Written by Nourman Hajar

Applied Data Science Capstone Project

Districts of Surabaya: Seeing Through A Culinary Perspective

Prologue

In the recent months, COVID-19 hit the world and brought most of the world into lockdowns. In this time, I used some of the additional free time to understand data science and learn thoroughly how to code in Python and using its diverse libraries, such as pandas, folium, numpy, geopandas, scikit-learn, etc.

I took IBM Data Science Professional Certificate course in Coursera and the final assignment of this course is this capstone project. The capstone project demands that I can analyze, prepare, and bring an insight about data on venues in cities using Foursquare API. I chose the city of Surabaya as a case. With this report, I also created a presentation and a Jupyter Notebook containing the technical source code. I hope everyone reading this report can enjoy my explanation.

Introduction/Business Problem

Surabaya is a city in the North Coast of Java. It is an important city for land and sea transport. It acts as a hub that connects cities of Indonesia, and is located in Java, the most densely populated island in the world.

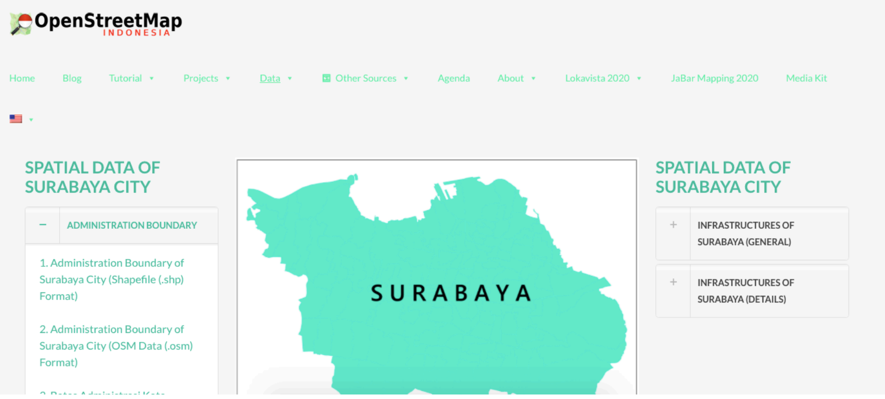
As a port city, Surabaya definitely is some kind of cultural pot from various different cultures, like the Javanese, Arab, Chinese, and European culture. It also attracts a huge number of tourists. Surabaya also has various cuisine. Finding the right spot to eat in Surabaya can be difficult especially for a person who wants to do a “culinary tourism”.

We want to find out if there is any pattern in the type of restaurants (including cafés, foodtrucks, etc) located in every district and neighborhood of Surabaya. Thus the problems are:

1. Do some districts represent certain types of cuisines/restaurants more than the others?
2. Is it possible to group districts together based on the sole criteria of food/restaurant types?

Description of the data

As requested by the assignment task, I will use Foursquare data about food venues in Surabaya. The data used is the name of the venue, latitude, longitude, venue category (in food section), etc. In addition, I will use the data provided by Openstreetmap Indonesia community to provide data for name, latitude, and longitude of geometric center of each district in the city of Surabaya. The community provides the data for free and is permissioned to be used for any project. The data is in .gpkg format, therefore I will convert it into a readable data first that will be explained in the full notebook. The link is here.



SPATIAL DATA OF SURABAYA CITY

− ADMINISTRATION BOUNDARY

1. Administration Boundary of Surabaya City (Shapefile (.shp) Format)

2. Administration Boundary of Surabaya City (OSM Data (.osm) Format)

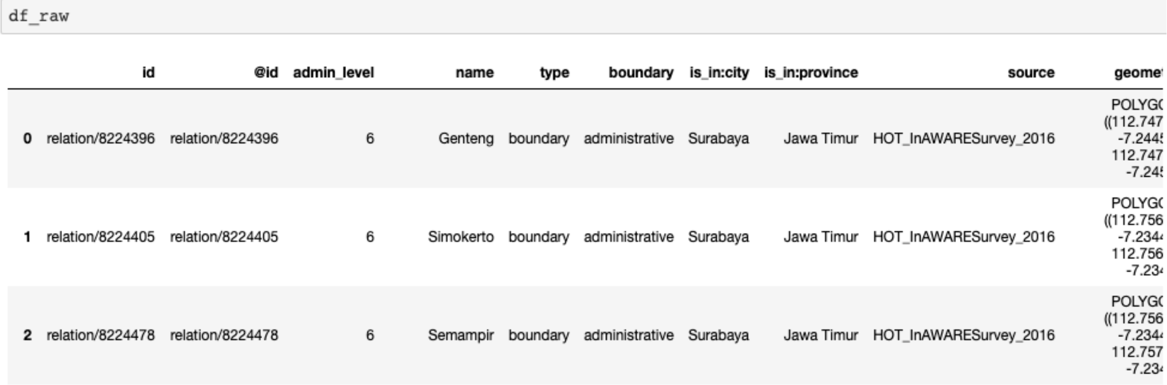
SPATIAL DATA OF SURABAYA CITY

+ INFRASTRUCTURES OF SURABAYA (GENERAL)

+ INFRASTRUCTURES OF SURABAYA (DETAILS)

Methodology

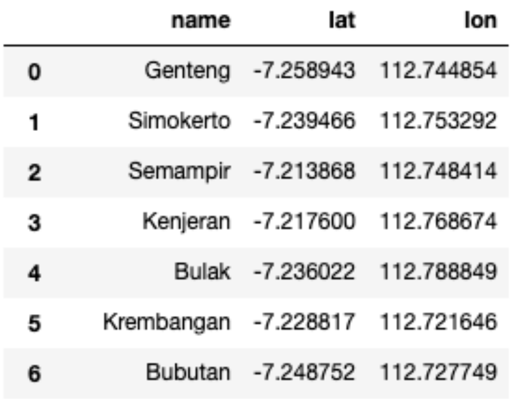
I used the data provided by Openstreetmap Indonesia community to provide data for name, latitude, and longitude. They provided data for districts boundaries (geometric shapes). I imported the data into a GeoDataFrame using geopandas.



```
df_raw
```

