Recommending Rental Houses in Izmir, Turkey

İhsan Köse

09.05.2020

1. Introduction

1.1. Background

Finding a suitable home has never been easy for any city in the World. Same applies to Izmir, Turkey. Furthermore, Izmir suffers from high real estate inflation as the city gets lots of immigration from big and small cities. Izmir has lots of house options for any budget or other preferences. It's the 3rd biggest city in Turkey, therefore, offers many boroughs. Each borough has its own characteristics, such as venue diversity, and housing choices.

1.2. Problem

It is a big challenge to find the correct house in Izmir. Parameters such as price, building age, house size, room numbers, price per square meter, borough, all has great effect over the selection. When renting a house considering all these parameters manually can be overwhelming and confusing.

1.3. Interest

Anyone who is moving to Izmir, or not happy with the current house can use this solution to find the correct house. Moreover, real estate companies can use solution to recommend houses to their customers.

2. Data Acquisition and Cleaning

2.1. Data Sources

Rental houses data is obtained from real estate website: Hürriyet Emlak. This is the second biggest website after sahibinden.com. But as that website didn't allow scraping, data is scraped from Hürriyet Emlak. The links that forwarding to ad webpage also scraped.

Borough names are also extracted from the same webpage as the rental houses.

IMAGE 1. REAL ESTATE WEBSITE VISUAL

Borough latitude and longitude values are obtained by Nominatim API.

	Borough	Latitude	Longitude
0	Ödemiş	38.231341	27.975181
1	Menderes	38.149147	27.106591
2	Balçova	38.380478	27.055728
3	Konak	38.410958	27.129453
4	Karabağlar	38.347205	27.041274

TABLE 1. BOROUGH COORDINATES

The venues in boroughs are obtained from Foursquare API. This data is used for clustering the boroughs and finding similar boroughs by venue categories.

2.2. Data Cleaning

Fortunately, to publish ads on the website, parameters such as *Price*, *Room#*, *Size* and *Borough* need to be filled completely. Only *Age* is not mandatory. But based on experiments, ads without *Age* information is around %0,1. So this is a low value and I choose to fill that missing age information with the mean of the other houses' ages.

But there are some people publishing house for sale ads on the rental houses section. Thus, resulting some prices in dataset 100 times more than it should be. Like, 350.000TL while that house should be maximum of 3.000TL.

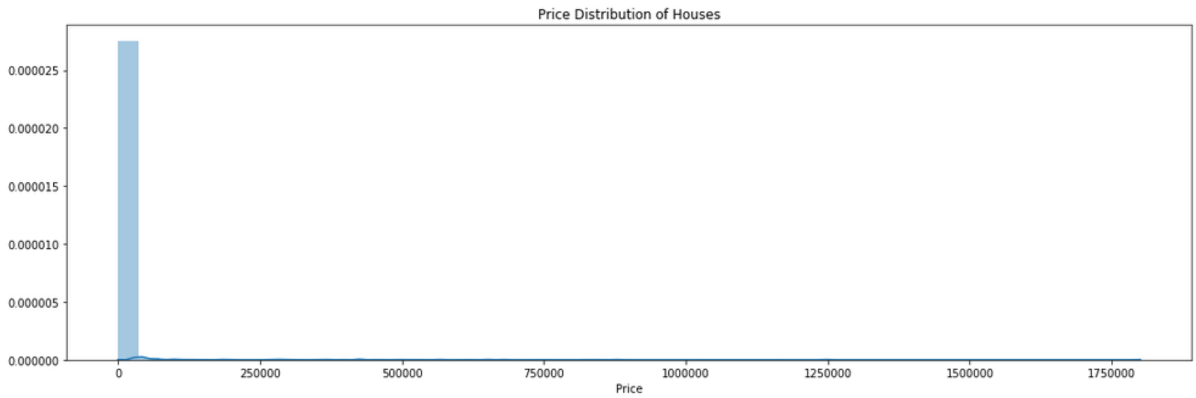


FIGURE 1. ADS PRICE DISTRIBUTION FROM WEBSITE

Some other people, especially for touristic boroughs, they are publishing annual prices, instead of monthly prices. Those will also be neglected.

Finally, the ads with more than 8.000TL rent is neglected considering above.

