CAPSTONE PROJECT Final Report

By Umut Kapucu

1. Introduction

The objective of my project is to help people in exploring similar district where they live and similar tastes of district which they rate. It will help people making decision on selecting similar and/or enjoyable districts in Istanbul, Turkey.

Istanbul, formerly known as Byzantium and Constantinople, is the most populous city in Turkey. With a total population of around 15 million residents, Istanbul is one of the world's largest cities by population. It is a transcontinental city (divided by Bosphorus, Europe and Asia) between the Sea of Marmara and the Black Sea located south-eastern Europe.

2. Problem

The purpose of this project, is to determine similar districts in Istanbul and recommend districts according to users ratings. By using data science methodology along with machine learning algorithms such as clustering and recommender system, the project aims to answer following questions:

1. In Istanbul, how many district clusters and what are their attributes?
2. According to users ratings for specific districts, which districts would be recommended to that user?

3. Target Audience

- Users who wants to learn about Istanbul districts and their similarities

- Users who initially rated the district for his/her own taste and wants to learn where could be interesting for him/her.

4. Data

- Istanbul is the most populous city in Turkey and Europe. It has a great metropolitan area and has a population of 15 million.

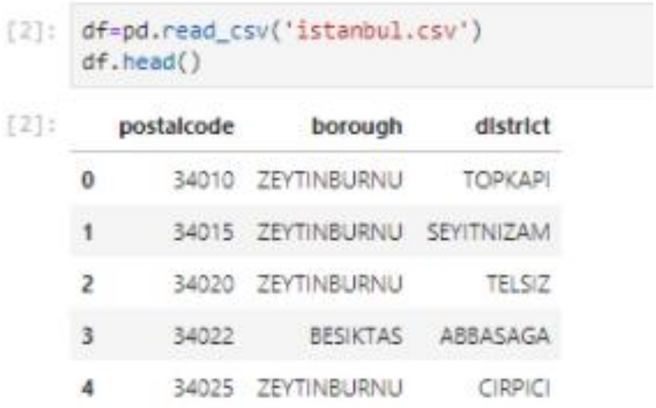
- The smallest available units (districts) of Istanbul are used for this study. The districts are represented as postal codes obtained from the government postal service. https://postakodu.ptt.gov.tr/dosyalar/pk_list.zip

- From link, all postal codes could be retrieved for all cities, boroughs and districts in Turkey. Istanbul is filtered from data and saved as CSV file for that study.

- Foursquare API, as data source, venues in each district are listed with the specified restrictions (with radius 500 and limit 100). The venues are categorized and determine the most common ones for analysis.

5. Methodology

- The districts are represented as postal codes obtained from the government postal service. Data is retrieved from csv file.



```
[2]: df=pd.read_csv('istanbul.csv')
     df.head()
```