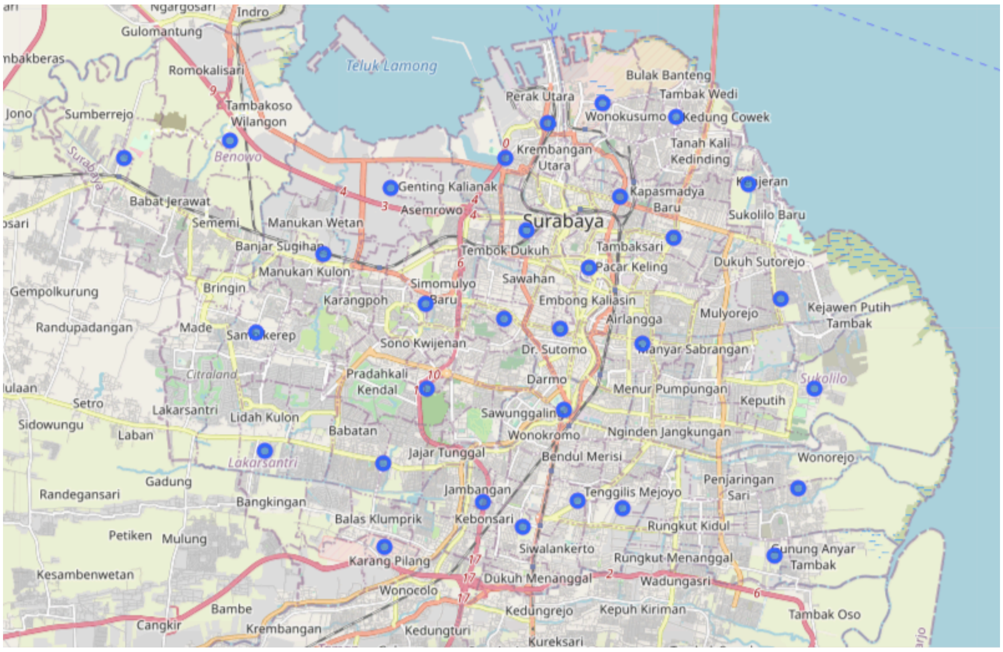
	id	@id	admin_level	name	type	boundary	is_in:city	is_in:province	source	geome
0	relation/8224396	relation/8224396	6	Genteng	boundary	administrative	Surabaya	Jawa Timur	HOT_InAWARESurvey_2016	POLYGO ((112.747 -7.2445 112.747 -7.245
1	relation/8224405	relation/8224405	6	Simokerto	boundary	administrative	Surabaya	Jawa Timur	HOT_InAWARESurvey_2016	POLYGO ((112.756 -7.234 112.756 -7.234
2	relation/8224478	relation/8224478	6	Semampir	boundary	administrative	Surabaya	Jawa Timur	HOT_InAWARESurvey_2016	POLYGO ((112.756 -7.234 112.757 -7.234

Then, I transformed the geometry to find the centroid of each district. Then I can transform them into a nice looking DataFrame.



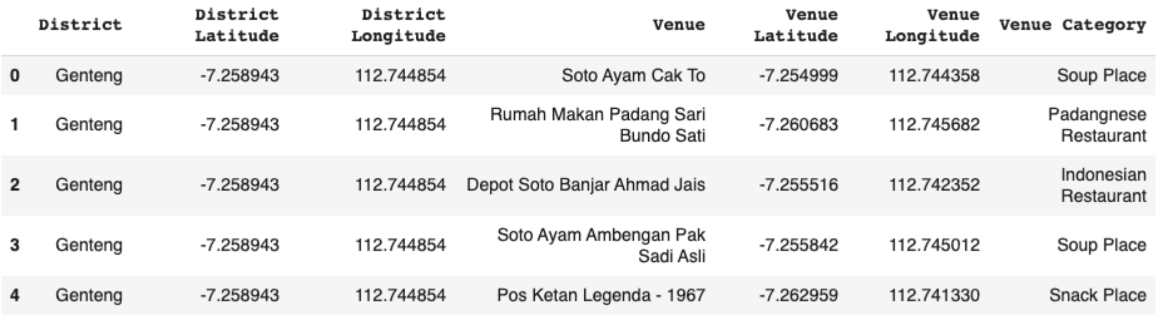
	name	lat	lon
0	Genteng	-7.258943	112.744854
1	Simokerto	-7.239466	112.753292
2	Semampir	-7.213868	112.748414
3	Kenjeran	-7.217600	112.768674
4	Bulak	-7.236022	112.788849
5	Krembangan	-7.228817	112.721646
6	Bubutan	-7.248752	112.727749

Then, using the coordinates on the DataFrame above, I can create a map that plots the districts' centers using Folium.



Using Foursquare, I created a function to return venue around a specific coordinate, which we will apply for each district. The API URL is an important thing to notice. Foursquare provided a parameter to return venues with specific categories, including food. So we will use that parameter. The URL will then be:

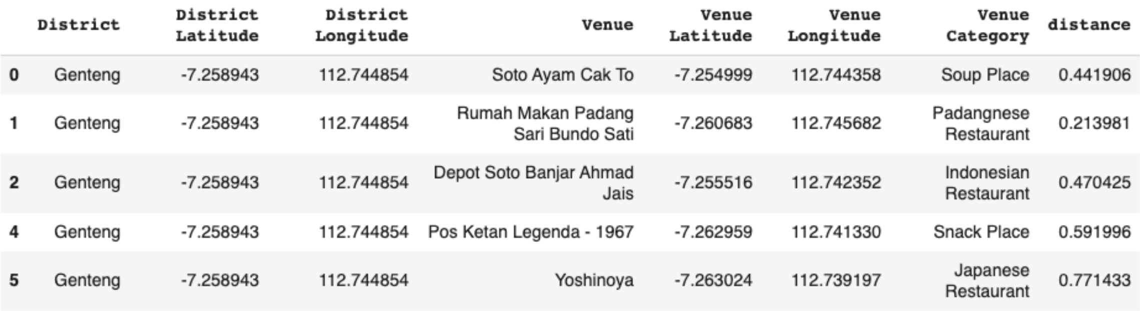
```
        https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret
={}&v={}&ll={},{}&radius={}&limit={}&section=food
```



