Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

I WAYAN NADIANTARA

I Wayan Nadiantara
BANDUNG | INDONESIA

TABLE OF CONTENTS

INTRODUCTION 2

BACKGROUND 2
PROBLEM 2
INTEREST 2

DATA 3

SOURCE 3
CLEANING AND WRANGLING 3

METHODOLOGY AND ANALYSIS 6

INITIAL ANALYSIS AND EXPECTATION 6
CLUSTERING AND VISUALIZATION 7
CLUSTER 0 7
CLUSTER 1 8
CLUSTER 2 8
CLUSTER 3 9

RESULT AND DISCUSSION 10

CONCLUSION 10

Capstone Project - The Battle of the Neighborhoods (Week 2)

I Wayan Nadiantara

26 July 2020

1. Introduction

1.1 Background

Every year, the city center of Bandung, Indonesia is flocked by students who try to get accomodation to study at Bandung Institut of Technology. Let say there is a new student, name him Budi. Budi is not from Bandung so he needs to get a flat or room there. He wants a neighborhood that is not quite far from the campus and around good, cheap restaurants or cafes. How could we help him to get a right neighborhood? So, let's go to the problem.

1.2 Problem

1. How could I determine which neighborhood is the best for Budi to live in Bandung?
2. Which neighborhood is the best for him?
3. What are the advantages and disadvantages of each neighborhood projected by the data?

1.3 Interest

This project targets everyone who wants to get the "right" neighborhood near Bandung Institut of Technology. This model could also easily developed to analyize other neighborhood in certain area with the given parameter.

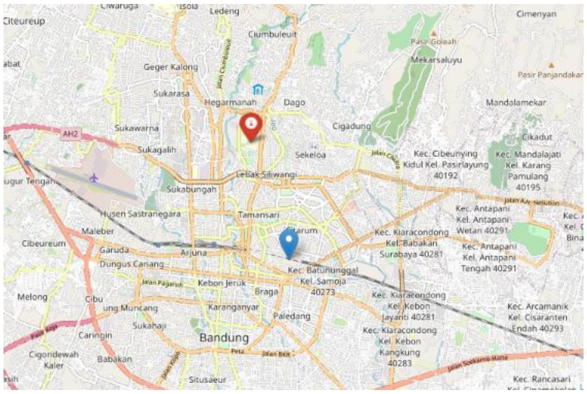


Figure 1. *Location of Bandung Institut of Technology (red balloon) relative to city center (blue balloon)*

2. Data

2.1 Data source

There are a lot of restaurants and cafes in the city, but I must give him the best possible neighborhood to live based on the data available. Ideally, I should use the Google places API to get better data, but due to the paywall limit regarding that API, my suggestion will be based on a combination of data from the Foursquare API and the Zomato API.

Foursquare City Guide, commonly known as Foursquare, is a local search-and-discovery mobile app developed by Foursquare Labs Inc. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history. Zomato is a restaurant aggregator and food delivery start-up which provides information, menus, and user-reviews of restaurants as well as food delivery options from partner restaurants in select cities. As of 2019, the service is available in 24 countries and in more than 10,000 cities, including Bandung.

I used Zomato’s (https://developers.zomato.com/api) and Foursquare’s API (https://foursquare.com/developers/apps) to retrieve the data that I need and then I combine it. From Zomato’s API, I got a data which consist of these properties of a restaurant:

1. Name
2. Address
3. Rating
4. Price range
5. Price for two
6. Latitude
7. Longitude

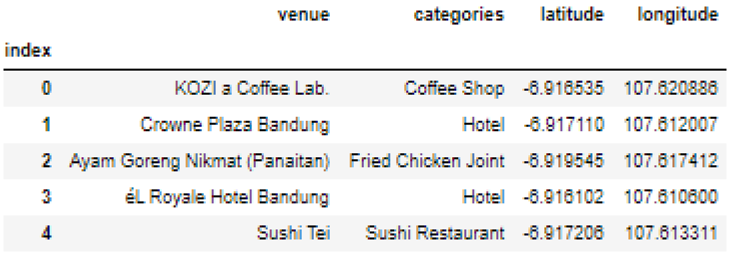
Meanwhile, from Foursquare I a got data, such as:

1. Name
2. Category
3. Latitude
4. Longitude

From all of those data, I can create clustering metrics and make some clusters with the K-Means algorithm. From each cluster, I will retrieve any points that reflect the advantages or disadvantages, visualize it, and then tell Budi as a suggestion.