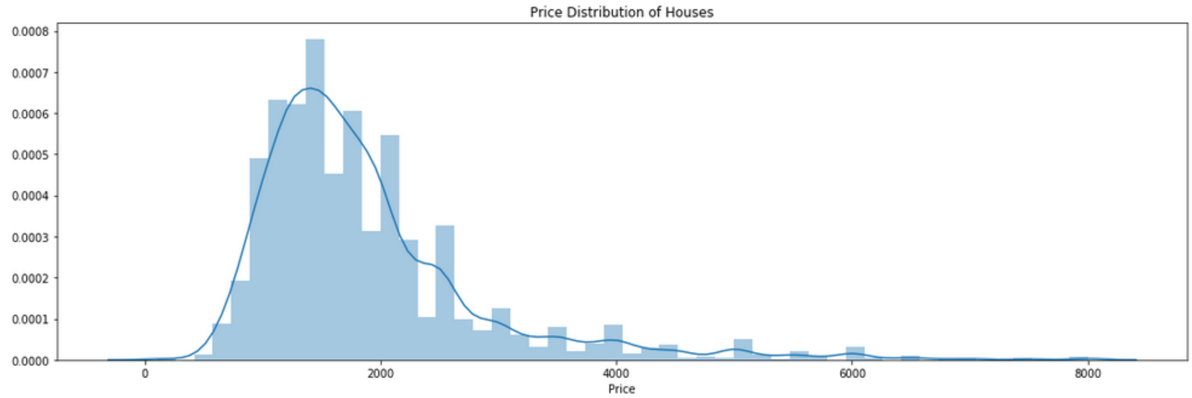


FIGURE 2. ADS PRICE DISTRIBUTION FILTERED BELOW 8.000TL

In order to analyze further in next sections, room numbers also changed to integers. In Turkey, its common to use 2+1, 3+1 etc. while naming room numbers. For example, 2+1 means 2 rooms, 1 saloon, so total of 3 rooms.

The coordinates for advertised houses are not given in website, so for each house borough coordinates are used for house location.

	Price	Size	Borough	Room#	Age	Link	P/m2	Latitude	Longitude
0	1850	86	Karşıyaka	2	7	www.hurriyetemlak.com/izmir-karsiyaka-ornekkoy...	22.0	38.503471	27.113483
1	1800	65	Karşıyaka	2	2	www.hurriyetemlak.com/izmir-karsiyaka-alaybey-...	28.0	38.503471	27.113483
2	2650	145	Karşıyaka	4	1	www.hurriyetemlak.com/izmir-karsiyaka-yali-kir...	18.0	38.503471	27.113483
3	3500	98	Karşıyaka	3	11	www.hurriyetemlak.com/izmir-karsiyaka-mavisehi...	36.0	38.503471	27.113483
4	3400	104	Karşıyaka	3	12	www.hurriyetemlak.com/izmir-karsiyaka-yali-kir...	33.0	38.503471	27.113483

TABLE 2. ADS FINAL DATA FRAME

Most popular venues are obtained and stored in a data frame from foursquare for each borough.

	name	categories	lat	lng	Borough
0	Ödemiş Yıldız Kent Arşivi ve Müzesi (ÖYKAM)	History Museum	38.226990	27.969595	Ödemiş
1	Köfteci Halil ÇARIKCI	Steakhouse	38.229363	27.970096	Ödemiş
2	Dostol Kebap Salonu	Comfort Food Restaurant	38.229145	27.973746	Ödemiş
3	Ödemiş Meşhur Köfteci İsmail	Steakhouse	38.229647	27.969806	Ödemiş
4	Kardelen Pide Ve Lahmacun Salonu	Bakery	38.228964	27.968276	Ödemiş
...	...	...	...	...	...
95	Gizli sahil	Beach	38.397508	26.984567	Narlıdere
96	Taka Balık Evi	Seafood Restaurant	38.406432	26.997929	Narlıdere
97	Balçova Termal Otel	Hotel	38.387449	27.034793	Narlıdere
98	Buckin Coffee Co.	Coffee Shop	38.393881	27.046481	Narlıdere
99	Wyndham Grand İzmir	Hotel	38.411814	27.032217	Narlıdere

2389 rows × 7 columns

TABLE 3. VENUES FINAL DATA FRAME

3. METHODOLOGY

3.1. Cluster House Ads

The houses are clustered by on their price, size, room#, age and p/m2(price per square meter). In order to find the optimal k value, cluster results are tested for k values from 1 to 15. Based on the two following graphs, k is selected as 6.