	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Genteng	-7.258943	112.744854	Soto Ayam Cak To	-7.254999	112.744358	Soup Place
1	Genteng	-7.258943	112.744854	Rumah Makan Padang Sari Bundo Sati	-7.260683	112.745682	Padangnese Restaurant
2	Genteng	-7.258943	112.744854	Depot Soto Banjar Ahmad Jais	-7.255516	112.742352	Indonesian Restaurant
3	Genteng	-7.258943	112.744854	Soto Ayam Ambengan Pak Sadi Asli	-7.255842	112.745012	Soup Place
4	Genteng	-7.258943	112.744854	Pos Ketan Legenda - 1967	-7.262959	112.741330	Snack Place

After retrieving the data for each district with a radius of 2000 meters, the result contains 1815 venues. But, this data is still not correct. This is because some venues might be in proximity of 2000 meters from more than one district point. This can create duplicates. Hence, I will remove duplicates by only keeping the venues with the closest distance to a district.

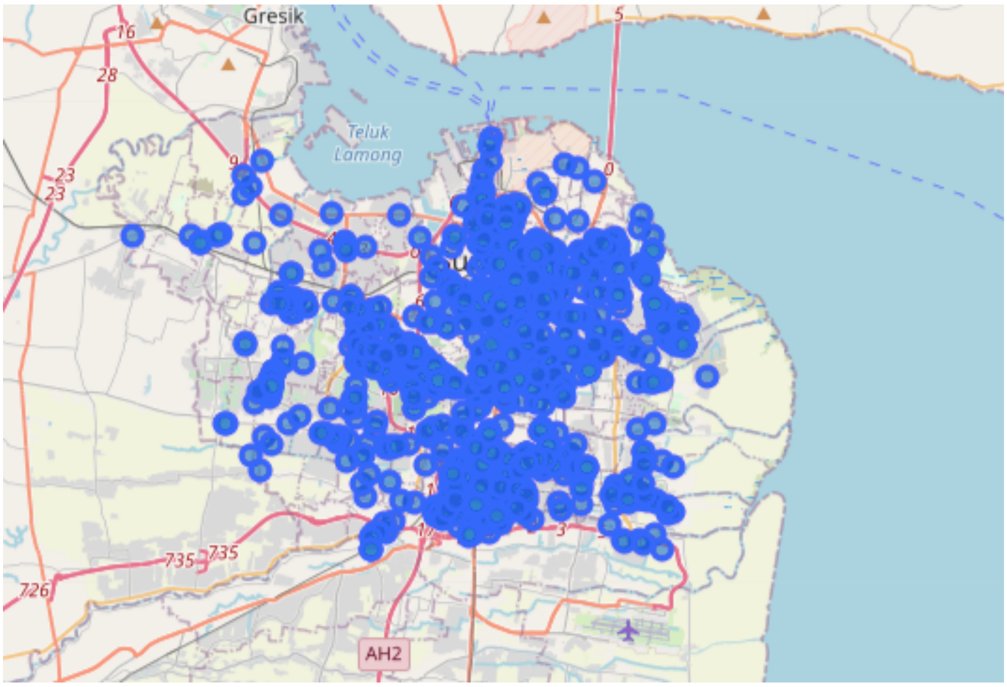
I calculated each duplicated venues' distance from the district center, which will produce this data:



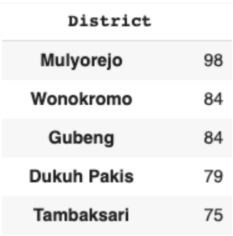
	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	distance
0	Genteng	-7.258943	112.744854	Soto Ayam Cak To	-7.254999	112.744358	Soup Place	0.441906
1	Genteng	-7.258943	112.744854	Rumah Makan Padang Sari Bundo Sati	-7.260683	112.745682	Padangnese Restaurant	0.213981
2	Genteng	-7.258943	112.744854	Depot Soto Banjar Ahmad Jais	-7.255516	112.742352	Indonesian Restaurant	0.470425
4	Genteng	-7.258943	112.744854	Pos Ketan Legenda - 1967	-7.262959	112.741330	Snack Place	0.591996
5	Genteng	-7.258943	112.744854	Yoshinoya	-7.263024	112.739197	Japanese Restaurant	0.771433

Then, I sorted the data ascendingly by distance, and used the drop_duplicate() function. This will drop the duplicate of data except the first item, which will always be the venue with the shortest distance because they are already sorted by distance ascendingly.

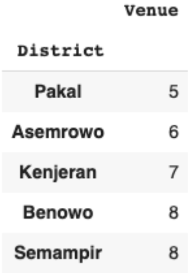
After cleaning duplicates, we can plot a nice-looking map containing all the food-related venues in the city of Surabaya.



We can also see the most and least venues-populated district. Here is the snippet:



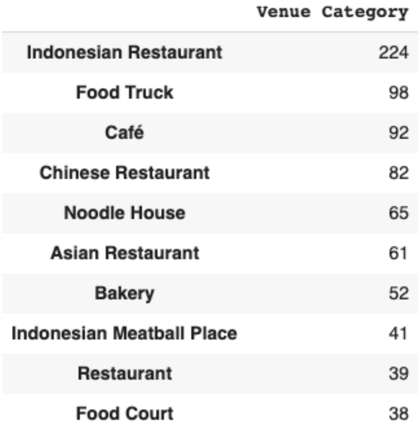
District	
Mulyorejo	98
Wonokromo	84
Gubeng	84
Dukuh Pakis	79
Tambaksari	75



	Venue
District	
Pakal	5
Asemrowo	6
Kenjeran	7
Benowo	8
Semampir	8

As you can see, the food venues are not spread evenly. Some districts have so many venues while other districts barely have any venues. We will use a minimum of top 5 venues category to cluster the districts later.