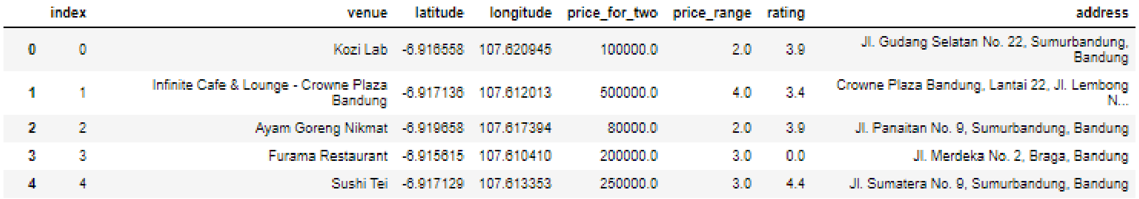
2.2 Cleaning and Wrangling

How do we use that data? Or why do we need two data from Foursquare and Zomato? Foursquare gave the category of each venues, merge it with Zomato which is a restaurant aggregator gave me only results that are food and beverages related category. Table 1 is the sample data that I get from using Foursquare's API and table 2 is from Zomato's API. Notice that I keep the index from each table as a Primary Key, this index will make it easier to merge or joins both tables.



	venue	categories	latitude	longitude
index				
0	KOZI a Coffee Lab.	Coffee Shop	-6.916535	107.620886
1	Crowne Plaza Bandung	Hotel	-6.917110	107.612007
2	Ayam Goreng Nikmat (Panaitan)	Fried Chicken Joint	-6.919545	107.617412
3	éL Royale Hotel Bandung	Hotel	-6.916102	107.610600
4	Sushi Tei	Sushi Restaurant	-6.917206	107.613311

Table 1*. Data sample from Foursquare's API.*



	index	venue	latitude	longitude	price_for_two	price_range	rating	address
0	0	Kozi Lab	-6.916558	107.620945	100000.0	2.0	3.9	Jl. Gudang Selatan No. 22, Sumurbandung, Bandung
1	1	Infinite Cafe & Lounge - Crowne Plaza Bandung	-6.917136	107.612013	500000.0	4.0	3.4	Crowne Plaza Bandung, Lantai 22, Jl. Lembong N...
2	2	Ayam Goreng Nikmat	-6.919658	107.617394	80000.0	2.0	3.9	Jl. Panaitan No. 9, Sumurbandung, Bandung
3	3	Furama Restaurant	-6.915615	107.610410	200000.0	3.0	0.0	Jl. Merdeka No. 2, Braga, Bandung
4	4	Sushi Tei	-6.917129	107.613353	250000.0	3.0	4.4	Jl. Sumatera No. 9, Sumurbandung, Bandung

Table 2*. Data sample from Zomato's API.*

If we check the data from Zomato's API, we will find that the latitude and longitude columns are object type variables. We need to change it as we need to use it to overlay each venue point on the folium map. After the cleaning, we can visualize each venue on the folium map, it showed by figure 2 and 3.

