1. Introduction / Business Problem

In this project, a certain part of the Kadikoy district in Istanbul, Turkey will be analyzed. The region to be analyzed is a dynamic and lively part of the city with lots of cafes, restaurants and bars, as well as cultural venues and also shops and some residential areas as well. It is quite a dense part of the city. The region to be analyzed is not homogenous in terms of the types of the venues though, so this study aims to find out the similar and different parts and to cluster the similar ones together.

This study can be interesting and beneficial for a number of groups. A tourist who would like to spend some time in this area can use it to have a better idea of what to expect from different parts. He may create his walking route to see the parts with different characteristics.

Besides, an investor who is planning to purchase or rent a property in this area can find it beneficial to look at this analysis to better understand the area. Depending on the category of his investment, he can eliminate some parts or focus on some other parts.

Someone living in Kadikoy or Istanbul can also be interested in this study to have a better understanding of this region and to learn more about the city that she is living in.

Additionally, this study can serve as a base for further analysis of this area or comparison with other areas.

Below is a map of the region to be analyzed. This area consists of parts of three neighborhoods (Caferaga, Osmanaga and Rasimpasa).

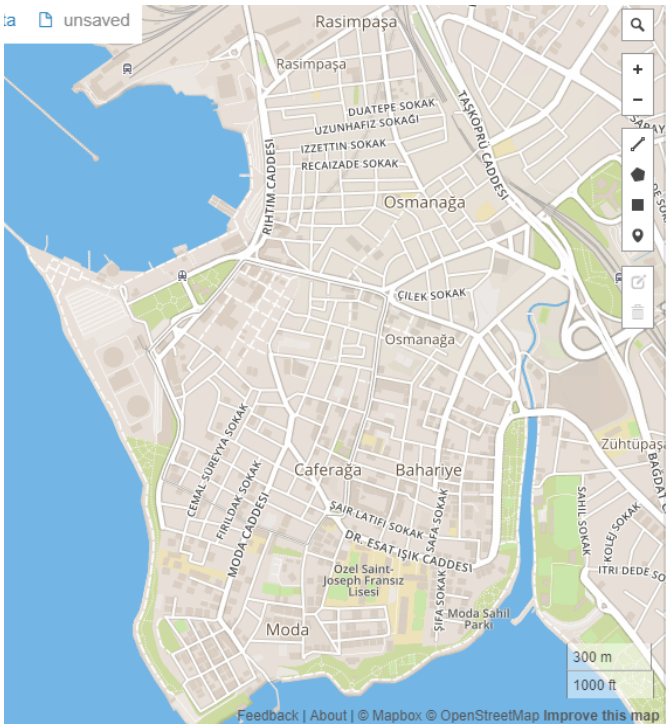


Image: The part of Kadikoy, Istanbul, Turkey to be analyzed in this study

(Image reference: http://geojson.io/#map=15/40.9886/29.0218)

2. Data

There are two main sources of data that is used in this study. One is the location file created using the geojson.io website and the other one is the Foursquare Places API.

In order to analyze the region, I needed a file with location points in it, in the region to be analyzed. I was not able to find such a file and decided to create by myself.

It is possible to draw markers on the desired location on a map using the geojson.io website. One thing to decide on is the distance between the markers. I decided it to be around (not less than and as close as possible to) 300 meters. In combination with this, I set the 'radius' parameter using the Foursquare Places API to be 150 meters. This way, I tried to make sure that:

- The area to be analyzed is covered as much as possible
- No venues are duplicated (a venue is counted and considered only once)

The resulting map is as below:

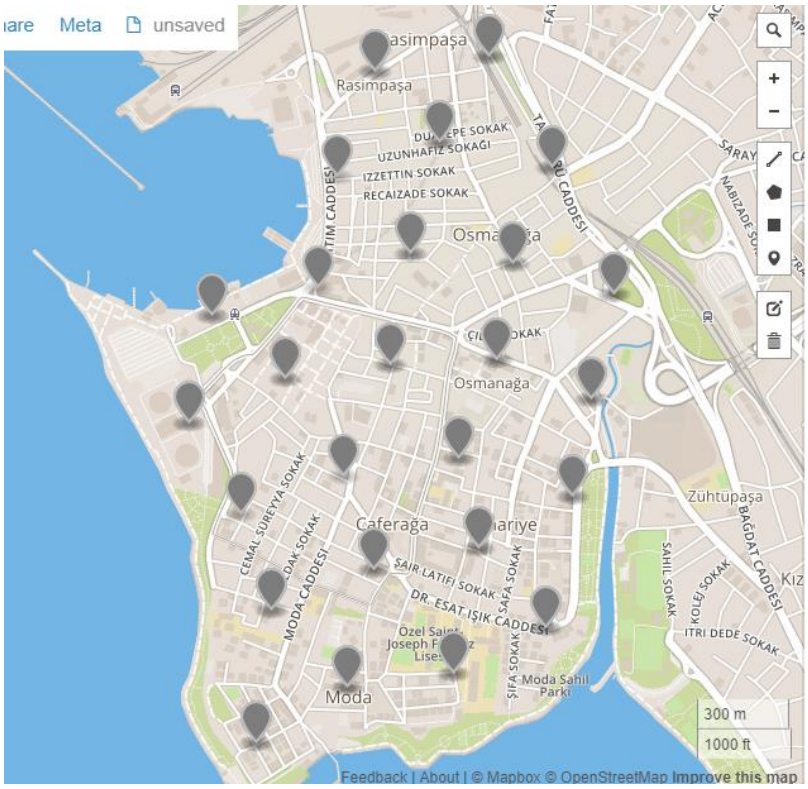


Image: Map of the region to be analyzed, with markers on it

Of course it is possible to select different distances instead of 300 meters (for min distance between two points) and as 150 meters as radius for the Foursquare Places API. I decided that this is suitable for the purposes of this study.

Another thing I had to do was to name these markers, as they do not correspond to any specific neighborhoods or location units. I named them as below, starting with A1 at the bottom-left and increasing as 2,3,4 to the right and as B,C,D upwards as you can see below.

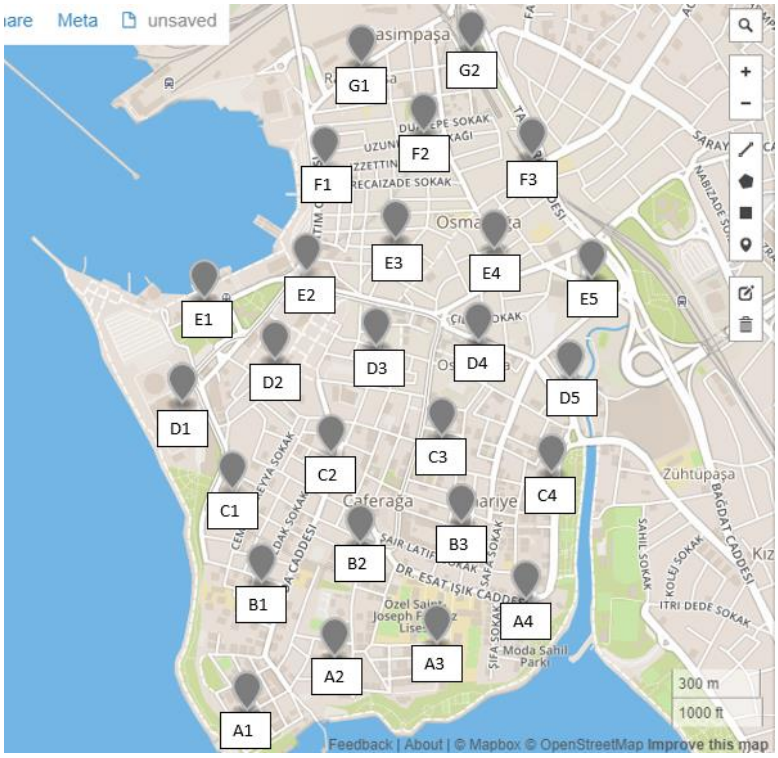


Image: Naming of the markers on the map

I saved this as a csv file where each line in the file contains the marker name, together with the latitudes and the longitudes of the marker, which is what is needed for the analysis.

