The Battle of neighborhoods

- Aravind Selvaraj

Business understanding

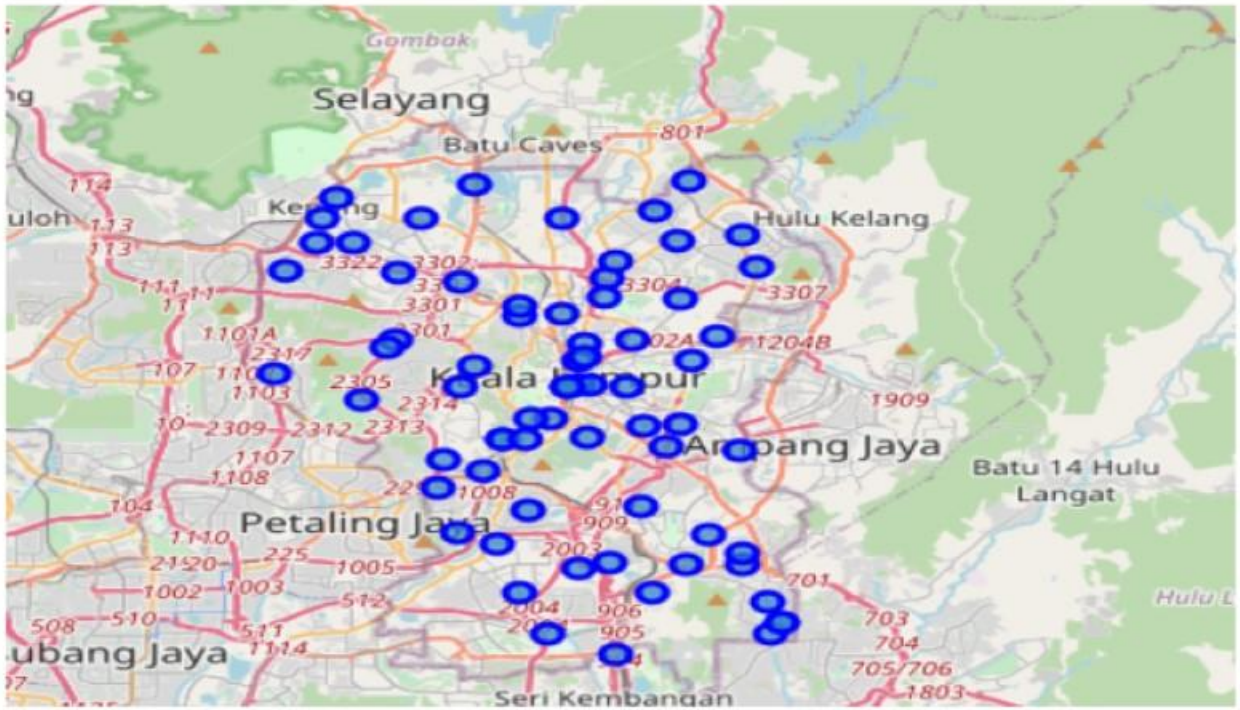
Finding an better venue in a kuala lumpur, Malaysia for an company to launch a new restaurant and coffee shop based on the count of existing shops available.

Data Collection

- There are lot of open source data sets available like kaggle.
- Here I am using an data from the Wikipedia .
- It consists of all the cities names in kuala lumpur, Malaysia.
- By using the library called **BeautifulSoup**, I scrapped those data and used it.