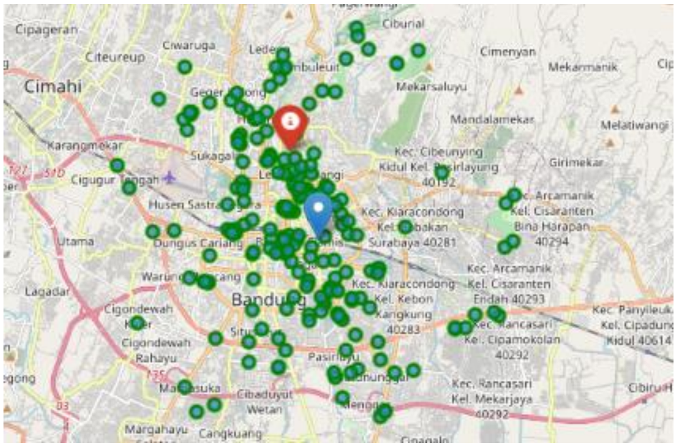

Cipageran
Citeureup
Ciwaruga
Cimahi
Ciburial
Cimenyan
Mekarmanik
Mekarsaluyu
Mandalamekar
Melatiwangi
Karangmekar
Sukagalih
Cigugur Tengah
Husen Sastranegara
Kec. Cibeunying Kidul Kel. Pasirlayung 40192
Girimekar
Kec. Kiaracondong Kel. Babakan Surabaya 40281
Kec. Arcamanik Kel. Cisaranten Bina Harapan 40294
Utama
Dungus Cariang
Lagadar
Bandung
Kec. Kiaracondong Kel. Kebon Kangkung 40283
Kec. Arcamanik Kel. Cisaranten Endah 40293
Kec. Panyileukan Kel. Cipadung Kidul 40614
Cigondewah Kaler
Cigondewah Rahayu
Kec. Rancasari Kel. Cipamokolan 40292
Pasirluyu
Kec. Rancasari Kel. Mekarjaya 40292
Margasuka
Cibaduyut Wetan
Margahayu
Cangkuang
Cibiru

Figure 2. *Location of each venues that retrieved using Foursquare API relative to the Bandung Institut of Technology (red balloon) and city center (blue balloon).*

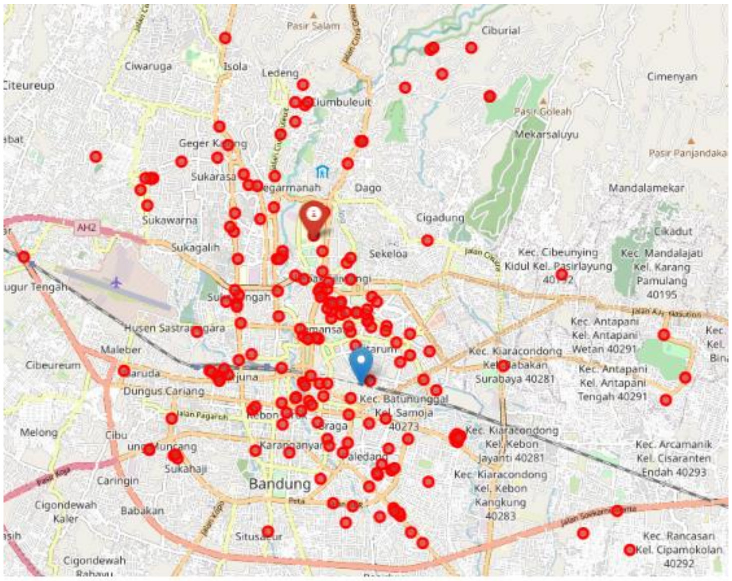
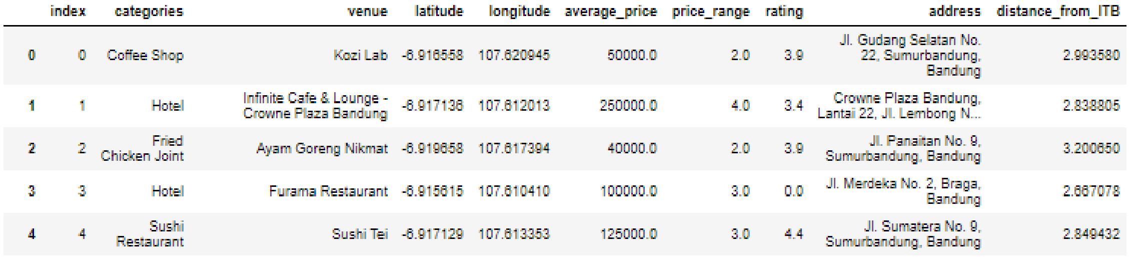


Figure 3. *Location of each venues that retrieved using Zomato's API relative to the Bandung Institut of Technology (red balloon) and city center (blue balloon).*

Notice that the data from Zomato's API contain less numbers of venues compared to data from Foursquare's API. It is because data from Foursquare's API does not only containing venues that related to food and beverages.

