Coursera Capstone Final Project : Applied Data Science

Emre Zafer Güney Istanbul, Turkey – 20.06.2020

emrezaferguney@gmail.com

Üsküdar – İSTANBUL / TURKEY

Overview

1) **Introduction & Business Problem**

2) **Data**
   - Neighborhoods
   - Geocoding
   - Venue Data

3) **Methodology**
   - Accuracy of the Geocoding API & Folium
   - One hot encoding & Top 5 most common venues
   - Optimal number of clusters & K-means clustering

4) **Results**

5) **Conclusion**

1) Introduction & Business Problem

Üsküdar is the one of the oldest districts of İstanbul with high density of population. According to data from Turkish Statistical Institute report of 2019, Üsküdar is the 8th biggest [1] among 38 other districts of İstanbul in terms of population with more than half million people.

The ratio of younger people living in Üsküdar is increasing year by year corelated with the jobs opportunities increase rate in Anatolian Side of İstanbul.

In the last decades, people are more aware of consumed goods in terms of being healthy. People are spending more time on to be on shape. They look for places close to Gym and healthy food restaurants or café. Thus, it might be a good idea open a “Stay Healthy Café” in Üsküdar.

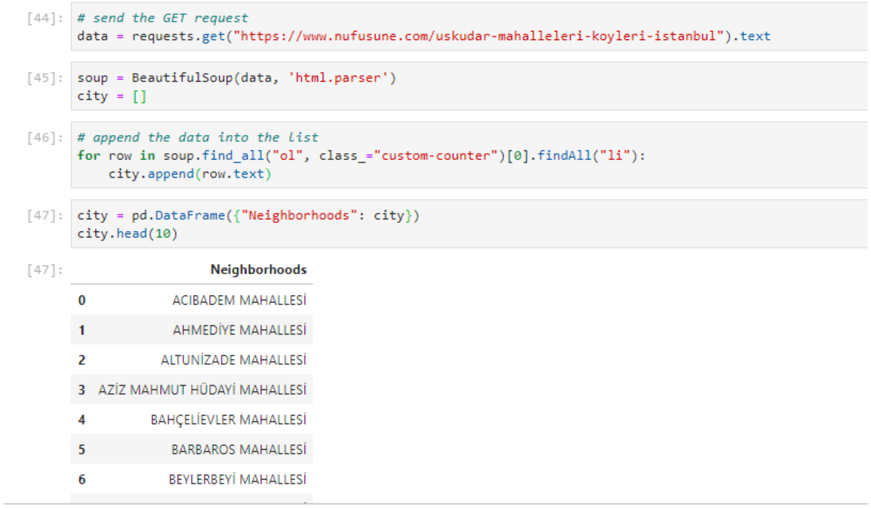
The mission of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the best neighborhood in Üsküdar (my hometown) to open a Stay Healthy Café, which offers beverages and foods mostly aiming for people who go to gym.

Therefore, I will analyze each neighborhood of Üsküdar in terms of venue categories (data from Foursquare) and using K-means clustering methodology, will aim to find a best cluster which are close to Gym, sports areas and café.

2) Data

Neighborhoods

Using requests.get function and with the help of website scratching BeautifulSoup function we get the 33 different neigborhods of Üsküdar.



```
[44]: # send the GET request
      data = requests.get("https://www.nufusune.com/uskudar-mahalleleri-koyleri-istanbul").text

[45]: soup = BeautifulSoup(data, 'html.parser')
      city = []

[46]: # append the data into the list
      for row in soup.find_all("ol", class_="custom-counter")[0].findAll("li"):
          city.append(row.text)

[47]: city = pd.DataFrame({"Neighborhoods": city})
      city.head(10)
```

