CAPSTONE PROJECT
THE BATTLE OF NEIGHBORHOOD

Hajid Naufal Atthousi, 2020

Introduction

Background

Jakarta is one of the cities in Indonesia which has a lot of coffee shops. From 2013 until the end of 2018, there have been several coffee shops spreads around every corner of the capital city of Jakarta even from several locations of offices, schools, or campuses.

Business Problem

The client is interested to open a specialty coffee shop in Jakarta. Unfortunately, he has issue on making a decision about the location to open the coffee shop. His first issue is that he wanted to know which place has lesser competition so that he can grow his business in a stable pace without fighting over customer, whether it is battle between coffee shops or other kind of cafes or restaurant. The second concern is that he wanted the place to not very far away from his supplier in the central Jakarta to minimize the time in retrieving the supply from the supplier. Finally, last but not least, he wanted the place to have an adequate population. So, where will I recommend the best place for him to open the coffee shop?

Target interest

Personal client who wants to gain insight about the best location to build a coffee shop in Jakarta according to his concerns.

The Battle of Neighborhoods
CAPSTONE PROJECT

Data acquisition and pre-processing

Data choice

To solve the problem, I need a precise data that can tell the population of each district. Furthermore, the data should also can tell the neighborhood within each district since that data will be used on the last section to see the distance on each neighborhood from the central Jakarta (supplier's place) and the population within the neighborhood. So, I will use the following data:

1. Dataset from Jakarta Open Data. I choose to use this data since it is the most up to date within the site. This data consists of:
   - The name of districts and neighborhoods
   - The spread of population based on gender (Male and Female)
   - The spread of population based on age (from 0 to above 75 years old with 4 years step)
   - The cities, districts and neighborhood of those population's spread
2. Latitude and Longitude from geopy.geocoders package that will be cast on each data
3. Venues list that I can get from real-time foursquare API

Data acquisition, cleaning and pre-processing

- From Jakarta Open Data, I'm going to use jupyter notebook and panda package to sort the data and group it by district. Before I sum the population to get a new column, I will drop the population whose age is in the range 0-4 since those population is rather out from the target market (in case growth hacking is needed).
- After that, I will group the dataframe by district, applying join function on the neighborhood and sum up the population for each district.
- The next step is to cast geocoder.arcgis function to retrieve all the location's latitude and longitude in a single for looping and then append it to the list and make a new column with the list of latitude and longitude of each districts.
- In order to make things easier for later analysis, I will retrieve the approximate distance from the supplier's location for each district by using haversine formula. The supplier's latitude and longitude are at (-6.171009, 106.852772).
- The last step is to get the data of nearby venues by using foursquare API.

Data acquisition, cleaning and pre-processing

Here's how the first five rows of the dataframe looks after I pre-processed It through first 4 steps.

