Using Machine Learning for Neighborhood Clustering Comparison of Capital City of Indonesia

This is my week 1 final capstone project for IBM Applied Data Science Capstone course in Coursera.

A. Introduction

A.1. Description & Disscusion of the Background

Jakarta, officially the **Special Capital Region of Jakarta** (Indonesian: *Daerah Khusus Ibukota Jakarta*), is the capital and largest city of Indonesia. Situated on the northwest coast of the world's most populous island of Java, it is the centre of economy, culture and politics of Indonesia with a population of more than **10 million** as of 2014. Officially, the area of the Jakarta Special District is 662 km^2 (256 sq mi) of land area and 6,977 km^2 (2,694 sq mi) of sea area. Jakarta consists of five Kota Administratif (Administrative cities/municipalities) and one Kabupaten Administratif (Administrative regency). Each city and regency is divided into districts/Kecamatan.

Jakarta's prime challenges include rapid urban growth, ecological breakdown, gridlocked traffic, congestion, and flooding. [1]. Additionally, Jakarta is sinking up to 17 cm (6.7 inches) per year, which, coupled with the rising of sea levels, has made the city more prone to flooding. It is also one of the fastest-sinking capitals in the world.[19] In August 2019, President Joko Widodo announced a move of the capital to the province of East Kalimantan on the island of Borneo.

East Kalimantan had a population of about 3.42 million at the 2015 Census; Its capital is Samarinda. The province will host the future capital city of Indonesia and its construction is projected to start in 2020, and conclude in 2024. East Kalimantan has a total area of 129,066.64 square kilometres (49,832.91 sq mi) and is the second least densely populated province in Kalimantan. East Kalimantan, is now divided into seven regencies and three cities, subdivided into 103 districts and 1,026 villages (kelurahan).

As a resident of Jakarta city and the announcement of capital city movement, I decided to use Jakarta for this capstone project. I would like to compare between Jakarta as the current capital city and East Kalimantan as the future capital city. The comparisons are about the neighborhoods and business prospects based on venues perspectives such as number of venues and veneus category. The result of this capstone project may become a reference to prepare East Kalimantan as the future capital city of Indonesia.

A.2. Target Audience

Government who want to prepare East Kalimantan as the future of capital city of Indonesia. Entrepreneur would like to create business or/and person or family who wants to move to the future of capital city of Indonesia.

A.3. Data Description

To consider the problem we can list the datas as below:

- ✓ The data which contains postal code, sub-district, district, and city of Jakarta and East Kalimantan.
- ✓ Based on data above, then locate the geocoding coordinates of each neighborhoods location using **Geopy** python library and **Google** Geocoding API.
- ✓ Used Forsquare API to get the most common venues of given neighborhoods (kelurahan) of Jakarta and East Kalimantan.

B. Methodology

As a database, I used GitHub repository for this capstone project. My master data which has the main components Neighborhoods/Urbans, Boroughs/Sub-District, Postal Code, Latitude and Longitude informations of Jakarta and East Kalimantan. To get the coordinates, I tried using Geocoder package and Google Geocoding API to match the coordinates of Jakarta's neighborhoods and East Kalimantan's neighborhoods.

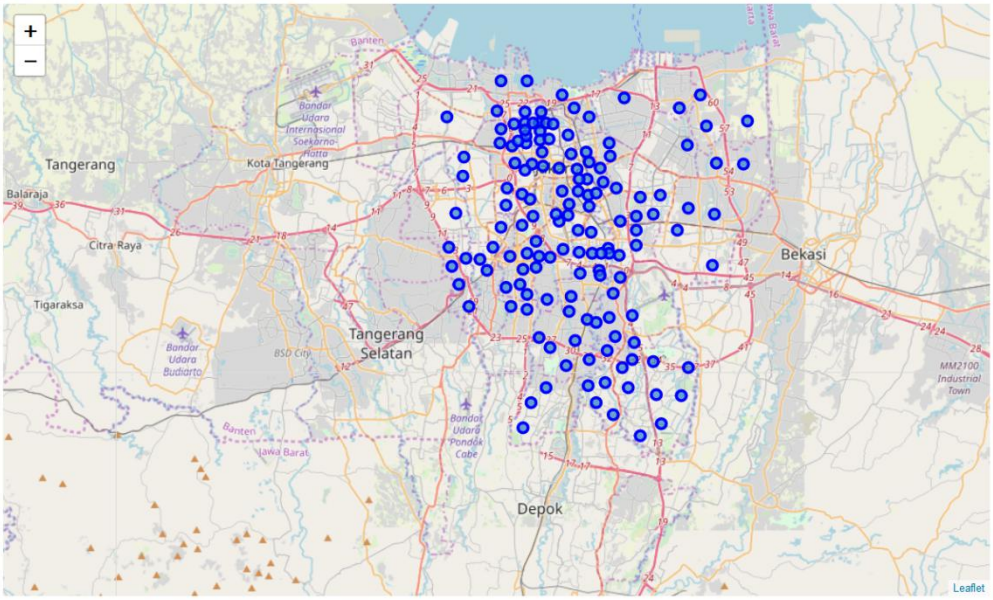
```
The jakarta dataframe shape is (261, 7)
```

