Capstone Project - The Battle of Neighborhoods Report

----Kuncoro Aji Nugroho----

1. Introduction

1.1 Background

In the past decade, the lifestyle of urban people has changed with the trends and habits of drinking coffee. Coffee, which was ancient, is identical to drinks commonly used by older men, now women and men of all ages are accustomed to drinking coffee. And not just enjoying coffee, but many people are looking for a place to drink coffee. The coffee shop has finally become a cool hangout with an internet connection while enjoying a variety of steeping coffee beans. This coffee drinking trend will become a big business opportunity. The business world is starting to work on places that serve specialty coffee.

Although the Indonesian people are not addicted to coffee, which means they have to drink every day, like in Melbourne. And the Coffe shop industry is still relatively new, but in big cities like Jakarta, Coffe shop has the opportunity to get a gross profit of Rp 100 million to Rp 1 billion. However, getting into the business world is not as easy as one might imagine.

If you already have the capital to open a coffee shop, then you must have the courage, start designing strategies and seeing the market. If you have long been in love with coffee and a hobby of drinking coffee, it means you can start a business with the right passion. Therefore I try to practice my learning at Coursera to answer relevant questions, namely designing strategies to determine which areas are suitable for opening coffee shops.

1.2 Problem

Finding data about the area and postcode in South Jakarta is a challenge that must be resolved. The price of renting a place to determine the exact location of a coffee shop is also one of the problems that must be resolved.

1.3 Interest

I believe this is a relevant challenge with a valid question for anyone who wants to open a coffee shop and determine the right location. The same methodology can be applied according to demands as applicable. This case also applies to anyone interested in exploring starting or finding new business in any city. Finally, this can also serve as a good practical exercise for developing Data Science skills.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

1. The data acquired for this project is a combination of data from two sources. The first data source of data is scraped from a wikipedia page that contains the list of Neighboorhod East Jakarta ---> https://id.wikipedia.org/wiki/Daftar_kecamatan_dan_kelurahan_di_Kota_Administrasi_Jakarta_Selatan. This page contains additional information about the boroughs, the following are the columns:
    - Kelurahan : Name of the urban village
    - Kecamatan : Name of the sub-district
    - Kota / Provinsi : Name of the province
2. The Second data source is the list of Logitude & Latitude from website longlat.net, and list of postcode from Wikipedia in East Jakartan, the following are columns:
    - Kelurahan : Name of the urban village.
    - PostCode : Number of the posccode area.
    - Latitude : Latitude of the urban village.
    - Longitude : Longitude of the urban village.
    -

2.2 Data Cleaning

The data preparation for each of the two sources of data is done separately. From the East Jakarta data on Wikipedia, The first data is scraped from a wikipedia page using the Beautiful Soup library in python. Using this library we can extract the data in the tabular format as shown in the website. After the web scraping, string manipulation is required to get the10 districts and 65 villages with postal codes 12110 to 12980.

	Kelurahan	Kecamatan	Kota
1	Cilandak Barat	Cilandak	Jakarta Selatan
2	Cipete Selatan	Cilandak	Jakarta Selatan
3	Gandaria Selatan	Cilandak	Jakarta Selatan
4	Lebak Bulus	Cilandak	Jakarta Selatan
5	Pondok Labu	Cilandak	Jakarta Selatan

Fig 2.1 East Jakarta data after preprocessing

The Second data source is the list of Logitude & Latitude from website longlat.net, and list of postcode from Wikipedia in East Jakartan.

	PostCode	Kecamatan	Kelurahan	Latitude	Longitude
0	12430	Cilandak	Cilandak Barat	-6.288289	106.796765
1	12410	Cilandak	Cipete Selatan	-6.271827	106.804876
2	12420	Cilandak	Gandaria Selatan	-6.272557	106.794845
3	12440	Cilandak	Lebak Bulus	-6.301837	106.779642
4	12450	Cilandak	Pondok Labu	-6.308832	106.797495
5	12620	Jagakarsa	Ciganjur	-6.335664	106.807269
6	12630	Jagakarsa	Cipedak	-6.351934	106.801658
7	12620	Jagakarsa	Jagakarsa	-6.324892	106.819768
8	12610	Jagakarsa	Lenteng Agung	-6.323122	106.836247
9	12640	Jagakarsa	Srengseng Sawah	-6.344935	106.826291
10	12530	Jagakarsa	Tanjung Barat	-6.306542	106.849476
11	12150	Kebayoran Baru	Cipete Utara	-6.262062	106.804211
12	12140	Kebayoran Baru	Gandaria Utara	-6.257878	106.790736
13	12120	Kebayoran Baru	Gunung	-6.236548	106.791315
14	12130	Kebayoran Baru	Kramat Pela	-6.245464	106.792431
15	12160	Kebayoran Baru	Melawai	-6.245165	106.802194

3. Modelling

Using the final dataset containing the neighborhoods in Kingston upon Thames along with the latitude and longitude, we can find all the venues within a 500 meter radius of each neighborhood by connecting to the Foursquare API. This returns a json file containing all the venues in each neighborhood which is converted to a pandas dataframe. This data frame contains all the venues along with their coordinates and category.

