Hafta 2 Ödevi

Grup 3 Sezgisel Analiz Raporu

Tarih:22/05/2022

Grup Üyeleri:

Arif Emre Karaduman

Baranalp Özkan

Eren Güneştaş

Furkan Güneştaş

Kumru Orkun

Mehmet Yavuz Gökmen

Muhammet Enes Bol

Ömer Topcu

Tunahan Demirkol

Yasin Tarakçı

House Price verisinde değişkenler incelendi ve elenmesi gerekenlere karar verildi. Kayıp veriler, outlierlar, istatistiksel dağılımlar, karşılıklı korelasyonlar incelendi ve alınabilecek aksiyonlar değerlendirildi.

İçindekiler

MSSubClass: Identifies the type of dwelling involved in the sale.

--> Kategorik Yapılacak

MSZoning: Identifies the general zoning classification of the sale.

-->Sadece 5 Kategori Var Datada

-->Sale price grafiğine göre gruplandırma yapılıp kategoriler azaltılabilir.

LotFrontage: Linear feet of street connected to property

-->%17 Missing_Value'lar mevcut. Yüksek Korele Olan Lotarea Baz Alınarak Doldurulabilir

LotArea: Lot size in square feet

-->Çok Yüksek Outlıer Değerler Var
-->Sale price grafiğine göre gruplandırma yapılıp kategoriler azaltılabilir.

Street: Type of road access to property

-->%99.6 Pave. Column Atılsın Mı???

Alley: Type of alley access to property

-->%93.8 Mıssıng. No Alley Şeklinde Doldurulmalı.

LotShape: General shape of property

--> Frekansı Düşük Olanları Tek Grup Yapmak Düşünülebilir

LandContour: Flatness of the property

--> Neighborhood İle High Korele

Utilities: Type of utilities available

--> Sil %99.9

LotConfig: Lot configuration

--> Saleprıce'a Karşı Korelasyon Ve Faktor Analızı Yapılabilir.

LandSlope: Slope of property

--> Landslope Ve Saleprice Grafiklerini Karşılaştırarak Eğer Korelasyon Yoksa Kategorilerin Birleştirilmesi Düşünülebilir. 11/81 Kolon Tamamlandı.

Neighborhood: Physical locations within Ames city limits

--> Saleprice Ve 35 Feature İle High Correlation Bulunuyor, Missing Yok.

Condition1: Proximity to various conditions

--> Norm and Other olarak sınıflandırılması düşünüldü.

--> Elimizdeki normal dağılmayan veri var bunları ggruplamak veri kaybetmiş olacağız.

Condition2: Proximity to various conditions (if more than one is present)

-->6 kolon ile yüksek korelasyon

-->%99 tek kategoriye yığılma var.

BldgType: Type of dwelling

--> BldgType ve HouseStyle featureları Salesprice'a aynı etkiyi göstereceğini düşürüp birleştirilmesini düşündük.

HouseStyle: Style of dwelling

-->8 kolon ile yüksek korelasyon var.

-->Kategorik veridir, SalePrice kolonuna göre faktör analizi yapılabilir.

OverallQual: Rates the overall material and finish of the house

--> Normal dağılıma çok yakın, SalePrice ile yüksek korelasyonu var bu feature çok değerli.

OverallCond: Rates the overall condition of the house

--> Normal dağılıma OverallQual gibi yakın, bu sebeple 2sinden birini seçmek gerekebilir.

--> OverallQual salesprice ile daha korele olduğu için bu feature kullanılmayabilir.

YearBuilt: Original construction date

--> Ggrafikte 3 farklı dağılım gözlenmiştir ve bu dağılımların birbrinden ayırılması düşünülmüştür.

--> 1900 öncesi inşa edilen yapılar outlier ggibi gözüküyor çöpe atılması düşünüldü ama emin değiliz. :D

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

--> 1950 yılı için bir anomaly bulunuuyor. 1950 ve öncesi eksik, atabiliriz ya da none şeklinde değiştirebiliriz.

--> Bu verilerin ne olduğu belli olmadığı için sağlıklı bir analiz yapılamamakta.

RoofStyle: Type of roof

--> Gable, Hip and Other şeklinde listelenmesi düşünüldü.

