IBM Applied Data Science Capstone on Coursera

Clustering Regions in Yogyakarta, Indonesia Based on It's Region Food Preferences

Rizky Oulia O. P. L.

July 27th, 2020

ouliaopl@gmail.com

1. Introduction

1.1. Background

Yogyakarta is a really diverse place in Indonesia. Food and beverage businesses is one of the most attractive and lucrative market.

All sort of people are here in Yogyakarta. Students, travellers and tourists, business owners, permananent residence and many others. These groups have a tendency to cluster to a certain location in Yogyakarta. Thus, this clustering also affects their food choices. Since their palates is affected, so does the success of a certain restaurants being open on certain locations.

Interests: This project is meant to help entrepreneurs, local cooks and chefs to better understand the correlation between the location and the best type of restaurant to be opened based on local palate. By choosing the best location based on the type of restaurant they're going to open, hopefully they could minimize the risk of bankruptcy and maximize profit.

Problems: To better understand the palate preferences in Yogyakarta, foursquare venue data will gives us relevant information about restaurant type and it's location. By using machine learning algorithm such as K-Modes clustering, we can group neighborhoods that have similar food palates based on type of restaurants thus giving us cluster of food preferences.

2. Data Acquisition and Cleaning

2.1. Data Sources

Data used in this project:

- Data retrieved from https://kodepos.nomor.net/_kodepos.php?_i=kota-kodepos&daerah=Provinsi&jobs=DI+Yogyakarta&perhal=400&urut=10&asc=00001111&sby=110000&no1=2 for region names and postal code information.
- Geopy library to find geographical location information.
- Using Foursquare API to find nearby restaurants information such as their location and types of dishes they sell.

2.2. Data Scraping, Cleaning and Preparation

Data from government website is used to retrieved Borough names, Regency names, Cities name and it's postal code. *BeutifulSoup* library is being used to retrieved these data. From this website I get 441 rows of location data.

To add location data such as Borough latitude and longitude I use *geopy* library to get the data from *Nominatim* OSM data.

For early visualization I use *Folium* library to visualize the data.

Figure 2.1 Visualization using *Folium* library

From the visualization I observed that some city that are labeled are outside of Yogyakarta. This might be because of the similar region names. So to exclude the region that are outside of Yogyakarta, I used limiting boundaries based on latitude and longitude of the places.

Some regions can't be detected by *geopy* resulting on a NaN values on some of the region latitude and longitude. Due to the importance of this attributes, datas with NaN values were dropped. 398 rows of data remained

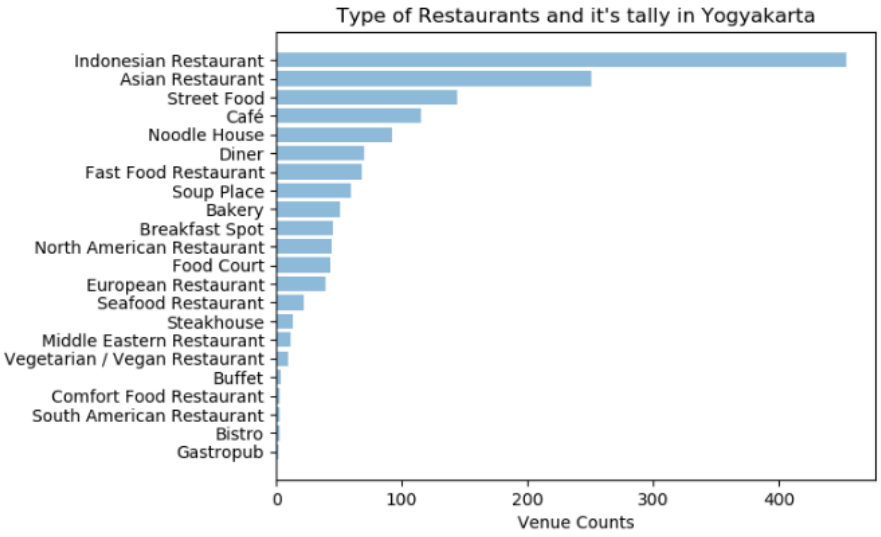
Collecting restaurants infromation were made by using foursquare API. The information that were requested are Restaurant Name's, Region where it is located, Geographical Location and Restaurant Category. 1547 rows of data attained.