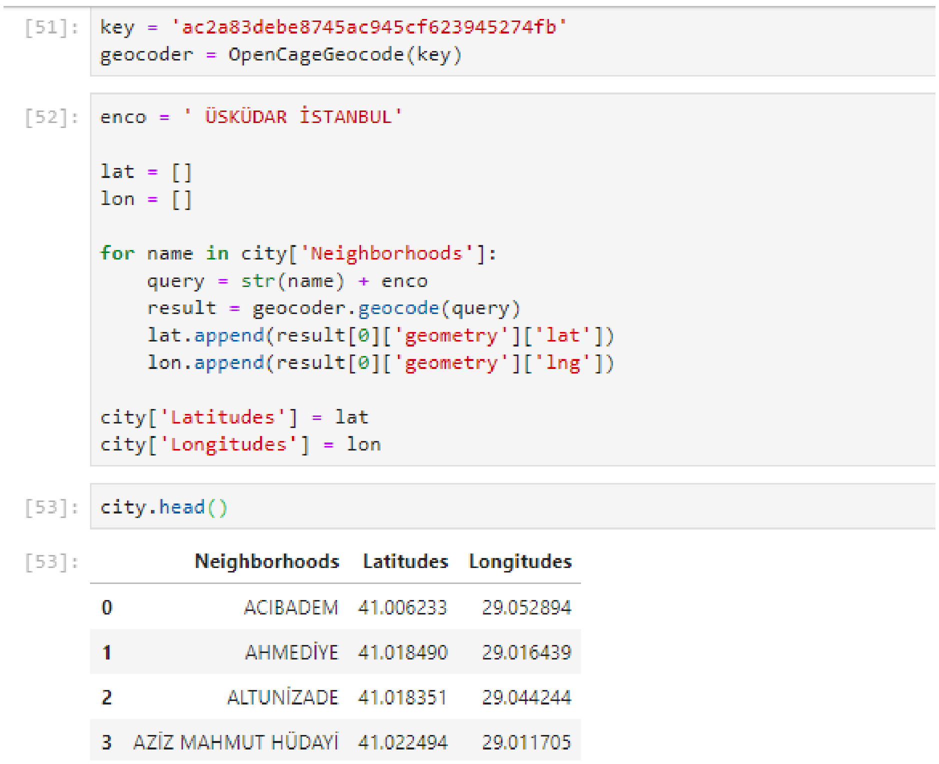
[47]:

	Neighborhoods
0	ACIBADEM MAHALLESİ
1	AHMEDİYE MAHALLESİ
2	ALTUNİZADE MAHALLESİ
3	AZİZ MAHMUT HÜDAYİ MAHALLESİ
4	BAHÇELİEVLER MAHALLESİ
5	BARBAROS MAHALLESİ
6	BEYLERBEYİ MAHALLESİ

[1] (AA, https://www.aa.com.tr/tr/turkiye/-istanbulun-en-kalabalik-10-ilcesinin-nufusu-22-avrupa-ulkesinden-daha-fazla/1724728)

Geocoding

Using Geocoding, the latitude and longitude of the neighborhoods are retrieved using OpenCage Geocoding API. The geometric location values are then stored into the initial data frame.



```
[51]: key = 'ac2a83debe8745ac945cf623945274fb'
      geocoder = OpenCageGeocode(key)
```

```
[52]: enco = ' ÜSKÜDAR İSTANBUL'

      lat = []
      lon = []

      for name in city['Neighborhoods']:
          query = str(name) + enco
          result = geocoder.geocode(query)
          lat.append(result[0]['geometry']['lat'])
          lon.append(result[0]['geometry']['lng'])

      city['Latitudes'] = lat
      city['Longitudes'] = lon
```

