NEIGHBORHOOD CLUSTERING FOR OPTIMAL BUSINESS LOCATION

Erdem Ünlü

1 – Introduction

1.1 – Background

Istanbul is the largest city in Europe by population and Turkey's economic, cultural and historic center. Along with its population and importance for the country, its geographic location, which connects Europe and Asia, makes it quite suitable for this project.

1.2 – Problem

One of the hardest choices for businesses (along with many others) is where to open an office / restaurant / store. This project aims to give businesses some insights by clustering boroughs of Istanbul by using data that includes age, income, and venues nearby.

1.3 – Interest

Any business that needs a settlement could be interested in this project.

2 – Data

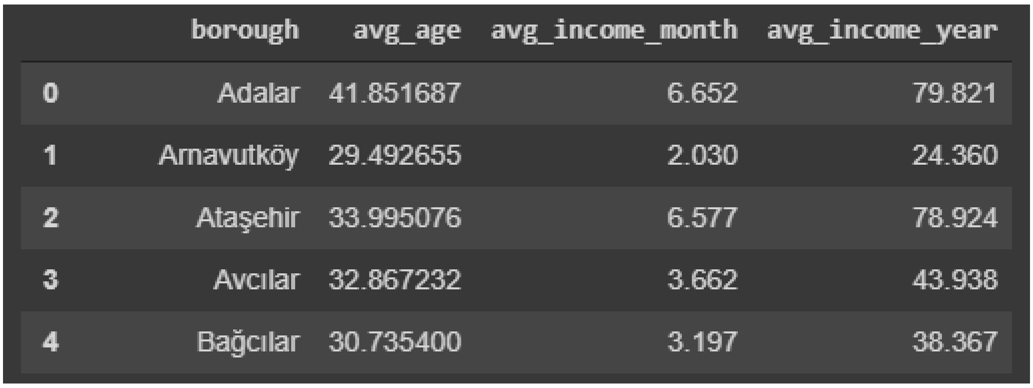
2.1 - Data Sources

- Grouped Ages for each borough could be accessed from TUIK's (Turkish Statistical Institute) website.
- JSON file which includes boroughs of Istanbul's coordinates could be found here.
- Alas, I couldn't find a ready to use data for average incomes for each borough, but I was able to find a picture from a research that could be found here.
- Venues are scraped using Foursquare's API.

2.2 – Data Preparation

- Downloaded file that store ages, needed some changes. NaN columns are dropped. The first column's data changed to the desired format. (For example: "Istanbul(Adalar)-1103" was changed to "Adalar")
- In the downloaded file, ages were grouped by 4 years (0-4, 5-9, 10-14, etc.). These groupings are replaced by average age using a weighted average as the method.
- The photo that contained average incomes for each borough is converted to a CSV file using an online converter. Little adjustments are made by hand to the data retrieved.
- These 2 datasets are merged into a data frame.

- JSON file needed some adjustments (like changing the name of a borough). These adjustments are done by hand.



	borough	avg_age	avg_income_month	avg_income_year
0	Adalar	41.851687	6.652	79.821
1	Arnavutköy	29.492655	2.030	24.360
2	Ataşehir	33.995076	6.577	78.924
3	Avcılar	32.867232	3.662	43.938
4	Bağcılar	30.735400	3.197	38.367

Figure 1. Final view of the data frame

3 – Methodology

3.1 – Mapping

Using Folium's Choropleth method, I've created a Choropleth map that uses average age to color the map which is centered around Istanbul.

3.2 – Foursquare API

To get the venues around each borough, I used Foursquare's API with a limit of 100 venues and a radius of 1 KM. While most of the boroughs exceeded the proposed limit, some of them didn't even reach it.