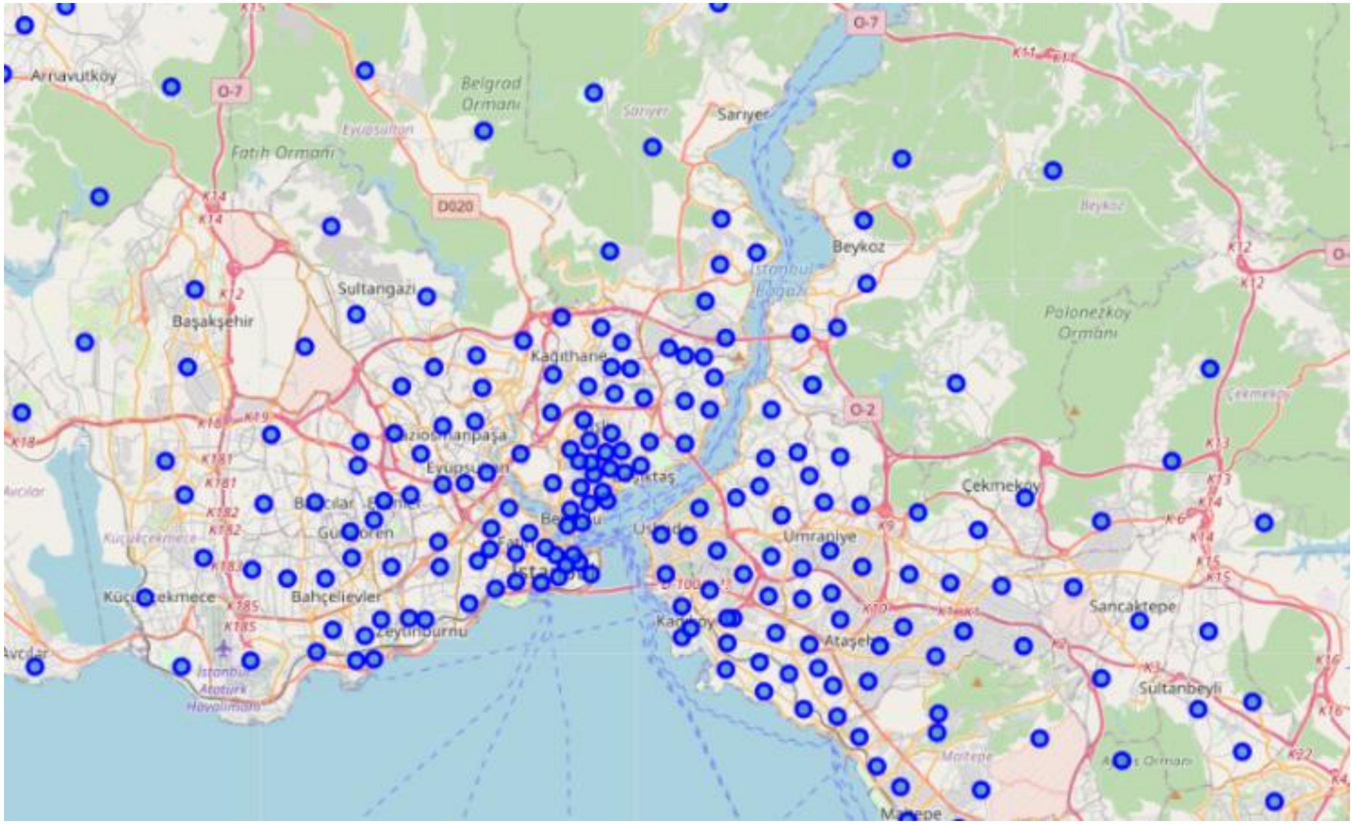
[2]:

	postalcode	borough	district
0	34010	ZEYTINBURNU	TOPKAPI
1	34015	ZEYTINBURNU	SEYITNIZAM
2	34020	ZEYTINBURNU	TELSIZ
3	34022	BESIKTAS	ABBASAGA
4	34025	ZEYTINBURNU	CIRPICI

- After define the latitude-longitude function, for each postal code, lat-lon values are obtained.