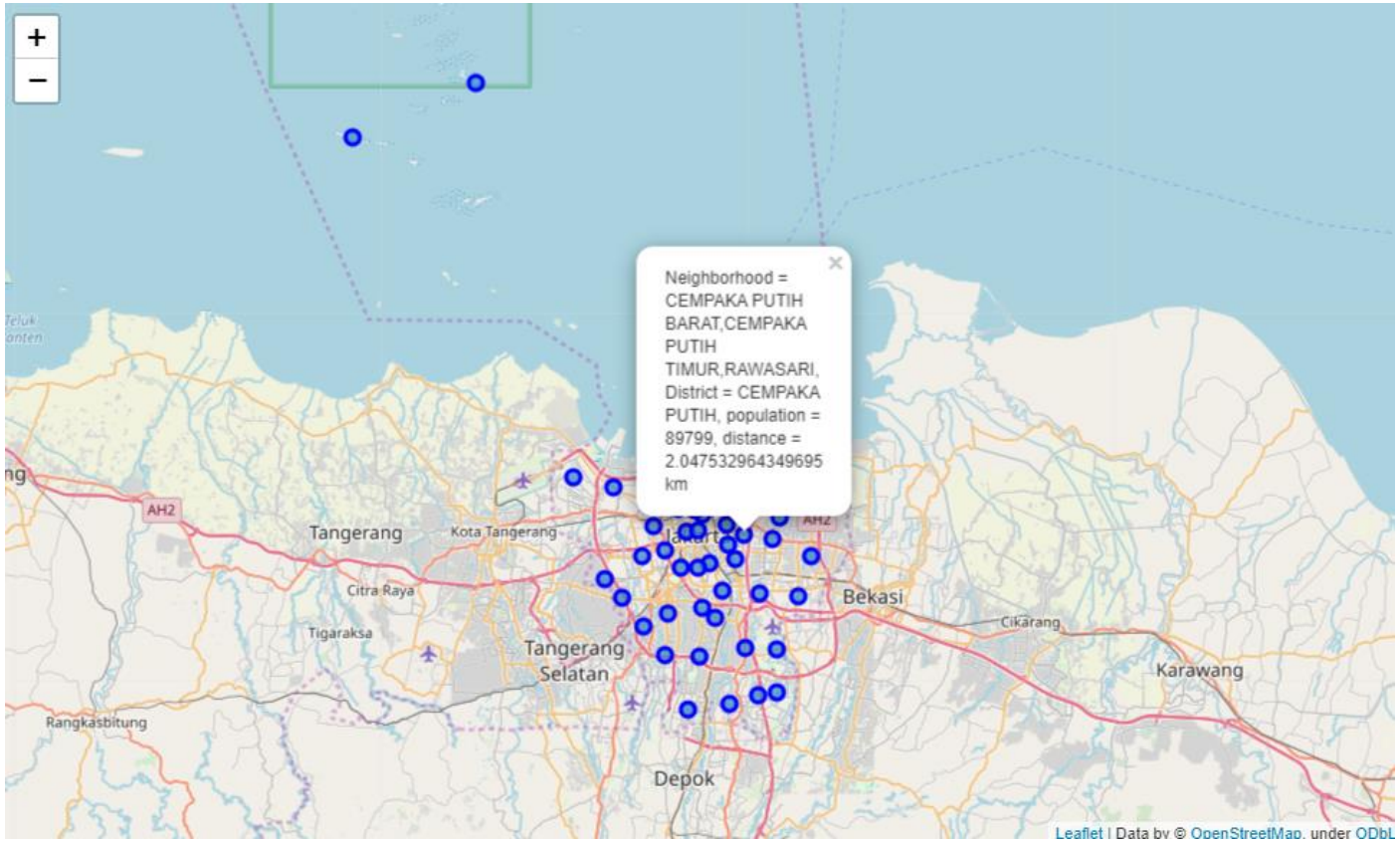
	District	Neighborhood	Total population	Latitude	Longitude	Approx distance
0	CAKUNG	CAKUNG BARAT,CAKUNG TIMUR,JATINEGARA,PENGGILIN...	461622	-6.19623	106.93522	9.536157
1	CEMPAKA PUTIH	CEMPAKA PUTIH BARAT,CEMPAKA PUTIH TIMUR,RAWASARI	89799	-6.17600	106.87060	2.047533
2	CENGKARENG	CENGKARENG BARAT,CENGKARENG TIMUR,DURI KOSAMBI...	457629	-6.13060	106.74559	12.672799
3	CILANDAK	CILANDAK BARAT,CIPETE SELATAN,GANDARIA SELATAN...	189780	-6.29051	106.79491	14.747089
4	CILINCING	CILINCING,KALIBARU,MARUNDA,ROROTAN,SEMPER BARA...	349075	-6.11358	106.94911	12.418460

This is the overview of the total number of venues returned by foursquare API. This dataset will be used to get the average frequency for each venue category within districts in the explanatory data analysis section.