RoofMatl: Roof material

--> Feature'un değersiz olduğunu düşünüyoruz.

Exterior1st: Exterior covering on house

--> Numerik olanları regresyon analizine dahil edebilyioru ama kategorikleri dahil edemiyoruz, değişirmek mi gerekiyor? nasıl yapabiliriz bilemiyoruz.

Exterior2nd: Exterior covering on house (if more than one material)

--> Conditiondaki gibi bir durum var aynı şekilde değerlendirilebilir.

MasVnrType: Masonry veneer type

--> Missingler'in fiyat karşısındaki etkisine göre düzenleme yapılcaktır. (bütün missing valuelar için geçerli)

MasVnrArea: Masonry veneer area in square feet

--> 0 ve others yapılabilir? 0'ın fiyat karşısındaki etkisine bakılmalı.

ExterQual: Evaluates the quality of the material on the exterior

--> OveralQuall gibi numerik hale çevrilebilir. Ya da kategorik veriler hesaplamada kullanılabilir.

BsmtQual: Evaluates the height of the basement

--> %2.5 NA, 0 Po, %2.4 Fa, %44.5 TA, %42.3 Gd ve %8.3 Ex. ten oluşuyor.

-->Po bulunmuyor, Ex-Gd ve Ta-Fa olarak eşleştirip one-hot encoding olabilir.

BsmtCond: Evaluates the general condition of the basement

-->%89.8 i TA, kolon elememiz gerektiğinden elenebilir bence.

BsmtExposure: Refers to walkout or garden level walls

--> %2.6 Na, %65.3 No, %7.8 Mn, %15.1 Av, %9.2 Gd den oluşuyor.

--> Na-No ve Mn-Av-Gd olarak eşleştirip one-hot encoding olabilir.

BsmtFinType1: Rating of basement finished area

-->Unf-NA ve geri kalanlar olarak eşleştirip one-hot encodingten yapılabilir.

BsmtFinSF1: Type 1 finished square feet

--> 0 değerleri haricinde normal dağılıma benziyor.

BsmtFinType2: Rating of basement finished area (if multiple types)

-->GLQ Good Living Quarters ALQ Average Living Quarters BLQ Below Average Living Quarters Rec Average Rec Room LwQ Low Quality Unf Unfinshed NA No Basement Missing'lerin dağılımdaki yüzdesi düşük ve no basement anlamına geldiği için Sales Price etkisine bakılarak, etkisi kısıtlı veya yok ise uçurulabilir. Veriler 2 gruba ayrılarak gruplandırılabilir. (UNF ve others şeklinde) Bu değişkenin raporda sales price ile yüksek korelasyonunun olmadığı gözükmekte fakat yine de korelasyon değerine bakılarak ilişkisi yorumlanabilir.

BsmtFinSF2: Type 2 finished square feet

-->BsmtFinType2 ile yüksek korelasyonlu olduğu için konusu da aynı olduğu için, bu veri numeric olduğundan modellemede bu konuda bir veri kullanıcaksak bunu tercih edebiliriz. Veri %88 0'lardan oluşmakta, sales price'a etkisine bakılarak kullanmamayı da tercih edebiliriz.

BsmtUnfSF: Unfinished square feet of basement area

-->Basement tip 1 ve total ile yüksek korelasyonlu olduğundan, bu konu için modellemede kullanabilir miyiz diye inceleyebiliriz.

TotalBsmtSF: Total square feet of basement area

-->Basement değişkenleriyle ve sales price ile yüksek korelasyonu olduğu için, modellemede kesinlikle bunu kullanmalıyız diye düşünüyoruz. Hem numeric bir veri seti, hem de veri dağılımı normal dağılıma yakın.

Heating: Type of heating

-->GasA ve others olarak 2 gruba ayırabiliriz, Sales price etkilerini de göz önünde bulundurarak.
Floor Floor Furnace GasA Gas forced warm air furnace GasW Gas hot water or steam heat Grav Gravity furnace OthW Hot water or steam heat other than gas Wall Wall furnace

HeatingQC: Heating quality and condition -> Dairenin ısınma kalitesi ve durumu.

