PR 2: Neteja i anàlisis de les dades

TIPOLOGIA I CILE DE VIDA DE LES DADES

JULI CREUS ESCANDELL

Índex:

1. RESOLUCIÓ:

1.1 Descripció del data set:

El conjunt de dades s'ha obtingut a les següents pàgines a partir de l'enllaç en Kaggle, el preu de les cases de diferents barris, per poder realitzar una anàlisis estadístic de les mateixes les següents variables per determinar quins factors determinen el preu del immoble:

1- *LotArea*: Area total del lot.
2- *Neighbornhood*: Barri en el que pertany l'immoble.
3- *OverallQual*: Qualitat general del immoble.
4- *OverallCond*: Condició general del immoble.
5- *YearBuild*: Any de construcció de l'immoble.
6- *TotalBsmtSF*: Àrea total de soterrani.
7- *CentralAir*: Si consta de sistema de aire centralitzat.
8- *X1stFlrSF*: Superfície del primer pis.
9- *X2ndFlrSF*: Superfície del segon pis.
10- *GrLivArea*: Àrea construïda a la planta baixa.
11- *FullBath*: Quantitat de banys complets.
12- *HalfBath*: Quantitat de banys petits.
13- *Fireplaces*: Quantitat de llars de focs.
14- *GarageArea*: Area de garatge.
15- *Pool* Area: Area de la piscina.
16- *EnclosedPorsh*: Àrea del porxo tancat.
17- *OpenPorch*: Àrea del porxo obert.
18- *SalePrice*: Preu de venta.

1.2 Objectius del anàlisis de les dades.

L'anàlisi del sector immobiliari, és un dels punts forts dins del sector financer. Un dels mercats més importants del sector és el de les hipoteques i crèdit immobiliari.
Per qualsevol entitat financera o agencia immobiliària és d'utilitat determinar quins factors fan variar el preu de un immoble en una regió determinada en un moment determinat per poder quantificar els seus actius o per planificar futures inversions, tipus de interès segons immoble o la creació i explotació de nous productes.

1.3 Importació de les dades

Per començar amb la anàlisis de les dades, ens farà falta importar els valors que després estudiarem. S'ha escollit el programa R per la seva solidesa i facilitat en analitzar de forma estadística les dades.

```
home <- read.csv("/Users/escac/OneDrive - Escola Superior de Cinema i Audiovisuals deCatalunya/ing_DataCience/01_tipologia_cicleDeVida/01_PACs/04_PR2/03_REC/data_housing/house_data.csv", header = TRUE)

head(home[,1:6])
```

Output:

```
      Id MSSubClass MSZoning LotFrontage LotArea Street
1461  20         RH       80       11622    Pave   <NA>
1462  20         RL       81       14267    Pave   <NA>
1463  60         RL       74       13830    Pave   <NA>
1464  60         RL       78        9978    Pave   <NA>
1465 120         RL       43        5005    Pave   <NA>
1466  60         RL       75       10000    Pave   <NA>
```

Amb la següent funció obtindrem quin valor emmagatzema cada una de les columnes de la base de dades:

```
sappy(home, function(x), class(x))
```