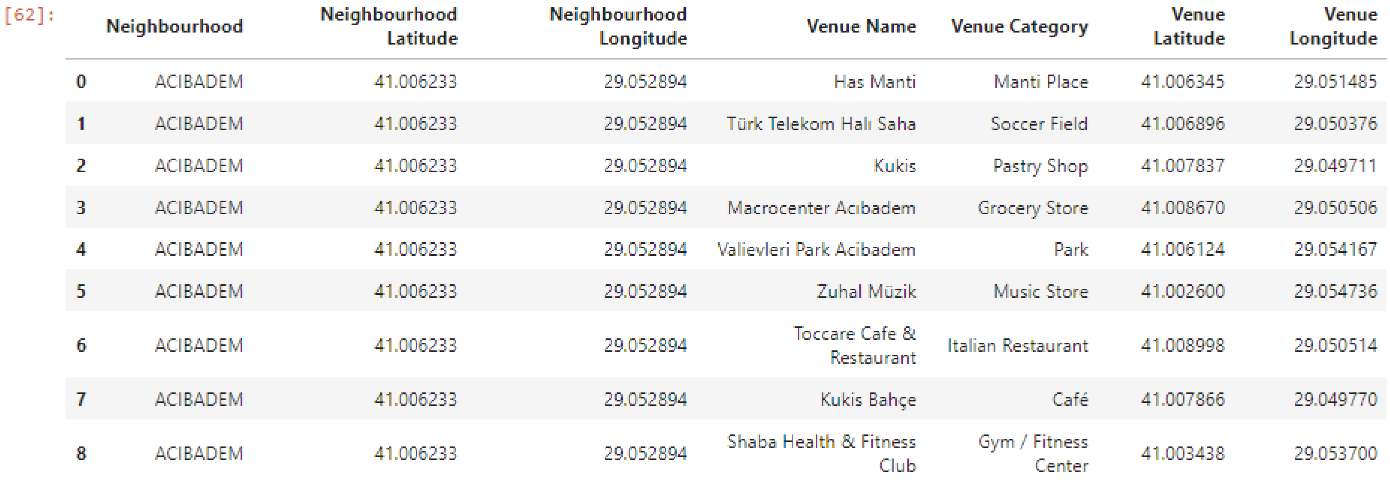
```
[53]: city.head()
```

[53]:

	Neighborhoods	Latitudes	Longitudes
0	ACIBADEM	41.006233	29.052894
1	AHMEDİYE	41.018490	29.016439
2	ALTUNİZADE	41.018351	29.044244
3	AZİZ MAHMUT HÜDAYİ	41.022494	29.011705

Venue Data

The data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another Data Frame to contain all the venue details along with the respective neighborhoods.



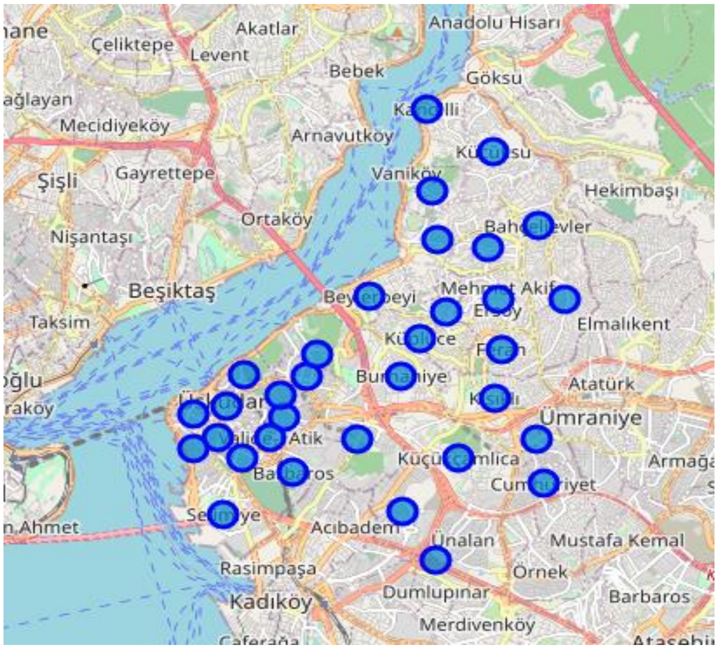
[62]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue Name	Venue Category	Venue Latitude	Venue Longitude
0	ACIBADEM	41.006233	29.052894	Has Manti	Manti Place	41.006345	29.051485
1	ACIBADEM	41.006233	29.052894	Türk Telekom Halı Saha	Soccer Field	41.006896	29.050376
2	ACIBADEM	41.006233	29.052894	Kukis	Pastry Shop	41.007837	29.049711
3	ACIBADEM	41.006233	29.052894	Macrocenter Acıbadem	Grocery Store	41.008670	29.050506
4	ACIBADEM	41.006233	29.052894	Valievleri Park Acibadem	Park	41.006124	29.054167
5	ACIBADEM	41.006233	29.052894	Zuhal Müzik	Music Store	41.002600	29.054736
6	ACIBADEM	41.006233	29.052894	Toccare Cafe & Restaurant	Italian Restaurant	41.008998	29.050514
7	ACIBADEM	41.006233	29.052894	Kukis Bahçe	Café	41.007866	29.049770
8	ACIBADEM	41.006233	29.052894	Shaba Health & Fitness Club	Gym / Fitness Center	41.003438	29.053700