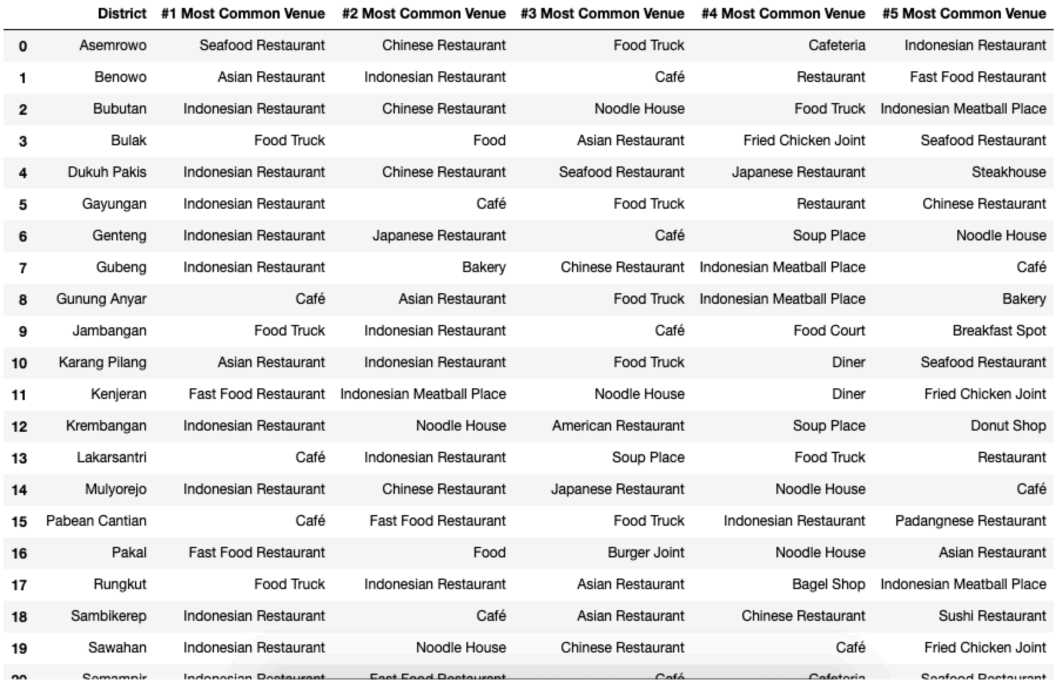
Using the data above, we can also see what are the most frequent food/restaurant types in Surabaya.



	Venue Category
Indonesian Restaurant	224
Food Truck	98
Café	92
Chinese Restaurant	82
Noodle House	65
Asian Restaurant	61
Bakery	52
Indonesian Meatball Place	41
Restaurant	39
Food Court	38

Unsurprisingly, Indonesian restaurant tops the list with over 200+ venues listed. Followed by food truck and cafe, which is obivously common in a metropolitan city like Surabaya. Chinese restaurant is sitting on #4, which can be explained considering that Surabaya is home to many Chinese-Indonesian descendants.

After transforming the data with calculating the mean of occurences of food venues in each district, we can see what is the first 5 of the most popular food category in each district.



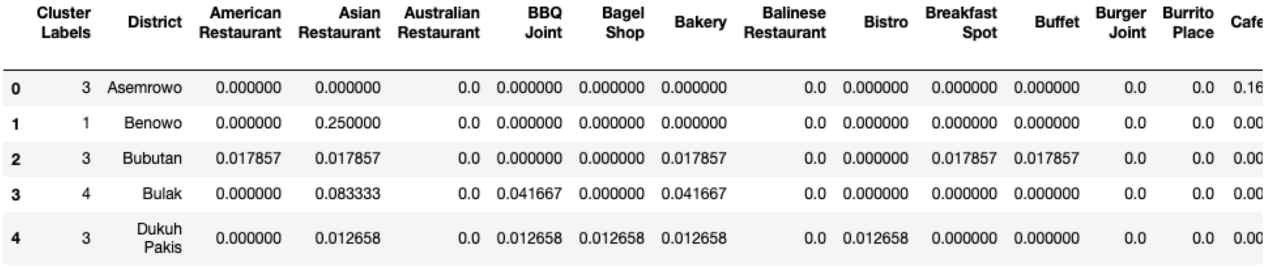
	District	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue
0	Asemrowo	Seafood Restaurant	Chinese Restaurant	Food Truck	Cafeteria	Indonesian Restaurant
1	Benowo	Asian Restaurant	Indonesian Restaurant	Café	Restaurant	Fast Food Restaurant
2	Bubutan	Indonesian Restaurant	Chinese Restaurant	Noodle House	Food Truck	Indonesian Meatball Place
3	Bulak	Food Truck	Food	Asian Restaurant	Fried Chicken Joint	Seafood Restaurant
4	Dukuh Pakis	Indonesian Restaurant	Chinese Restaurant	Seafood Restaurant	Japanese Restaurant	Steakhouse
5	Gayungan	Indonesian Restaurant	Café	Food Truck	Restaurant	Chinese Restaurant
6	Genteng	Indonesian Restaurant	Japanese Restaurant	Café	Soup Place	Noodle House
7	Gubeng	Indonesian Restaurant	Bakery	Chinese Restaurant	Indonesian Meatball Place	Café
8	Gunung Anyar	Café	Asian Restaurant	Food Truck	Indonesian Meatball Place	Bakery
9	Jambangan	Food Truck	Indonesian Restaurant	Café	Food Court	Breakfast Spot
10	Karang Pilang	Asian Restaurant	Indonesian Restaurant	Food Truck	Diner	Seafood Restaurant
11	Kenjeran	Fast Food Restaurant	Indonesian Meatball Place	Noodle House	Diner	Fried Chicken Joint
12	Krembangan	Indonesian Restaurant	Noodle House	American Restaurant	Soup Place	Donut Shop
13	Lakarsantri	Café	Indonesian Restaurant	Soup Place	Food Truck	Restaurant
14	Mulyorejo	Indonesian Restaurant	Chinese Restaurant	Japanese Restaurant	Noodle House	Café
15	Pabean Cantian	Café	Fast Food Restaurant	Food Truck	Indonesian Restaurant	Padangnese Restaurant
16	Pakal	Fast Food Restaurant	Food	Burger Joint	Noodle House	Asian Restaurant
17	Rungkut	Food Truck	Indonesian Restaurant	Asian Restaurant	Bagel Shop	Indonesian Meatball Place
18	Sambikerep	Indonesian Restaurant	Café	Asian Restaurant	Chinese Restaurant	Sushi Restaurant
19	Sawahan	Indonesian Restaurant	Noodle House	Chinese Restaurant	Café	Fried Chicken Joint
20	Semampir	Indonesian Restaurant	Fast Food Restaurant	Café	Cafeteria	Seafood Restaurant

