PRA2-Tipologia i cicle de vida de les dades - HousePrices

Damián Martínez & Eusebi Garcia

09/6/2020

1 Repositori Github

Repositori on es troba tota la documentació de la pràctica: https://github.com/egarciare/House-Prices-Kaggle (https://github.com/egarciare/House-Prices-Kaggle)

Link a la competició de Kaggle en que es basa la pràctica: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview)

2 Descripició i Objectiu

Aquest dataset conté les característiques d'un conjunt de cases amb 79 variables que descriuen gairebé completament cada aspecte d'aquestes cases residencials ubicades en Ames, Iowa. Aquest data set forma part d'un challenge en Kaggle anomenat: "HousePrices: Advance Regression Techniques" on el repte és predir el preu final de cada casa.

En el dataset Hi ha 1460 observacions amb 81 atributs, on la variable a predir és contínua (SalePrice).

Addicionalment a la predicció dels preus es pretén donar resposta a les següents preguntes:

- Predicció dels preus de venda (SalePrice) de les cases en funció de els seves característiques (creació d'un model de regressió lineal).
- Anàlisi per barris: diferència en el preu de venda en funció del barri on es troba la casa.
- Comparació dels preus de venda entre els anys 80 i els 90: hi ha diferències significatives entre els preus de venda en aquestes dues dècades?
- Anàlisi del preu de venda al llarg del temps: com evoluciona el preu de venda de les cases en funció de les variables temporals del dataset?