Output:

```
          Id    MSSubClass      MSZoning   LotFrontage       LotArea
   "integer"      "factor"     "integer"     "integer"      "factor"
   Utilities     LotConfig     LandSlope  Neighborhood    Condition1
    "factor"      "factor"      "factor"      "factor"      "factor"
 OverallCond     YearBuilt  YearRemodAdd     RoofStyle      RoofMatl
   "integer"     "integer"      "factor"      "factor"      "factor"
   ExterQual     ExterCond    Foundation      BsmtQual      BsmtCond
    "factor"      "factor"      "factor"      "factor"      "factor"
  BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
   "integer"     "integer"      "factor"      "factor"      "factor"
LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
   "integer"     "integer"     "integer"     "integer"     "integer"
TotRmsAbvGrd    Functional    Fireplaces   FireplaceQu    GarageType
    "factor"     "integer"      "factor"      "factor"     "integer"
  GarageQual    GarageCond    PavedDrive    WoodDeckSF   OpenPorchSF
    "factor"      "factor"     "integer"     "integer"     "integer"
      PoolQC         Fence   MiscFeature       MiscVal        MoSold
    "factor"      "factor"     "integer"     "integer"     "integer"
```

```
      Street         Alley      LotShape   LandContour
    "factor"      "factor"      "factor"      "factor"
  Condition2      BldgType    HouseStyle   OverallQual
    "factor"      "factor"     "integer"     "integer"
 Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
    "factor"      "factor"     "integer"      "factor"
BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
    "factor"     "integer"      "factor"     "integer"
  CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
    "factor"     "integer"     "integer"     "integer"
    HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual
   "integer"     "integer"      "factor"     "integer"
 GarageYrBlt  GarageFinish    GarageCars    GarageArea
    "factor"     "integer"     "integer"      "factor"
EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
   "integer"     "integer"     "integer"      "factor"
      YrSold      SaleType SaleCondition     SalePrice
    "factor"      "factor"     "integer"     "integer"
```

1.4 Neteja de les dades buides

Abans de continuar, seleccionarem les dades que ens interessen analitzar generant una segona taula:

```
house_analytic <- select(habitatges, LotArea, Neighborhood, OverallQual,
OverallCond, YearBuilt, TotalBsmtSF, CentralAir, X1stFlrSF, X2ndFlrSF, GrLivArea,
FullBath, HalfBath, Fireplaces, GarageArea, PoolArea, EnclosedPorch, OpenPorchSF,
SalePrice  )
```

Per treballar correctament amb les dades, quins valors de les dades no presenten informació.
Obtenim els valors vuits de les variables, dels que s'han eliminat prèviament els valors NA i s'han substituït per el vuit, amb la següent funció:

```
sapply(house_analytic, function(x) sum(is.na(x)))
```

Otput:

```
  LotArea  Neighborhood   OverallQual   OverallCond     YearBuilt   TotalBsmtSF    CentralAir     X1stFlrSF     X2ndFlrSF
        0             0             0             0             0            42             0             0           430
GrLivArea      FullBath      HalfBath    Fireplaces    GarageArea      PoolArea EnclosedPorch   OpenPorchSF     SalePrice
        0             0             0             0             1             0             0             0             0
```

Estudiarem si els valors que no tenen contingut, no el tenen per algun motiu aparent o realment son valors que al ser zero, seran valors que no haurem de completar.

X2ndFlrSF: Es poden donar si no existeix primer pis, en aquest cas veiem que en les dades tenim valors que son 0 i valors

Tots els valors en les dades originals tenen un valor numèric igual a 0.
Utilitzarem el valor de els veïns per completar les dades que son buides.

```
house_analytic$X2ndFlrSF     <- kNN( house_analytic )$  X2ndFlrSF
```

1.5 Valors extrems

Procedim a la eliminació de les valors extrems dins de R, per tal de fer-ho possible utilitzarem la expressió de boxplot.

Command:

```
boxplot.stats(house_analytic$OverallQual)$out
```

Output:

```
[1] 1 1
```

Command:

```
boxplot.stats(house_analytic$OverallCond)$out
```

Output:

```
[1] 8 8 2 9 9 3 8 8 3 1 3 3 8 8 2 2 8 8 8 9 8 8 8 8 9 9 8 8 8 8 3 3 8 1 9 3 2 8 8 9 8 8 2 8 3 9 8 3 8
9 1 8 9 1 9 8 8 8 3 8
 [61] 8 8 3 8 8 8 8 9 3 9 8 8 9 1 3 8 9 8 9 3 8 8 8 9 8 8 8 8 3 9 1 8 3 8 8 8 3 8 8 8 9 8 3 3 8 8 8 3
8 8 8 8 8 8 3 8 3 8 8 8
[121] 8 3 3 8 8 8 3
```

