CAPSTONE PROJECT
REPORT

Tuğçe KARAGÖL
April 15, 2021

1. Introduction

1.1 Background

Istanbul is the most populous city in Turkey and is also one of the most populated cities in the World. It is a unique city that combines historic fabric and modernity at the same time. It has a special location as a bridge between Asia and Europe. This city, which has a high population density due to its small surface area relatively its population, consists of 39 districts.

Istanbul is also one of the most advanced cities in Turkey. For this reason, it attracts a lot of people and lets in immigrants. The range of income distribution is very wide in İstanbul. There are people of all income levels. It offers different opportunities to people with different income distributions. Social and cultural structuring in the districts of Istanbul reflects these differences.

1.2 Problem

As mentioned before, Istanbul has constantly allowed immigrants.This study aims to be a guide for people who want to settle in İstanbul and having own real estate there. For this purpose, in this study, real estate prices for each district will be investigated and comparisons will be made on square meter costs. Besides, district-based venues will be analyzed to give an idea about the social opportunities of districts. People who want to have a house in Istanbul are expected to decide on the district that is suitable for both their budget and their social life by taking advantage of the results of this study.

2. Data acquisition and cleaning

2.1 Data sources

First of all , the district list of Istanbul data is obtained from Wikipedia https://en.wikipedia.org/wiki/List_of_districts_of_Istanbul

Latitude and longitude information are obtained using GeoPy library.

Average house sales price data is obtained from local real estate sales web page (https://www.mahalligundem.com/hurriyet-emlak-2021-endeksi-aciklandi/47283/). It is converted to csv file to use. (https://github.com/TugceKaragol/Coursera_Capstone/blob/6f276377cbf42239dd622efedce673b1251fb3f4/Avg_House_Sales_Price.csv)

Finally Foursquare API is used to get the most common venues of Districts of Istanbul.

2.2 Data cleaning

District data was scraped from the web and it was formed a data table after removing unnecessary columns.

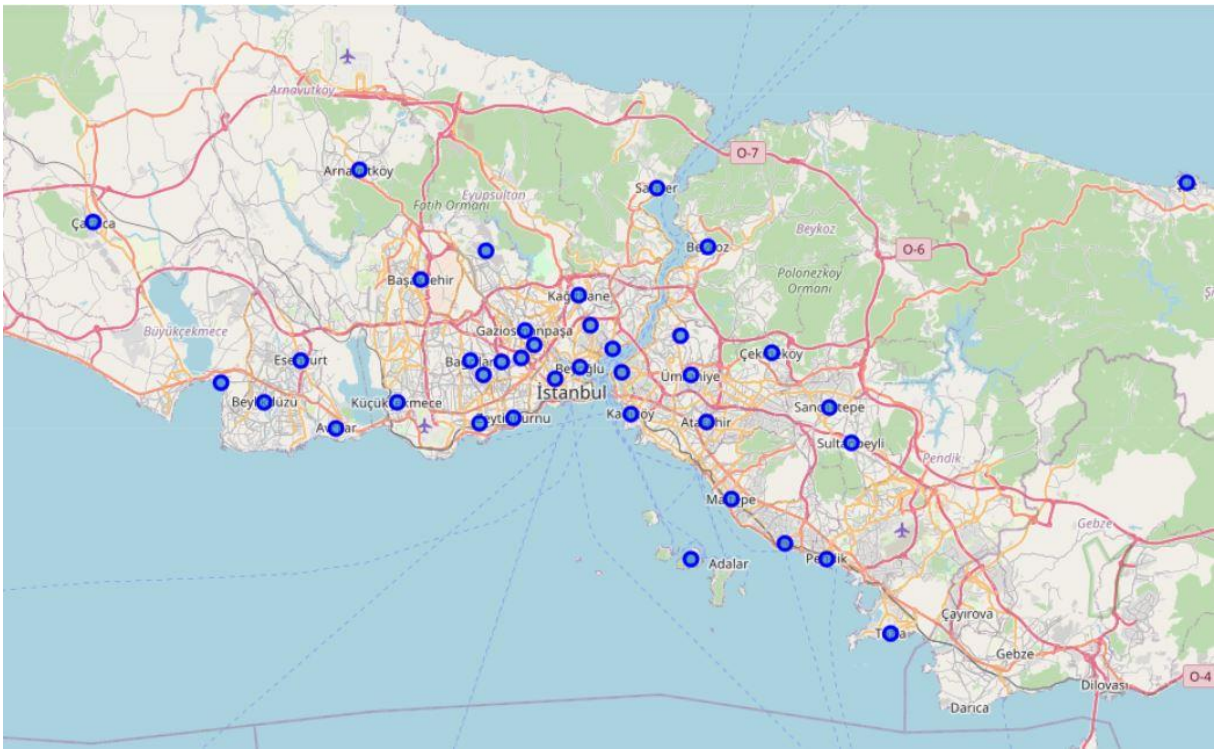
After the data was collected, it is first formed a data table. Later unnecessary columns are deleted. When it is ready for use, it is converted to csv file.

