Data Science

Capstone Project

THE BATTLE OF NEIGHBOURHOODS

FINAL

Introduction- Business opportunity in Selangor

Selangor also known by its Arabic honorific Darul Ehsan, or "Abode of Sincerity", is one of the 13 states of Malaysia. It is on the west coast of Peninsular Malaysia and is bordered by Perak to the north, Pahang to the east, Negeri Sembilan to the south and the Strait of Malacca to the west. Selangor surrounds the federal territories of Kuala Lumpur and Putrajaya, both of which were previously part of it.

The state capital of Selangor is Shah Alam and its royal capital is Klang. Petaling Jaya and Subang Jaya received city status in 2006 and 2019, respectively.Selangor is one of four Malaysian states that contain more than one city with official city status; the others are Sarawak, Johor, and Penang.

The state of Selangor has the largest economy in Malaysia in terms of gross domestic product (GDP), with RM 239.968 billion (roughly US$55.5 billion) in 2015, comprising 22.6 percent of the country's GDP. It is the most developed state in Malaysia and has the largest population and the lowest poverty rate in the country.

Business Problem and Sources

The Wikipedia page Selangor-List of districts is the major source of data that is being used to obtain all the districts of Selangor and Foursquare API to access the venues in the neighbourhoods.

List all the major parts and restaurant in Selangor. Which area are the most restaurant located. Folium visualization library can be used to visualize the clusters superimposed on the map of Selangor.

These clusters will be further analyzed to help business owners selecting a potential location to open-up Hotels, Shopping Malls, Restaurants or Coffee shops.

Target Audience

This analyst aims to people who want to start run their first business in Selangor. The below dataset will give them an idea on how to select a better place and what type of business is better.

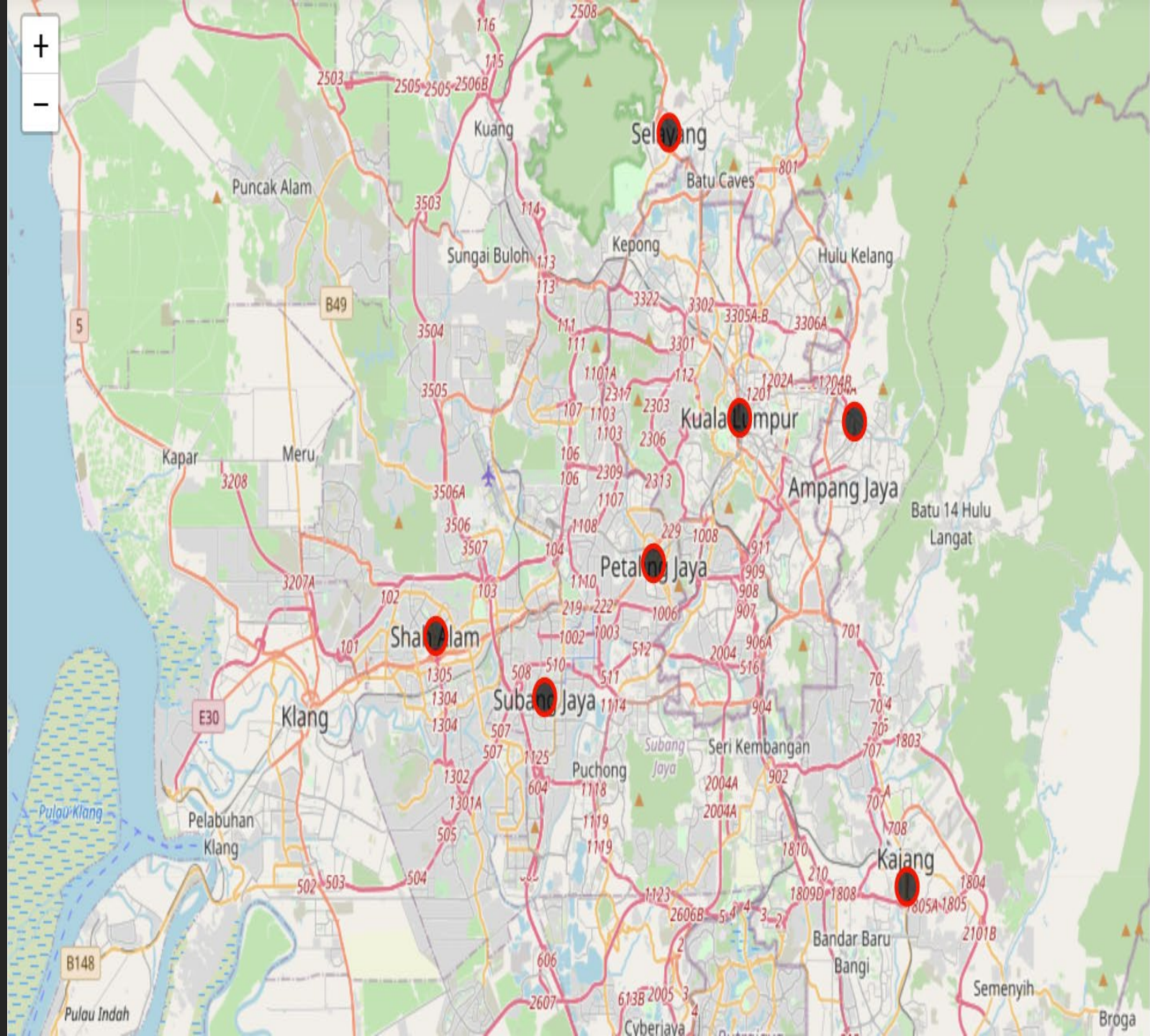
This report also suits people who already own a business and want to make a change for their business. The below data set will provide them some valuable information for them to make decision.