3 Exploració del dataset

S'assignen a variables el dataset de train i el de test

```
#summary(cars)
train <- read_csv('HousePrices/train.csv' )
test <- read_csv("HousePrices/test.csv")
```

```
names(train)
```

```
##  [1] "Id"           "MSSubClass"    "MSZoning"      "LotFrontage"
##  [5] "LotArea"      "Street"        "Alley"         "LotShape"
##  [9] "LandContour"  "Utilities"     "LotConfig"     "LandSlope"
## [13] "Neighborhood" "Condition1"    "Condition2"    "BldgType"
## [17] "HouseStyle"   "OverallQual"   "OverallCond"   "YearBuilt"
## [21] "YearRemodAdd" "RoofStyle"     "RoofMatl"      "Exterior1st"
## [25] "Exterior2nd"  "MasVnrType"    "MasVnrArea"    "ExterQual"
## [29] "ExterCond"    "Foundation"    "BsmtQual"      "BsmtCond"
## [33] "BsmtExposure" "BsmtFinType1"  "BsmtFinSF1"    "BsmtFinType2"
## [37] "BsmtFinSF2"   "BsmtUnfSF"     "TotalBsmtSF"   "Heating"
## [41] "HeatingQC"    "CentralAir"    "Electrical"    "1stFlrSF"
## [45] "2ndFlrSF"     "LowQualFinSF"  "GrLivArea"     "BsmtFullBath"
## [49] "BsmtHalfBath" "FullBath"      "HalfBath"      "BedroomAbvGr"
## [53] "KitchenAbvGr" "KitchenQual"   "TotRmsAbvGrd"  "Functional"
## [57] "Fireplaces"   "FireplaceQu"   "GarageType"    "GarageYrBlt"
## [61] "GarageFinish" "GarageCars"    "GarageArea"    "GarageQual"
## [65] "GarageCond"   "PavedDrive"    "WoodDeckSF"    "OpenPorchSF"
## [69] "EnclosedPorch" "3SsnPorch"    "ScreenPorch"   "PoolArea"
## [73] "PoolQC"       "Fence"         "MiscFeature"   "MiscVal"
## [77] "MoSold"       "YrSold"        "SaleType"      "SaleCondition"
## [81] "SalePrice"
```

```
summary(train)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##                                                      NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   LandSlope         Neighborhood        Condition1
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Condition2          BldgType          HouseStyle         OverallQual
##  Length:1460        Length:1460        Length:1460        Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 5.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.000
##                                                           Mean   : 6.099
##                                                           3rd Qu.: 7.000
##                                                           Max.   :10.000
##
##   OverallCond      YearBuilt     YearRemodAdd   RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Length:1460
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Class :character
##  Median :5.000   Median :1973   Median :1994   Mode  :character
##  Mean   :5.575   Mean   :1971   Mean   :1985
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004
##  Max.   :9.000   Max.   :2010   Max.   :2010
##
##    RoofMatl         Exterior1st        Exterior2nd
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   MasVnrType          MasVnrArea      ExterQual          ExterCond
##  Length:1460        Min.   :   0.0   Length:1460        Length:1460
##  Class :character   1st Qu.:   0.0   Class :character   Class :character
##  Mode  :character   Median :   0.0   Mode  :character   Mode  :character
##                     Mean   : 103.7
##                     3rd Qu.: 166.0
##                     Max.   :1600.0
##                     NA's   :8
##   Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##   HeatingQC          CentralAir         Electrical          1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##     2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
```

```
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
##  Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.047                      Mean   : 6.518
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :3.000                      Max.   :14.000
##
##    Fireplaces    FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.000   Length:1460        Length:1460        Min.   :1900
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1961
##  Median :1.000   Mode  :character   Mode  :character   Median :1980
##  Mean   :0.613                                         Mean   :1979
##  3rd Qu.:1.000                                         3rd Qu.:2002
##  Max.   :3.000                                         Max.   :2010
##                                                        NA's   :81
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460
##  Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.767   Mean   : 473.0
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :4.000   Max.   :1418.0
##
##   GarageCond          PavedDrive          WoodDeckSF      OpenPorchSF
##  Length:1460        Length:1460        Min.   :  0.00   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median : 25.00
##                                        Mean   : 94.24   Mean   : 46.66
##                                        3rd Qu.:168.00   3rd Qu.: 68.00
##                                        Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      3SsnPorch       ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     PoolQC             Fence           MiscFeature
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     MiscVal             MoSold           YrSold       SaleType
##  Min.   :    0.00   Min.   : 1.000   Min.   :2006   Length:1460
##  1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   Class :character
##  Median :    0.00   Median : 6.000   Median :2008   Mode  :character
##  Mean   :   43.49   Mean   : 6.322   Mean   :2008
##  3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :15500.00   Max.   :12.000   Max.   :2010
##
##  SaleCondition        SalePrice
##  Length:1460        Min.   : 34900
##  Class :character   1st Qu.:129975
##  Mode  :character   Median :163000
##                     Mean   :180921
##                     3rd Qu.:214000
##                     Max.   :755000
##
```

```
str(train)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1460 obs. of  81 va
riables:
##  $ Id           : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : num  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ LotFrontage  : num  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : num  8450 9600 11250 9550 14260 ...
##  $ Street       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  NA NA NA NA ...
##  $ LotShape     : chr  "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr  "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : num  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : num  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : num  2003 1976 2001 1915 2000 ...
##  $ YearRemodAdd : num  2003 1976 2002 1970 2000 ...
##  $ RoofStyle    : chr  "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : num  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : chr  "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : chr  "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr  "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr  "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : num  706 978 486 216 655 ...
##  $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ BsmtFinSF2   : num  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : num  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : num  856 1262 920 756 1145 ...
##  $ Heating      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 1stFlrSF     : num  856 1262 920 961 1145 ...
##  $ 2ndFlrSF     : num  854 0 866 756 1053 ...
##  $ LowQualFinSF : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : num  1710 1262 1786 1717 2198 ...
##  $ BsmtFullBath : num  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : num  0 1 0 0 0 0 0 0 0 0 ...
```

```
##  $ FullBath     : num  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : num  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : num  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : num  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : num  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : num  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : num  2003 1976 2001 1998 2000 ...
##  $ GarageFinish : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars   : num  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : num  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ WoodDeckSF   : num  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : num  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: num  0 0 0 272 0 0 0 228 205 0 ...
##  $ 3SsnPorch    : num  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA NA NA ...
##  $ MiscVal      : num  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : num  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : num  2008 2007 2008 2006 2008 ...
##  $ SaleType     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : num  208500 181500 223500 140000 250000 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   MSSubClass = col_double(),
##   ..   MSZoning = col_character(),
##   ..   LotFrontage = col_double(),
##   ..   LotArea = col_double(),
##   ..   Street = col_character(),
##   ..   Alley = col_character(),
##   ..   LotShape = col_character(),
##   ..   LandContour = col_character(),
##   ..   Utilities = col_character(),
##   ..   LotConfig = col_character(),
##   ..   LandSlope = col_character(),
##   ..   Neighborhood = col_character(),
##   ..   Condition1 = col_character(),
##   ..   Condition2 = col_character(),
##   ..   BldgType = col_character(),
##   ..   HouseStyle = col_character(),
##   ..   OverallQual = col_double(),
```

```
##   ..   OverallCond = col_double(),
##   ..   YearBuilt = col_double(),
##   ..   YearRemodAdd = col_double(),
##   ..   RoofStyle = col_character(),
##   ..   RoofMatl = col_character(),
##   ..   Exterior1st = col_character(),
##   ..   Exterior2nd = col_character(),
##   ..   MasVnrType = col_character(),
##   ..   MasVnrArea = col_double(),
##   ..   ExterQual = col_character(),
##   ..   ExterCond = col_character(),
##   ..   Foundation = col_character(),
##   ..   BsmtQual = col_character(),
##   ..   BsmtCond = col_character(),
##   ..   BsmtExposure = col_character(),
##   ..   BsmtFinType1 = col_character(),
##   ..   BsmtFinSF1 = col_double(),
##   ..   BsmtFinType2 = col_character(),
##   ..   BsmtFinSF2 = col_double(),
##   ..   BsmtUnfSF = col_double(),
##   ..   TotalBsmtSF = col_double(),
##   ..   Heating = col_character(),
##   ..   HeatingQC = col_character(),
##   ..   CentralAir = col_character(),
##   ..   Electrical = col_character(),
##   ..   `1stFlrSF` = col_double(),
##   ..   `2ndFlrSF` = col_double(),
##   ..   LowQualFinSF = col_double(),
##   ..   GrLivArea = col_double(),
##   ..   BsmtFullBath = col_double(),
##   ..   BsmtHalfBath = col_double(),
##   ..   FullBath = col_double(),
##   ..   HalfBath = col_double(),
##   ..   BedroomAbvGr = col_double(),
##   ..   KitchenAbvGr = col_double(),
##   ..   KitchenQual = col_character(),
##   ..   TotRmsAbvGrd = col_double(),
##   ..   Functional = col_character(),
##   ..   Fireplaces = col_double(),
##   ..   FireplaceQu = col_character(),
##   ..   GarageType = col_character(),
##   ..   GarageYrBlt = col_double(),
##   ..   GarageFinish = col_character(),
##   ..   GarageCars = col_double(),
##   ..   GarageArea = col_double(),
##   ..   GarageQual = col_character(),
##   ..   GarageCond = col_character(),
##   ..   PavedDrive = col_character(),
##   ..   WoodDeckSF = col_double(),
##   ..   OpenPorchSF = col_double(),
##   ..   EnclosedPorch = col_double(),
##   ..   `3SsnPorch` = col_double(),
```

```
##   ..   ScreenPorch = col_double(),
##   ..   PoolArea = col_double(),
##   ..   PoolQC = col_character(),
##   ..   Fence = col_character(),
##   ..   MiscFeature = col_character(),
##   ..   MiscVal = col_double(),
##   ..   MoSold = col_double(),
##   ..   YrSold = col_double(),
##   ..   SaleType = col_character(),
##   ..   SaleCondition = col_character(),
##   ..   SalePrice = col_double()
##   .. )
```