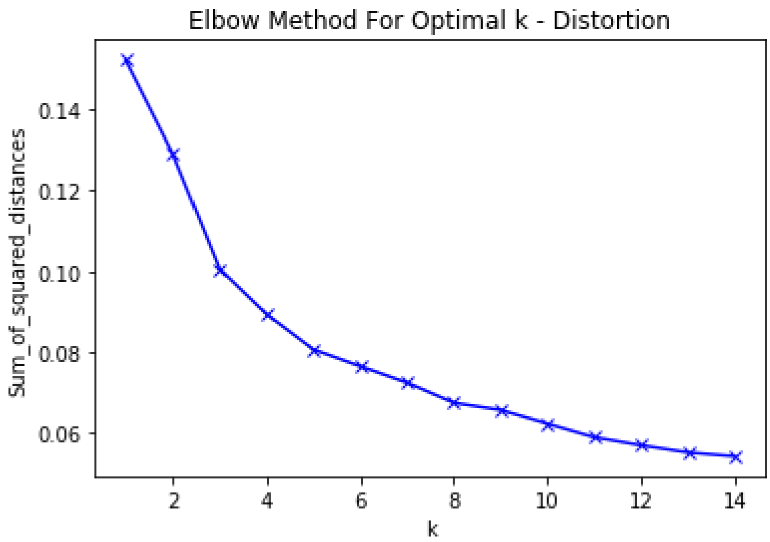


FIGURE 3. ELBOW METHOD FOR OPTIMAL K - DISTORTION

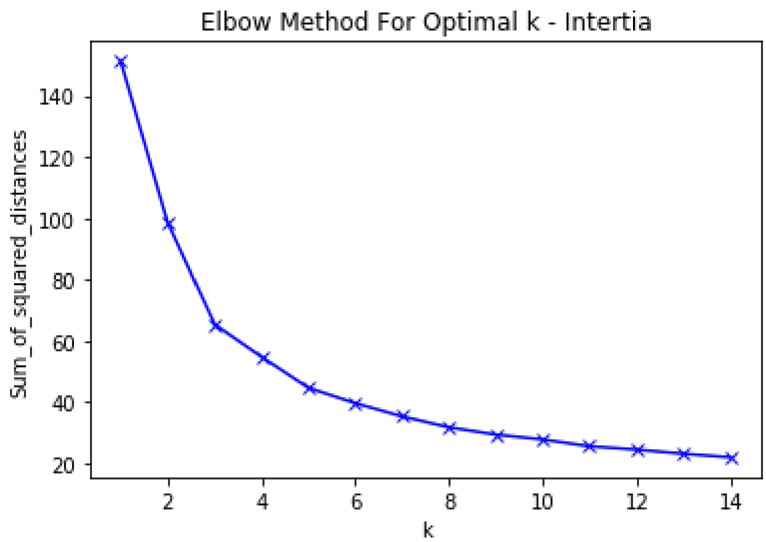
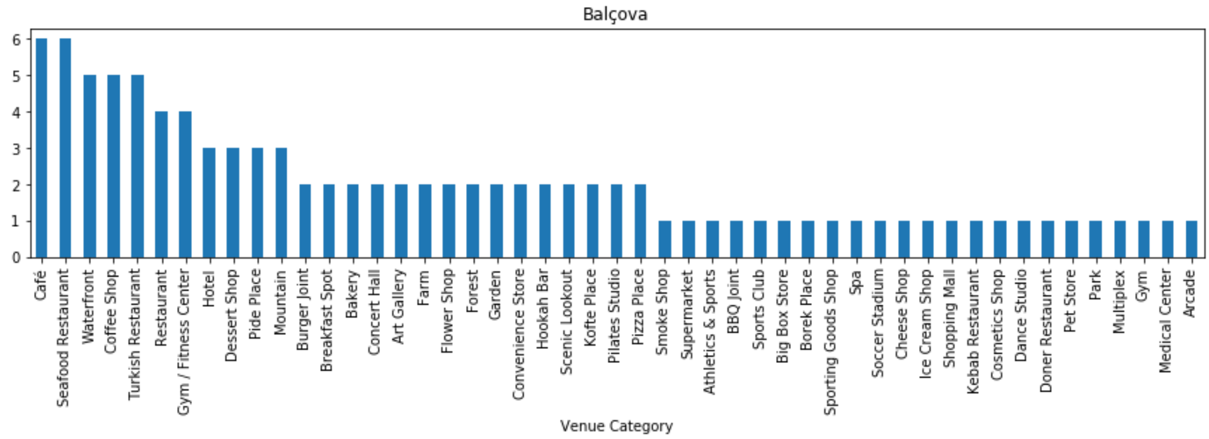


FIGURE 4. ELBOW METHOD FOR OPTIMAL K – INERTIA

3.2. Cluster Boroughs

For each borough most common venues are determined based on their categories' occurrence. This data frame will be used to cluster boroughs.



