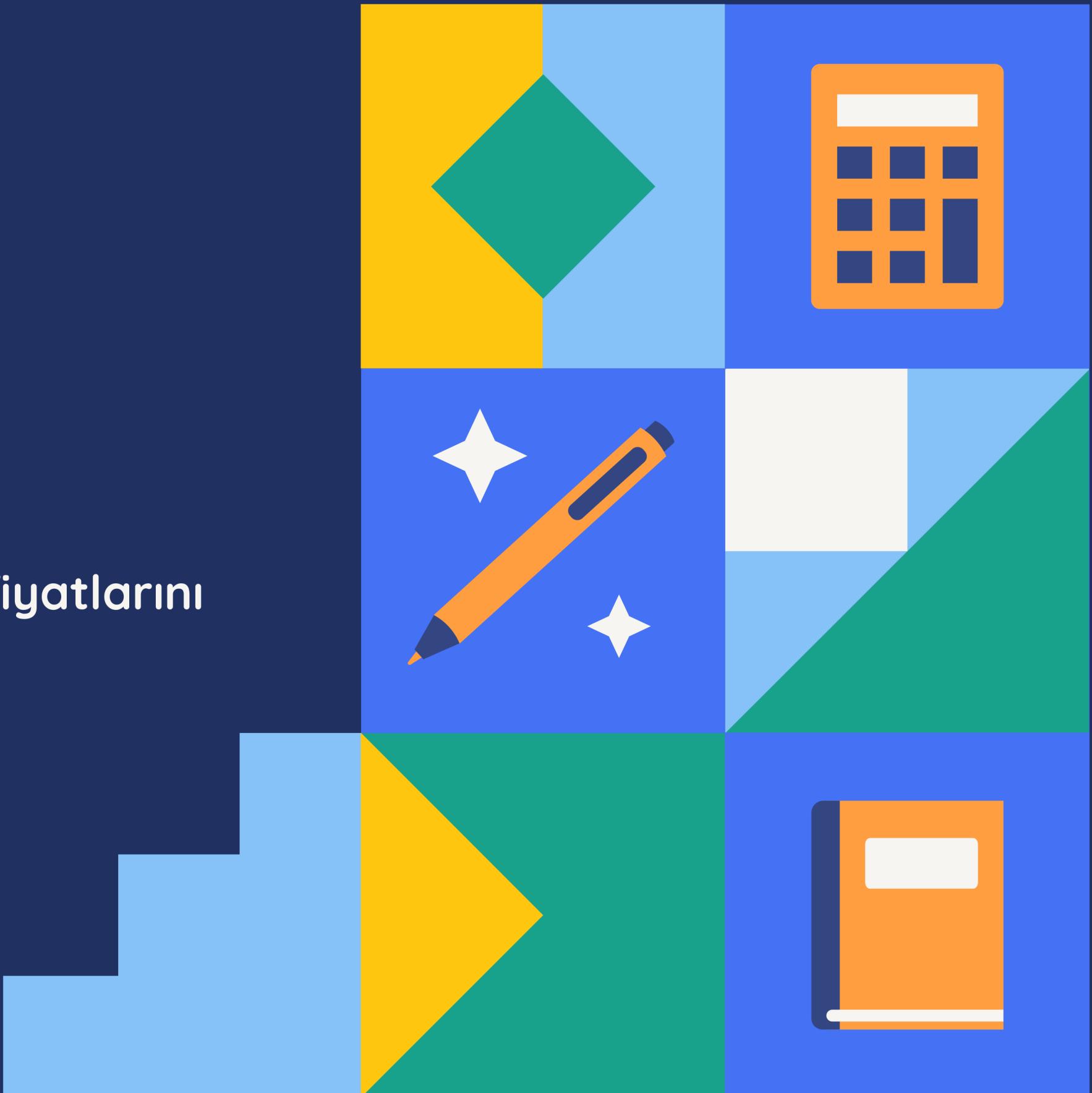


# Home Data Analysis

Evin özelliklerine dayanarak yeni satışların fiyatlarını tahmin ediyoruz.

