Happy New Year 2020. This is my capstone project for IBM Applied Data Science Capstone course in Coursera. Hope this can be an insight for applied data science.

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).[1]

Jakarta's prime challenges include rapid urban growth, ecological breakdown, gridlocked traffic, congestion, and flooding. 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. 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 divided into 7 regencies and 3 cities, subdivided into 103 districts and 1,026 villages (kelurahan).[2]

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 East Kalimantan.

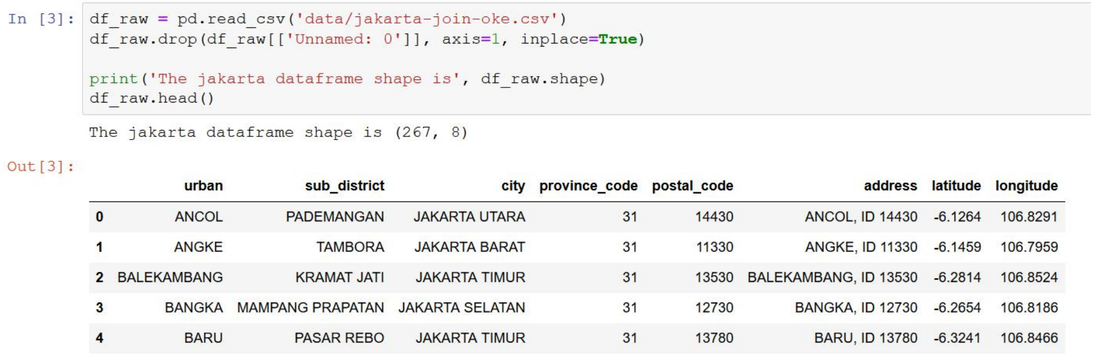
A.3. Data Description

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

- ✓ The data which contains postal code, urban/neighborhood, 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[3] and **Google** Geocoding API [4].
- ✓ Used Forsquare API[5] to get the most common venues of given neighborhoods (kelurahan) of Jakarta and East Kalimantan.

B. Methodology

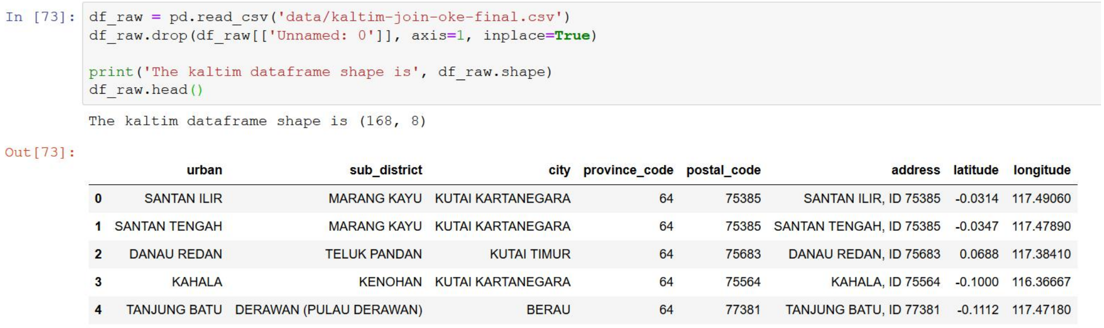
As a database, I used GitHub repository for this capstone project. My master data which has the main components Neighborhoods/Urbans, Boroughs/Sub-District, Postal Code, Latitude and Longitude informations of Jakarta and East Kalimantan. This data is gained from postal code data then combine with coordinate data. The challenge is the coordinate data is not complete, therefore I use geocoding using geopy python library and google geocoding API. Below are the master data.



```
In [3]: df_raw = pd.read_csv('data/jakarta-join-oke.csv')
        df_raw.drop(df_raw[['Unnamed: 0']], axis=1, inplace=True)

        print('The jakarta dataframe shape is', df_raw.shape)
        df_raw.head()

        The jakarta dataframe shape is (267, 8)
Out[3]:
```

	urban	sub_district	city	province_code	postal_code	address	latitude	longitude
0	ANCOL	PADEMANGAN	JAKARTA UTARA	31	14430	ANCOL, ID 14430	-6.1264	106.8291
1	ANGKE	TAMBORA	JAKARTA BARAT	31	11330	ANGKE, ID 11330	-6.1459	106.7959
2	BALEKAMBANG	KRAMAT JATI	JAKARTA TIMUR	31	13530	BALEKAMBANG, ID 13530	-6.2814	106.8524
3	BANGKA	MAMPANG PRAPATAN	JAKARTA SELATAN	31	12730	BANGKA, ID 12730	-6.2654	106.8186
4	BARU	PASAR REBO	JAKARTA TIMUR	31	13780	BARU, ID 13780	-6.3241	106.8466