4 Anàlisi

4.1 Atributs del dataset

S'identifiquen inicialment els atributs numèrics que poden ser d'interès per l'estimació del preu de la casa. Identificació atributs numèrics:

- OverallQual: Rates the overall material and finish of the house

```
10   Very Excellent
9    Excellent
8    Very Good
7    Good
6    Above Average
5    Average
4    Below Average
3    Fair
2    Poor
1    Very Poor
```

- OverallCond: Rates the overall condition of the house

```
10   Very Excellent
9    Excellent
8    Very Good
7    Good
6    Above Average
5    Average
4    Below Average
3    Fair
2    Poor
1    Very Poor
```

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

- MasVnrArea: Masonry veneer area in square feet
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
- Kitchen: Kitchens above grade
- Fireplaces: Number of fireplaces
- GarageYrBlt: Year garage was built
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold (MM)
- YrSold: Year Sold (YYYY)

A continuació, llistem també els atributs categòrics que poden influir en el preu:

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

```
 20  1-STORY 1946 & NEWER ALL STYLES
 30  1-STORY 1945 & OLDER
 40  1-STORY W/FINISHED ATTIC ALL AGES
 45  1-1/2 STORY - UNFINISHED ALL AGES
 50  1-1/2 STORY FINISHED ALL AGES
 60  2-STORY 1946 & NEWER
 70  2-STORY 1945 & OLDER
 75  2-1/2 STORY ALL AGES
 80  SPLIT OR MULTI-LEVEL
 85  SPLIT FOYER
 90  DUPLEX - ALL STYLES AND AGES
120  1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150  1-1/2 STORY PUD - ALL AGES
160  2-STORY PUD - 1946 & NEWER
180  PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190  2 FAMILY CONVERSION - ALL STYLES AND AGES
```

- MSZoning: Identifies the general zoning classification of the sale.

```
A    Agriculture
C    Commercial
FV   Floating Village Residential
I    Industrial
RH   Residential High Density
RL   Residential Low Density
RP   Residential Low Density Park
RM   Residential Medium Density
```

- Street: Type of road access to property

```
Grvl Gravel
Pave Paved
```

- Alley: Type of alley access to property

```
Grvl Gravel
Pave Paved
NA   No alley access
```

- LotShape: General shape of property

```
Reg  Regular
IR1  Slightly irregular
IR2  Moderately Irregular
IR3  Irregular
```

- LandContour: Flatness of the property

```
Lvl  Near Flat/Level
Bnk  Banked - Quick and significant rise from street grade to building
HLS  Hillside - Significant slope from side to side
Low  Depression
```

- Utilities: Type of utilities available

```
AllPub   All public Utilities (E,G,W,& S)
NoSewr   Electricity, Gas, and Water (Septic Tank)
NoSeWa   Electricity and Gas Only
ELO  Electricity only
```

- LotConfig: Lot configuration

```
Inside   Inside lot
Corner   Corner lot
CulDSac  Cul-de-sac
FR2  Frontage on 2 sides of property
FR3  Frontage on 3 sides of property
```

- LandSlope: Slope of property

```
Gtl  Gentle slope
Mod  Moderate Slope
Sev  Severe Slope
```

