Prediction Potential Areas For Investment By Analyzing Venues Data And Sales Price Of Houses Sold Data

Onur Çakı
02.03.2020

1. Introduction: Business Problem

Turkish metropolitan city Izmir has population of 4.36 million. It makes Izmir the third most populous city in Turkey and the fourth in Mediterranean Sea. Izmir consist of 11 urban districts – namely Balcova, Bayrakli, Bornova, Buca, Cigli, Gaziemir, Guzelbahce, Karabaglar, Karsiyaka, Konak, and Narlidere – and additionally 19 rural districts – namely Aliaga, Bayindir, Bergama, Beydag, Cesme, Dikili, Foca, Karaburun, Kemalpasa, Kinik, Kiraz, Menderes, Menemen, Odemis, Seferihisar, Selcuk, Tire, Torbalı, Urla –. [1]

Located on the west coast of Turkey, Izmir includes two ports – namely Aliaga Port and Izmir Alsancak Port – that are the primary ports of Turkey for exports. Particularly Izmir Alsancak Port is the biggest container port of Turkey with a capacity of 10 million tons. Thanks to capacity of exports and imports, there are vast industrial zones from various branches of industry, especially high-tech industry. Apart from industry, Izmir is one of the centers of modern agriculture and animal husbandry because of its fertile lands, favorable climatic conditions, rich water resources, and biodiversity. Today, Izmir is the second largest commercial hub in Turkey. [2] According to Brookings Institution Global Metro Monitor, Izmir is the world's second fastest growing metropolitan economy. [3] Moreover, hundreds of thousands of tourists come to Izmir every year to vacation in coastal towns such as Cesme, Seferihisar, Urla, and Foca, and to visit ancient heritages and religious areas in Selcuk and Bergama. [4]

All those factors mentioned above make Izmir attractive for investment. As a result of this, the number of Turkish citizens and foreign people who immigrate to Izmir rises day by day. [2] This situation induces sharply increase property prices in Izmir. Izmir has become the most profitable city of Turkey in which house rents increase most recently. [5] Prediction of new areas that may have the potential to appreciate in value is vital for investors or for those who think to buy a new house. I aimed to meet this requirement in this project. Boroughs of Izmir were clustered based on the venues that they have. Then, the clusters were compared with sold house prices to reveal patterns behind house prices.

2. Data

In project four data sources will be used:

- An open-source API, which is publicly available in GitHub, is used to retrieve longitude and latitude coordinates of the center of boroughs. [6] The API includes coordinates of all cities and boroughs in Turkey. The information of Izmir is parsed through URL query.
- I mentioned above that similar boroughs were clustered by exploiting venues that they have. By using the center coordinates, the most common venues with their features in boroughs are retrieved from Foursquare API through URL query. [7]
- The average sales price of houses sold in Izmir are taken from Endeksa that is one of the most popular property price index companies in Turkey. [8]
- The geolocation data of polygons that draw the boundary of boroughs of Izmir are retrieved from Second-level Administrative Divisions of the Turkey from Spatial Data Repository of NYU. [9] The data contains the geolocation of whole cities and boroughs in Turkey. The data of Izmir was parsed from all geojson files. It will be used to create choropleth map.

3. Methodology

First, geolocations of the borough of Izmir -i.e. longitude and latitude coordinates- are retrieved from a open-source API [6] by URL query that is designed for Izmir specifically. The map which is drawn by using this data can be shown in Figure 1.

