Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera
F. Gonca SAHIN - 2020

A. Introduction

A.1. A survey on the most popular venues around the main hospitals in Istanbul

Istanbul is the most populous city in Turkey and the country's economic, cultural and historic centre with 15 million population and 2.813 people per square kilo meter. The city is divided into 39 districts and most of the districts differ from each other with different characteristics [1].

According to the last data of TUIK, there are 33.052 doctors, 34.502 nurses, and 27.392 other health care workers currently working in Istanbul [2].

This analysis based on 22 university hospitals in Istanbul aims to be a guide for health care workers who are about to move to Istanbul and need information about the neighbourhood of hospital where they will work in.

A.2. Data description

Following data sources will be used to extract and generate the required information:

- The dataset of health centres in Istanbul will be obtained from Istanbul Metropolitan Municipality [3]

- The number of venues and their type and location in every hospital neighbourhood will be obtained using Foursquare API [4]

B. Data acquisition and cleaning

B.1. Data Sources

I obtain the dataset of health centres in Istanbul from Istanbul Metropolitan Municipality. To visualise the Turkish alphabet properly, 'windows-1254' encoding is used.

B.2. Data Cleaning and Feature Selection

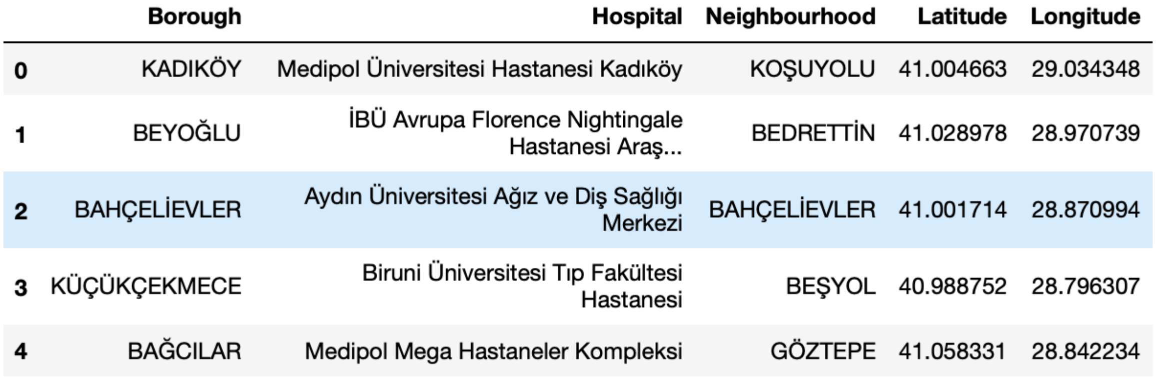
To clean the dataset:

- The dataset includes all categories of health centres like veterinarians, dental practitioners or primary care physician. Firstly the results are filtered according to category and only university hospitals are selected for the project.

- The unneeded columns are dropped and only hospital name, borough, neighbourhood, latitude and longitude data are kept.