Exploratory Data analysis

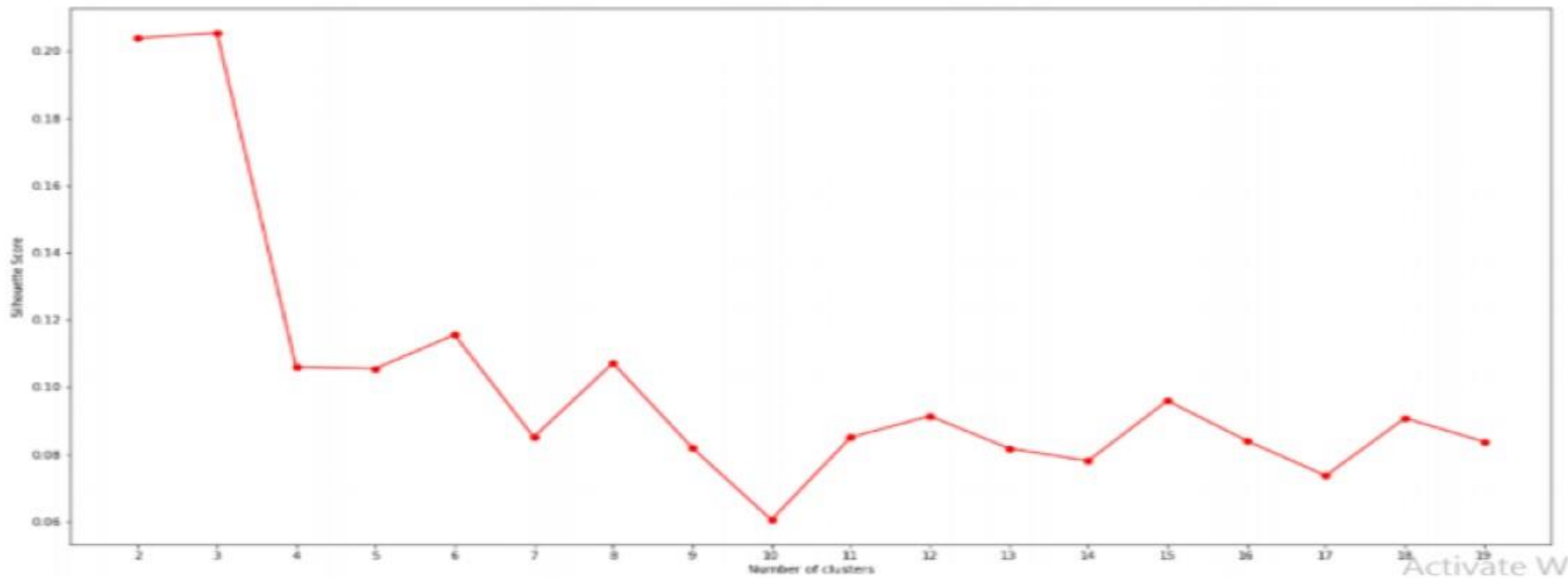
- For every data science projects, we need to visualize the data either by using any chart or plots.
- In this projects, I used city data which is not possible to do any chart/plots.
- So I used a library called **folium** to visualize the data we scrapped.



Predictive Modelling

- There are many 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

Predictive Modelling cont'd



While seeing this chart, after the point 3, the data points are not much increased. So kcluster as 3 could be the better one.

