Housing Sales Prices & Venues Data Analysis of Istanbul

A. Introduction

A.1. Description & Disscusion of the Background

Istanbul is one of the largest metropolises in the world where over **15 million** people live and it has a population density of **2.813** people per square kilometer. As a resident of this city, I decided to use Istanbul in my project. The city is divided into 39 districts in total. However, the fact that the districts are squeezed into an area of approximately **72** square kilometers causes the city to have a very intertwined and mixed structure [1].

As you can see from the figures, Istanbul is a city with a high population and population density. Being such a crowded city leads the owners of shops and social sharing places in the city where the population is dense. When we think of it by the investor, we expect from them to prefer the districts where there is a lower real estate cost and the type of business they want to install is less intense. If we think of the city residents, they may want to choose the regions where real estate values are lower, too. At the same time, they may want to choose the district according to the social places density. However, it is difficult to obtain information that will guide investors in this direction, nowadays.

When we consider all these problems, we can create a map and information chart where the real estate index is placed on Istanbul and each district is clustered according to the venue density.

A.2. Data Description

To consider the problem we can list the datas as below:

- I found the Second-level Administrative Divisions of the Turkey from Spatial Data Repository of NYU [2]. The .json file has coordinates of the all city of Turkey. I cleaned the data and reduced it to city of Istanbul where I used it to create choropleth map of Housing Sales Price Index of Istanbul.
- I used **Forsquare API** to get the most common venues of given Borough of Istanbul [3].
- There are not too many public datas related to demographic and social parameters for the city of Istanbul. Therefor you must set-up your own data tables in most cases. In this case, I collected latest per

square meter Housing Sales Price (HSP) Averages for each Borough of Istanbul from housing retail web page [4].

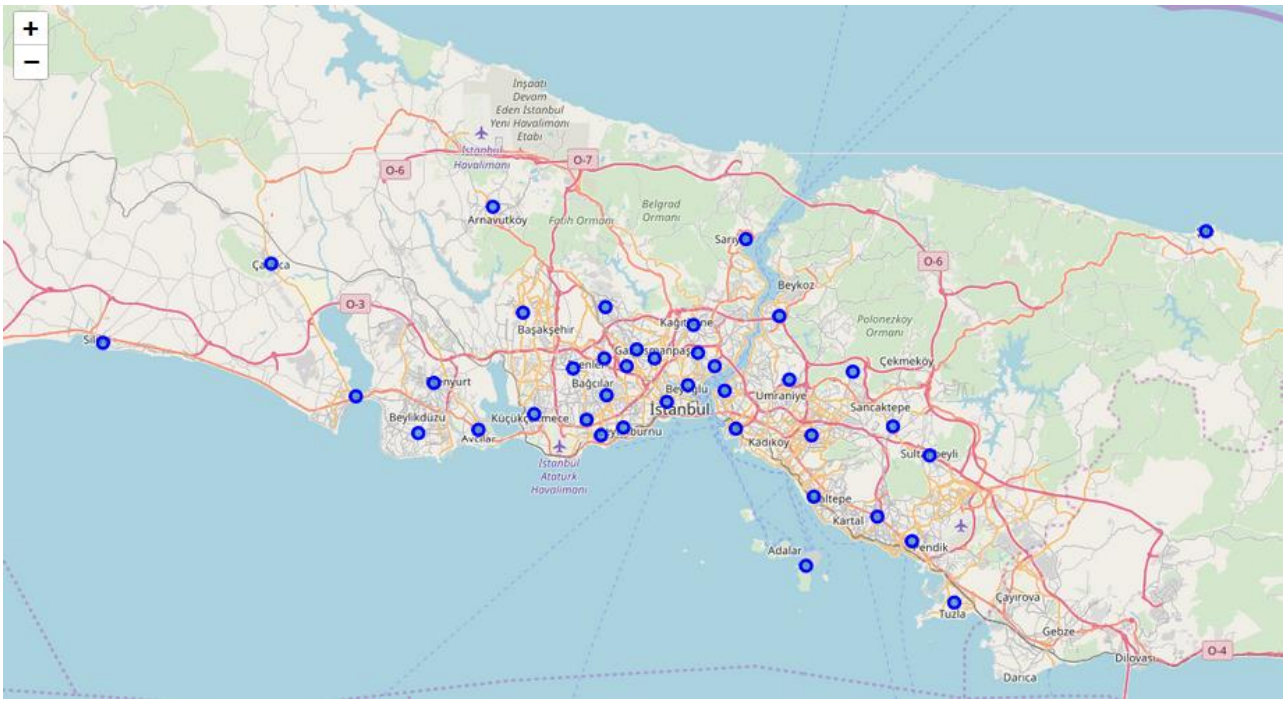
- I used Google Map, 'Search Nearby' option to get the center coordinates of the each Borough. [5].