Jakarta’s Master Data



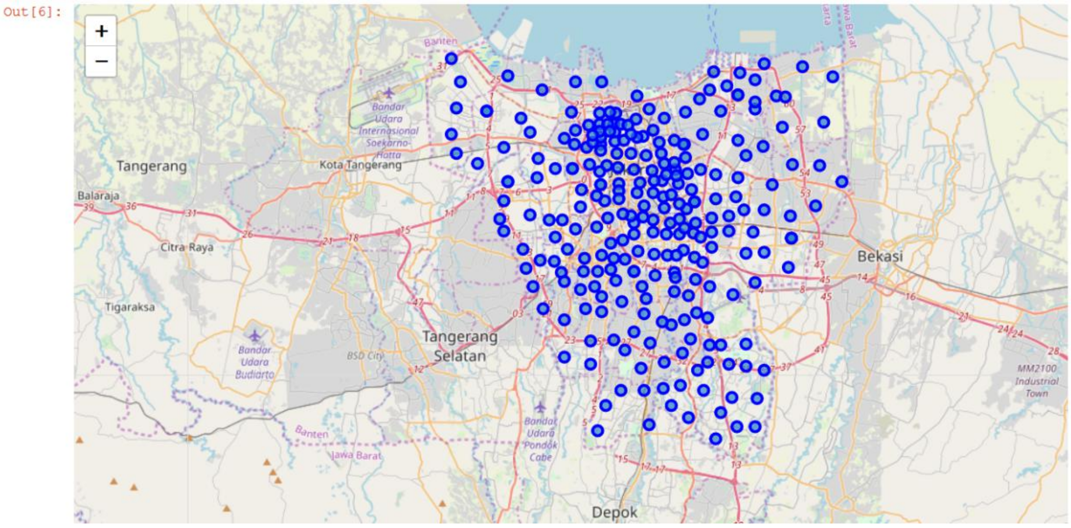
```
In [73]: df_raw = pd.read_csv('data/kaltim-join-oke-final.csv')
         df_raw.drop(df_raw[['Unnamed: 0']], axis=1, inplace=True)

         print('The kaltim dataframe shape is', df_raw.shape)
         df_raw.head()

         The kaltim dataframe shape is (168, 8)
Out[73]:
```

	urban	sub_district	city	province_code	postal_code	address	latitude	longitude
0	SANTAN ILIR	MARANG KAYU	KUTAI KARTANEGARA	64	75385	SANTAN ILIR, ID 75385	-0.0314	117.49060
1	SANTAN TENGAH	MARANG KAYU	KUTAI KARTANEGARA	64	75385	SANTAN TENGAH, ID 75385	-0.0347	117.47890
2	DANAU REDAN	TELUK PANDAN	KUTAI TIMUR	64	75683	DANAU REDAN, ID 75683	0.0688	117.38410
3	KAHALA	KENOHAN	KUTAI KARTANEGARA	64	75564	KAHALA, ID 75564	-0.1000	116.36667
4	TANJUNG BATU	DERAWAN (PULAU DERAWAN)	BERAU	64	77381	TANJUNG BATU, ID 77381	-0.1112	117.47180

