

IBM – Coursera
Data Science Specialization

Final report – Capstone project
Jakarta Business Clustering

Dionisius Darryl Hermansyah
June 2020

I. Introduction

This report is written to fulfill the capstone project for Coursera IBM Data Science Specialization



(Source: Global Government Forum)

The main goal of this project is to explore and cluster the neighborhoods in Jakarta, Indonesia and also to determine the best location for various business placement. This idea comes from the process of business modeling which often require a strategic placement for their office or headquarter. It is very common for business contractor to do some analysis before setting up their business office in an area. So, which areas in Jakarta is most suitable for a specific kind of business? What are the factors affecting a strategic business placement? The target audience for this project are:

- Potential constructor who wants to start their own business and need a recommendation for setting up their office in Jakarta
- Existing company who wants to renew their place in a different area of neighborhoods in Jakarta
- Anyone who is interested in neighborhood clustering using Python as a data science tools



II. Data Description

The observation target for this project is Jakarta neighborhoods which is located in Indonesia, and It is chosen due to:

- The abundant amount of business company
- The diversity of the places and neighborhoods
- The capital city and the heart of the economy for Indonesia
- The availability of the geo data

The data will be collected from:

- Wikipedia that provides list of Jakarta neighborhoods (https://en.wikipedia.org/wiki/List_of_districts_of_Jakarta)
- FourSquare API that provides surrounding venues of a given coordinates.

The data will be processed as follows:

- Scrape the Wikipedia webpage to obtain the neighborhoods data
- Find the geographic data of the neighborhoods. Both their center coordinates and their border.
- For each neighborhood, pass the obtained coordinates to FourSquare API to explore surrounding venues in a pre-defined radius.
- Count the occurrence of each venue type in a neighborhood.
- Cluster the data.
- Find the best cluster from each cluster that has been created.



III. Methodology

Web scraping

The dataset was collected through web scraping. This project implemented BeautifulSoup which is a library in Python to scrape a list of Jakarta Neighborhoods in Wikipedia (https://en.wikipedia.org/wiki/List_of_districts_of_Jakarta). Furthermore, the library was also used to clean the scraped dataset.

Getting the coordinates

To get all of the coordinates needed for visualizing the neighborhood in a map, the project used a library called GeoPy, which is based on the OpenCage Geocoder API. By implementing this method, we can get the latitude and longitude for each neighborhood in Jakarta.

Data visualization

Folium was used to visualize the data in a map and filled with markers of each neighborhood with a popup-description.

Venues exploration using FourSquare API

To explore all the venues that existed near the neighborhood, the FourSquare API was used to get all the venues' names and categories based on radius to the specific neighborhood.

One-hot encoding

One-hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.

K-Means Clustering

The neighborhood was clustered using the K-Means method and scored using silhouette score. Silhouette Score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

IV. Results

Here are the result of web scraping with BeautifulSoup:

```
Scraping the wikipedia website

In [7]: url = 'https://en.wikipedia.org/wiki/List_of_districts_of_Jakarta'
        res = requests.get(url).text
        soup = BeautifulSoup(res, 'html.parser')

Getting the neighborhood name

n [49]: lis = soup.find_all('li')
        lis = lis[5:51] # Removing the redundant elements

n [61]: neighborhoods = []
        districts = []

        for li in lis:
            li = li.get_text() # Getting the text from <li> elements
            neighborhoods.append(li)

n [66]: for neighborhood in neighborhoods:
        if 'Jakarta' in neighborhood:
            districts.append(neighborhood)
            neighborhoods.remove(neighborhood)

        neighborhoods[:5]

Out[66]: ['Cengkareng', 'Grogol Petamburan', 'Kalideres', 'Kebon Jeruk', 'Kembangan']
```

