Finding the Most Convenience Place to Stay in Jakarta

M. Subagja Sastra Wardaya

October 31, 2020

1. Introduction

1.1 Background

If you have ever considered making a change and moving to live near MRT Station, you have probably, at some point, considered a variety of factors to determine where best to move. There are a lot of factors about where you live that can affect your quality of life and your happiness. One of most important factor is neighborhoods around the stations should be suitable with your needs and people needs might be different. For example if you are yoga enthusiast, you might be consider to live near yoga studio, or if you are movie addict, you might be consider to live near cinema, and so on.

1.2 Problem

Data that might contribute to determining locations to stay might include list of MRT Station in Jakarta, detail of mrt station's address like latitude and longitude, list all of venues arround MRT Station, and preferenced venues from user. This project aims to find the most convenience place to stay near MRT Station based on user preferenced venues.

1.3 Interest

People who might be intersted to use this recomendation system are office workers which had workplace around MRT Stations, or people who are tired of Jakarta's traffic and deciding to use MRT as primary transportation mode and they have a lot of preferred neighborhood arround their home.

2. Data Acquisition and Cleaning

2.1 Data Sources

The data that I'm going to use for this project come from MRT Jakarta Official Web Page which had published list all of MRT Stations in Jakarta and it had map track too, but they did not have any details about latitude and longitude so after get list all of station, I need to fetch the latitude and longitude using Geopy. Afterwards, I fetch all of neighborhood venues using Foursquare API. And the last data is preferred venues which come from user, for the sake of simplicity, I will use dummy user called Nancy.

2.2 Data Cleaning

I have got no problem while scraping from MRT Jakarta Official Web Page, the problem came from geopy which stop and return attribute error because there is no latitude and logitude for some station, so I must using try and except for preventing stop and transform attribute error into NaN values.

Table 1. Sucessful and Nan Values from Geopy

	Station Name	Latitude	Longitude
0	Stasiun Lebak Bulus Grab	NaN	NaN
1	Stasiun Fatmawati	-6.256329	106.797123
2	Stasiun Cipete Raya	-6.307817	106.838940
3	Stasiun Haji Nawi	NaN	NaN
4	Stasiun Blok A	-6.256329	106.797123
5	Stasiun Blok M BCA	-6.913684	107.602550
6	Stasiun ASEAN	NaN	NaN
7	Stasiun Senayan	-6.228269	106.800492
8	Stasiun Istora Mandiri	NaN	NaN
9	Stasiun Bendungan Hilir	-6.207213	106.797583
10	Stasiun Setiabudi Astra	NaN	NaN
11	Stasiun Dukuh Atas BNI	NaN	NaN
12	Stasiun Bundaran HI	-6.191864	106.822988

So, for better machine learning algorithm and result, I decided to fill NaN values manually using Google Map's coordinate.

Table 2. Completed Location of MRT Stations

	Station Name	Latitude	Longitude
0	Stasiun Lebak Bulus Grab	-6.289271	106.770543
1	Stasiun Fatmawati	-6.256329	106.797123
2	Stasiun Cipete Raya	-6.307817	106.838940
3	Stasiun Haji Nawi	-6.266690	106.792932
4	Stasiun Blok A	-6.256329	106.797123
5	Stasiun Blok M BCA	-6.913684	107.602550
6	Stasiun ASEAN	-6.238770	106.794022
7	Stasiun Senayan	-6.228269	106.800492
8	Stasiun Istora Mandiri	-6.222360	106.806396
9	Stasiun Bendungan Hilir	-6.207213	106.797583
10	Stasiun Setiabudi Astra	-6.209090	106.817352
11	Stasiun Dukuh Atas BNI	-6.201160	106.823575
12	Stasiun Bundaran HI	-6.191864	106.822988

For checking latitude and longitude I decide to use Folium to interpretate my finding, and afterwards I using free foursquare developer account, which give limit to developer to not request more than 5000 results hourly so I divide it to 13 station, and radius 2000 m for better finding. And the results is 1639 venues with had null values arround 700 in categories, so I decided to fill it manually again by Google Maps, not all of them, only for known place and drop for the rest.

