House Price Prediction

Abdurrahman Bulut

İçerik

- Veri Seti Hikayesi ve Problem Tanımı
- Keşifçi Veri Analizi
- Veri Ön işleme ve Özellik Mühendisliği
- Modelleme
- Bulgular ve İş Önerileri

İş Problemi

- Ev fiyat bilgileri ve çeşitli ev özellikleri ile ilgili veriler kullanılarak, belirli bölgelerdeki evlerin fiyatlarını tahmin edebilen bir makine öğrenmesi modeli geliştirilmesi.

Veri Seti Hikayesi

- Bu veri seti orijinal olarak Amerika Birleşik Devletleri'nin konut piyasası verilerinden derlenmiştir. Veri seti, ABD'deki farklı eyaletlerde bulunan konutların fiyatları ve bu konutlara ait çeşitli özellikleri içermektedir. Veriler, konutların bulunduğu mahallelerin demografik bilgileri, konutların fiziksel özellikleri ve satış tarihleri gibi bilgileri içermektedir. Veri seti, emlak danışmanları ve veri analistleri tarafından analiz edilmek üzere derlenmiştir ve konut piyasasında trendleri analiz etmek ve gelecekteki ev fiyatlarını tahmin etmek amacıyla kullanılmaktadır.

Değişkenler

81 Değişken

1460 Gözlem

Değişken Adı	Açıklaması
SalePrice	Mülkün dolar cinsinden satış fiyatı. Bu, tahmin etmeye çalıştığınız hedef değişkendir.
MSSubClass	Bina sınıfı
MSZoning	Genel bölgeleme sınıflandırması
LotFrontage	Mülkle bağlantılı cadde uzunluğu (lineer feet olarak)
LotArea	Arazi büyüklüğü (kare feet olarak)
Street	Yol erişim türü
Alley	Ara yol erişim türü
LotShape	Mülkün genel şekli
LandContour	Mülkün düzlüğü
Utilities	Mevcut tesisat türleri
LotConfig	Arazi yapılandırması
LandSlope	Mülkün eğimi
Neighborhood	Ames şehri sınırları içindeki fiziksel konumlar
Condition1	Ana yol veya demiryoluna yakınlık
Condition2	Ana yol veya demiryoluna yakınlık (eğer ikinci bir yol mevcutsa)
BldgType	Konut türü
HouseStyle	Konut stili
OverallQual	Genel malzeme ve bitiş kalitesi
OverallCond	Genel durum derecesi
YearBuilt	Orijinal inşaat tarihi
YearRemodAdd	Yenileme tarihi
RoofStyle	Çatı türü
RoofMatl	Çatı malzemesi
Exterior1st	Evin dış kaplaması
Exterior2nd	Evin dış kaplaması (eğer birden fazla malzeme kullanılmışsa)
MasVnrType	Taş kaplama türü
MasVnrArea	Taş kaplama alanı (kare feet olarak)
ExterQual	Dış malzeme kalitesi
ExterCond	Dış malzemenin mevcut durumu
Foundation	Temel türü
BsmtQual	Bodrum yüksekliği
BsmtCond	Bodrumun genel durumu
BsmtExposure	Yarı veya bahçe seviyesinde bodrum duvarları
BsmtFinType1	Bodrum bitmiş alan kalitesi
BsmtFinSF1	Tür 1 bitmiş alan (kare feet olarak)
BsmtFinType2	İkinci bitmiş alan kalitesi (varsa)

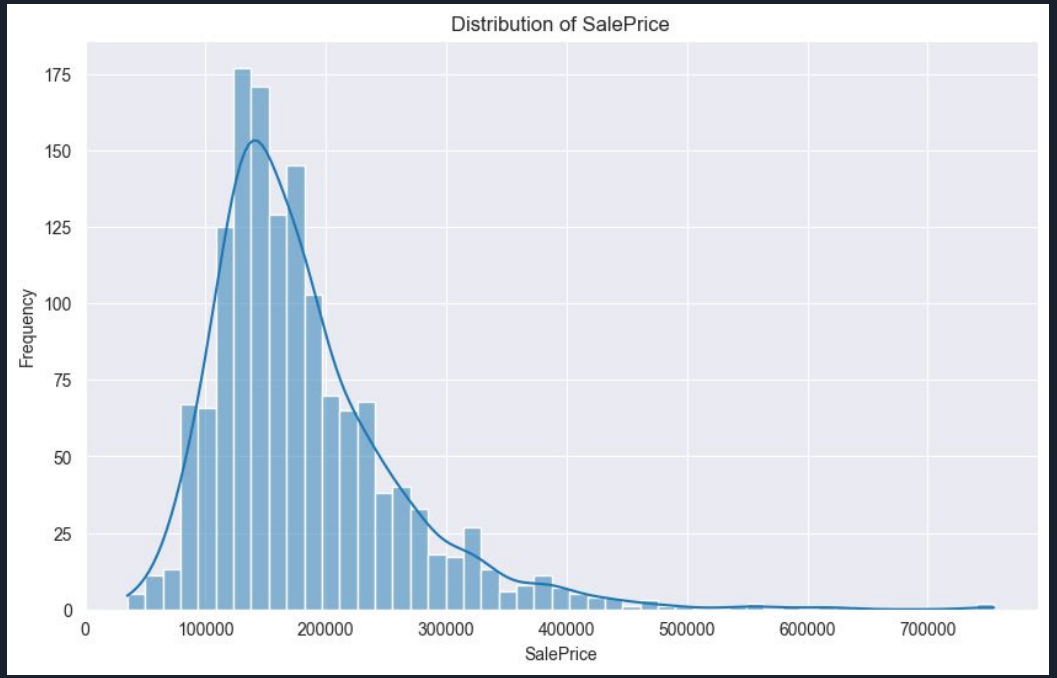
Değişkenler

Değişken Adı	Açıklaması
BsmtFinSF2	Tür 2 bitmiş alan (kare feet olarak)
BsmtUnfSF	Bitmemiş bodrum alanı (kare feet olarak)
TotalBsmtSF	Toplam bodrum alanı (kare feet olarak)
Heating	Isıtma türü
HeatingQC	Isıtma kalitesi ve durumu
CentralAir	Merkezi klima
Electrical	Elektrik sistemi
1stFlrSF	Birinci kat alanı (kare feet olarak)
2ndFlrSF	İkinci kat alanı (kare feet olarak)
LowQualFinSF	Düşük kaliteli bitmiş alan (tüm katlar, kare feet olarak)
GrLivArea	Zemin üstü yaşam alanı (kare feet olarak)
BsmtFullBath	Bodrum tam banyoları
BsmtHalfBath	Bodrum yarım banyoları
FullBath	Zemin üstü tam banyolar
HalfBath	Zemin üstü yarım banyolar
Bedroom	Bodrum seviyesi üzerindeki yatak odası sayısı
Kitchen	Mutfak sayısı
KitchenQual	Mutfak kalitesi
TotRmsAbvGrd	Zemin üstü toplam oda sayısı (banyolar hariç)
Functional	Evin fonksiyonellik derecesi
Fireplaces	Şömine sayısı
FireplaceQu	Şömine kalitesi
GarageType	Garaj konumu
GarageYrBlt	Garajın yapım yılı
GarageFinish	Garajın iç bitiş durumu
GarageCars	Garajın araç kapasitesi
GarageArea	Garaj alanı (kare feet olarak)
GarageQual	Garaj kalitesi
GarageCond	Garaj durumu
PavedDrive	Asfalt yol
WoodDeckSF	Ahşap güverte alanı (kare feet olarak)
OpenPorchSF	Açık veranda alanı (kare feet olarak)
EnclosedPorch	Kapalı veranda alanı (kare feet olarak)
3SsnPorch	Üç mevsim verandası alanı (kare feet olarak)
ScreenPorch	Kapalı veranda alanı (kare feet olarak)
PoolArea	Havuz alanı (kare feet olarak)

Değişkenler

Değişken Adı	Açıklaması
PoolQC	Havuz kalitesi
Fence	Çit kalitesi
MiscFeature	Diğer kategorilere dahil olmayan çeşitli özellikler
MiscVal	Diğer özelliklerin $ değeri
MoSold	Satış ayı
YrSold	Satış yılı
SaleType	Satış türü
SaleCondition	Satış koşulu