Methodology

Data retrieval, exploration, and wrangling

List of postal code and district name is scrapped from Wikipedia. The latitudes and longitudes for the above neighbourhoods are obtained using a python library named, Geocoder.

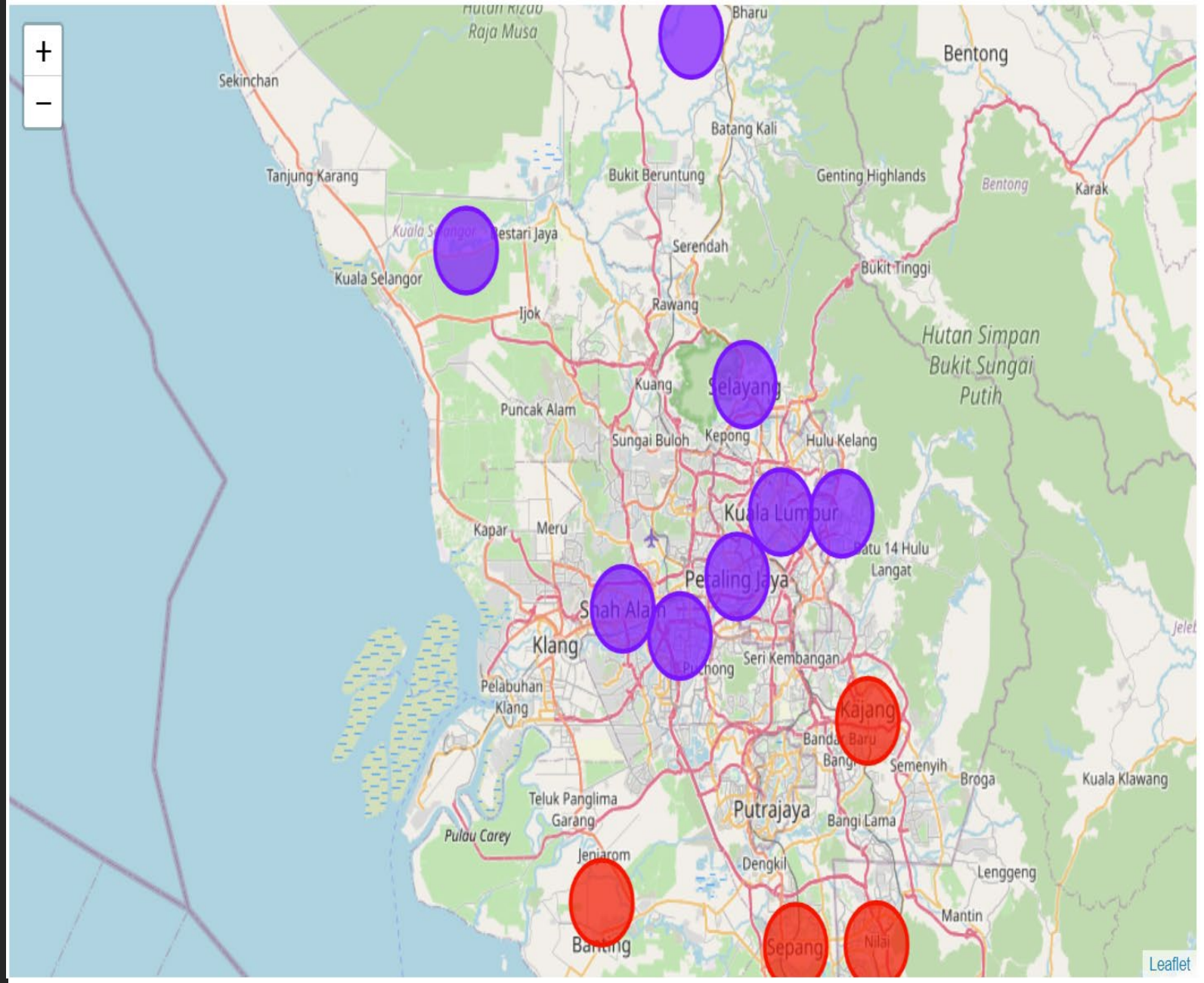
This library uses various APIs from different platforms like google, ArcGISetc to obtain the required location information. The location data thus obtained after utilizing the above library, is then fed into the previous dataframe.



Performing K-means clustering algorithm to segment neighbourhoods

Perform clustering on the data by using k-means clustering. K-Means clustering is an unsupervised learning algorithm which identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

It is one of the simplest and popular machine learning algorithms and is particularly suited to solve the problem for this project.



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

Data Cleaning

Wikipedia

Data cleaning the dataset of Selangor districts has been downloaded, we must edit the dataset provided to only have information, necessary for our problem. Wikipedia provided information on rank, notes, and urban area etc. that will not be required. After cleaning, we will only be left with Local government area, State, Total population, and Status of the districts.