- Neighborhood: Physical locations within Ames city limits

```
Blmngtn  Bloomington Heights
Blueste  Bluestem
BrDale   Briardale
BrkSide  Brookside
ClearCr  Clear Creek
CollgCr  College Creek
Crawfor  Crawford
Edwards  Edwards
Gilbert  Gilbert
IDOTRR   Iowa DOT and Rail Road
MeadowV  Meadow Village
Mitchel  Mitchell
Names    North Ames
NoRidge  Northridge
NPkVill  Northpark Villa
NridgHt  Northridge Heights
NWAmes   Northwest Ames
OldTown  Old Town
SWISU    South & West of Iowa State University
Sawyer   Sawyer
SawyerW  Sawyer West
Somerst  Somerset
StoneBr  Stone Brook
Timber   Timberland
Veenker  Veenker
```

- Condition1: Proximity to various conditions

```
Artery   Adjacent to arterial street
Feedr    Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad
```

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

```
Artery   Adjacent to arterial street
Feedr    Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad
```

- BldgType: Type of dwelling

```
1Fam Single-family Detached
2FmCon   Two-family Conversion; originally built as one-family dwelling
Duplx    Duplex
TwnhsE   Townhouse End Unit
TwnhsI   Townhouse Inside Unit
```

- HouseStyle: Style of dwelling

```
1Story   One story
1.5Fin   One and one-half story: 2nd level finished
1.5Unf   One and one-half story: 2nd level unfinished
2Story   Two story
2.5Fin   Two and one-half story: 2nd level finished
2.5Unf   Two and one-half story: 2nd level unfinished
SFoyer   Split Foyer
SLvl Split Level
```

- RoofStyle: Type of roof

```
Flat Flat
Gable    Gable
Gambrel  Gabrel (Barn)
Hip  Hip
Mansard  Mansard
Shed Shed
```

- RoofMatl: Roof material

```
ClyTile  Clay or Tile
CompShg  Standard (Composite) Shingle
Membran  Membrane
Metal    Metal
Roll Roll
Tar&Grv  Gravel & Tar
WdShake  Wood Shakes
WdShngl  Wood Shingles
```

- Exterior1st: Exterior covering on house

```
AsbShng  Asbestos Shingles
AsphShn  Asphalt Shingles
BrkComm  Brick Common
BrkFace  Brick Face
CBlock   Cinder Block
CemntBd  Cement Board
HdBoard  Hard Board
ImStucc  Imitation Stucco
MetalSd  Metal Siding
Other    Other
Plywood  Plywood
PreCast  PreCast
Stone    Stone
Stucco   Stucco
VinylSd  Vinyl Siding
Wd Sdng  Wood Siding
WdShing  Wood Shingles
```

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

```
AsbShng  Asbestos Shingles
AsphShn  Asphalt Shingles
BrkComm  Brick Common
BrkFace  Brick Face
CBlock   Cinder Block
CemntBd  Cement Board
HdBoard  Hard Board
ImStucc  Imitation Stucco
MetalSd  Metal Siding
Other    Other
Plywood  Plywood
PreCast  PreCast
Stone    Stone
Stucco   Stucco
VinylSd  Vinyl Siding
Wd Sdng  Wood Siding
WdShing  Wood Shingles
```

- MasVnrType: Masonry veneer type

```
BrkCmn   Brick Common
BrkFace  Brick Face
CBlock   Cinder Block
None None
Stone    Stone
```

- ExterQual: Evaluates the quality of the material on the exterior

```
 Ex   Excellent
 Gd   Good
 TA   Average/Typical
 Fa   Fair
 Po   Poor
```

- ExterCond: Evaluates the present condition of the material on the exterior

```
 Ex   Excellent
 Gd   Good
 TA   Average/Typical
 Fa   Fair
 Po   Poor
```

- Foundation: Type of foundation

```
 BrkTil   Brick & Tile
 CBlock   Cinder Block
 PConc    Poured Contrete
 Slab Slab
 Stone    Stone
 Wood Wood
```

- BsmtQual: Evaluates the height of the basement

```
 Ex   Excellent (100+ inches)
 Gd   Good (90-99 inches)
 TA   Typical (80-89 inches)
 Fa   Fair (70-79 inches)
 Po   Poor (<70 inches
 NA   No Basement
```

- BsmtCond: Evaluates the general condition of the basement

```
 Ex   Excellent
 Gd   Good
 TA   Typical - slight dampness allowed
 Fa   Fair - dampness or some cracking or settling
 Po   Poor - Severe cracking, settling, or wetness
 NA   No Basement
```

- BsmtExposure: Refers to walkout or garden level walls

```
 Gd   Good Exposure
 Av   Average Exposure (split levels or foyers typically score average or
above)
 Mn   Mimimum Exposure
 No   No Exposure
 NA   No Basement
```

- BsmtFinType1: Rating of basement finished area

```
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
```

- 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
```

- Heating: Type of heating

```
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

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
Po   Poor
```

- CentralAir: Central air conditioning

```
N    No
Y    Yes
```

- Electrical: Electrical system

```
SBrkr    Standard Circuit Breakers & Romex
FuseA    Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF    60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP    60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix  Mixed
```

- KitchenQual: Kitchen quality

```
 Ex   Excellent
 Gd   Good
 TA   Typical/Average
 Fa   Fair
 Po   Poor
```

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

