# IBM Data Science Coursera Capstone Project

## Introduction: Business Problem

Jakarta is the capital of Indonesia with a population of 10,5 million, and is the heart of the second densest metropolitan area in the world behind Tokyo, Japan. Having hosted ASEAN games recently back in 2018, it has witnessed heavy investment in transportational infrastructure with the opening of the first MRT line in Indonesia just this year.

Given that, the city will surely see more growth of which information regarding the lay of the land will be invaluable for investors and entrepreneurs to make strategic decision for investment or choosing locations for business operation.

This project will attempt to explore patterns of subdistricts within Jakarta by categorizing them into clusters in order to identify existing trends within neighborhoods of Jakarta. From there on, recommendations can be made on which category of neighborhood will be most suitable for a certain type of venue to be opened.

The result of this project is aimed at general entrepreneur but may be most useful for entrepreneurs on the food and beverage sector given that location can be the deciding factor for a success.

## Data

To analyze trends in Jakarta's subdistrict, the list of subdistrict is obtained from [Jakarta subdistrict wikipedia page](https://en.wikipedia.org/wiki/Subdistricts_of_Jakarta).

Venue queries will then be made by subdistricts using FourSquare APIs. The resulting data regarding venue category will be used to observe commonality between subdistricts. The commonality clusters can then provide insight on which type of venue will thrive better on which cluster. K-means clustering algorithm will be used to find pattern between the subdistricts.

In summary, the following data is required to meet the objective:

* Subdistricts of Jakarta
* Coordinates of these subdistricts
* Trending Venues on the area
* Venue categories

## Methodology

Given that our objective is to generally categorize the subdistricts, we will use K-means clustering algorithm to categorize each of the subdistricts within Jakarta.

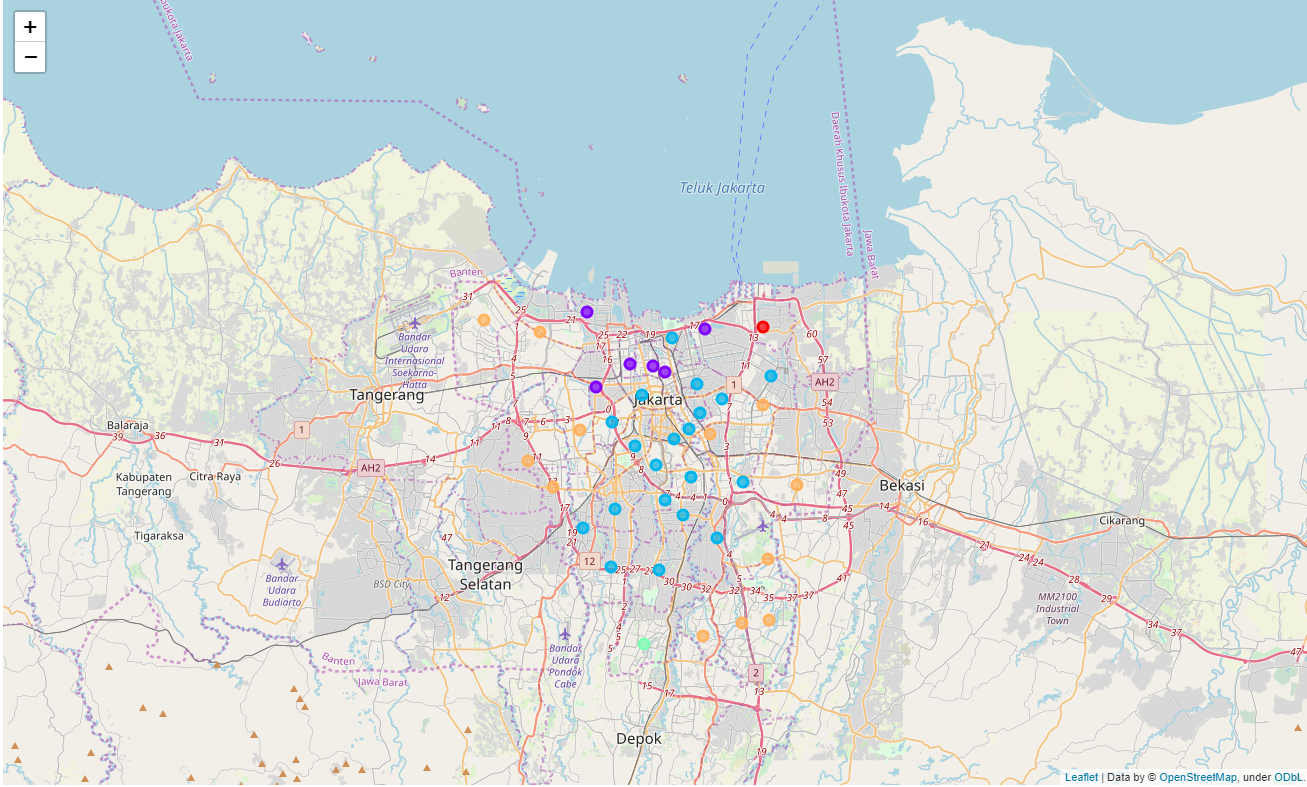
A one-hot encoding will be done on the venue dataframe and it will be grouped by subdistrict. The encoding will return venue categories as column per subdistrict, which will then be grouped to provide weighting of venue type occurence on each subdistrict.

The encoded dataframe will be further filtered into top venues before the K-means clustering algorithm will be run over it. This will return cluster labels over the subdistricts. The clusters will be observed one by one manually to determine its content.

Recommendation will be made based on the clusterring.

## Analysis

### Map of Clusters



**Cluster 1**

Cluster 1 contains a higher concentration of noodle house and chinese restaurant. Located north of Jakarta, this might be the cluster for Chinese restaurants.



**Cluster 2**

With the most member, cluster 2 seems to contain a good amount of coffee shops and hotels. Located central-south of the city, it is expected to have a high concentration of places to hang out and stay.



**Cluster 3**

Containing only 1 subdistrict, cluster 3 seems to be in its own group located at the south of the city. This might be due to the low count of trending venues of 10. Nothing much can be gathered from this cluster.



**Cluster 4**

Having Indonesian restaurants as its common venue, cluster 4 have higher count of fast food and convenience store in general. The subdistricts within this cluster are more scatered, being dispersed on around the border surrounding cluster 2.



**Cluster 5**

Similar with cluster 3 with only 1 subdistrict, cluster 5 seems to be in its own group located at the north east of the city. This might be due to the low count of trending venues of 15. Nothing much can be gathered from this cluster.



## Result and Discussion

Groupings as a result of K-means clustering algorithm tallies with how Jakarta historically develops. Having most of cluster 1, containing a high count of chinese restaurant, at the north side of the city fits the chinatown part of the city. Cluster 2 being the dominant type of subdistrict which is located in the middle also fits the reality. North eastern part being quite sparse in trending venue also fits the reality that the area is more of an industrial area, thus having less venues.

There are definite limitation with using the FourSquare API as the 100 venues limit might skew the result of the more densely populated subdistrict. Also, some subdistricts have low count of venues that it might be considered to be insufficient in determining its characteristics. It might also be the case that FourSquare user base are skewed to the foodie type, which might explain the limited trending venues on the north east part of the city.

For most of the subdistricts, restaurants and coffee shops are the dominant venue type with cluster 2 having more variation in terms of cuisine.

## Conclusion

Opening of new western restaurant may be best done in cluster 5 where there are less of such restaurant to compete. Business which does not rely on foot traffic may choose to locate themself in the north east of Jakarta.