Some Region in Yogyakarta didn't have any data of restaurants. So this regions were dropped. Some of it also had duplicates, so the duplicated one's were also dropped.

62 unique categories of restaurants were obtained from the data. Further observations showed that some of this restaurant is categorized as 'Food' and 'Restaurant'. Since these classification was too broad, I replace the following category with the mode of the dataset and that's 'Indonesian Restaurant'. From the observation I also saw some of the categories are basically similar for example DimSum and Dumpling restaurants or Food Court and Cafeteria. So to

eliminate this overlapping, I grouped those categories into one category. Furthermore from the data I observed, categories outside of Indonesian and Asian Restaurants were really low on frequencies. For example, there were lots of European Restaurants in the data but due to some of them are being categorized as European, Germans, French and so on the number of European Restaurant in the data was low. So to simplify the classification and also to make sure that some categories were well represented I grouped the categories into few bigger categories. Hence, only 22 categories left.

From that 22 categories I visualize them using *matplotlib* and obtained this result.



Graph 2.1. Type of Restaurants and It's Tally in Yogyakarta

Since Indonesian restaurant and Asian restaurant were way too frequent in this dataset, I then chose not to include them on the analysis. Building a restaurants with too much competitor isn't a good idea either.

3. Data Analysis

3.1. One-Hot Encoding

Using One-Hot encoding I assign binary values on the category of restaurants. 1 if it had the right category and 0 for everything else. I used the Borough name as the key

	Borough	Asian Restaurant	Bakery	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	Diner	European Restaurant	Fast Food Restaurant	Food Court	Gastropub	Indonesian Restaurant	Middle Eastern Restaurant
0	Ambarketawang	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1	Ambarketawang	0	0	0	0	0	0	0	0	0	0	0	0	1	0
2	Ambarketawang	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Ambarketawang	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	Argomulyo	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 3.1. One-Hot encoded the dataframe

Then I grouped the dataframe by the same Borough and find the mean values of the same categories and the same Borough.

	Borough	Bakery	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	Diner	European Restaurant	Fast Food Restaurant	Food Court	Gastropub	Middle Eastern Restaurant	Noodle House	North American Restaurant	S Res
0	Argomulyo	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.5	0.000000	0.000000	
1	Argomulyo (Argo Mulyo)	0.0	0.0	0.0	0.0	0.0	0.0	0.200000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.400000	
2	Argorejo	0.0	0.0	0.0	0.0	0.5	0.0	0.250000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	
3	Argosari	0.0	0.0	0.0	0.0	1.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.000000	
4	Baciro	0.0	0.0	0.0	0.0	0.0	0.0	0.142857	0.0	0.0	0.0	0.142857	0.0	0.285714	0.142857	

Figure 3.2. Grouping the dataframe by Borough

3.2. Clustering the Restaurants

I then sort the categories of restaurant based on it's most common venues to the tenth most common venues.

	Borough	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	Argomulyo (Argo Mulyo)	North American Restaurant	Street Food	Diner	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant
2	Argorejo	Café	Soup Place	Diner	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Comfort Food Restaurant	European Restaurant
3	Argosari	Café	Vegetarian / Vegan Restaurant	Street Food	Bistro	Breakfast Spot	Buffet	Comfort Food Restaurant	Diner	European Restaurant	Fast Food Restaurant
4	Baciro	Noodle House	Vegetarian / Vegan Restaurant	North American Restaurant	Gastropub	Street Food	Diner	European Restaurant	Bistro	Breakfast Spot	Buffet
5	Balecatur	Soup Place	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	Diner	European Restaurant
6	Baleharjo	Café	Noodle House	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Comfort Food Restaurant	Diner	European Restaurant
7	Banaran	Noodle House	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	Diner	European Restaurant
8	Banguncipto	Diner	Vegetarian / Vegan Restaurant	Street Food	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	European Restaurant	Fast Food Restaurant
9	Bangunjiwo	Breakfast Spot	Noodle House	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Buffet	Café	Comfort Food Restaurant	Diner