From the introduction, I already explained that the distance from the University's campus will be matter for the solution. Here, I will use *geopy.distance* module to compute the distance when a latitude and longitude of two points are known. After we find the distance from each venue to the campus, merge that data, and do more cleaning, here is the sample table of combination from Foursqare, Zomato data, and venue distance from the campus:



	index	categories	venue	latitude	longitude	average_price	price_range	rating	address	distance_from_ITB
0	0	Coffee Shop	Kozi Lab	-6.916558	107.620945	50000.0	2.0	3.9	Jl. Gudang Selatan No. 22, Sumurbandung, Bandung	2.993580
1	1	Hotel	Infinite Cafe & Lounge - Crowne Plaza Bandung	-6.917136	107.612013	250000.0	4.0	3.4	Crowne Plaza Bandung, Lantai 22, Jl. Lembong N...	2.838805
2	2	Fried Chicken Joint	Ayam Goreng Nikmat	-6.919658	107.617394	40000.0	2.0	3.9	Jl. Panaitan No. 9, Sumurbandung, Bandung	3.200650
3	3	Hotel	Furama Restaurant	-6.915615	107.610410	100000.0	3.0	0.0	Jl. Merdeka No. 2, Braga, Bandung	2.667078
4	4	Sushi Restaurant	Sushi Tei	-6.917129	107.613353	125000.0	3.0	4.4	Jl. Sumatera No. 9, Sumurbandung, Bandung	2.849432

Table 3. *Data combination from Foursqare, Zomato data, and venue distance from the campus*.

3. Methodology and Analysis

3.1 Initial analysis and expectation

The data from Zomato's API is retrieved using Foursquare's parameters, this will make some rows or values is missing in Zomato Data. We already clean it at "Data" section. The data that already cleaned, merged with the data from Foursquare’s API. I also droped all data from Foursquare’s API that have no match with its Zomato counterparts.

After initial visualization, I will make a table with relevant columns as metrics for K-Means clustering, in this process I normalize the average price column because it’s based on Indonesian currency which is relatively has larger numerical value (1 USD = 15.000 IDR) and could easily dominate the algorithm metric if we do not normalize it. After clustering process, I will do more advanced visualizations and make conclusions for solving the problem.

First, I will do a binning process for two columns — distance and average price this binnings will make us easier to intuitively analyzing each venue. I divide distance column to three categories: “nearby”, “not far”, and “quite far”, and average columns to four categories: "low","low-medium","medium-high", and "high". Here is the result from those binnings and scatter plot regarding those variables:

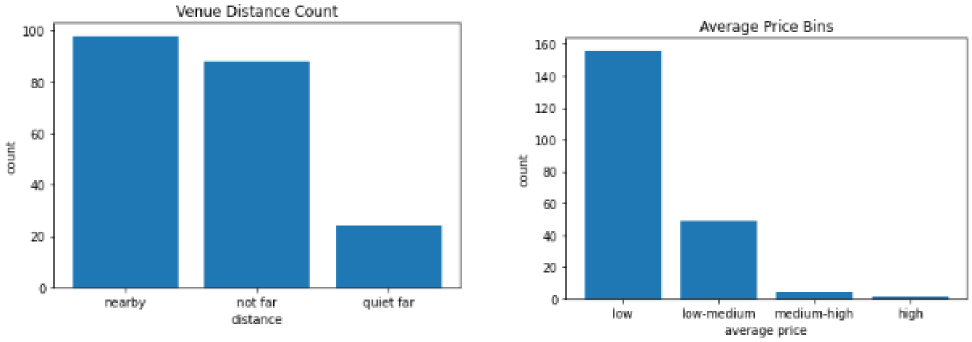


Figure 4. (a-left) *Distance binning and* **(b-right)** *average price binning.*

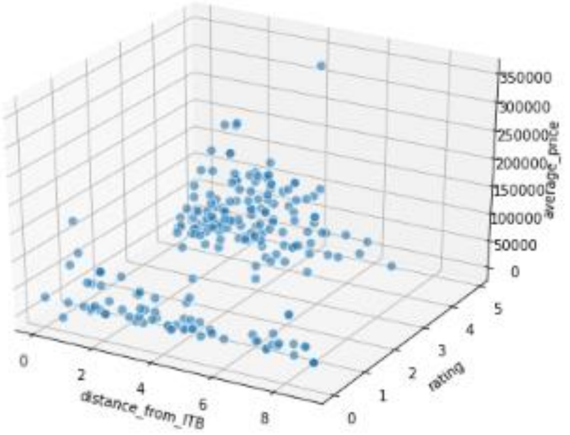


Figure 5. *Scatter plot with x-axis is distance, y-axis is rating, and z-axis is average price.*