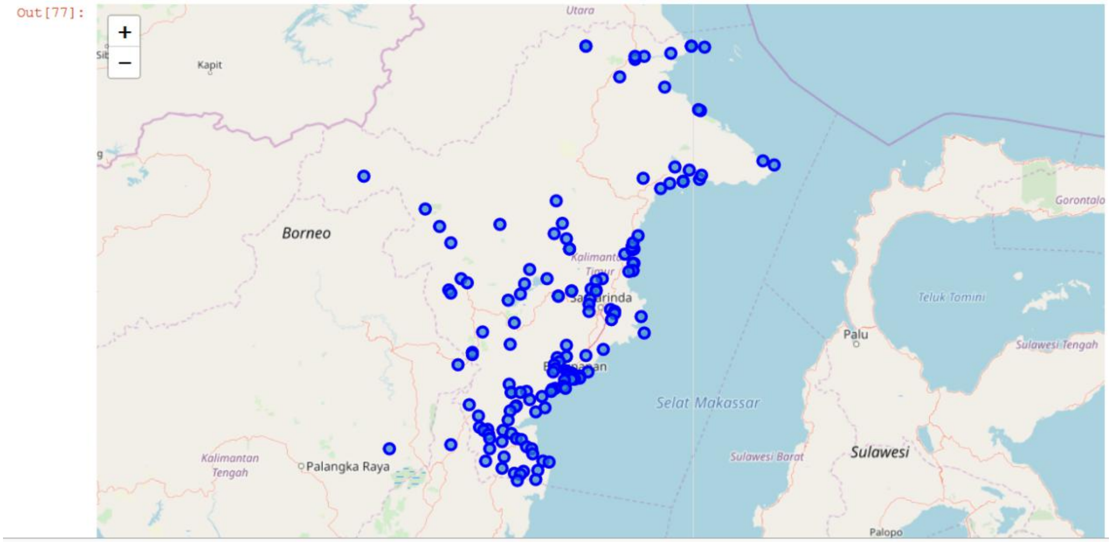
East Kalimantan’s Master Data

Based on data above, then i used **folium** python library to visualize geographic details of Jakarta and East kalimantan. I created a map of Jakarta and East

Kalimantan with Neighborhoods/Urbans superimposed on top. I used latitude and longitude values to get the visual as below:

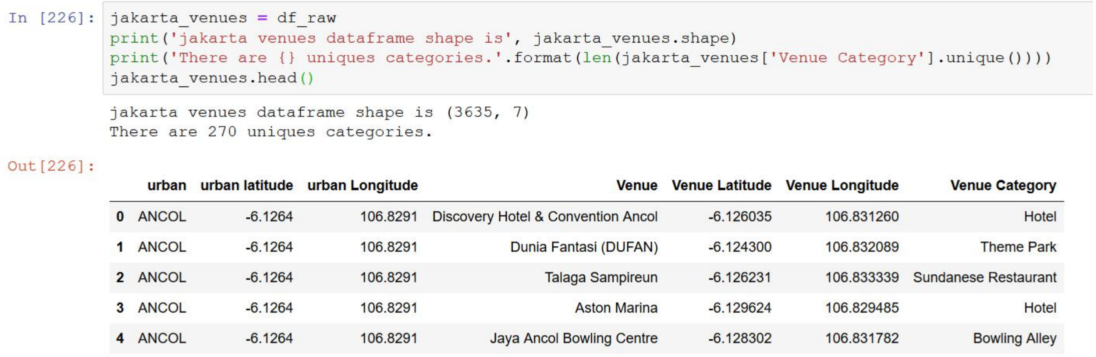


Jakarta’s Map of neighborhoods



East Kalimantan’s Map of neighborhoods

I utilized the Foursquare API to explore the neighborhoods and segment them. I designed the limit as **100 venue** and the radius **500 meter** for each neighborhoods from their given latitude and longitude informations. Here is a head of the list Venues name, category, latitude and longitude informations from Forsquare API. Here is a merged table of neighborhoods and venues.

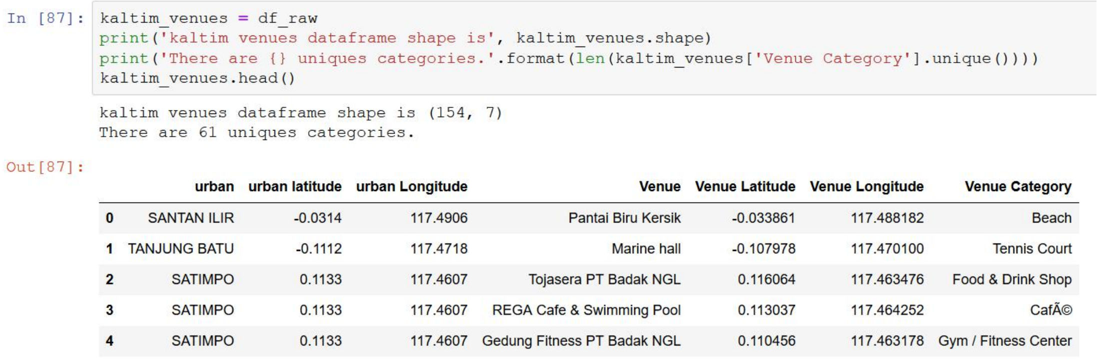


```
In [226]: jakarta_venues = df_raw
          print('jakarta venues dataframe shape is', jakarta_venues.shape)
          print('There are {} uniques categories.'.format(len(jakarta_venues['Venue Category'].unique())))
          jakarta_venues.head()

          jakarta venues dataframe shape is (3635, 7)
          There are 270 uniques categories.
```