Figure 3.3. Sorting the most common to the tenth common venue on the region

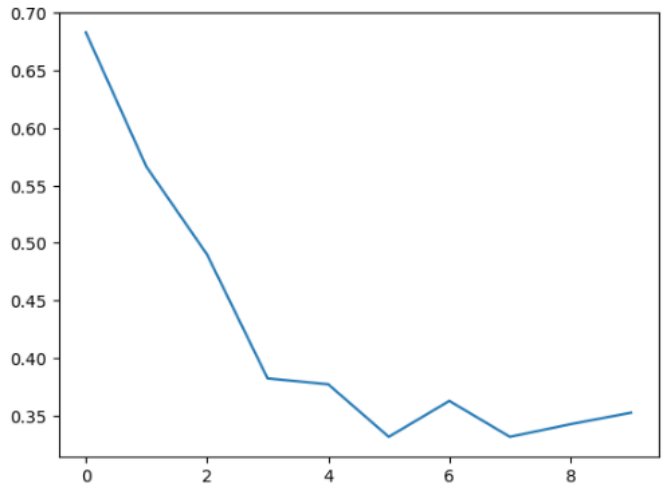
3.3. Using K-Modes to cluster the restaurants

Since the latest dataframe was a categorical data, I cannot use K-Means to cluster the data. So I used K-Modes.

First, I need to find the best K(number of clusters) before clustering the data.

By iterating several values of K (1 to 10) I obtained found different values of Error Means for each different K's. By using *matplotlib* I obtained,

```
[<matplotlib.lines.Line2D at 0xf78dcc8>]
```



Graph 3.1. K vs. Mean Error Relation

From the graph I concluded that the elbow point was on K=4 (K=3 on the graph for unknown reason). So the optimum number of clusters would be 4.

From K-Mode clustering I got the following centroid of clusters,

```
array([['Noodle House', 'Vegetarian / Vegan Restaurant',
        'Fast Food Restaurant', 'Bistro', 'Breakfast Spot', 'Buffet',
        'Café', 'Comfort Food Restaurant', 'Diner',
        'European Restaurant'],
       ['Café', 'Vegetarian / Vegan Restaurant', 'Street Food', 'Bistro',
        'Breakfast Spot', 'Buffet', 'Comfort Food Restaurant', 'Diner',
        'European Restaurant', 'Fast Food Restaurant'],
       ['Street Food', 'Vegetarian / Vegan Restaurant', 'Bistro',
        'Breakfast Spot', 'Buffet', 'Café', 'Comfort Food Restaurant',
        'Diner', 'European Restaurant', 'Fast Food Restaurant'],
       ['Café', 'Café', 'Vegetarian / Vegan Restaurant',
        'Fast Food Restaurant', 'Bistro', 'Breakfast Spot', 'Buffet',
        'Café', 'Comfort Food Restaurant', 'European Restaurant']],
      dtype='<U29')
```

Figure 3.4. Centroid of Clusters

After labelling each rows with it's cluster number, I then joined the dataframe with sorted most-common category data and the dataframe that have the location data of each Borough.

	Postal Code	Borough	Regency	City	Latitude	Longitude	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
0	55294	Ambarketawang	Gamping	Sleman	-7.805396	110.317874	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	55752	Argomulyo	Sedayu	Bantul	-7.670610	110.457459	Middle Eastern Restaurant	Street Food	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Café
2	55583	Argomulyo (Argo Mulyo)	Cangkringan	Sleman	-7.664847	110.463854	North American Restaurant	Street Food	Diner	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet
3	55752	Argorejo	Sedayu	Bantul	-7.821616	110.262355	Café	Soup Place	Diner	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet
4	55752	Argosari	Sedayu	Bantul	-7.813184	110.243059	Café	Vegetarian / Vegan Restaurant	Street Food	Bistro	Breakfast Spot	Buffet	Comfort Food Restaurant	Diner

Figure 3.5. Each Borough and it's important attributes

And the merged dataframe is now ready to be visualized.

