# Istanbul Rental Housing Prices and Venues

- Data Analysis

ABSTRACT
This study is a peer-graded assignment, and a part of capstone project called the battle of neighbourhoods

Sami Nejmeh
02/11/2020

# 1. Introduction

## 1.1 Description and Discussion 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 2,902 people per square kilometre. The city is divided into 39 districts in total, 14 of them belong to the Asian side and the other 25 belong to the European side. However, the fact that the districts are squeezed into an area of approximately 72 square kilometres causes the city to have a very intertwined and mixed structure.

The city has been always a destination for students around the world to start or continue their studies, where it has 52 universities out of 175 universities in all Turkey. When we think of the city students, they probably want to choose the regions where rental housing values are low. At the same time, they may want to choose the district according to the social places density.

When we consider the previous problem, we can create a map and information chart where the rental housing index is placed on Istanbul and each district is clustered according to the venues density.

## 1.2 Data Description

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

1) I found a list of districts of Istanbul, which I cleaned and convert to a dataframe. Then I used the `geopy` library to get the districts coordinates where I used it to create choropleth map of Rental Housing Price Index of Istanbul.

2) There are not too many public data 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 Rental Housing Price Averages for each district of Istanbul from housing retail web page.

3) I used the .json file of the Second-level Administrative Divisions of the Turkey from Spatial Data Repository of NYU, which I borrowed from *Sercan Yıldız* GitHub repository, to create a choropleth map of Average Rental Housing Index of Istanbul.

4) I used **Foursquare API** to get the most common venues of given district of Istanbul.

# 2. Methodology

## 2.1 Preparing the Districts Dataset

I started by obtaining a list of districts of Istanbul and their population from Wikipedia [1], which I cleaned and converted to a pandas dataframe as shown on the right:

| | District | Population (2019) | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Adalar | 15238 | 40.876259 | 29.091027 |
| 1 | Arnavutköy | 282488 | 41.184182 | 28.740729 |
| 2 | Ataşehir | 425094 | 40.984749 | 29.106720 |
| 3 | Avcılar | 448882 | 40.980135 | 28.717547 |
| 4 | Bağcılar | 745125 | 41.033899 | 28.857898 |

*Table 1. Districts of Istanbul and their population*

## 2.2 Preparing the Rental Housing Prices Dataset

I collected latest Rental Housing Price Averages for each district of Istanbul from housing retail web page [2] and defined the side of each district.

Then I put the previous information in a table as shown below:

| | District | District Side | Avg-Rent (Lira) | Avg-Rent (USD) |
|---|---|---|---|---|
| 0 | Adalar | Asian | 1874 | 244 |
| 1 | Arnavutkoy | European | 857 | 111 |
| 2 | Atasehir | Asian | 1632 | 212 |
| 3 | Avcilar | European | 1364 | 177 |
| 4 | Bagcilar | European | 1337 | 174 |

*Table 2. Rental Housing Prices and District Sides*

## 2.3 Preparing the Master Dataset

After creating the Districts and Rental Housing Prices dataframes, I merged them into one main dataframe which I used for later analysis.

| | District | District Side | Avg-Rent (Lira) | Avg-Rent (USD) | Population (2019) | Latitude | Longitude |
|---|---|---|---|---|---|---|---|
| 0 | Adalar | Asian | 1874 | 244 | 15238 | 40.876259 | 29.091027 |
| 1 | Arnavutkoy | European | 857 | 111 | 282488 | 41.184182 | 28.740729 |
| 2 | Atasehir | Asian | 1632 | 212 | 425094 | 40.984749 | 29.106720 |
| 3 | Avcilar | European | 1364 | 177 | 448882 | 40.980135 | 28.717547 |
| 4 | Bagcilar | European | 1337 | 174 | 745125 | 41.033899 | 28.857898 |