2.3 Feature Selection

After data cleaning, I have got 2 main dataframe 1 dataframe for stations list containing 13 stations name, latitude and longitude, 1 dataframe venues from foursqare API containing 1.094 completed venues category from 1.639 venues, and need to examining which feature to need kept and discarded. After examining the feature I have got 5 features to kept from foursquare.

Table 3. Feature Selection

Completed Features	Feature to Kept	Dropped Feature	Reason for Drop
Id, name, categories, referralId, hasPerk, location.address, location.lat location.lng,location.labeledLatLngs, location.distance, location.cc, location.city, location.state, location.country, location.formattedAddress, location.crossStreet, location.postalCode, location.neighborhood.	name, categories location.lat, location.lng.	Id, referralId, hasPerk, location.address, location.labeledLatLngs, location.distance, location.cc, location.city, location.state, location.country, location.formattedAddress, location.crossStreet, location.postalCode, location.neighborhood.	For creating recomender system, I only need the venues name, categories, latitude and longitude.

With 4 feature before, I merge it with station dataframe and only choose station name, and then I have 5 feature for my recomender system.

Table 4. Choosen Feature

Station Name	Venue	Categories	Latitude	Longitude
Stasiun Lebak Bulus Grab	Perapatan Pasar Jumat	Bridge	-6.289241	106.769998
Stasiun Lebak Bulus Grab	Park And Ride Lebak Bulus	Parking	-6.289630	106.770286
Stasiun Lebak Bulus Grab	Komplek Sepolwan	General Travel	-6.289224	106.770614
Stasiun Lebak Bulus Grab	Sate Talago Biru, Lebak Bulus	Indonesian Restaurant	-6.289457	106.770885
Stasiun Lebak Bulus Grab	Ayam Bakar Mas Mono (Ciputat)	BBQ Joint	-6.289260	106.770900
...	...	...	...	...
Stasiun Bundaran HI	Jenius center Plaza Indonesia	Business Center	-6.191840	106.822865
Stasiun Bundaran HI	Sate Senayan Grand Indonesia	Indonesian Restaurant	-6.191610	106.822905
Stasiun Bundaran HI	Point Indomaret	Department Store	-6.191943	106.822982
Stasiun Bundaran HI	Toba Room	Meeting Room	-6.191972	106.822974
Stasiun Bundaran HI	Plaza PT. Bank Mandiri .Tbk	Office	-6.191825	106.822982

3. Exploratory Data Analysis

3.1 Relationship Between Station and Number of Venues

If a station has many places, we can hypothesize that the station maybe can match to everyone's need. Here I show the number of venues for each station, from the most to the least .