3. Methodology

I scraped Wikipedia to obtain the list of districts of Istanbul and other information and I created a dataframe. Then I got latitude and longitude information for each district using the GeoPy library of Pytho and added this information to the data frame.

	District	Population (2020)	Area (km²)	Density (per km²)	Latitude	Longitude
0	Adalar	16033	11.05	1451	40.876259	29.091027
1	Arnavutköy	296709	450.35	659	41.184471	28.741245
2	Ataşehir	422594	25.23	16750	40.984749	29.106720
3	Avcılar	436897	42.01	10400	40.980135	28.717547
4	Bağcılar	737206	22.36	32970	41.033899	28.857898

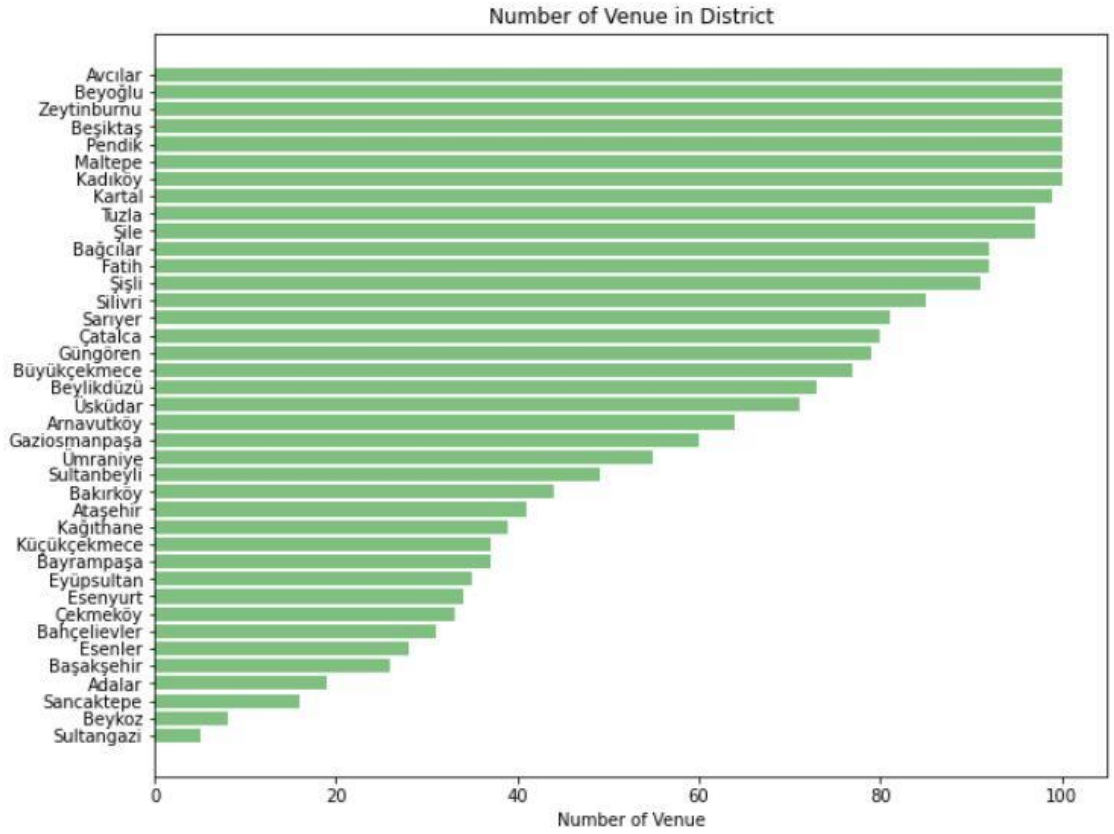
I used the Folium library to create an Istanbul map and visualize districts.