*Table 3. Master Dataset*

## 2.4 Visualizing the Districts Geographically

I used python `folium` library to visualize geographic details of Istanbul and its districts, and I created a map of Istanbul with districts and their population superimposed on top. I used latitude and longitude values to get the visual as shown below:

*Map 1. Districts of Istanbul*

## 2.5 Analysing the Master dataset

### *2.5.1 Visualizing the Rental Housing Prices*

I visualized the districts and compared the two district sides based on the average rental price in US dollars using `bar charts` as follows:

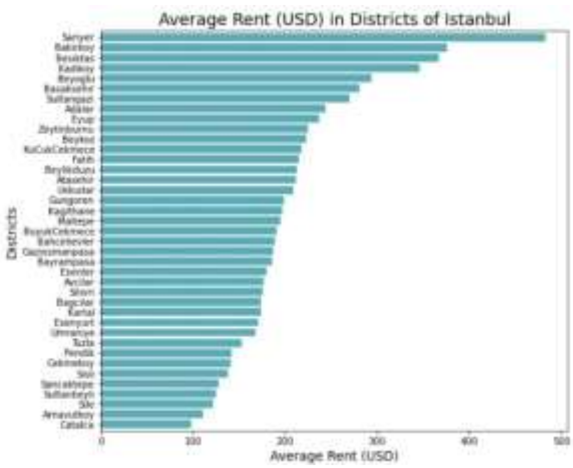



*Chart 1. Average Rent (USD) in Districts of Istanbul*

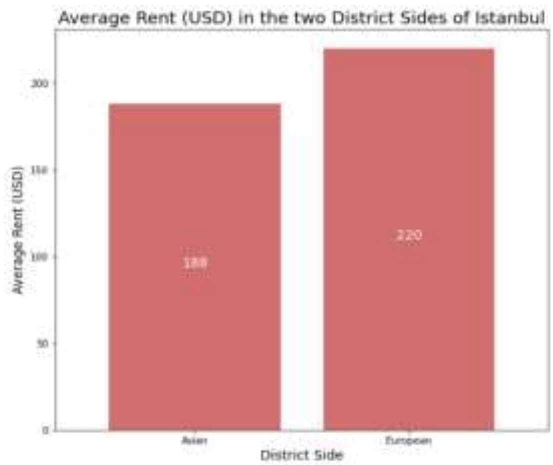



*Chart 2. Average Rent (USD) in the two District Sides of Istanbul*

As we can see from the previous charts that **Sariyer** has the highest average rental housing price equals to 483 USD, and **Catalca** has the lowest average rental housing price equals to 98 USD.

Also, we can see that there is no significant difference between the average rental housing price in the Asian side and the European side. But in general, we can say that the European side is more expensive than the Asian side.

After that I visualized the frequency of rental housing prices in different ranges using `histogram`. Thus, I could create labels which involve pricing features as shown on the right:

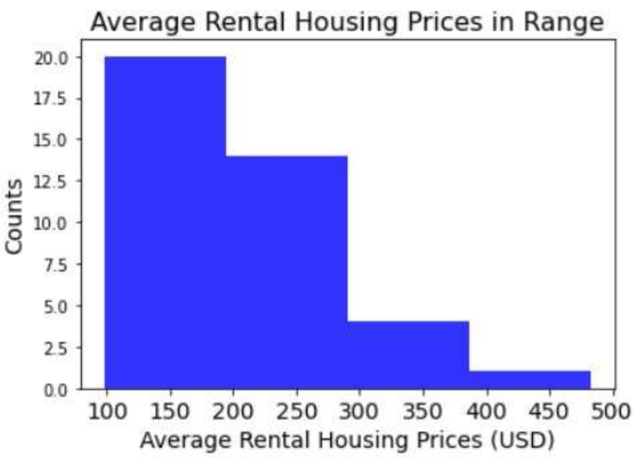



*Chart 3. Average Rental Housing Prices (USD) in Range*

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

- 100-200 ARHP: "Low Level RHP"
- 200-300 ARHP: "Mid-1 Level RHP"
- 300-400 ARHP: "Mid-2 Level RHP"
- '>' 400 ARHP: "High Level RHP"