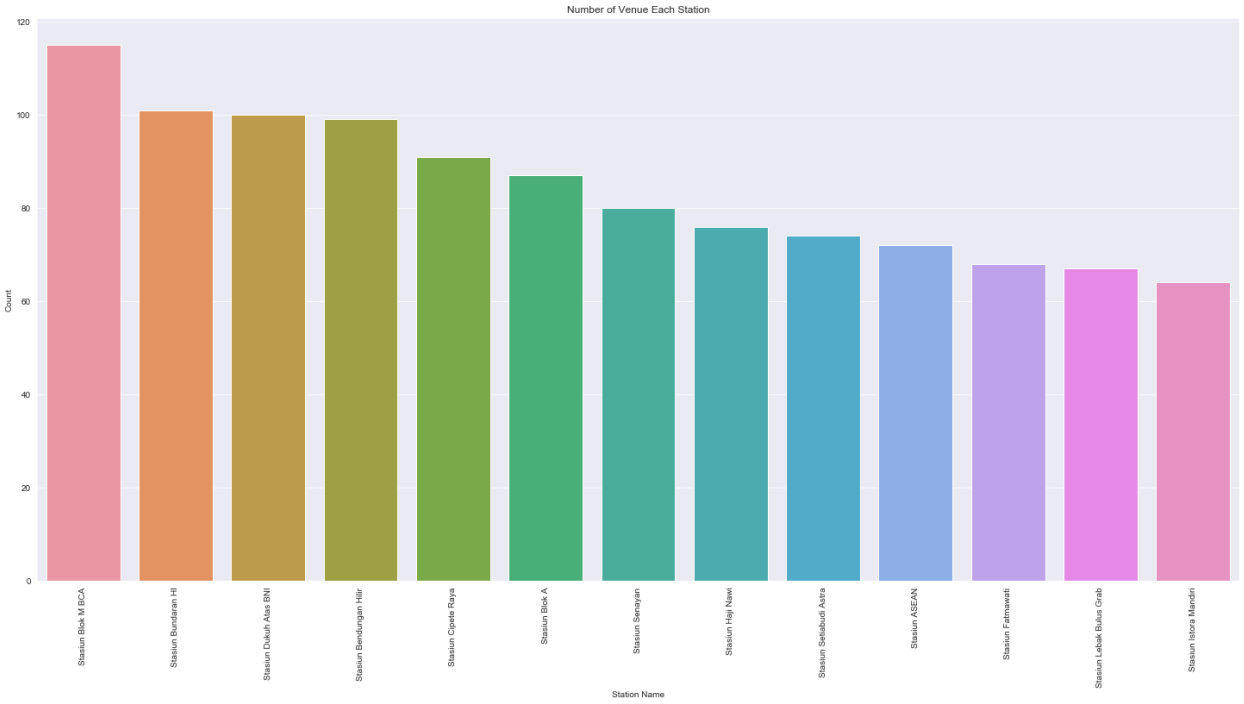

Figure 1. Number of Venues Each Station

Each of the station had a lot of venues, which is Stasiun Blok M BCA had the most number of venues, which had 117 venues and then second position is Stasiun Bundaran HI, which had 100 venues, the third is Stasiun Dukuh Atas BNI, which had 100 venues and the least is Stasiun Lebak Bulus Grab with arround 69 venues and Stasiun Istora Mandiri with 65 venues.

3.2 Relationship Categories Name and Number of Categories

If a category has a large number, we can hypothesize that the greater the number, the easier it is to find at each station. Here I show the number of categories from the most to the least .