I used the Foursquare API to explore the venues in each district. I chose a radius of 500 meters and a limit of 100 for the query. A data frame is obtained consist of venue name, its categories, and its latitude and longitude values as a result of the query. Then I merged this dataframe and districts in a table.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adalar	40.876259	29.091027	Merit Halki Palace Hotel	40.878802	29.090974	Hotel
1	Adalar	40.876259	29.091027	İnönü Evi Müzesi	40.878251	29.093647	History Museum
2	Adalar	40.876259	29.091027	L'isola Guesthouse	40.877038	29.096136	Bed & Breakfast
3	Adalar	40.876259	29.091027	Aqua Green Beach	40.880498	29.090354	Beach
4	Adalar	40.876259	29.091027	Asaf Beach Club	40.879211	29.088317	Surf Spot

I grouped the venues by district and counted them. Then I decided to visualize in to see the results better. I created a horizontal bar chart using matplotlib library.



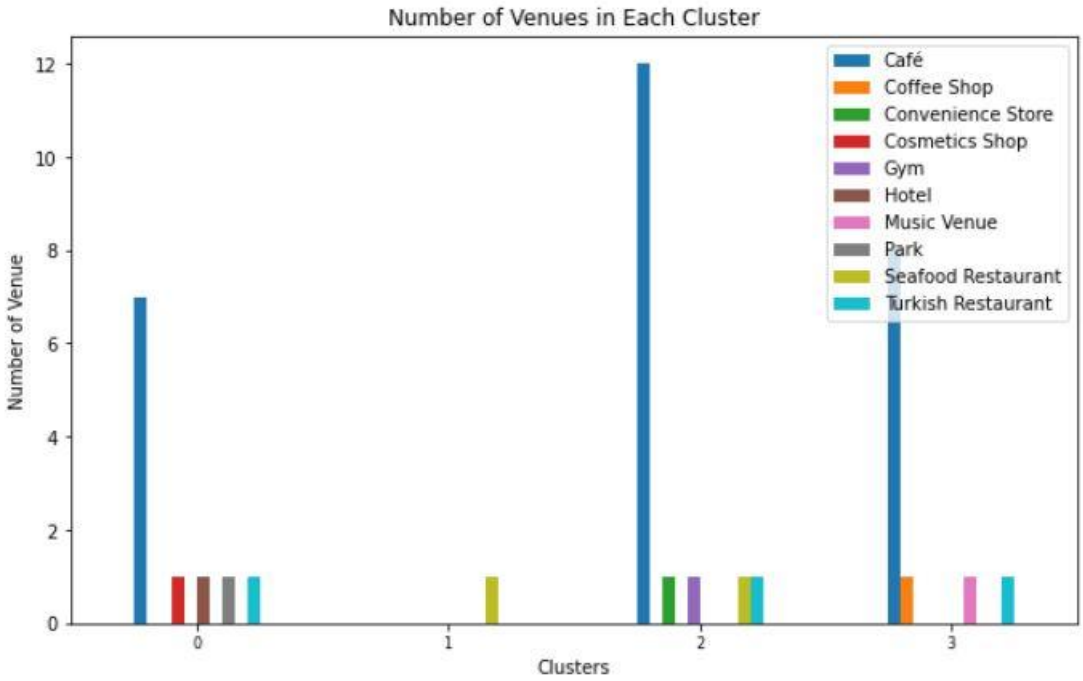
The bar chart shows that 7 districts which are Avcılar, Beyoğlu, Zeytinburnu, Beşiktaş, Pendik, Maltepe and Kadıköy reach 100 limit. 4 Districts - Adalar, Sancaktepe, Beykoz, Sultangazi- are below 20 venues and in the other districts number of venues change between 20 and 90.

I analyzed each district according to the venue information and grouped rows by district and by taking the mean of the frequency of occurrence of each category. Then I generated a function to explore the most common 10 venues in each district and created a table to show that.

