Applied Data Science Capstone

Final Assignment

Suggesting a nice neighborhood in Kadikoy

Gokhan Ince

09/17/2019

INTRODUCTION

I am a data scientist in a consultancy company. My project targets the people who wants to move to Istanbul, Turkey. I will try to give them many suggestions based on their preferences and expectations. I will try to find them the neighborhoods that they can enjoy to live.

BUSINESS PROBLEM

Istanbul is a big city and it is very multicultural. In this project, i will focus on a client's need and try to find a good neighborhood where he can enjoy. But also, I will cluster the neighborhoods based on the venues they have. We will be able to suggest about which neighborhoods can make them happier. The neighborhoods should satisfy the needs of my clients. The clients who has different backgrounds would like live in a neighborhoods which can feel them comfortable. I should be able to give them suggestions to satisfy their needs.

My client Hans is currently living in Berlin, Germany. His company promoted him to be the general manager for one of their branch office. The company office is in Kadikoy, Istanbul. Hans wants to live close to the office and he prefers to live in Kadikoy. He is asking for our suggestions about neighborhoods. He wants to live in an enjoyable neighborhood. He likes bars, pubs, restaurants etc. and we must give him the suggestions. Also, he wants to be close to public transportation like bus, metro and trains.

DATA AND PREPARATION

I have some ideas about detecting the required data. First, I will detect the zip code, latitude and longitude values of the neighborhoods. I will convert and keep them in a data frame. After that, I will use some loops to get information from Foursquare API by using the latitude and longitude information of the neighborhoods. After that, i will merge those data frames based on the postal codes. Then I will create new categories based on venue categories. At last, I will detect the client needs and then I will be able to suggest them neighborhoods to move based on their needs.

First, I will import the zip code, latitude and longitude information from a csv file. I got the information from the internet and converted to the csv format.

```
In [4]: kadikoy_df.head()
```

Out[4]:

	PostCode	Borough	Neighbourhood	Latitude	Longitude
0	34744	KADIKOY	BOSTANCI ...	40.957850	29.095760
1	34728	KADIKOY	CADDEBOSTAN ...	40.966740	29.062889
2	34710	KADIKOY	CAFERAGA ...	40.985741	29.024500
3	34722	KADIKOY	EGITIM ...	40.989441	29.049490
4	34722	KADIKOY	HASANPASA ...	40.996230	29.044650

This data includes the 21 neighborhoods in Kadikoy. It has latitude and longitude information. It is important to pull data from Foursquare. I will use them to create endpoints to pull data.

After that, I prepared the endpoints and pull the necessary info by using Foursquare API. Then, I pull the data from Foursquare and merged it with my first data frame.

This data frame has venue name, categories, latitude and longitude information and some other information. Now, our data was ready. I will also work on the data during the analysis stage. It is just the raw data to use for the analysis.

```
In [14]: venue_data.head(5)
```

Out[14]:

	Name	Categories	Latitude	Longitude	Distance	Zip Code	Borough	Neighborhood	Center Latitude
0	Safranbolu Fırını	Bakery	40.95879801089139	29.094185061159358	169	34744	KADIKOY	BOSTANCI	40.95785
1	Bakıroğlu Gurme	Breakfast Spot	40.95776765987196	29.097471013753243	144	34744	KADIKOY	BOSTANCI	40.95785
2	Dondurmacı Yaşar Usta	Ice Cream Shop	40.9571934765289	29.09702105815656	128	34744	KADIKOY	BOSTANCI	40.95785
3	Studyo pilates	Athletics & Sports	40.95801339918345	29.09609239972061	33	34744	KADIKOY	BOSTANCI	40.95785
4	Ekler İstanbul Bostancı	Dessert Shop	40.95770450717472	29.096905420764813	97	34744	KADIKOY	BOSTANCI	40.95785

METHODOLOGY

In this project I will use Foursquare data, analyze it and try to get meaningful results. The venues will be within 500 meters to the neighborhood centers.

In first step, I pulled the data from Foursquare and matched them with neighborhood data. Also, I created some filtered data frames. There is a purpose for each data frame.

In second step, I will start to analyze the data. I will try to get the most popular venue categories, frequency, filtering with desired venue types and then visualizing them. I will also, show the findings on a map.