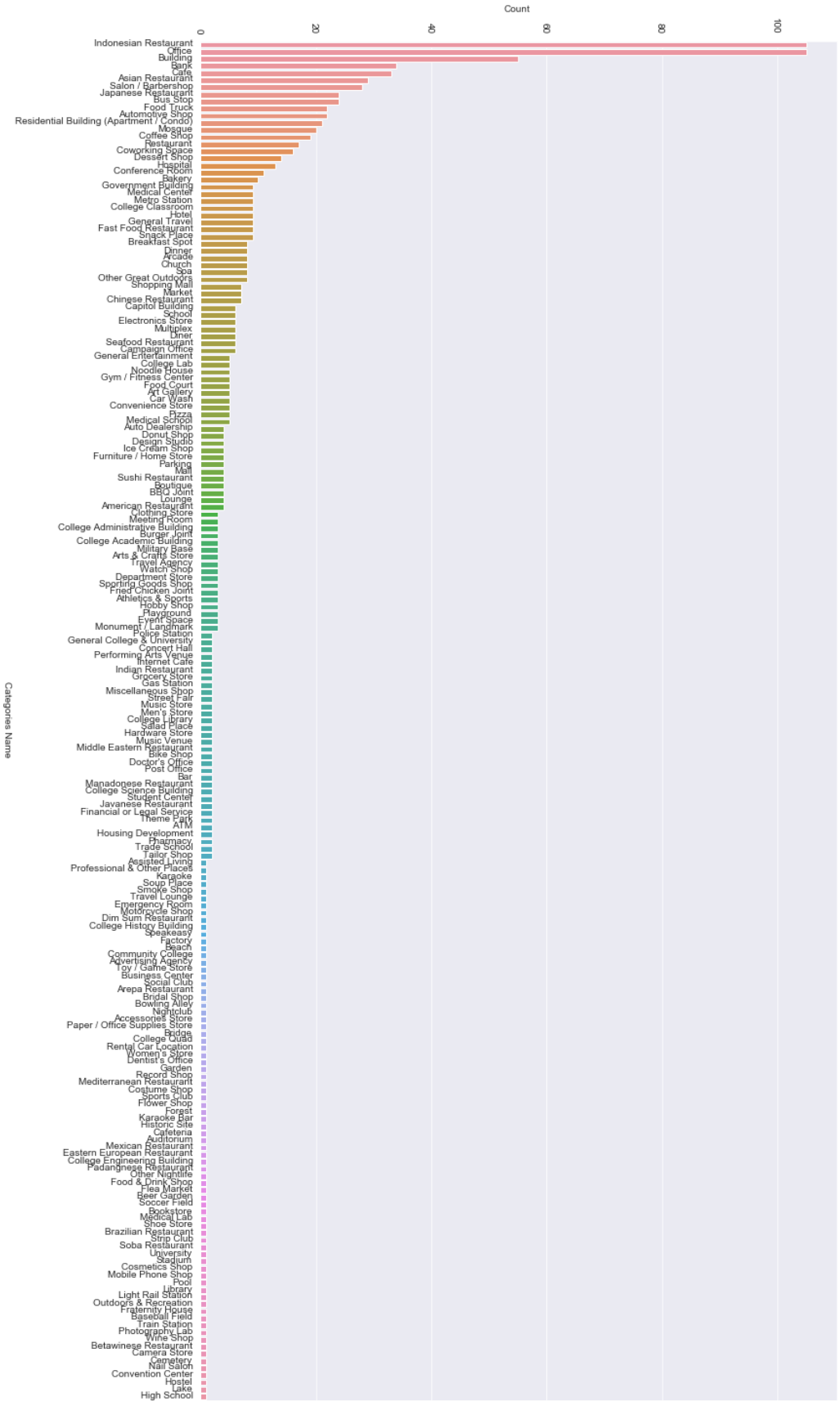

Figure 2. Number of Categories

From the picture above, it can be concluded that the most place category is Indonesian Restaurant followed by offices and buildings. Meanwhile, the least category position is occupied by hostel, lake and high school.

3.3 Venues Arround MRT Station

Based on Relationship Between Station and Number of Venues before, we know that Stasiun Blok M BCA had the most number of venues, and Stasiun Istora Mandiri had the least number of venues. Venues are intepreted with red dot, the more red dots, the more venues and vice versa.

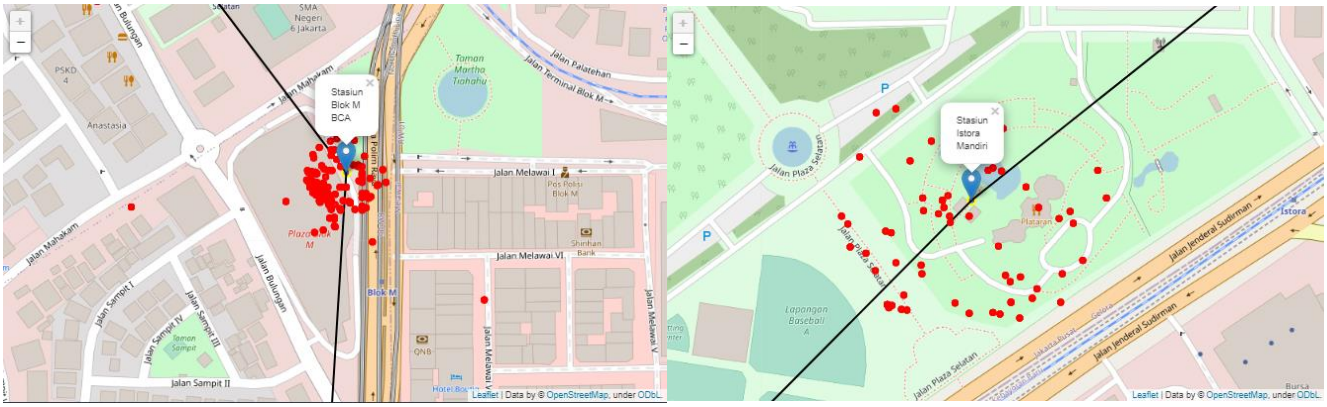

Figure 3. Venues Arround MRT Station

4. Predictive Modeling

4.1 Recomender System

There are a lot of technique for getting recomender system. We can use Euclidian Distance, Cosine Similarity and so on. For this project, I will use Cosine Similarity technique because the simplicity.