B. Methodology

As a database, I used GitHub repository in my study. My master data which has the main components *Borough, Average House Price, Latitude* and *Longitude* informations of the city.

	Borough	Avg-HousePrice	Latitude	Longitude
0	Adalar	5568	40.8619	29.1208
1	Arnavutkoy	2265	41.1956	28.7352
2	Atasehir	5512	40.9831	29.1279
3	Avcilar	2454	40.9880	28.7170
4	Bagcilar	3264	41.0450	28.8338

I used python **folium** library to visualize geographic details of Istanbul and its boroughs and I created a map of Istanbul with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:



I utilized the Foursquare API to explore the boroughs and segment them. I designed the limit as **100 venue** and the radius **750 meter** for each borough from their given latitude and longitude informations. Here is a head of the list Venues name, category, latitude and longitude informations from Forsquare API.

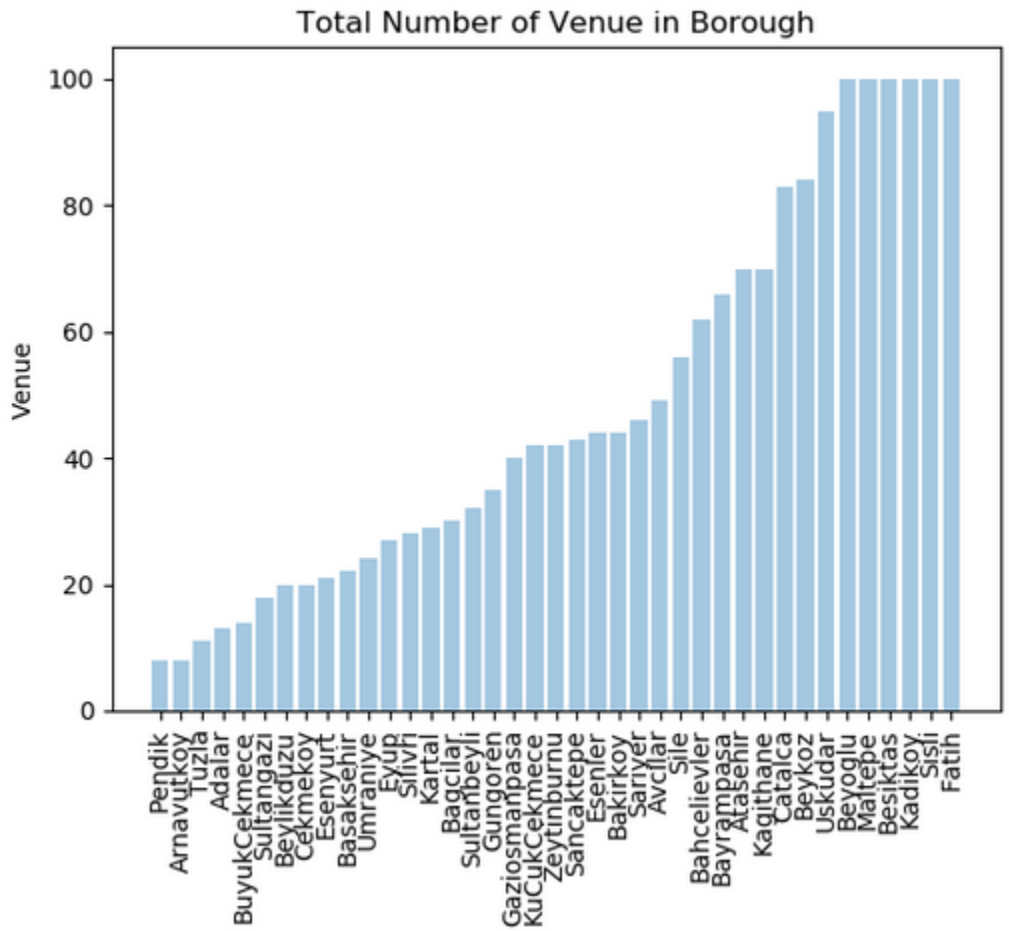
	name	categories	lat	lng
0	Büyükada Tepesi	Mountain	40.861107	29.117418
1	Eski Rum Yetimhanesi	Historic Site	40.861705	29.123323
2	Büyükada Bisiklet Parkuru	Bike Trail	40.865000	29.116861
3	Büyükada Lale köşkü	Bed & Breakfast	40.865657	29.125223
4	Nizam Butik Otel & Bistro	Bed & Breakfast	40.863322	29.116257