-->Kategorik veridir. Mükemmel,güzel,ortalama,makul ve zayıf şeklinde sıralı veriler olduğundan ordinal kategorik veridir. Ordinal encoding yapmak mantıklı olabilir.

CentralAir: Central air conditioning ->Merkezi klima bulunması veya bulunmaması.

-->Kategorik veridir.Verilerin %93.5'i Y %6.5'u N. Cinsiyet gibi düşünüp Label Encoding yapılabilir.

Electrical: Electrical system -> Elektrik sistemi

-->Kategorik veridir.FuseP ve Mix %0.3 lük kısım olduğu için, FuseF FuseA SBrkr ve Other şeklinde rare encoding yapılabilir. 4 sınıfa indirilir.

1stFlrSF: First Floor square feet ->1.kat feet^2 (1 m^2 yaklaşık 10.76 feet^2 dir.)

-->Nümerik veridir.2300den sonraki veriler outlier olarak kabul edilip veriden atılırsa histogramı normal dağılıma benziyor.Normal dağılım olduktan sonra modelde iyi çalışır.

2ndFlrSF: Second floor square feet 2.kat feet^2

-->Nümerik veridir.%56.8 i 0. Bunlar 2.katı olmayan tek katlı evler olabilir.0 olanların dışında normal dağılıma uyuyor. pandas.qcut ile kategorik olarak ayrılabilir.

LowQualFinSF

-->Sayısal veri

-->%98.2 sıfır değer

-->hiç bir veri ile korelasyonu yok

-->kayıp veri yok

-->Pek kullanışlı değil gibi.

GrLivArea (yaşam alanı)

-->Sayısal veri

-->hiç bir veri ile korelasyonu yok

-->genel dağılım var

-->kayıp veri yok

BsmtFullBath (bodrum banyolar)

-->Sayısal veri

-->hiç bir veri ile korelasyonu yok

-->%58 0

-->%42 1 den fazla

BsmtHalfBath (bodrum yarım banyo)

-->Sayısal veri

-->hiç bir veri ile korelasyonu yok

-->%94 0

-->%6 1 den fazla

-->Pek kullanışlı değil gibi.

FullBath(banyo sayısı)

-->Sayısal veri

-->hiç bir veri ile korelasyonu yok

-->%53 2 adet banyo

-->%45 1 adet banyo

-->Her evde 1 banyo kesin vardır bir veya daha fazlası olarak sınıflandırıldığında belki korelasyon görülebilir.

HalfBath

-->Eksik veri yok.

-->Bodrum disinda bulunan dusu olmayan banyo sayisini(yarim banyo) belirtiyor.

-->Range 2. Evlerde min 0 ve max 2 yarim banyo var.

-->Numerik ve kesikli bir degisken.

-->Cogunlukla evlerde 0 daha az ihtimalle 1 adet var. 2 adet olan cok az ev var.

-->Evin fiyati ile bir korelasyonu yok.

BedroomAbvGr

-->Eksik veri yok.

-->Bodrum disinda bulunan yatak odasi sayisi.

-->Numerik ve kesikli bir degisken.

-->Range 8. min 0 max 8 yatak odasi olan ev var.

-->Evlerin buyuk bir cogunlugunda 3 adet daha az ihtimalle 2 ya da 4 adet var.

-->Evlerin fiyatina belirleyici bir etkisi yok. Tum yatak odasi adetlerine sahip evlerin fiyatlarinin ortanca degerleri neredeyse ayni.

-->Ev fiyati ile korelasyon yok.

KitchenAbvGr

-->Eksik veri yok.

-->Numerik ve kesikli bir degisken.

-->Bodrum disinda evdeki mutfak sayisini veren degisken.

-->Veri setinin yuzde 96sinda bir adet mutfak var.

-->Fiyat ile bir korelasyon gostermemekler birlikte cogu evde bir adet mutfak olmasindan dolayi belirleyici bir etkiye sahip degil.

KitchenQual

-->5 adet deger uzerinden mutfagin kalitesini belirten degisken.

-->Kategorik bir degisken. Eksik veri yok.

-->Datanin yarisi TA neredeyse diger yarisi da GD almis.

-->Fiyata etkileri bakildiginda EX alan evlerin bariz cok daha pahali oldugu goruluyor.