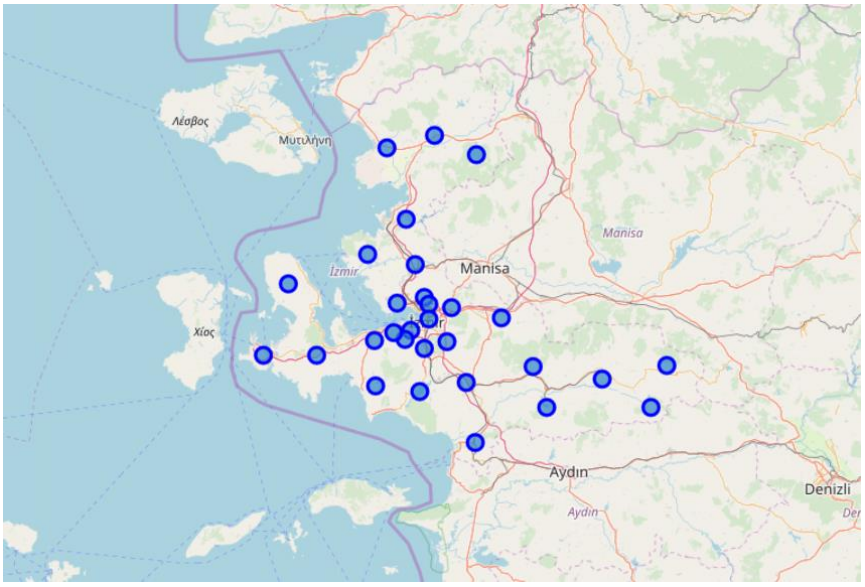

Figure 1 Boroughs of Izmir

The idea behind of this project to cluster boroughs based on venues that they have. Then, compare clusters with the average sales price of houses sold. So, we obtain the venues for each borough from Foursquare API. [7] Besides URL query, Foursquare API requires radius and limit parameters as inputs. Radius is used to determines the range in which the venues are retrieved. Limit is the maximum number of returned venues. Radius is chosen as 10 km because of many wide rural areas. Because we need to many distances to capture sufficient venues. The limit of the number of venues returned by Foursquare API is decided as 200 so as to avoid overlapping of urban boroughs.

Actually, we describe borough by using category of venues. So that, I applied one hot encoding over categories. Then, I grouped rows by boroughs and by taking the mean of the frequency of occurrence of each category. Those numerical values will be used as input of k-means algorithms.

After the obtain frequency of categories of venues, we need remove some features that are not discriminative. I dropped Café, Breakfast Spot, Turkish Restaurant, Restaurant, and Turkish restaurants because they are not discriminative features. I came back this step again and again after clustering.
Apart from numerical feature values which describe each borough, k-means algorithm also requires cluster number -k. I used to Elbow to choose optimum cluster number k. However, it just gives a recommendation not an exact result. Therefore, one who wants to tune k-value can return to this step again and again after the obtained clusters.

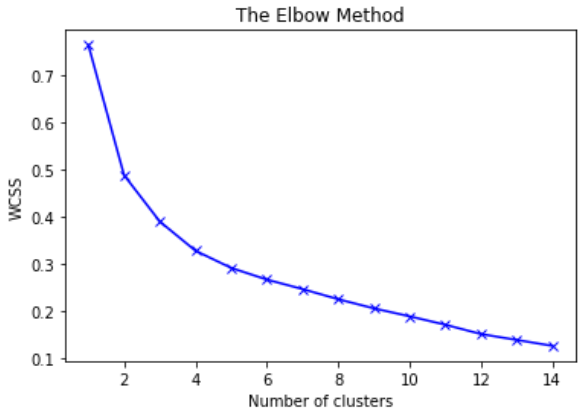

Figure 2

I chose number of cluster as 5 because the derivative of curve starts to decrease at this point. Besides that, the results obtained after clustering was sensible. Python library which is named sckit-learn is used to k-mean algorithm. As result of the clustering, I obtained 5 clusters with labels 0,1,2,3, and 4.

All methodology is summarized in Figure 3.