Command:

```
boxplot.stats(house_analytic$SalePrice)$out
```

Output:

```
[1] 220079 139647 138607 232991 140791 139010 221648 225810 214217 228156 220449
220522 224118 215785 221813 255629 228617
[18] 223581 218676 215911 139214 281643 139169 280618 277936 249768 216855 223055
220514 230841 218108 140196 216032 234707
[35] 135751 228148 222406 218648 228684 234492 259423 224323 218200 222438 235419
221134 245283 223667 137402 219222
```

Command:

```
boxplot.stats(house_analytic$LotArea)$out
```

Output:

```
[1] 18494 18837 20062 18600 19645 23303 19255 26400 21780 31220 47280 18559 19508
24572 20270 19550 47007 26073 23730 18265
[21] 17979 56600 18160 51974 41600 19522 17778 22002 21281 22692 17808 23920 20064
39290 25485 21579 17871 20693 18044 19958
[41] 43500 33983 27697 39384 18062 18261 21299 22136 18275 33120 21370 20355 19950
19800 21533 21780 23580 50102 31250 20000
```

2 Anàlisis de les dades

2.1 Selecció per grups de les dades

Separem les dades segons el barri:

```
house_analytic.IDOTRR <- house_analytic[house_analytic$Neighborhood =="IDOTRR",]
house_analytic.NAmes <- house_analytic[house_analytic$Neighborhood =="NAmes",]
house_analytic.Giblert <- house_analytic[house_analytic$Neighborhood =="Giblert",]
house_analytic.StoneBr <- house_analytic[house_analytic$Neighborhood =="StoneBr",]
house_analytic.Blmngtn <- house_analytic[house_analytic$Neighborhood =="Blmngtn",]
house_analytic.Blueste <- house_analytic[house_analytic$Neighborhood =="Blueste",]
house_analytic.BrDale <- house_analytic[house_analytic$Neighborhood =="BrDale",]
house_analytic.BrkSide <- house_analytic[house_analytic$Neighborhood =="BrkSide",]
house_analytic.ClearCr <- house_analytic[house_analytic$Neighborhood =="ClearCr",]
house_analytic.CollgCr <- house_analytic[house_analytic$Neighborhood =="CollgCr",]
house_analytic.Crawfor <- house_analytic[house_analytic$Neighborhood =="Crawfor",]
house_analytic.Edwards <- house_analytic[house_analytic$Neighborhood =="Edwards",]
house_analytic.Gibert <- house_analytic[house_analytic$Neighborhood =="Girbert",]
house_analytic.MeadowV <- house_analytic[house_analytic$Neighborhood =="MeadowV",]
house_analytic.Mitchel <- house_analytic[house_analytic$Neighborhood =="Mitchel",]
house_analytic.NoRidge <- house_analytic[house_analytic$Neighborhood =="NoRidge",]
house_analytic.NPkVill <- house_analytic[house_analytic$Neighborhood =="NPkVill",]
house_analytic.NridgHt <- house_analytic[house_analytic$Neighborhood =="NridgHt",]
house_analytic.NWAmes <- house_analytic[house_analytic$Neighborhood =="NWAmes",]
house_analytic.OldTown <- house_analytic[house_analytic$Neighborhood =="OldTown",]
house_analytic.Sawyer <- house_analytic[house_analytic$Neighborhood =="Sawyer",]
house_analytic.SawyerW <- house_analytic[house_analytic$Neighborhood =="SawyerW",]
house_analytic.Somerst <- house_analytic[house_analytic$Neighborhood =="Somerst",]
house_analytic.StoneBr <- house_analytic[house_analytic$Neighborhood =="StoneBr",]
house_analytic.SWISU <- house_analytic[house_analytic$Neighborhood =="SWISU",]
house_analytic.Timber <- house_analytic[house_analytic$Neighborhood =="Timber",]
house_analytic.Veenker <- house_analytic[house_analytic$Neighborhood =="Veenker",]
```