Finally, we can gain more insight from this data with running an unsupervised machine learning algorithm to study and group districts based on the sole criteria of food/restaurant types.

I chose K-Means Clustering Algorithm because it is the most suitable algorithm to use with a data that has high dimensionality. It is also fast to run. Although we can use the elbow method to determine the k, I simply chose the k number to be 5 after trying many times with other k values, because it looks like it makes the most sense for this data.

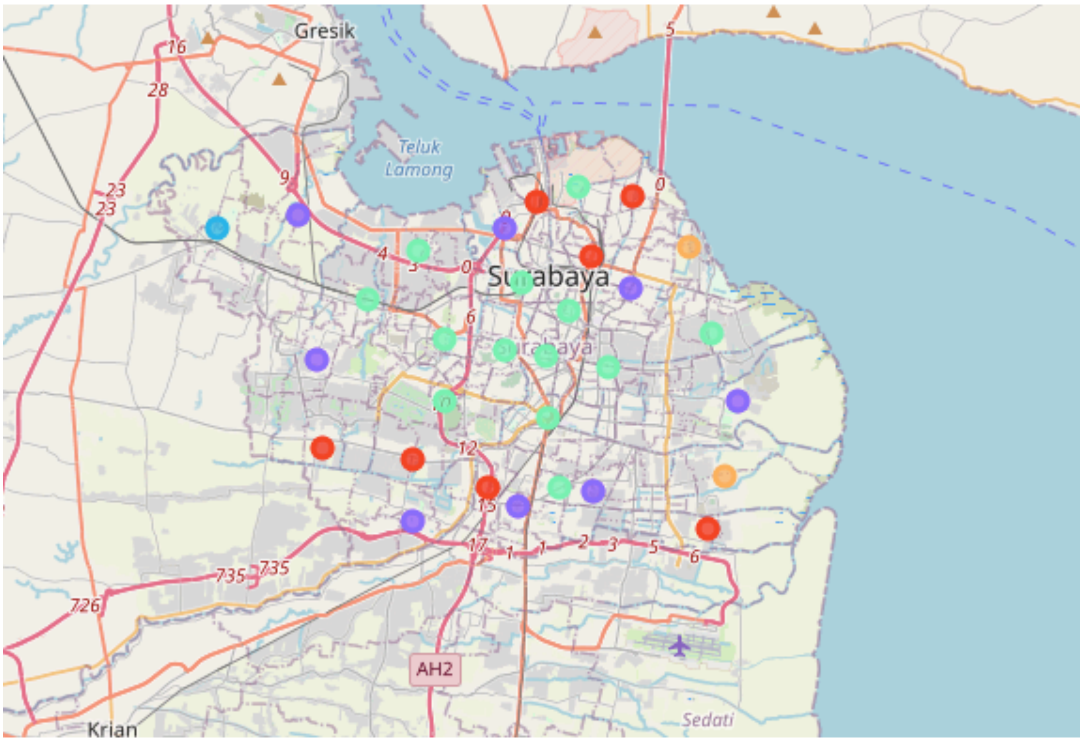
Results

After running the algorithm with our data, we can group the districts of Surabaya into 5 different clusters. Here is the snippet:



	Cluster Labels	District	American Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Balinese Restaurant	Bistro	Breakfast Spot	Buffet	Burger Joint	Burrito Place	Cafe
0	3	Asemrowo	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.16
1	1	Benowo	0.000000	0.250000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.00
2	3	Bubutan	0.017857	0.017857	0.0	0.000000	0.000000	0.017857	0.0	0.000000	0.017857	0.017857	0.0	0.0	0.00
3	4	Bulak	0.000000	0.083333	0.0	0.041667	0.000000	0.041667	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.00
4	3	Dukuh Pakis	0.000000	0.012658	0.0	0.012658	0.012658	0.012658	0.0	0.012658	0.000000	0.000000	0.0	0.0	0.00

We can plot the clusters into a map using Folium. We then can assign a different color to each district to show the cluster difference.