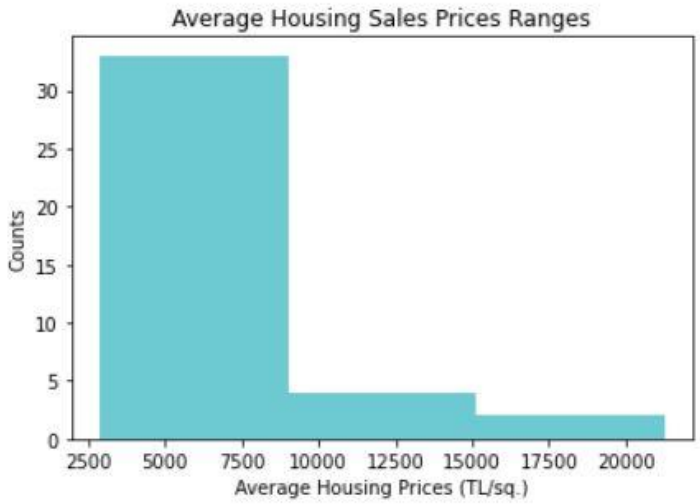
	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adalar	Café	Scenic Lookout	Mountain	Bike Trail	Museum	Tennis Court	Beach	Bed & Breakfast	Surf Spot	University
1	Arnavutköy	Café	Dessert Shop	Restaurant	Coffee Shop	Steakhouse	Turkish Restaurant	Electronics Store	Fast Food Restaurant	Cafeteria	Kofte Place
2	Ataşehir	Café	Restaurant	Doner Restaurant	Coffee Shop	Steakhouse	Gym Pool	Bistro	Lounge	Bakery	Kebab Restaurant
3	Avcılar	Café	Gym / Fitness Center	Turkish Restaurant	Dessert Shop	Bar	Restaurant	Breakfast Spot	Pizza Place	Coffee Shop	Bistro
4	Bahçelievler	Park	Café	Restaurant	Trail	Middle Eastern Restaurant	Bakery	General Entertainment	Garden	Supermarket	Steakhouse

I used the K-means algorithm to cluster the districts and determined the ' k ' value as 4. Then, added a Cluster Label column to show each district's cluster. To visualize clusters, I used the matplotlib library and created a bar chart.

	District	Population (2020)	Area (km²)	Density (per km²)	Latitude	Longitude	Cluster Labels	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	Adalar	16033	11.05	1451	40.876259	29.091027	2	Café	Scenic Lookout	Mountain	Bike Trail	Museum	Tennis Court	Beach
1	Arnavutköy	296709	450.35	659	41.184471	28.741245	3	Café	Dessert Shop	Restaurant	Coffee Shop	Steakhouse	Turkish Restaurant	Electronics Store
2	Ataşehir	422594	25.23	16750	40.984749	29.106720	2	Café	Restaurant	Doner Restaurant	Coffee Shop	Steakhouse	Gym Pool	Bistro
3	Avcılar	436897	42.01	10400	40.980135	28.717547	3	Café	Gym / Fitness Center	Turkish Restaurant	Dessert Shop	Bar	Restaurant	Breakfast Spot
4	Bağcılar	737206	22.36	32970	41.033899	28.857898	0	Café	Gym	Coffee Shop	Dessert Shop	Hookah Bar	Turkish Restaurant	Music Venue



I read the average house sales price data from a CSV file and converted it into a dataframe. Then, I divided the average house sales prices into 3 groups as low level, medium level, and high level, and create a histogram to see the results.



According to the histogram, the average house sales price below 9000 TL/m2 represents the low level, 9000 TL/m2 to 15000 TL/m2 the mid-level, and above the 15000 TL/m2 the high level.