What should I expect from these visualizations? From the bins bar graph and 3D scatterplot above, I expect that it should be easy for Budi to get the neighborhood that he wants. First distance bins give me indication that a lot of venues are nearby the campus. The average price bins give me indication that most venues are at "low" average price bins. The 3D scatter plot below gave me an insight, that there are relatively a lot of venues nearby the campus, which has "low" average prices and quite high ratings.

3.2 Clustering and Visualization

By using K-Means algorithm, I divide this data to four clusters labeled by 0 - 3. The first cluster, cluster 0 contains most venues with total of 97 venues. The second cluster is cluster 1 with the least number of venues with total 25 venues, the third clusters—cluster 2 contains 39 venues, and the last cluster— cluster 3 contains 39 venues.

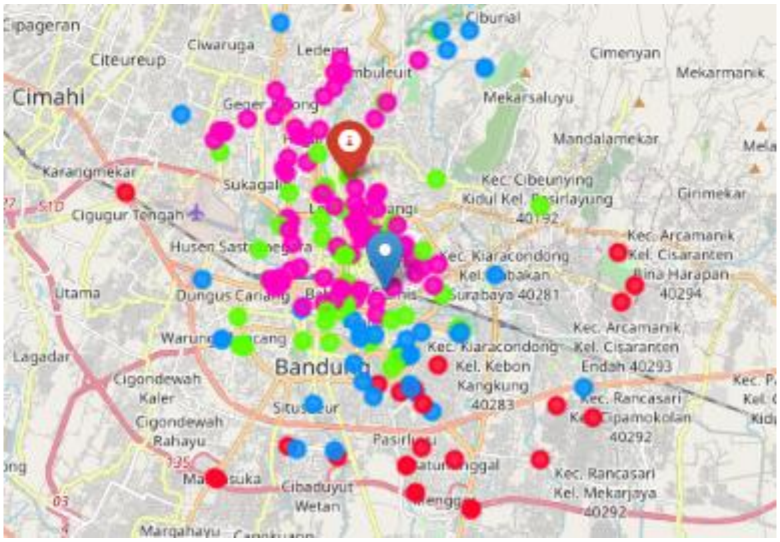

Cipageran
Citeureup
Ciwaruga
Cimahi
Ciburial
Cimenyan
Mekarmanik
Mekarsaluyu
Mandalamekar
Karangmekar
Cigugur Tengah
Sukagalih
Kec. Cibeunying Kidul Kel. Pasirlayung 40192
Girimekar
Husen Sastranegara
Kec. Arcamanik Kel. Cisaranten Bina Harapan 40294
Utama
Dungus Cariang
Kec. Kiaracondong Kel. Babakan Surabaya 40281
Lagadar
Bandung
Kec. Kiaracondong Kel. Kebon Kangkung 40283
Kec. Arcamanik Kel. Cisaranten Endah 40293
Cigondewah Kaler
Cigondewah Rahayu
Situsaeur
Kec. Rancasari Kel. Cipamokolan 40292
Pasirluyu
Kec. Rancasari Kel. Mekarjaya 40292
Cibaduyut Wetan
Margahayu

Figure 6. *All Venues that are member of* ***Cluster 0*** *depicted by* ***"dark pink"*** *points scattered on the map. All Venues that are member of* ***Cluster 1*** *depicted by* ***"red"*** *points scattered on the map. All Venues that are member of* ***Cluster 2*** *depicted by* ***"light green"*** *points scattered on the map. All Venues that are member of Cluster 3 depicted by* ***"light blue"*** *points scattered on the map.*

3.2.1 Cluster 0

Cluster 0 has the most venues. With scatter plots at figure 7 we can see that there are relatively many venues nearby ITB, has cheap price, and also relatively high rating. We can see below that the average normalized price of this cluster is 0.2198, the average distance from this cluster to ITB is 2.1739 km, and the average rating of this cluster is 3.8567.