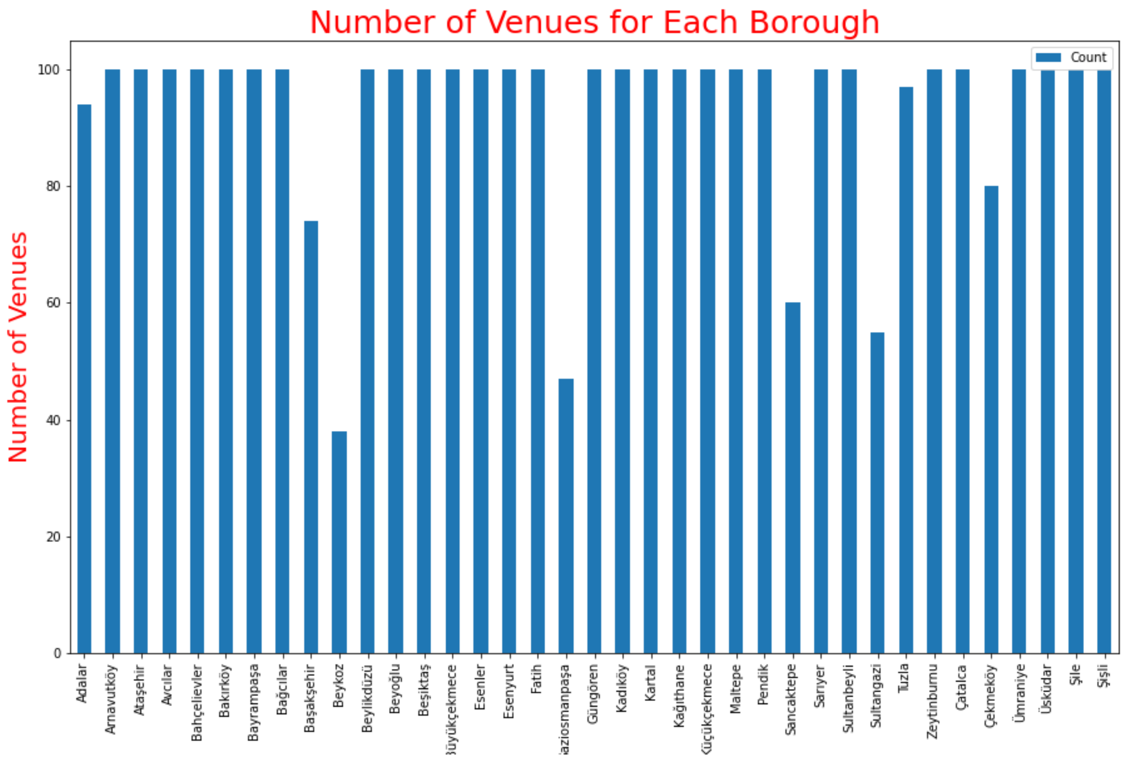


Figure 2. Number of venues returned for each borough

3.3 – Scaling

I've used the min-max scaling approach to scale the columns "avg_age", "avg_income_month" and "avg_income_year" between 0 and 1.

3.4 – Modeling

After preparing everything that is needed, I've used sci-kit learn's K-means algorithm to cluster boroughs. To choose the number of clusters K-means algorithm return, I've used the elbow method and decided on K = 5.