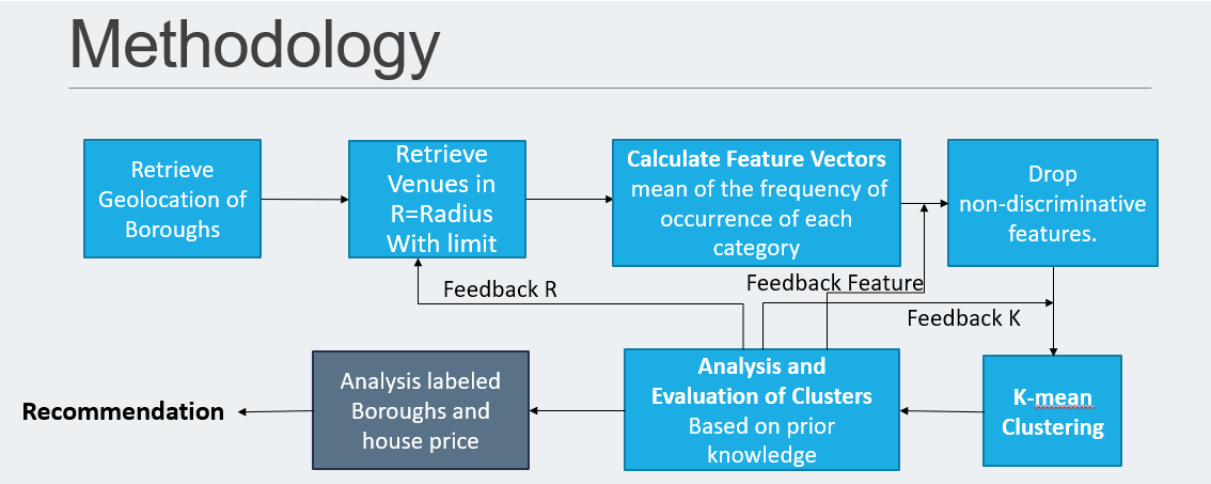

Figure 3

After the clustering, I combined labeled boroughs data with the data which contains average sales price of houses sold. I visualize it by using choropleth in folium python library as shown as figure 4.

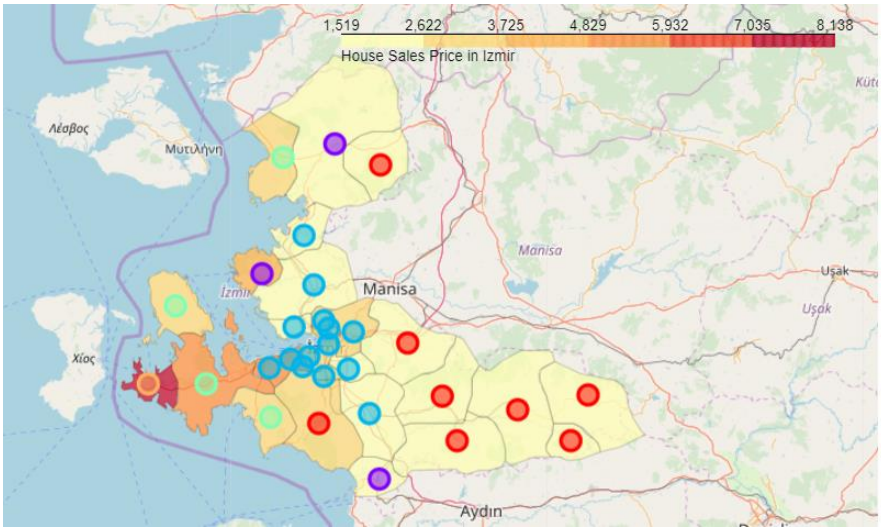

Figure 4

4. Analysis

Before the start analysis correlation between price and clusters, we must look in more detailed venues of clusters to understand characteristics of clusters.

Cluster 0 is indicated with a red label on map. When you look at the map, you can easily see that those boroughs with red labels are countryside rural areas. Also, most ten common venues of those borough support this result. The list of the cluster is dominated by Parks, Forests, Mountains, Farms.

	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
3	Kiraz	Park	Seafood Restaurant	Mountain	Boarding House	Snack Place	Beach Bar	Fish & Chips Shop	Fast Food Restaurant	Pharmacy	Farmers Market
4	Kemalpasa	BBQ Joint	Park	Bakery	Forest	Pizza Place	Dessert Shop	Mountain	Kebab Restaurant	Steakhouse	Shopping Mall
5	Bayindir	Diner	Mountain	Tea Room	Flower Shop	Botanical Garden	Food Court	Garden	Turkish Coffeehouse	Pizza Place	Dive Bar
7	Odemis	Bakery	Dessert Shop	Kebab Restaurant	Sandwich Place	Gym / Fitness Center	Steakhouse	Clothing Store	Middle Eastern Restaurant	Tea Room	History Museum
11	Kinik	Botanical Garden	Plaza	Buffet	Forest	Bar	Pizza Place	Pide Place	Beach	Comfort Food Restaurant	Beer Garden
19	Menderes	Steakhouse	Arcade	Bakery	BBQ Joint	Forest	Trail	Athletics & Sports	Buffet	Farm	Diner
22	Beydag	Mountain	Steakhouse	Park	Convenience Store	Pharmacy	Tea Room	Farm	Lake	Turkish Coffeehouse	Plaza
23	Tire	Dessert Shop	Pizza Place	Snack Place	Historic Site	Clothing Store	Pub	Mountain	Sandwich Place	Arcade	Gym / Fitness Center