Below you can see a screenshot from this file:

	lon	lat	marker-color	marker-size	marker-symbol	name
2	29.02231693267822	40.979962866583264	#7e7e7e	medium		A1
3	29.02538537979126	40.9814045825477	#7e7e7e	medium		A2
4	29.028990268707275	40.98174475812385	#7e7e7e	medium		A3
5	29.03214454650879	40.98291106106472	#7e7e7e	medium		A4
6	29.029870033264157	40.984952041582275	#7e7e7e	medium		B3
7	29.026308059692383	40.9844013070302 6	#7e7e7e	medium		B2

Image 3: Screenshot of the csv file with location data

The other data source that is used in this study is the Foursquare Places API. The code generated with the required parameters (which includes User credentials, coordinates of the point whose surrounding will be searched and the radius of the search) sends a request to this API and receives back the requested data about the venues. This data includes the name, address, category and certain other properties about the venue. Afterwards, this received data is further processed to extract meaningful information in the analysis.

3. Methodology

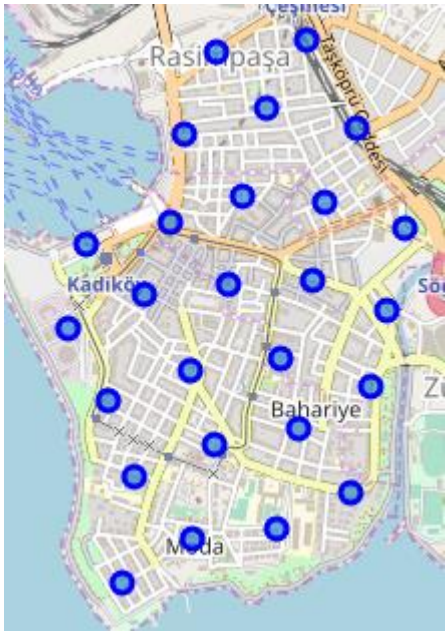
In the .csv file which is explained in the Data section, there were 26 points including the names and coordinate information. I checked to make sure that these are imported correctly:

```
(26, 3)
```

Out[3]:

	lon	lat	name
0	29.022317	40.979963	A1
1	29.025385	40.981405	A2
2	29.028990	40.981745	A3
3	29.032145	40.982911	A4
4	29.029870	40.984952	B3

The points are located on the map created by Folium correctly:



Initially, the Foursquare information for only 1 location point (A1) was checked:

	name	categories	lat	lng
0	Cafe MOON	Café	40.980442	29.021340
1	Cafe Los Manços	Café	40.981140	29.022968
2	Moda Parkı	Park	40.980748	29.021341
3	Nefess İstanbul	Yoga Studio	40.980128	29.022858
4	Tatlı Mesai	Café	40.980285	29.023170

```
33 venues were returned by Foursquare.
```

It is seen that 33 venues are retrived (within the radius of 150m of the given location point). It is reasonable.

Then the data for all of the location points is retrieved from the Foursquare Places API.

```
(687, 7)
```

Out[23]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A1	40.979963	29.022317	Cafe MOON	40.980442	29.021340	Café
1	A1	40.979963	29.022317	Cafe Los Manços	40.981140	29.022968	Café
2	A1	40.979963	29.022317	Moda Parkı	40.980748	29.021341	Park
3	A1	40.979963	29.022317	Nefess İstanbul	40.980128	29.022858	Yoga Studio
4	A1	40.979963	29.022317	Tatlı Mesai	40.980285	29.023170	Café

A total of 687 venues were retrived. (The limit was set to 10.000, so it was not reached)

In [25]: `kadikoy_venues`

Out[25]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A1	40.979963	29.022317	Cafe MOON	40.980442	29.021340	Café
1	A1	40.979963	29.022317	Cafe Los Manços	40.981140	29.022968	Café
2	A1	40.979963	29.022317	Moda Parkı	40.980748	29.021341	Park
3	A1	40.979963	29.022317	Nefess İstanbul	40.980128	29.022858	Yoga Studio
4	A1	40.979963	29.022317	Tatlı Mesai	40.980285	29.023170	Café
5	A1	40.979963	29.022317	Meze Moda	40.980325	29.023182	Mediterranean Restaurant
6	A1	40.979963	29.022317	MEKAN Cafe	40.980362	29.021181	Restaurant
7	A1	40.979963	29.022317	Pizza Locale	40.980556	29.023853	Pizza Place
8	A1	40.979963	29.022317	MODACTIVE	40.978970	29.023144	Gym