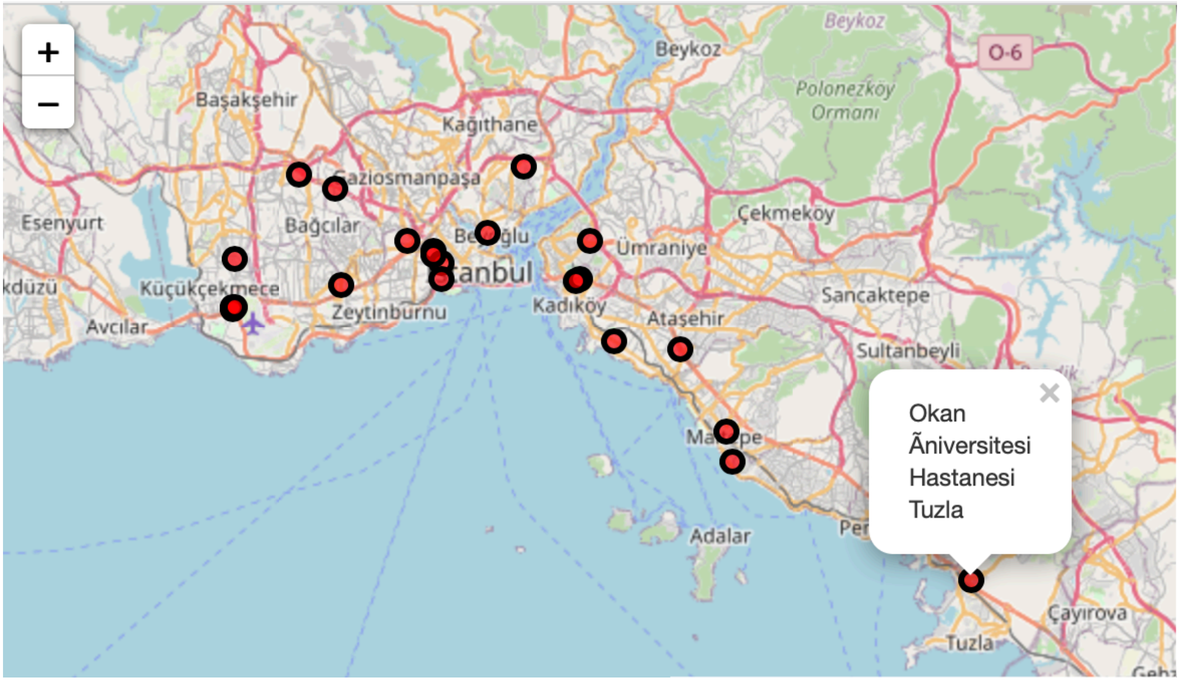
- The column names are converted to English since the original dataset is in Turkish.

When we check the shape of the data frame, there are 22 university hospitals in Istanbul.

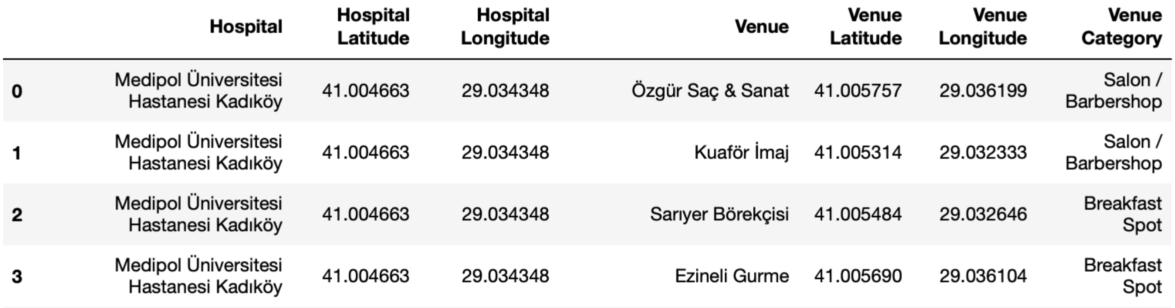


	Borough	Hospital	Neighbourhood	Latitude	Longitude
0	KADIKÖY	Medipol Üniversitesi Hastanesi Kadıköy	KOŞUYOLU	41.004663	29.034348
1	BEYOĞLU	İBÜ Avrupa Florence Nightingale Hastanesi Araş...	BEDRETTİN	41.028978	28.970739
2	BAHÇELİEVLER	Aydın Üniversitesi Ağız ve Diş Sağlığı Merkezi	BAHÇELİEVLER	41.001714	28.870994
3	KÜÇÜKÇEKMECE	Biruni Üniversitesi Tıp Fakültesi Hastanesi	BEŞYOL	40.988752	28.796307
4	BAĞCILAR	Medipol Mega Hastaneler Kompleksi	GÖZTEPE	41.058331	28.842234

We can see these hospitals on map:



Since I have the location of hospitals now, I use Foursquare API to get information on venues in each neighbourhood. I define a function and use it to extract the category of the venues (max:100) around hospital locations in a circle with 500m radius. As a result, I obtain a data frame including the hospitals, hospital location, venue names, categories and locations. There are 1450 venues in total which are extracted from Foursquare API.



