KODLUYORUZ

HOUSE PRICE DATASET'İ

SEZGİSEL ANALİZ RAPORU

GRUP 1

Batuhan Özbay	Melisa Gündüz	Aybüke Akçay
Çiğdem Taş	Ali Şenyurt	Ercan Tuncay
Fuat Akdemir	Uğur Can Kıvanç	İzel Çelikkaya
Yiğit Hakverdi	Serenay Ardahanlı	Mehmet Özmen

Kodluyoruz ve İstanbul Büyükşehir Belediyesi, İstanbul Veri Bilimi Bootcamp kapsamında House Price Dataset üzerinde Sale Price değerinin tahminlenmesi için bir makine öğrenmesi modeli oluşturulması planlanmaktadır. Bu kapsamda, veri setinin sezgisel analizi gerçekleştirilmiştir.

Veri seti incelenirken ilk olarak kayıp değer olup olmadığı kontrol edilmiştir. Kategorik değişkenlerdeki kayıp olarak görülen değerler aslında bir kayıp değer değil, mevcut kategorik değişkenin özelliklerinin taşınıp taşınmadığını göstermektedir. Örneğin, PoolQC değişkenindeki kayıp değer olarak görülen değerler aslında o ev için havuzun bulunmadığını dolayısıyla da bir kalite olmayacağı anlamına gelmektedir.

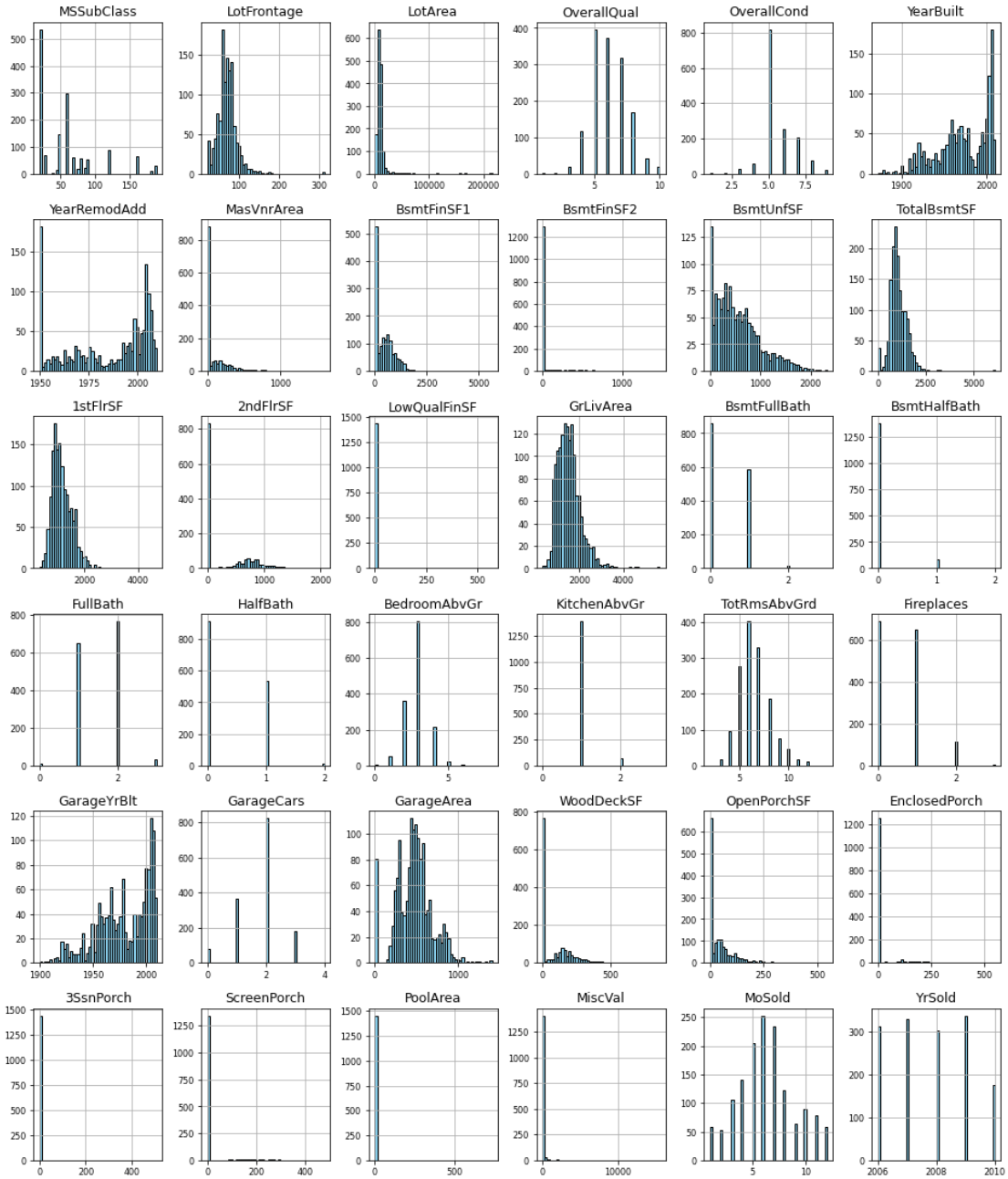
Aşağıda verilen değişkenlerde kayıp değer olarak görülen değerler kayıp değer değildir. Bu değerler 'None' olarak atanarak kayıp değerler yok edilebilir.