4. Results

4.1. Cluster 1

Cluster 1 had these attributes,

	Borough	Longitude	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	Cluster Labels
count	52	52.000000	52	52	52	52	52	52	52	52	52	52	52.0
unique	52	NaN	12	15	12	12	12	13	10	11	10	10	NaN
top	Pendowoharjo	NaN	Noodle House	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	Diner	European Restaurant	NaN

Figure 4.1. Cluster 1 Descriptive Statistics

4.2. Cluster 2

Cluster 2 had these attributes,

	Borough	Longitude	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	Cluster Labels
count	37	37.000000	37	37	37	37	37	37	37	37	37	37	37.0
unique	37	NaN	7	6	5	7	7	8	8	9	7	6	NaN
top	Kepuharjo (Kepuh Harjo)	NaN	Café	Vegetarian / Vegan Restaurant	Street Food	Bistro	Breakfast Spot	Buffet	Comfort Food Restaurant	Diner	European Restaurant	Fast Food Restaurant	NaN

Figure 4.2. Cluster 2 Descriptive Statistics

4.3. Cluster 3

Cluster 3 had these attributes,

	Borough	Longitude	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	Cluster Labels
count	29	29.000000	29	29	29	29	29	29	29	29	29	29	29.0
unique	29	NaN	4	10	10	7	9	8	8	8	7	8	NaN
top	Bantul	NaN	Street Food	Vegetarian / Vegan Restaurant	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	Diner	European Restaurant	Fast Food Restaurant	NaN

Figure 4.3. Cluster 3 Descriptive Statistics

4.4. Cluster 4

Cluster 4 had these attributes,

	Borough	Longitude	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	Cluster Labels
count	40	40.000000	40	40	40	40	40	40	40	40	40	40	40.0
unique	40	NaN	14	15	13	8	8	6	7	8	7	11	NaN
top	Piyaman	NaN	Café	Café	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bistro	Breakfast Spot	Buffet	Café	Comfort Food Restaurant	European Restaurant	NaN

Figure 4.4. Cluster 4 Descriptive Statistics

Then from the data of every region geographic location, I then visualize it on map using *Folium* map library assigning different colors to each clusters.

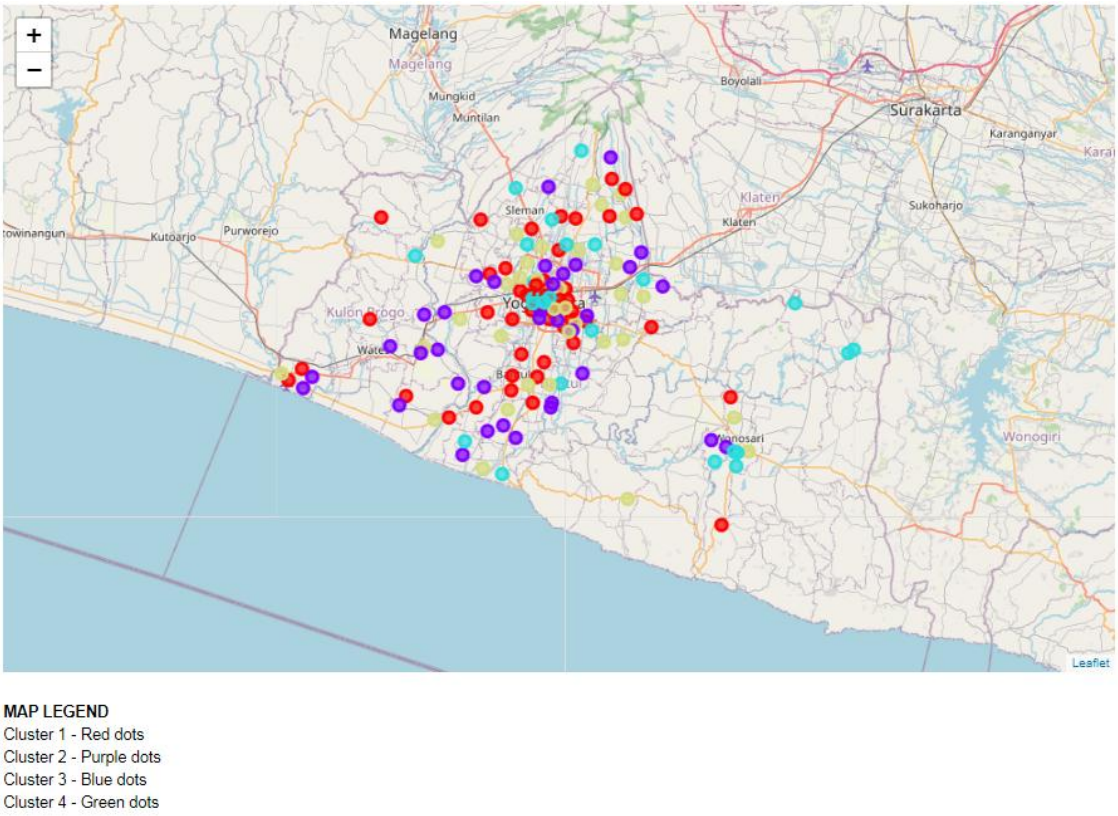


Figure 4.5. Map of Yogyakarta and The Different Regions Based on Food Preferences