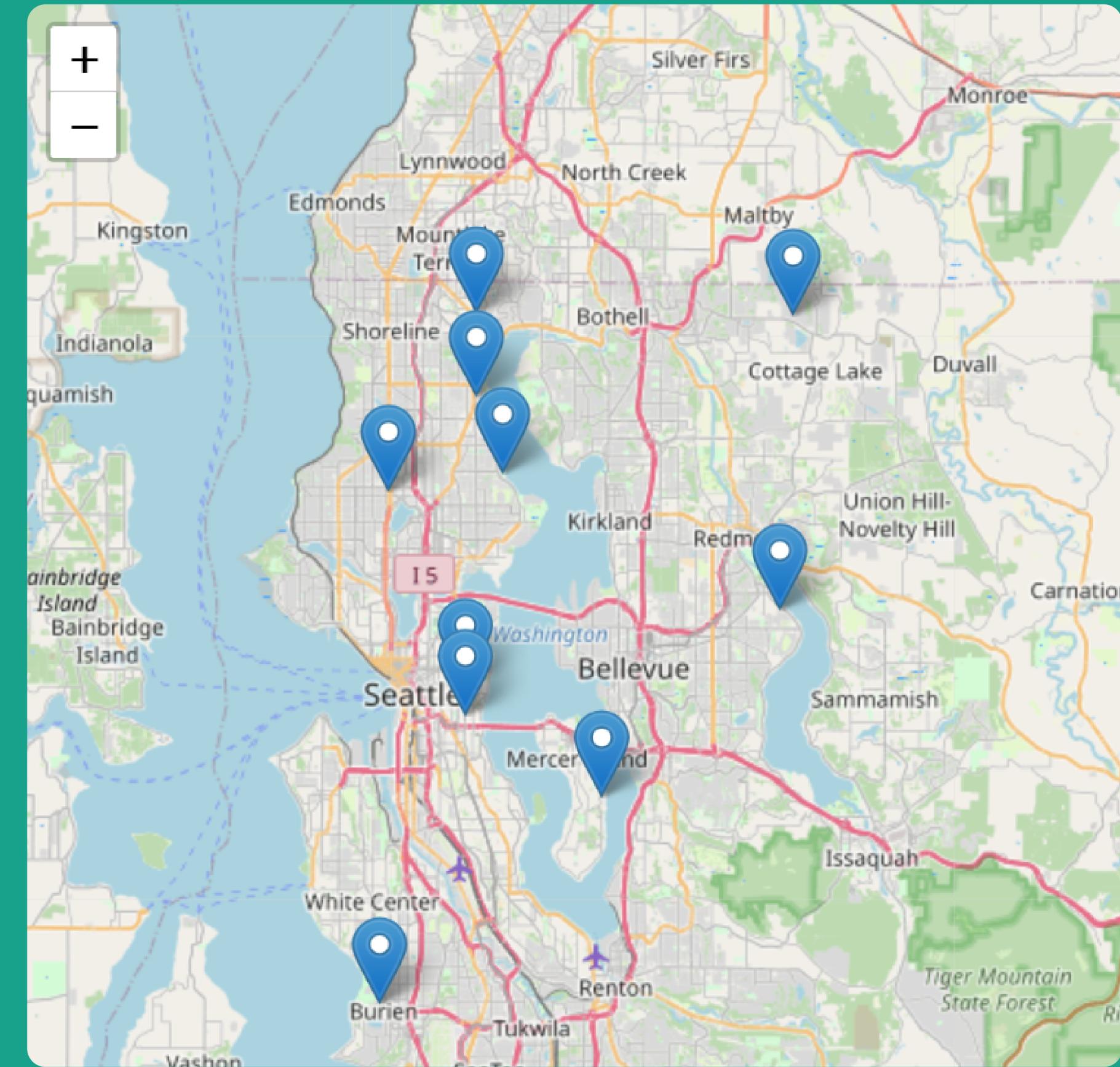


```
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   id                21613 non-null   int64  
 1   date              21613 non-null   object  
 2   price              21613 non-null   float64 
 3   bedrooms           21613 non-null   int64  
 4   bathrooms          21613 non-null   float64 
 5   sqft_living        21613 non-null   int64  
 6   sqft_lot            21613 non-null   int64  
 7   floors              21613 non-null   float64 
 8   waterfront          21613 non-null   int64  
 9   view               21613 non-null   int64  
 10  condition           21613 non-null   int64  
 11  grade              21613 non-null   int64  
 12  sqft_above          21613 non-null   int64  
 13  sqft_basement       21613 non-null   int64  
 14  yr_built            21613 non-null   int64  
 15  yr_renovated        21613 non-null   int64  
 16  zipcode             21613 non-null   int64  
 17  lat                 21613 non-null   float64 
 18  long                21613 non-null   float64 
 19  sqft_living15       21613 non-null   int64  
 20  sqft_lot15           21613 non-null   int64  
dtypes: float64(5), int64(15), object(1)
```