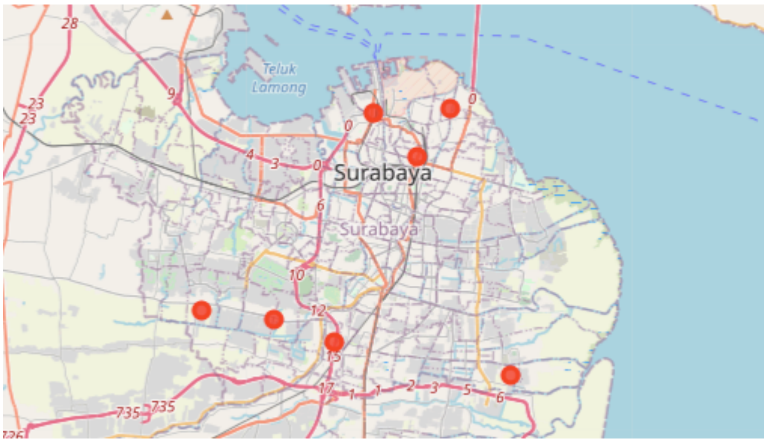


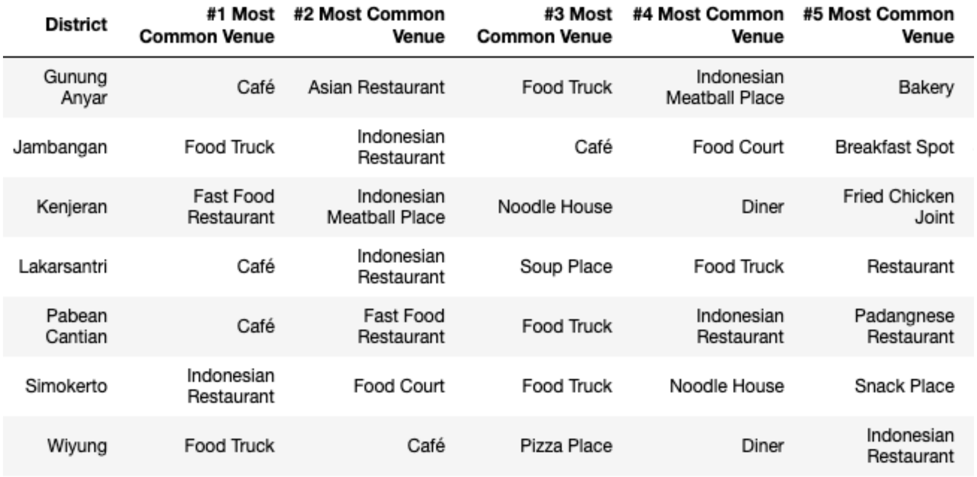
On the map above, we can see that a district's location and venue count number plays a big role to which group it belongs. For example, Pakal district is the only district with the cluster label 2 (labeled as blue), and it is the district with the least venues and the most distant from the city center.

I looked into the data closely and found that the cluster can be divided based on location and the most common food types available.

Cluster 0 – The Café and Food Truck Cluster

This cluster is located in the suburbs of Surabaya near the city edge and dominated by food trucks and cafes.

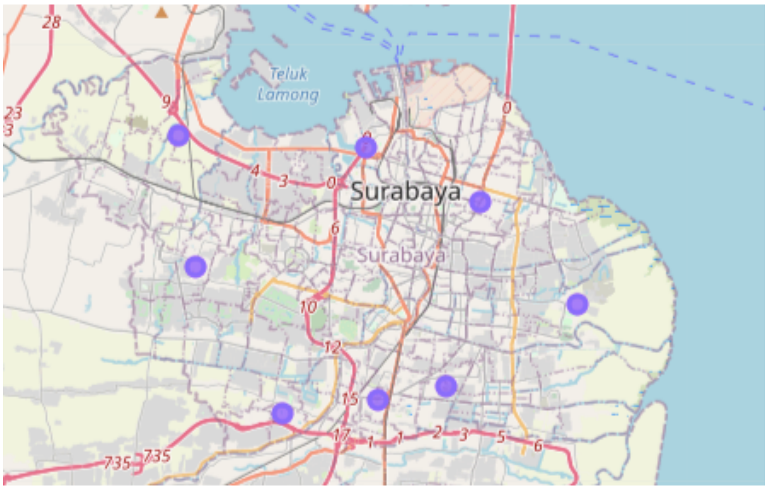


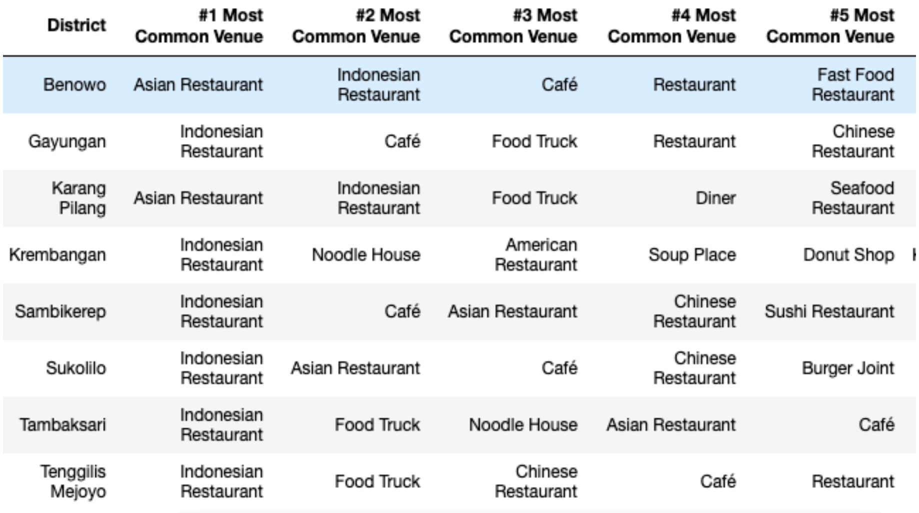


District	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue
Gunung Anyar	Café	Asian Restaurant	Food Truck	Indonesian Meatball Place	Bakery
Jambangan	Food Truck	Indonesian Restaurant	Café	Food Court	Breakfast Spot
Kenjeran	Fast Food Restaurant	Indonesian Meatball Place	Noodle House	Diner	Fried Chicken Joint
Lakarsantri	Café	Indonesian Restaurant	Soup Place	Food Truck	Restaurant
Pabean Cantian	Café	Fast Food Restaurant	Food Truck	Indonesian Restaurant	Padangnese Restaurant
Simokerto	Indonesian Restaurant	Food Court	Food Truck	Noodle House	Snack Place
Wiyung	Food Truck	Café	Pizza Place	Diner	Indonesian Restaurant

Cluster 1 – The Classic Suburban Cluster

This cluster is dominated by various restaurant types, but mostly local types of restaurant. We will call it the classic suburban cluster.