Keşifçi Veri Analizi

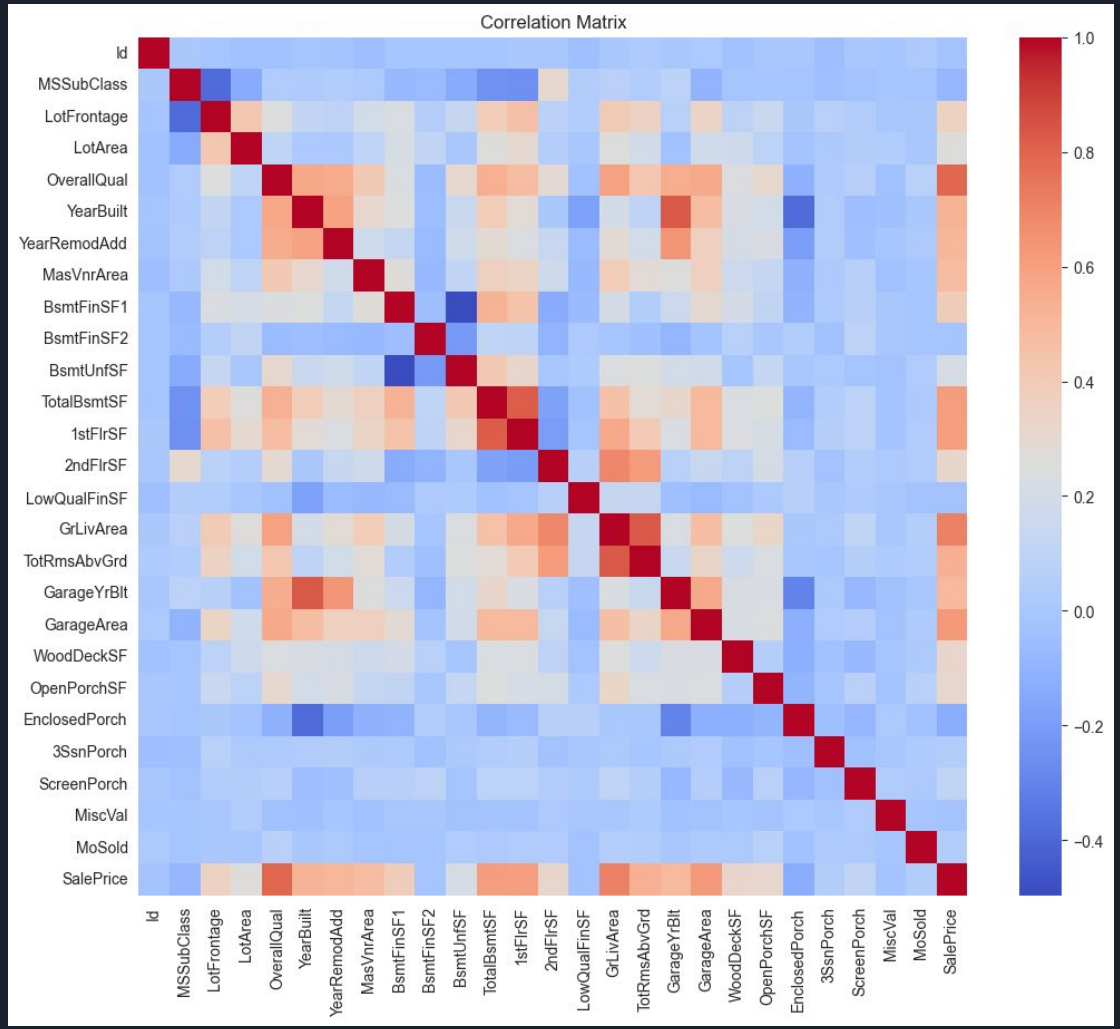


Hedef Değişken Dağılımı

Keşifçi Veri Analizi

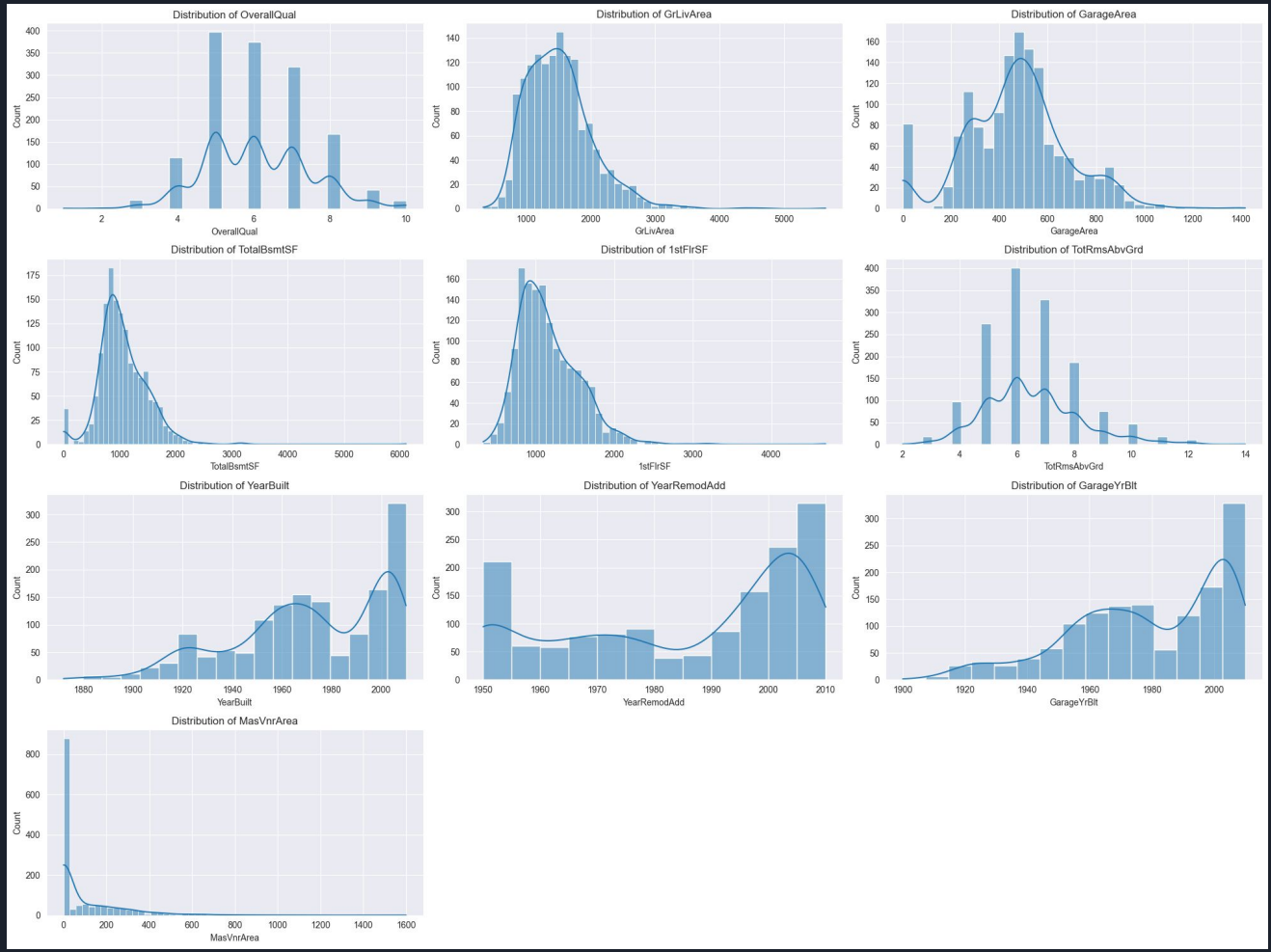
'SalePrice' ile en yüksek korelasyona sahip ilk 10 sayısal değişken

- OverallQual
- GrLivArea
- GarageArea
- TotalBsmtSF
- 1stFlrSF
- TotRmsAbvGrd
- YearBuilt
- YearRemodAdd
- GarageYrBlt
- MasVnrArea