The coordinates obtained from GeoPy are as follows:

```
Using GeoPy to get the neighborhood's coordinates

In [80]: geolocator = Nominatim(user_agent="jakarta_explore")
        latitudes = []
        longitudes = []

        for neighborhood in df['Neighborhood']:
            location = geolocator.geocode(neighborhood)
            latitudes.append(location.latitude)
            longitudes.append(location.longitude)

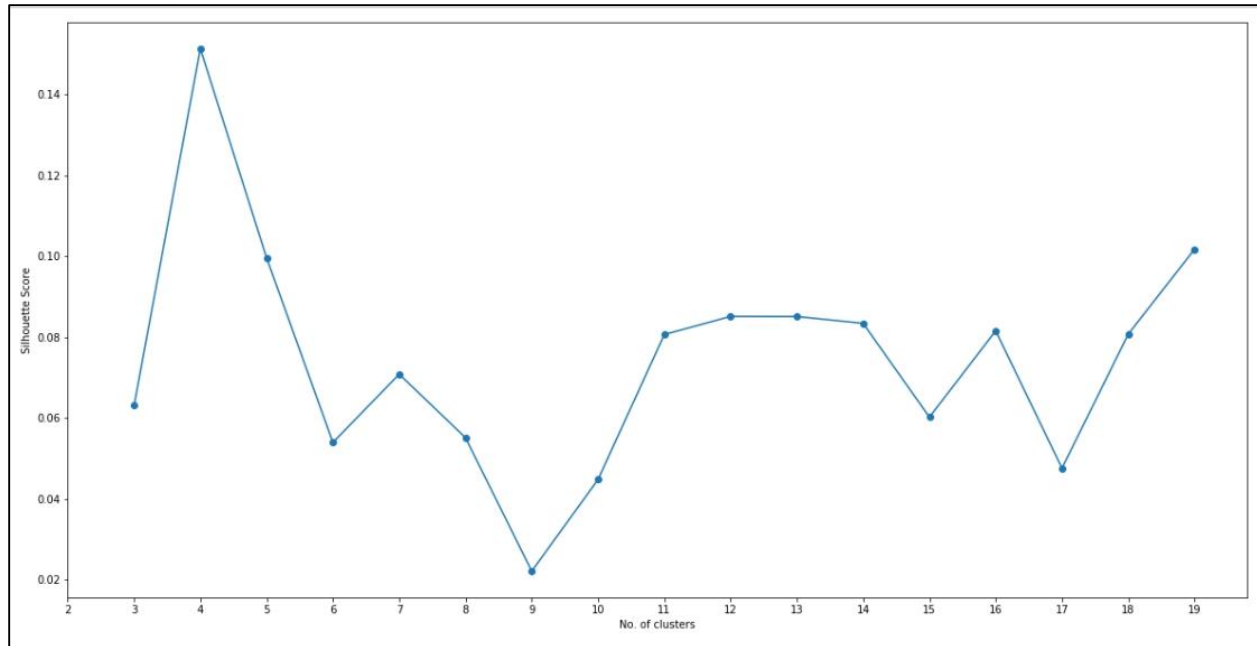
Adding the coordinates to the dataframe

In [87]: df['Longitude'] = longitudes
        df['Latitude'] = latitudes
        df.head()

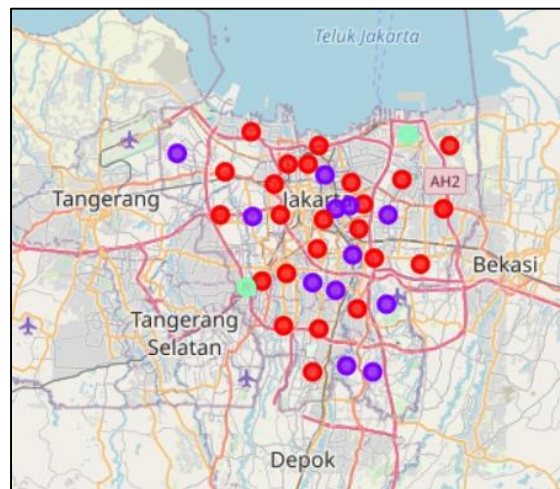
Out[87]:
```

	Neighborhood	Longitude	Latitude
0	Cengkareng	106.744718	-6.152899
1	Grogol Petamburan	106.788317	-6.164188
2	Kalideres	106.701594	-6.137006
3	Kebon Jeruk	106.769725	-6.192572
4	Kembangan	106.740586	-6.191395

After the clustering process, we get that the best scoring was shown when n-clusters = 3.



Furthermore, here is the map of Jakarta neighborhoods based on their own clusters, where cluster 0 is purple, cluster 1 is red, and cluster 2 is blue.