	Kelurahan	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cilandak Barat	-6.288289	106.796765	Total Buah Segar	-6.287626	106.795414	Farmers Market
1	Cilandak Barat	-6.288289	106.796765	Roti Bakar Wiwied	-6.289398	106.795378	Sandwich Place
2	Cilandak Barat	-6.288289	106.796765	Maxima Fitness	-6.287644	106.795485	Gym
3	Cilandak Barat	-6.288289	106.796765	Mars Kitchen	-6.287163	106.795424	Café
4	Cilandak Barat	-6.288289	106.796765	TOUS les JOURS	-6.291655	106.799730	Bakery

One hot encoding is done on the venues data. (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). The Venues data is then grouped by the

Neighborhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighborhoods. To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster the 15 neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

4. Results

After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters and drop coffe and caffe venue because we will search best location for open coffe shop . Looking into the neighborhoods in the first cluster.

	Cluster Labels	Kelurahan	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
3	0	Ciganjur	Bakery	Arcade	Grocery Store	Convenience Store	Flea Market	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand
10	0	Cipulir	Food Truck	Asian Restaurant	Sporting Goods Shop	Department Store	Hotel	Restaurant	Hobby Shop	Supermarket	Fish & Chips Shop	French Restaurant
26	0	Kebayoran Lama Utara	Golf Course	Donut Shop	Art Gallery	Music Venue	Tailor Shop	Women's Store	Food	Fried Chicken Joint	French Restaurant	Food Truck
40	0	Pasar Manggis	Betawinese Restaurant	Bakery	Food Truck	Flea Market	Market	Grocery Store	Electronics Store	Hotel	Art Gallery	Pizza Place
41	0	Pasar Minggu	Noodle House	Asian Restaurant	Bakery	Department Store	Indonesian Meatball Place	Donut Shop	Dumpling Restaurant	Convenience Store	Food Truck	Steakhouse
49	0	Petukangan Utara	Asian Restaurant	Noodle House	Food Court	Food & Drink Shop	Women's Store	Flea Market	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Truck

Cluster 1

Cluster one contains some food truck and food stand that can be sell some coffe drink, lets saw another cluster.

	Cluster Labels	Kelurahan	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
1	1	Bintaro	Indonesian Restaurant	Food Truck	Bakery	Pizza Place	Restaurant	Clothing Store	Burger Joint	Motorcycle Shop	Asian Restaurant	Hospital
2	1	Bukit Duri	Indonesian Restaurant	Asian Restaurant	Flea Market	Garden	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food Court
9	1	Cipete Utara	Bakery	Snack Place	Food Truck	Dessert Shop	Playground	Dog Run	Donut Shop	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant
14	1	Grogol Selatan	Indonesian Restaurant	Soup Place	Food Court	Noodle House	Flea Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food & Drink Shop
18	1	Jagakarsa	Acehnese Restaurant	Food Truck	Convenience Store	Pharmacy	Fish & Chips Shop	Fried Chicken Joint	French Restaurant	Food Stand	Food Court	Food & Drink Shop
19	1	Jati Padang	Breakfast Spot	Japanese Restaurant	Indonesian Restaurant	Music Store	Asian Restaurant	Convenience Store	Food Truck	Fast Food Restaurant	High School	Food & Drink Shop
27	1	Kebon Baru	Indonesian Meatball Place	Grocery Store	Restaurant	Noodle House	Bridal Shop	Indonesian Restaurant	Housing Development	Hot Dog Joint	French Restaurant	Food Stand
34	1	Manggarai	Arcade	Hotel	Light Rail Station	Women's Store	Flea Market	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand
35	1	Manggarai Selatan	Clothing Store	Indonesian Restaurant	Arcade	Diner	Indian Restaurant	Garden	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand
39	1	Pancoran	Indonesian Restaurant	Hotel	Acehnese Restaurant	Office	Food Stand	Record Shop	Pizza Place	Food Court	Restaurant	Gas Station
44	1	Pela Mampang	Indonesian Meatball Place	Airport	Music Venue	Food Truck	Snack Place	Hobby Shop	Noodle House	Food	Flea Market	French Restaurant
45	1	Pengadegan	BBQ Joint	Soup Place	Food Truck	Car Wash	Flea Market	Fried Chicken Joint	French Restaurant	Food Stand	Food Court	Food & Drink Shop
46	1	Pesanggrahan	Indonesian Restaurant	Arcade	Bus Station	Food Truck	Food	Garden	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Stand
48	1	Petukangan Selatan	Indonesian Meatball Place	Soccer Stadium	Soccer Field	Noodle House	Flea Market	French Restaurant	Food Truck	Food Stand	Food Court	Food & Drink Shop

Cluster 2

Looking into the neighborhoods in the second, third and fifth clusters, we can see these clusters have only one neighborhood in each. This is because of the unique venues in each of the neighborhoods, hence they couldn't be clustered into similar neighborhoods .