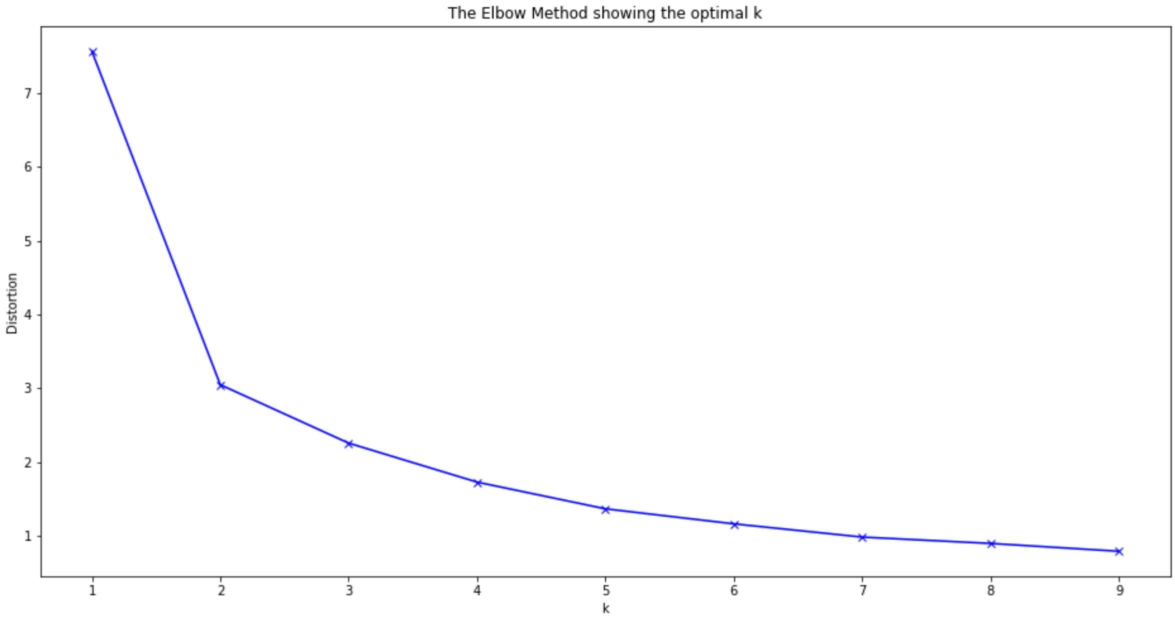


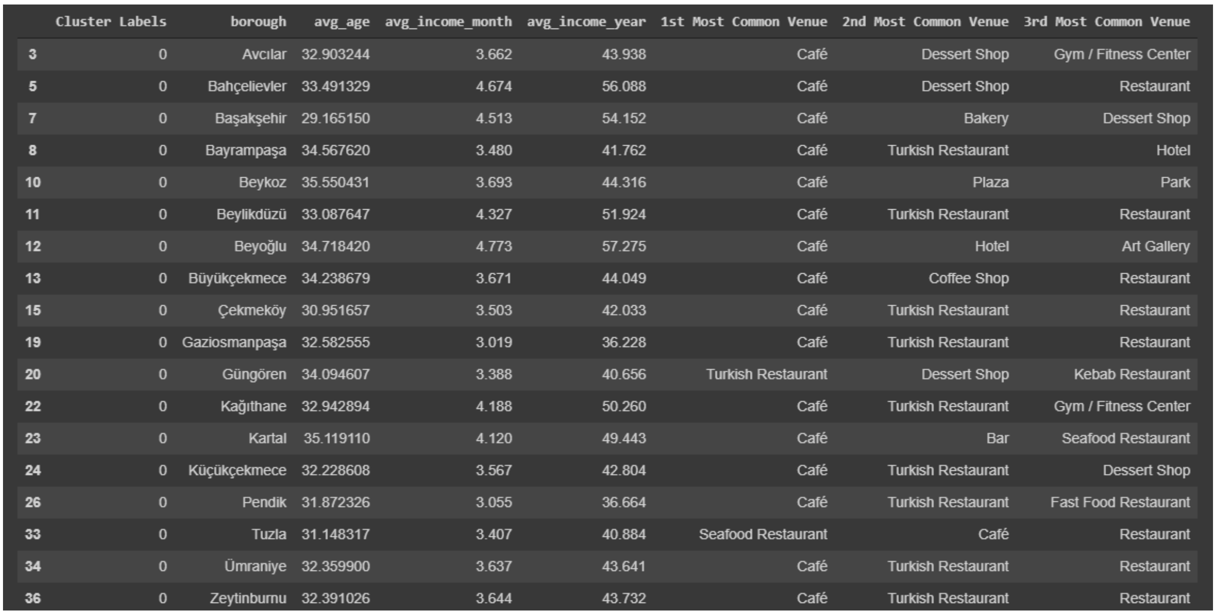
Figure 3 Elbow method for finding optimal number of clusters

3.5 – Mapping V2

The last thing to do is, adding markers with their reflecting colours to the Choropleth map created before.

4 – Results

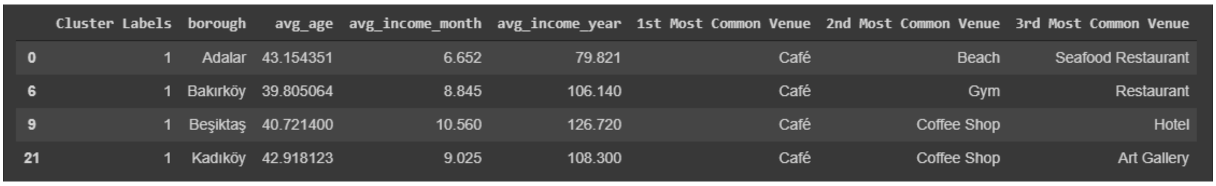
K-means algorithm returned 5 clusters. Every borough belonging to the same clusters have some common characteristics. Boroughs in Cluster 0, doesn't have an average age that differentiates them from other clusters. The same applies to the average income by month and average income by year. One common characteristic that includes almost all of them is that the most common venue Café. Thus, this cluster will be named "Moderate age & income, with mostly Cafés."



