which consist of *6 to 9 cafés*. This is the biggest cluster with total of 33 neighbourhoods and consist of almost half of all neighbourhoods in Kuala Lumpur (47%). Most of the neighbourhoods in this cluster are located at the central area of Kuala Lumpur which also consists a lot of commercial buildings and offices. Maybe the target market of cafés in that area is the worker from those commercial buildings.



Cluster 2 has the highest count of café in each neighbourhood which are *more than 10 cafés*. This cluster consists of 12 neigbourhoods out of 70 neighbourhoods which equivalet of only 17% of neighbourhoods in Kuala Lumpur. From the Kuala Lumpur map, we can see that its located in area of Mont Kiara, Bukit Jalil, Bukit Tunku and Damansara which are categories as high end neighbourhood in Kuala Lumpur.

Limitation & future research

In this project, we only consider one factor i.e. frequency of occurrence of cafés, there are other factors such as land title, population and income of residents that could influence the location decision of a new café. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new café. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

As observations noted from the map in the Results section, most of the cafés are concentrated in the high end area of Kuala Lumpur, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number of cafés in the neighbourhoods. This represents a great opportunity and high potential areas to open new café as there is very little to no competition from existing cafés. Meanwhile, cafés in cluster 2 are likely suffering from intense competition high concentration of coffee shop in that area. From another perspective, the results also show that the oversupply of cafés mostly happened in the central area of the city, with the suburb area still have very cafés. Therefore, this project recommends investor to capitalize on these findings to open new café in neighbourhoods in cluster 0 with little competition. Investor with unique selling propositions to stand out from the competition can also open new cafés in neighbourhoods in cluster 1 with moderate competition and also nearby to commercial buidings and office in Kuala Lumpur which is a good market for café. Lastly, investors are advised to avoid neighbourhoods in cluster 2 which already have high concentration of cafés and suffering from intense competition.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new café. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new café.