```
 Typ  Typical Functionality
 Min1 Minor Deductions 1
 Min2 Minor Deductions 2
 Mod  Moderate Deductions
 Maj1 Major Deductions 1
 Maj2 Major Deductions 2
 Sev  Severely Damaged
 Sal  Salvage only
```

- FireplaceQu: Fireplace quality

```
 Ex   Excellent - Exceptional Masonry Fireplace
 Gd   Good - Masonry Fireplace in main level
 TA   Average - Prefabricated Fireplace in main living area or Masonry Fir
eplace in basement
 Fa   Fair - Prefabricated Fireplace in basement
 Po   Poor - Ben Franklin Stove
 NA   No Fireplace
```

- GarageType: Garage location

```
 2Types   More than one type of garage
 Attchd   Attached to home
 Basment  Basement Garage
 BuiltIn  Built-In (Garage part of house - typically has room above garag
e)
 CarPort  Car Port
 Detchd   Detached from home
 NA   No Garage
```

- GarageFinish: Interior finish of the garage

```
 Fin  Finished
 RFn  Rough Finished
 Unf  Unfinished
 NA   No Garage
```

- GarageQual: Garage quality

```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
NA   No Garage
```

- GarageCond: Garage condition

```
Ex   Excellent
Gd   Good
TA   Typical/Average
Fa   Fair
Po   Poor
NA   No Garage
```

- PavedDrive: Paved driveway

```
Y    Paved
P    Partial Pavement
N    Dirt/Gravel
```

- PoolQC: Pool quality

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
NA   No Pool
```

- Fence: Fence quality

```
GdPrv    Good Privacy
MnPrv    Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA   No Fence
```

- MiscFeature: Miscellaneous feature not covered in other categories

```
Elev Elevator
Gar2 2nd Garage (if not described in garage section)
Othr Other
Shed Shed (over 100 SF)
TenC Tennis Court
NA   None
```

- SaleType: Type of sale

```
 WD   Warranty Deed - Conventional
 CWD  Warranty Deed - Cash
 VWD  Warranty Deed - VA Loan
 New  Home just constructed and sold
 COD  Court Officer Deed/Estate
 Con  Contract 15% Down payment regular terms
 ConLw    Contract Low Down payment and low interest
 ConLI    Contract Low Interest
 ConLD    Contract Low Down
 Oth  Other
```

- SaleCondition: Condition of sale

```
 Normal   Normal Sale
 Abnorml  Abnormal Sale -  trade, foreclosure, short sale
 AdjLand  Adjoining Land Purchase
 Alloca   Allocation - two linked properties with separate deeds, typicall
y condo with a garage unit
 Family   Sale between family members
 Partial  Home was not completed when last assessed (associated with New H
omes)
```

4.2 Valors NA

Anàlisi dels valors NA en el total d'atributs del dataset de training:

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

```
##             Id     MSSubClass       MSZoning    LotFrontage        LotArea
##              0              0              0            259              0
##         Street          Alley      LotShape    LandContour      Utilities
##              0           1369              0              0              0
##     LotConfig      LandSlope   Neighborhood     Condition1     Condition2
##              0              0              0              0              0
##      BldgType     HouseStyle    OverallQual    OverallCond      YearBuilt
##              0              0              0              0              0
##  YearRemodAdd      RoofStyle      RoofMatl    Exterior1st    Exterior2nd
##              0              0              0              0              0
##    MasVnrType     MasVnrArea     ExterQual     ExterCond     Foundation
##              8              8              0              0              0
##      BsmtQual       BsmtCond   BsmtExposure   BsmtFinType1     BsmtFinSF1
##             37             37             38             37              0
##  BsmtFinType2     BsmtFinSF2     BsmtUnfSF    TotalBsmtSF        Heating
##             38              0              0              0              0
##     HeatingQC     CentralAir     Electrical      1stFlrSF      2ndFlrSF
##              0              0              1              0              0
##  LowQualFinSF      GrLivArea   BsmtFullBath   BsmtHalfBath       FullBath
##              0              0              0              0              0
##      HalfBath   BedroomAbvGr   KitchenAbvGr    KitchenQual   TotRmsAbvGrd
##              0              0              0              0              0
##    Functional     Fireplaces    FireplaceQu     GarageType    GarageYrBlt
##              0              0            690             81             81
##  GarageFinish     GarageCars     GarageArea     GarageQual     GarageCond
##             81              0              0             81             81
##    PavedDrive     WoodDeckSF    OpenPorchSF  EnclosedPorch      3SsnPorch
##              0              0              0              0              0
##   ScreenPorch       PoolArea         PoolQC          Fence    MiscFeature
##              0              0           1453           1179           1406
##       MiscVal         MoSold         YrSold       SaleType  SaleCondition
##              0              0              0              0              0
##     SalePrice
##              0
```

Dels atributs numèrics seleccionats, s'observa que no hi ha valors NA, tot i que poden haver-hi 0, per exemple en els anys. En la variable SalePrice tampoc hi ha NA.