In summary of this data **43** venues were returned by Foursquare. Here is a merged table of boroughs and venues.

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adalar	40.8619	29.1208	Büyükada Tepesi	40.861107	29.117418	Mountain
1	Adalar	40.8619	29.1208	Eski Rum Yetimhanesi	40.861705	29.123323	Historic Site
2	Adalar	40.8619	29.1208	Nizam Butik Otel & Bistro	40.863322	29.116257	Bed & Breakfast
3	Adalar	40.8619	29.1208	Büyükada Bisiklet Parkuru	40.865000	29.116861	Bike Trail
4	Adalar	40.8619	29.1208	Asiklar Cay Bahcesi	40.860402	29.116640	Café

We can see that Kadikoy, Maltepe, Beyoglu, Besiktas, Sisli and Fatih how reached the **100** limit of venues. On the other hand; Pendik, Arnavutkoy, Tuzla, Adalar, Buyukcekmece, Sultangazi, Cekmekoy, Beylikduzu, Sultangazi boroughs are below **20** venues in our given coordinates with Latitude and Longitude, in below graph.

The result doesn't mean that inquiry run all the possible results in boroughs. Actually, it depends on given Latitude and Longitude informations and here is we just run single Latitude and Longitude pair for each borough. We can increase the possibilities with Neighborhood informations with more Latitude and Longitude informations.

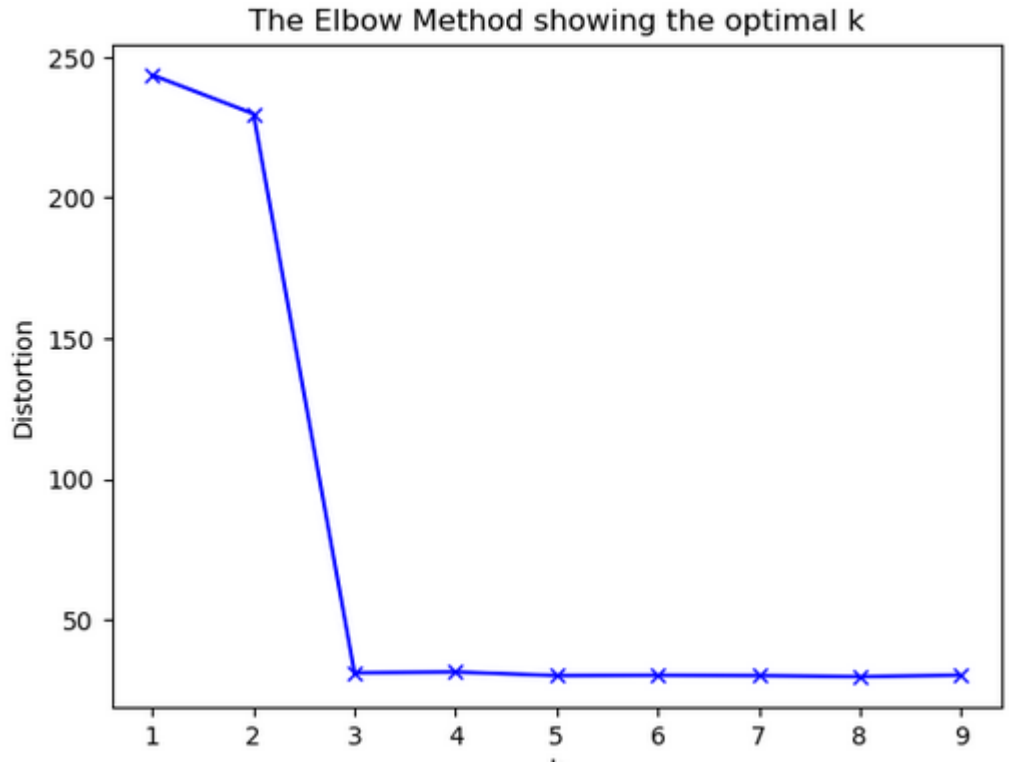


In summary of this graph **256** unique categories were returned by Foursquare, then I created a table which shows list of top 10 venue category for each borough in below table.