In third and final step, I will cluster the neighborhoods by using the most popular venue types and frequencies in the neighborhood, detect the clusters and show them on a map. Wish me luck.

ANALYSIS

In this stage, I will analyze my data in detail. I will perform some basic explanatory data analysis and derive some additional info from my raw data.

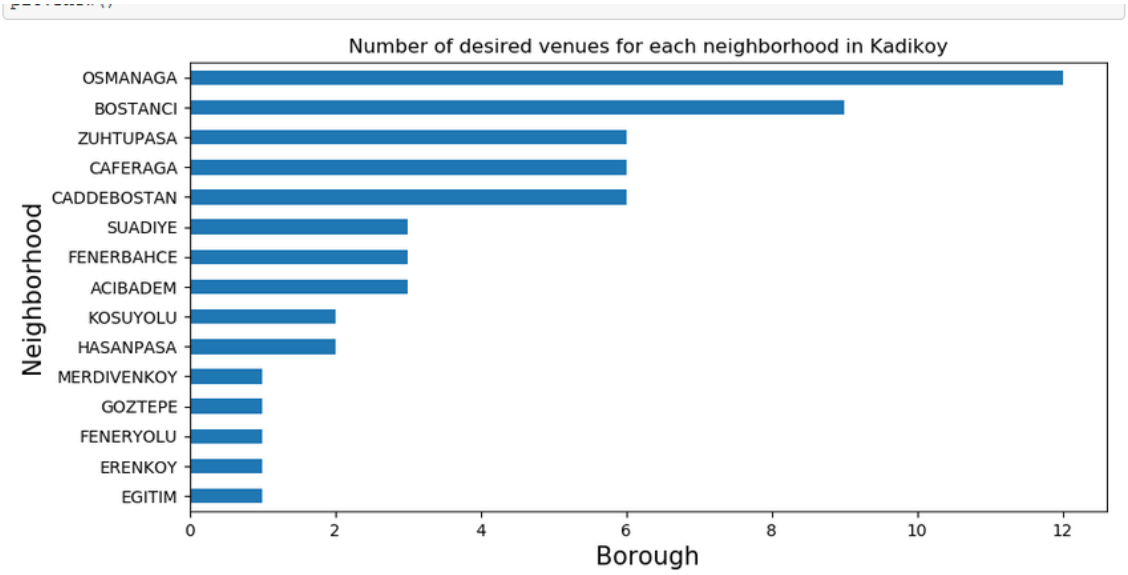
For example, I will get the venue number of each categories in the neighborhoods.