Separem les dades per si tenen sistema de aire central:

```
house_analytic.CentralAir_Y <- house_analytic[house_analytic$CentralAir =="Y",]
house_analytic.CentralAir_N <- house_analytic[house_analytic$CentralAir =="N",]
```

2.3 Cerca de les variables que expressen normalitat

Utilitzarem el comando shapiro per coneixre si les variables expressen normalitat:

```
shapiro.test(house_analytic$PoolArea)
```

```
        Shapiro-Wilk normality test

data:  house_analytic$PoolArea
W = 0.031674, p-value < 2.2e-16

shapiro.test(house_analytic$LotArea)

        Shapiro-Wilk normality test

data:  house_analytic$LotArea
W = 0.79058, p-value < 2.2e-16

shapiro.test(house_analytic$OverallQual)

        Shapiro-Wilk normality test

data:  house_analytic$OverallQual
W = 0.94699, p-value < 2.2e-16

 shapiro.test(house_analytic$OverallCond)

        Shapiro-Wilk normality test

data:  house_analytic$OverallCond
W = 0.82219, p-value < 2.2e-16

 shapiro.test(house_analytic$YearBuilt)

        Shapiro-Wilk normality test

data:  house_analytic$YearBuilt
W = 0.92147, p-value < 2.2e-16

shapiro.test(house_analytic$TotalBsmtSF)

        Shapiro-Wilk normality test

data:  house_analytic$TotalBsmtSF
W = 0.93971, p-value < 2.2e-16

shapiro.test(house_analytic$X1stFlrSF)

        Shapiro-Wilk normality test

data:  house_analytic$X1stFlrSF
W = 0.9189, p-value < 2.2e-16

shapiro.test(house_analytic$X2ndFlrSF)

        Shapiro-Wilk normality test

data:  house_analytic$X2ndFlrSF
W = 0.87009, p-value < 2.2e-16
```

```
shapiro.test(house_analytic$GrLivArea)

        Shapiro-Wilk normality test

data:  house_analytic$GrLivArea
W = 0.94065, p-value < 2.2e-16

 shapiro.test(house_analytic$FullBath)

        Shapiro-Wilk normality test

data:  house_analytic$FullBath
W = 0.71026, p-value < 2.2e-16

 shapiro.test(house_analytic$HalfBath)

        Shapiro-Wilk normality test

data:  house_analytic$HalfBath
W = 0.63675, p-value < 2.2e-16

 shapiro.test(house_analytic$Fireplaces)

        Shapiro-Wilk normality test

data:  house_analytic$Fireplaces
W = 0.74448, p-value < 2.2e-16

shapiro.test(house_analytic$GarageArea)

        Shapiro-Wilk normality test

data:  house_analytic$GarageArea
W = 0.97537, p-value = 4.287e-15

shapiro.test(house_analytic$EnclosedPorch)

        Shapiro-Wilk normality test

data:  house_analytic$EnclosedPorch
W = 0.41376, p-value < 2.2e-16

shapiro.test(house_analytic$OpenPorchSF)

        Shapiro-Wilk normality test

data:  house_analytic$OpenPorchSF
W = 0.71705, p-value < 2.2e-16

shapiro.test(house_analytic$SalePrice)

        Shapiro-Wilk normality test
```

```
data:  house_analytic$SalePrice
W = 0.95287, p-value < 2.2e-16
```

Com s'observa en totes les proves estadístiques que el valor *p-value* és major de *0,05* per tan donem totes les hipòtesis per nul·les i considerem que cap de les variables estudiades expressa normalitat.
Es decideix no alterar els valors de les variables originals sense forçar a la normalitat.