	Borough	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	Bed & Breakfast	Garden Center	Mountain	BBQ Joint	Hotel	Road	History Museum	Historic Site	Tea Room
1	Arnavutkoy	Arcade	Pharmacy	Restaurant	Kofte Place	Diner	Leather Goods Store	Convenience Store	Farmers Market	Electronics Store	Entertainment Service
2	Atasehir	Café	Restaurant	Pool	Spa	Clothing Store	Çöp Şiş Place	Farmers Market	Soccer Stadium	Park	Hotel
3	Avcilar	Café	Fast Food Restaurant	Turkish Restaurant	Restaurant	Coffee Shop	Shoe Store	Donut Shop	Mobile Phone Shop	Modern European Restaurant	Molecular Gastronomy Restaurant
4	Bagcilar	Café	Gym	Turkish Restaurant	Snack Place	Men's Store	Dessert Shop	Tennis Court	Tea Room	Fried Chicken Joint	Restaurant

We have some common venue categories in boroughs. In this reason I used unsupervised learning **K-means algorithm** to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning.

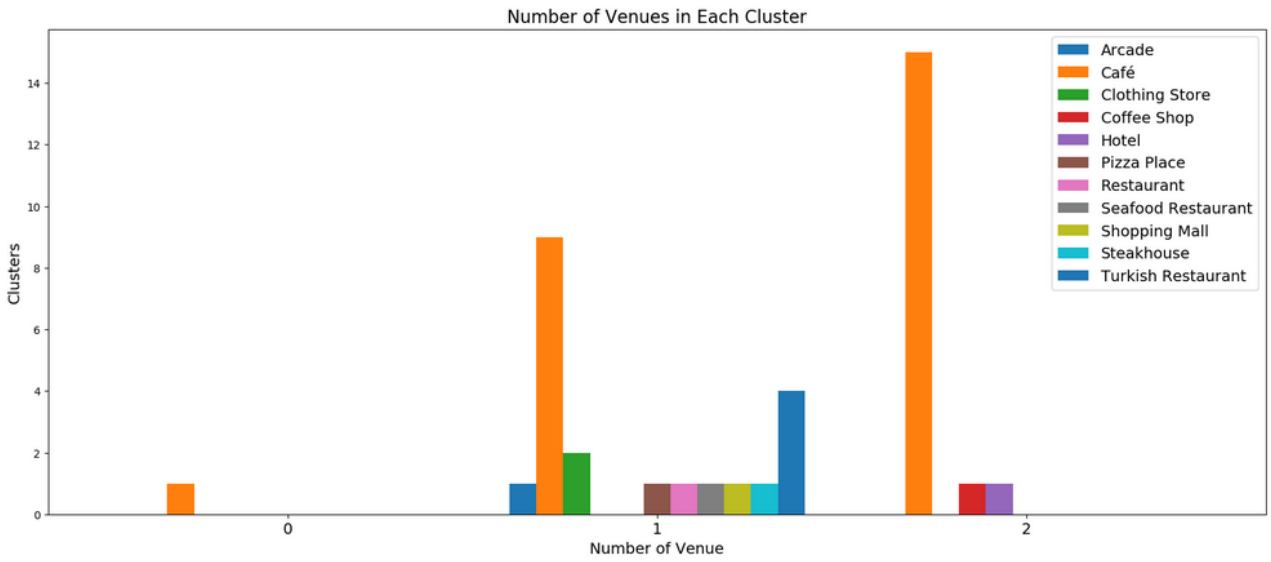
First, I will run K-Means to cluster the boroughs into **3** clusters because when I analyze the K-Means with elbow method it ensured me the 3 degree for optimum k of the K-Means.



Here is my merged table with cluster labels for each borough.