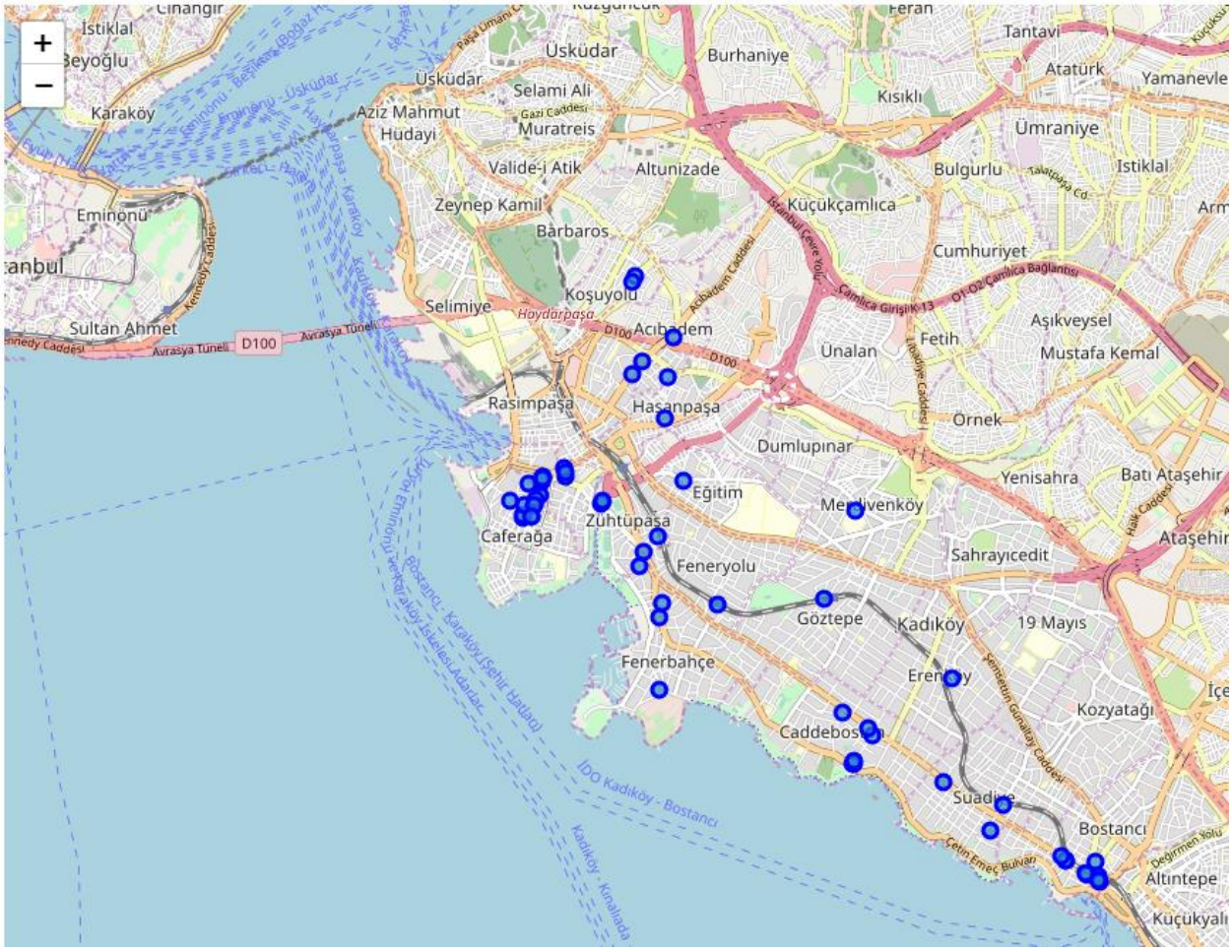
```
In [18]: filtered_venue_df.groupby(['Neighborhood'])['Categories'].value_counts()
```

```
Out[18]: Neighborhood  Categories
         ACIBADEM      Bar              2
                       Pub              1
         BOSTANCI      Bar              3
                       Nightclub        3
                       Pub              2
                       Train Station    1
         CADDEBOSTAN   Bar              3
                       Nightclub        3
         CAFERAGA      Bar              3
                       Pub              3
         EGITIM        Pub              1
         ERENKOY       Train Station    1
         FENERBAHCE    Pub              2
                       Bar              1
         FENERYOLU     Train Station    1
         GOZTEPE       Train Station    1
         HASANPASA     Bus Station      1
                       Nightclub        1
         KOSUYOLU      Bar              1
                       Pub              1
         MERDIVENKOY   Bus Station      1
         OSMANAGA      Bar              6
                       Pub              5
                       Nightclub        1
         SUADIYE       Pub              2
                       Bar              1
         ZUHTUPASA     Nightclub        3
                       Pub              2
                       Bar              1
         Name: Categories, dtype: int64
```

Our client has some expectations. His house should be entertainment venues like bars, clubs and it should be close to public transportation like metro. I filtered the data and visualized it.

I will visualize the desired venues on a horizontal bar plot and maps.





Also, I used one-hot encoding to detect the frequencies of the venue categories.

```
: venue_c
```

:

	Categories_Accessories Store	Categories_Advertising Agency	Categories_African Restaurant	Categories_Arcade
Neighborhood				
19MAYIS	0	0	0	0
ACIBADEM	0	0	0	0
BOSTANCI	0	0	0	0
CADDEBOSTAN	0	0	0	0
CAFERAGA	0	0	0	0
DUMLUPINAR	0	0	1	1
EGITIM	0	0	0	0
ERENKOY	0	0	0	1
FENERBAHCE	1	1	0	1
FENERYOLU	0	0	0	1

