# Coursera Capstone

IBM Applied Data Science Course

Building A Park in Kuala Lumpur, Malaysia

By: Nizam Asmari

October 2019

## Introduction

As countries aiming to reach a developed country, lot of projects have been made in order to meet the requirement. Developing the residential area, infrastructures as well as the improving of towns and cities management were made mainly in big cities. Therefore, in recent years, we can see a lot of concrete jungles in big cities like in Kuala Lumpur. As a result, it helps increasing the carbon emission rate. Furthermore, it also leads to a physical and mental health problem to cities resident, as there are not many places to socialize or to getaway from the crowded and hectic life. Building a park is a way to solve this matter and it also line with Malaysia government who is currently study on low-carbon footprint in an effort to create more green spaces and linkages in Kuala Lumpur. However, it is very difficult to find a suitable place to build it because it requires serious consideration and a lot more complicated than it seems. Particularly, the location of the park is one of the important decisions that will determine its function towards the environment and citizens.

## Business Problem

The objective of this capstone project is to analyze and explore the best and suitable place to build a park in Kuala Lumpur. Using the data science methodology and machine learning techniques such as clustering, in order to solve the business question; If your government, developers or any related NGOs what to build a park in Kuala Lumpur metropolitan city, where will you recommend it?

## Target Audience of the project

This project is very useful for the environmentalist NGOs, developers and government who want to solve the environment problem in Kuala Lumpur. This project indeed is taking time, as Kuala Lumpur is crowded and bustled with concretes and bricks. Thus, the target audience of this project includes developers, NGOs, constructors and anyone who deals with environment.

## Data

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

1. The list of Kuala Lumpur neigbourhood. This define the project which is confined the city of Kuala Lumpur, the capital city of Malaysia.

2. The longitude and latitude of Kuala Lumpur. This is necessary in order to plot the map and to get the venues data.

3. Venues data, particularly the one who related for building a park. We will use this data to perform clustering on the neighbourhood.

## Sources of data and methods to extract them

( https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur )

Refer the above Wikipedia link, it is the list of Kuala Lumpur neighborhood. We will be using the link and scrap the webpage by using BeautifulSoup Package and Request Package in order to obtain the necessary data for this project. Then, using the Python Geocoder Package in order to get the longitudes and latitudes of each neighbourhood data.

After that, we will use Foursquare Application Programming Interface (API) to get the venues data of each neighbourhood. Foursquare has one of the largest database, with 105 millions places and 150 thousands developers. Foursquare will provide many categories on each of the venues data; particularly, we are interested in park, shopping mall and residential area in order to find the best solution in this matter. This project will make use of many data science skills; from web scrapping, API, data wrangling, data develop and data evaluation to machine learning and map visualization. In the next section, we will discuss about the methodology of the project.