Then I created a new column "Rental_labels" with those levels as shown below:

| | District | Avg-Rent (USD) | Rental_labels |
|---|---|---|---|
| 0 | Catalca | 98 | Low Level RHP |
| 1 | Arnavutkoy | 111 | Low Level RHP |
| 2 | Sile | 122 | Low Level RHP |
| 3 | Sultanbeyli | 125 | Low Level RHP |
| 4 | Sancaktepe | 128 | Low Level RHP |

*Table 4. Rental Labels of each District*

Finally, using the JSON file from *Sercan Yıldız* GitHub repository, I visualized the districts and their population based on the average rental price in US dollars, but this time using a choropleth map:


*Map 2. Average Rental Housing Prices of Istanbul Districts and their Population*

### *2.5.2 Exploring the Districts*

Next, I started utilizing the `Foursquare API` [3] to explore the districts and segment them.

I designed the limit as 100 venue and the radius 2000 meter for each district from their given latitude and longitude information. In summary **3837** venues were returned by Foursquare. Here is the head of the table of districts name, district latitude and longitude, venues name, venues category, venues latitude and longitude information from Foursquare API.

| | District | District Latitude | District Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Adalar | 40.876259 | 29.091027 | İnönü Evi Müzesi | 40.878251 | 29.093647 | History Museum |
| 1 | Adalar | 40.876259 | 29.091027 | Merit Halki Palace Hotel | 40.878802 | 29.090974 | Hotel |
| 2 | Adalar | 40.876259 | 29.091027 | L'isola Guesthouse | 40.877038 | 29.096136 | Bed & Breakfast |
| 3 | Adalar | 40.876259 | 29.091027 | Heybeliada Su Sporları Kulübü | 40.882365 | 29.089167 | Pool |
| 4 | Adalar | 40.876259 | 29.091027 | Luz Café | 40.877528 | 29.097877 | Café |

*Table 5. Districts and Venues*

In the below chart, we can see that all districts except *Catalca* and *Arnavutkoy* reached the 100 limit of venues.

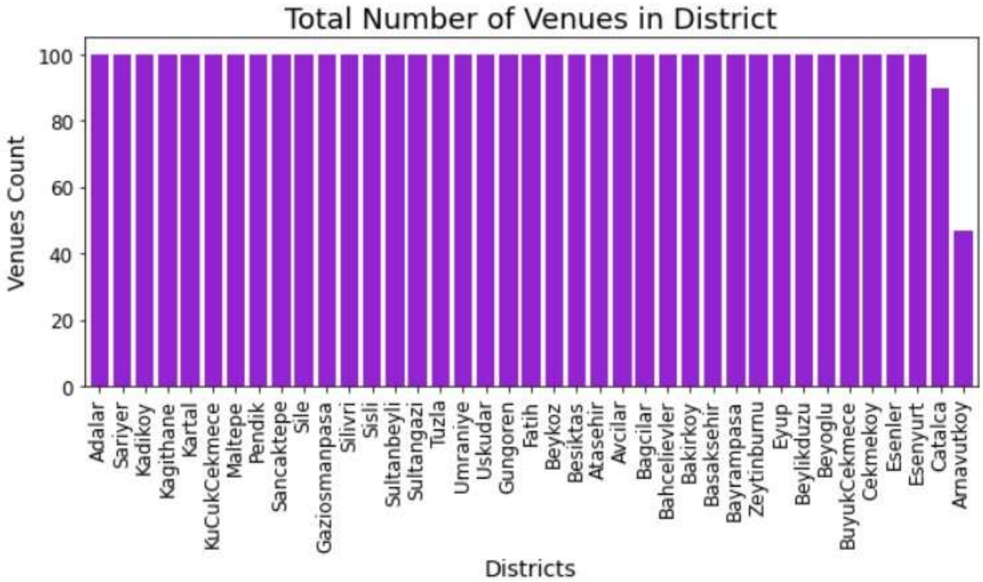



*Chart 4. Number of Venues in each District*

The result does not mean that inquiry run all the possible results in districts. It depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each district. We can increase the possibilities of venues information in each district with more Latitude and Longitude information.