I just had a look through the list and it seems ok.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
A1	33	33	33	33	33	33
A2	15	15	15	15	15	15
A3	1	1	1	1	1	1
A4	9	9	9	9	9	9
B1	21	21	21	21	21	21
B2	71	71	71	71	71	71
B3	2	2	2	2	2	2
C1	16	16	16	16	16	16
C2	77	77	77	77	77	77
C3	26	26	26	26	26	26
C4	7	7	7	7	7	7
D1	8	8	8	8	8	8
D2	66	66	66	66	66	66

There are variations between the number of venues, such as some are quite small as 1 or 2 while some are in the range of 60-70's. But this is the data we received from Foursquare and since the Limit is not reached, this is not an issue for the analysis.

```
There are 143 unique categories.
```

We can also see that there are a total of 143 unique categories.

The next step would be to use the k-Means Clustering machine learning algorithm to cluster the similar regions together. K-Means Clustering is selected since it is a suitable unsupervised algorithm for the purpose of clustering similar items in this case.

First, one hot encoding is applied on the dataset to be able to apply k-Means Clustering later on.

	Neighborhood	Accessories Store	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Bagel Shop	Bakery	...
0	A1	0	0	0	0	0	0	0	0	0	...
1	A1	0	0	0	0	0	0	0	0	0	...
2	A1	0	0	0	0	0	0	0	0	0	...
3	A1	0	0	0	0	0	0	0	0	0	...
4	A1	0	0	0	0	0	0	0	0	0	...

Then the rows are grouped together so that each row represents a location point (header Neighborhood) and the values represent the means.

	Neighborhood	Accessories Store	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	Asi Restaura
0	A1	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
1	A2	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
2	A3	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
3	A4	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
4	B1	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
5	B2	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
6	B3	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
7	C1	0.000000	0.000000	0.000000	0.062500	0.000000	0.0000
8	C2	0.000000	0.000000	0.000000	0.012987	0.000000	0.0000
9	C3	0.000000	0.000000	0.153846	0.038462	0.000000	0.0000
10	C4	0.000000	0.000000	0.000000	0.142857	0.000000	0.0000
11	D1	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000

The top 5 venues in each neighborhood is listed (sample below):

```
----A1----
           venue  freq
0           Café  0.18
1    Coffee Shop  0.15
2            Pub  0.09
3    Yoga Studio  0.06
4  Hot Dog Joint  0.06


----A2----
          venue  freq
0          Café  0.13
1  Music Venue  0.07
2     Boutique  0.07
3  Coffee Shop  0.07
4  Flower Shop  0.07
```

The top 10 most common venues are listed in each neighborhood:

	Neighborhood	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
0	A1	Café	Coffee Shop	Pub	Yoga Studio	Tea Room	Diner	Hot Dog Joint	Mediterranean Restaurant	Gym	Fast Food Restaurant
1	A2	Café	Music Venue	Seafood Restaurant	Boutique	Park	Turkish Restaurant	Museum	Hookah Bar	Flower Shop	Thrift / Vintage Store
2	A3	Restaurant	Yoga Studio	Fast Food Restaurant	Furniture / Home Store	Fried Chicken Joint	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Falafel Restaurant
3	A4	Pool	Café	Hotel	Athletics & Sports	Restaurant	Fast Food Restaurant	Health & Beauty Service	Park	Tennis Court	Falafel Restaurant
4	B1	Café	Coffee Shop	Bakery	Dessert Shop	Chocolate Shop	Sausage Shop	Steakhouse	Food & Drink Shop	Music Venue	Sushi Restaurant

The neighborhoods are then grouped into 5 clusters by k-means Clustering algorithm based on their similarities. K is selected to be 5 as a suitable value here.