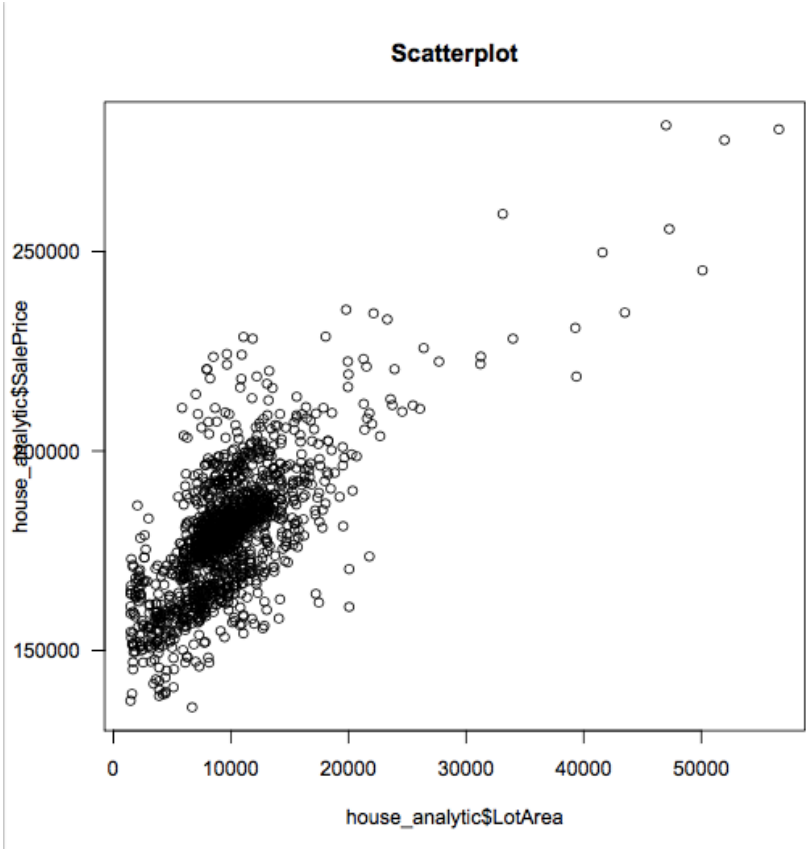
3 Proves estadístiques

Compararem cada un dels valors de les dades gracies a un càlcul de correlació, de tal forma que podrem trobar quins son els factors més significatius que afecten al preu.

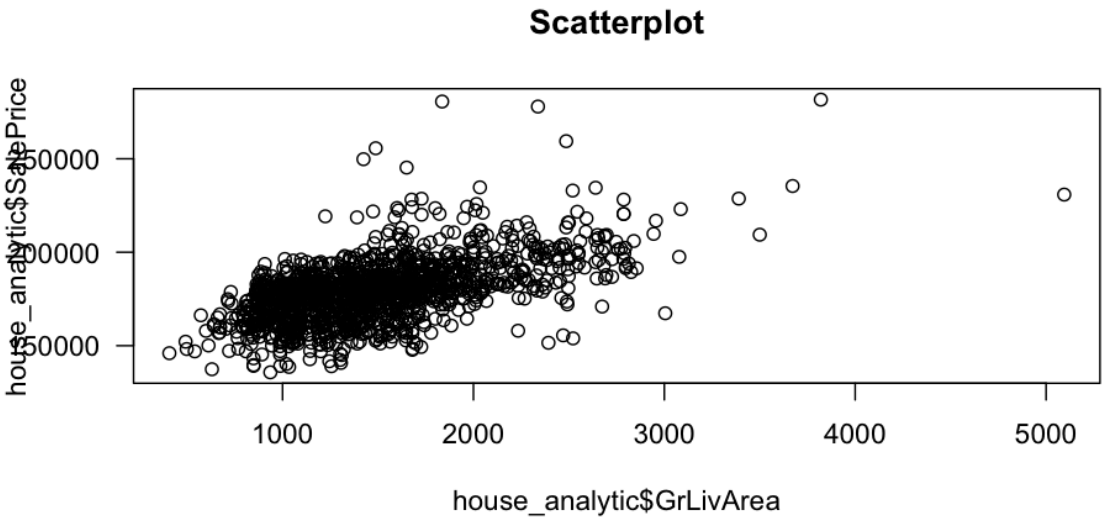
```
cor(house_analytic$OverallQual,house_analytic$SalePrice, method="pearson" )
[1] 0.09364365
cor(house_analytic$YearBuilt,house_analytic$SalePrice, method="pearson" )
[1] 0.008186861
cor(house_analytic$TotalBsmtSF,house_analytic$SalePrice, method="pearson" )
[1] NA
cor(house_analytic$X1stFlrSF,house_analytic$SalePrice, method="pearson" )
[1] 0.3393351
cor(house_analytic$X2ndFlrSF,house_analytic$SalePrice, method="pearson" )
[1] NA
cor(house_analytic$GrLivArea,house_analytic$SalePrice, method="pearson" )
[1] 0.566654
cor(house_analytic$FullBath,house_analytic$SalePrice, method="pearson" )
[1] 0.3263128
cor(house_analytic$HalfBath,house_analytic$SalePrice, method="pearson" )
[1] 0.2230017
cor(house_analytic$Fireplaces,house_analytic$SalePrice, method="pearson" )
[1] 0.213041
cor(house_analytic$GarageArea,house_analytic$SalePrice, method="pearson" )
[1] NA
cor(house_analytic$PoolArea,house_analytic$SalePrice, method="pearson" )
[1] 0.07127058
cor(house_analytic$EnclosedPorch,house_analytic$SalePrice, method="pearson" )
[1] 0.09465828
cor(house_analytic$OpenPorchSF,house_analytic$SalePrice, method="pearson" )
[1] 0.1513134

cor(house_analytic$LotArea,house_analytic$SalePrice, method="pearson" )
[1] 0.7157962
```