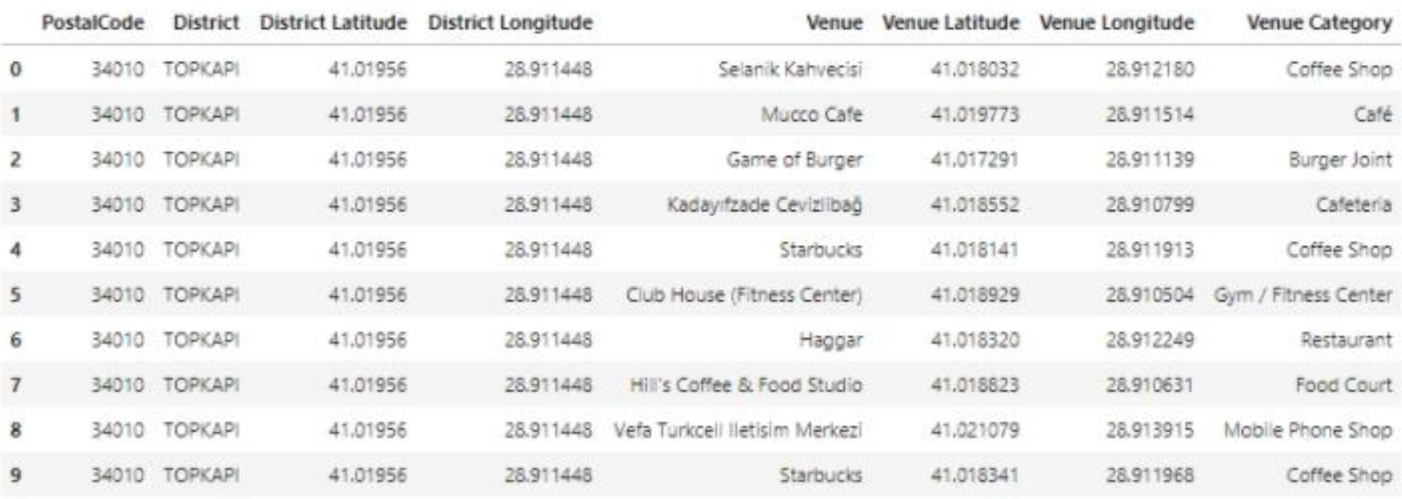
5. Methodology

- All districts are displayed on a folium map.



5. Methodology

- By using Foursquare API, all venues are listed for each district with parameters of radius=500 m and a limit of 100.