As a result of the clustering, the 26 regions are grouped into 5 cluster based on their similarities and differences. The column “Cluster Labels” shows the Clusters ranging from 0 to 4, in the sample table below:

	lon	lat	name	Cluster Labels	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
0	29.022317	40.979963	A1	2	Café	Coffee Shop	Pub	Yoga Studio	Tea Room	Diner	Hot Dog Joint	Mediterranean Restaurant	Gym	Fast Food Restaurant
1	29.025385	40.981405	A2	2	Café	Music Venue	Seafood Restaurant	Boutique	Park	Turkish Restaurant	Museum	Hookah Bar	Flower Shop	Thrift / Vintage Store
2	29.028990	40.981745	A3	4	Restaurant	Yoga Studio	Fast Food Restaurant	Furniture / Home Store	Fried Chicken Joint	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Falafel Restaurant
3	29.032145	40.982911	A4	2	Pool	Café	Hotel	Athletics & Sports	Restaurant	Fast Food Restaurant	Health & Beauty Service	Park	Tennis Court	Falafel Restaurant
4	29.029870	40.984952	B3	3	Music Venue	Tattoo Parlor	Yoga Studio	Gastropub	Fried Chicken Joint	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Fast Food Restaurant

4. Results

As a result of the k-means clustering algorithm, the 26 areas are grouped into 5 clusters. These clusters can be seen in the map below:

The clusters can be differentiated by their colors, namely blue, red, purple, green and orange.

Just as a reminder, please note that each point represents the center of the area with a radius of 150 meters around it. The circles created by this definition do not overlap, yet the area that is not covered in between the circles is quite small, since the distance between the points is slightly more than 300 meters.

We can see that majority -19 out of the 26 areas- are blue. Then there are 3 red, 2 purple, 1 green and 1 orange areas.

We can see the characteristics of each area as below:

Blue (Cluster 2):

	name	Cluster Labels	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
0	A1	2	Café	Coffee Shop	Pub	Yoga Studio	Tea Room	Diner	Hot Dog Joint	Mediterranean Restaurant	Gym	Fast Food Restaurant
1	A2	2	Café	Music Venue	Seafood Restaurant	Boutique	Park	Turkish Restaurant	Museum	Hookah Bar	Flower Shop	Thrift / Vintage Store
3	A4	2	Pool	Café	Hotel	Athletics & Sports	Restaurant	Fast Food Restaurant	Health & Beauty Service	Park	Tennis Court	Falafel Restaurant
5	B2	2	Café	Coffee Shop	Dessert Shop	Chocolate Shop	Restaurant	Breakfast Spot	Italian Restaurant	Comfort Food Restaurant	Bookstore	Gym / Fitness Center
6	B1	2	Café	Coffee Shop	Bakery	Dessert Shop	Chocolate Shop	Sausage Shop	Steakhouse	Food & Drink Shop	Music Venue	Sushi Restaurant
7	C3	2	Café	Arcade	Theater	Restaurant	Coffee Shop	Korean Restaurant	Tea Room	Opera House	Movie Theater	Furniture / Home Store
8	D1	2	Hotel	Café	Hotel Bar	Restaurant	Bistro	Spa	Park	Yoga Studio	Flower Shop	Food Truck
9	D2	2	Bar	Coffee Shop	Pub	Café	Bookstore	Comfort Food Restaurant	Meyhane	Kebab Restaurant	Turkish Restaurant	Music Venue
10	D3	2	Café	Pub	Bar	Meyhane	Coffee Shop	Music Venue	Pide Place	Theater	Bistro	Tattoo Parlor
11	D4	2	Café	Pub	Coffee Shop	Performing Arts Venue	Breakfast Spot	Pizza Place	Bar	Comfort Food Restaurant	Turkish Restaurant	Clothing Store
12	D5	2	Theater	Nightclub	Café	Sports Club	Coffee Shop	Outdoor Sculpture	Bakery	Restaurant	Burrito Place	Meyhane
13	E5	2	Kebab Restaurant	Mobile Phone Shop	Café	Dessert Shop	Bakery	Pilavcı	Miscellaneous Shop	Theater	Pub	Electronics Store

The dominant venue in this cluster is Cafe as we can see. Also, coffee shops, pubs and bars are also common. As this is the biggest cluster, we can say that the whole area is mostly a heaven of Cafes.