Out[226]:

	urban	urban latitude	urban Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ANCOL	-6.1264	106.8291	Discovery Hotel & Convention Ancol	-6.126035	106.831260	Hotel
1	ANCOL	-6.1264	106.8291	Dunia Fantasi (DUFAN)	-6.124300	106.832089	Theme Park
2	ANCOL	-6.1264	106.8291	Talaga Sampireun	-6.126231	106.833339	Sundanese Restaurant
3	ANCOL	-6.1264	106.8291	Aston Marina	-6.129624	106.829485	Hotel
4	ANCOL	-6.1264	106.8291	Jaya Ancol Bowling Centre	-6.128302	106.831782	Bowling Alley

Jakarta's Venues



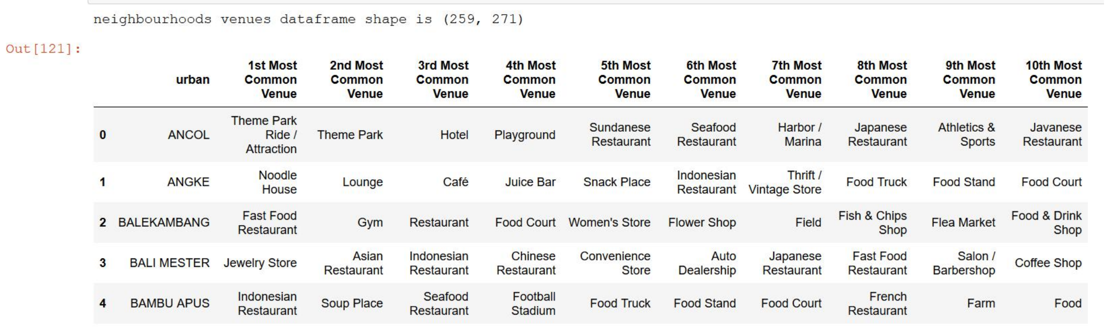
```
In [87]: kaltim_venues = df_raw
         print('kaltim venues dataframe shape is', kaltim_venues.shape)
         print('There are {} uniques categories.'.format(len(kaltim_venues['Venue Category'].unique())))
         kaltim_venues.head()

         kaltim venues dataframe shape is (154, 7)
         There are 61 uniques categories.
```

Out[87]:

	urban	urban latitude	urban Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	SANTAN ILIR	-0.0314	117.4906	Pantai Biru Kersik	-0.033861	117.488182	Beach
1	TANJUNG BATU	-0.1112	117.4718	Marine hall	-0.107978	117.470100	Tennis Court
2	SATIMPO	0.1133	117.4607	Tojasera PT Badak NGL	0.116064	117.463476	Food & Drink Shop
3	SATIMPO	0.1133	117.4607	REGA Cafe & Swimming Pool	0.113037	117.464252	CafÃ©
4	SATIMPO	0.1133	117.4607	Gedung Fitness PT Badak NGL	0.110456	117.463178	Gym / Fitness Center

East Kalimantan's Venues

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

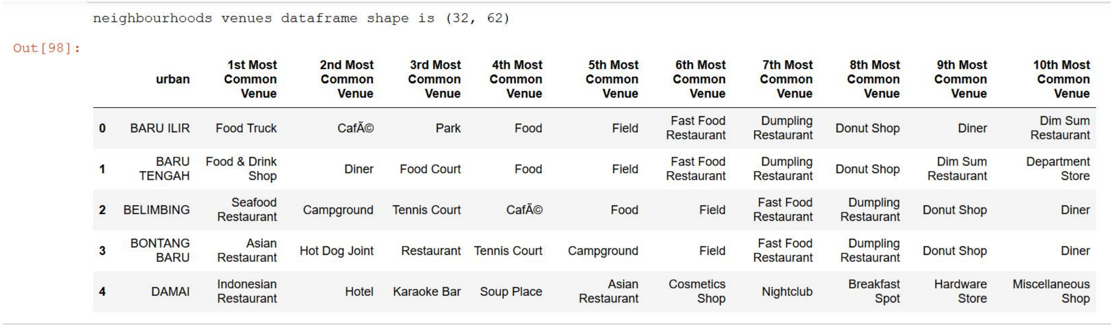


```
neighbourhoods venues dataframe shape is (259, 271)
```

Out[121]:

	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 Ride / Attraction	Theme Park	Hotel	Playground	Sundanese Restaurant	Seafood Restaurant	Harbor / Marina	Japanese Restaurant	Athletics & Sports	Javanese Restaurant
1	ANGKE	Noodle House	Lounge	Café	Juice Bar	Snack Place	Indonesian Restaurant	Thrift / Vintage Store	Food Truck	Food Stand	Food Court
2	BALEKAMBANG	Fast Food Restaurant	Gym	Restaurant	Food Court	Women's Store	Flower Shop	Field	Fish & Chips Shop	Flea Market	Food & Drink Shop
3	BALI MESTER	Jewelry Store	Asian Restaurant	Indonesian Restaurant	Chinese Restaurant	Convenience Store	Auto Dealership	Japanese Restaurant	Fast Food Restaurant	Salon / Barbershop	Coffee Shop
4	BAMBU APUS	Indonesian Restaurant	Soup Place	Seafood Restaurant	Football Stadium	Food Truck	Food Stand	Food Court	French Restaurant	Farm	Food

Jakarta's top 10 venues category



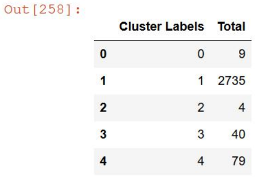
```
neighbourhoods venues dataframe shape is (32, 62)
```

Out[98]:

	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	BARU ILIR	Food Truck	CafÃ©	Park	Food	Field	Fast Food Restaurant	Dumpling Restaurant	Donut Shop	Diner	Dim Sum Restaurant
1	BARU TENGAH	Food & Drink Shop	Diner	Food Court	Food	Field	Fast Food Restaurant	Dumpling Restaurant	Donut Shop	Dim Sum Restaurant	Department Store
2	BELIMBING	Seafood Restaurant	Campground	Tennis Court	CafÃ©	Food	Field	Fast Food Restaurant	Dumpling Restaurant	Donut Shop	Diner
3	BONTANG BARU	Asian Restaurant	Hot Dog Joint	Restaurant	Tennis Court	Campground	Field	Fast Food Restaurant	Dumpling Restaurant	Donut Shop	Diner
4	DAMAI	Indonesian Restaurant	Hotel	Karaoke Bar	Soup Place	Asian Restaurant	Cosmetics Shop	Nightclub	Breakfast Spot	Hardware Store	Miscellaneous Shop

East Kalimantan's top 10 venues category

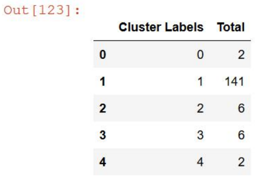
We have some common venue categories in neighborhoods. In this reason I used unsupervised learning **K-means algorithm** to cluster the neighborhoods. K-Means algorithm is one of the most common cluster method of unsupervised learning. First, I will run K-Means to cluster the neighborhoods into **5** clusters, And next is merged table with cluster labels for each neighborhoods in below pictures.



Out[258]:

	Cluster Labels	Total
0	0	9
1	1	2735
2	2	4
3	3	40
4	4	79

Jakarta's K-Means Cluster



Out[123]:

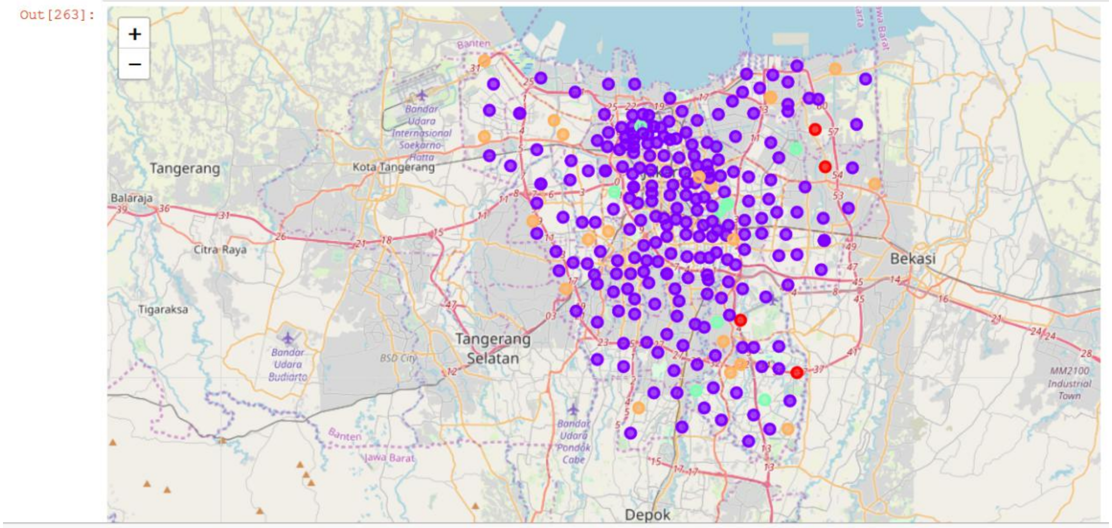
	Cluster Labels	Total
0	0	2
1	1	141
2	2	6
3	3	6
4	4	2

