ISTANBUL – SATILIK EV FIYAT TAHMINI

emlakjet.com

SENA NUR BALCIOĞLU

7/10/2023

1. Projenin Amacı

Emlakjet, satılık/kiralık veya yatırımlık olsun aranan emlağı en hızlı ve en kolay şekilde bulma fırsatı sunmak üzere 2006 yılında kurulan bir platformdur. Projenin amacı, Türkiyedeki satılık konutların analizini çıkartmak ve fiyat tahminini yapmaktır.

2. Veri seti Bilgileri

19-23 Eylül 2023 tarihleri arasındaki emlakjet sitesindeki bütün illerin satılık konut bilgilerini içeren bir verisetidir. Yaklaşık 155.000 veri ve 20 adet parametre içermektedir

Fiyat : Konutun fiyat bilgisi
Adres : Konum bilgisi (İl/İlçe/Mahalle)
Oda Sayısı : Konut oda bilgisi (3+1, Stüdyo...)
Bulunduğu Kat : Kat sayısı (5.Kat, Çatı Katı, Bahçe Katı, Yüksek Giriş...)
Isıtma Tipi : Sobalı / Doğalgaz ...
Krediye Uygunluk : Krediye Uygun/Krediye Uygun Değil
Yapı Durumu : Sıfır / Yapım Aşamasında / İkinci El
Tapu Durumu : Kat Mülkiyeti / Kat İrtifakı ...
Eşya Durumu : Eşyalı / Eşyasız
Site İçerisinde : Evet / Hayır
Türü : Verisetinde tür sadece "Konut" bilgisini içermektedir
Tipi : Konutun tipi (Bina, Daire, Residence...)
Brüt Metrekare : Konutun brüt metrekare değeri
Binanın Yaşı : Bina kaç yıldır var
Binanın Kat Sayısı : Konut kaç katlı binada
Kullanım Durumu : Boş / Kiracı Oturuyor / Mülk sahibi Oturuyor
Yatırıma Uygunluk : Uygun / Uygun Değil
Banyo Sayısı : Konutta kaç banyo var
Balkon Sayısı : Konutta kaç balkon var
WC Sayısı : Konutta kaç wc var

```
RangeIndex: 154802 entries, 0 to 154801
Data columns (total 20 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Fiyat               154802 non-null  object
 1   Adres               154802 non-null  object
 2   Oda Sayısı          154802 non-null  object
 3   Bulunduğu Kat       154802 non-null  object
 4   Isıtma Tipi         154802 non-null  object
 5   Krediye Uygunluk    154802 non-null  object
 6   Yapı Durumu         154802 non-null  object
 7   Tapu Durumu         154802 non-null  object
 8   Eşya Durumu         154802 non-null  object
 9   Site İçerisinde     154802 non-null  object
 10  Türü                154802 non-null  object
 11  Tipi                154802 non-null  object
 12  Brüt Metrekare      154802 non-null  object
 13  Binanın Yaşı        154802 non-null  object
 14  Binanın Kat Sayısı  154802 non-null  int64
 15  Kullanım Durumu     154802 non-null  object
 16  Yatırıma Uygunluk   154802 non-null  object
 17  Banyo Sayısı        154802 non-null  int64
 18  Balkon Sayısı       154802 non-null  object
 19  WC Sayısı           154802 non-null  object
dtypes: int64(2), object(18)
```

3. Sorular

1- Veri kümesinde kaç ev bulunmaktadır?

2- Her ilden kaç ev bulunmaktadır?

3- Hangi illerde 10binden fazla ev vardır?

4- En fazla ev bulunan ilin en yüksek ve en düşük fiyatları nedir?

5- Evlerin ortalama fiyatı nedir?

6- Her ildeki evlerin ortalama fiyatı nedir?

7- Yeni yapılan evlerin ortalama fiyatı nedir?

8- İstanbuldaki ev sayısının yüzdesi kaçtır?

9- Tip değişkenine göre grafik çıkartın.

10- İstanbul ilçe bazında grafiği çıkartın.

11- Farklı bina yaşına sahip evlerin ortalama fiyatı nedir? (İl bazında)

12- İstanbulda ilçelerdeki ortalama fiyat nedir?

13- Ev tipine göre fiyat dağılımı nedir?

14- Yatırıma uygun olup olmama durumunun fiyata kayda değer bir etkisi var mıdır?