PoolQC	1453
MiscFeature	1406
Alley	1369
Fence	1179
FireplaceQu	690
GarageType	81
GarageCond	81
GarageQual	81
GarageFinish	81
GarageYrBlt	81
BsmtFinType2	38
BsmtExposure	38
BsmtFinType1	37
BsmtQual	37
BsmtCond	37
MasVnrArea	8
LotFrontage	259

Aşağıda verilen değerlerin karşısında kayıp değer sayısı verilmiştir.

MasVnrType	8
Electrical	1

Numeric değerlerin histogramları, buradan değerlerin discrete valuelarının dağılımına göre hangi kolonların kendi içinde baskın davrandığı ve düşürülmesi gerektiği görülebilir. Örnek olarak ‘LowQualFinSF’ histogramına bakılabilir.



Veri seti detaylı bir şekilde incelendiğinde aşağıdaki bulgulara ulaşılmıştır:

MSSubClass değişkeninin türü numerik olarak gösterilmiştir ancak değişken kategorik değişkendir. Evler yapıldıkları tarihlere göre gruplandırılmışlardır. MSSubClass değişkeninin değeri arttıkça evin yapıldığı tarih de günümüze yaklaşmıştır. Herhangi bir encoding gerekmeden numerik hali ile modelde kullanılabilir.

Bazı değişkenler çok yakın değerlere sahip olduğundan tek değer olarak ele alınabilir. Birlikte kullanılmak üzere birleştirilecek değerler seçilirken yüksek korelasyona sahip olup olmadıkları incelenecektir.

Overall Quality yükseldikçe ev fiyatları da yükselmiştir. Bu durum Overall Condition değişkeni için geçerli değildir.

Yearbuilt ve Yearremodeled değişkenleri SalePrice değişkeninde benzer bir artışa sebep olmuştur. 2000 ve sonrası için değerler yüksektir. 2000-2006 yılları arasında evlerde daha çok tadilat yapılmıştır. O dönem evlerde tadilat için bir talep olduğu söylenebilir.

Evler %99.0 ile Condition2 durumunda ve evlerin %49.7si tek katlı; %30.5i ise 2 katlı. Evlerin %83.6sı tek ailelik evler / sonra %7.8 ile ikiz evler geliyor

Year remodeled — neighborhood değişkenleri birbirleri ile yüksek oranda ilişkilidir. İyi mahallelerde insanların evlerine tadilat yaptırmayı tercih ettiklerini söyleyebiliriz.

Saleprice'ın median değeri 163.000'dir.

KitchenQual değişkeni OvrQual ve Saleprice ile koreledir. Model içinde kullanılmaya uygun olduğu söylenebilir. Benzer şekilde TotRmsAbvGrd değişkeni BedroomabvGrd ve Saleprice ile koreledir. Değişken unimodal bir dağılıma sahiptir. Model kurulurken kullanılması tercih edilebilir.

Functional değişkeninin %93'lük kısım typical ve herhangi bir değişkenle korele değil, kullanılması tercih edilmeyebilir. Ancak hasarlı evlerin fiyatlarının düşük olabileceği hipotezinden yola çıkılarak kullanılması tercih edilebilir.

FirePlaces değişkeninin modelde kullanılmaya uygun bir dağılımı vardır. Aynı zamanda SalePrice ile korele olduğu için kullanılabilir.

FireplaceQu: Missing value olmayan değerler kullanılabilir.