This data will be used for clustering. For example, you can see the venue categories which has high frequencies in each neighborhood.

```
----19MAYIS----
                          venue  freq
0              Categories_Bakery  0.10
1                Categories_Café  0.09
2                Categories_Park  0.08
3                 Categories_Gym  0.04
4  Categories_Turkish Restaurant  0.04


----ACIBADEM----
                        venue  freq
0              Categories_Café  0.17
1    Categories_Breakfast Spot  0.05
2      Categories_Dessert Shop  0.05
3  Categories_Kebab Restaurant  0.05
4      Categories_Burger Joint  0.05
```

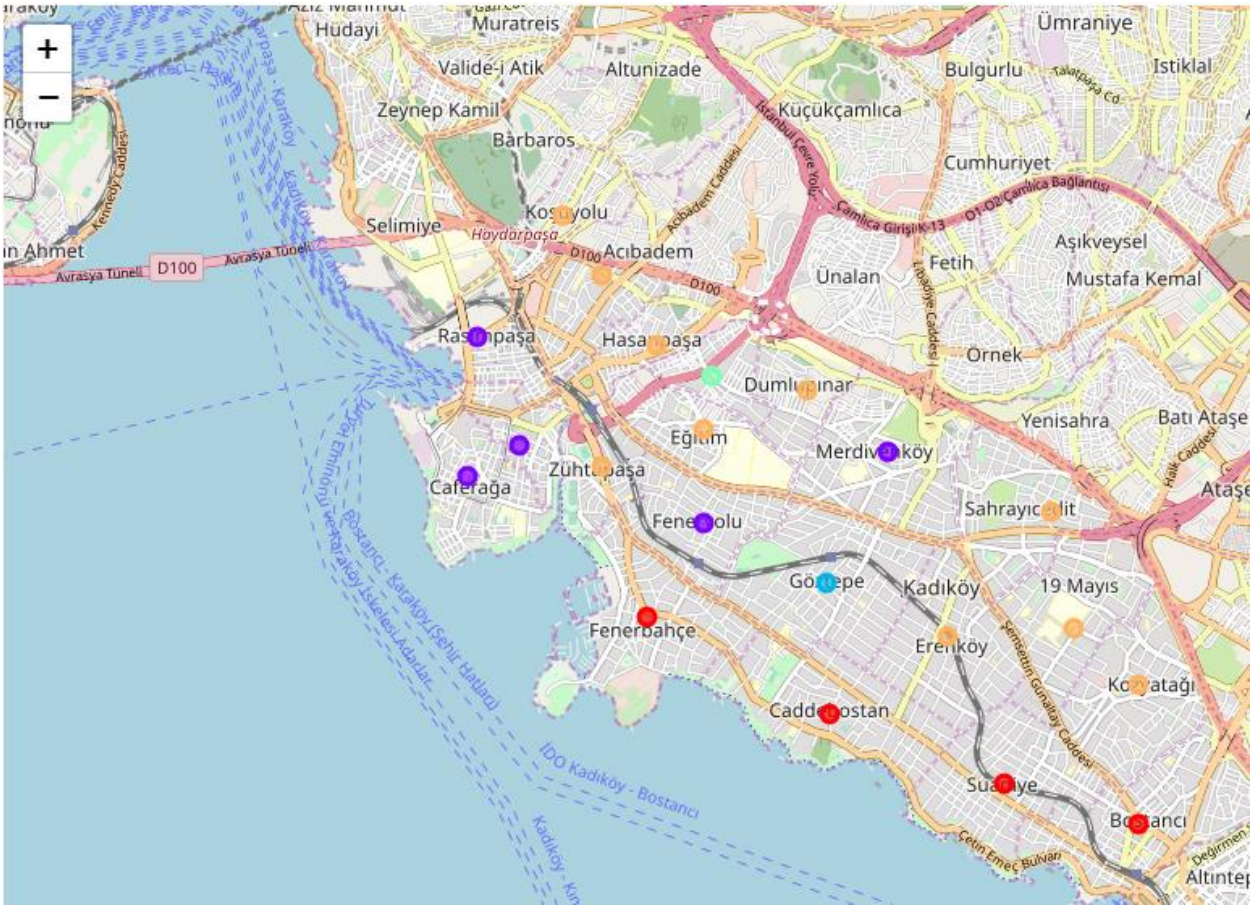
Also, I detected 10 most common venue categories for each neighborhood. It will be used for clustering.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most C Venue
0	19MAYIS	Categories_Bakery	Categories_Café	Categories_Park	Categories_Furniture / Home Store	Categories
1	ACIBADEM	Categories_Café	Categories_Turkish Restaurant	Categories_Breakfast Spot	Categories_Burger Joint	Categories Restaurant
2	BOSTANCI	Categories_Café	Categories_Restaurant	Categories_Coffee Shop	Categories_Nightclub	Categories Shop
3	CADDEBOSTAN	Categories_Coffee Shop	Categories_Café	Categories_Restaurant	Categories_Seafood Restaurant	Categories Cream Shc
4	CAFERAGA	Categories_Café	Categories_Coffee Shop	Categories_Dessert Shop	Categories_Pub	Categories
5	DUMLUPINAR	Categories_Café	Categories_Fast Food Restaurant	Categories_Park	Categories_Gym / Fitness Center	Categories
6	EGITIM	Categories_Café	Categories_Grocery Store	Categories_Bakery	Categories_Restaurant	Categories Shop
7	ERENKOY	Categories_Bakery	Categories_Dessert Shop	Categories_Turkish Restaurant	Categories_Coffee Shop	Categories
8	FENERBAHCE	Categories_Restaurant	Categories_Café	Categories_Seafood Restaurant	Categories_Dessert Shop	Categories Restaurant
9	FENERYOLU	Categories_Café	Categories_Coffee Shop	Categories_Restaurant	Categories_Fast Food Restaurant	Categories Restaurant
10	FIKIRTEPE	Categories_Restaurant	Categories_Automotive Shop	Categories_Comedy Club	Categories_Fast Food Restaurant	Categories