4.1.1. Cosine Similarities

Compute cosine similarity between samples in X and Y. Cosine similarity, or the cosine kernel, computes similarity as the normalized dot product of X and Y.

4.1.1. Defining Dummy User

Nancy is my friend. She is a healthy girl. She start her morning routine with something healthy like apple

vinegar, ant then she prepares her meal and go to coffee shop to drink coffee and do some exercise in a fitness center before she goes to work. Her office is close to MRT Station, so she uses MRT as her mode of transportation.

Now her rent is almost over, she want to moves to another place that might as match as she wants. But she had no idea where should be. She need some recomendation based on her preferenced venues. So where Nancy should live?

4.1.2. Preparation

Before computing similarity, I must transform categories to one hot encoding. I convert categorical values in category become integer representation and grouping for venues table and user.

Table 5. Venues Table One Hot Encoding

Station Name	ATM	Accessories Store	Advertising Agency	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	& Crafts Store	Asian Restaurant	Assisted Living	...
Stasiun ASEAN	0	0	0	0	0	0	0	0	0	0	...
Stasiun Bendungan Hilir	0	0	0	0	1	0	0	1	1	0	...
Stasiun Blok A	0	0	0	0	1	0	1	1	1	0	...
Stasiun Blok M BCA	0	1	0	1	1	0	0	1	1	0	...
Stasiun Bundaran HI	1	0	1	1	1	0	1	0	1	0	...
Stasiun Cipete Raya	0	0	0	0	0	0	0	0	0	0	...
Stasiun Dukuh Atas BNI	0	0	0	0	0	1	0	0	1	0	...
Stasiun Fatmawati	1	0	0	0	1	0	0	0	1	1	...
Stasiun Haji Nawi	0	0	0	0	0	0	0	0	1	0	...
Stasiun Istora Mandiri	0	0	0	1	0	0	1	0	0	0	...
Stasiun Lebak Bulus Grab	0	0	0	0	0	0	0	0	1	0	...
Stasiun Senayan	0	0	0	0	0	0	1	0	1	0	...
Stasiun Setiabudi Astra	0	0	0	0	0	0	0	0	1	0	...

Table 6. User Table One Hot Encoding

	Username	Accessories Store	Advertising Agency	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	Arts & Crafts Store	Asian Restaurant	Assisted Living	...
0	0	0	0	0	0	0	0	0	0	0	...

4.1.3. Modelling

After the grouped, we define user table as X with cotaining feature Gym / Fitness Center, Pool, Market and Coffee Shop

Table 7. Defining X

	Username	Gym / Fitness Center	Pool	Market	Coffee Shop
0	Nancy	1	1	1	1

I also defining Y with cotaining feature Gym / Fitness Center, Pool, Market and Coffee Shop from venues table.

Table 8. Defining Y

Station Name	Gym / Fitness Center	Pool	Market	Coffee Shop
Stasiun ASEAN	0	0	1	0
Stasiun Bendungan Hilir	0	0	1	1
Stasiun Blok A	0	0	1	0
Stasiun Blok M BCA	0	0	0	1
Stasiun Bundaran HI	1	0	0	1
Stasiun Cipete Raya	0	0	0	1
Stasiun Dukuh Atas BNI	0	0	1	1
Stasiun Fatmawati	0	0	0	1
Stasiun Haji Nawi	0	0	1	1
Stasiun Istora Mandiri	0	0	0	0
Stasiun Lebak Bulus Grab	1	1	1	1
Stasiun Senayan	0	0	0	1
Stasiun Setiabudi Astra	0	0	0	1

The result after modelling X and Y with cosine similarity is become array in similarity like Table 9, but for improving ease of application, I put it on table

Table 9. Similarity Result

	Station Name	Similarity
10	Stasiun Lebak Bulus Grab	100.000000
1	Stasiun Bendungan Hilir	70.710678
4	Stasiun Bundaran HI	70.710678
6	Stasiun Dukuh Atas BNI	70.710678
8	Stasiun Haji Nawi	70.710678
0	Stasiun ASEAN	50.000000
2	Stasiun Blok A	50.000000
3	Stasiun Blok M BCA	50.000000
5	Stasiun Cipete Raya	50.000000
7	Stasiun Fatmawati	50.000000
11	Stasiun Senayan	50.000000
12	Stasiun Setiabudi Astra	50.000000
9	Stasiun Istora Mandiri	0.000000

4.1.4. Evaluation

Never trust your found! Since Recomender System is unsupervise machine learning, we should check the result can be accepted or not. let use Pandas to check!