	Hospital	Hospital Latitude	Hospital Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Medipol Üniversitesi Hastanesi Kadıköy	41.004663	29.034348	Özgür Saç & Sanat	41.005757	29.036199	Salon / Barbershop
1	Medipol Üniversitesi Hastanesi Kadıköy	41.004663	29.034348	Kuaför İmaj	41.005314	29.032333	Salon / Barbershop
2	Medipol Üniversitesi Hastanesi Kadıköy	41.004663	29.034348	Sarıyer Börekçisi	41.005484	29.032646	Breakfast Spot
3	Medipol Üniversitesi Hastanesi Kadıköy	41.004663	29.034348	Ezineli Gurme	41.005690	29.036104	Breakfast Spot

C.Methodology

In this project I direct my efforts on creating a guide for health care workers who are about to move to Istanbul and need information about their working environment.I aim to define the most popular venues around main hospitals in Istanbul and cluster them.

In first step I have collected the required data: name and location of the hospitals, category and location of the venues within 500m from the defined hospitals.

Now, we have some common venue categories around defined hospitals. Second step in our analysis will be clustering the data.

I decided to use K-means algorithm to cluster the hospitals, which is one of the most common clustering method of unsupervised learning. K-means is a partition-based clustering which is relatively efficient on medium and large sized data sets. Despite it is considered the one of the simplest models, k-means is especially useful for quick insights from unlabelled data.

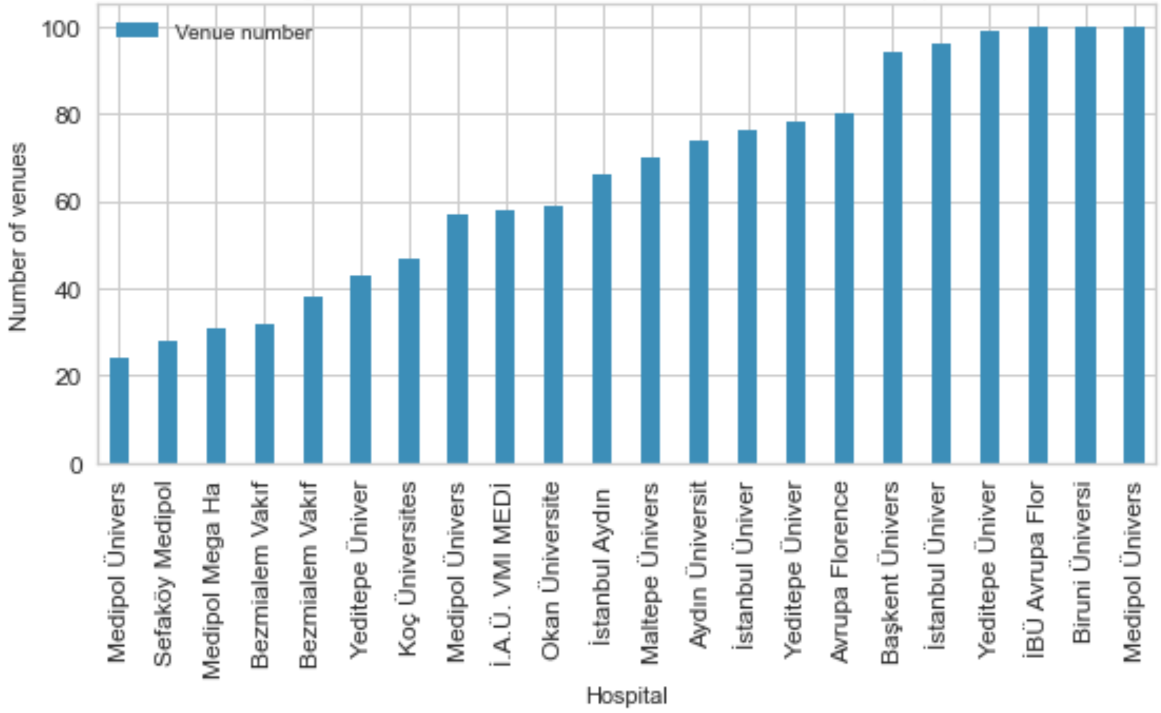
It produces sphere-like clusters because the clusters are shaped around the centroids and, its drawback is that we should pre-specify the number of clusters. To define this number, I will analyse the K-Means with elbow method.