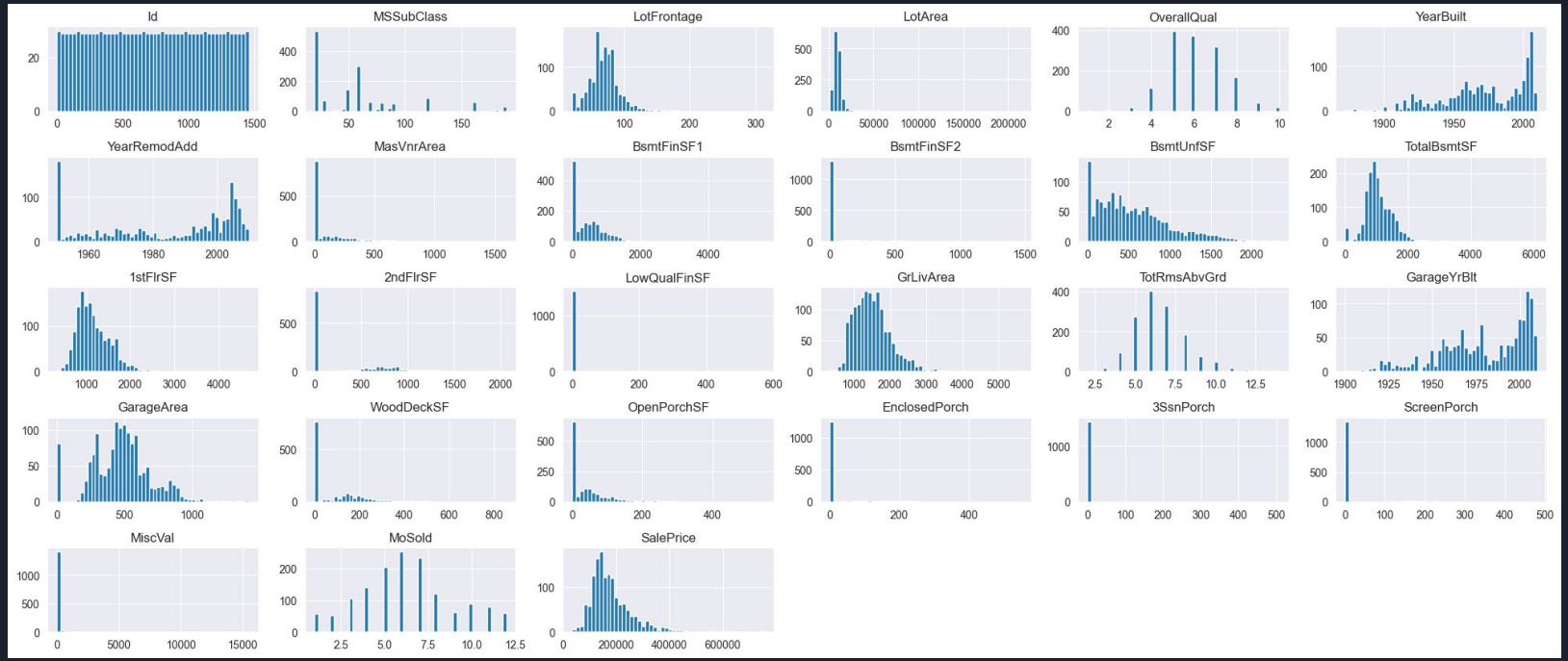
Keşifçi Veri Analizi

- En çok etki eden 10 sayısal özellik



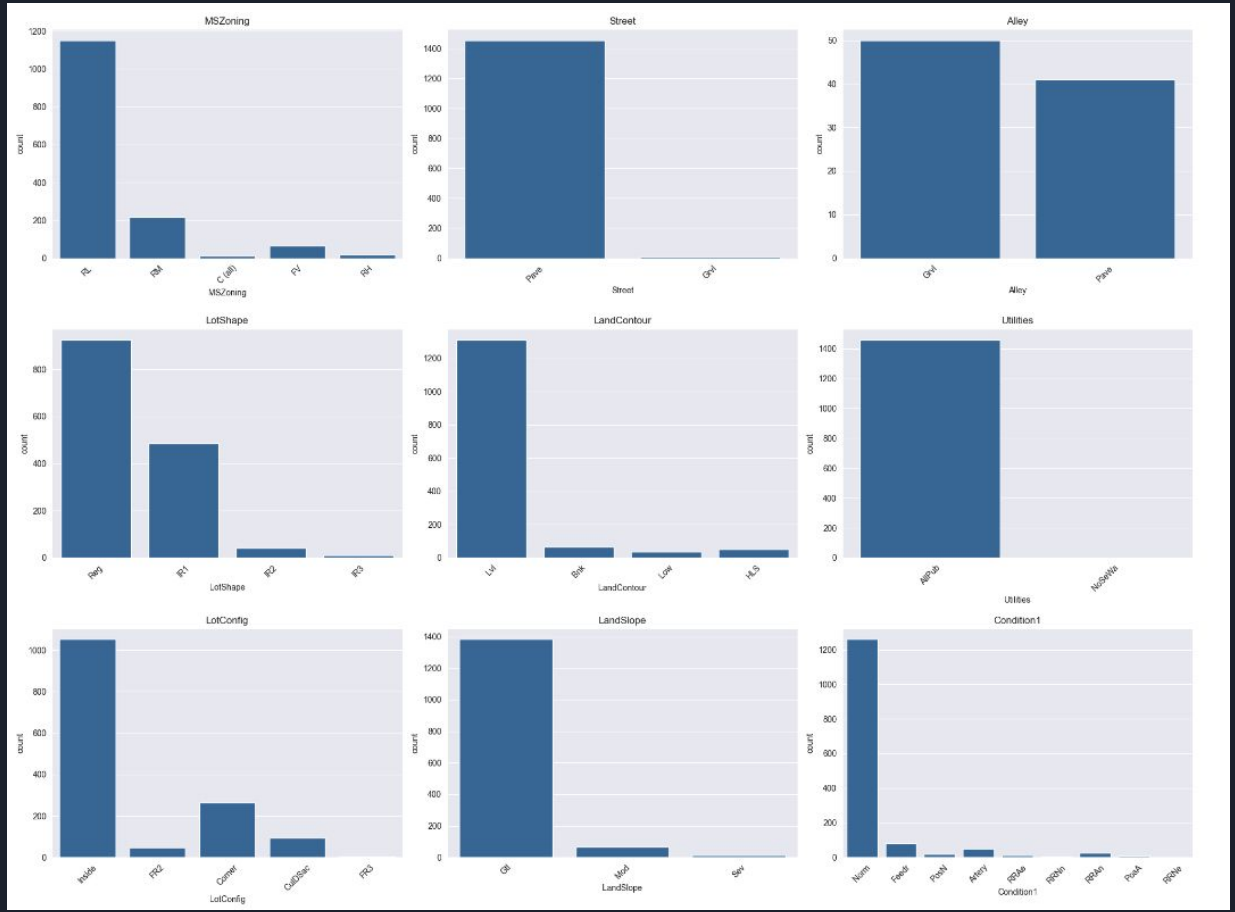
Keşifçi Veri Analizi

- Tüm Sayısal Özelliklerin dağılımı



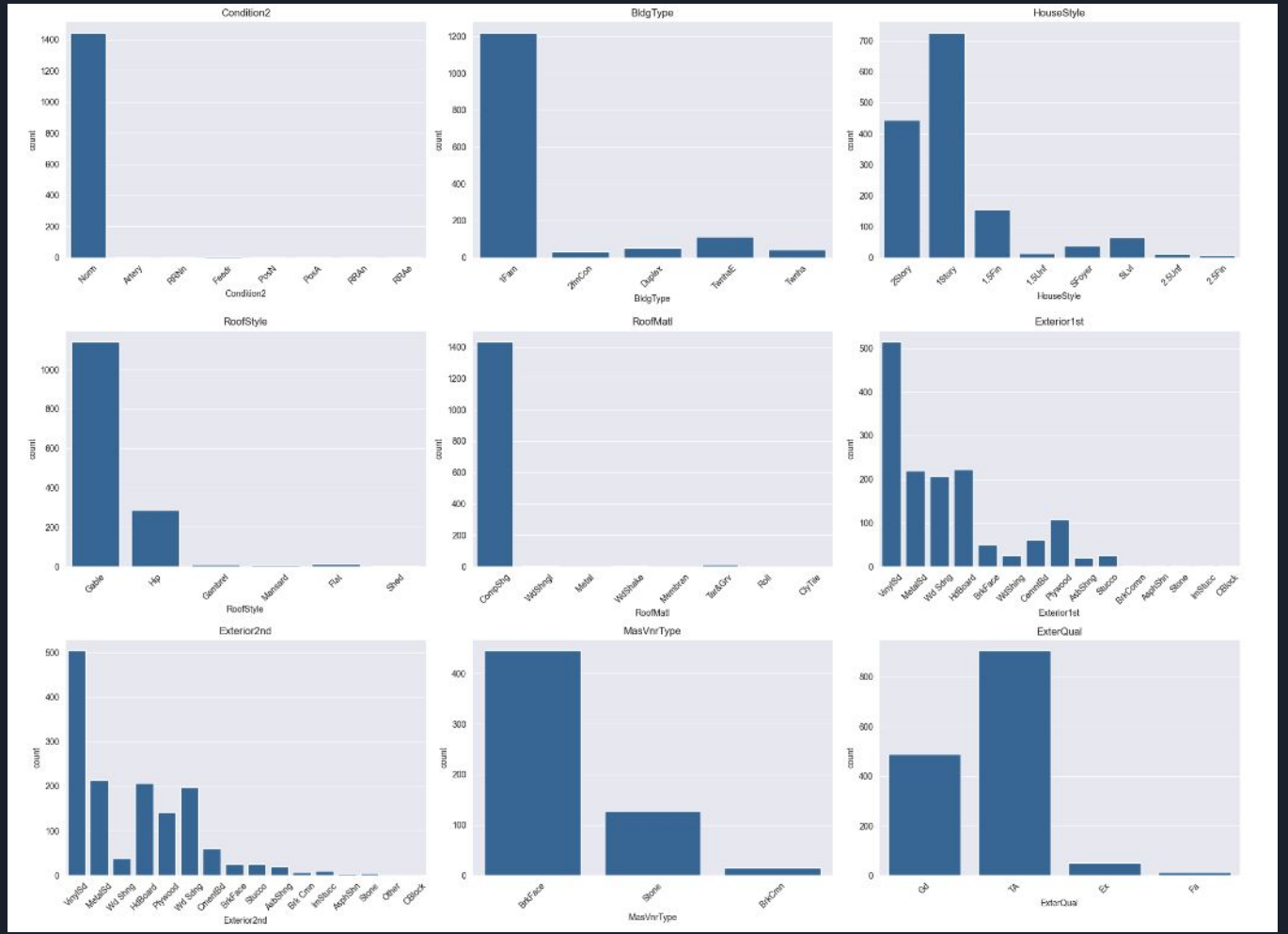
Keşifçi Veri Analizi

- Tüm Kategorik Özelliklerin dağılımı - 1



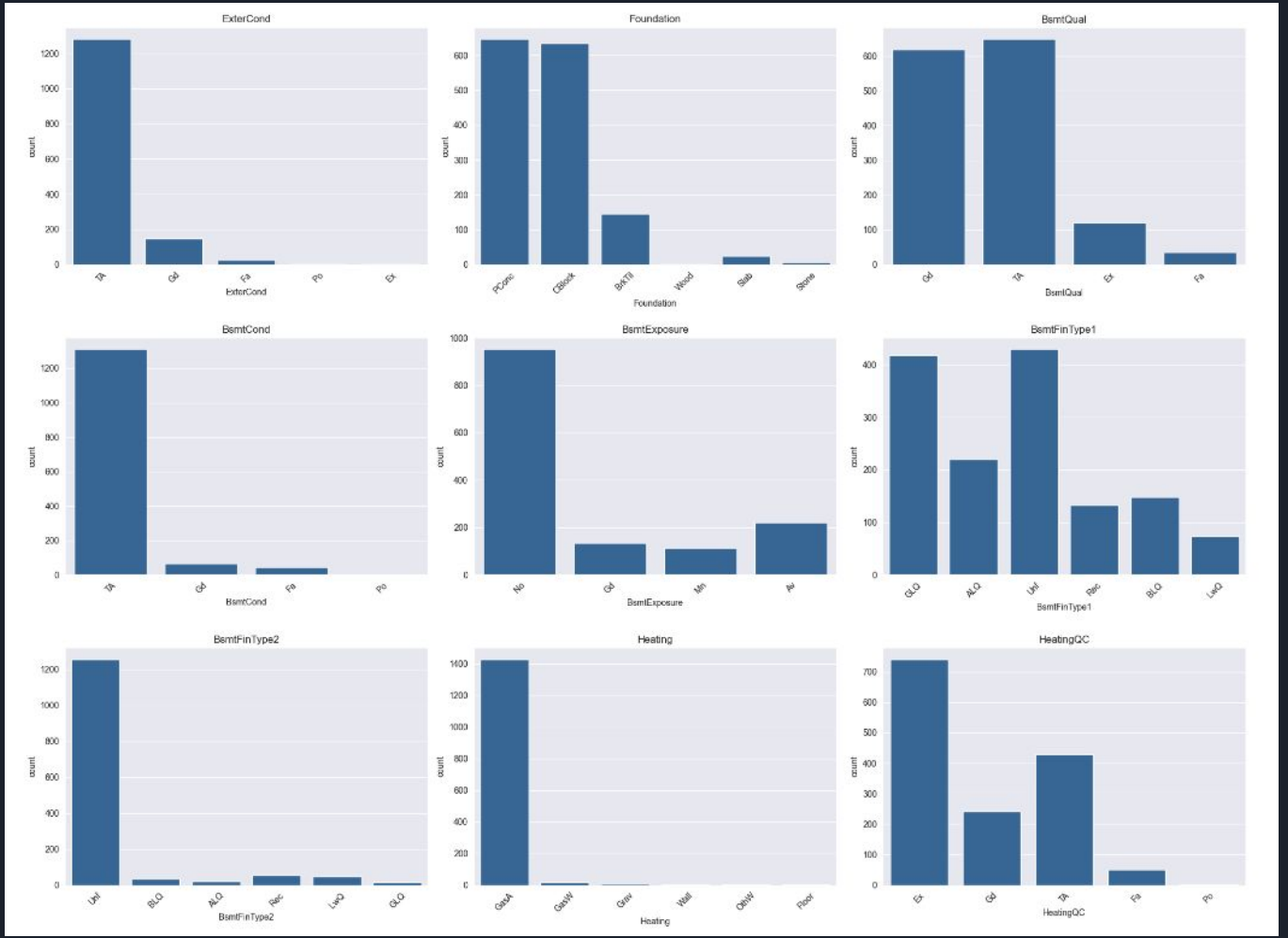
Keşifçi Veri Analizi

- Tüm Kategorik Özelliklerin dağılımı - 2



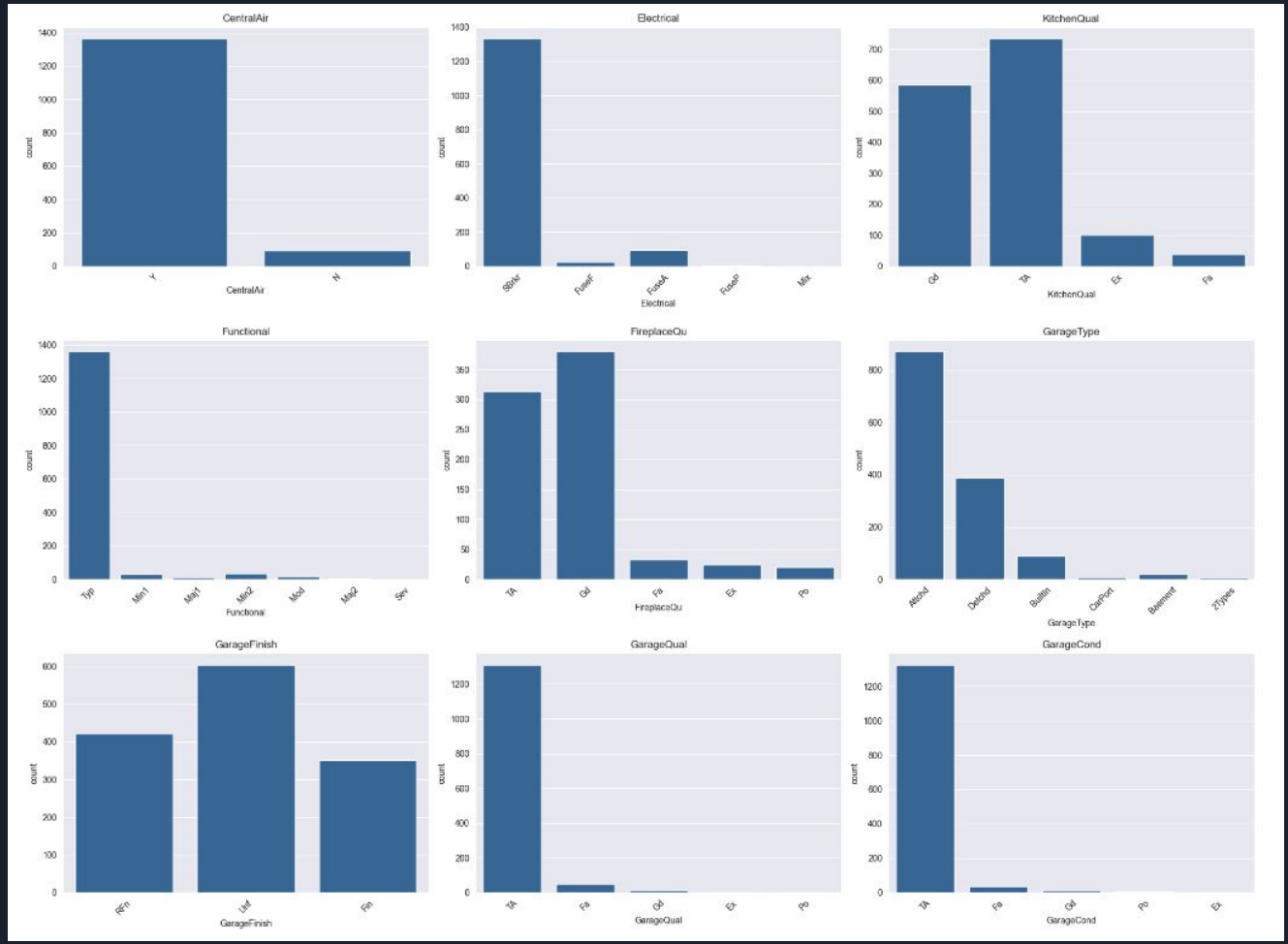
Keşifçi Veri Analizi

- Tüm Kategorik Özelliklerin dağılımı - 3



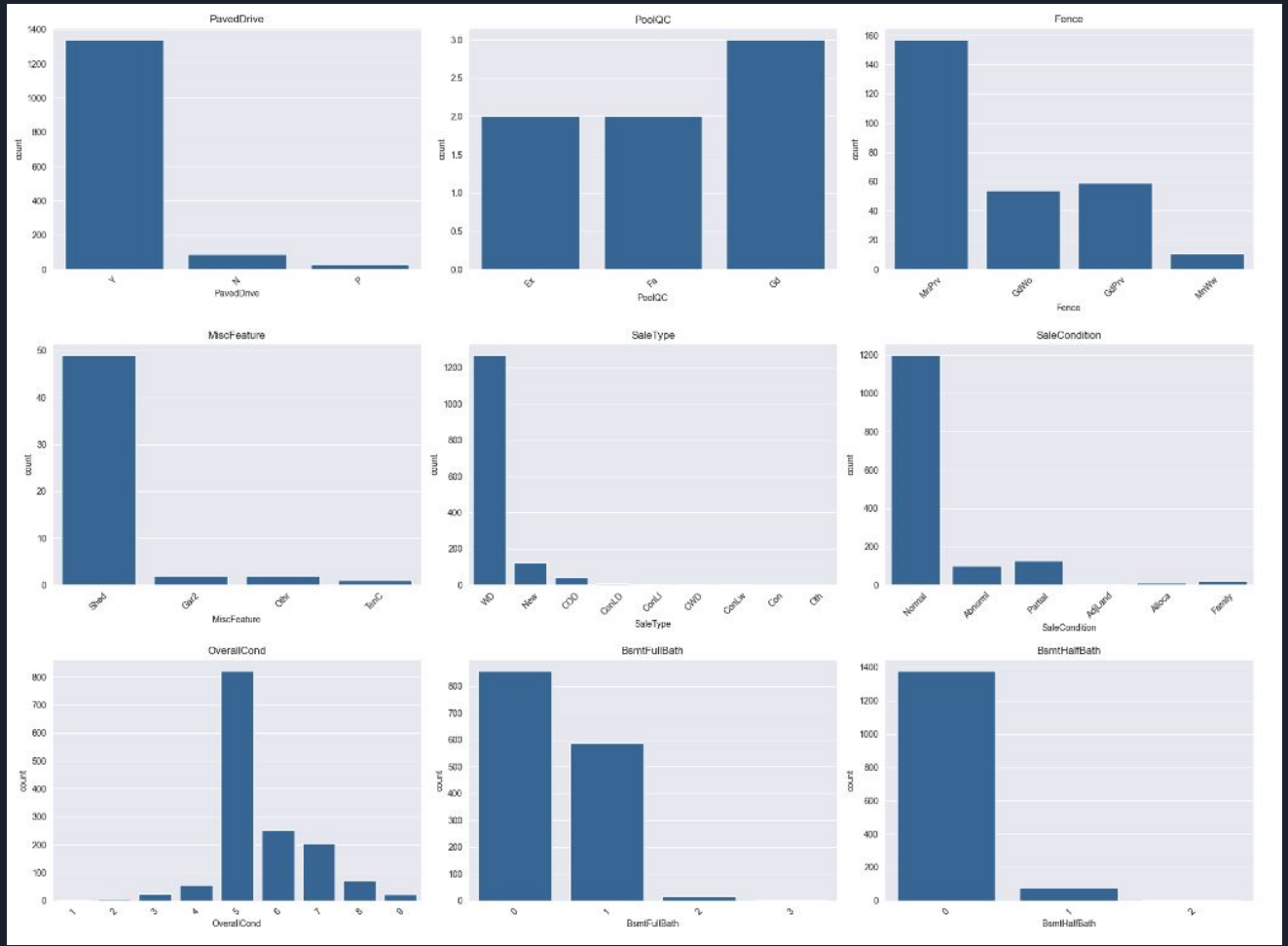
Keşifçi Veri Analizi

- Tüm Kategorik Özelliklerin dağılımı - 4



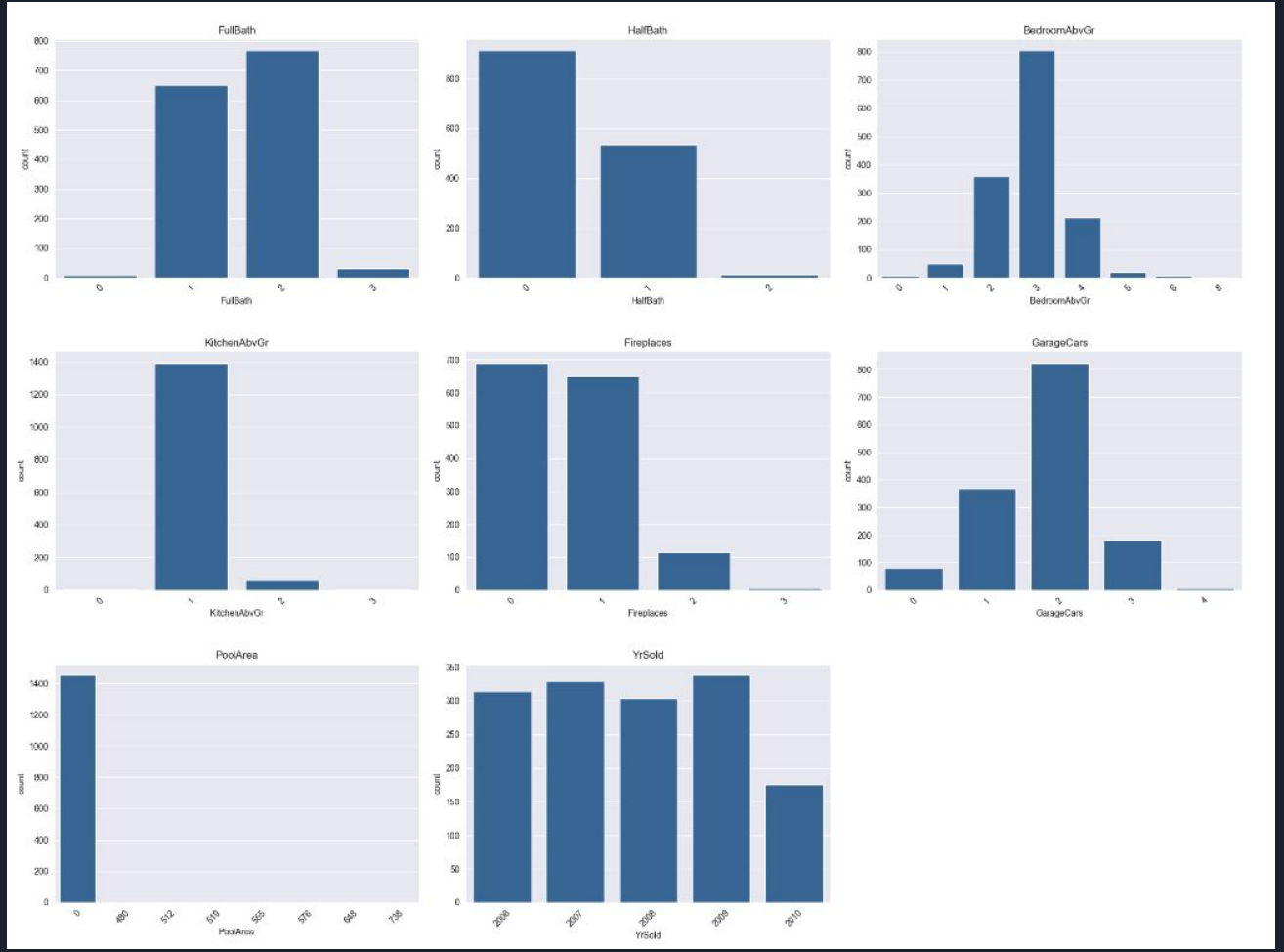
Keşifçi Veri Analizi

- Tüm Kategorik Özelliklerin dağılımı - 5



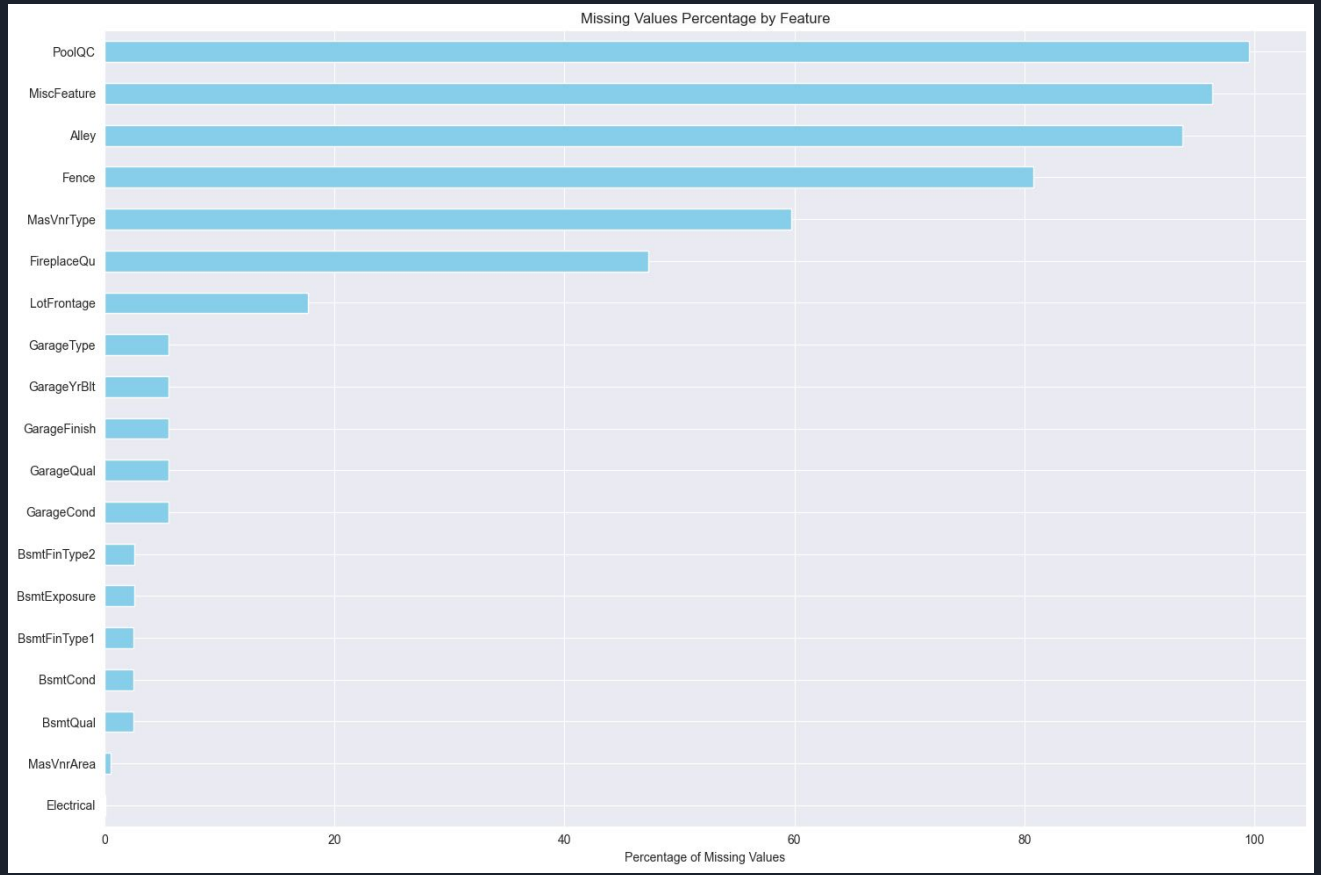
Keşifçi Veri Analizi

- Tüm Kategorik Özelliklerin dağılımı - 6