Figure 5 -> Cluster 0

Cluster 1 is indicated with a purple label on map. The venue list of the borough belonging to this cluster is dominated by Hotels and Historic sites. As I mentioned in the introduction, many tourists come to Izmir to visit Selcuk and Bergama every year. There are Ephesus Ancient City and House of Virgin Mary in Selcuk, and there is Pergamon Ancient City in Bergama. Foca is also appealing to tourists with its beaches and antic heritages.

	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
13	Selcuk	Hotel	Historic Site	Coffee Shop	Bed & Breakfast	Steakhouse	Wine Shop	Fast Food Restaurant	Bar	Gym	Park
14	Bergama	Historic Site	Hotel	Bar	Lounge	Arcade	Turkish Home Cooking Restaurant	History Museum	Dessert Shop	Snack Place	Soccer Field
25	Foca	Hotel	Seafood Restaurant	Bed & Breakfast	Steakhouse	Bar	Beach	Historic Site	Pide Place	Harbor / Marina	Resort

Figure 6 -> Cluster 1

Cluster 2 is indicated with a blue label on map. This cluster consists of urban areas of Izmir. The boroughs in these clusters include a great number of coffee shops, sports, and art facilities. Industry zones of Izmir are built in those boroughs and the workforce of Izmir's Industry lives in those areas.

	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
1	Menemen	Coffee Shop	Gym	Gym / Fitness Center	Steakhouse	Kofte Place	Kokoreç Restaurant	Tea Room	Accessories Store	Bakery	Food Truck
2	Cigli	Gym / Fitness Center	Coffee Shop	Waterfront	Clothing Store	Scenic Lookout	Supermarket	Shopping Mall	Bakery	Bookstore	Basketball Stadium
6	Buca	Gym / Fitness Center	Bakery	Coffee Shop	Gym	Hookah Bar	Pub	Fast Food Restaurant	Kebab Restaurant	Turkish Home Cooking Restaurant	Bar
9	Karsiyaka	Coffee Shop	Gym / Fitness Center	Bookstore	Waterfront	Bar	Bakery	Scenic Lookout	Meyhane	Manti Place	Motorcycle Shop
10	Aliaga	Seafood Restaurant	Beach	Bakery	Fast Food Restaurant	Gym / Fitness Center	Beer Garden	Hotel	Waterfront	Steakhouse	Burger Joint
15	Bornova	Coffee Shop	Dessert Shop	Gym / Fitness Center	Clothing Store	Pastry Shop	Brewery	Bar	Salon / Barbershop	Bakery	Burger Joint
16	Torbali	Coffee Shop	Steakhouse	Gym	Diner	Hookah Bar	Beer Garden	Bar	Clothing Store	Plaza	Art Gallery
17	Balcova	Coffee Shop	Theater	Hotel	Waterfront	Art Gallery	Seafood Restaurant	Chocolate Shop	Historic Site	Pastry Shop	Bakery
18	Gaziemir	Gym / Fitness Center	Coffee Shop	Hotel	Arcade	Bar	Kebab Restaurant	Bakery	Dessert Shop	Sandwich Place	Baby Store
21	Bayrakli	Coffee Shop	Meyhane	Gym / Fitness Center	Pastry Shop	Art Gallery	Chocolate Shop	Seafood Restaurant	Dessert Shop	Dance Studio	Performing Arts Venue
24	Guzelbahce	Seafood Restaurant	Coffee Shop	Bakery	Harbor / Marina	Hookah Bar	Beach	Burger Joint	Market	Pool	Supermarket
27	Konak	Theater	Dance Studio	Meyhane	Art Gallery	Pizza Place	Pastry Shop	Coffee Shop	Bakery	Turkish Home Cooking Restaurant	Chocolate Shop
28	Karabaglar	Coffee Shop	Waterfront	Seafood Restaurant	Gym	Bakery	Cosmetics Shop	Gym / Fitness Center	Hotel	Pizza Place	Concert Hall
29	Narlidere	Seafood Restaurant	Waterfront	Gym	Gym / Fitness Center	Coffee Shop	Pizza Place	Steakhouse	Scenic Lookout	Art Gallery	Hotel