Determining the number of clusters in a data set is a frequent problem in data clustering. The correct choice of K is very dependent on the shape and scale of the distribution of points in a dataset. Elbow method runs the clustering across the different values of K. But the problem is that with increasing the number of clusters, the distance of centroids to data points will always reduce. This means increasing K will always decrease the error. So, the value of the metric as a function of K is plotted and the elbow point is determined where the rate of decrease sharply shifts. It is the right K for clustering. This method is called the elbow method [5].

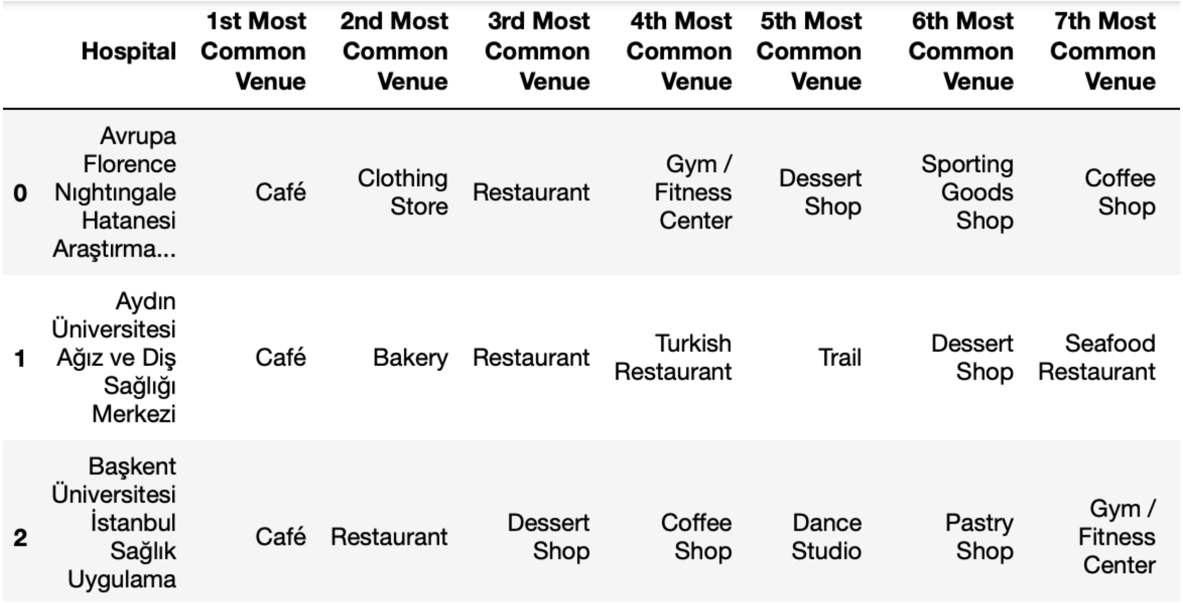
C.1. Analyse

Before starting to cluster, I convert the data frame by using get_dummies method of pandas. This data frame is used for clustering. I also create a bar chart to show the number of venues around each hospital. We can see that the total number of venues around just 3 hospital reach the number of 100, which I defined as a limit.