District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
CAKUNG	6	6	6	6	6	6
CEMPAKA PUTIH	23	23	23	23	23	23
CENGKARENG	4	4	4	4	4	4
CILANDAK	19	19	19	19	19	19
CILINCING	1	1	1	1	1	1
CIPAYUNG	1	1	1	1	1	1
CIRACAS	4	4	4	4	4	4
DUREN SAWIT	9	9	9	9	9	9
GAMBIR	19	19	19	19	19	19
GROGOL PETAMBURAN	50	50	50	50	50	50
JAGAKARSA	3	3	3	3	3	3
JATINEGARA	4	4	4	4	4	4
JOHAR BARU	2	2	2	2	2	2
KALI DERES	6	6	6	6	6	6
KEBAYORAN BARU	39	39	39	39	39	39
KEBAYORAN LAMA	4	4	4	4	4	4

Data acquisition, cleaning and pre-processing

In order to make sure that my dataframe can be plotted into a map, I will use folium package to make the map from my current dataframe. The coordinate of Jakarta can be found by using Nominatim.



The Battle of Neighborhoods
CAPSTONE PROJECT


Methodology

Explanatory data analysis

1. One hot encoding and frequency

One hot encoding of venues will be retrieved by using get dummies from panda package. The result of one hot encoding will be used to get average frequency for each venue categories by using mean function. This is the data that will be used for modeling

	District	Accessories Store	Acehnese Restaurant	African Restaurant	Airport	American Restaurant	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Automotive Shop	BBQ Joint	Bakery	Basketb Cou
0	CAKUNG	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	0.0000
1	CEMPAKA PUTIH	0.0	0.043478	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.086957	0.0	0.00	0.0	0.0000
2	CENGKARENG	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.25	0.0	0.0000
3	CILANDAK	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	0.0526
4	CILINCING	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.00	0.0	0.0000

2. Checking top 5 venues for general overview

To get the general overview of the frequency we can use for looping code. This step is used for further analysis when K-Means cluster has finished generating its result. This will improve my understanding of why the cluster is leaned to be labeled that way.

I use the code below to see the top 5 venues within each district to see the general overview

```
In [183]: num_top_venues = 5

for hood in jakarta_grouped['District']:
    print("----"+hood+"----")
    temp = jakarta_grouped[jakarta_grouped['District'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
----CAKUNG----
                venue  freq
0          Restaurant  0.17
1         Video Store  0.17
2  Fast Food Restaurant  0.17
3   Chinese Restaurant  0.17
4        Noodle House  0.17


----CEMPAKA PUTIH----
                venue  freq
0  Indonesian Restaurant  0.17
1                 Café  0.13
2          Pizza Place  0.13
3  Fast Food Restaurant  0.09
4     Asian Restaurant  0.09
```

The Battle of Neighborhoods
CAPSTONE PROJECT

Modeling

The algorithm that I choose is K-Means. Based on the frequency, I found that choosing K-Means is actually preferable in this problem rather than DBScan.

1. Finding the best K with silhouette method

Before I run my K-Means model, I search for the best K first. This can be done by either using elbow method or silhouette method. For this problem, I choose to do silhouette method. The graphs show the best K for this problem after I run the silhouette method. The result shows that the best K is 4.

2. K-Means Algorithm for clustering

After I got the best K, I will pass it to the K-Means algorithm provided by scikit learn cluster. The dataset that will be fitted in this algorithm is the dataset of average frequency of each districts in Jakarta. This will produce cluster labels list for each district.


Results

Results – Clustered Data

In order to make it easier for the client to see the result, I will plot the map that can show the cluster and its description. There will be some steps to achieve my desired map.