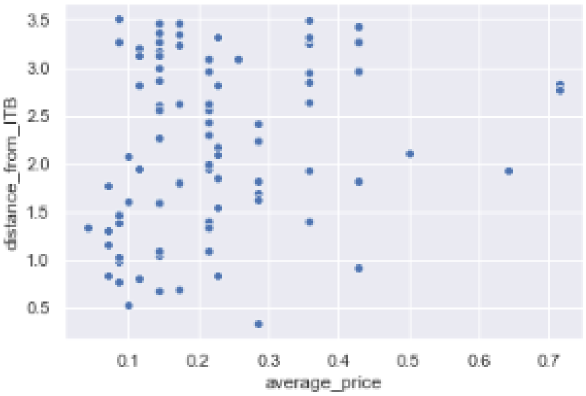


Figure 7. *Scatter plot of cluster 0 with normalized average price as x-axis and distance as y-axis.*

3.2.2 Cluster 1

Cluster 1 has the least number of venues. With scatter plots below we can see that there are relatively low venues that is nearby ITB. We can see below that the normalized average price of this cluster is 0.1371, the average distance from this cluster to ITB is 7.3494 km, and the average rating of this cluster is 0.0. I personally do not find that 0.0 rating is an indication that all venues of this cluster are bad, merely it caused by incomplete data that gave by the user and affect the K-Means algorithm metrics. We can use 0.0 rating as an indication that this cluster not quite popular enough to get any review. From price and distance persepective, this cluster has lower average price from cluster 0 but also quite far from the campus.

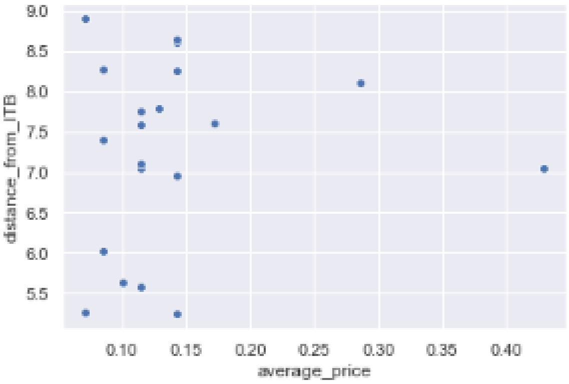


Figure 8. *Scatter plot of cluster 1 with normalized average price as x-axis and distance as y-axis.*

3.2.3 Cluster 2

Cluster 2 has medium number of venues. With scatter plots below we can see that there are relatively low venues that is nearby ITB. We can see below that the normalized average price of this cluster is 0.1574, the average distance from this cluster to ITB is 2.9629 km, and the average rating of this cluster is 0.0. I personally do not find that 0.0 rating is an indication that all venues of this cluster are bad, merely it caused by incomplete data that gave by the user and affect the K-Means algorithm metrics. We can use 0.0 rating as an indication that this cluster not quite popular enough to get any review. From price and distance persepective, this cluster has lower average price from cluster 0 also not realy further away from it.

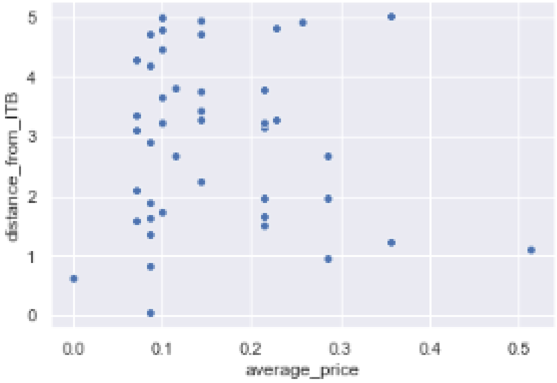


Figure 9. *Scatter plot of cluster 2 with normalized average price as x-axis and distance as y-axis.*

3.2.4 Cluster 3

Cluster 3 has quite low number of venues. With scatter plots below we can see that there are relatively low venues that is nearby ITB. We can see below that the normalized average price of this cluster is 0.1371, the average distance from this cluster to ITB is 7.3494 km, and the average rating of this cluster is 0.0. I personally do not find that 0.0 rating is an indication that all venues of this cluster are bad, merely it caused by incomplete data that gave by the user and affect the K-Means algorithm metrics. We can use 0.0 rating as an indication that this cluster not quite popular enough to get any review. From price and distance persepective, this cluster has lower average price from cluster 0 also quite further away from it.