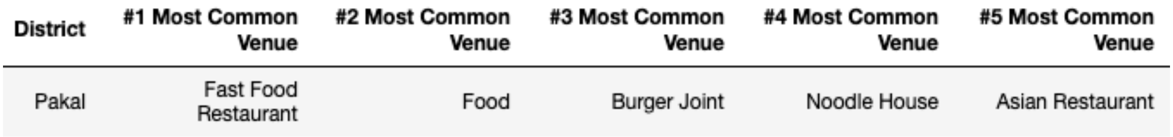


District	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue
Benowo	Asian Restaurant	Indonesian Restaurant	Café	Restaurant	Fast Food Restaurant
Gayungan	Indonesian Restaurant	Café	Food Truck	Restaurant	Chinese Restaurant
Karang Pilang	Asian Restaurant	Indonesian Restaurant	Food Truck	Diner	Seafood Restaurant
Krembangan	Indonesian Restaurant	Noodle House	American Restaurant	Soup Place	Donut Shop
Sambikerep	Indonesian Restaurant	Café	Asian Restaurant	Chinese Restaurant	Sushi Restaurant
Sukolilo	Indonesian Restaurant	Asian Restaurant	Café	Chinese Restaurant	Burger Joint
Tambaksari	Indonesian Restaurant	Food Truck	Noodle House	Asian Restaurant	Café
Tenggilis Mejoyo	Indonesian Restaurant	Food Truck	Chinese Restaurant	Café	Restaurant

Cluster 2 – The Pakal Cluster

This cluster only contains Pakal district. The district's location is very far from the city center and the venue choice is also "strange" relative to other clusters, dominated by Western and fast food restaurants. This district might be an outlier.

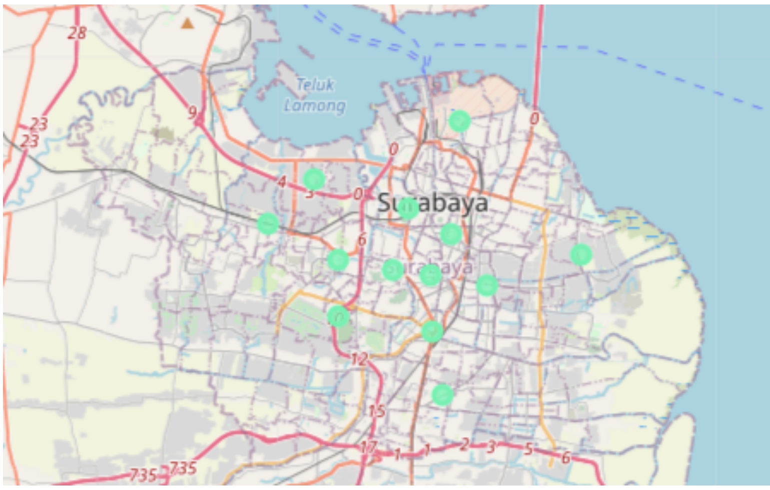
Teluk
Lamong
Surabaya
Surabaya

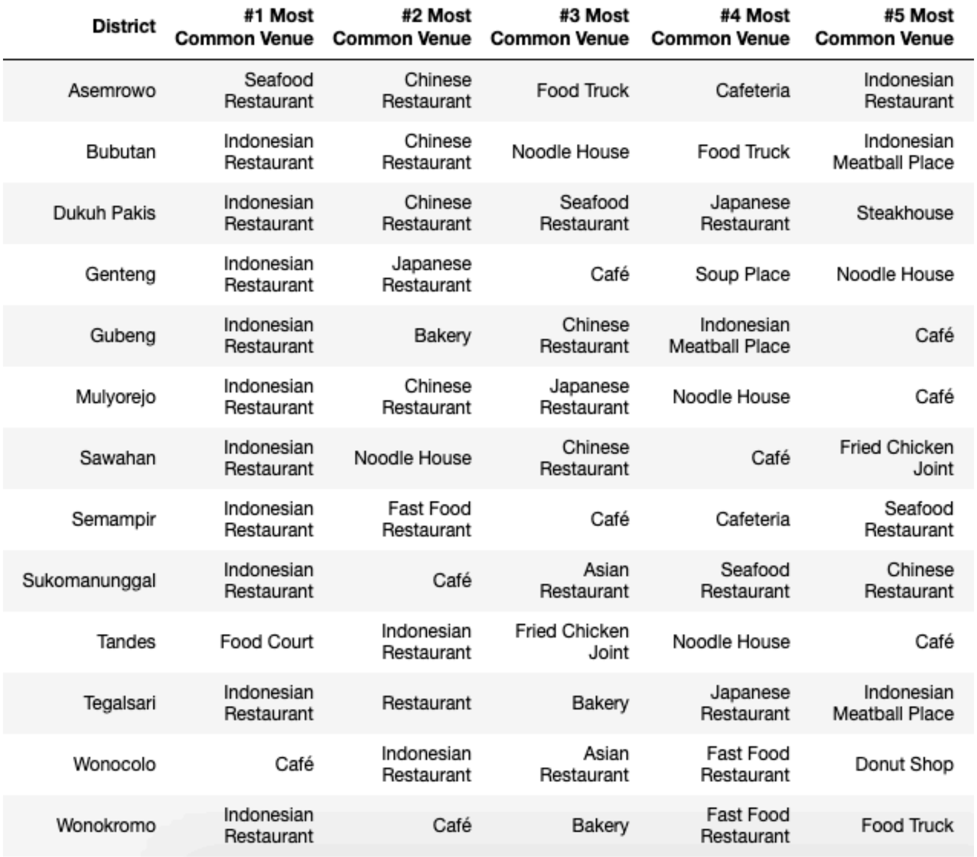


District	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue
Pakal	Fast Food Restaurant	Food	Burger Joint	Noodle House	Asian Restaurant

Cluster 3 – The Downtown Cuisine Cluster

This cluster has the most districts and most of them are located in the city center, hence the most popular venue types are restaurants, foreign cuisine restaurants, cafes, foodcourts, and other "modern metropolitan" food palace.