	Borough	Avg-HousePrice	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adalar	5568	40.8619	29.1208	2	Café	Bed & Breakfast	Garden Center	Mountain	BBQ Joint	Hotel	Road	History Museum	Historic Site	Tea Room
1	Arnavutkoy	2265	41.1956	28.7352	1	Arcade	Pharmacy	Restaurant	Kofte Place	Diner	Leather Goods Store	Convenience Store	Farmers Market	Electronics Store	Entertainment Service
2	Atasehir	5512	40.9831	29.1279	1	Café	Restaurant	Pool	Spa	Clothing Store	Çöp Şiş Place	Farmers Market	Soccer Stadium	Park	Hotel
3	Avcilar	2454	40.9880	28.7170	2	Café	Fast Food Restaurant	Turkish Restaurant	Restaurant	Coffee Shop	Shoe Store	Donut Shop	Mobile Phone Shop	Modern European Restaurant	Molecular Gastronomy Restaurant
4	Bagcilar	3264	41.0450	28.8338	1	Café	Gym	Turkish Restaurant	Snack Place	Men's Store	Dessert Shop	Tennis Court	Tea Room	Fried Chicken Joint	Restaurant

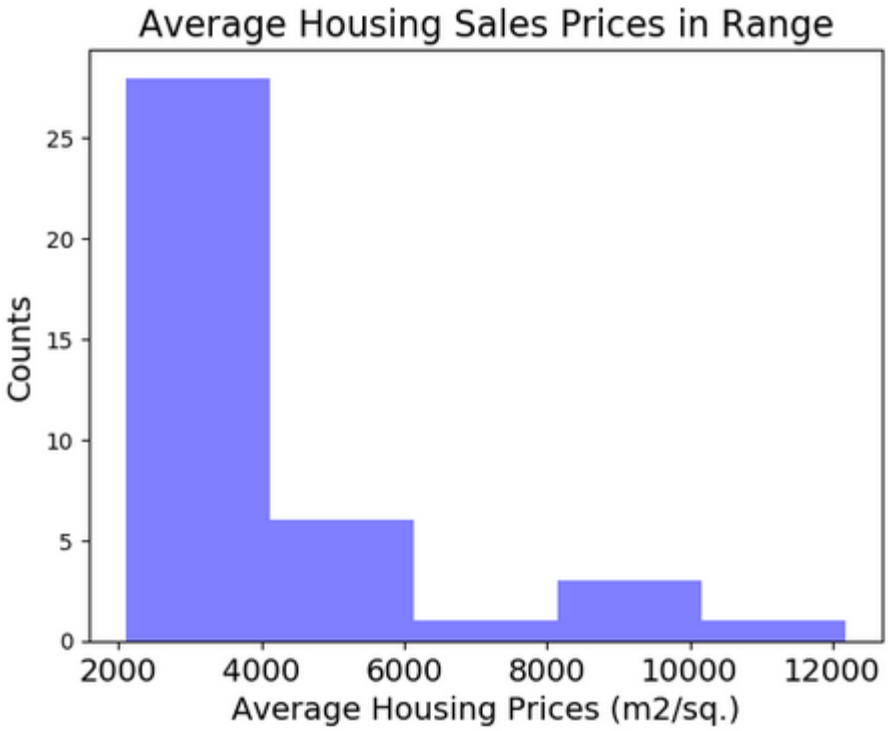
We can also estimate the number of **1st Most Common Venue** in each cluster. Thus, we can create a bar chart which may help us to find proper labels for each cluster.



When we examine above graph we can label each cluster as follows:

- Cluster 0 : “Cafe Venues”
- Cluster 1 : “Multiple Social Venues”
- Cluster 2 : “Accommodation & Intensive Cafe Venues”

We can also examine that what is the frequency of average housing sales prices in different ranges. Thus, histogram can help to visualization:



As it seems in above histogram, we can define the ranges as below:

- 4000 AHP : “Low Level HSP”
- 4000–6000 AHP : “Mid-1 Level HSP”
- 6000–8000 AHP : “Mid-2 Level HSP”
- 8000–10000 AHP : “High-1 Level HSP”
- > 10000 AHP : “High-2 Level HSP”

One of my aim was also show the number of top 3 venues information for each borough on the map. Thus, I grouped each borough by the number of top 3 venues and I combined those informations in **Join** column.