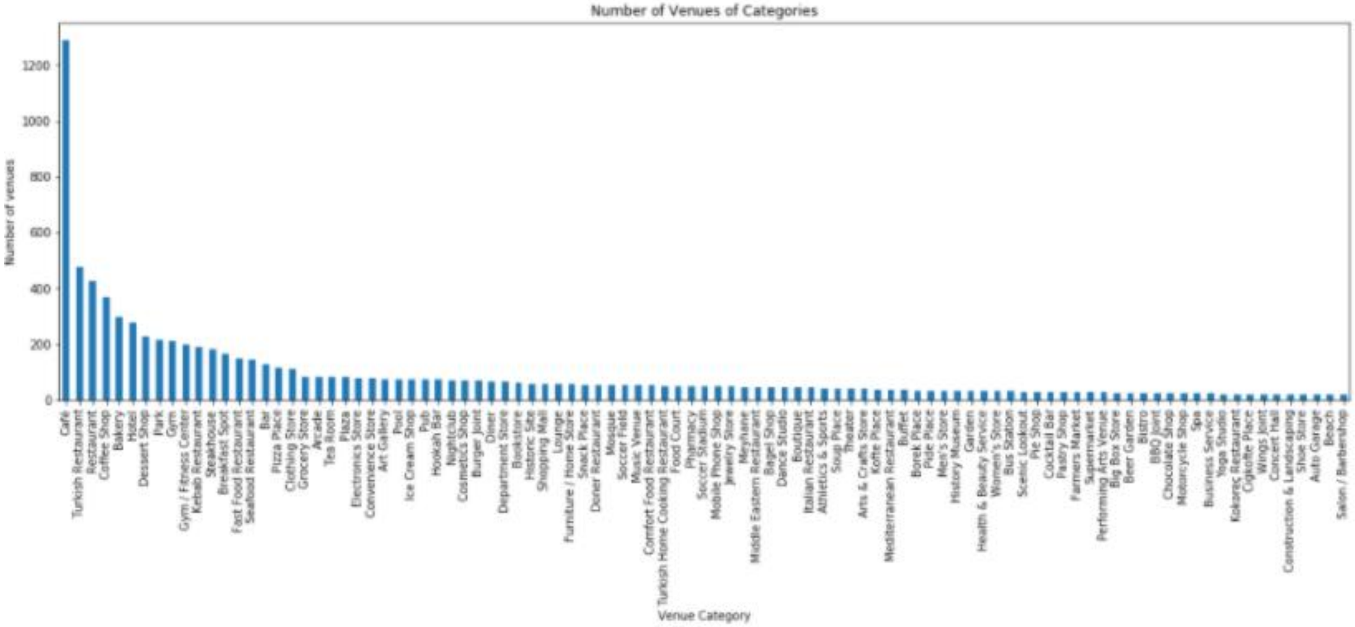


	PostalCode	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	34010	TOPKAPI	41.01956	28.911448	Selanik Kahvecisi	41.018032	28.912180	Coffee Shop
1	34010	TOPKAPI	41.01956	28.911448	Mucco Cafe	41.019773	28.911514	Café
2	34010	TOPKAPI	41.01956	28.911448	Game of Burger	41.017291	28.911139	Burger Joint
3	34010	TOPKAPI	41.01956	28.911448	Kadayıfzade Cevizlibağ	41.018552	28.910799	Cafeteria
4	34010	TOPKAPI	41.01956	28.911448	Starbucks	41.018141	28.911913	Coffee Shop
5	34010	TOPKAPI	41.01956	28.911448	Club House (Fitness Center)	41.018929	28.910504	Gym / Fitness Center
6	34010	TOPKAPI	41.01956	28.911448	Haggar	41.018320	28.912249	Restaurant
7	34010	TOPKAPI	41.01956	28.911448	Hill's Coffee & Food Studio	41.018823	28.910631	Food Court
8	34010	TOPKAPI	41.01956	28.911448	Vefa Turkcell Iletisim Merkezi	41.021079	28.913915	Mobile Phone Shop
9	34010	TOPKAPI	41.01956	28.911448	Starbucks	41.018341	28.911968	Coffee Shop

5. Methodology

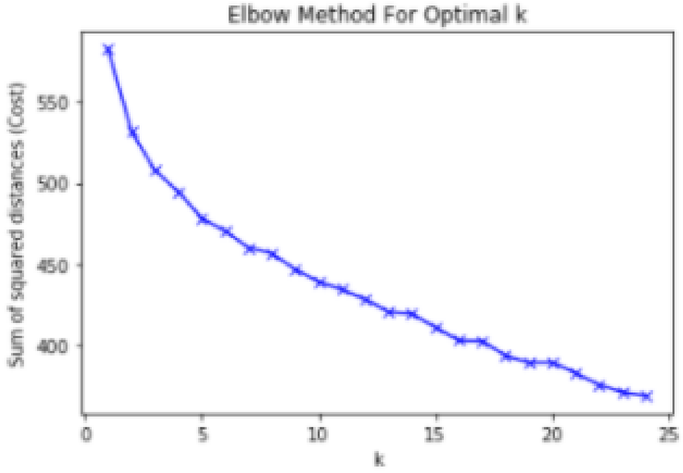
- We have number of categories which are not enough for analysis and drop those categories with a number of less than or equal to 20.

- According to bar plot of distributions, it is observed that there are mostly cafes, restaurants, bakeries and hotels with high distribution.



5. Methodology

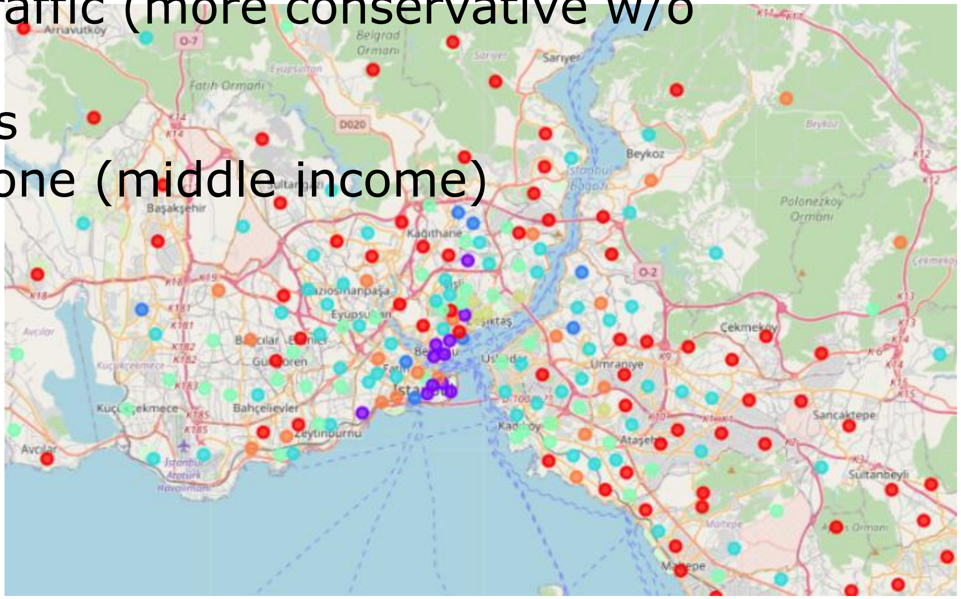
- In order to segment the districts, one-hot encoding is applied on data. In addition to that, all districts are labeled as their most common venue categories.
- In order to find the clusters, firstly min-max scaler is applied to data. After that, k-means is applied for clustering. For finding the optimal k, we examine the elbow shape of sum of squared distances. We take 7 as k (someone could take 13, 16, 19 or something different, but it generates clusters with 1 or 2 districts)