15- Tapu durumunun fiyata kayda değer bir etkisi var mıdır?

16- Ev tiplerinin illere göre dağılımı nasıldır?

17- Tipi daire olan konutların fiyat dağılımı illere göre nasıldır?

18- Dairelerin bulunduğu kat sayısının ve binanın kaç katlı olduğunun fiyata etkisi nasıldır?

19- Isıtma tipini fiyat üzerindeki etkisi nasıldır?

20- Ortalama fiyatı ülke genelindeki ortalama ev fiyatından daha yüksek olan illerde konut tipine göre nasıl bir dağılım vardır?

4. Veri seti Düzenleme Aşamaları

Data_prep fonksiyonuyla;

- Veri tiplerinde düzenlemeler yapıldı.
- Belirli sütunlar kaldırıldı.
- Boş değerler belirlendi ve nan olarak değiştirildi.
- Adres bilgileri şehir-ilçe-mahalle olmak üzere 3 ayrı sütuna ayrıldı.
- Veri setinde kullanıcılar tarafından yanlış girilen veriler vardı. Mesela bina olup tipi daire, oda sayısı 2+1 şeklinde girilen hatalı veriler kaldırıldı.
- Oda sayısını sayısal değişken olarak değiştirmek için oda sayıları toplandı. En fazla +9 şeklinde olan 9 olarak değiştirildi.

- Veri siteden çekilirken binanın yaşı 5-10 ve 11-15 olan değerler "10-May" ve "15-Nov" şeklinde alındığı için düzeltildi.
- Konutun bulunduğu kat sayısını sayısal değere çevrildi. Belirtilmemiş değerler nan olarak atandı.
- Bulunduğu kat değişkeninde "Villa Tipi" ve "Müstakil" olup, binanın kat sayısı 4ten büyük olan evler kaldırıldı.
- Bulunduğu kat değişkeninde "Villa Tipi","Müstakil","Çatı Katı","Çatı Dubleks" yazan değerleri binanın kat sayısı ile değiştirildi.

5. Veri seti İnceleme ve Temizleme Aşamaları

Yazılan sorularla veri seti hakkında genel bir bilgiye sahip olmuş olduk. İlk başta tüm şehirler için ev fiyat tahmini yapmayı planlamıştım. Bu yüzden şehirdeki ilan sayısı 500den az olan illeri sildim. Çünkü model az veriyle iyi bir öğrenme gerçekleştiremez. Eşik değerini kendim belirledim.

Veri setinin sayısal değerlerin dağılımına bakacak olursak fiyat ve brüt metrekare değerlerinde sapmalar olduğunu farkettim. Aykırı değerleri belirleyebilmek için IQR yöntemini kullandım. Çeyrek değerler olarak 25 ve 75 sınırlarını belirledim. Aslında ilk baktığımda 10-90 ya da 5-95 yapmayı tercih etmiştim. Yani aralığı genişletmiş oldum ve böylece daha az değer aykırı olacaktı. Fakat 10-90 yaptığımda sağlıklı bir sonuç elde edemedim. 25-75de aykırı olan değerler silindiğinde daha sağlıklı bir model elde ettim.

Veri setine baktığımda fiyatı 30milyondan fazla 1369 olduğunu görüyoruz. Fakat bu değerlerin içinde yalı, daire, residence, çiftlik evi gibi bütün tipte konutlar var. Bu yüzden fiyat aralığı geniş olarak gözüküyor. Bu yüzden sadece “daire” olan veriler üzerinden bir çalışma gerçekleştirdim.

Fiyatı 500.000den düşük olanları sildim. IQR yöntemi ile illerin kendi içindeki alt ve üst sınırlarını belirleyip, bu sınırların dışında olan verileri sildim. Ek olarak brüt metrekaresi 1000den fazla olan ve 20den az olan verileri de eledim.

6. Makine Öğrenmesi

İlçe ve mahalle sütunlarını kaldırdım. Kategorik değişkenlere one hot encoder işlemi gerçekleştirdikten sonra xgboost algoritmasını kullanarak model oluşturdum. R2 değeri 0.57 geldi ve daha sonra bütün illeri kendi içinde modellemeye karar verdim. Her bir şehri ayrı excel dosyasına böldüm ve mahalle sütununu kaldırdım. Sonra her şehri tek tek modelledim ve hata oranlarına baktığımda bazı illerde hata oranı düşükken bazı illerde yüksek olduğunu gördüm. Bu yüzden detaylı çalışma yapmak için veri setinin sadece “İstanbul” kısmını almaya karar verdim.