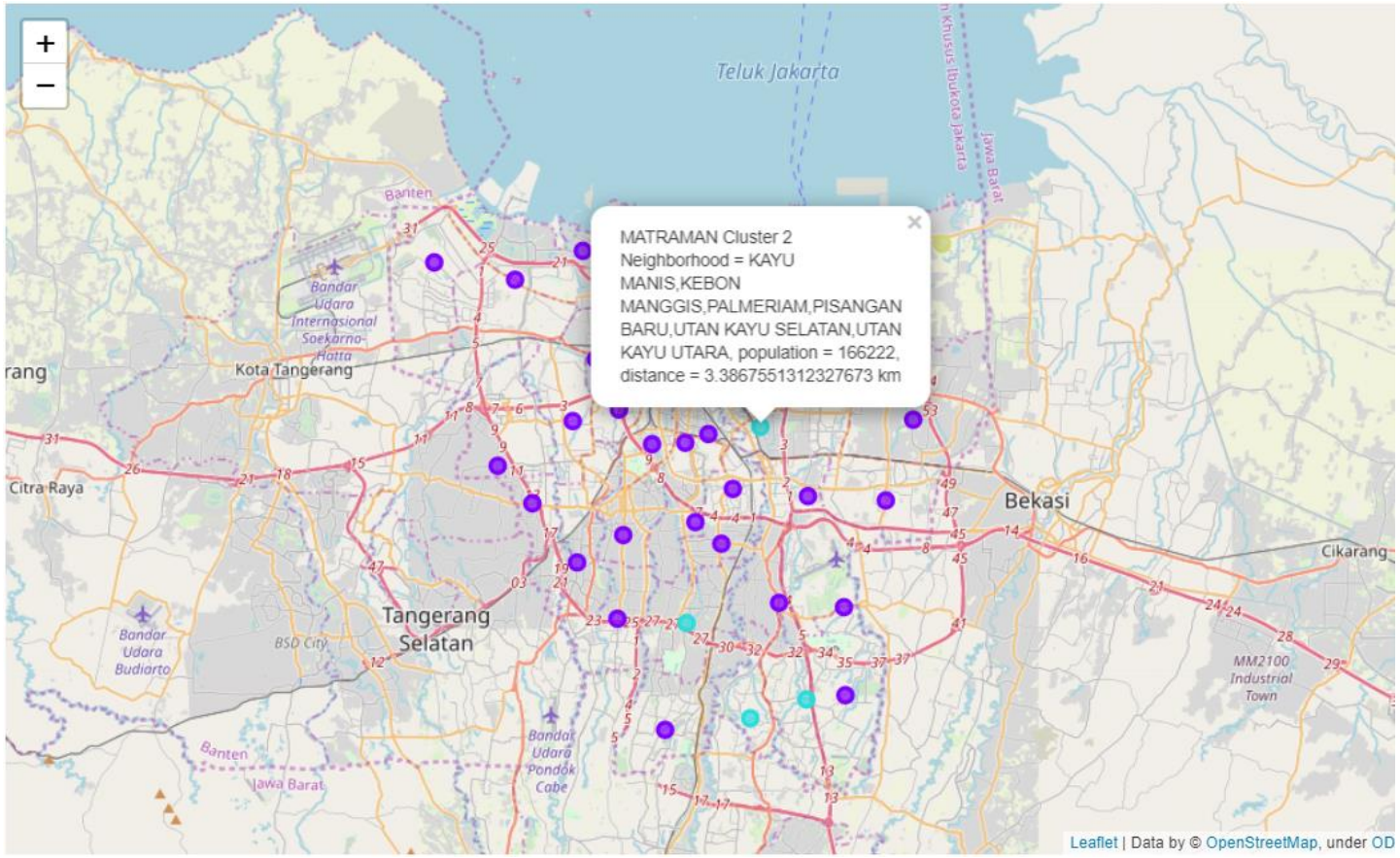
1. Constructing dataframe that shows top common venues for each district
2. Then, append the cluster labels to the dataframe

After that, I will have a dataframe with its cluster labels attached.