Figure 1: The number of venues around each hospital

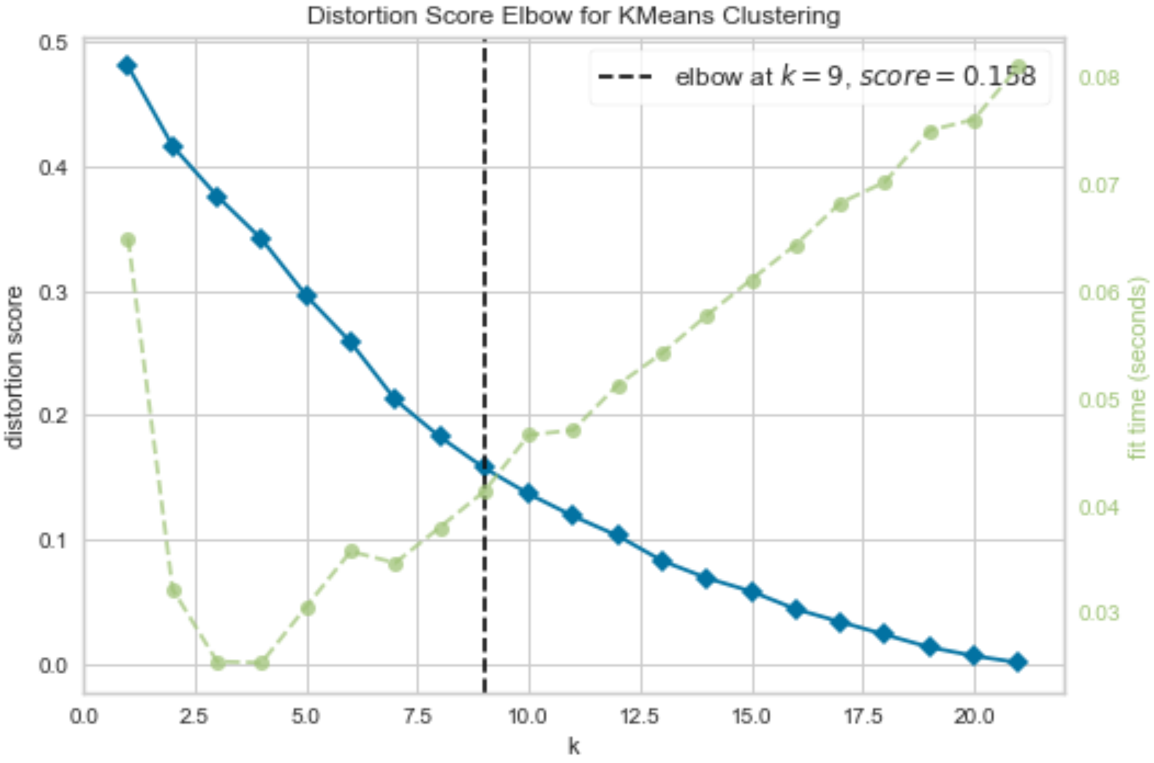


Next, I group rows by 'Hospital' and by taking the mean of the frequency of occurrence of each category. By defining a new function, I create a data frame showing the most common venues around each hospital.



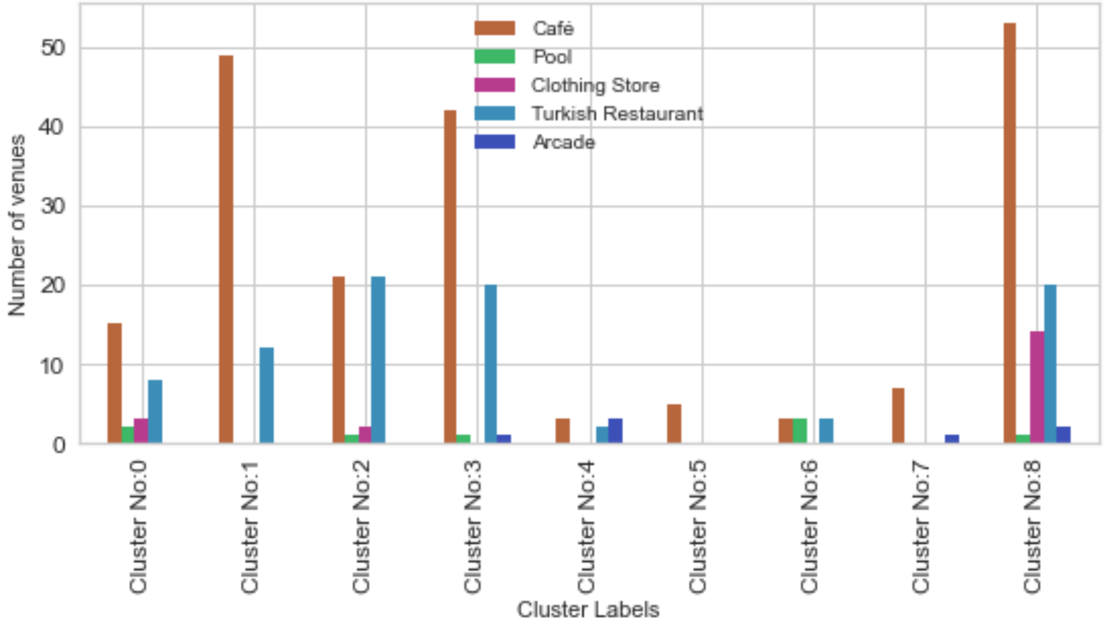
	Hospital	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Avrupa Florence Nıghtıngale Hatanesi Araştırma...	Café	Clothing Store	Restaurant	Gym / Fitness Center	Dessert Shop	Sporting Goods Shop	Coffee Shop
1	Aydın Üniversitesi Ağız ve Diş Sağlığı Merkezi	Café	Bakery	Restaurant	Turkish Restaurant	Trail	Dessert Shop	Seafood Restaurant
2	Başkent Üniversitesi İstanbul Sağlık Uygulama	Café	Restaurant	Dessert Shop	Coffee Shop	Dance Studio	Pastry Shop	Gym / Fitness Center