Com es pot observar, els 6 valors que son més representatius en la correlació de *Pearson son els factors de: LotArea, GrLivingArea, X1stFlrSF, FullBath, HalfBath i Fireplace.*

Podem mostrar unes gràfiques en les que observem visualment les dos correlacions més importants:



Il·lustració 1: Àrea del lot y preu de venta



Il·lustració 2: Mida planta baixa y preu de venta

Per realitzar un anàlisis més profund, podem analitzar quin serà el valor de correlació amb el preu sobre el valor més significatiu, que n'és la àrea del lot, en el cas de que no existeixi cap llar de foc i també cap bany complet:

```
house_analytic.NoFPlace <- house_analytic[house_analytic$Fireplace ==0,]
cor(house_analytic.NoFPlace$LotArea, house_analytic.NoFPlace$SalePrice,
method="pearson" )
[1] 0.6473336

house_analytic.NoFBath <- house_analytic[house_analytic$FullBath ==0,]
cor(house_analytic.NoFBath$LotArea, house_analytic.NoFBath$SalePrice, method="pearson" )
[1] 0.5118943
```

4. Conclusions

Els valors que es consideren més representatius en el anàlisis correlatiu tenen a veure amb la quantitat de metres del lot, pràcticament podem dir que existeix una relació directe dels metres quadrats del lot i el preu del immoble, la següent funció de correlació son els metres quadrats de casa construïts, sobretot pel que fa la planta baixa de la mateixa.
Els factors que influencien en el preu de venta del immoble però per altre banda no son valors que millorin la grandesa del immoble, sinó que son valors afegits seran les llars de foc i també la quantitat de banys complets a l'habitatge.

Per altre banda, observem que son un factor de correlació important el fet de que existeixi o no un bany complet, ja que el fet de que no existeixi fa que la correlació del preu amb l'àrea total del lot baixa significativament, el mateix succeeix en el cas de que no tinguem llar de foc.

Altres conclusions que podem extreure, es que els anys que tingui l'habitatge no son un factor suficientment significatiu, tampoc ho es la qualitat general del immoble i les condicions en les que es troba.

Link de GitHub: https://github.com/jcreuses/uoc-AhousePriceAnalytic