The Battle of Neighborhoods

Aravind Selvaraj

June 22, 2021

1. Introduction:

1.1 Background

The Battle of neighborhood is one of the capstone projects of the certifcation offered by coursera. In this project, I took an dataset, analyze it and concluded it with the result I got.

1.2 Problem Statement

There is an company, who wants to launch a coffee shop and restaurant in an kuala lumpur, Malaysia. They don't know which could be the better place to launch it so that it will be profitable to them.
As a Data Scientist, need to analyze the city data and suggest some place to them so that they can launch their Restaurant and coffee shop.

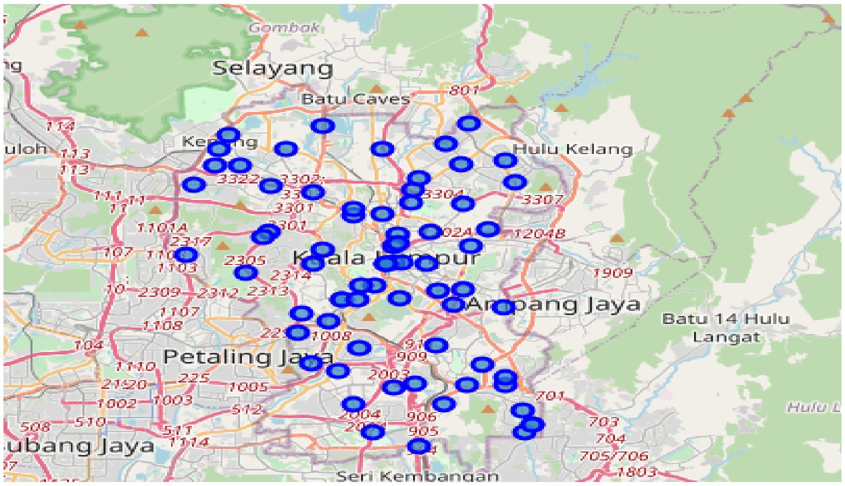
2. Data acquisition and cleaning:

2.1 Data sources

As a project in data science field, data would be the major source. I took the data from the wikipedia page which consisits of the various places in the kuala lumpur, Malaysia.
You can get the data from here

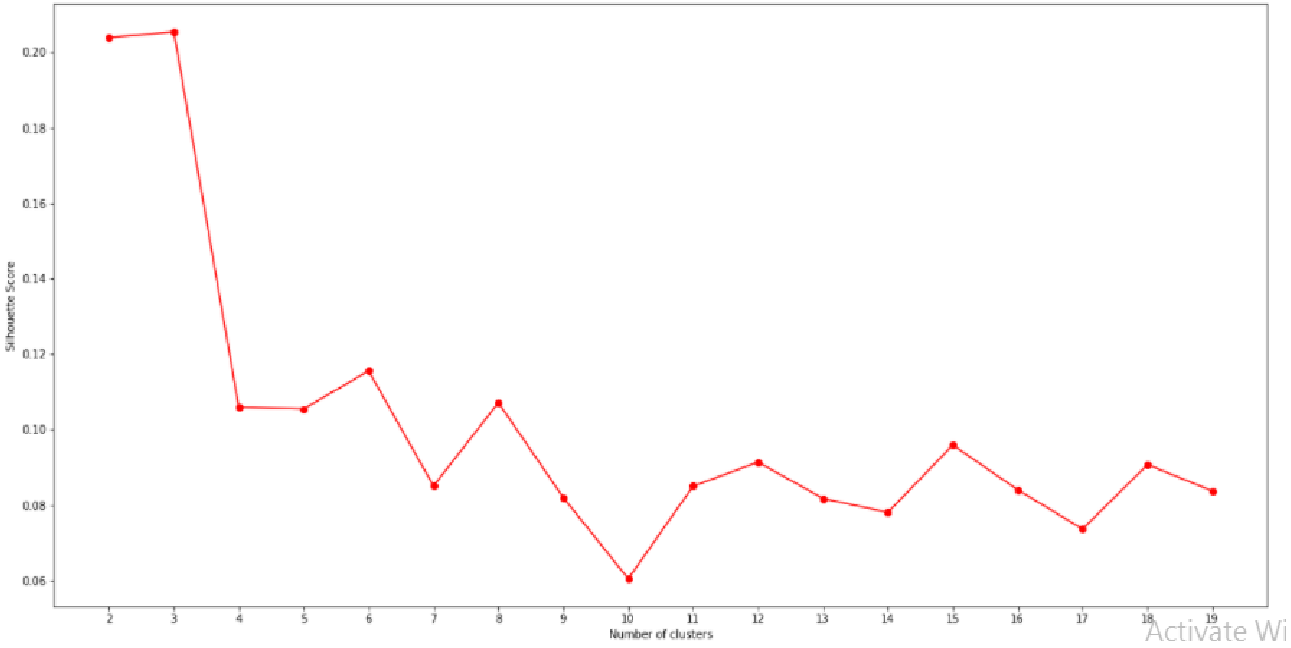
3. Exploratory Data Analysis:

Once, we get the data, I needed to visualize it by using some python library. This project is totally based on the city data. So I have used library called **Geolocator, Folium.**
Geolocator is used to get the latitude and longitude for a particular address we know. In this case, it is for **kuala lumpur, malaysia**.
Once we get the latitude and longitude, we can pass it to the function in folium which plot the city map with all the data we have scrapped.

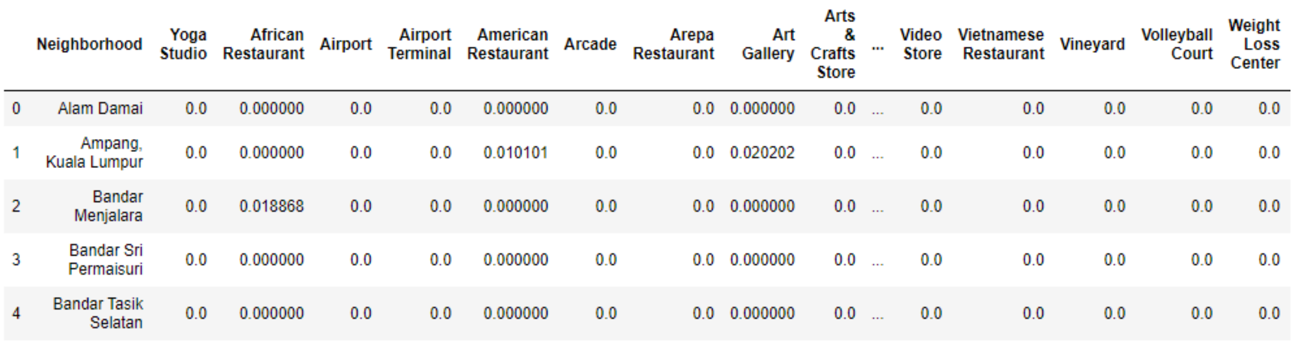


In that image, all the blue marks are the area we have scrapped and it plots in the map of kuala lumpur, malaysia.
This is one of the best way of visualizing the city data.

4. Predictive Modelling:
There are any techniques are available in data science for modelling. Here I choose **K-Means** clustering.
K-Means clustering is a un-supervised learning technique. Since we have unlabelled data, this method could be the better one to use.
Whereas when it comes to K-Means clustering choosing right K value would be the tricky one. So I have used a function and plot over many K values and choose the right one.



While seeing this chart, after the point 3, the data points are not much increased. So kcluster as 3 could be the better one.
After choosing the right k value one thing we need to consider is the data. In this city data we have lot of categorical data where as for clustering we cannot use the categorical data as it is. So need to convert those categorical data into an numeric data.
One approach could be converting the data into one-hot data and finding the mean value for each cell