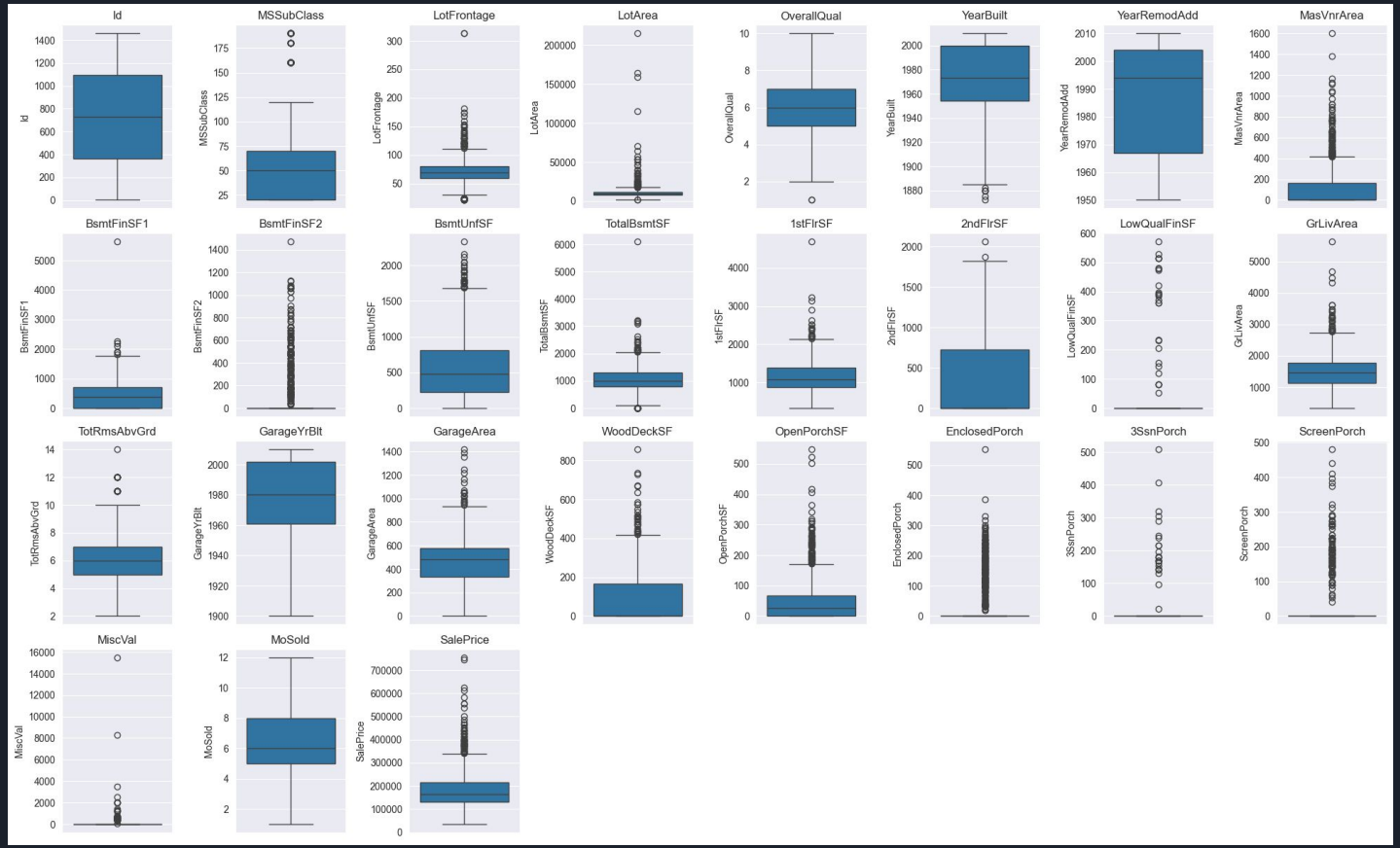
Keşifçi Veri Analizi

- Missing Values

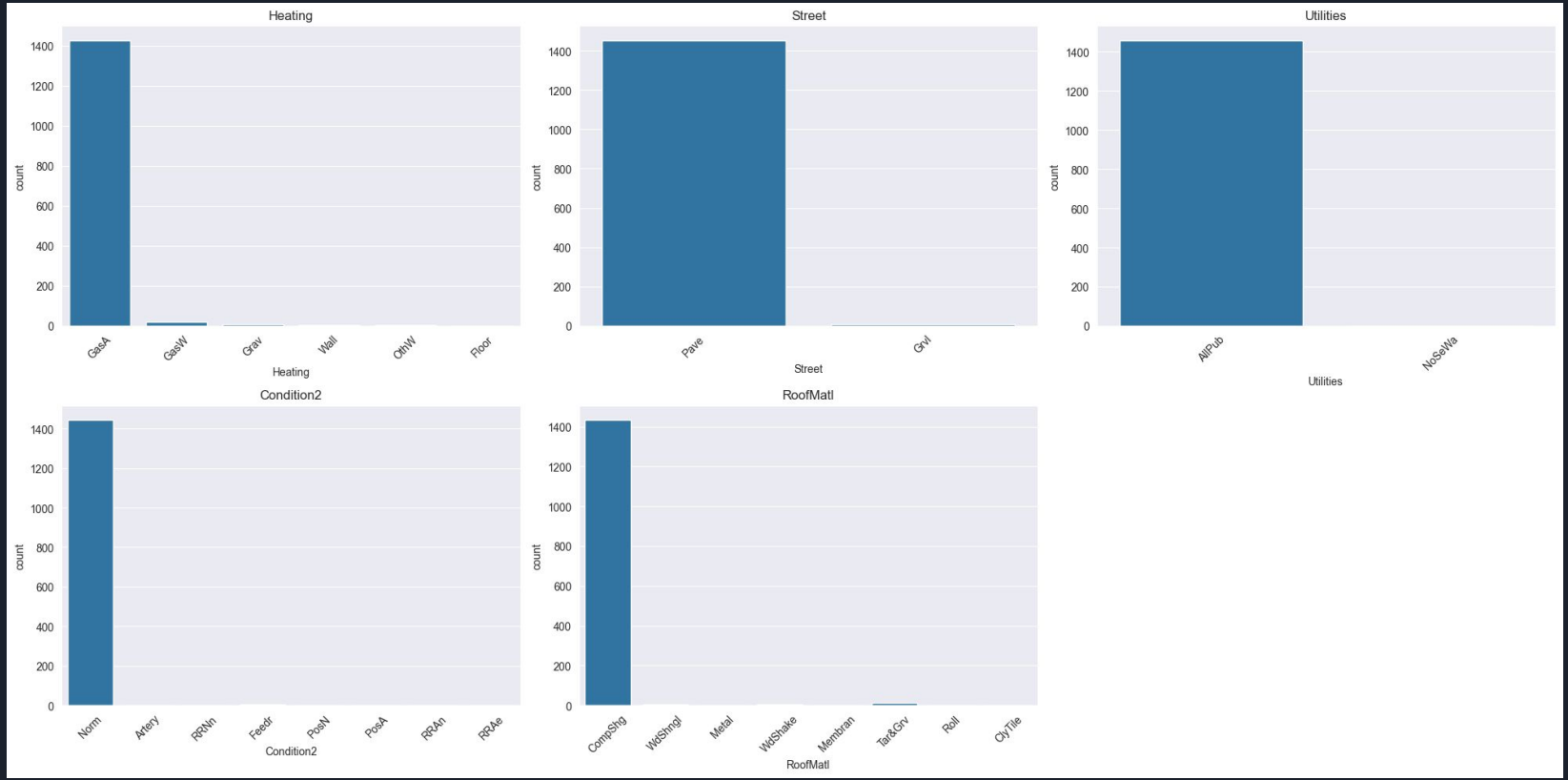


Keşifçi Veri Analizi

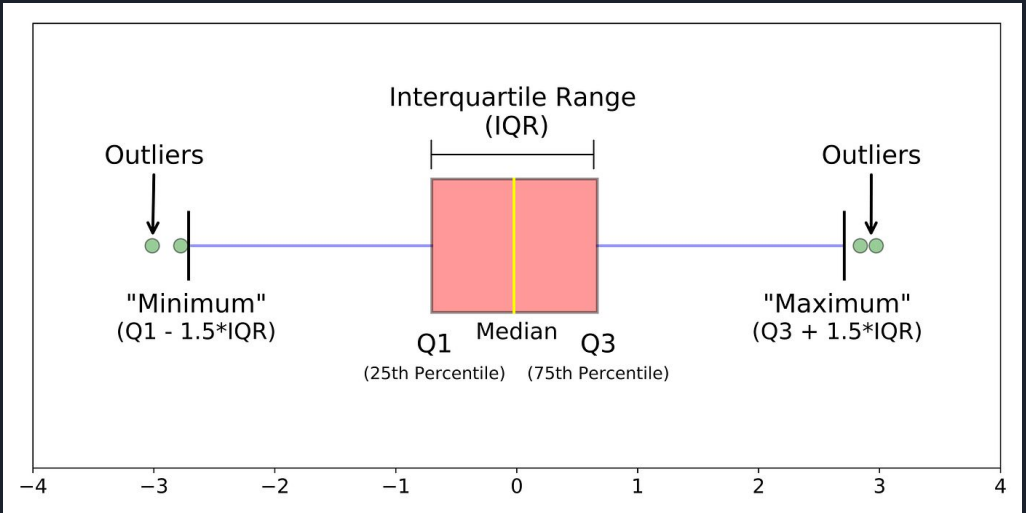
Boxplot Gösterimi



Veri Ön İşleme – Düzensiz Veriler



Veri Ön İşleme – Outliers

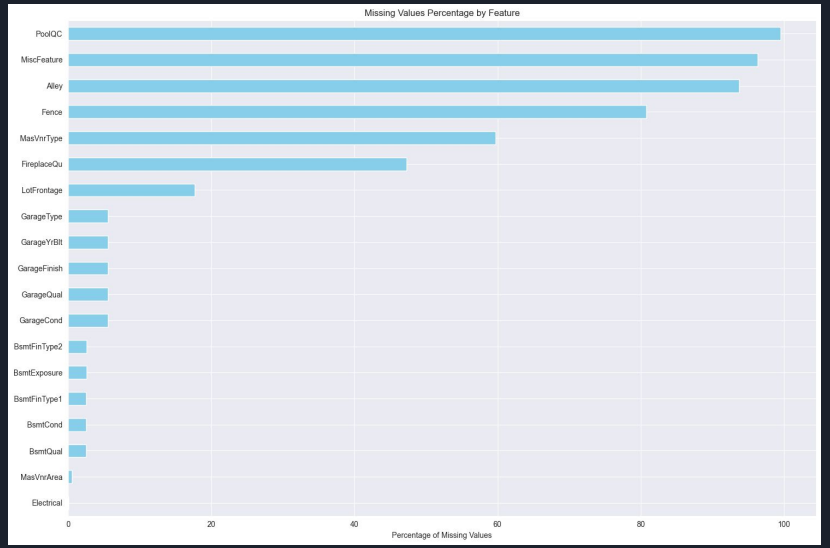



```
# Define the outlier threshold function
def outlier_thresholds(dataframe, col_name, q1=0.25, q3=0.75):
    quartile1 = dataframe[col_name].quantile(q1)
    quartile3 = dataframe[col_name].quantile(q3)
    interquantile_range = quartile3 - quartile1
    up_limit = quartile3 + 1.5 * interquantile_range
    low_limit = quartile1 - 1.5 * interquantile_range
    return low_limit, up_limit
```

Veri Ön İşleme – Missing Values

- %80 den yüksek eksik değer olan özellikler datasetten kaldırıldı.