Figure 7 -> Cluster 2

Cluster 3 is indicated with a green label on the map. There are popular coastal holiday towns of Izmir in this cluster. Those boroughs which are intertwined with nature and sea includes many number of bars, beaches, hotels as well as farms, mountains in their lists.

	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	Seferihisar	Beach	Bed & Breakfast	Seafood Restaurant	Hotel	Historic Site	Coffee Shop	Grocery Store	Furniture / Home Store	Lounge	Turkish Home Cooking Restaurant
8	Karaburun	Beach	Mountain	Other Great Outdoors	Scenic Lookout	Farm	Tea Room	Harbor / Marina	Bakery	Coffee Shop	Plaza
20	Dikili	Beach	Bar	Diner	Seafood Restaurant	Coffee Shop	Fast Food Restaurant	Pide Place	Music Venue	Farm	Plaza
26	Urla	Beach	Surf Spot	Hotel	Plaza	Farm	Campground	Tea Room	Pool	Scenic Lookout	Bar

Figure 8 -> Cluster 3

According to one of Turkey's most famous estate website, [10] Cesme is the fifth most expensive district of Turkey. There are villas of the richest men of Turkey and ultra-luxury hotels. So, it's the outlier instance in our data. It's not surprising that cluster 4 contains only Cesme.

	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
12	Cesme	Hotel	Beach	Bed & Breakfast	Lounge	Bar	Motel	Surf Spot	Seafood Restaurant	Farmers Market	Bakery

Figure 9 -> Cluster 4

5. Results and Discussion

My project aims to obtain a correlation between borough clusters and house selling prices. And, my ultimate goal is to predict potential boroughs to invest by using this correlation. Before the discuss our results, I analyzed the correlation between clusters and prices by visualizing to reveals results more clear. To do so, I divided the price into 5 levels first. I should note that Cesme is an outlier instance due to its extremely high average house sold price. So that, I dropped Cesme first, and then add it with level-5. I merge cluster labels, price and price level into one data frame as shown as Figure 10. I consider cluster labels and price levels as independent categorical attributes and consider prices as dependent value. While cluster labels are drawing as one categorical attribute by scattering against dependent continuous price value, price level that is another categorical attribute is combined as a color with this plot. The plot can be shown in figure 11.

	Cluster Labels	Borough	Price	Price Levels
2	0	Bayindir	1795	level-1
5	0	Beydag	1519	level-1
17	0	Kemalpasa	2153	level-1
18	0	Kinik	1568	level-1
19	0	Kiraz	1707	level-1
21	0	Menderes	3021	level-2
24	0	Odemis	1666	level-1
27	0	Tire	2091	level-1

	Cluster Labels	Borough	Price	Price Levels
4	1	Bergama	1821	level-1
11	1	Foca	3893	level-3
26	1	Selcuk	2303	level-1

	Cluster Labels	Borough	Price	Price Levels
0	2	Aliaga	2311	level-1
1	2	Balcova	3383	level-3
3	2	Bayrakli	3445	level-3
6	2	Bornova	3564	level-3
7	2	Buca	2560	level-2
9	2	Cigli	2520	level-2
12	2	Gaziemir	3357	level-2
13	2	Guzelbahce	5222	level-4
14	2	Karabaglar	2400	level-1
16	2	Karsiyaka	3552	level-3
20	2	Konak	3271	level-2
22	2	Menemen	2215	level-1
23	2	Narlidere	5130	level-4
28	2	Torbali	1760	level-1

	Cluster Labels	Borough	Price	Price Levels
10	3	Dikili	2995	level-2
15	3	Karaburun	3714	level-3
25	3	Seferihisar	2954	level-2
29	3	Urla	5033	level-4