Foursquare

Foursquare provides a dataset of venues around the specific coordinates or venues, if we use the “Explore “function in the Developer tab. Once requested, we get a full breakdown of all recorder venues around the boroughs of interest.

Table

Rank	Local government area	State	Total population	Status	Notes	Urban area
1	Kuala Lumpur	Federal Territories	1588750	City	Federal capital of Malaysia	Greater Kuala Lumpur (Klang Valley)
2	Seberang Perai	Penang	818197	City	Includes Butterworth, Bukit Mertajam, Batu Kaw...	Greater Penang
3	Kajang	Selangor	795522	Municipality	NaN	Greater Kuala Lumpur (Klang Valley)
4	Klang	Selangor	744062	Municipality	NaN	Greater Kuala Lumpur (Klang Valley)
5	Subang Jaya	Selangor	708296	City	Part of Petaling District	Greater Kuala Lumpur (Klang Valley)
6	Penang Island	Penang	708127	City	Includes George Town, the capital of Penang	Greater Penang
7	Ipoh	Perak	657892	City	Capital of Perak	Greater Ipoh (Kinta Valley)
8	Petaling Jaya	Selangor	613977	City	Part of Petaling District	Greater Kuala Lumpur (Klang Valley)
9	Selayang	Selangor	542409	Municipality	NaN	Greater Kuala Lumpur (Klang Valley)
10	Shah Alam	Selangor	541306	City	Capital of Selangor	Greater Kuala Lumpur (Klang