Then, I used K-means clustering to cluster the neighborhoods. Each neighborhood has a cluster label now.

Neighborhood_y	Latitude	Longitude	Cluster Labels
19 MAYIS ...	40.973510	29.088960	4
ACIBADEM ...	41.001780	29.038740	4
BOSTANCI ...	40.957850	29.095760	0

To show the clustered neighborhoods. I used Folium library again. You can see the clustered neighborhoods on the map.



It can be clearly seen that there are 5 different clusters. I also have been in Kadikoy many times and this map makes so much sense. I detected that Osmanaga is the best option for our client, but also Caferaga, Rasimpasa, Feneryolu and Merdivenkoy can be alternative solutions for him. Because they are also in the same cluster.

DISCUSSION AND RESULTS

Now, let's discuss about the issue and results. I had to analyze the neighborhoods in Kadikoy and detect what kind of venues these neighborhoods have. After that i should group them based on client needs. Different kind of people wants to be close to different type of places. I need to detect their needs and find the perfect neighborhoods for them to satisfy their needs.

First, I detected latitude and longitude information of the neighborhoods. After that I used Foursquare API to pull the venue to the neighborhoods closer than 500 meters. Then, I used some filters to detect the best neighborhood for my client. My client already told us that he likes to be close to bars, pubs, restaurant and the entertainment venues. So, we should suggest the neighborhoods that can satisfy his needs. Then I used one-hot encoding to detect the frequency of the venues. I detected 10 most popular venue categories in the neighborhoods. I used K-means clustering to cluster the neighborhoods and used Folium library to show them on a map. I detected 5 clusters.

My analysis shows that Osmanaga is the best neighborhood in Kadikoy for our client. There are many entertainment venues around. Hans can find many places like bars, pubs, restaurants etc. But also, Rasimpasa, Caferaga, Feneryolu and Merdivenkoy can be good for him. These neighborhoods are also in the same cluster.

CONCLUSION

Purpose of the project was to suggest the best neighborhood for our client. In order to do it, I clustered the neighborhoods by using Foursquare data. This data helped me to cluster the neighborhoods.

I detected that Osmanaga is the best neighborhood for the client. But also, we can suggest him Rasimpasa and Caferaga neighborhoods. They are in the same clusters.

As a final decision, we will offer him to move to Osmanaga and give Rasimaga and Caferaga neighborhoods as alternative suggestions.