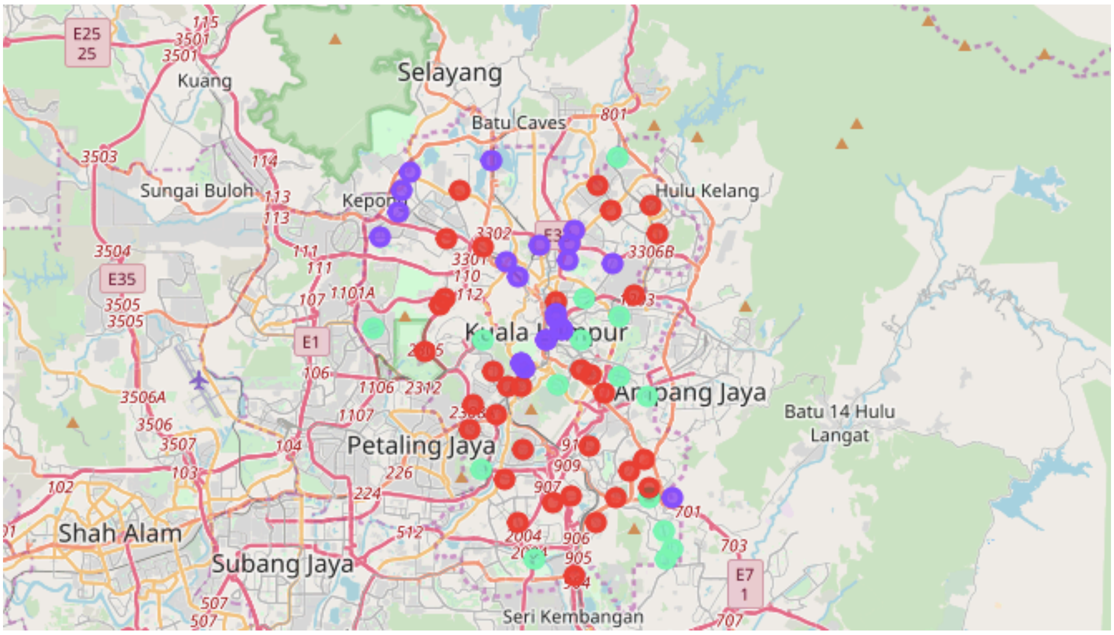
## Methodology

As our first step, we will search the neighbourhoods of Kuala Lumpur by using the wikipedia page link: ( https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur ). The next step is scrapping the webpage by using python library Request and BeautifulSoup to get the list of the neighbourhoods. In order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighbourhood. To do so, we will use the Arcgis Geocoding Application Programming Interface (API) which it will detect the list of neighbourhood to get their geographical coordinates in the form of longitude and latitude. After done, create a dataframe from the data using Pandas and plot a map using Folium Package. This allows us to check the geographical coordinates data returned by Geocoder is correctly plotted on the map of Kuala Lumpur.

Next, we will use Foursquare API to explore the neighbourhood of Kuala Lumpur. First, we set it to the radius of 2000 metres and limit it to 100 list number of venues. In order to use Foursquare API, we need to register Foursquare Developer Account to get the Foursquare ID and Foursquare secret key. Then, we make API calls to Foursquare passing the geographical coordinates of the neighbourhood using python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. By referring to the data, we can check how many venues were returned for each neighbourhood and can examine how many unique categories can be curated from all the returned venues. Then, we will analyse all the neighbourhood by grouping the rows of the same neighbourhood by using the mean value of each venue categories. Since, we are analysing the “Park” data, we will filter the “park category of the neighbourhood to be use in our clustering method.

As for our last step, we will perform the clustering on the data by using k-mean clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms and is suitable for our project. In this project, we will cluster the data into 3 cluster based on the frequency of the value of “Park” category in each neighbourhood. The result will allows us to identify the availability of “Park” in each neighbourhood, which helps us in finding the solution of the problem.

## Results



Refer to the picture map above, there are many dots on the map of Kuala Lumpur. The dots have been divided into 3 colours; Red, Purple and Aqua.

1. Red is referring to cluster 0 that shows the neighbourhood that do not have park.
2. Purple shows cluster 1 that have low number of parks.
3. Aqua indicates the cluster 2 which have a high number of parks.

## Discussions

As observations, most of the park is located in the Center and West of Kuala Lumpur, with the highest number of park in cluster 2. While, the low number of park is located mainly in the East of Kuala Lumpur which indicated in cluster 1. From the result, we can see that the low number of park that showed in cluster 0 is 33 out of 70 listed of neighbourhood that is more than the others. This mean that there are stills many neighbourhoods that do not have park within the range of 2000 metres. Here, we can conclude that the suitable place to build a park is the area in cluster 0 and cluster 1 as there are still lack number of park. However, it would be great if we could focus on building the park somewhere in the south area of Kuala Lumpur such as Salak South, Desa Petaling and Bandar Tun Razak. This is because the area has very low number of park compare to other area. From another perspective, this finding also help the cities resident from focus on certain park which can lead to traffic jam and crowd problem. Thus, help government in solving the physical and mental health of citizen as well as reducing the carbon emission rate by making Kuala Lumpur greener. Lastly, government and developer are advised not to build a park in cluster 2 or any area near to it, as there are already sufficient park and avoiding citizen from focus on certain area.

## Limitation and Suggestions for Future Research

The core in this project is to find the suitable location of building a park in Kuala Lumpur. There are lots of factors that give affect on finding a location, which are the price, access and the space of the land. These factors are not within our analyzation as the data that we receive from Foursquare did not cover these types of data. In order to obtain the complete data, we may need to deal and ask permission from the government which we are unable to do that in this project. In the future, we hope that we are able to provide these data to improve this project.

## Conclusion

This project aims to help government, NGOs and developers to find a suitable place to build a park in Kuala Lumpur. In order to extract the data, we are using python libraries, arcgis API and Foursquare API. Then, we analyse it by using the clustering method and plot a map using Folium package. From the result, we can conclude that the south of Kuala Lumpur is the most suitable place to build a park as it have low number of it. Lastly, we hope that we can provide not only the availability of park in certain location but other factors as well; price, access and size of the land which can help our project in the future.