  - 'Alley', 'PoolQC', 'Fence', 'MiscFeature'

- Sayısal null değerler median ile,
- Kategorik null değerler mod ile doldurulduu.
- Neighborhood özelliği kaldırıldı

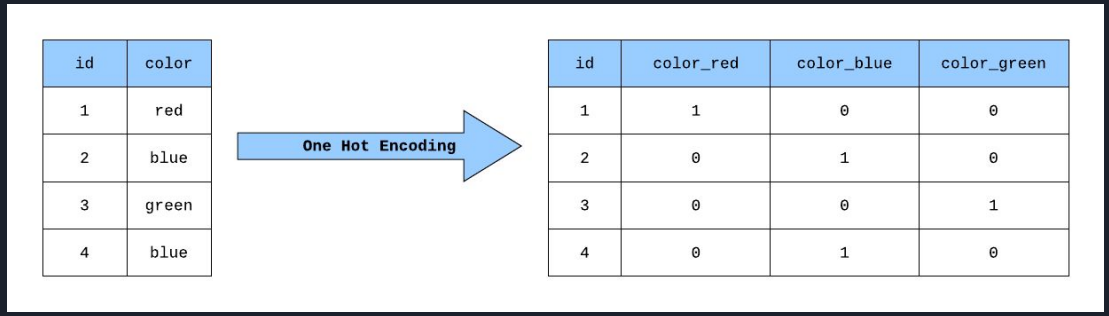


Veri Ön İşleme – Özellik Mühendisliği

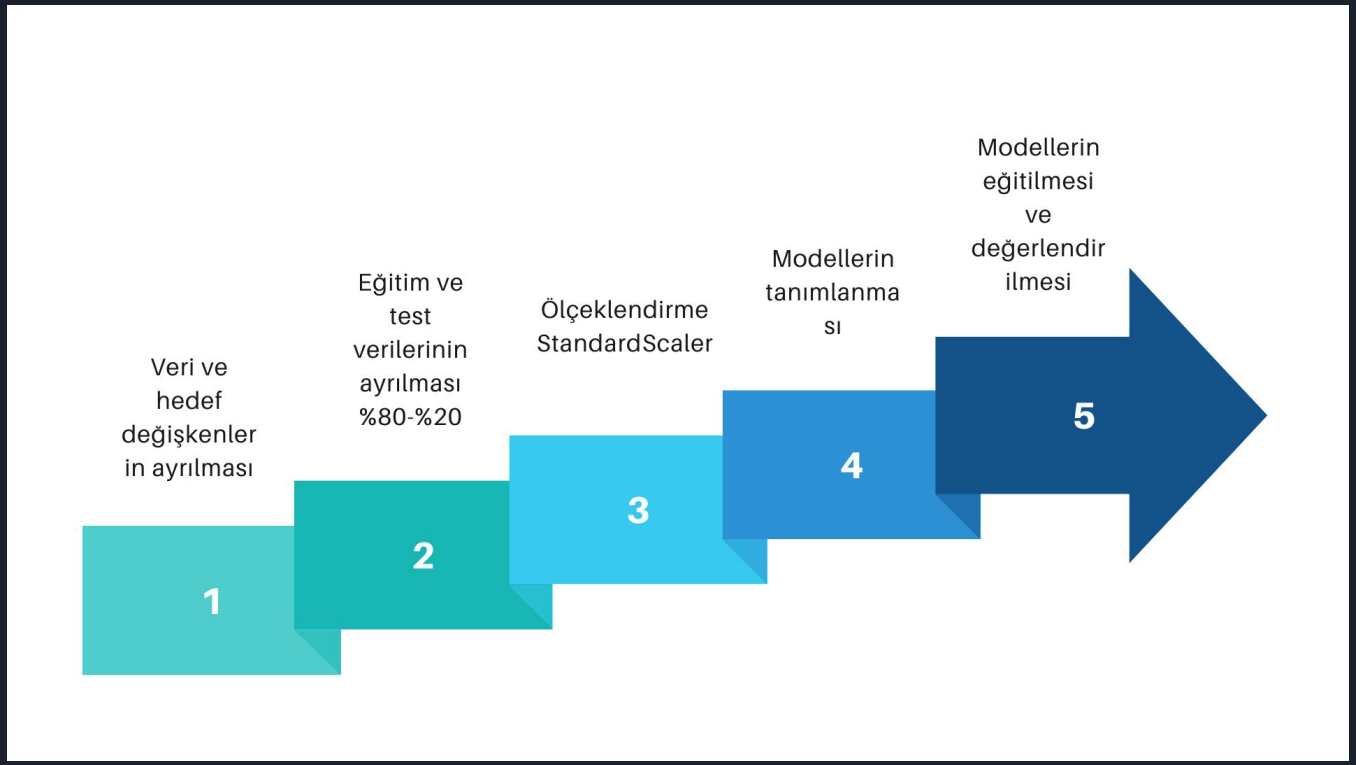
Yeni Özellik	Açıklama	Hesaplama
TotalLivArea	Yaşam alanı, bir evin toplam yaşam alanını ifade eder ve bodrum katı, zemin katı ve ikinci kat alanlarını içerebilir.	data['TotalBsmtSF'] + data['1stFlrSF'] + data['2ndFlrSF']
HouseAge	Evin yaşını hesaplamak, evin yeni veya eski olup olmadığını belirlemeye yardımcı olabilir.	data['YrSold'] - data['YearBuilt']
RemodAge	Evin en son yenilendiği tarihten itibaren geçen süreyi hesaplamak, yenilemelerin evin değerine etkisini anlamaya yardımcı olabilir.	data['YrSold'] - data['YearRemodAdd']