Red (Cluster 0):

	name	Cluster Labels	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
16	E1	0	Café	Plaza	Seafood Restaurant	City Hall	Hot Dog Joint	Tea Room	Yoga Studio	Fast Food Restaurant	Food Court	Food & Drink Shop
17	F2	0	Café	Breakfast Spot	Coffee Shop	Restaurant	Art Gallery	Pub	Seafood Restaurant	Food & Drink Shop	Bistro	Dessert Shop
19	G2	0	Café	Vegetarian / Vegan Restaurant	Dance Studio	Fast Food Restaurant	Furniture / Home Store	Fried Chicken Joint	Food Truck	Food Court	Food & Drink Shop	Flower Shop

The Cafes are still the most common in this cluster, but the other common venues differ from the Blue (Cluster 2) regions.

Purple (Cluster 1):

	name	Cluster Labels	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
22	E2	1	Coffee Shop	Electronics Store	Burger Joint	Mac & Cheese Joint	Doner Restaurant	Pastry Shop	Pizza Place	Sandwich Place	Clothing Store	Candy Store
23	F1	1	Bed & Breakfast	Medical Center	Café	Burger Joint	Sandwich Place	Steakhouse	Breakfast Spot	Food Truck	Bookstore	Coffee Shop

This is a different cluster where we dont see Cafes as the most common venue, but there are other common venus like Bed&Breakfast and Electronics Store which we dont see much in the other clusters.

Green (Cluster 3):

	name	Cluster Labels	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
4	B3	3	Music Venue	Tattoo Parlor	Yoga Studio	Gastropub	Fried Chicken Joint	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Fast Food Restaurant

This is a cluster with only one member. Actually we also know that there are only two venues in this cluster, as can be seen in a list in the Methodology section. Those are Music Venue and Tattoo Parlor. So this is a specific cluster with only these two venues.

Orange (Cluster 4):

	name	Cluster Labels	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
2	A3	4	Restaurant	Yoga Studio	Fast Food Restaurant	Furniture / Home Store	Fried Chicken Joint	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Falafel Restaurant

This one is also a specific cluster with only one venue in it, and it is a Restaurant. This makes it unique and different from the other clusters.

Looking at the results, in general we can say that the analyzed area is dominated mainly by Cafes. There are relatively few areas which do not have much Cafes. The Cafes are followed by Coffee Shops, Restaurants, Pubs/Bars and Dessert Shops.

5. Discussion

We obtained one very big cluster and then 4 smaller clusters. This shows that most of the parts have more common characteristics than differences. If we had, for instance 5 clusters with 5-6 parts in each, it would have been a different case.

Another observation is that, the venue categories are not completely standardized. For instance, there venue categories like Mediterranean Restaurant, Seafood Restaurant, Falafel Restaurant and Restaurant. At the end, these are all restaurants, but they are treated as different categories. Or the difference betweeen a Cafe and a Coffee Shop is not very clear. By further preprocessing, such cases can be classified into one group and this may change the results.

Some areas have as small as one or two venues while some have 60-70 venues. It can be worthwhile to check further whether there are really one or two values, or these are missing in Foursquare Places API.

In terms of recommendations, for a tourist or for someone who is interested in exploring this area, it may be worthwhile for him to see the smaller clusters, since these are areas which are relatively different from the other parts.

For an investor, depending on the type of investment, he can use this study as a basis to deepen his analysis based on the type of investment he is planning.

6. Conclusion

In this study, we analyzed part of the Kadikoy district in Istanbul, Turkey.

26 points, each with about 300 meters to each other, are selected. The data about the venues within the circles, where these points are the centers and the radius is 150 meters is retrieved from the Foursquare Places API.

A total of 687 venues were retrived and this corresponds to 143 unique categories.

The venues in each region are sorted according to their frequencies, and then they are grouped into 5 clusters using the k-means clustering machine learning algorithm.

The characteristics of the resulting clusters are defined and these clusters are plotted on a map.

This study shows the powerful potential of using the location information data. The analysis done here can be deepened further down, by taking other venue characteristics such as user ratings into account, or by focusing on a certain venue such as Restaurants, by increasing the number of location points or extending the analyzed area.

I hope it will give a different perspective of looking at this area and to the usage of location data to the reader.

Mehmet Aydinel

26.06.2019

Istanbul, Turkey