In summary of this graph **304** unique categories were returned by Foursquare.

Then I created a table which shows list of top 10 venue category for each district as shown below:

| | 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 | Seafood Restaurant | Café | Beach | Restaurant | Fast Food Restaurant | Ice Cream Shop | Turkish Restaurant | Trail | Campground | Bar |
| 1 | Arnavutkoy | Café | Restaurant | Turkish Restaurant | Kofte Place | Dessert Shop | Arcade | Fish & Chips Shop | Gym | Steakhouse | Breakfast Spot |
| 2 | Atasehir | Restaurant | Coffee Shop | Seafood Restaurant | Steakhouse | Kebab Restaurant | Hotel | Gym / Fitness Center | Park | Café | Shopping Mall |
| 3 | Avcilar | Café | Dessert Shop | Coffee Shop | Restaurant | Bar | Burger Joint | Gym / Fitness Center | Pizza Place | Pub | Mobile Phone Shop |
| 4 | Bagcilar | Café | Dessert Shop | Steakhouse | Gym | Turkish Restaurant | Kebab Restaurant | Coffee Shop | Hookah Bar | Gym / Fitness Center | Pizza Place |

*Table 6. Top 10 Venues Category*

We have some common venue categories in districts. For this reason, I used unsupervised learning `k-means` algorithm to cluster the districts. K-means algorithm is one of the most common cluster method of unsupervised learning.

First, I run k-means to cluster the districts into **3** clusters because when I analyse the k-Means with elbow method it ensured me that the optimum k of the k-means equals 3 degree.

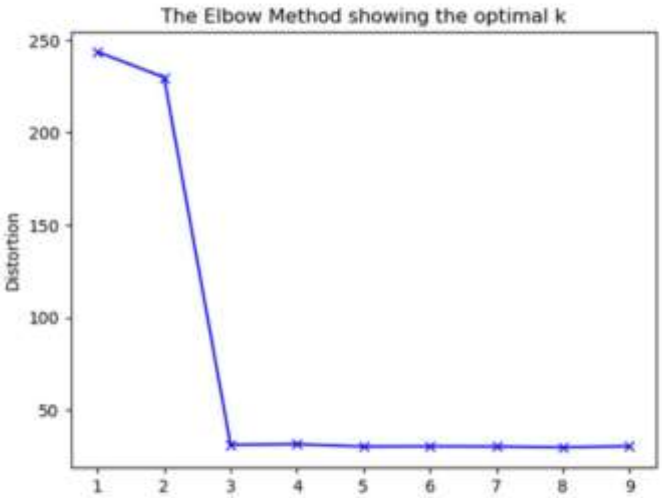



*Chart 5. Elbow Method*

Here is my merged table with cluster labels for each district:

| | District | District Side | Avg-Rent (Lira) | Avg-Rent (USD) | Population (2019) | 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 | Asian | 1874 | 244 | 15238 | 40.876259 | 29.091027 | 0 | Seafood Restaurant | Café | Beach | Restaurant | Fast Food Restaurant | Ice Cream Shop | Turkish Restaurant | Trail | Campground | Bar |
| 1 | Arnavutkoy | European | 857 | 111 | 262488 | 41.184182 | 28.740729 | 2 | Café | Restaurant | Turkish Restaurant | Kofte Place | Dessert Shop | Arcade | Fish & Chips Shop | Gym | Steakhouse | Breakfast Spot |
| 2 | Atasehir | Asian | 1632 | 212 | 425094 | 40.984749 | 29.106720 | 1 | Restaurant | Coffee Shop | Seafood Restaurant | Steakhouse | Kebab Restaurant | Hotel | Gym / Fitness Center | Park | Café | Shopping Mall |
| 3 | Avcilar | European | 1364 | 177 | 448882 | 40.980135 | 28.717547 | 2 | Café | Dessert Shop | Coffee Shop | Restaurant | Bar | Burger Joint | Gym / Fitness Center | Pizza Place | Pub | Mobile Phone Shop |
| 4 | Bagcilar | European | 1337 | 174 | 745125 | 41.033899 | 28.857898 | 1 | Café | Dessert Shop | Steakhouse | Gym | Turkish Restaurant | Kebab Restaurant | Coffee Shop | Hookah Bar | Gym / Fitness Center | Pizza Place |

*Table 7. Merged Table with Clusters Labels*

We can also estimate the number of **1st Most Common Venue** in each cluster. Thus, I created a `bar chart` which may help us to find proper label names for each cluster.

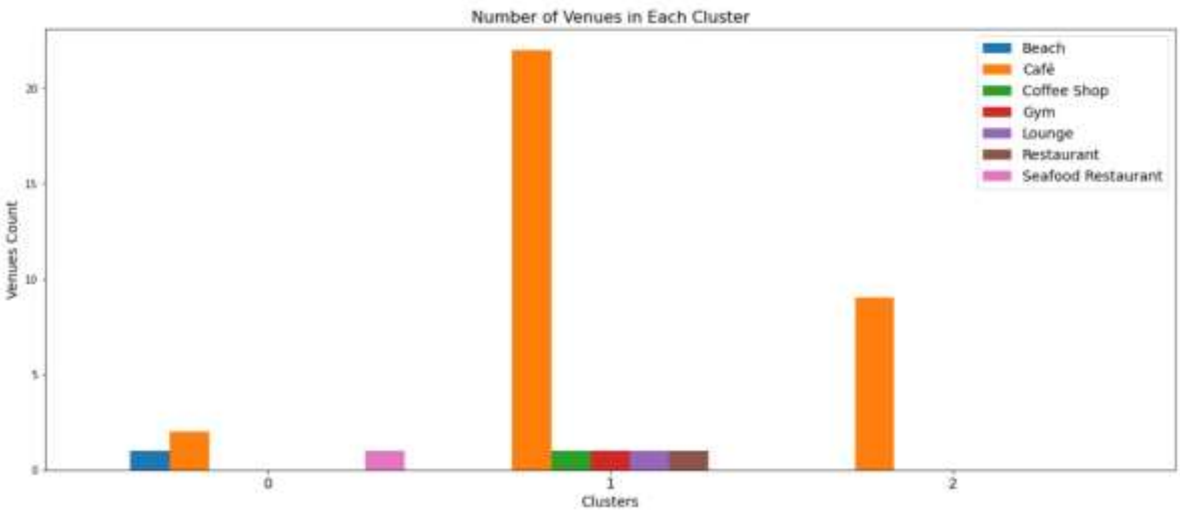



*Chart 6. Number of Venues in each Cluster*

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

- Cluster 0: "Coastal Venues"
- Cluster 1: "Multiple Social Venues"
- Cluster 2: "Cafe Venues"

Then I assigned those new labels to existing label of clusters:

| | Clusters | Labels |
|---|---|---|
| **0** | 0 | Coastal Venues |
| **1** | 1 | Multiple Social Venues |
| **2** | 2 | Cafe Venues |