The picture below shows the neighborhoods based on their clusters:

Cluster	Count	Neighbourhoods
0	24	Cengkareng, Grogol Petamburan, Kembangan, Palmerah, Taman Sari, Tambora, Cempaka Putih, Kemayoran, Cilandak, Jagakarsa, Kebayoran Baru, Kebayoran Lama, Pasar Minggu, Setiabudi, Cakung, Duren Sawit, Jatinegara, Kramat Jati, Matraman, Menteng, Cilandak, Kelapa Gading, Pademangan, Penjaringan
1	12	Kalideres, Kebon Jeruk, Johar Baru, Sawah Besar, Senen, Mampang Prapatan, Pancoran, Tebet, Ciracas, Makasar, Pasar Rebo, Pulo Gadung
2	2	Pesanggrahan, Koja

V. Discussion


After the final clustering, we combined all of the dataframe into one final dataframe as follows:

	Neighborhood	Longitude	Latitude	Cluster Labels	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	
0	Cengkareng	106.744718	-6.152899	0	Coffee Shop	Japanese Restaurant	Sandwich Place	Food Truck	Chinese Restaurant	Women's Store	American Restaurant	Airport Terminal	Acehnese Restaurant	A
1	Grogol Petamburan	106.788317	-6.164188	0	Chinese Restaurant	Noodle House	Asian Restaurant	Indonesian Restaurant	Hotel	Pizza Place	Fast Food Restaurant	Coffee Shop	Seafood Restaurant	
2	Kalideres	106.701594	-6.137006	1	Indonesian Restaurant	Bookstore	Optical Shop	Fried Chicken Joint	Food Truck	Convenience Store	Women's Store	American Restaurant	Airport Terminal	f
3	Kebon Jeruk	106.769725	-6.192572	1	Noodle House	Asian Restaurant	Indonesian Restaurant	Concert Hall	Convenience Store	Café	Art Museum	Ice Cream Shop	Coffee Shop	
4	Kembangan	106.740586	-6.191395	0	Asian Restaurant	Chinese Restaurant	Japanese Restaurant	Coffee Shop	Café	Food Court	Clothing Store	Pizza Place	Seafood Restaurant	

According to the dataframe, the most frequent venue that we may look at for each neighborhood is food industries such as restaurant, food court, café, etc. It is recommended for business owner that want to start their businesses to get away from food since there are abundant venues already.

The dataframe was also grouped by the cluster labels into:

Cluster Labels	1st Most Common Venue
0	Arcade
1	Art Gallery
2	Asian Restaurant
3	Bakery
4	Bar
5	Chinese Restaurant
6	Coffee Shop
7	Food Truck
8	Gas Station
9	Hotel
10	Indonesian Meatball Place
11	Indonesian Restaurant
12	Italian Restaurant
13	Japanese Restaurant
14	Park
15	Pharmacy
16	Pizza Place
17	Hotel
18	Indonesian Restaurant
19	Noodle House
20	Pizza Place



By analyzing the clusters and their most common venue. We get that cluster 0 is perfect for most of the businesses including food, health, hospitality, and entertainment, whereas cluster 0 is suitable for food and hospitality. For cluster 2, it is inconclusive since the cluster is only consisting of 2 neighborhood.

All in all, cluster 0 would be the best spot for all kind of businesses to be placed in, that includes 24 neighborhoods: Cengkareng, Grogol Petamburan, Kembangan, Palmerah, Taman Sari, Tambora, Cempaka Putih, Kemayoran, Cilandak, Jagakarsa, Kebayoran Baru, Kebayoran Lama, Pasar Minggu, Setiabudi, Cakung, Duren Sawit, Jatinegara, Kramat Jati, Matraman, Menteng, Cilincing, Kelapa Gading, Pademangan, and Penjaringan.

On the other hand, cluster 1 and 2 are worse than cluster 0, that includes 14 neighborhoods: Kalideres, Kebon Jeruk, Johar Baru, Sawah Besar, Senen, Mampang Prapatan, Pancoran, Tebet, Ciracas, Makasar, Pasar Rebo, Pulo Gadung, Pesanggrahan, and Koja

VI. Conclusion

It is unfortunate that some labels in the model was still inconclusive. Apart from that we can still get some meaningful and logical insights from the result.

We get to know that most of the neighborhoods in Jakarta are placed in cluster 0 which is filled with various businesses, mostly in food industries.

In conclusion, the best clusters for businesses placement in Jakarta are neighborhoods in cluster 0 that contains mostly food, health, hospitality, and entertainment industries.