-->Ancak EX alan ev sayisi cok az oldugundan genel fiyat durumuna etkisi cok az olur. Diger durumlarda ise fiyatlar birbirine yakin.

TotRmsAbvGrd

-->Bodrum disinda kalan toplam oda sayisini veren kesikli ve numerik degisken. Eksik veri yok.

-->Evlerde minimum 2 maksimum 14 adet oda var.

-->Dataki sikliga gore yogunluk 5,6,7,8 oda sayisinda.

-->Fiyata etkisi ise oda sayisi ile direkt dogru orantili. Toplam oda sayisi artikca fiyat dogrusal sekilde artiyor.

-->Fiyat degiskeni ile korelasyonu yuksek.

-->Aykiri degerler var.

Functional: Home functionality (Assume typical unless deductions are warranted)

--> Kategorik.Min1 ve min2 1 grup,kalanlar ayrı 1 grup,Typical 1 grup olarak azaltılma yapılabilir

Fireplaces: Number of fireplaces

-->SalePrive ile High Correlation.Aslında yarısında yok ancak olduğu zaman eve + değer ekliyor(+ değer mi ekliyor ondan tam emin değilim)

FireplaceQu: Fireplace quality

-->Missing (%) 47.3% değeri var. Yani olmayanlar boş olarak yazılmış. Bu göz önünde bulundurulmalı.Olanların büyük çoğunluğu Gd ve TA olarak yazılmış.

-->SalePrice ile High Correlation yok!

GarageType: Garage location

-->Missing (%) 5.5%.Yani olmayanlar boş olarak yazılmış

-->3 kategori yapmak daha mantıklı olabilir. Eve dahil ,Eve dahil değil,Her ikiside olarak.zaten.Attchd 870 59.6%.Detchd 387 26.5% olarak bir baskınlık var.

GarageYrBlt: Year garage was built

-->Missing (%) 5.5%.Yani olmayanlar boş olarak yazılmış

-->SalePrice üzerindeki etkisi GarageType'dan daha yüksek!

GarageFinish: Interior finish of the garage

-->%5 missing var ama NA değeri yok missingler NA yapılabilir.

GarageCars: Size of garage in car capacity

-->General distributiona benziyen bir yapı mevcut.

GarageArea: Size of garage in square feet

-->0 değeri çok ayrık gibi duruyor yokmuş gibi varsayılırsa ggrafik farklı ve değerli bir correlation yakalayabilr

GarageQual: Garage quality

-->81 evin garajının olmadaığı açık. Grafikten bir mana çıkarmak mümkün değil.

GarageCond: Garage condition

-->Garage qual ile benzer yapıya sahip bir grafik.

PavedDrive: Paved driveway

--> Çok fazla (91.8%) "Paved" değeri olduğu için model yanılabilir. Paved ve diğerleri olarak gruplayabiliriz, eğer P ve N benzer özellikler gösteriyorsa.

WoodDeckSF: Wood deck area in square feet

--> Çok fazla 0 olan değer var (52.1%). Bunun yanında çok uç değerler de var. Bunları daha düzgün ele almak için 0 olanlar, örneğin 400'den büyük olanlar ve aradaka kalanlar olarak veriyi 3'e bölüp 3 model geliştirilebilir.

OpenPorchSF: Open porch area in square feet

--> (44.9% oranında 0 var) WoodDeckSF ile aynı şeyleri uygulayabiliriz.

EnclosedPorch: Enclosed porch area in square feet

--> Burada da 0 olan değerler %85. SalesPrice ile arasındaki ilişkiye bakılarak 0 olanlar ve diğerleri olarak gruplanabilir ya da atılabilir.

3SsnPorch: Three season porch area in square feet

--> %98 oranında 0 olan değer var. Atılacak.

ScreenPorch: Screen porch area in square feet

-->KAPALI SUNDURMA BÜYÜKLÜĞÜ

-->Zeros (%) 92.1%

-->0.111447 korelasyon var. Düşük korelasyon

-->Diğer veriler var yok diye kategorize edilip bakıldığında korelasyon iyice düşüyor.

-->Sales Price'a etki etmediği için modele eklenmesin.