Table 9. Nancy's Preferenced Categories in Stasiun Lebak Bulus Grab

	Station Name	Venue	Categories	Latitude	Longitude
12	Stasiun Lebak Bulus Grab	Coffe toffee Pejaten village	Coffee Shop	-6.289001	106.770839
18	Stasiun Lebak Bulus Grab	Pasar Jumat	Market	-6.289220	106.771202
58	Stasiun Lebak Bulus Grab	Golds Gym Bxc	Gym / Fitness Center	-6.289728	106.771191
65	Stasiun Lebak Bulus Grab	Selapa/Sespimma Gym	Gym / Fitness Center	-6.288732	106.770908
85	Stasiun Lebak Bulus Grab	sport club pondok chandra indah	Pool	-6.289001	106.770640
115	Stasiun Lebak Bulus Grab	Starbucks Coffee rest. area KM 19 Tol Jakarta ...	Coffee Shop	-6.289543	106.771012

As the show above, Venues Table showing us all of Nancy's preferred venues. How about in maps?

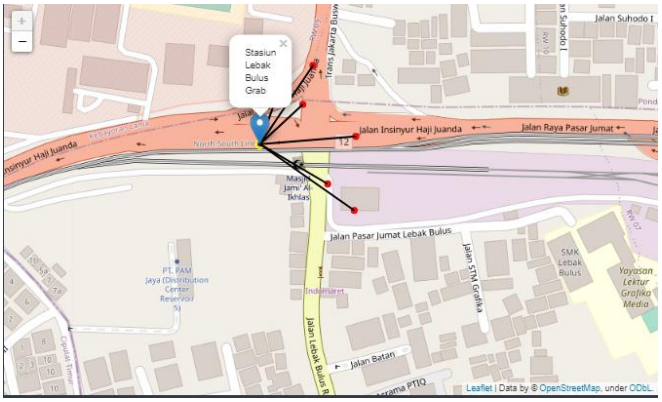

Figure 4. Nancy's Preferred Venues arround Stasiun Lebak Bulus Grab

As the figure above, are very close from Stasiun Lebak Bulus Grab! Now let see Stasiun Istora Mandiri which had 0% similarity.

Table 10. Nancy's Preferenced Categories in Stasiun Istora Mandiri

Station Name	Venue	Categories	Latitude	Longitude

As we see it return with zero result, so, the map should be empty

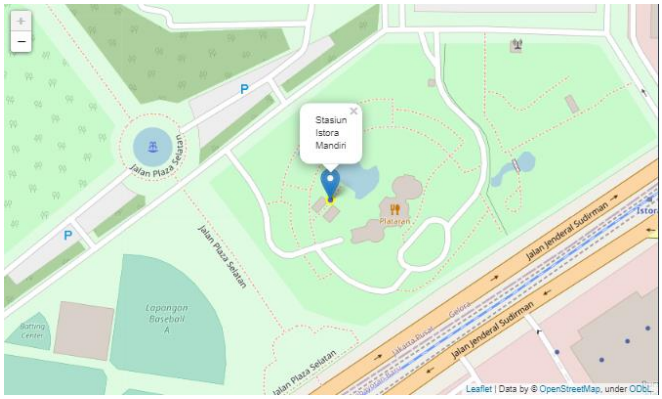

Figure 4. Nancy's Preferred Venues arround Stasiun Istora Mandiri

5. Conclusions

Based on my recomender system, Nancy should live near Stasiun Lebak Bulus Grab, because it have all of Nancy's preferred venues with 100% similarity, and Nancy should not choose to live near Stasiun Istora Mandiri because it had 0% similarity.

6. Future Directions

Based on my project, I was able to fetch venues category and with a lot of null values, so for the future directions, using different API besides foursquare like Google Maps, might be better result and more complete than foursquare.