5. Methodology

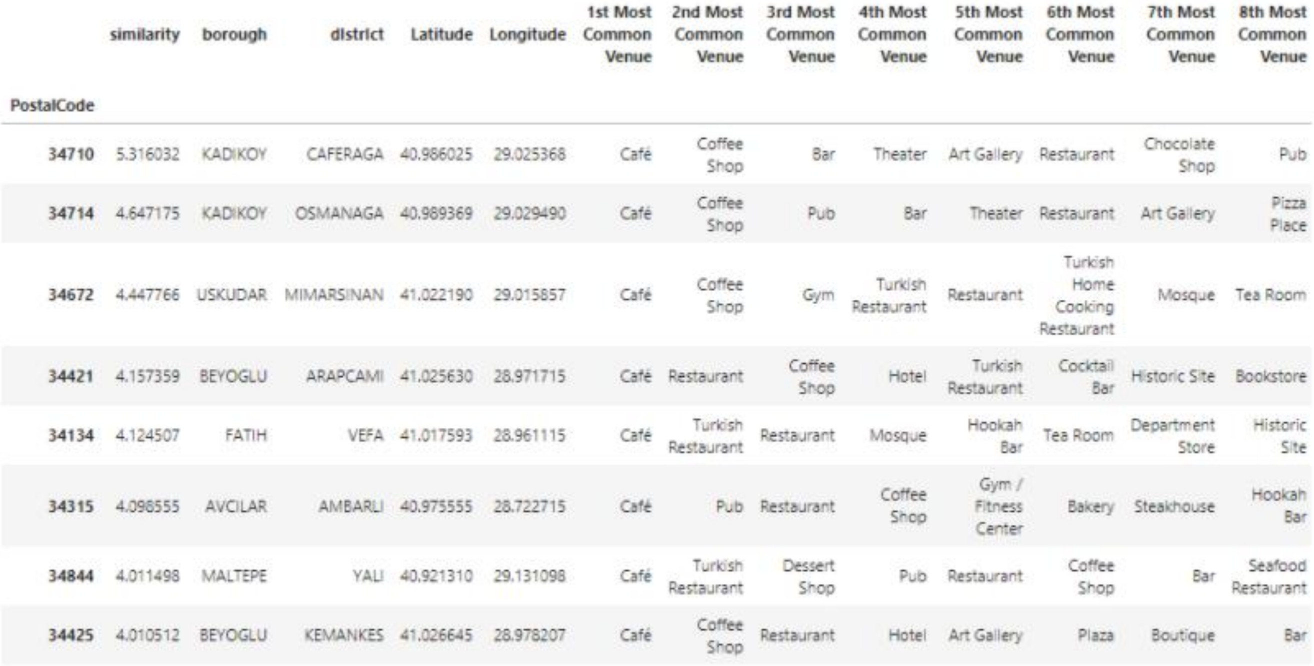
- According to clustering; the attributes of 7 clusters:

- Cluster-0: standart district with cafes, parks and restaurants. Generally located on sub-urban areas of Istanbul
- Cluster-1: hotel zones, historic places, touristic places
- Cluster-2: Touristic places, bar/pub zones
- Cluster-3: High pedestrian traffic (mostly cafes/restaurants)
- Cluster-4: High pedestrian traffic (more conservative w/o bars/pubs)
- Cluster-5: Popular cafe zones
- Cluster-6: Cafe/restaurant zone (middle income)