GarageType değişkeni Garagefinish değişkeni ile koreledir. Ancak kullanılması hakkında soru işaretleri mevcuttur.

GarageYrBlt değişkeni SalePrice'la koreledir. Bu değişken GarageCars ve GarageArea değişkenleri ile de koreledir. Kullanılması tercih edilebilir. Kullanıldığı takdirde yüksek korelasyona sahip olduğu diğer değişkenlerle birleştirilerek kullanılabilir.

GarageFinish değişkeni kategorik değişken olduğu için satış fiyatıyla olan korelasyonu değerlendirilememiştir. Ayrıca kategorideki 'No Garage' seçeneği missing value olarak işlenmiştir, bu hata düzeltilerek modellemeye ekleme yapılabilir.

GarageCars değişkeni nümerik değişken tipinde tanımlanmıştır. Satış fiyatıyla yüksek korelasyona sahiptir.

GarageArea değişkeni SalePrice ile yüksek korelasyona sahip nümerik değişkendir.

GarageQual ve GarageCond değişken tipleri çok yüksek korelasyona sahiptir bu sebeple beraber tek bir değişken olarak incelenebilirler.

PavedDrive, Paved değişkeninde yığılma olan kategorik değişken türü. Drop edilebilir.

Alley değişkeninde NA olarak adlandırılmış bir kategori de mevcuttur. Bu nedenle, eksik değer bulunmamaktadır. Değişken NA dışında iki tür kategori bilgisi içermektedir: paved ve gravel(çakıl).

LotFrontage değişkeni "Linear feet of street connected to property" olarak tanımlanmıştır. Bu tanımdan yola çıkılarak LotFrontage eksik değerleri bu tanıma göre LotFrontage eksik değerlerin bulunması mümkün değil gibi görünüyor

WoodDeckSF (Wood deck area in square feet) değişkeninin Phik korelasyonu sadece "PoolQC" ile bulunmaktadır. Wood Deck Area olmayan evler 0 değeri almıştır. 761 adet 0 değeri bulunmaktadır

OpenPorchSF: (Open porch area in square feet) değişkeninin Phik korelasyonu Condition2:0.51 BsmtFinSF1:0.59 GrLivArea:0.56 değerleri almaktadır. Open Porch Area bulunmayan evler 0 değeri almamıştır. (656 adet)

EnclosedPorch (Enclosed porch area in square feet) değişkeninin Phik korelasyonu PoolArea:0.51.Enclosed porch area olmayan evler 0 değeri almıştır. 1252 adet 0 değeri vardır.

PoolArea (Pool area in square feet) değişkeninin Phik korelasyon GrLivArea: 0.76, MiscFeature: 1.00. 1453 adet evde havuz yoktur.

PoolQC değişkenine ait bilgiler şu şekilde özetlenebilir: Pool quality Ex Excellent= 2 Gd Good :3 TA Average/Typical:0 Fa Fair : 2 adet NA No Pool: 1453 Null(NA) değeri yani havuz olmayan ev sayısı 1453 adettir. NA değerler eksik veri değil havuz olmamasıdır.

PoolQC Phik korelasyon değerleri : LotConfig: 0.81 Neighborhood: 1.00 HouseStyle: 0.80 YearBuilt: 0.63 YearRemodAdd: 0.83 MasVnrType: 0.92 ExterQual: 0.81 BsmtExposure: 0.72 BsmtFinSF1: 0.63 BsmtFinSF2: 0.65 BsmtUnfSF: 0.82 GrLivArea: 0.83 BedroomAbvGr: 0.81 FireplaceQu: 0.67 GarageYrBlt: 0.83 GarageFinish: 0.87 GarageArea: 0.68 WoodDeckSF:0.67 SaleCondition: 0.68 SalePrice: 0.63

ExterQual (Dış cephedeki malzemenin kalitesini değerlendirir): 17 farklı başlıkla yüksek korelasyonlu (ExterCond dahil değil). %62.1'i Tipik, %33.4'ü İyi seviyede.

ExterCond (Malzemenin dış cephedeki mevcut durumunu değerlendirir): OverallCond ve MiscFeature ile yüksek korelasyonlu. %87.8'i Tipik, %10'u İyi seviyede.

Foundation (Yapı Temeli: Yapı temeli türü): 12 alanla yüksek korelasyonlu. %44.3'ü Dökme Beton, %43.4'ü Cinder Bloğu (bir çeşit tuğla), %10'u ise Tuğla, Fayans'tan oluşmaktadır.