	District	Neighborhood	Total population	Latitude	Longitude	Approx distance	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Mo Comm Ven
0	CAKUNG	CAKUNG BARAT,CAKUNG TIMUR,JATINEGARA,PENGGILIN...	461622	-6.19623	106.93522	9.536157	1	Fast Food Restaurant	Video Store	Restaura
1	CEMPAKA PUTIH	CEMPAKA PUTIH BARAT,CEMPAKA PUTIH TIMUR,RAWASARI	89799	-6.17600	106.87060	2.047533	1	Indonesian Restaurant	Café	Pizza Pla
2	CENGKARENG	CENGKARENG BARAT,CENGKARENG TIMUR,DURI KOSAMBI...	457629	-6.13060	106.74559	12.672799	1	Internet Cafe	Vegetarian / Vegan Restaurant	Ca
3	CILANDAK	CILANDAK BARAT,CIPETE SELATAN,GANDARIA SELATAN...	189780	-6.29051	106.79491	14.747089	1	Donut Shop	Pizza Place	Indonesi Restaura
4	CILINCING	CILINCING,KALIBARU,MARUNDA,ROROTAN,SEMPER BARA...	349075	-6.11358	106.94911	12.418460	3	Seafood Restaurant	Women's Store	Farme Mark
5	CIPAYUNG	BAMBU APUS,CEGER,CILANGKAP,CIPAYUNG,LUBANG BUA...	237204	-6.32673	106.90298	18.183007	1	Food Truck	Women's Store	Fast Fo Restaura
6	CIRACAS	CIBUBUR,CIRACAS,KELAPA DUA WETAN,RAMBUTAN,SUSUKAN	263342	-6.32879	106.88475	17.896966	2	Pizza Place	Vegetarian / Vegan Restaurant	Hi Scho
7	DUREN SAWIT	DUREN SAWIT,KLENDER,MALAKA JAYA,MALAKA SARI,PO...	372835	-6.23501	106.92261	10.499846	1	Indonesian Meatball Place	Gym	Pet Sto
8	GAMBIR	CIDENG,DURI PULO,GAMBIR,KEBON KELAPA,PETOJO SE...	94310	-6.17299	106.81571	4.103136	1	Indonesian Restaurant	Camera Store	Coff Sh
9	GROGOL PETAMBURAN	GROGOL,JELAMBAR,JELAMBAR BARU,TANJUNG DUREN SE...	205166	-6.16777	106.78460	7.545080	1	Noodle House	Fast Food Restaurant	Coff Sh
10	JAGAKARSA	CIGANJUR,CIPEDAK,JAGAKARSA,LENTENG AGUNG,SRENG...	297685	-6.34308	106.81745	19.527694	1	Gift Shop	Convenience Store	Coff Sh

Results - Visualization

This map will be the visualization for the cluster map for each district in the dataframe with their own description. This visualization will help the client to easily understand the cluster spread in Jakarta.



The Battle of Neighborhoods
CAPSTONE PROJECT

Analysis and Discussion

Cluster analysis

I can access the district within each cluster in the last dataframe to see the result. Thanks to the silhouette method the number of clusters are 4 with no empty nodes (label 0,1,2,3). Here's the district in cluster label 0,1,2 and 3. (please note I only call the district column and its top venues)

cluster label 0

	District	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
42	TANJUNG PRIOK	Asian Restaurant	Donut Shop	Women's Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand

cluster label 1 (first 5 rows)

	District	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
0	CAKUNG	Fast Food Restaurant	Video Store	Restaurant	Donut Shop	Noodle House	Chinese Restaurant	Women's Store
1	CEMPAKA PUTIH	Indonesian Restaurant	Café	Pizza Place	Noodle House	Fast Food Restaurant	Asian Restaurant	Indonesian Meatball Place
2	CENGKARENG	Internet Cafe	Vegetarian / Vegan Restaurant	Café	BBQ Joint	Women's Store	Fish & Chips Shop	French Restaurant
3	CILANDAK	Donut Shop	Pizza Place	Indonesian Restaurant	Farmers Market	Sandwich Place	Food Court	Bookstore
5	CIPAYUNG	Food Truck	Women's Store	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Stand	Food Court

cluster label 2

	District	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
6	CIRACAS	Pizza Place	Vegetarian / Vegan Restaurant	High School	Women's Store	Farmers Market	Food Truck	Food Stand
26	MATRAMAN	Pizza Place	Women's Store	Farmers Market	French Restaurant	Food Truck	Food Stand	Food Court
31	PASAR MINGGU	Gas Station	Pizza Place	Farmers Market	French Restaurant	Food Truck	Food Stand	Food Court
32	PASAR REBO	Diner	Pizza Place	Women's Store	Fast Food Restaurant	French Restaurant	Food Truck	Food Stand

cluster label 3

	District	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
4	CILINCING	Seafood Restaurant	Women's Store	Farmers Market	French Restaurant	Food Truck	Food Stand	Food Court

The Battle of Neighborhoods
CAPSTONE PROJECT

Cluster analysis – frequency check

Why there are so many districts labeled in cluster label 1? The answer to that question is within the frequency. To progress further, I will pick cluster label 0, 2 and 3 since it seems those clusters will have less competition if my client wants to open a specialty coffee shop. The frequency check within these clusters hopefully will show the general idea of why the clusters leaned to labeled that way.