PoolArea: Pool area in square feet

-->HAVUZ BÜYÜKLÜĞÜ

-->0.092404 korelasyon var. Düşük korelasyon

-->Zeros (%) 99.5%

-->Direk silinebilir.

PoolQC: Pool quality

-->HAVUZ KALİTESİ

-->Missing (%) 99.5%

-->Bu da direk silinebilir

Fence: Fence quality

-->ÇİT KALİTESİ

-->Missing (%) 80.8%

-->GoodPrivacy ve Missing Valuelar'ın ortalaması benzer olduğunu için bu ikisi bir grup diğerleri bir grup yapıldığında correlation 0.140'tan -0.186'a çıkıyor. Anova f score da yükseliyor. Modelde denenebilir.

"""

Fence	Fence
GdPrv mean : 178927 ± 56274 count : 59	Missed+GdPrv mean : 187183 ± 80616 count : 1238
GdWo mean : 140379 ± 53094 count : 54	Other mean : 145997 ± 61618 count : 222
MnPrv mean : 148751 ± 65885 count : 157	Fence Missing : 0.0%
MnWw mean : 134286 ± 20768 count : 11	Fence p : 0.000000
Fence Missing : 80.8%	Fence f : 52.376623
Fence p : 0.002313	
Fence f : 4.948159	

"""

MiscFeature: Miscellaneous feature not covered in other categories

--> Diğer kategorilerde belirtilmeyen ek özellikler

--> Missing (%) 96.3%

--> Anova'da p value anlamsız olduğu ve missing value çok olduğu için direk modele alınmayabilir.

"""

MiscFeature

Gar2 mean : 170750 ± 19250 count : 2

Othr mean : 94000 ± 39000 count : 2

Shed mean : 151187 ± 51113 count : 49

TenC mean : 250000 ± 0 count : 1

MiscFeature Missing : 96.3%

MiscFeature p : 0.104728

MiscFeature f : 2.157324

"""

MiscVal :

-->Nümerik değişken (Sale Price üzerinde High Corelation bulunmyor)

--> Çeşitli özellik değerlerinin hangi anlam da olduğunun çıkarımını yapamadım(dolar işareti mevcut)

-->Çok fazla zeros değer içermekte, zeros değerler harici atılabilir. (Ekstra outlayer mevcut olsa da count 1 olduğu için silinebilir diye düşünüyorum)

MoSold:

--> Nümerik değişken (herhangi bir değişkenler high corelation durumu yok.)

--> Histogramı düzenli dağılım gösteriyor. (6.ayda max durumdayken sağa ve sola doğru düşüş gözlemleniyor.)

YrSold:

-->Categorik değişken (Sale Price üzerinde high corelation durumu bulunmuyor.)

--> 2009'da max count mevcut ama 2006-2009 arası countlar çok yakın 2010 da düşüş var nispeten düzenli dağılım mevcut (2006-2007-2008-2009 bir grup 2010 bir grup olarak düşünülebilir.)

SaleType:

-->Categorik değişken (Sale Price üzerinde high corelation durumu bulunmuyor.)

-->Warranty Deed - Conventional -> Garanti Belgesi - Konvansiyonel(anlaşmalı) % 87 ile 1. durumda

-->New -> Ev yeni yapıldı ve satıldı (Temelden) olabilir. %8.4 ile 2. durumda

-->COD -> Mahkeme Memuru Tapu/Emlak %2.9 ile 3. durumda

-->Belitrilenlerin harici yüzde 3lük dilimi oluşturuyor 3 ayrı grup yapılp kalanlar silinebilir. Ya da dağıtılabilir.

SaleCondition:

-->Categorik değişken (Sale Price üzerinde high corelation durumu bulunmuyor.)

-->Normal %82.1 ile 1.durumda

-->Partial %8.6 ile 2. durumda (Son değerlendirme yapıldığında ev tamamlanmamıştı, yeni evler için geçerli)

-->Abnorml %6.9 ile 3. durumda (Anormal Satış - ticaret, haciz, açığa satış)

--> 3ü ayrı grup olarak yapılıp kalanları atılabilir çünkü yüzde 97lik durum sağlanıyor ama aile satışlarında fiyat etkisi olabilir.