Cluster 3

```
In [36]: 1 Location_Recomendation.loc[Location_Recomendation['Cluster Labels'] == 2]
```

Out[36]:

	Cluster Labels	Kelurahan	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
59	2	Srengseng Sawah	College Residence Hall	Women's Store	Flea Market	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food Court	Food & Drink Shop

Cluster 4

```
In [37]: 1 Location_Recomendation.loc[Location_Recomendation['Cluster Labels'] == 3]
```

Out[37]:

	Cluster Labels	Kelurahan	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
15	3	Grogol Utara	Noodle House	Women's Store	Dessert Shop	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food Court	Food & Drink Shop

Cluster 5

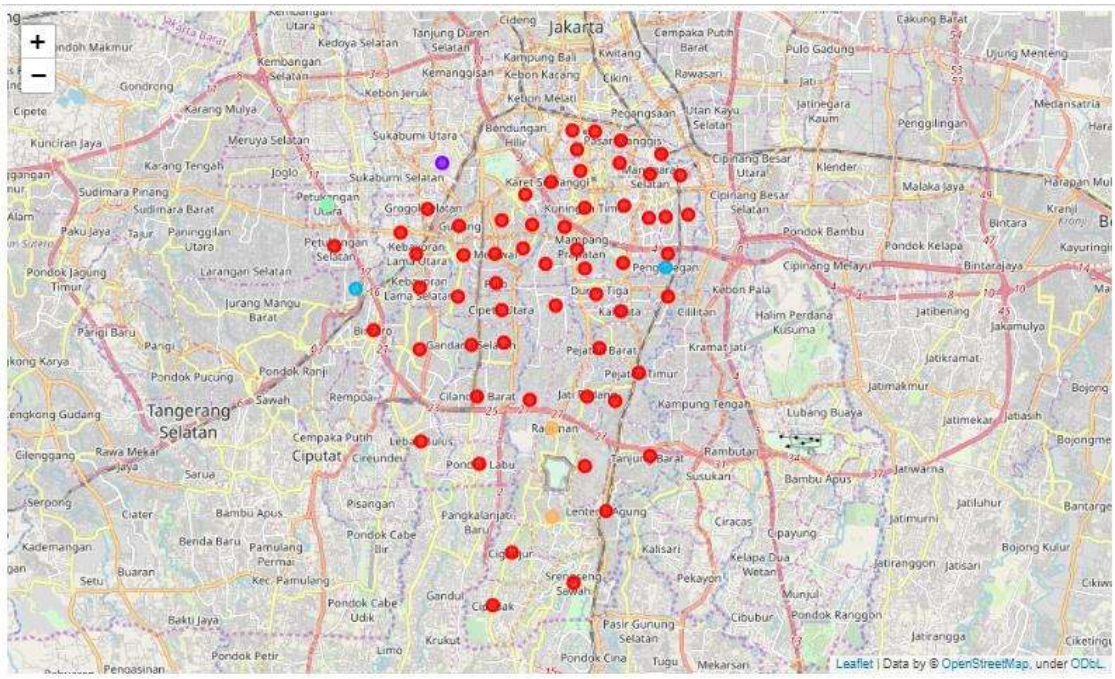
```
In [38]: 1 Location_Recomendation.loc[Location_Recomendation['Cluster Labels'] == 4]
```

Out[38]:

	Cluster Labels	Kelurahan	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
7	4	Cipedak	Grocery Store	Department Store	Pharmacy	Farmers Market	Flea Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food Court
50	4	Pondok Labu	Grocery Store	Fast Food Restaurant	Rental Car Location	Javanese Restaurant	Flea Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Stand	Food Court

Cluster 3-5

Visualising the clustered neighborhoods on a map using the folium library.



Each cluster is color coded for the ease of presentation, we can see that majority of the neighborhood falls in the red cluster which is the first cluster. Three neighborhoods have their own cluster (Blue, Purple and Yellow), these are clusters two three and five. The green cluster consists of two neighborhoods which is the 4th cluster.

5. Discussion

The purpose of this project is to help people or coffee shop owners who want to open a new shop in an area by comparing the number of coffee shops in the area. The right area to open coffee shops for the first time is in clusters 3, 4 or 5 because the venue categories such as Food Trucks, Food Stands and Food Court indirectly there is a possibility of coffee menus in the venue and still a little even mostly above the 5th Most Common Venue. But if you want to open branches or add franchises to clusters 1 and 2 can be a consideration.

6. Conclusion

This project helps one get a better understanding of the environment in relation to the most suitable place to open coffee shops. The future of this project includes considering other factors such as the cost of renting a place, the price of land to open a new coffee shop or even the work and salaries of each person in the area to be able to more accurately determine the price of coffee to be sold.