- Notice that cluster 2 (CIRACAS, MATRAMAN, PASAR MINGGU, PASAR REBO) are leaned to be clustered to one of its most recurring venue which is pizza place
- While there are some labels in cluster 1 that also has pizza place, the frequency might differ in the second most recurring venue (or perharps the first).
- On the other hand, the CILINCING and TANJUNG PRIOK district were also in different cluster, if you notice their first and second venue frequency were quite unique from other clusters.

```
num_top_venues = 7

for hood in freq_check['District']:
    print("----"+hood+"----")
    temp = freq_check[freq_check['District'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
----CILINCING----
                  venue  freq
0     Seafood Restaurant   1.0
1      Accessories Store   0.0
2              Pet Store   0.0
3          Noodle House   0.0
4                Office   0.0
5  Padangnese Restaurant   0.0
6                Palace   0.0


----CIRACAS----
                       venue  freq
0                Pizza Place  0.50
1                High School  0.25
2  Vegetarian / Vegan Restaurant  0.25
3                       Park  0.00
4                 Nightclub  0.00
5               Noodle House  0.00
6                     Office  0.00


----MATRAMAN----
                  venue  freq
0           Pizza Place   1.0
1      Accessories Store   0.0
2                   Park   0.0
3              Nightclub   0.0
4          Noodle House   0.0
5                Office   0.0
6  Padangnese Restaurant   0.0


----PASAR MINGGU----
                  venue  freq
0           Gas Station   0.5
1           Pizza Place   0.5
2                   Park   0.0
3              Nightclub   0.0
4          Noodle House   0.0
5                Office   0.0
6  Padangnese Restaurant   0.0


----PASAR REBO----
                  venue  freq
0                 Diner   0.5
1           Pizza Place   0.5
2              Pet Store   0.0
3          Noodle House   0.0
4                Office   0.0
5  Padangnese Restaurant   0.0
6                Palace   0.0


----TANJUNG PRIOK----
                  venue  freq
0       Asian Restaurant   0.5
1           Donut Shop   0.5
2              Pet Store   0.0
3          Noodle House   0.0
4                Office   0.0
5  Padangnese Restaurant   0.0
6                Palace   0.0
```

The Battle of Neighborhoods
CAPSTONE PROJECT

Cluster analysis – Label description

If I want to make a descriptive label from the result and analysis, it would be:

different description for each label, it would be:

- Label 0 : Districts with moderate level competition with Asian and Donut shop as its main competitor
- Label 1 : Districts with moderate to high level competition with various unique venues as its main competitor
- Label 2 : Districts with low to moderate level competition with pizza place as its main competitor
- Label 3 : Districts with low level competition with seafood restaurant as its main competitor.

Picking suitable place – distance

From the results, I can see from the district MATRAMAN (cluster 2) and CILINCING (cluster 3) has less competition. The other district from cluster 2 will be dropped, since from the map I can see that district MATRAMAN was close to the supplier place compared to other districts in the same cluster. The next step is to compare the distance between MATRAMAN and CILINCING from the supplier. Here's the comparison between the two of them.