En els atributs categòrics sí s'aprecien valors NA, en el cas de Fence, els NA equivaldrien a no fence, i s'observa que en tots els casos els valors NA formarien part del domini. Per exemple en els següents casos: NA= No Fence, NA=No Miscellaneous o que la casa no disposa de Garatge.

- Fence: Fence quality

```
GdPrv    Good Privacy
MnPrv    Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA   No Fence
```

- MiscFeature: Miscellaneous feature not covered in other categories

```
 Elev Elevator
 Gar2 2nd Garage (if not described in garage section)
 Othr Other
 Shed Shed (over 100 SF)
 TenC Tennis Court
 NA   None
```

- GarageType:81 NA Garage location 81, el quan indica que l'any del garatge no pot omplir-se perquè no hi ha garatge. –GarageYrBlt: Year garage was built

```
 2Types   More than one type of garage
 Attchd   Attached to home
 Basment  Basement Garage
 BuiltIn  Built-In (Garage part of house - typically has room above garag
e)
 CarPort  Car Port
 Detchd   Detached from home
 NA   No Garage
```

4.3 Reducció de dimensionalitat

Donat que es té una gran quantitat d'atributs: 81 atributs es fa una reducció de la dimensionalitat.

1.Donat que tenim 1460 observacions, es prescindeix dels atributs amb un nombre molt elevat de NA. (where NA >1000 NA) –> Alley, PoolQC, Fence, MiscFeature

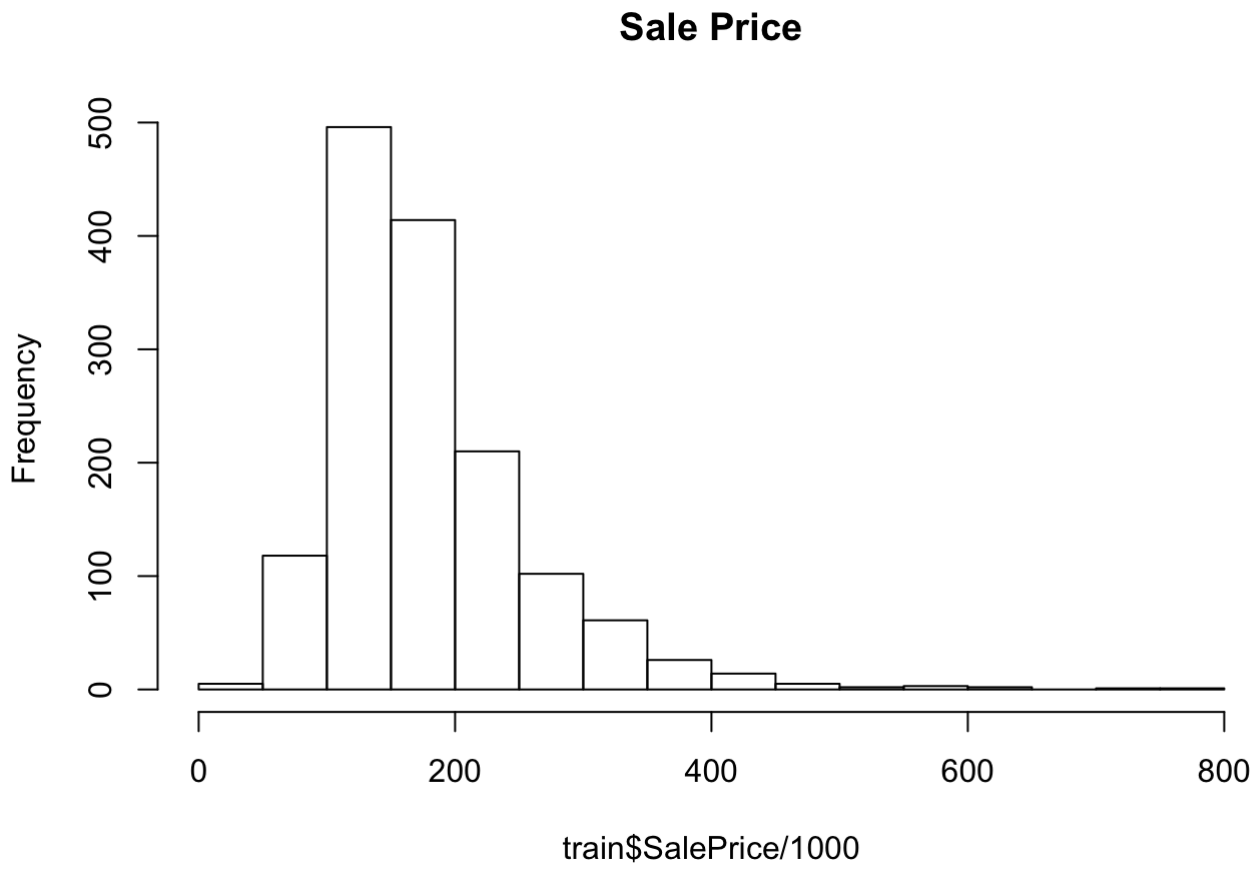
```
train = subset(train, select = -c(Alley,PoolQC,Fence,MiscFeature) )
```

2.Abans de fer matrius de correlació per a identificar i prescidir d'atributs amb una alta correlació entre ells (colinealitat) o amb molt poca correlació amb la variable Sale Price, es fa una prova de normalitat.