TotalBath	Evin toplam banyo sayısını hesaplamak, evin büyüklüğünü ve rahatlığını ifade edebilir.	data['FullBath'] + (0.5 * data['HalfBath']) + data['BsmtFullBath'] + (0.5 * data['BsmtHalfBath'])
TotalRooms	Evin toplam oda sayısını hesaplamak, evin büyüklüğünü ifade edebilir.	data['TotRmsAbvGrd'] + data['BedroomAbvGr']
GarageAge	Garajın yaşını hesaplamak, garajın evin yaşına oranla ne kadar yeni veya eski olduğunu belirlemek için kullanılabilir.	data['YrSold'] - data['GarageYrBlt']
TotalPorchSF	Evin toplam sundurma alanını hesaplamak, evin dış mekan olanaklarını ifade edebilir.	data['OpenPorchSF'] + data['EnclosedPorch'] + data['3SsnPorch'] + data['ScreenPorch']
LivAreaPerRoom	Toplam yaşam alanını toplam oda sayısına bölerek, her oda başına düşen yaşam alanını hesaplayabiliriz. Bu, evin ne kadar ferah olduğunu ifade edebilir.	data['TotalLivArea'] / data['TotalRooms']
GarageCarsPerArea	Garaj alanını garajdaki araba sayısına bölerek, garajın ne kadar verimli kullanıldığını gösterebiliriz.	data['GarageArea'] / data['GarageCars']
TotalConstructionArea	Toplam bodrum alanı, birinci kat alanı, ikinci kat alanı ve garaj alanını birleştirerek toplam inşaat alanını hesaplayabiliriz.	data['TotalBsmtSF'] + data['1stFlrSF'] + data['2ndFlrSF'] + data['GarageArea']
MonthlySeasonalIndex	Her ayın mevsimsel etkisini göstermek için bir indeks oluşturabiliriz. Bu, satışların yıl içindeki dağılımını ve mevsimsel etkilerini belirlemeye yardımcı olabilir.	12, 1, 2 -> 1 # Winter 3, 4, 5 -> 2 # Spring 6, 7, 8 -> 3 # Summer else -> 4 # Fall
AgeCategory	#Ev yaşını kategorilere ayırarak (yeni, orta yaşlı, eski gibi), evlerin yaşını daha anlamlı bir şekilde ifade edebiliriz.	age < 10 -> 'New' age < 50 -> 'MidAge' else -> 'Old'