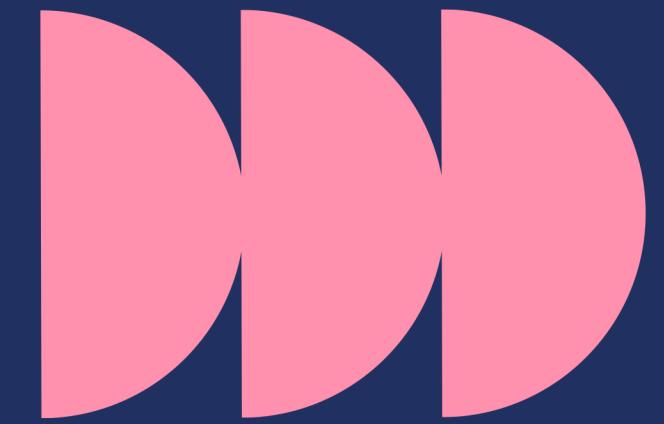
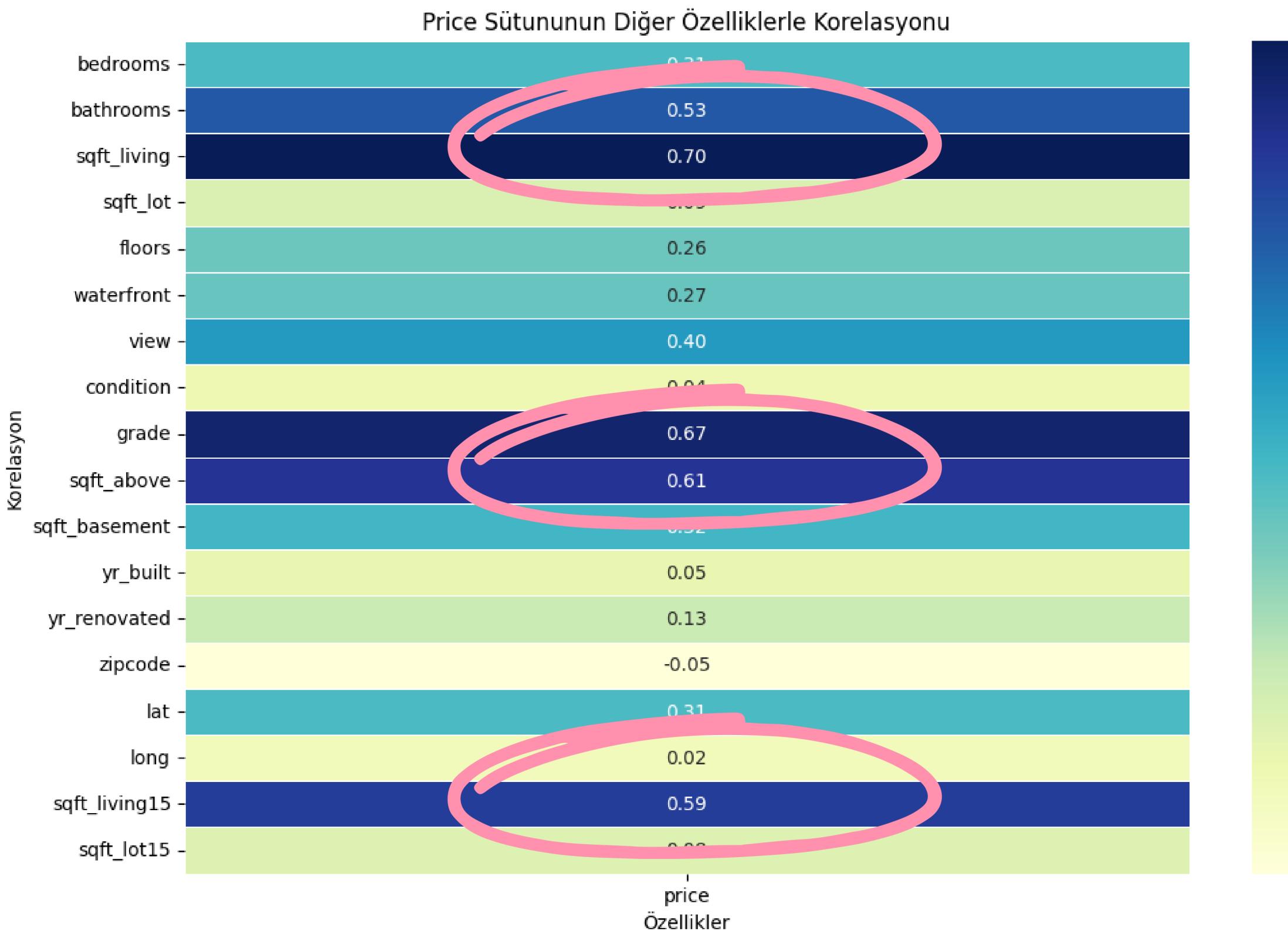
# Veri Setimizi Taniyalım

Veri kümesi 21613 satır ve 21 sütun içermektedir. Amacımız, diğer sütunlardaki değerleri kullanarak "price" sütunundaki değeri tahmin etmektir. Bu sayede, evin özelliklerine dayanarak yeni satışların fiyatlarını tahmin edebiliriz.