	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	Aliağa	Restaurant	Café	Beach	Seafood Restaurant	Turkish Restaurant	Fast Food Restaurant	Hotel	Coffee Shop	Pide Place	Gym / Fitness Center
1	Balçova	Seafood Restaurant	Café	Turkish Restaurant	Waterfront	Coffee Shop	Gym / Fitness Center	Restaurant	Pide Place	Mountain	Dessert Shop
2	Bayraklı	Café	Bakery	Arcade	Restaurant	Convenience Store	Pharmacy	Pide Place	Waterfront	Farmers Market	Breakfast Spot
3	Bergama	Café	Historic Site	Hotel	Bar	Beer Garden	Breakfast Spot	Dessert Shop	Restaurant	Turkish Restaurant	Jewelry Store
4	Bornova	Café	Restaurant	Turkish Restaurant	Meyhane	Dessert Shop	Coffee Shop	Gym	Gym / Fitness Center	Pastry Shop	Park

TABLE 4. MOST COMMON VENUES DATA FRAME

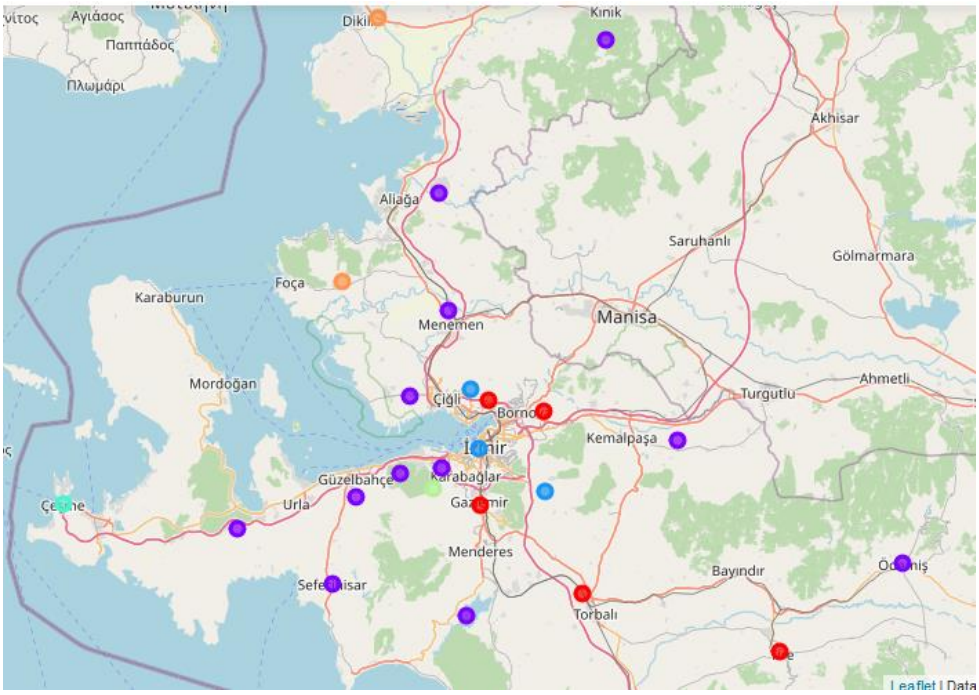


IMAGE 2. CLUSTERED BOROUGHS

4. RESULTS

In the end program asks for user input for search criteria from user. After the user provides input, previously optimized K-Means cluster fits the input and finds its cluster.

```
Please enter your search criterias approximately:
RENT[TL]:3000
SIZE[m2]:130
AGE[Years]:5
ROOM#[3+1=4]:4
Enter your favourite borough:Bornova
user cluster: 2
```

Then based on user cluster and user's favorite borough cluster, program offers 5 house recommendation to the user on the same house cluster and borough cluster. Also, provides links to access detailed data.

```
One of the 5 houses below is suitable for you
```

Recommended House 1

Recommended House 2

Recommended House 3

Recommended House 4

Recommended House 5

	House Cluster	Price	Size	Borough	Room#	Age	P/m2	Latitude	Longitude	Venue Cluster
1184	2	3750	170	Bornova	4	12	22.0	38.467898	27.260805	0
3098	2	3300	182	Gaziemir	4	0	18.0	38.321982	27.132703	0
3061	2	2500	135	Gaziemir	4	13	19.0	38.321982	27.132703	0
1415	2	2700	150	Bornova	4	10	18.0	38.467898	27.260805	0
3049	2	2200	155	Gaziemir	4	5	14.0	38.321982	27.132703	0

5. CONCLUSION

In this study, I scraped rental house ads from real estate website. K-means clustering is used for classifying both houses and boroughs. After clustering boroughs and ads separately and merging, final report gives user an idea about suitable houses in suitable boroughs based on users' preferences. This report can be used by users who is looking for a new house to move, or users who are curious about their current house's specifications. Moreover, companies can use this solution in order to offer their customers appropriate houses.