3) Methodology

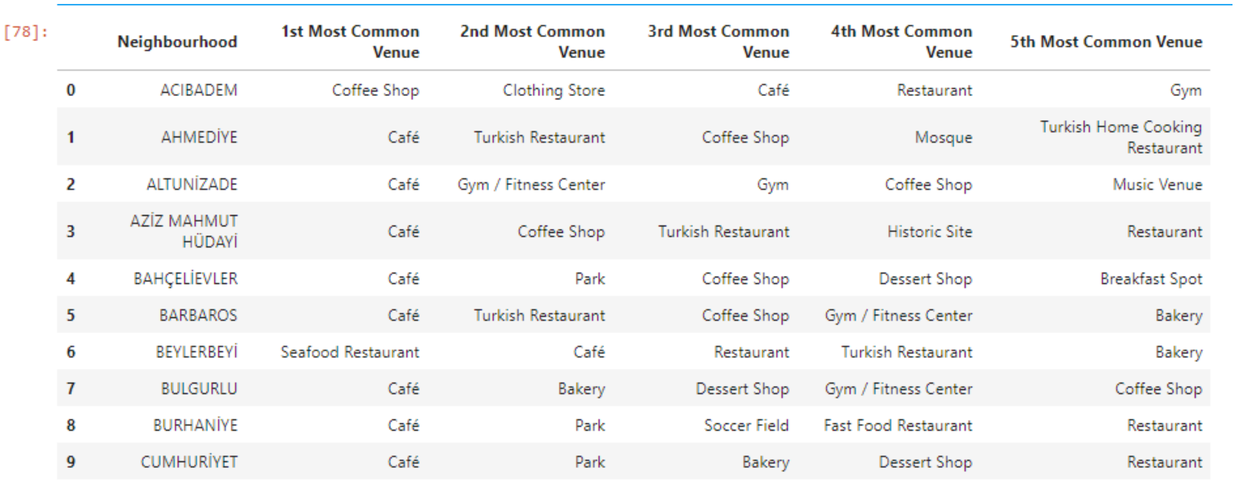
Accuracy of the Geocoding API & Folium

Having Geo data of 33 different neighborhood, we can show it with the help of folium.