Veri Ön İşleme – One-Hot Encoding

- kategorik verileri sayısal verilere dönüştürdük.



Modelleme

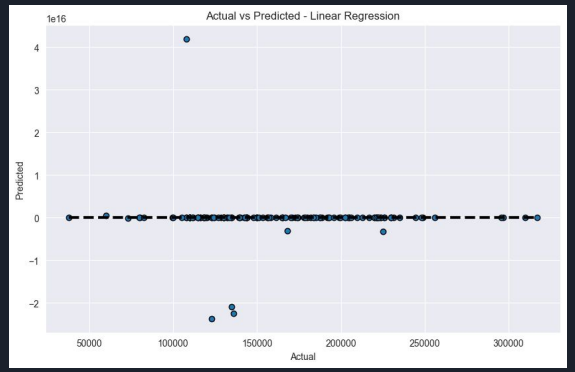

Veri ve hedef değişkenler in ayrılması
1
Eğitim ve test verilerinin ayrılması %80-%20
2
Ölçeklendirme StandardScaler
3
Modellerin tanımlanma sı
4
Modellerin eğitilmesi ve değerlendir ilmesi
5

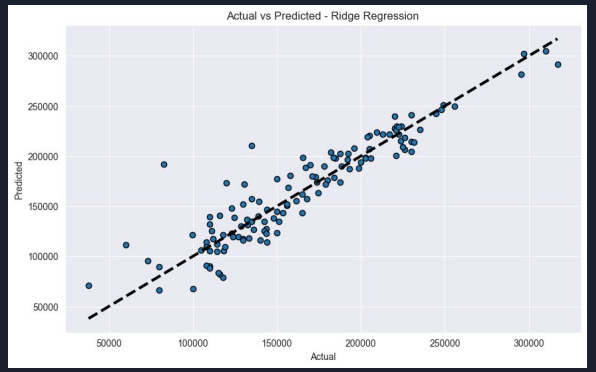
Modelleme

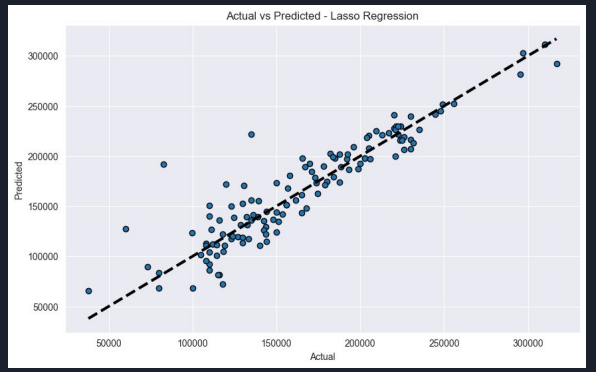
- **Linear Regression**: Sonuçlar anormal, model muhtemelen uygun değil.
- **Ridge Regression ve Lasso Regression**: İkisi de iyi performans gösteriyor, Ridge biraz daha iyi.
- **Decision Tree**: Orta düzeyde performans, genellikle Random Forest'a kıyasla daha düşük.
- **Random Forest**: İyi performans gösteriyor, ancak Ridge veya Lasso Regression kadar iyi değil.

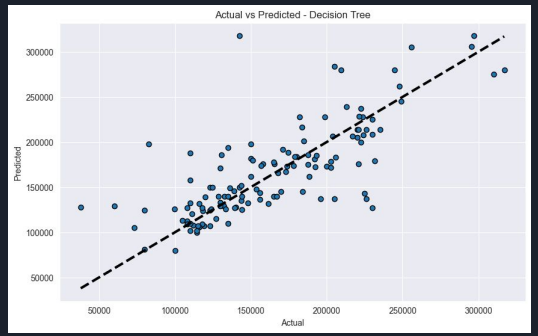
Model	MSE	R^2
Linear Regression	2.58E+31	-9.46E+21
Ridge Regression	4.22E+08	0.8454921
Lasso Regression	4.60E+08	0.8315289
Decision Tree	1.42E+09	0.4804507
Random Forest	5.61E+08	0.7944721

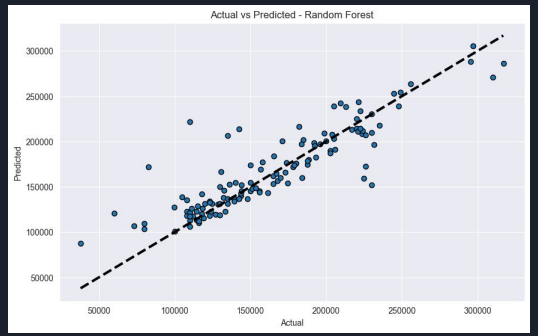
Modelleme – Tahmin Edilen ve Gerçek Fiyat Grafiği











Modelleme – Hyperparameter Tuning

- **Ridge Regression ve Lasso Regression**: İkisinde performans iyileşmesi oldu.
- **Random Forest**: Performansı negatif etkilendi.

Model	MSE	R^2
Ridge Regression	4.06E+08	8.51E-01
Lasso Regression	3.83E+08	0.859602
Random Forest	5.88E+08	0.784775

Modelleme - PCA

- **Linear Regression ve Decision Tree**: İkisinde de performans iyileşmesi oldu.
- Diğerlerinde performan negatif etkilendi

Model	MSE	R^2
Linear Regression	4.78E+08	8.25E-01
Ridge Regression	4.78E+08	0.825002
Lasso Regression	4.78E+08	0.825008
Decision Tree	7.15E+08	0.738187
Random Forest	5.25E+08	0.807643

Bulgular ve İş Önerileri

- Lasso Regression Modeli bu problem için en uygun olan model oldu.
- PCA bazı modeller için faydalı, bazıları için faydasız oldu.
- Dataset iyileştirmesi yapılabilir. Çok fazla etkisiz özellik bulunuyor.
- Tahmin modeli entegrasyonu ile hem satıcılar için karar verme süreçlerine katkıda bulunabilir, hem de alıcıların fiyat karşılaştırmasına gerek kalmadan alacakları yerleri piyasa değerinde almaları sağlanabilir.

Teşekkürlerr