5. Discussion

From the map we could see some interesting clustering of cuisines in Yogyakarta province. At the city center it include all cluster. Most of cluster 1 are on the city center, Cluster 2 on the southern part near the beaches, Cluster 3 on the northern part near the highlands, and the 4th Cluster mainly located on the city centers or residential areas.

By looking at every clusters most common venue:
Cluster 1:Noodle House, Vegan/Vegetarian Restaurant, Fast Food
Cluster 2:Cafe, Vegan/Vegetarian Restaurant, Street Food
Cluster 3:Street Food, Vegan/Vegetarian Restaurant, Bistro
Cluster 4:Cafe, Cafe, Vegan/Vegetarian Restaurant

We may classify these cluster based on their food price range and group of people qualitatively to make a better and more complete reccomendation. But further research would be very useful to justify this qualitative claim.

Since restaurant businesses could be categorized as a red ocean markets, opening similar restaurant as the most common restaurant found in that region might be a challenge. But opening a restaurant with the least common type might be a poor decision as well since there might be lack of demands. So we suggest to choose the 4th to 6th most common venue as the

safe bet as the region is not crowded by the similar type of restaurants but also doesn't show the lack of demands.

To summarize the discussion, we will show the recommended type of restaurant to be opened on a certain cluster based on the data analysis. To complete the reccomendation, we will add some qualitative reasonings. This qualitative reasonings is only based on one's experience living in Yogyakarta.

Restaurant Reccomendations:

Cluster 1:
Types: Bistro, Breakfast Spot, Buffet;
Group of People: All Kinds;
Price: Low-Medium-High
Cluster 2:
Types: Bistro, Breakfast Spot, Buffet;
Group of People: Tourist;
Price: Medium-High
Cluster 3:
Types: Breakfast Spot, Buffet, Cafe;
Group of People: Tourists;
Price: Medium-High
Cluster 4:
Types: Fast Food, Bistro and Breakfast Spot;
Group of People: Students, Residents;
Price: Low-Medium

6. Conclusion

In this study I cluster regions in Yogyakarta province, Indonesia based on food preferences. The food preferences are based on the most common type of restaurants found on 500 meter radius and rank them. I found that the most popular types of restaurant in Yogyakarta with the exception of Indonesian and Asian Restaurant are Street Food, Cafe and Noodle House. But these types of restaurant does not always be the most common types of restaurant on all clusters, this may shows that these restaurants are popular only on certain regions. And Vegan/Vegetarian restaurants are all at the top of the charts on every cluster this may indicate that eventhough the number of Vegan/Vegetarian restaurants are low, they are widely spread all over Yogyakarta. By using K-modes clustering, I clustered all of the borough in Yogyakarta into 4 cluster of food preferences. I chose 4 clusters as I saw it as the elbow point and it gives 62 percent accuracy. I hope that this clustering could be useful to determine the types of restaurants to be built and where it should be built. And furthermore maximizing profit.

This model only gives 62% of accuracy, even by increasing the number of clusters the accuracy only increases to about 65%, although using really big K's would give much better accuracy but the clustering would be too specific. I think there may be better approaches other than the K-Modes that was used in this study. Furthermore, more data might be the answer since I

observed that small and rural regions does not have many data or even worse, no data at all from the Foursquare API.

This study also excluded the frequency/number of restaurants on each region since I only cluster the region by ranking. This may cause overvalue-ing regions with lack of data and undervalue-ing regions with lots of data. And more study about other attributes such as prices and group of people is needed to better understand the clustering. Since the foursquare API have limited access to such data, I did not include it in the study.

Notes: The inspiration and most of the algorithm are from this person https://github.com/neofluar/coursera_capstone/blob/master/opening_restaurant_in_london.ipynb. **Some algorithm and methods are modified to feed my needs, but a lot of thanks Mr. Neofluar (github account) you inspired and helped me learn a lot from your project ☺**