One hot encoding & Top 5 most common venues

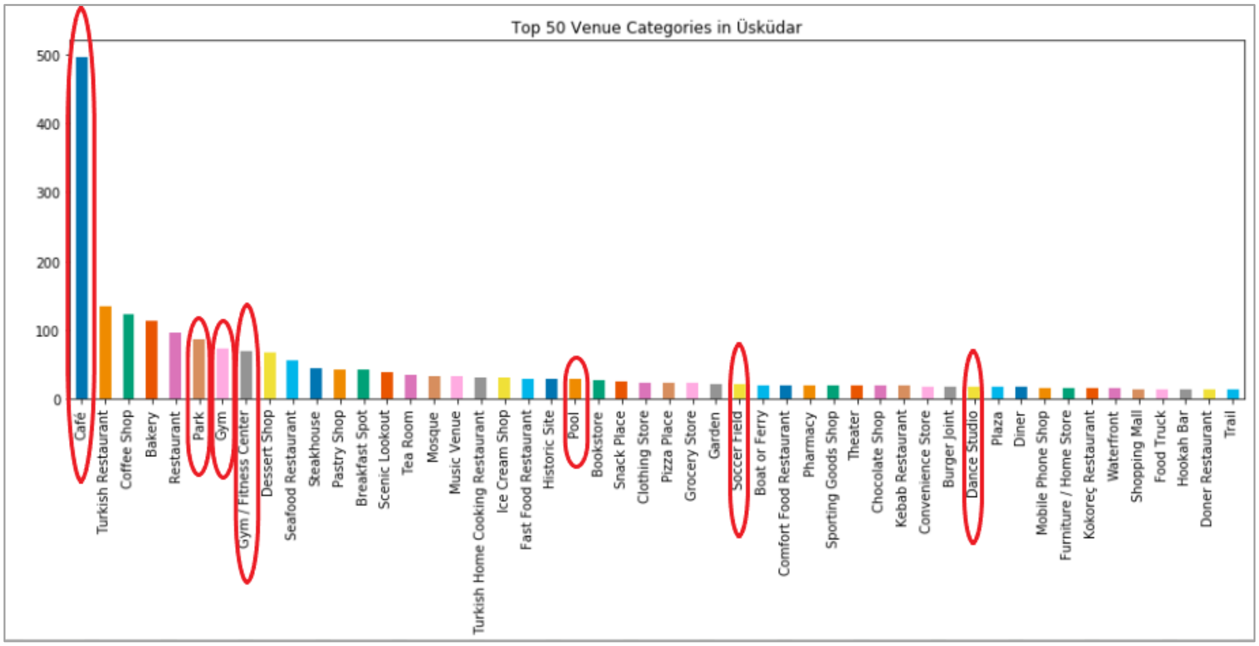
Using Foursquare data, we get at most 100 venues for each neighborhood and find the 5 most common venues in the neighborhoods.



[78]:

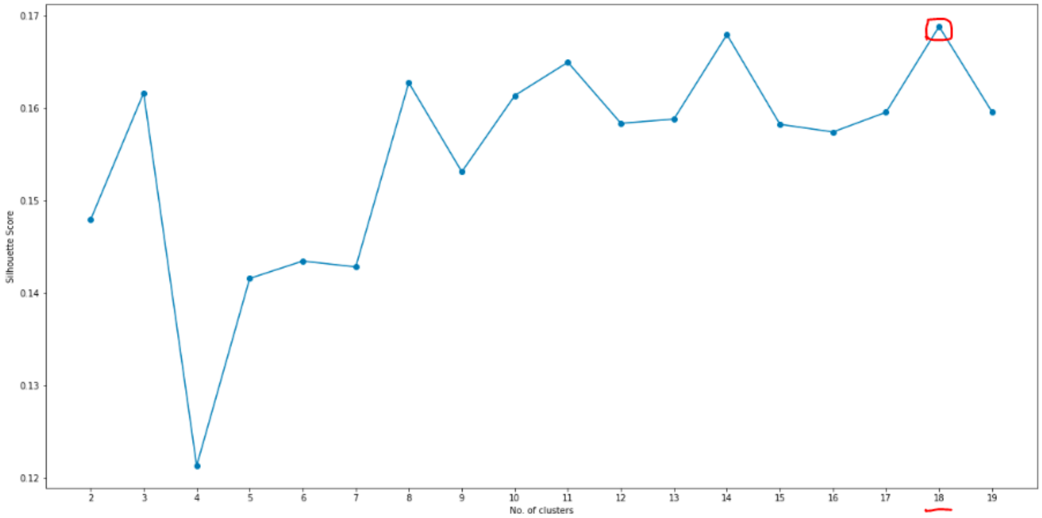
	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	ACIBADEM	Coffee Shop	Clothing Store	Café	Restaurant	Gym
1	AHMEDİYE	Café	Turkish Restaurant	Coffee Shop	Mosque	Turkish Home Cooking Restaurant
2	ALTUNİZADE	Café	Gym / Fitness Center	Gym	Coffee Shop	Music Venue
3	AZİZ MAHMUT HÜDAYİ	Café	Coffee Shop	Turkish Restaurant	Historic Site	Restaurant
4	BAHÇELİEVLER	Café	Park	Coffee Shop	Dessert Shop	Breakfast Spot
5	BARBAROS	Café	Turkish Restaurant	Coffee Shop	Gym / Fitness Center	Bakery
6	BEYLERBEYİ	Seafood Restaurant	Café	Restaurant	Turkish Restaurant	Bakery
7	BULGURLU	Café	Bakery	Dessert Shop	Gym / Fitness Center	Coffee Shop
8	BURHANİYE	Café	Park	Soccer Field	Fast Food Restaurant	Restaurant
9	CUMHURİYET	Café	Park	Bakery	Dessert Shop	Restaurant