	Cluster Labels	Borough	Price	Price Levels
8	4	Cesme	8138	level-5

Figure 10

Correlation between class labels that are constructed by k-means algorithm and price level can be seen easily above the plot. It's obvious that investment in cluster 0 which consists of rural boroughs won't be a good idea. Selcuk and Bergama in Cluster 1 can be a great opportunity for investment. Although they are very similar to Foca, the prices in these boroughs are far cheaper than Foca. I should emphasize that those two boroughs have beach although not as popular as Foca. Those beaches may be appreciated in value in the near future. Cluster-2 require more discriminative feature because of their complex structure. But if we remove expensive and cheap outliers, we can see that Aliaga, Cigli, Buca may be a profitable investment for the future. Because they are cheaper than other boroughs in the cluster. Maybe Cluster 3 contains the boroughs which have the most potential. Urla is one of the most valuable boroughs not only of Izmir but in Turkey. However, Karaburun, Dikili, and Seferhisar are affordable despite their similarity to Urla. They can be considered for real estate projects.

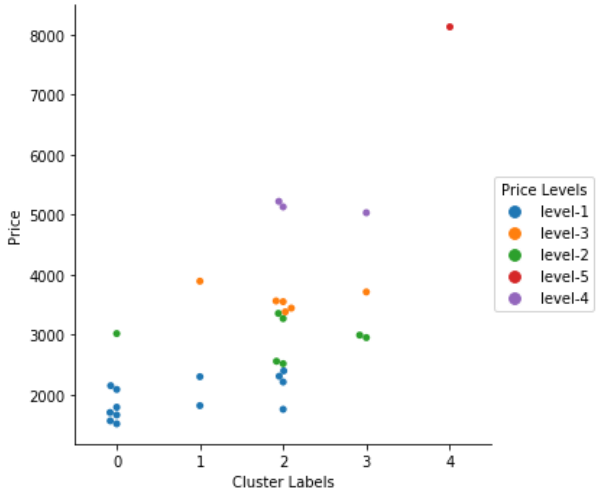

Figure 11

6. Conclusion

In this project, I demonstrated a method to predict the best areas for investors or for those who think to buy a new home. In this method, boroughs were clustered by k-means algorithm based on venues that they have. Then, cluster labels were compared to house-prices. By analyzing both data, one can recommend the areas may have the potential to invest. This method can be developed by contributing other data sets into venues. Because many factors can affect boroughs similarity. Especially, urban boroughs need more features to be distinguished. Thereby, more accurate results can be obtained.

References

[1] “İzmir.” Wikipedia. Wikimedia Foundation, February 28, 2020. https://en.wikipedia.org/wiki/İzmir.
[2] “Economy Of İzmir.” Izto. Accessed March 1, 2020. http://www.izto.org.tr/en/izmir-ekonomisi.
[3] “Tourism of İzmir.” Izto. Accessed March 1, 2020. http://www.izto.org.tr/en/izmir-turizmi.
[4] Trujillo, Jesus Leal, and Joseph Parilla. “The World's 10 Fastest Growing Metropolitan Areas.” Brookings. Brookings, August 8, 2019. https://www.brookings.edu/blog/the-avenue/2015/02/10/the-worlds-10-fastest-growing-metropolitan-areas/.
[5] “REIDIN Emlak Endeks 2018 Temmuz Ayı Sonuçları.” REIDIN, September 17, 2018. https://blog.reidin.com/reidin-emlak-endeks-2018-temmuz-ayi-sonuclari/.
[6] https://github.com/melihkorkmaz/il-ilce-mahalle-geolocation-rest-api
[7] https://developer.foursquare.com/
[8] https://www.endeksa.com/en/analiz/izmir/endeks/for-sale/house
[9] Second-level Administrative Divisions, 2. (2020). Second-level Administrative Divisions, Turkey, 2015 - NYU Spatial Data Repository. [online] Geo.nyu.edu. https://geo.nyu.edu/catalog/stanford-nj696zj1674
[10] Özdalgıç, Aytaç. “Türkiye'nin En Pahalı 50 Ilçesi.” Tapusor.com, October 19, 2018. https://tapusor.com/blog/turkiyenin-en-pahali-50-ilcesi/.