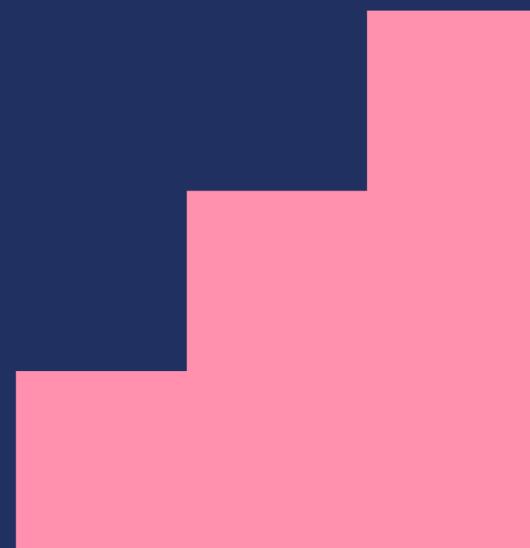
# Evlerin Kordinat Dağılımı



# Price Sütunu ile Diğer Sütunlar Arasında ki Korelasyonu

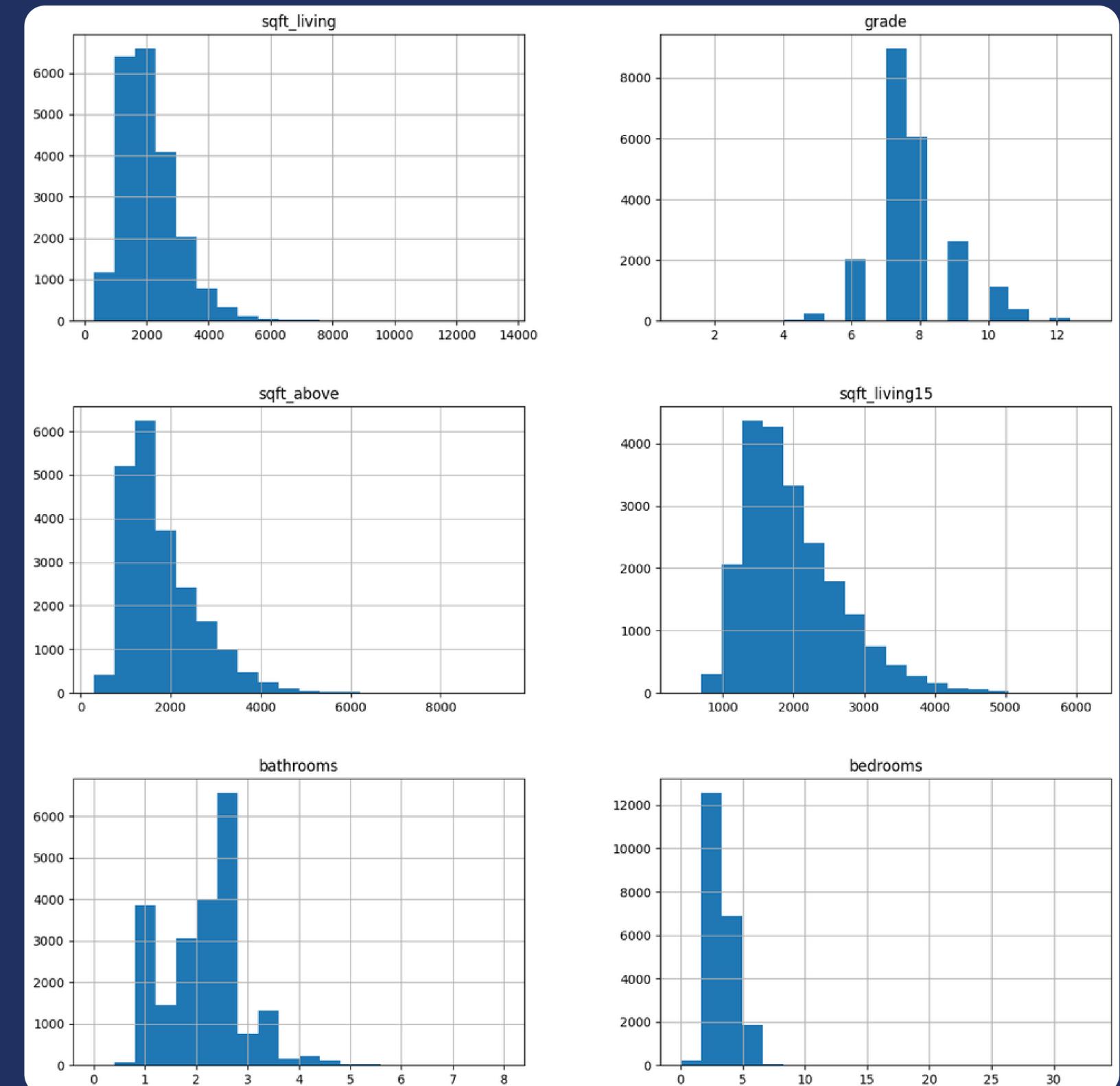


Korelasyon tablosuna göre, ev fiyatları en çok sqft\_living ve grade özellikleri ile pozitif korelasyon göstermektedir.