7. İstanbul

İstanbul için önce tekrar veri temizleme ve inceleme işlemlerini gerçekleştirdim.

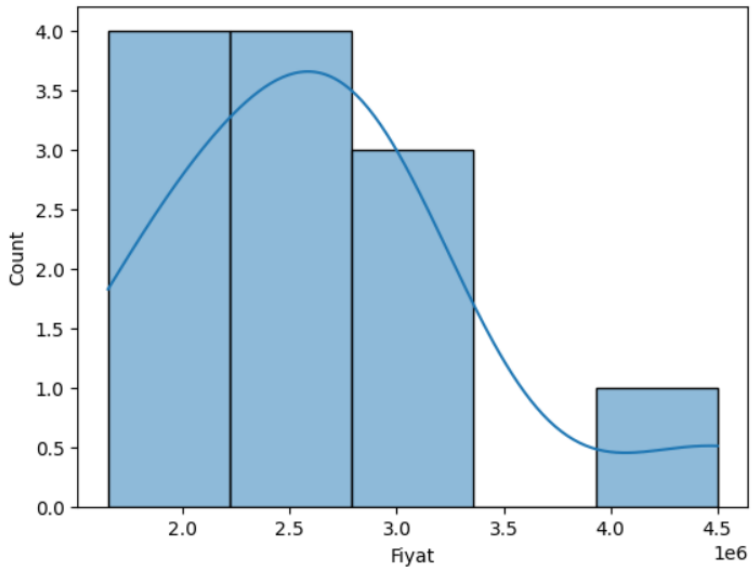
Genel veri setinde fiyatı 1.000.000dan düşük olan değerleri kaldırdım. Fiyat dağılımına histplot grafiği ile baktığımda sağa çarpık bir dağılım görmekteyim. İstanbul için grafik yine sağa çarpıktır.

Fiyat dağılımı1.000.000 – 19.600.000 aralığındadır. 27361 verinin 24588 adeti 8milyondan düşüktür.

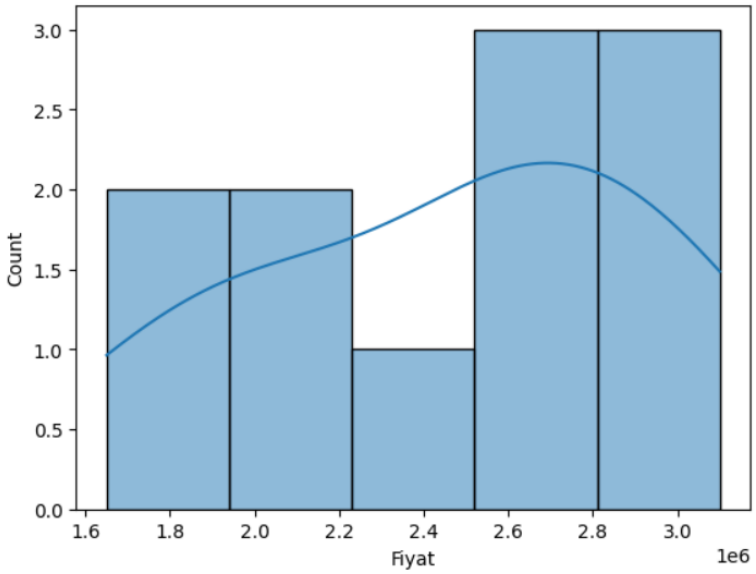
İlçe özelinde aykırı değer analizini yine IQR yöntemini kullanarak gerçekleştirdim. Aslında bazı ilçelere özel olarak baktığımda yine bir çarpıklık olduğunu fark ettim. Mesela Beylikdüzü ilçesi de sağa çarpıktır. İlçelerin fiyat dağılımlarını scatter grafiği ile baktığımda aykırı değerleri daha net görebildim ve daha sonra silme işlemlerini yaptım. Çatalca ilçesinde bir adet aykırı değer vardı onu ayrıca kaldırdım. Burada önemli olan o ilçeden kaç adet verinin olduğu ve bu verilerin fiyat dağılımının nasıl olduğudur. Mesela çatalca ilçesinden 13 veri var ve bu yorum yapmak için yeterli değildir.