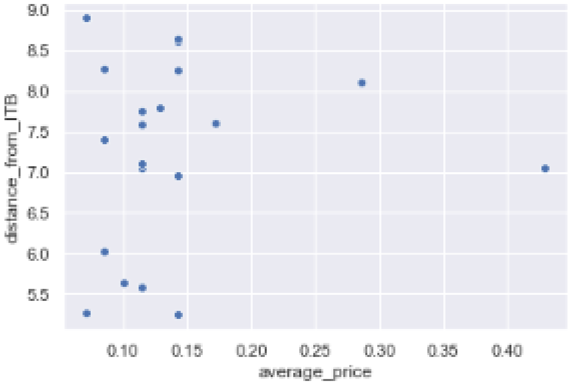


Figure 10. *Scatter plot of cluster 3 with normalized average price as x-axis and distance as y-axis.*

4. Results and Discussion

From the data and analysis, it shows that there are relatively a lot of restaurants and cafes nearby the campus, we got total of 210 venues that later divided to 4 clusters. Initialy, I got 229 venues data using Foursquare’s API, but with Zomato’s API, I can filter these results to get data that match restaurants or cafe related category that I want.

The campus location is relatively near the city center, this is a factor that make the average price of further restaurant is cheaper than the one that is nearby. The data also shown us that there are a lot of 0.0 ratings at each venue. I personally do not find that 0.0 rating is an indication that all venues of this cluster are bad, merely it caused by incomplete data that gave by the user and affect the K-Means algorithm metrics. I choose to not delete this 0.0 rating because it is still useful as popularity metric. Notice that nearby restaurants tends to have a rating than those who are further away. We can see from 4 cluster (cluster 0 - 3), that only one cluster; cluster 0 that has rating on each venue. This indicates that venues on cluster 0 is more popular than other clusters.

From each cluster we can see that every cluster has their own advantages and disadvandtages, cluster 1 and 3 by average is cheaper than 0 and 2, but those clusters is quite far from the campus. Cluster 0 and 1 have more expensive average price but those clusters are not far from the campus, moreover cluster 0 has more popular venues compare to other clusters.

5. Conslusion and Future Development

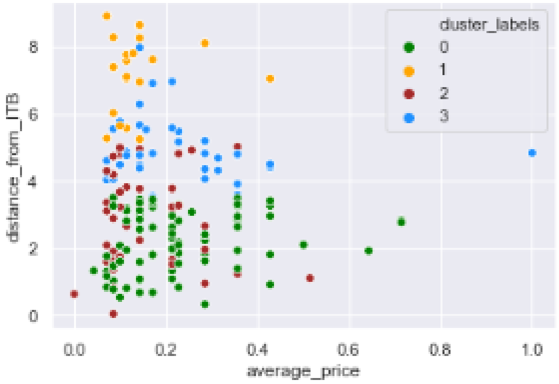


Figure 11. *Scatter plot of all clusters with normalized average price as x-axis and distance as y-axis.*

Now it is time for me to give Budi suggestion based on those data and analysis using K-Means clustering alghoritm. So according to the data, cluster 0 and 2 is the best for him. Cluster 2 which represent by scatters of "light green" points on the folium map at **Figure 7** has lower average price, quite near the campus, but not realy popular. Meanwhile, cluster 0 is the most popular and the nearest cluster by average ratings and distances to the campus, but also has the highest average price. I personally recommend him cluster 0. If we see the scatter plot from **Figure 11**, the highest average price is also caused by the numbers of venues that exist on cluster 0. Cluster 0 has most venues, more than twice that cluster 2 has. Cluster 0 has a lot of venues that in same range with cluster 2 has, also a lot of venues in more expensive range. This make cluster 0 average price is more expensive from the cluster 1.

I consider this data analysis need more development, for example, the accomodation near the campus usualy has higher price than the rest, it should be a metric to consider. But, due to the limitance of getting that data, my suggestion is limited based on data from Foursquare’s and Zomato’s API.