2.1.Normalitat de Sale Price que és la variable dependent.

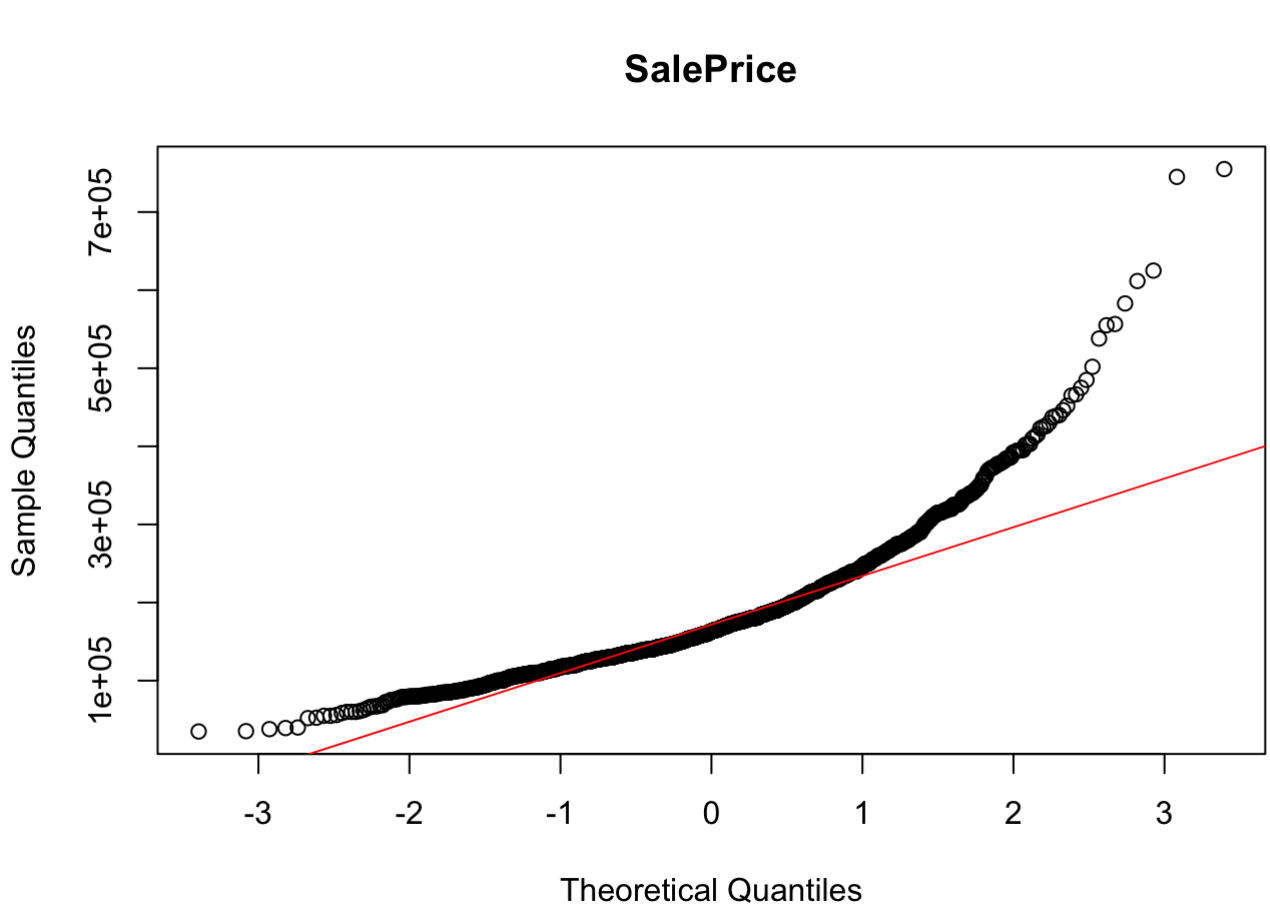
Histograma: Per a veure les frequências. Es divideix per 1000 per a fer llegibles els preus. Els preus més freqüents estan entre 100k i 200k

```
hist(train$SalePrice/1000, main="Sale Price")
```



Visualment no sembla seguir una distribució normal

```
qqnorm(train$SalePrice, main="SalePrice")
qqline(train$SalePrice,col=2)
```



Es confirma la manca de normalitat amb els tests de Kolmogorov-Smirnov i Shapiro-Wilk.

```
ks.test(train$SalePrice, pnorm, mean(train$SalePrice), sd(train$SalePrice))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  train$SalePrice
## D = 0.12369, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
shapiro.test(train$SalePrice)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train$SalePrice
## W = 0.86967, p-value < 2.2e-16
```

veiem com el p-value és menor a 0.05, el qual vol dir que hem de refusa la Hipòtesi 0 i per tant podem confirmar que hi ha un manca de normalitat.