BsmtQual (Bodrumun yüksekliğini değerlendirir): 16 alanla yüksek korelasyonlu. %44.5'i Tipik, %42.3'ü İyi, %8.3'ü Çok İyi seviyede. NA'ları (No basement) missing olarak görülmektedir. Ancak missing değer değildir.

BsmtCond (Bodrumun genel durumunu değerlendirir): OverallQual ve OverallCond ile yüksek korelasyonlu. %89.8'i Tipik. NA'ları (No basement) missing olarak görülmektedir. Ancak missing değer değildir.

BsmtExposure (Walkout veya bahçe seviyesindeki duvarları ifade eder): PoolQC ile yüksek korelasyonlu. %65.3'ü No Exposure, %15.1'i Average Exposure, %9.2'si Good Exposure'dur. NA'ları (No basement) missing olarak görülmektedir. Ancak missing değer değildir.

BsmtFinType1 (Bodrum bitmiş alanın değerlendirmesi): 7 alanla korelasyonu var. %29.5'i Unfinshed, %28.6'sı Good Living Quarters, %15.1'i Average Living Quarters, %10.1'i Below Average Living Quarters, %9.1'i Average Rec Room'dan oluşmakta. NA'ları (No basement) missing olarak görülmektedir. Ancak Missing değildir.

Evin içerisinde bulunan bodrum ile ilgili birden fazla değişken mevcuttur. Bu değişkenlerden SalePrice ile en yüksek korelasyona sahip olanının modelde kullanılması yeterli olacaktır. Modelle ilgili yapılacak geliştirmelerde birden fazla bsmt değişkeni birleştirilerek modele eklenebilir.

HeatingQC, CentralAir, Electrical değişkenleri kategorik değişkenlerdir.

Electrical kolonunda sadece 1 adet null değer bulunmaktadır. Bu null değerin etkisi olmadığından en çok bulunan değere atanmasında bir problem oluşmayacaktır.

Ancak CentralAir ve Electrical kolonlarındaki değişkenler histogramda incelendiğinde ayrık değişkenlerin dağılımının %90 ‘nından fazlasının ayrık değişkenlerden sadece birine kaymış

olması, yani diğer değerler yüksek oranda baskılanması, bu kolonların SalesPrice üzerinde etkisinin çok az olacağını gösterir ve bu kolonların drop edilmesi gerekmektedir.

CentralAir Bool değerden oluşan, %93.5 ‘i ‘True’ olan bir kolon, SalePrice’a etkisi düşük bir değişkendir.

Electrical 5 ayrık değerden oluşan, %91.4 ‘ü ‘SBrkr’ olan bir kolon, SalePrice’a etkisi düşük bir değişkendir.

HeatingQC değişkeni 5 ayrık değerden oluşuyor. Ancak HeatingQC böyle bir durum olmadığında labeling yapılarak devam edilebilir.

1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea değişkenleri numerik değişkenlerdir.

1stFlrSF, 2ndFlrSF, GrLivArea kolonlarının SalesPrice ile yüksek korelasyon ilişkisi bulunan kolonlardır.

Ancak LowQualFinSF ise SalesPrice le korelasyon ilişkisi yoktur. Histogramında ise %98.2 oranında ‘0’ değerinde bulunmakta ve bu yüzden drop edilmesi gerekmekte.

Full bathroom: “lavabo, tuvalet ve küvet veya duş içeren banyo”
Half bathroom: “lavabo ve tuvalet içeren ancak küvet veya duş içermeyen banyo” şeklinde tanımlanmaktadır.

BsmtFullBath değişkeni incelendiğinde, bodrum katındaki full banyo sayısı %58.6’sında 0, %40.3’ünde 1, kalanlarda 2 veya 3 olduğu görülür. BsmtFullBath SalePrice ile Pearsons correlation: 0.23’tür.

BsmtHalfBath değişkeni incelendiğinde, bodrum katındaki yarım banyo sayısı %94.4’ünde 0, kalanlarda 1 veya 2 olduğu görülür. BsmtHalfBath SalePrice ile Pearsons correlation: -0.02’dir.

FullBath değişkeni incelendiğinde, bodrum katı haricindeki full banyo sayısı %52.6’sında 2, %44.5’inde 1, kalanlarda 0 veya 3 olduğu görülür. FullBath değişkeni SalePrice ile yüksek korelasyona sahiptir. Pearsons korelasyonları 0.56’dır.