*Table 8. Labels of Clusters*

One of the aims of this project was also to show the number of the top 3 venues information for each district on the map. Thus, I grouped each district by the number of top 3 venues, and I combined those information in *Venues* column as shown below:

| | District | Venues |
|---|---|---|
| **0** | Adalar | 19 Seafood Restaurant, 15 Café, 13 Beach |
| **1** | Arnavutkoy | 11 Café, 4 Restaurant, 4 Turkish Restaurant |
| **2** | Atasehir | 10 Restaurant, 6 Coffee Shop, 5 Seafood Restau... |
| **3** | Avcilar | 30 Café, 6 Dessert Shop, 5 Coffee Shop |
| **4** | Bagcilar | 18 Café, 7 Dessert Shop, 6 Gym |

*Table 9. Number of top 3 Venues for each District*

# 3. Results

## 3.1 The Master dataframe with Results

Let's merge those new variables with related cluster information in our main *Master Table*.

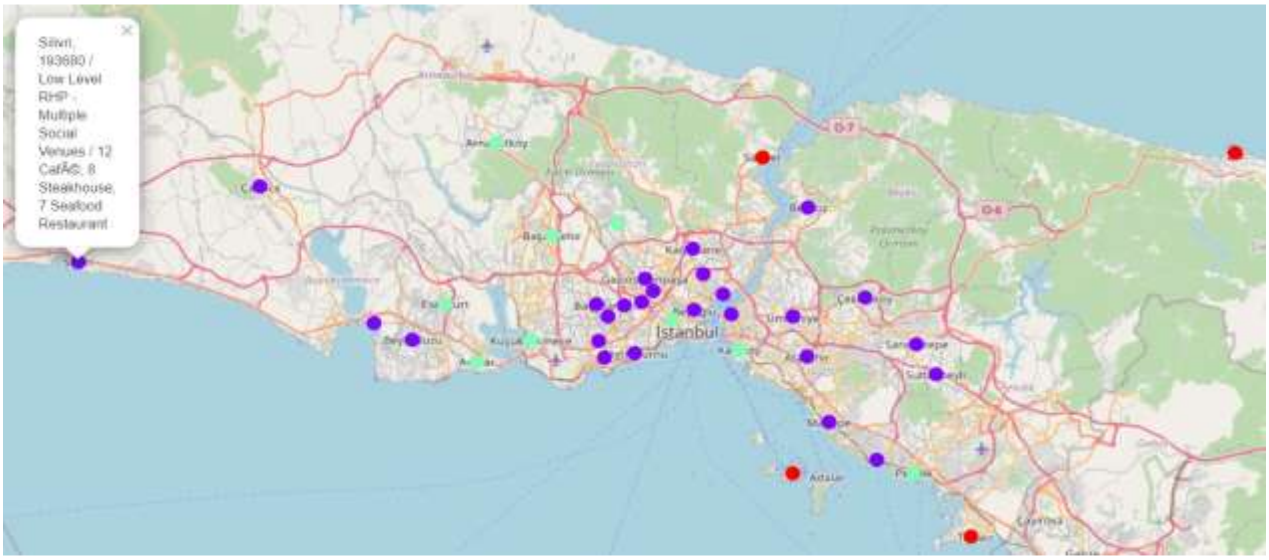
| Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | ... | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue | Venues | Labels | Rental_labels |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Seafood Restaurant | Café | ... | Restaurant | Fast Food Restaurant | Ice Cream Shop | Turkish Restaurant | Trail | Campground | Bar | 19 Seafood Restaurant, 15 Café, 13 Beach | Coastal Venues | Mid-1 Level RHP |
| 2 | Café | Restaurant | ... | Kofte Place | Dessert Shop | Arcade | Fish & Chips Shop | Gym | Steakhouse | Breakfast Spot | 11 Café, 4 Restaurant, 4 Turkish Restaurant | Cafe Venues | Low Level RHP |
| 1 | Restaurant | Coffee Shop | ... | Steakhouse | Kebab Restaurant | Hotel | Gym / Fitness Center | Park | Café | Shopping Mall | 10 Restaurant, 6 Coffee Shop, 5 Seafood Restau... | Multiple Social Venues | Mid-1 Level RHP |
| 2 | Café | Dessert Shop | ... | Restaurant | Bar | Burger Joint | Gym / Fitness Center | Pizza Place | Pub | Mobile Phone Shop | 30 Café, 6 Dessert Shop, 5 Coffee Shop | Cafe Venues | Low Level RHP |

*Table 10. Master Table*

You can now see Venues, Labels and Rental_labels columns as the last three ones in the above table.