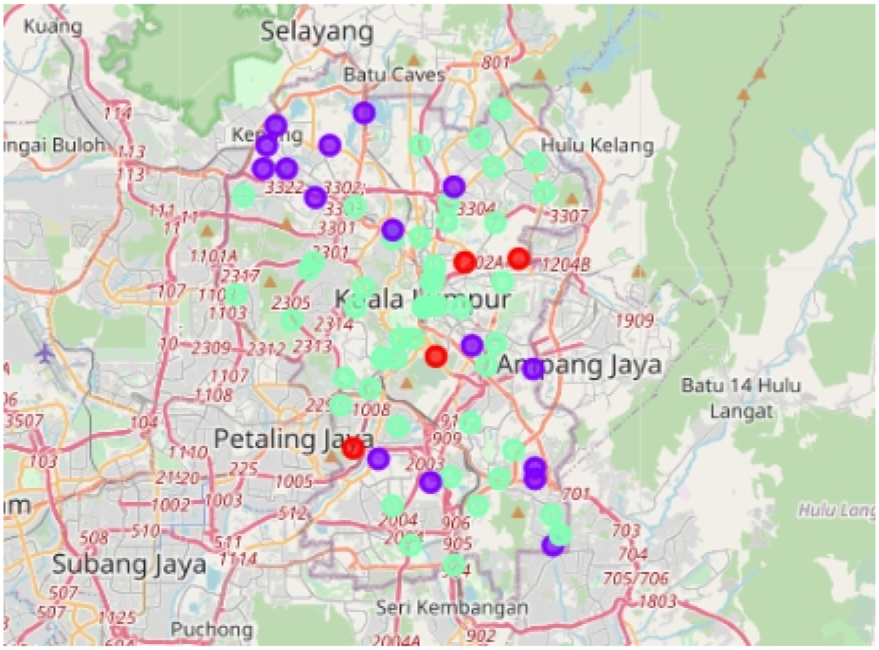


	Neighborhood	Yoga Studio	African Restaurant	Airport	Airport Terminal	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	Arts & Crafts Store	...	Video Store	Vietnamese Restaurant	Vineyard	Volleyball Court	Weight Loss Center
0	Alam Damai	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0
1	Ampang, Kuala Lumpur	0.0	0.000000	0.0	0.0	0.010101	0.0	0.0	0.020202	0.0	...	0.0	0.0	0.0	0.0	0.0
2	Bandar Menjalara	0.0	0.018868	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0
3	Bandar Sri Permaisuri	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0
4	Bandar Tasik Selatan	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0

Once we convert into one-hot data, our original data looks like this.
We can pass this data to the model and train it.

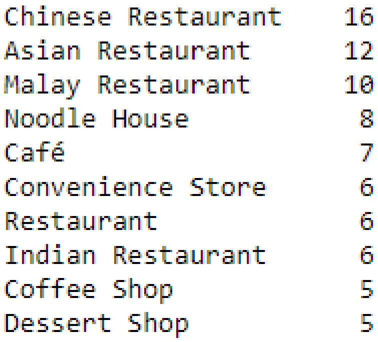
5.Conclusion:

After fitting the model, we need to analyse it to find which place could be the better one to launch the restaurant and coffee shop.



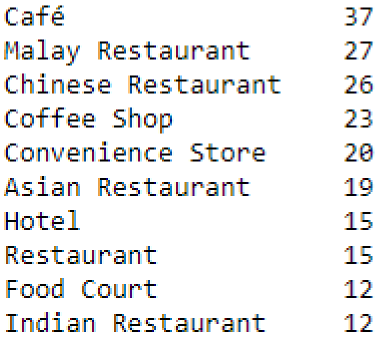
This is the model we trained and all the data we had were placed according to their cluster.
Also, these are the shops and counts of such shops available in those cluster.

Cluster0:



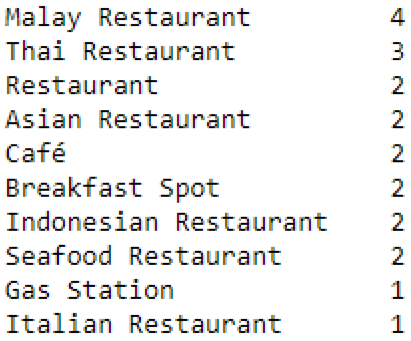
```
Chinese Restaurant    16
Asian Restaurant      12
Malay Restaurant      10
Noodle House           8
Café                   7
Convenience Store      6
Restaurant             6
Indian Restaurant      6
Coffee Shop            5
Dessert Shop           5
```

Cluster1



```
Café                      37
Malay Restaurant          27
Chinese Restaurant        26
Coffee Shop               23
Convenience Store         20
Asian Restaurant          19
Hotel                     15
Restaurant                15
Food Court                12
Indian Restaurant         12
```

Cluster2



```
Malay Restaurant          4
Thai Restaurant           3
Restaurant                2
Asian Restaurant          2
Café                      2
Breakfast Spot            2
Indonesian Restaurant     2
Seafood Restaurant        2
Gas Station               1
Italian Restaurant        1
```

As we can see, in cluster0 and cluster2, there is no much of coffee shops and as in of same cluster0 and cluster2 not much of restaurant.

So for the company who wants to launch a new restaurant and coffee shops, the areas which is in cluster0 and cluster2 will be the best option.