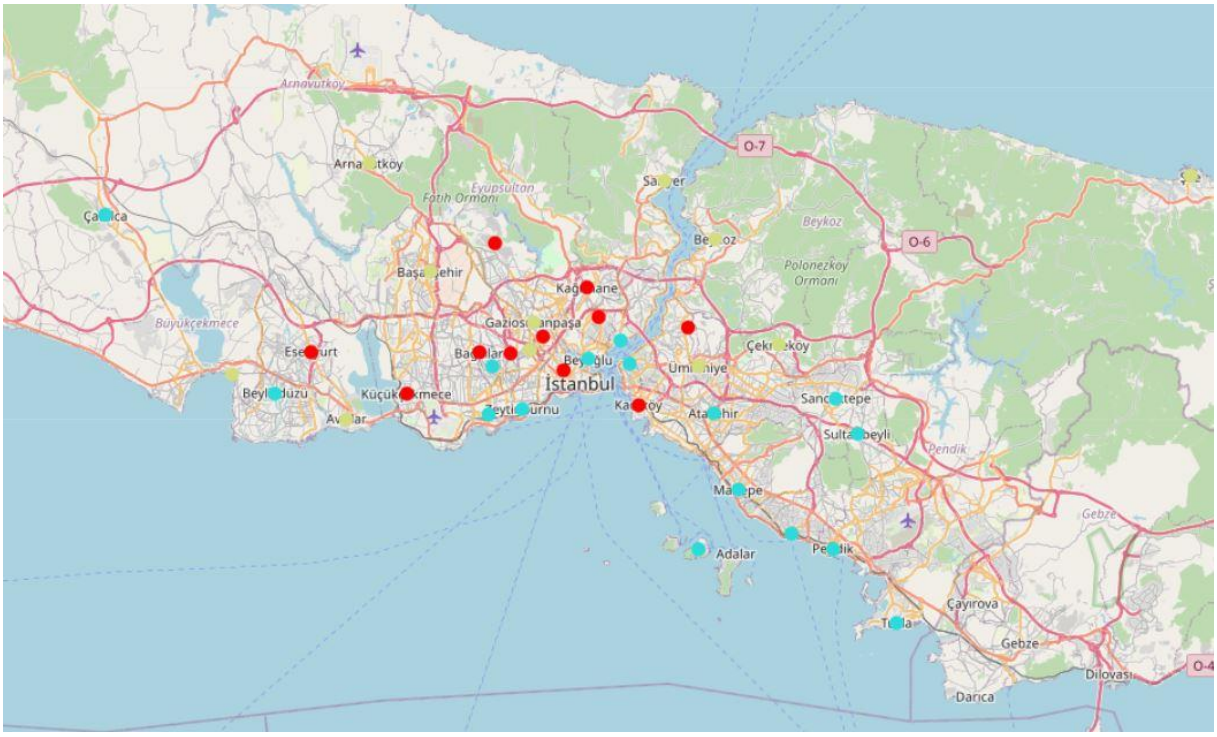
I added a level labels column into the average house sales price dataframe to show levels.

	District	Avg House Sales Price (TL/m2)	Level_labels
0	Esenyurt	2875	Low Level House Price
1	Esenler	3090	Low Level House Price
2	Sultangazi	3125	Low Level House Price
3	Silivri	3252	Low Level House Price
4	Çatalca	3294	Low Level House Price

Finally, I explored the top 3 venues for each district, combined them, and added a combined column.

	District	Combine
0	Adalar	3 Café,2 Mountain,2 Scenic Lookout
1	Arnavutköy	12 Café,3 Dessert Shop,3 Restaurant
2	Ataşehir	5 Café,5 Restaurant,3 Coffee Shop
3	Avcılar	27 Café,4 Bar,4 Dessert Shop
4	Bahçelievler	3 Park,2 Café,2 Restaurant

After I merged all data into a data table, I created an Istanbul map again to visualize the analysis results.



4. Discussion

In this study, average house sales prices for 39 districts of Istanbul were examined. Also, the venues were explored to understand the social structures of the district and to have an idea about the possibilities they offer.

Clustering work has been done for venues in each district using the K-means algorithm. In this work, the radius is determined as 500 meters. This reduces the quality of the study. Different values to be used will affect the result. Also, the value of k was chosen as 4. When the optimum k value is calculated with different methods, the accuracy of the study can increase.

A clustering study has been carried out on average house sales prices. It is divided into 3 clusters as low level, medium level, and high level. It is seen as a result of the study that the price ranges of these clusters are wide. Increasing the number of clusters and lowering the price ranges will increase the quality of the study.

5. Conclusion

As mentioned before, Istanbul is a city that receives a lot of immigration. The range of income distribution in Istanbul is wide and it affects the structure of the districts. People who intend to settle in Istanbul and own a house can benefit from these and similar studies while determining the districts that are suitable for their social life and budget.

6. References

- Wikipedia – List of Districts of Istanbul
- Average House Sales Price
- Foursquare API