### 3.2 Map the Cluster Results

Finally, I visualized the resulting clusters as shown below:



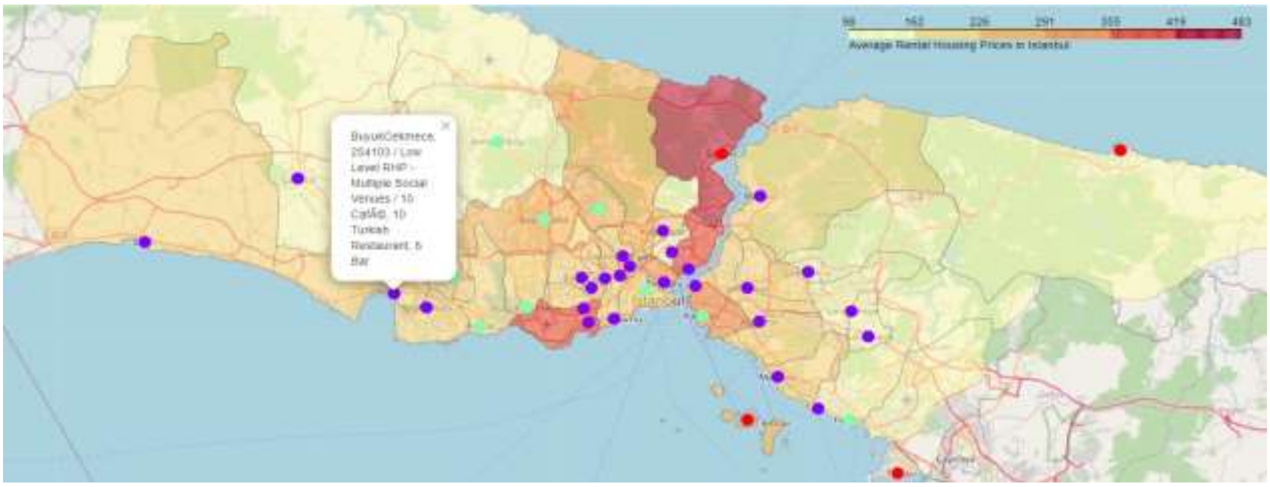

*Map 3. Clustered Districts Map of Istanbul*

### 3.3 Map the Rental Housing Prices with Cluster Results

One of my aim was also to visualize the Average Rental Housing Prices using a choropleth map. Thus, first I borrowed a json file of Second-level Administrative Divisions of the Turkey [4] from *Sercan Yıldız* GitHub repository.

Then In the final section, I created a `choropleth map` which has the below information for each district:

- District name
- Population
- Cluster name,
- Rental Housing Price (RHP) level
- Top 3 most common venues




*Map 4. Choropleth Map of Istanbul with Final Data*

# 3. 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, quite 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 metropole

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 neighbourhood or street can also be drilled.

I also performed data analysis through this information by using the coordinates of districts from a Wikipedia page and rental home price averages from a housing retail web page.

I ended the study by visualizing the data and clustering information on the Istanbul map.

# 4. Conclusion

As a result, students are turning to new cities to start or continue their studies. For this reason, students can achieve better outcomes through their access to the platforms where such information is provided.

Not only for students but also city managers can manage the city more regularly by using similar data analysis types or platforms.

# 5. References


[1] List of Districts of Istanbul

[2] Rental Housing Prices of Istanbul

[3] Foursquare API

[4] Second-level Administrative Divisions of the Turkey