The top 50 venue categories are examined in terms of # of venues and our focus categories which's customers are assumed to be our customer as well for Stay Heathy Café.

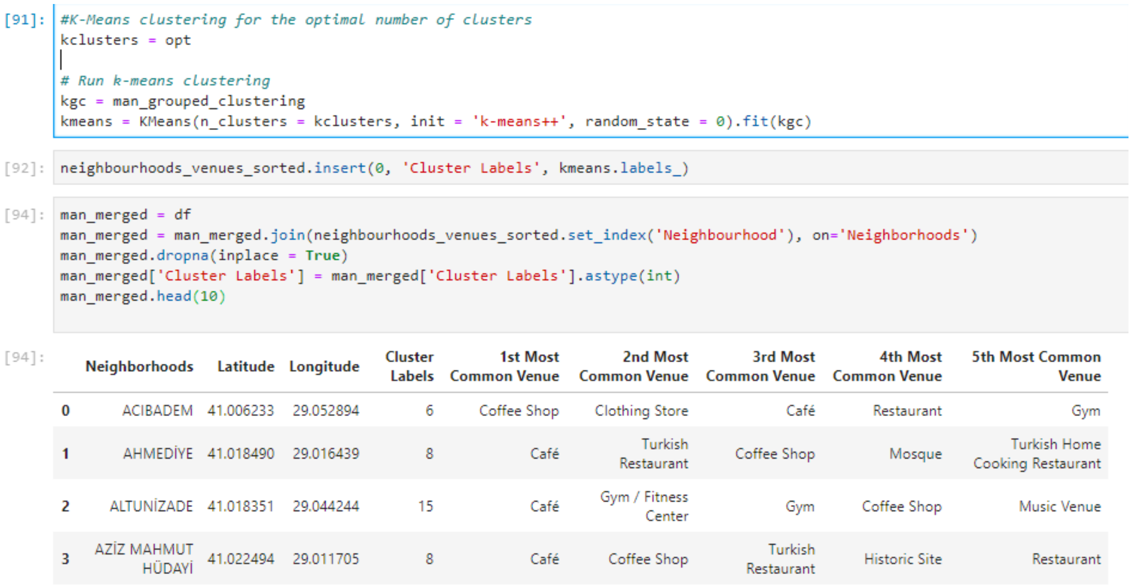


Optimal number of clusters & K-means clustering

Silhouette Score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Based on the Silhouette Score of various clusters below 18, the optimal cluster size is determined



The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base the analysis on. K-means was chosen as the variables (Venue Categories) are huge, and in such situations K-means will be computationally faster than other clustering algorithms.



```
[91]: #K-Means clustering for the optimal number of clusters
      kclusters = opt
      |
      # Run k-means clustering
      kgc = man_grouped_clustering
      kmeans = KMeans(n_clusters = kclusters, init = 'k-means++', random_state = 0).fit(kgc)
```

```
[92]: neighbourhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
```