East Kalimantan's K-Means Cluster

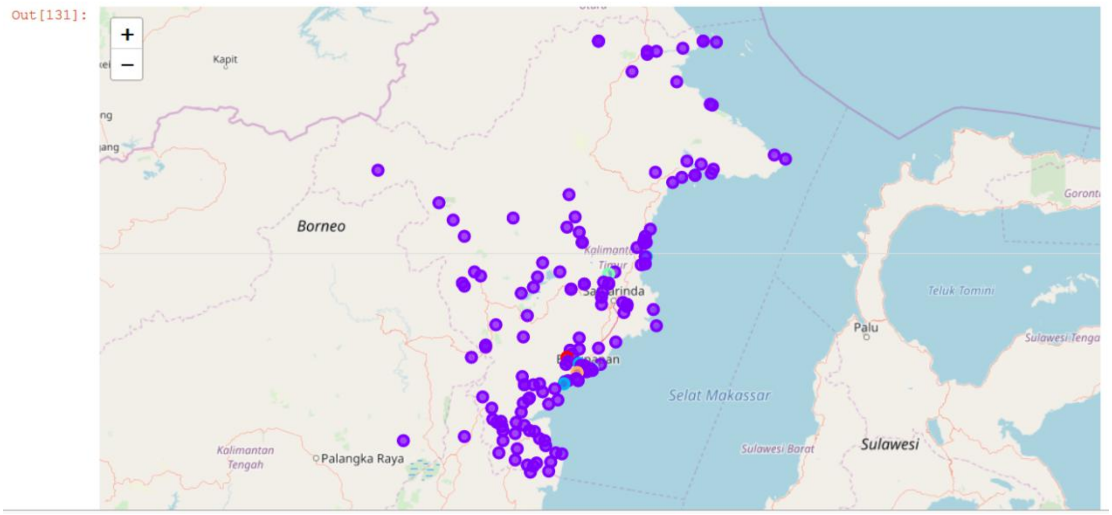
One of my aim was also show the number of top 5 venues information for each neighborhoods on the map. Thus, I grouped each neighborhoods by the number of top 10 venues and I combined those informations in **Join** column.

C. Results

Clustering the neighborhoods using k = 5 gives us a clustered map neighborhoods of Jakarta and East Kalimantan in the below pictures.



Jakarta's Clustered Map



East Kalimantan's Clustered Map

The final comparison between Jakarta and East Kalimantan is in the below table.

No	Item	Jakarta	East Kalimantan
1	Area	661.5 km^2 (255.4 sq mi)	129,066.64 km^2 (49,832.91 sq mi)
2	Population	10,075,310 (2014)	3,619,700 (Mid 2019)
3	Administrative divisions	5 Kota Administratif 1 Kabupaten Administratif	7 regencies 3 cities
4	Master data shape	44 boroughs 267 neighborhoods.	63 boroughs 168 neighborhoods.
5	Venues	(3635, 7) 270 uniques categories.	(154, 7) 61 uniques categories.

D. Discussion

As I mentioned earlier, Jakarta is a big city with a high population density in a narrow area and has more venues. While East Kalimantan is a low population density in wider area and has less venues. Therefore as the future of capital city of Indonesia, there should be more venues on East Kalimantan especially for public veneus.

As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for these cities.

I used the Kmeans algorithm as part of this clustering study. For more detailed and accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled. I ended the study by visualizing the data and clustering information on the Jakarta and East Kalimantan map. In future studies, web or telephone applications can be carried out to direct investors.

E. Conclusion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided. Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

Regards,

Bermansyah DY

F. References:

[1]. Jakarta — Wikipedia
[2]. East Kalimantan - Wikipedia
[3]. Indonesia Postal Code
[4]. Geonames Repository
[5]. Forsquare API
[6]. Google Geocoding API