Çatalca ilçesi;

```
In [134]: sns.histplot(data=df_ist[df_ist["Ilce"] == "Çatalca"], x="Fiyat", kde=True)
Out[134]: <Axes: xlabel='Fiyat', ylabel='Count'>
```

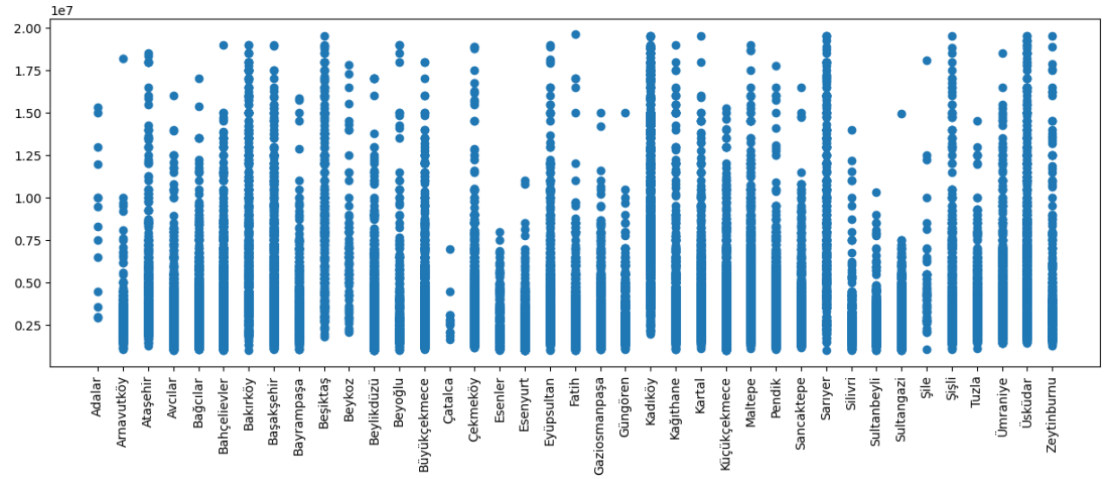



```
In [137]: sns.histplot(data=df_ist[df_ist["Ilce"] == "Çatalca"], x="Fiyat", kde=True)
Out[137]: <Axes: xlabel='Fiyat', ylabel='Count'>
```



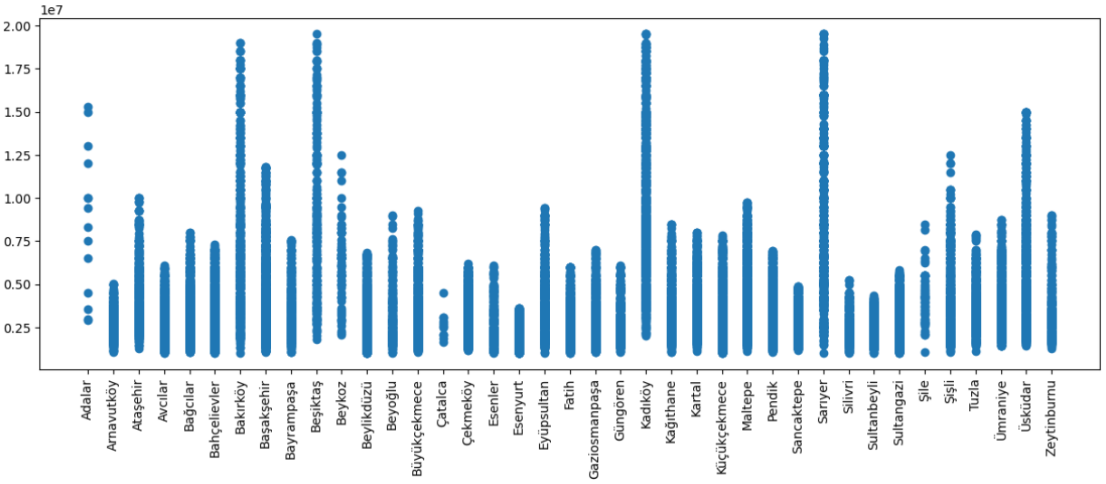
Silmeden Önce;

```
In [122]: plt.figure(figsize=(15,5))
          plt.scatter(df_ist["Ilce"],df_ist["Fiyat"])
          plt.xticks(rotation=90)
          plt.show()
```