HalfBath değişkeni incelendiğinde, bodrum katı haricindeki yarım banyo sayısı
%62.5’inde 0, %36.6’sında 1, kalanlarda 2 olduğu görülür. HalfBath SalePrice ile Pearsons correlation: 0.28’dir.

BedroomAbvGr değişkeni incelendiğinde Bodrum katı haricindeki yatak odası sayısı %55.1’inde 3, %24.5’inde 2, kalanlarda 0, 1, 4, 5, 6, 7, 8 değerleri görülür. BedroomAbvGr SalePrice ile Pearsons correlation: 0.17’dir.

KitchenAbvGr değişkeni incelendiğinde bodrum katı haricindeki Mutfak sayısı % 95.3'ünde 1, kalanlarda 0, 2, 3 olduğu görülür. KitchenAbvGr SalePrice ile Pearsons correlation: -0.14'tür.

Fence degiskeni %80'nin ustunde N/A'dir, bu da 'No Fence' anlamina geliyor. Verilerin cogunlugunda N/A oldugu icin bu degiskten modelde tercih edilmemesi daha doğru olabilir.

MiscFeature, her bir verinin çeşitli özelliklerini içeriyor; bunlar, asansör, ikinci bir garaj var mi veya tenis kortu var mi şeklinde ayrılmıştır.

MiscVal, her bir miscellaneous değerini içeren özellik; bu özelliğin kullanılması çok doğru olmayabilir çünkü bu özellikte çok fazla 0 değer var.

MoSold, Hangi ay satildigi

YrSold, Hangi yil satildigi

SaleType, Bu feature satisin tipini iceren veriler bulunuyor. (Kategorik) Verinin dagilimina bakildiginda en fazla Warranty Deed bulunmaktadir tahmin yapilmadan once veriler bu feature gore ayiklanip bunun WD uzerinden yapilabilir

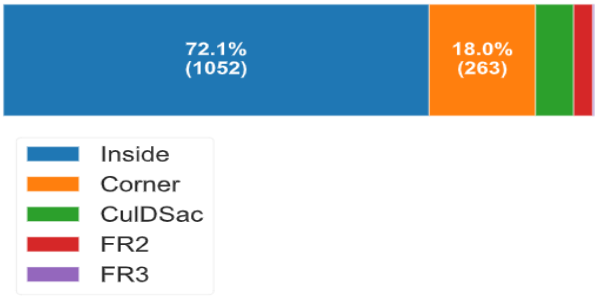
SaleCondition Ayni sekilde satisin kosulu, en fazla Normal deger icermektedir (Kategorik degisken) SaleType ile yuksek korelasyona sahiptir. Veriler preprocessing isleminden gecirilmedne once bu iki degisken goz onune alinarak bir ayiklama yapilabilir ve model bu iki kategorik degerin uzerine kurulabilir

Landcontour: Binanın bulunduğu arazinin düzlüğünü ifade etmektedir. Missing value bulundurmamaktadır. Neighbourhood ile yüksek korelasyon içermektedir. Landcontour için dört derece atanmıştır lvl(zemin seviyesi) binaların yüzde 89.8'inin toplandığı alandır. Diğer dereceler banked,hillsed ve low'dan oluşmaktadır.Missing value yoktur.

Utilities: Binada ulaşılabilir kamu hizmetlerini ifade etmektedir. Binaların %99.9'u tüm kamu hizmetlerine ulaşılabilmektedir. Yüksek korelasyonu olan hiçbir değişken bulunmamaktadır. (Bence bu veriyi göz önünde bulundurmamalıyız.)

LotConfig: Parsel yapısı

Mülkün cephelerini tanımlayan bir değişkendir. Bu değişkenler inside,corner, cu-de-sac(çıkmaz), FR2 (mülkün iki tarafı cephe ve FR3(mülkün üç tarafı cephe)den oluşmaktadır. LotConfig PoolQc ile yüksek korelasyon göstermektedir.Missing value yoktur



LandSlope: Arazinin eğimi

LotArea, Roofstyle ve Neighbourhood ile yüksek korelasyon içindedir. (belki bu değişkenlerle birlikte incelenebilir) Gentle, moderate ve severe olmak üzere üç dereceye ayrılmıştır. Binaların büyük bir kısmı (%94.7) Gentle slope'tur. Missing value yoktur.

Neighbourhood: 36 değişkenle birden yüksek korelayon içerisindedir. (Saleprice da buna dahildir.) Missing value yoktur.