Out[125]:

	urban	postal_code	latitude	longitude	sub_district	city	address
0	ANCOL	14430	-6.127243	107.222605	PADEMANGAN	JAKARTA UTARA	ANCOL, ID 14430
1	ANGKE	11330	-6.145900	106.795900	TAMBORA	JAKARTA BARAT	ANGKE, ID 11330
2	BALEKAMBANG	13530	-6.281400	106.852400	KRAMAT JATI	JAKARTA TIMUR	BALEKAMBANG, ID 13530
3	BALI MESTER	13310	-6.220472	106.866717	JATINEGARA	JAKARTA TIMUR	BALI MESTER, ID 13310
4	BAMBU APUS	13890	-6.313792	106.900130	CIPAYUNG	JAKARTA TIMUR	BAMBU APUS, ID 13890

After gathering all these coordinates, I visualized the map of Jakarta and East Kalimantan using Folium package to verify whether these are correct coordinates.

Out[126]:



Next, I utilized the Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

In [132]:
```
print('jakarta venues dataframe shape is', jakarta_venues.shape)
print('There are {} uniques categories.'.format(len(jakarta_venues['Venue Category'].unique())))
jakarta_venues.head()
```

```
jakarta venues dataframe shape is (2293, 7)
There are 240 uniques categories.
```

Out[132]:

	urban	urban latitude	urban Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ANCOL	-6.127243	106.8291	Discovery Hotel & Convention Ancol	-6.126035	106.831260	Hotel
1	ANCOL	-6.127243	106.8291	Dunia Fantasi (DUFAN)	-6.124300	106.832089	Theme Park
2	ANCOL	-6.127243	106.8291	Aston Marina	-6.129624	106.829485	Hotel
3	ANCOL	-6.127243	106.8291	Jaya Ancol Bowling Centre	-6.128302	106.831782	Bowling Alley
4	ANCOL	-6.127243	106.8291	Talaga Sampireun	-6.126231	106.833339	Sundanese Restaurant

Based on data above, then I created a table which shows list of top 10 venue category for each neighborhoods in below table.

```
neighbourhoods venues dataframe shape is (158, 241)
```

Out[138]:

	urban	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	ANCOL	Theme Park	Hotel	Sundanese Restaurant	Bowling Alley	Japanese Restaurant	BBQ Joint	Seafood Restaurant	Beach	Athletics & Sports	Food Truck
1	ANGKE	Noodle House	Juice Bar	Café	Lounge	Women's Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food Court
2	BALEKAMBANG	Restaurant	Noodle House	Gym	Women's Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food Court	Food & Drink Shop
3	BALI MESTER	Asian Restaurant	Indonesian Restaurant	Convenience Store	Chinese Restaurant	Pizza Place	Bar	Fast Food Restaurant	Bakery	Coffee Shop	Seafood Restaurant
4	BAMBU APUS	Indonesian Restaurant	Seafood Restaurant	Soup Place	High School	Restaurant	Flea Market	Farmers Market	Fast Food Restaurant	Field	Fish & Chips Shop

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, 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 one of the most common cluster method of unsupervised machine learning algorithms and it is highly suited for this capstone project as well. I will run K-Means to cluster the neighborhoods into **5** clusters, And next is merged table with cluster labels for each neighborhoods. We can also estimate the number of **1st Most Common Venue** in each cluster. Thus, we can create a chart which may help us to find proper labels for each cluster of Jakarta and East Kalimantan.

F. References:

[1]. Jakarta — Wikipedia
[2]. Indonesia Postal Code
[3]. Forsquare API
[4]. Google Geocoding API

bdy