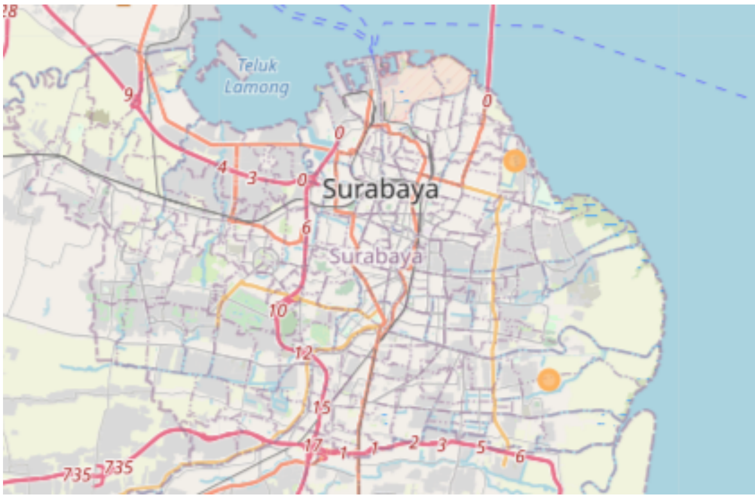


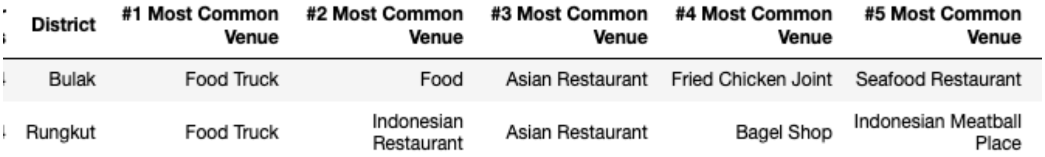


District	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue
Asemrowo	Seafood Restaurant	Chinese Restaurant	Food Truck	Cafeteria	Indonesian Restaurant
Bubutan	Indonesian Restaurant	Chinese Restaurant	Noodle House	Food Truck	Indonesian Meatball Place
Dukuh Pakis	Indonesian Restaurant	Chinese Restaurant	Seafood Restaurant	Japanese Restaurant	Steakhouse
Genteng	Indonesian Restaurant	Japanese Restaurant	Café	Soup Place	Noodle House
Gubeng	Indonesian Restaurant	Bakery	Chinese Restaurant	Indonesian Meatball Place	Café
Mulyorejo	Indonesian Restaurant	Chinese Restaurant	Japanese Restaurant	Noodle House	Café
Sawahan	Indonesian Restaurant	Noodle House	Chinese Restaurant	Café	Fried Chicken Joint
Semampir	Indonesian Restaurant	Fast Food Restaurant	Café	Cafeteria	Seafood Restaurant
Sukomanunggal	Indonesian Restaurant	Café	Asian Restaurant	Seafood Restaurant	Chinese Restaurant
Tandes	Food Court	Indonesian Restaurant	Fried Chicken Joint	Noodle House	Café
Tegalsari	Indonesian Restaurant	Restaurant	Bakery	Japanese Restaurant	Indonesian Meatball Place
Wonocolo	Café	Indonesian Restaurant	Asian Restaurant	Fast Food Restaurant	Donut Shop
Wonokromo	Indonesian Restaurant	Café	Bakery	Fast Food Restaurant	Food Truck

Cluster 4 – The East Food Trucks Cluster

This cluster is dominated by food trucks near the east coast of Surabaya. Hence we will name it the east food trucks cluster.





District	#1 Most Common Venue	#2 Most Common Venue	#3 Most Common Venue	#4 Most Common Venue	#5 Most Common Venue
Bulak	Food Truck	Food	Asian Restaurant	Fried Chicken Joint	Seafood Restaurant
Rungkut	Food Truck	Indonesian Restaurant	Asian Restaurant	Bagel Shop	Indonesian Meatball Place

Discussion

In working with the data, the longest part in data science is not the modelling or exploring. But by simply having a data that is hard to get and clean, you will spend more time to adjust and clean the data to get the best result. Data preparation really decides the quality of data to be explored and modelled.

I'm also amazed by the availability of tools out there to do data science. Most of them are open source and of course free of charge. I'm becoming more understanding that in data science, the best asset is not the machine or tools, but our brains and the quality of the data.

Conclusion

Based on the problem we defined earlier, we wanted to find out if there is any pattern in the type of restaurants (including cafés, foodtrucks, etc) located in every district and neighborhood of Surabaya. In this report, we found that:

1. Some districts represent certain types of cuisines/restaurants more than the others. For example, Pakal district has more western and fast-food restaurant, but Sukomanunggal district has more Asian, Chinese, and Indonesian restaurants.

2. Surabaya districts can be grouped based on criteria of food/restaurant types. There is 5 clusters.

Acknowledgement & sources

I used many code from IBM Data Science Professional Certificate courses. Especially in the Week 2 of the Applied DS Capstone course regarding Foursquare API.

I also get a lot of inspiration from the publication of Minh Nguyen. I also would like to thank Towardsdatascience, Stackoverflow, Machine Learning Mastery, and other data science tutorial contributors. Thank you for making learning data science easier than ever!