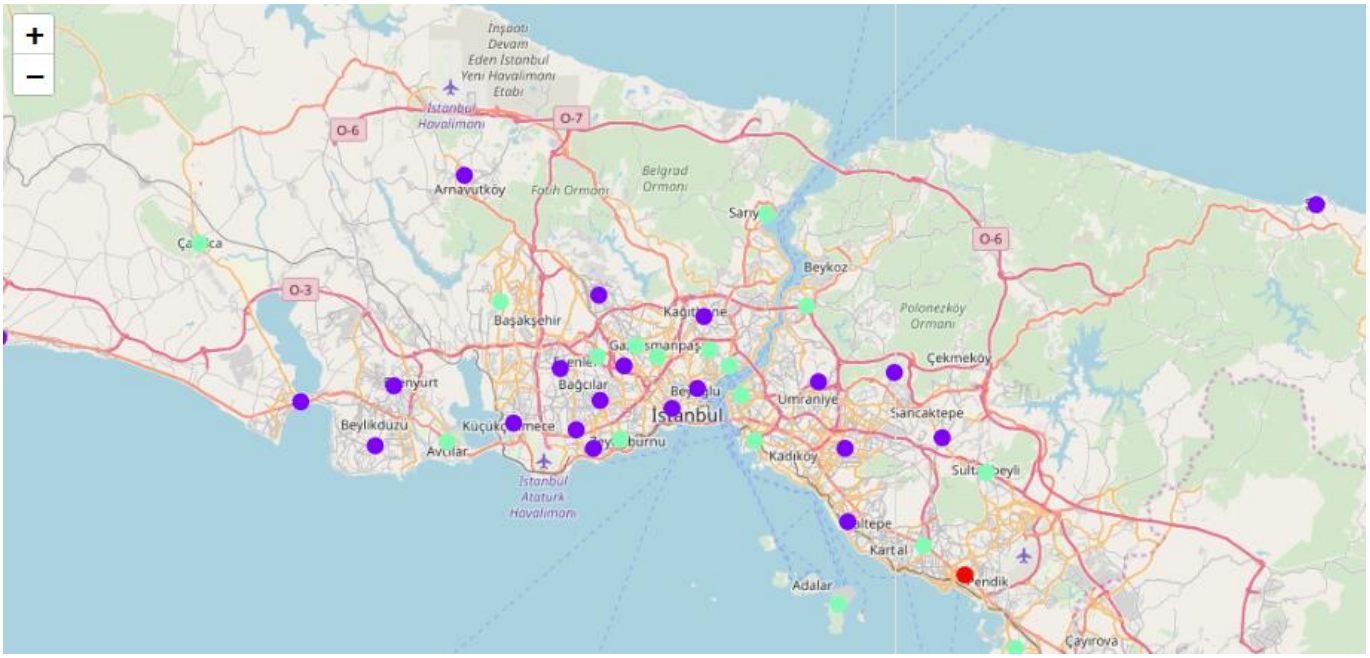
	Borough	Join
0	Adalar	2 Bed & Breakfast, 2 Café, 1 BBQ Joint
1	Arnavutkoy	2 Arcade, 1 Convenience Store, 1 Diner
2	Atasehir	4 Café, 3 Clothing Store, 3 Pool
3	Avcilar	15 Café, 3 Fast Food Restaurant, 3 Restaurant
4	Bagcilar	3 Café, 2 Gym, 2 Turkish Restaurant

C. Results

Let’s merge those new variables with related cluster informations in our main **master table**.

st on ıe	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	Join	Labels	Level_labels
fé	Bed & Breakfast	Garden Center	Mountain	BBQ Joint	Hotel	Road	History Museum	Historic Site	Tea Room	2 Bed & Breakfast, 2 Café, 1 BBQ Joint	Accommodation & Intensive Cafe Venues	Mid-1 Level HSP
de	Pharmacy	Restaurant	Kofte Place	Diner	Leather Goods Store	Convenience Store	Farmers Market	Electronics Store	Entertainment Service	2 Arcade, 1 Convenience Store, 1 Diner	Multiple Social Venues	Low Level HSP
fé	Restaurant	Pool	Spa	Clothing Store	Çöp Şiş Place	Farmers Market	Soccer Stadium	Park	Hotel	4 Café, 3 Clothing Store, 3 Pool	Multiple Social Venues	Mid-1 Level HSP