	Cluster Labels	borough	avg_age	avg_income_month	avg_income_year	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
3	0	Avcılar	32.903244	3.662	43.938	Café	Dessert Shop	Gym / Fitness Center
5	0	Bahçelievler	33.491329	4.674	56.088	Café	Dessert Shop	Restaurant
7	0	Başakşehir	29.165150	4.513	54.152	Café	Bakery	Dessert Shop
8	0	Bayrampaşa	34.567620	3.480	41.762	Café	Turkish Restaurant	Hotel
10	0	Beykoz	35.550431	3.693	44.316	Café	Plaza	Park
11	0	Beylikdüzü	33.087647	4.327	51.924	Café	Turkish Restaurant	Restaurant
12	0	Beyoğlu	34.718420	4.773	57.275	Café	Hotel	Art Gallery
13	0	Büyükçekmece	34.238679	3.671	44.049	Café	Coffee Shop	Restaurant
15	0	Çekmeköy	30.951657	3.503	42.033	Café	Turkish Restaurant	Restaurant
19	0	Gaziosmanpaşa	32.582555	3.019	36.228	Café	Turkish Restaurant	Restaurant
20	0	Güngören	34.094607	3.388	40.656	Turkish Restaurant	Dessert Shop	Kebab Restaurant
22	0	Kağıthane	32.942894	4.188	50.260	Café	Turkish Restaurant	Gym / Fitness Center
23	0	Kartal	35.119110	4.120	49.443	Café	Bar	Seafood Restaurant
24	0	Küçükçekmece	32.228608	3.567	42.804	Café	Turkish Restaurant	Dessert Shop
26	0	Pendik	31.872326	3.055	36.664	Café	Turkish Restaurant	Fast Food Restaurant
33	0	Tuzla	31.148317	3.407	40.884	Seafood Restaurant	Café	Restaurant
34	0	Ümraniye	32.359900	3.637	43.641	Café	Turkish Restaurant	Restaurant
36	0	Zeytinburnu	32.391026	3.644	43.732	Café	Turkish Restaurant	Restaurant

Figure 4 Elements of Cluster 0

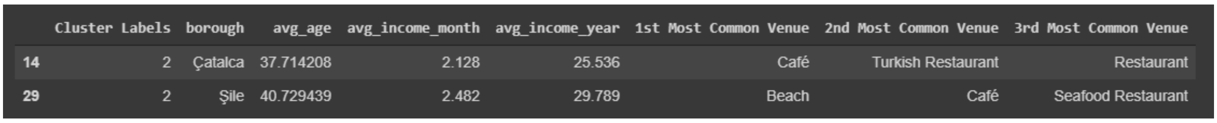
Each borough in Cluster 1 has high age and income averages. And in each one them, the most common venue is Café. Thus, this cluster will be named "High age & income, with most commonly Cafés."



	Cluster Labels	borough	avg_age	avg_income_month	avg_income_year	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	1	Adalar	43.154351	6.652	79.821	Café	Beach	Seafood Restaurant
6	1	Bakırköy	39.805064	8.845	106.140	Café	Gym	Restaurant
9	1	Beşiktaş	40.721400	10.560	126.720	Café	Coffee Shop	Hotel
21	1	Kadıköy	42.918123	9.025	108.300	Café	Coffee Shop	Art Gallery

Figure 5 Elements of Cluster 1

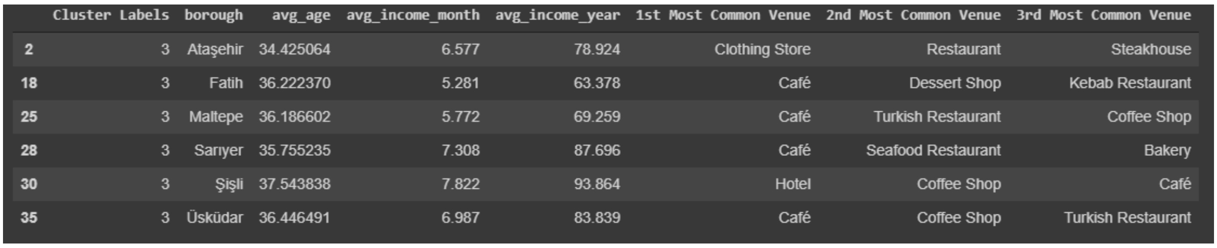
Boroughs in Cluster 2 (there are only two) have high ages with low-income averages. There doesn't seem to be any common venue that differs. Thus, it will be named "High age and Low income"