One-Hot encoding

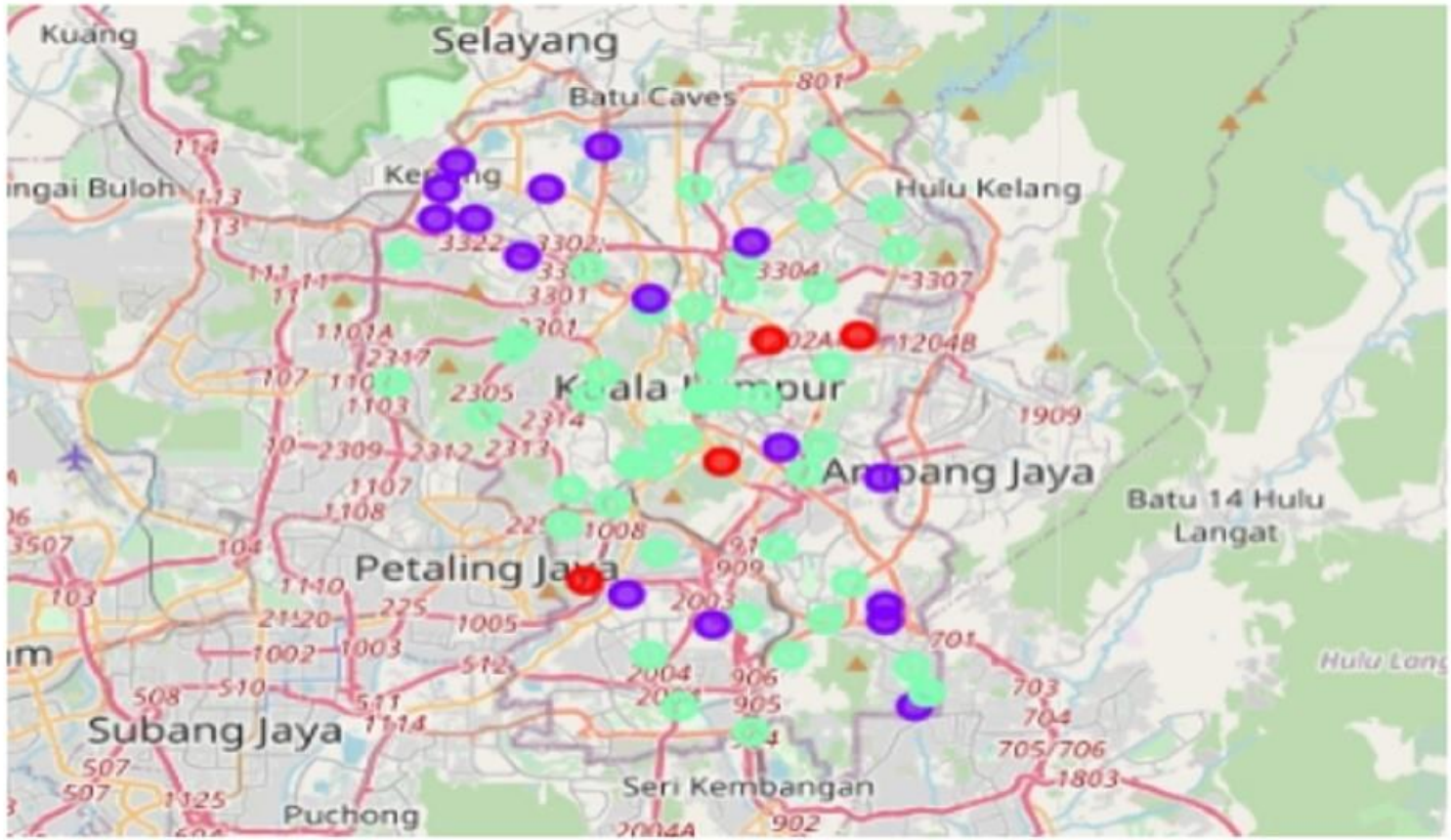
- After choosing the right k value one thing we need to consider is the data.
- In this city data we have lot of categorical data where as for clustering we cannot use the categorical data as it is. So need to convert those categorical data into an numeric data.
- One approach could be converting the 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	[illegible] & 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.

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.



Conclusion

- Once 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:		Cluster1		Cluster2	
Chinese Restaurant	16	Café	37	Malay Restaurant	4
Asian Restaurant	12	Malay Restaurant	27	Thai Restaurant	3
Malay Restaurant	10	Chinese Restaurant	26	Restaurant	2
Noodle House	8	Coffee Shop	23	Asian Restaurant	2
Café	7	Convenience Store	20	Café	2
Convenience Store	6	Asian Restaurant	19	Breakfast Spot	2
Restaurant	6	Hotel	15	Indonesian Restaurant	2
Indian Restaurant	6	Restaurant	15	Seafood Restaurant	2
Coffee Shop	5	Food Court	12	Gas Station	1
Dessert Shop	5	Indian Restaurant	12	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.