Before fit the data in K-Means clustering, I use Elbow method to define the optimal cluster number and the score. According to the results, elbow method ensured me the 9 degree for optimum k of the K-Means.

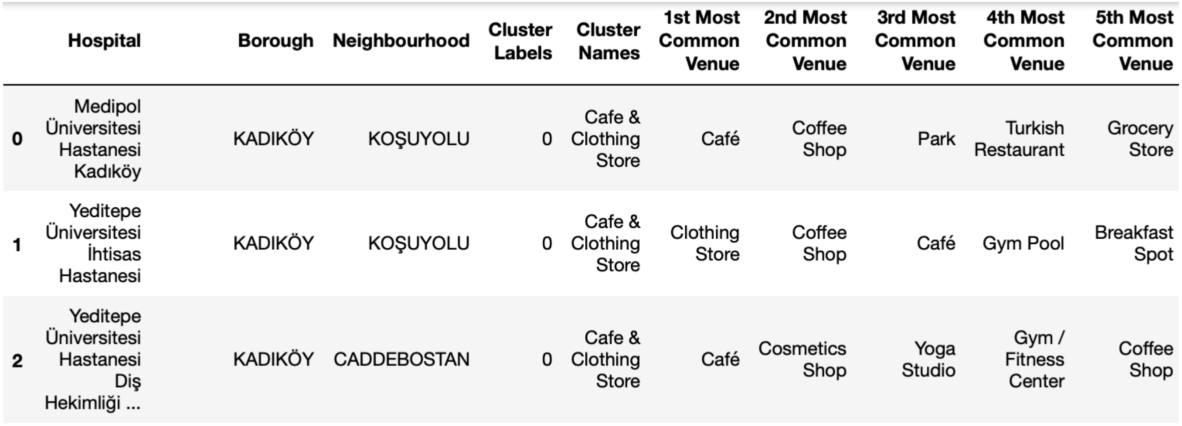


I use K-Means method and cluster the hospitals according to the venues around them. Now, I have labels for each cluster.
At this point, I create a bar chart showing the number of 1st Most Common Venues in each cluster. It would be helpful to use together with the table showing the most common venues around each hospital to find proper label names for each cluster.