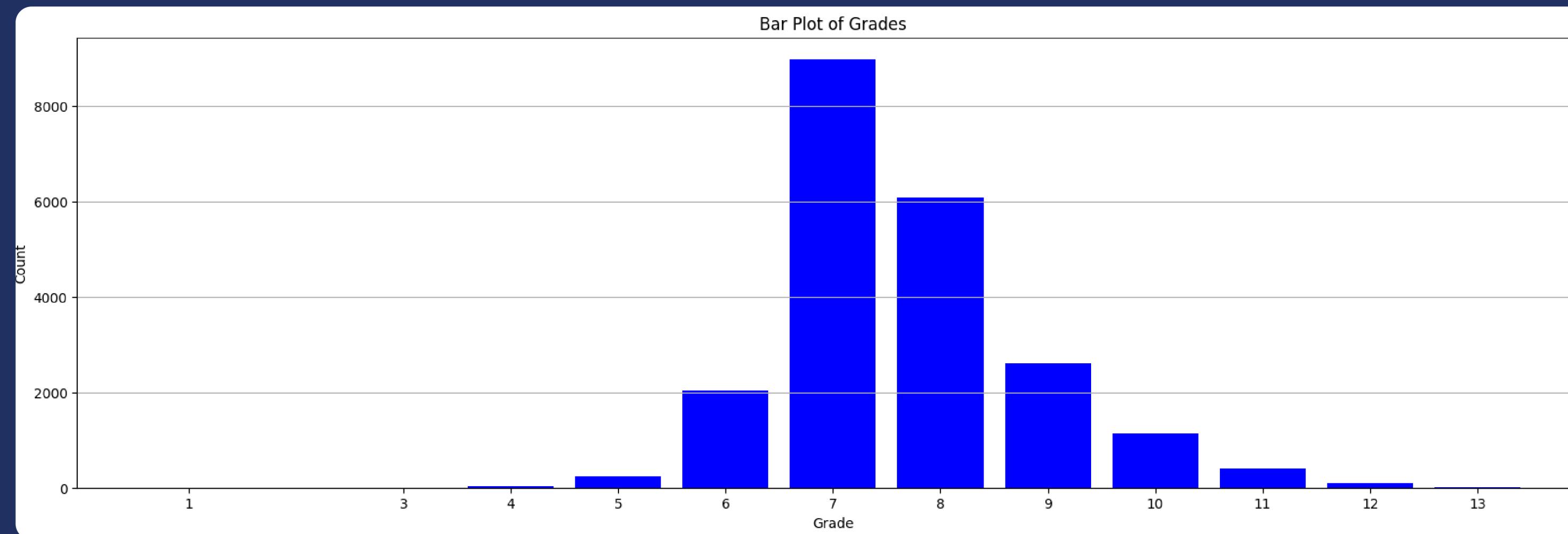
# Korelasyonu Yüksek Olan Verilerin Dağılımı

- Yaşam Alanı (Sqft\_living): Çoğu ev 500-4000 sqft aralığında. Dağılım sağa çarpık; yüksek değerler nadir.
- Derece (Grade): Evlerin çoğu 6-9 aralığında derecelendirilmiş. Ortalama kaliteyi gösteriyor.
- Zemin Üstü Alan (Sqft\_above): Çoğu ev 500-3000 sqft aralığında. Yüksek değerler nadir, sağa çarpık dağılım.
- Sonraki 15 Yıl Yaşam Alanı (Sqft\_living15): Çoğu ev 1000-3000 sqft aralığında. Dağılım sağa çarpık.
- Banyolar (Bathrooms): Çoğu evde 1-3 banyo var. 3'ten fazla banyolu evler nadir.
- Yatak Odaları (Bedrooms): Çoğu evde 2-5 yatak odası var. 5'ten fazla yatak odaklı evler nadir, ekstrem değerler mevcut.



## Grade Sütun Dağılımı

"**Grade**" sütunundaki farklı değerlerin dağılımı model performansını olumsuz etkiliyordu. Bu nedenle, değerleri "**iyi**", "**orta**" ve "**kötü**" olarak kategorilere ayırdık ve One Hot Encoder ile kodladık. Bu sayede, modelimizin performansını önemli ölçüde iyileştirdik.