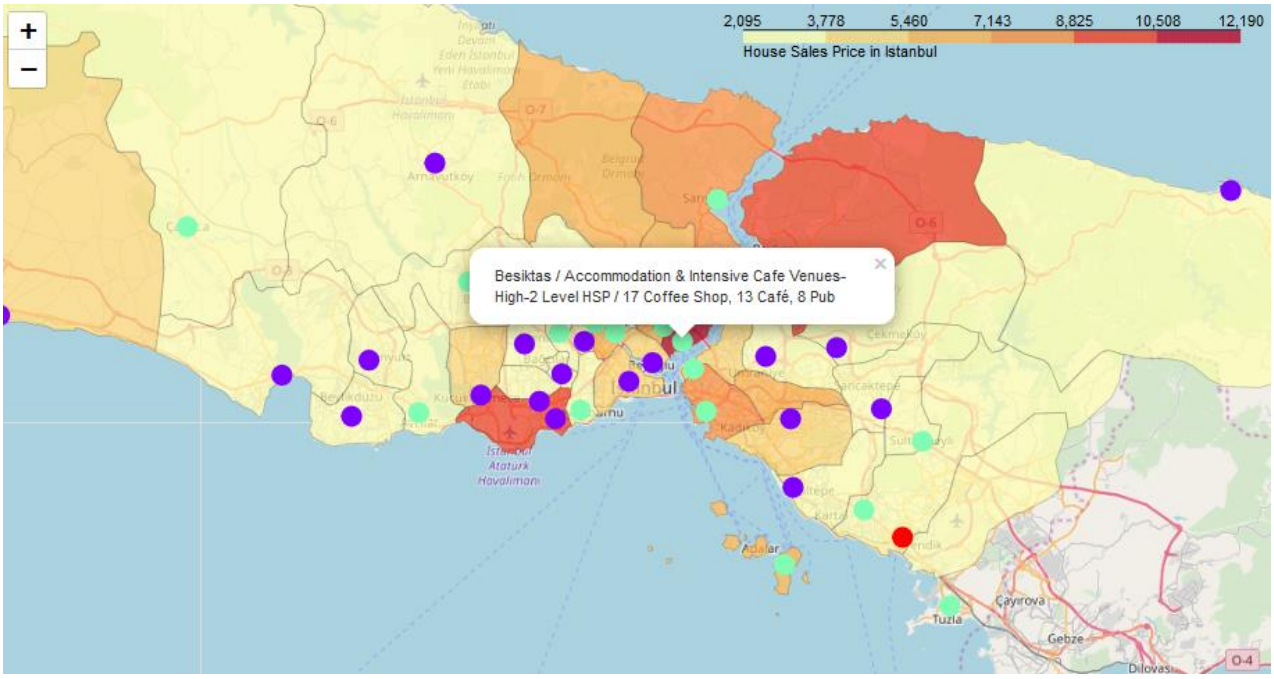
You can now see Join, Labels and Level_labels columns as the last three ones in above table.
You can also see a clustered map boroughs of Istanbul in the below.



In summary section, one of my aim was also visualize the Average Housing Sale Prices for per square meter with choropleth style map. Thus, first I downloaded a json file of Second-level Administrative Divisions of the Turkey from Spatial Data Repository of NYU [2]. I cleaned the json file and pull out only city of Istanbul.

In final section, I created choropleth map which also has the below informations for each borough:

- Borough name,
- Cluster name,
- Housing Sales Price (HSP) Levels,
- Top 3 number of venue



D. Discussion

As I mentioned before, Istanbul is a big city with a high population density in a narrow area. The total number of measurements and population densities of the 39 districts in total can vary. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high quality results for this metropol.

I used the K means algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 3. However, only 39 district coordinates were used. For more detailed and accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled.

I also performed data analysis through this information by adding the coordinates of districts and home sales price averages as static data on GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and clustering information on the Istanbul map. In future studies, web or telephone applications can be carried out to direct investors.

F. Conclusion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

For not only investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

G. References:

- [1] **Istanbul — Wikipedia**
- [2] **Second-level Administrative Divisions of the Turkey**
- [3] **Forsquare API**
- [4] **Housing Sales Prices of Each Borough from "Hurriyet Retail Index for 2018"**
- [5] **Google Map**