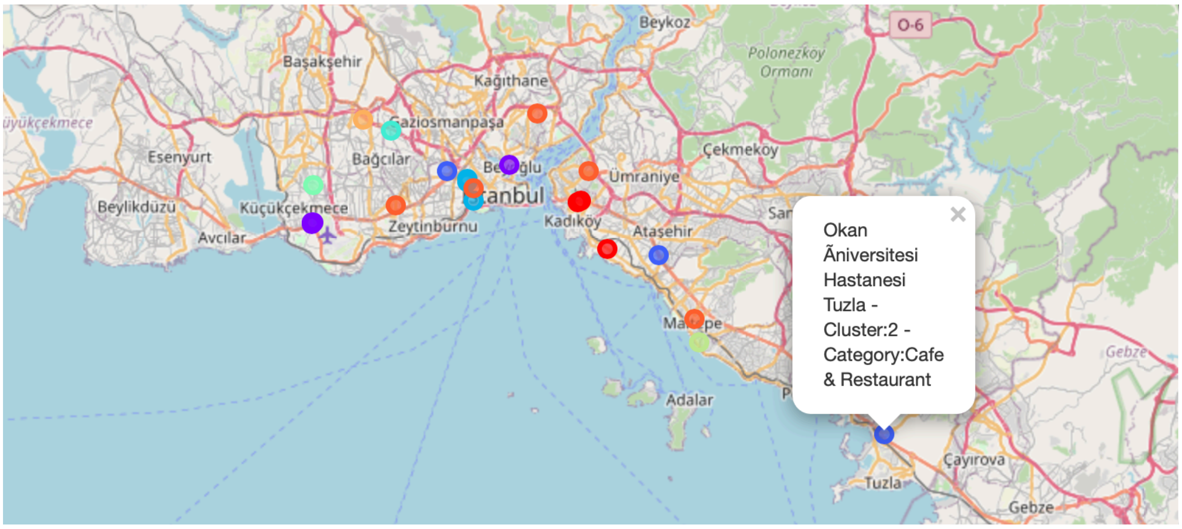
Figure 2: Distribution of the '1st most common venues' around the hospitals in each cluster



Finally, I create a master table to combine all relevant information and use this table to create a map showing hospital locations, names, cluster numbers and defined cluster names.



	Hospital	Borough	Neighbourhood	Cluster Labels	Cluster Names	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Medipol Üniversitesi Hastanesi Kadıköy	KADIKÖY	KOŞUYOLU	0	Cafe & Clothing Store	Café	Coffee Shop	Park	Turkish Restaurant	Grocery Store
1	Yeditepe Üniversitesi İhtisas Hastanesi	KADIKÖY	KOŞUYOLU	0	Cafe & Clothing Store	Clothing Store	Coffee Shop	Café	Gym Pool	Breakfast Spot
2	Yeditepe Üniversitesi Hastanesi Diş Hekimliği ...	KADIKÖY	CADDEBOSTAN	0	Cafe & Clothing Store	Café	Cosmetics Shop	Yoga Studio	Gym / Fitness Center	Coffee Shop



D. Results and Discussion

Aa a big city and economic, cultural and historic centre of the country, Istanbul has a high population and density in a narrow area. The city is divided into 39 districts and most of the districts differ from each other with different characteristics. Thus, it is always challenging for the new residents to understand the dynamics of the city.

This analysis aims to be a guide for health care workers who are about to move to Istanbul and need information about the neighbourhood of hospital where they will work in. There are more than 95 thousands of health care workers, currently working in Istanbul.

This analysis based on 22 university hospitals in Istanbul. The K-means algorithm was preferred as part of this clustering study. According to the results of Elbow method, the optimum k value was set to 9.

I created bar charts showing the number of venues around each hospital and the distribution of the '1st most common venues' around the hospitals in each cluster.

I ended the study by visualising the data and clustering information of the hospitals on the Istanbul map including hospital locations, names, cluster numbers and defined cluster names.

E. Conclusion

In this study, I analysed the venues around the main hospitals in Istanbul/Turkey to guide new residents of Istanbul as a health care worker. For more detailed and accurate guidance, the data set can be expanded so other hospitals can also be drilled, and different approaches can be tried in clustering and classification.

F. References

[1] https://en.wikipedia.org/wiki/Istanbul
[2] https://data.tuik.gov.tr/Bulten/DownloadIstatistikselTablo?p=1sDY/9oQZO8DamcSJ9zgtnOrS10JgqV1stehK2Rz9SbQD33xL9rhoK4BN8742etP
[3] https://data.ibb.gov.tr/dataset/istanbul-saglik-kurum-ve-kuruluslari-verisi
[4] https://developer.foursquare.com
[5] https://www.coursera.org/learn/machine-learning-with-python/lecture/rLcgP/more-on-k-means