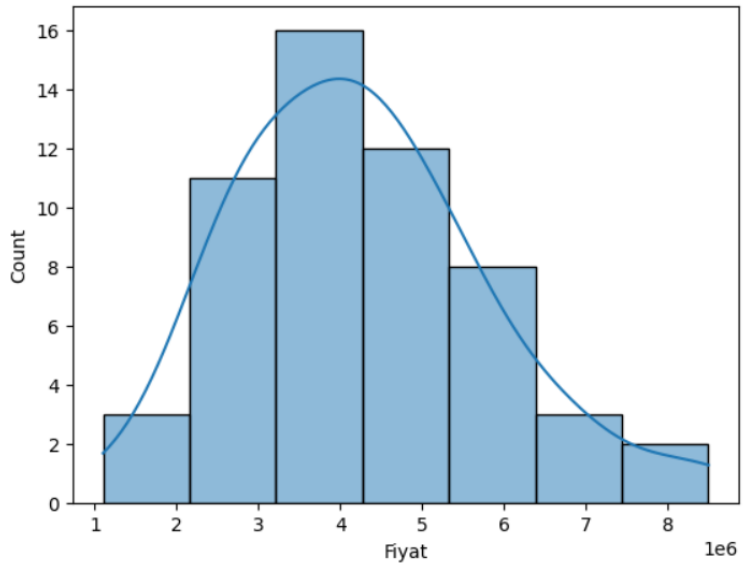
Sildikten sonra;

```
In [124]: plt.figure(figsize=(15,5))
          plt.scatter(df_ist["Ilce"],df_ist["Fiyat"])
          plt.xticks(rotation=90)
          plt.show()
```



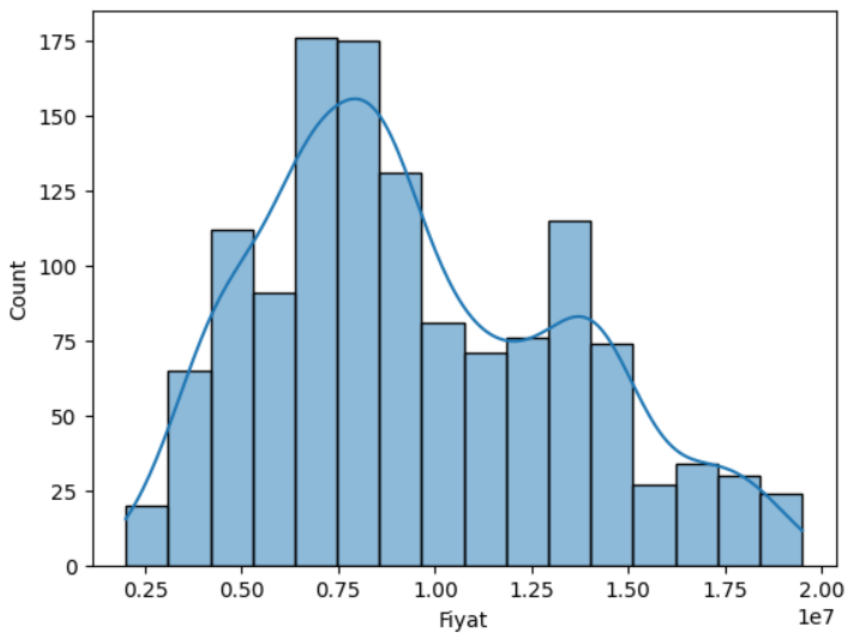
Şile ilçesinde de aykırı değerler varmış gibi gözüküyor fakat kendi içinde baktığımda normal dağılımı görüyorum.

```
In [138]: sns.histplot(data=df_ist[df_ist["Ilce"] == "Şile"], x="Fiyat", kde=True)
Out[138]: <Axes: xlabel='Fiyat', ylabel='Count'>
```



Yüksek fiyatların olduğu Kadıköy ilçesine baktığımda yine normal dağılıma yakın bir grafik görüyorum.

```
In [144]: sns.histplot(data=df_ist[df_ist["Ilce"] == "Kadıköy"], x="Fiyat", kde=True)
Out[144]: <Axes: xlabel='Fiyat', ylabel='Count'>
```