## Aykırı Değerler

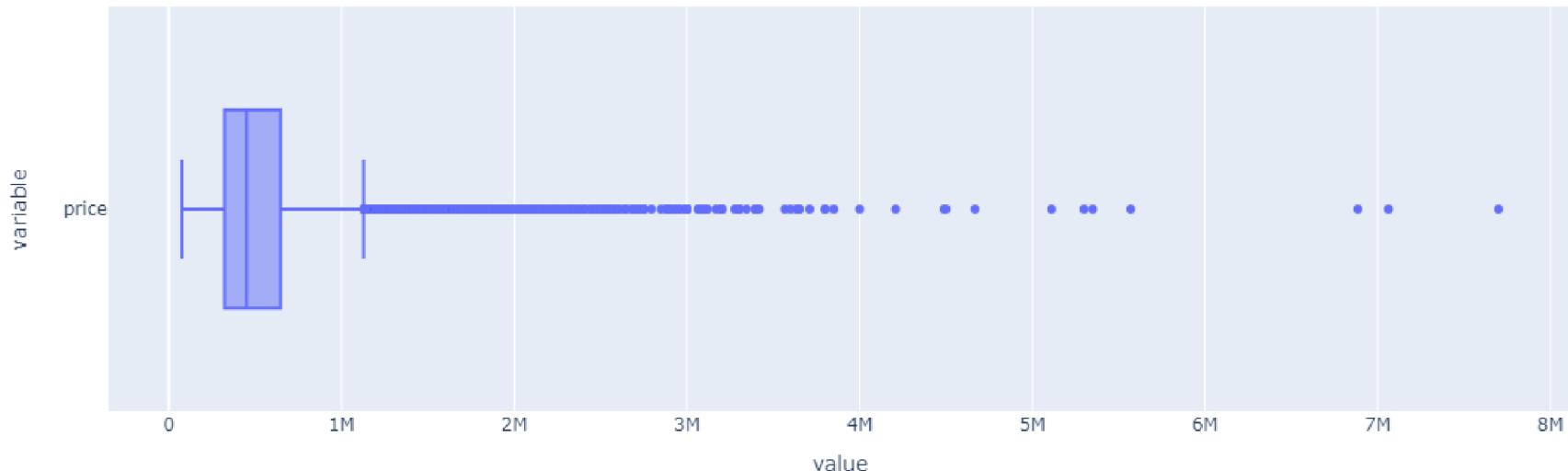
# IQR

2 milyon ile 8 milyon arasında yer almaktır ve genel veriyi etkilemektedir. Veri setindeki genel eğilimden saparak analizleri yanlıltabilir ve istatistiksel sonuçları bozabilir.

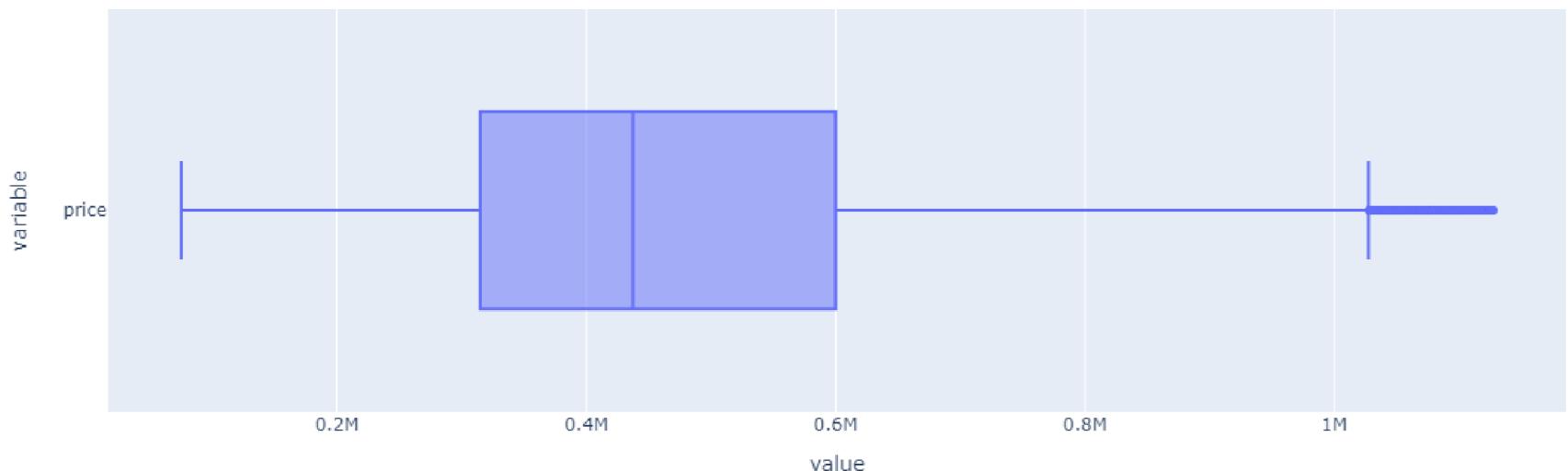
Aykırı değerler çıkarıldığında, veriler daha dar bir aralıkta yoğunlaşır. Bu durum, analizlerin güvenilirliğini artırır ve model performansını iyileştirir.

**1146 veri çıkarıldı ve 20467 verimiz kaldı.**

Aykırı Değerler ile



Aykırı Değerler Çıkarıldıkten Sonra





# Veri Dengesizliği

(**Synthetic Minority Over-sampling Technique**)

SMOTE

Verilerimiz oldukça dağınık olduğu için veri çeşitliliğimizi artırmak amacıyla **SMOTE** yöntemini denedik. Ancak, bu yöntemin modellerimiz üzerinde olumlu bir etkisi olmadığını gözlemledik. Bunun yanı sıra, **XGBoost** ve **Gradient Boosting** modellerinin büyük veri setleriyle daha iyi performans gösterdiğini fark ettik.