:

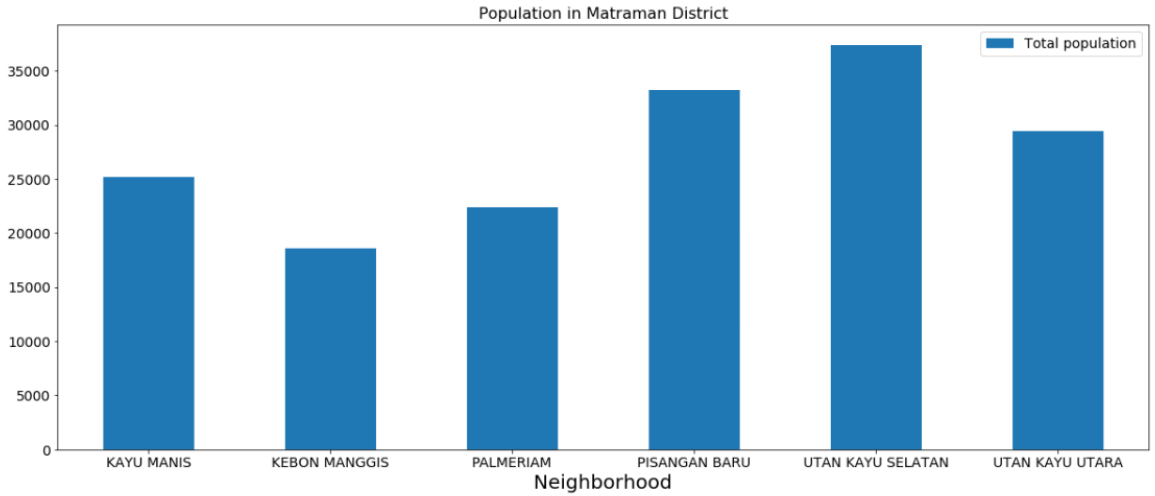
	District	Neighborhood	Total population	Latitude	Longitude	Approx distance	Cluster Labels
0	CILINCING	CILINCING,KALIBARU,MARUNDA,ROROTAN,SEMPER BARA...	349075	-6.11358	106.94911	12.418460	3
1	MATRAMAN	KAYU MANIS,KEBON MANGGIS,PALMERIAM,PISANGAN BA...	166222	-6.19983	106.86268	3.386755	2

Picking suitable place – population

We have a clear winner which is MATRAMAN district. After this, I will move on to the client's third concern which is the place with adequate population. By iterating the very first dataframe again, I can obtain the population within each neighborhood of MATRAMAN district.

	Neighborhood	Total population
0	KAYU MANIS	25205
1	KEBON MANGGIS	18624
2	PALMERIAM	22415
3	PISANGAN BARU	33192
4	UTAN KAYU SELATAN	37406
5	UTAN KAYU UTARA	29380

In order for our client to get better understanding with ease, I will use bar graph to plot the result above so that the comparison can be interpreted visually.



The Battle of Neighborhoods
CAPSTONE PROJECT

Conclusion & Future DIrection

Conclusion

In this project, I have analyzed the frequency of venues within each district in Jakarta. I used the K-Means algorithm to make clusters of those districts. This algorithm is very useful for clustering and plotting a cluster map in order to help the client to gain better understanding of the market competition within each district.

From the results, I will recommend the client to open a specialty coffee shop in UTAN KAYU SELATAN neighborhood which resides in MATRAMAN district. The reasons are:

- MATRAMAN district has less competition compared with other districts.
- MATRAMAN district is closer to the supplier's place compared with other district that also has less competition.
- UTAN KAYU SELATAN neighborhood in MATRAMAN district is the recommended place to open the specialty coffee shop because the population within that area is the highest compared with other neighborhoods in the MATRAMAN district.

Future Direction

From this project, there are some improvements that can be made to gain a better model and analysis:

- The analysis above will have different results if you use Google Maps API instead. Personally, I think Gmaps has more comprehensive data set of Indonesia compared with foursquare, but the price is too expensive if you just want to do a one time project like this.
- If, somehow, you use google maps API and see the results have few differences in densities and you want to have more accurate results (to see whether there is cluster within clusters), DBScan might be preferred to solve it.
- Elbow method can also be used to retrieve optimum K in Kmeans, this may produce slightly different result but it is worth to try. You can also set the number iteration in the KMeans function, the default is 10. If you perhaps want to play with the code, you can tweak this variable alongside the random state.

Thank You