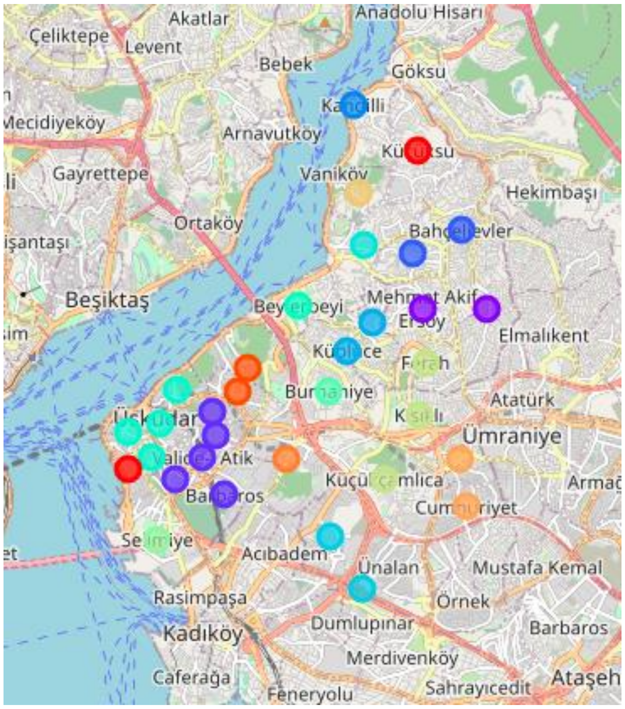
```
[94]: man_merged = df
      man_merged = man_merged.join(neighbourhoods_venues_sorted.set_index('Neighbourhood'), on='Neighborhoods')
      man_merged.dropna(inplace = True)
      man_merged['Cluster Labels'] = man_merged['Cluster Labels'].astype(int)
      man_merged.head(10)
```

[94]:

	Neighborhoods	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	ACIBADEM	41.006233	29.052894	6	Coffee Shop	Clothing Store	Café	Restaurant	Gym
1	AHMEDİYE	41.018490	29.016439	8	Café	Turkish Restaurant	Coffee Shop	Mosque	Turkish Home Cooking Restaurant
2	ALTUNİZADE	41.018351	29.044244	15	Café	Gym / Fitness Center	Gym	Coffee Shop	Music Venue
3	AZİZ MAHMUT HÜDAYİ	41.022494	29.011705	8	Café	Coffee Shop	Turkish Restaurant	Historic Site	Restaurant

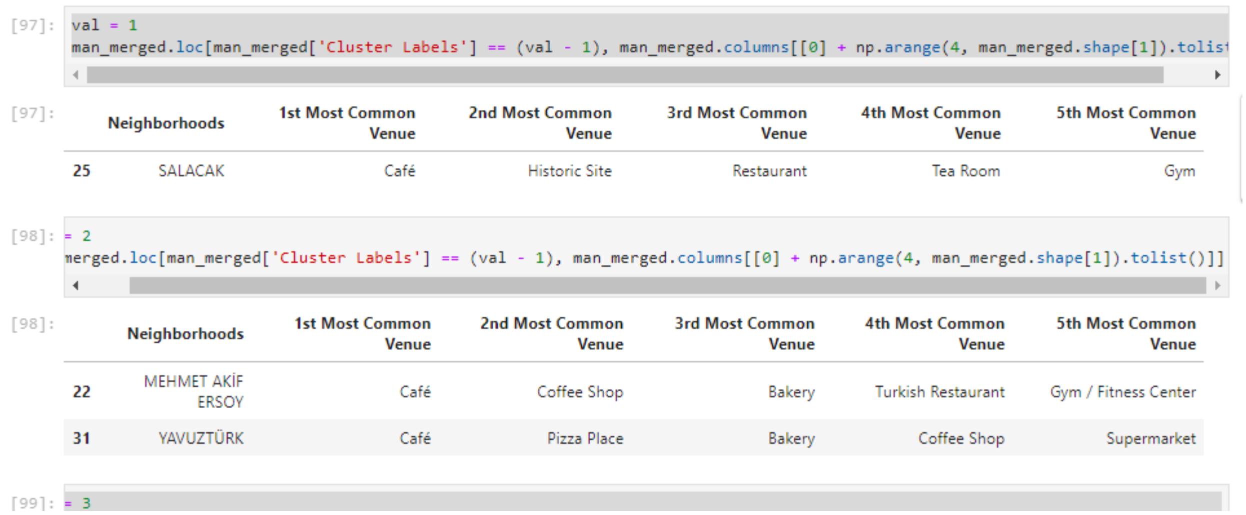
4) Results

The neighborhoods are divided into 18 clusters using the optimal approach. The clustered neighborhoods are visualized using different colors so as to make them distinguishable