5. Methodology

- As the second step of our project, we try to find the most similar districts according to our ratings for specific ones.
- We calculate the similarity scores for each districts according to our scores and district venue category attributes. We just want to join the whole table, sort them and drop the rows which are rated before. Results are interesting! We like 6 out of top 10 districts. We will visit and take a tour at other 4 :)



	similarity	borough	district	Latitude	Longitude	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
PostalCode													
34710	5.316032	KADIKOY	CAFERAGA	40.986025	29.025368	Café	Coffee Shop	Bar	Theater	Art Gallery	Restaurant	Chocolate Shop	Pub
34714	4.647175	KADIKOY	OSMANAGA	40.989369	29.029490	Café	Coffee Shop	Pub	Bar	Theater	Restaurant	Art Gallery	Pizza Place
34672	4.447766	USKUDAR	MIMARSINAN	41.022190	29.015857	Café	Coffee Shop	Gym	Turkish Restaurant	Restaurant	Turkish Home Cooking Restaurant	Mosque	Tea Room
34421	4.157359	BEYOGLU	ARAPCAMI	41.025630	28.971715	Café	Restaurant	Coffee Shop	Hotel	Turkish Restaurant	Cocktail Bar	Historic Site	Bookstore
34134	4.124507	FATIH	VEFA	41.017593	28.961115	Café	Turkish Restaurant	Restaurant	Mosque	Hookah Bar	Tea Room	Department Store	Historic Site
34315	4.098555	AVCILAR	AMBARLI	40.975555	28.722715	Café	Pub	Restaurant	Coffee Shop	Gym / Fitness Center	Bakery	Steakhouse	Hookah Bar
34844	4.011498	MALTEPE	YALI	40.921310	29.131098	Café	Turkish Restaurant	Dessert Shop	Pub	Restaurant	Coffee Shop	Bar	Seafood Restaurant
34425	4.010512	BEYOGLU	KEMANKES	41.026645	28.978207	Café	Coffee Shop	Restaurant	Hotel	Art Gallery	Plaza	Boutique	Bar

6. Conclusion and Further Studies

- Our project aims to answer following questions:

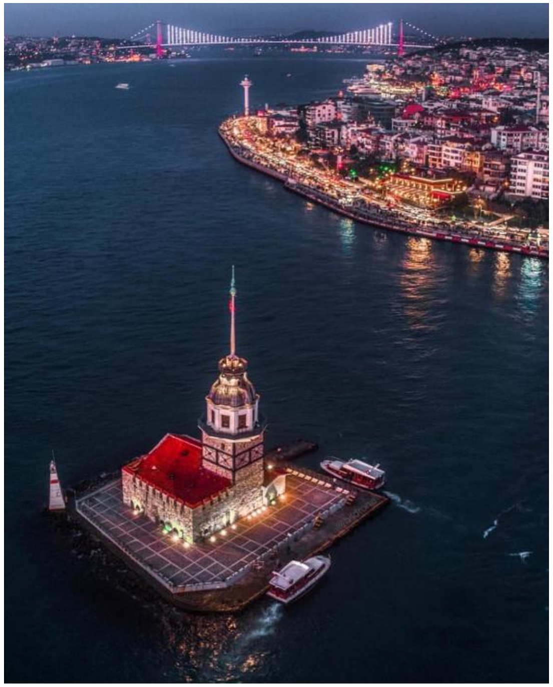
1. In Istanbul, how many district clusters and what are their attributes?
2. According to users ratings for specific districts, which districts would be recommended to that user?

As it is detailed in methodolgy section; 7 clusters are found for districts in Istanbul according to most common venue categories. Those are:

- Cluster-0: standart district with cafes, parks and restaurants. Generally located on sub-urban areas of Istanbul
- Cluster-1: hotel zones, historic places, touristic places
- Cluster-2: Touristic places, bar/pub zones
- Cluster-3: High pedestrian traffic (mostly cafes/restaurants)
- Cluster-4: High pedestrian traffic (more conservative w/o bars/pubs)
- Cluster-5: Popular cafe zones
- Cluster-6: Cafe/restaurant zone (middle income)

6. Conclusion and Further Studies

Also, according to user ratings; we could recommend different districts to discover.



For further studies; district data could be enriched with different types of data such as distance to public transport, accessibility, resident/pedestrian demographics and so on. It also helps people to find similar locations for their own tastes.