Aşağıdaki sonuca bakarsak ortanca değer ile ortalama değer birbirine yakın ve bu bize verinin düzgün dağıldığına dair bir çıkarım yapmamızı sağlar.

```
In [143]: df_ist[df_ist["Ilce"]=="Kadıköy"].describe([0.05,0.25, 0.75,0.90,0.99]).T
```

Out[143]:

	count	mean	std	min	5%	25%	50%	75%	90%	99%	max
Fiyat	1302.000	9480479.279	3967326.328	2000000.000	3901750.000	6512500.000	8750000.000	12500000.000	14950000.000	18750000.000	19500000.000
Oda Sayısı	1302.000	3.773	0.780	1.000	3.000	3.000	4.000	4.000	5.000	6.000	7.000
Bulunduğu Kat	1302.000	5.049	3.863	-1.000	1.000	2.000	4.000	7.000	10.000	16.990	24.000
Brüt Metrekare	1302.000	127.730	35.091	45.000	75.000	105.000	125.000	143.250	175.000	240.000	300.000
Binanın Kat Sayısı	1302.000	9.751	3.998	3.000	4.000	7.000	9.000	12.000	14.000	24.000	30.000
Banyo Sayısı	1302.000	1.627	0.547	1.000	1.000	1.000	2.000	2.000	2.000	3.000	4.000
Balkon Sayısı	160.000	1.337	0.571	1.000	1.000	1.000	1.000	2.000	2.000	3.000	3.000
WC Sayısı	282.000	1.635	0.545	1.000	1.000	1.000	2.000	2.000	2.000	3.000	3.000

Genel İstanbul fiyat dağılımına bakarsak ortanca değer ve ortalamanın yine o kadar yakın olmadıklarını ve %90dan sonra fiyat artış hızının yükseldiğini anlıyorum.

```
In [132]: df_ist.describe([0.05,0.25, 0.75,0.90,0.99]).T
```

Out[132]:

	count	mean	std	min	5%	25%	50%	75%	90%	99%	max
Fiyat	25909.000	3793196.155	2653345.867	1000000.000	1395000.000	2150000.000	3000000.000	4500000.000	6750000.000	15000000.000	19500000.000
Oda Sayısı	25909.000	3.487	0.988	1.000	2.000	3.000	3.000	4.000	5.000	7.000	10.000
Bulunduğu Kat	25909.000	3.120	3.224	-4.000	0.000	1.000	2.000	4.000	6.000	16.000	40.000
Brüt Metrekare	25909.000	117.138	41.122	27.000	70.000	90.000	110.000	135.000	168.200	250.000	1000.000
Binanın Kat Sayısı	25909.000	6.580	4.842	1.000	3.000	4.000	5.000	7.000	12.000	30.000	99.000
Banyo Sayısı	25909.000	1.332	0.526	1.000	1.000	1.000	1.000	2.000	2.000	3.000	5.000
Balkon Sayısı	6275.000	1.301	0.525	1.000	1.000	1.000	1.000	2.000	2.000	3.000	3.000
WC Sayısı	9364.000	1.374	0.539	1.000	1.000	1.000	1.000	2.000	2.000	3.000	4.000

Aşağıdaki sonuca bakarsak fiyatı 15milyondan fazla olan değerlerin çoğunun Kadıköy ilçesinde olduğunu ve dağılımı etkilediğini söyleyebiliriz. Bu değerler silinebilir. Fakat ben silmeyi tercih etmedim. Çünkü Kadıköy ilçesinde normal olarak adlandırılır. Aynı şekilde Sarıyer, Beşiktaş gibi ilçelerde bu aralıkta fiyatlar olması normaldir.

```
In [133]: df_ist[df_ist["Fiyat"] > 15000000].groupby("Ilce")["Fiyat"].agg(["count","mean","max","min"])
```

Out[133]:

Ilce	count	mean	max	min
Adalar	1	15300000.000	15300000	15300000
Bakırköy	39	16965384.615	18500000	15500000
Beşiktaş	32	17203125.000	19500000	15250000
Kadıköy	115	17259348.017	19500000	15200000
Sarıyer	41	17005414.634	19500000	15247000

Makine öğrenmesi aşamasında ise ilçe ve mahalle sütunlarını kaldırmadım. R2 değeri 0.82 sonucunu elde ettim.