Top 10 population area in Selangor

TOP 5 COMMON VENUE IN SELANGOR

Local government area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Alor Gajah	Malay Restaurant	Asian Restaurant	Café	Pizza Place	Coffee Shop
Alor Setar	Malay Restaurant	Asian Restaurant	Café	Thai Restaurant	Food Truck
Ampang Jaya	Malay Restaurant	Hotel	Coffee Shop	Italian Restaurant	Chinese Restaurant
Batu Pahat	Malay Restaurant	Steakhouse	Bistro	Asian Restaurant	Toll Plaza
Bintulu	Café	Hotel	Beach	Coffee Shop	Asian Restaurant
Hulu Selangor	Food Truck	Night Market	Farm	Motel	Beach
Ipoh	Coffee Shop	Chinese Restaurant	Café	Dessert Shop	Indian Restaurant
Iskandar Puteri	Theme Park Ride / Attraction	Café	Coffee Shop	Theme Park	Indian Restaurant

TOP 5 VENUES BY POPULATION

Local government area	State	Total population	Status	Co-ordinates	Latitude	Longitude	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Kuala Lumpur	Selangor	1588750	City	(3.1516964, 101.6942371)	3.151696	101.694237	0	Hotel	Café	Boutique	Hotel Bar	Shopping Mall
Seberang Perai	Selangor	818197	City	(5.35397935, 100.45734643656445)	5.353979	100.457346	0	Chinese Restaurant	Coffee Shop	Malay Restaurant	Food Truck	Asian Restaurant
Kajang	Selangor	795522	Municipality	(2.98320905, 101.79035124931443)	2.983209	101.790351	2	Malay Restaurant	Restaurant	Chinese Restaurant	Asian Restaurant	Indian Restaurant
Klang	Selangor	744062	Municipality	(49.3198, 6.3722)	49.319800	6.372200	0	Bakery	Vacation Rental	French Restaurant	Resort	History Museum
Subang Jaya	Selangor	708296	City	(3.051487, 101.5823339)	3.051487	101.582334	0	Coffee Shop	Japanese Restaurant	Convenience Store	Bakery	Ice Cream Shop
Penang Island	Selangor	708127	City	(5.3696832, 100.25281058745169)	5.369683	100.252811	0	Chinese Restaurant	Asian Restaurant	Café	Food Truck	Seafood Restaurant

Results

Results of the above analysis and clustering cam be summarized:

1. The most common venue in Selangor is Malay Restaurant, Asian Restaurant and Cafe.

2. In the second list we can see the most common results is Pizza Shop, follow by Zoo and Flower Shop.

3. Refer to above data analyst, we can confirm Selangor residents more prefer to go to restaurant rather than others.

Discussion

Looking at the data, Alor Gajah and Alor Setar are the best places outside of Kuala Lumpur where a new venue is worth opening. However, a lot of information is not considered, and cannot be obtained from Foursquare Developer.

Higher ethnic presence in each borough can and will influence the popularity of a given cuisine. Closer proximity to Inner boroughs and better transport links allows people to travel to the neighbouring borough and impact the measurements.

Many small venues are not registered in Foursquare and are marketed via word-of-mouth, and are not considered Regardless, the analysis provided an insight into what people like and opt for, when it comes to going out in their own neighbourhoods.

Conclusion

Objective of this project was to analyze the neighbourhoods of Selangor and create a clustering model to suggest potential places to start a new business based on the category. The neighbourhood's data was obtained from an online source and the Foursquare API was used to find the major venues in each neighbourhood.

A few examples for the applications that the clusters can be used for have also been discussed. A map showing the clusters have been provided. Both these can be used by stakeholders to decide the location for the business. But the data have some of the information was not reliable due to Foursquare API not popular using in Selangor, so we need more data and other parts to take in for consideration before making any decision.