	Cluster Labels	borough	avg_age	avg_income_month	avg_income_year	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
14	2	Çatalca	37.714208	2.128	25.536	Café	Turkish Restaurant	Restaurant
29	2	Şile	40.729439	2.482	29.789	Beach	Café	Seafood Restaurant

Figure 6 Elements of Cluster 2

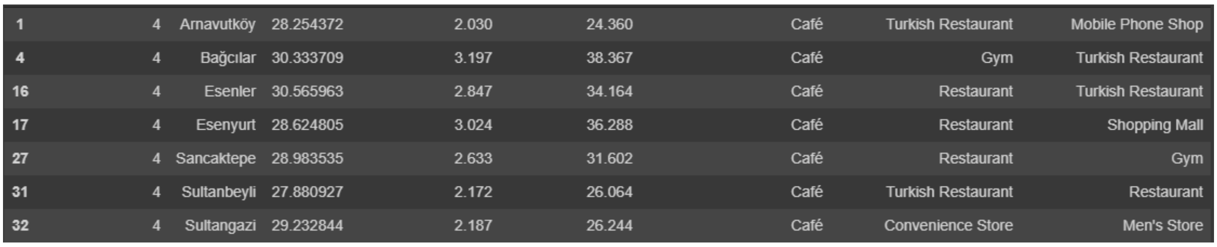
Average ages and incomes in Cluster 3 are not as high as Cluster 1, but still pretty high. Most common venues are mostly again Cafés. Thus, it will be named "Above moderate age and income, with mostly Cafés."



	Cluster Labels	borough	avg_age	avg_income_month	avg_income_year	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
2	3	Ataşehir	34.425064	6.577	78.924	Clothing Store	Restaurant	Steakhouse
18	3	Fatih	36.222370	5.281	63.378	Café	Dessert Shop	Kebab Restaurant
25	3	Maltepe	36.186602	5.772	69.259	Café	Turkish Restaurant	Coffee Shop
28	3	Sarıyer	35.755235	7.308	87.696	Café	Seafood Restaurant	Bakery
30	3	Şişli	37.543838	7.822	93.864	Hotel	Coffee Shop	Café
35	3	Üsküdar	36.446491	6.987	83.839	Café	Coffee Shop	Turkish Restaurant

Figure 7 Element of Cluster 3

The last cluster consists of boroughs with low age & income averages. In each one of the boroughs', the most common venue is Café. Thus, it will be named "Low age & income, with most commonly Cafés."



1	4	Arnavutköy	28.254372	2.030	24.360	Café	Turkish Restaurant	Mobile Phone Shop
4	4	Bağcılar	30.333709	3.197	38.367	Café	Gym	Turkish Restaurant
16	4	Esenler	30.565963	2.847	34.164	Café	Restaurant	Turkish Restaurant
17	4	Esenyurt	28.624805	3.024	36.288	Café	Restaurant	Shopping Mall
27	4	Sancaktepe	28.983535	2.633	31.602	Café	Restaurant	Gym
31	4	Sultanbeyli	27.880927	2.172	26.064	Café	Turkish Restaurant	Restaurant
32	4	Sultangazi	29.232844	2.187	26.244	Café	Convenience Store	Men's Store

Figure 8 Elements of Cluster 4

5 – Discussion

Istanbul has a high population with high density. In these conditions, conducting the research could be pretty hard. Even though there are more than 35 boroughs, almost in all of them, Cafés were the most common venue. The reason for that could be geographic, cultural, or financial. Nonetheless, Café's being the most common venue should not affect businesses' decisions about their location. However, the clusters are mostly not affected by the venues and could be the base for deciding a settlement's location.

6 - Conclusion

In this project, I used the K-Means algorithm to cluster boroughs in Istanbul with features including age, income, and most common venues. These clusters are also marked in a Choropleth map created, for better and easier assessment.

This project is just research. Businesses who want to have insights about where to open an office/restaurant/store, could look into this research and make their own decisions.