Each 18 clusters are examined according to common venues of Gym, Café, Park, Gym, Fitness Center, Pool, Soccer Field, Dance Studio.

Example;



```
[97]: val = 1
      man_merged.loc[man_merged['Cluster Labels'] == (val - 1), man_merged.columns[[0] + np.arange(4, man_merged.shape[1]).tolist
```

[97]:

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
25	SALACAK	Café	Historic Site	Restaurant	Tea Room	Gym

```
[98]: = 2
      erged.loc[man_merged['Cluster Labels'] == (val - 1), man_merged.columns[[0] + np.arange(4, man_merged.shape[1]).tolist()]]
```

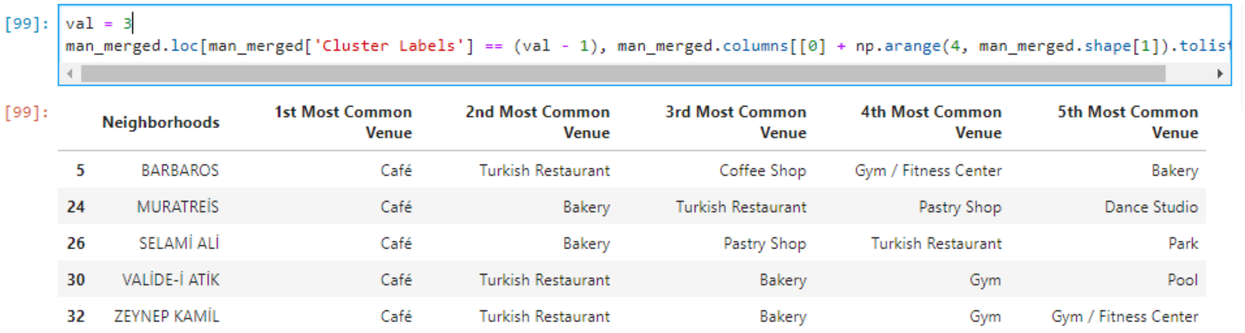
[98]:

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
22	MEHMET AKİF ERSOY	Café	Coffee Shop	Bakery	Turkish Restaurant	Gym / Fitness Center
31	YAVUZTÜRK	Café	Pizza Place	Bakery	Coffee Shop	Supermarket

```
[99]: = 3
```

5) Conclusion

As a result, the cluster group of 3 with the neighborhood of Barbaros, Valide-I Atik and Zeynep Kamil are good options to open a “Stay Healthy Café”, hence their most common venues are more related to our business approach.



```
[99]: val = 3
      man_merged.loc[man_merged['Cluster Labels'] == (val - 1), man_merged.columns[[0] + np.arange(4, man_merged.shape[1]).tolis
```

[99]:

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
5	BARBAROS	Café	Turkish Restaurant	Coffee Shop	Gym / Fitness Center	Bakery
24	MURATREİS	Café	Bakery	Turkish Restaurant	Pastry Shop	Dance Studio
26	SELAMİ ALİ	Café	Bakery	Pastry Shop	Turkish Restaurant	Park
30	VALİDE-İ ATİK	Café	Turkish Restaurant	Bakery	Gym	Pool
32	ZEYNEP KAMİL	Café	Turkish Restaurant	Bakery	Gym	Gym / Fitness Center

P.S: Find the .ipynb version of the report in the link

https://github.com/emrezaferguney/github-zafer/blob/master/CAPSTONE%20PROJECT%20BATTLE%20OF%20THE%20NEIGHBORHOODS%20-%20Final%20Project.ipynb