# Basit Modellerimiz

## Median Model

MSE: 204816.98

RMSE: 452.57

MAPE: 40 %

r2: -0.032205504147433395

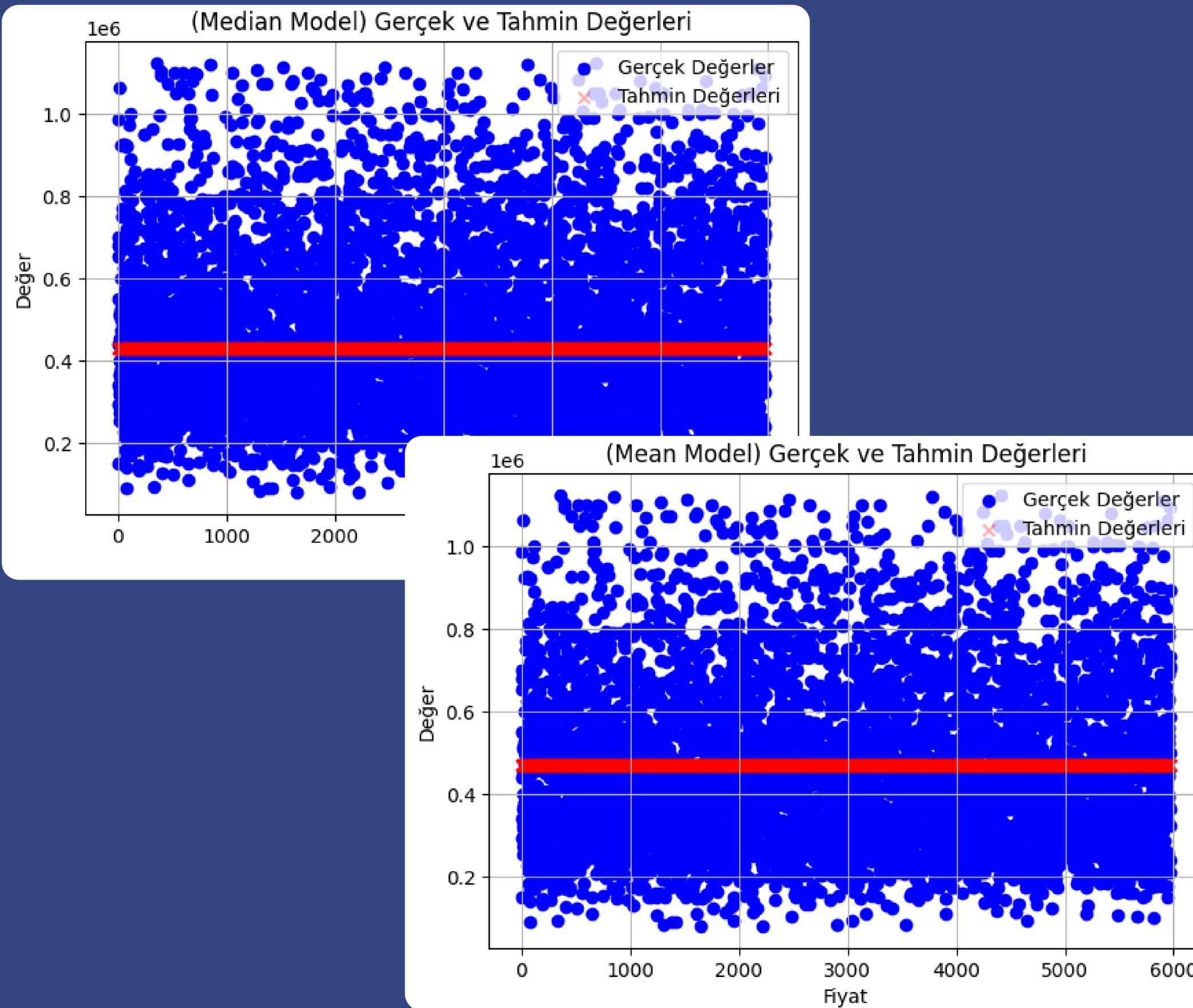
## Mean Model

MSE: 201596.45

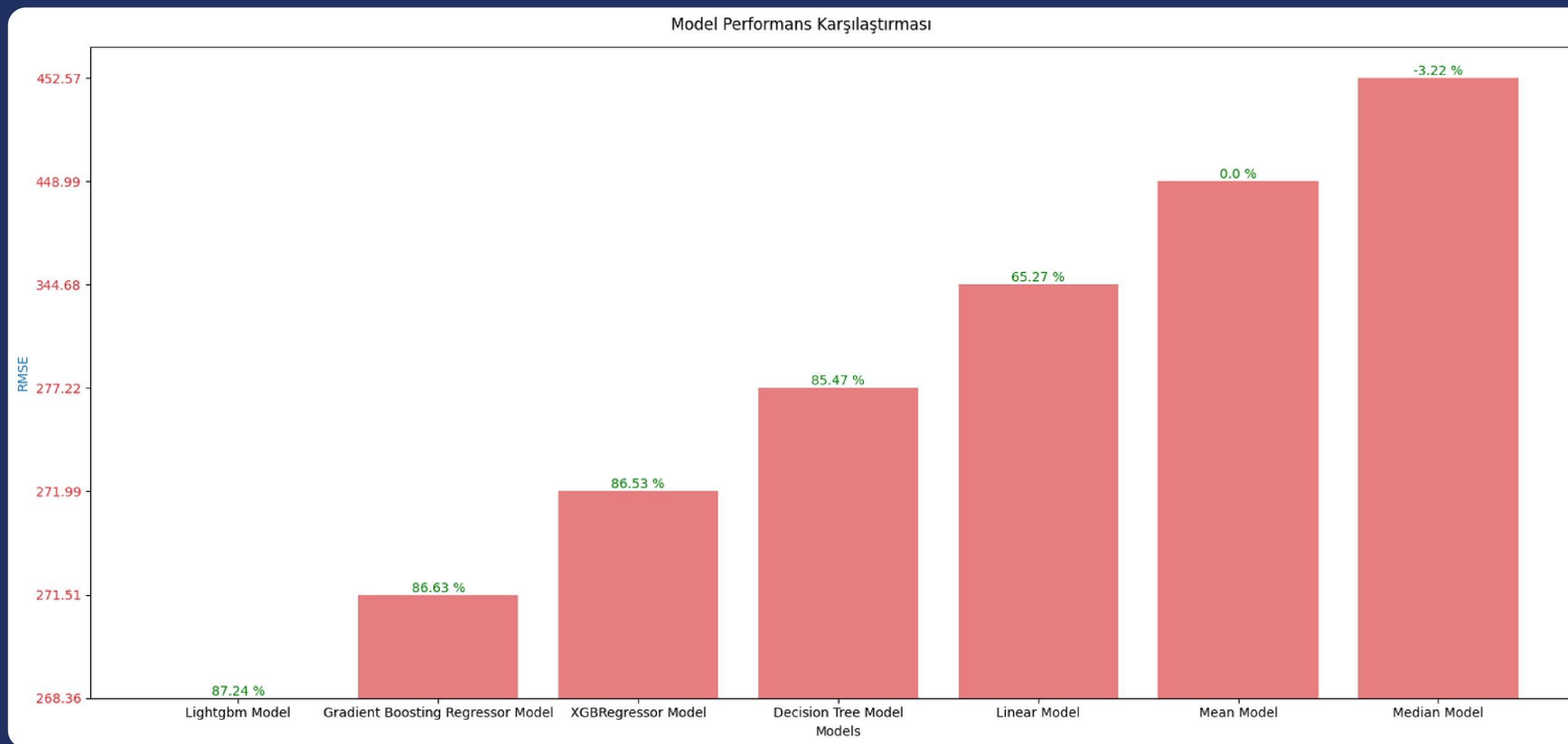
RMSE: 448.99

MAPE: 44 %

R2: 0.0



# Modellerin Karşılaştırması



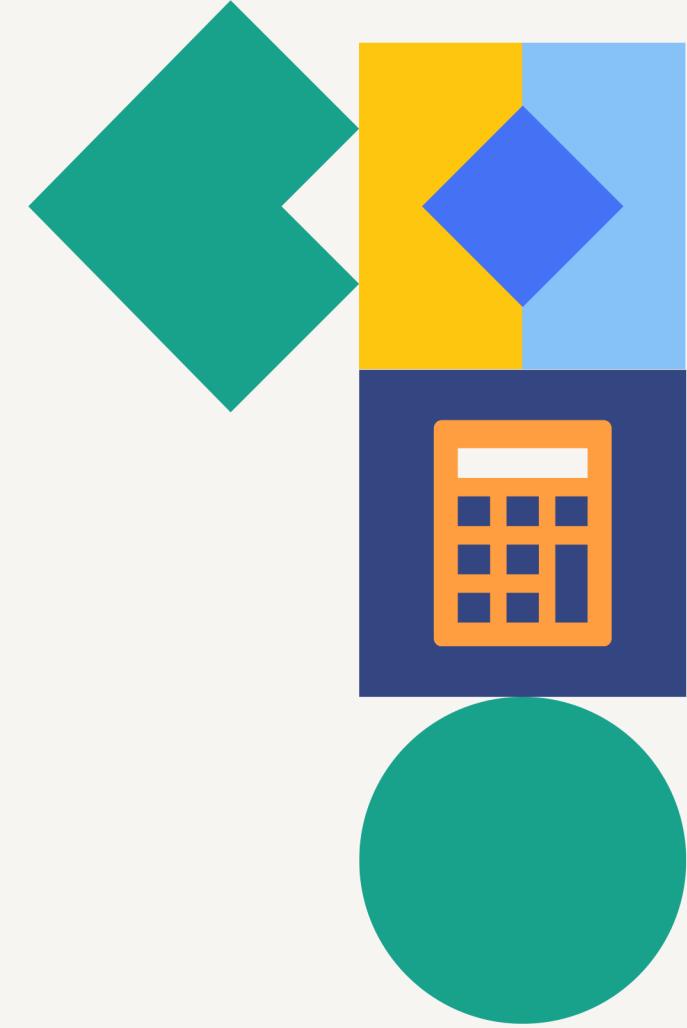
# Hiper-Parametre Turing

## RandomizedSearchCV

RandomizedSearchCV'yi, **Gradient Boosting Regressor** modelimiz için en iyi hiperparametreleri hızlı ve etkili bir şekilde bulmak amacıyla kullandık. Bu yöntem, rastgele kombinasyonlar deneyerek daha hızlı sonuç verir ve çapraz doğrulama ile modelin genellemeye yeteneğini artırır. Bu sayede, model performansını optimize etmek için verimli bir hiperparametre araması yapmamızı sağlar.



```
1 Gradient Boosting Regression RMSE: 275.99
2 Gradient Boosting Regression mape: 13 %
3 En İyi Hiperparametreler: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_depth': 4, 'learning_rate': 0.1}
4 En İyi Skor: 0.8515702619110677
```





# Teşekkürler

İSMAİL CAN KARATAŞ

SİMGE KARABUĞA

EBRU BAYRAM