2.2.A continuació es fa la prova de normalitat de les variables numèriques usant el mètode de Shapiro-Wilk:

```
library(nortest)
train <- as.data.frame(train)
##---- **** encara no funciona!
alpha = 0.05
col.names = colnames(train)
for (i in 1:ncol(train)) {
        if (i == 1) cat("Variables que no segueixen una distribució normal:\n")

        if (is.integer(train[,i]) | is.numeric(train[,i])) {

                #p_val = ad.test(train[,i])$p.value
                 p_val = shapiro.test(train[,i])$p.value
                 #  print(col.names[i])
                # print (p_val)
                 if (p_val < alpha) {
                         cat (paste("No Normal: -- ", col.names[i]),"\r\n")
# Format output if (i < ncol(automoviles) - 1) cat(", ") if (i %% 3 == 0) cat
("\n")
                 }
                else{
                        cat (paste("Normal: -- ", col.names[i]),"\r\n")
                }

        }

}
```

```
## Variables que no segueixen una distribució normal:
## No Normal: --  Id
## No Normal: --  MSSubClass
## No Normal: --  LotFrontage
## No Normal: --  LotArea
## No Normal: --  OverallQual
## No Normal: --  OverallCond
## No Normal: --  YearBuilt
## No Normal: --  YearRemodAdd
## No Normal: --  MasVnrArea
## No Normal: --  BsmtFinSF1
## No Normal: --  BsmtFinSF2
## No Normal: --  BsmtUnfSF
## No Normal: --  TotalBsmtSF
## No Normal: --  1stFlrSF
## No Normal: --  2ndFlrSF
## No Normal: --  LowQualFinSF
## No Normal: --  GrLivArea
## No Normal: --  BsmtFullBath
## No Normal: --  BsmtHalfBath
## No Normal: --  FullBath
## No Normal: --  HalfBath
## No Normal: --  BedroomAbvGr
## No Normal: --  KitchenAbvGr
## No Normal: --  TotRmsAbvGrd
## No Normal: --  Fireplaces
## No Normal: --  GarageYrBlt
## No Normal: --  GarageCars
## No Normal: --  GarageArea
## No Normal: --  WoodDeckSF
## No Normal: --  OpenPorchSF
## No Normal: --  EnclosedPorch
## No Normal: --  3SsnPorch
## No Normal: --  ScreenPorch
## No Normal: --  PoolArea
## No Normal: --  MiscVal
## No Normal: --  MoSold
## No Normal: --  YrSold
## No Normal: --  SalePrice
```

Per tant sembla que cap de les variables numèriques no segueix una distribució normal.

2.3.Tot seguit, comprovem la correlació entre els diferents atributs numèrics respecte a l'atribut SalePrice, aplicant la funció cor() que genera una matriu amb els percentatges de correlació entre les variables seleccionades.

```
cor(train[,c('OverallQual','OverallCond','YearBuilt','YearRemodAdd','MasVnrArea','BsmtFinSF1','BsmtFinSF2','BsmtUnfSF','TotalBsmtSF','1stFlrSF','2ndFlrSF','LowQualFinSF','GrLivArea','BsmtFullBath','BsmtHalfBath','FullBath','HalfBath','BedroomAbvGr','KitchenAbvGr','TotRmsAbvGrd','Fireplaces','GarageYrBlt','GarageCars','GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','3SsnPorch','ScreenPorch','PoolArea','MiscVal','MoSold','YrSold','SalePrice')], train['SalePrice'])
```

```
##                SalePrice
## OverallQual    0.79098160
## OverallCond   -0.07785589
## YearBuilt      0.52289733
## YearRemodAdd   0.50710097
## MasVnrArea             NA
## BsmtFinSF1     0.38641981
## BsmtFinSF2    -0.01137812
## BsmtUnfSF      0.21447911
## TotalBsmtSF    0.61358055
## 1stFlrSF       0.60585218
## 2ndFlrSF       0.31933380
## LowQualFinSF  -0.02560613
## GrLivArea      0.70862448
## BsmtFullBath   0.22712223
## BsmtHalfBath  -0.01684415
## FullBath       0.56066376
## HalfBath       0.28410768
## BedroomAbvGr   0.16821315
## KitchenAbvGr  -0.13590737
## TotRmsAbvGrd   0.53372316
## Fireplaces     0.46692884
## GarageYrBlt            NA
## GarageCars     0.64040920
## GarageArea     0.62343144
## WoodDeckSF     0.32441344
## OpenPorchSF    0.31585623
## EnclosedPorch -0.12857796
## 3SsnPorch      0.04458367
## ScreenPorch    0.11144657
## PoolArea       0.09240355
## MiscVal       -0.02118958
## MoSold         0.04643225
## YrSold        -0.02892259
## SalePrice      1.00000000
```

En aquest cas com els atributs numèrics no segueixen una distribució normal, s'usa Spearman per a fer una matriu de correlació de les variables quantitatives amb la variable Sale Price. Prèviament es normalitzen els valors usant el mètode scale que usa una normalització de tipus z-score.

```
train.scaled <- scale(train[,c('OverallQual','OverallCond','YearBuilt','YearRem
odAdd','MasVnrArea','BsmtFinSF1','BsmtFinSF2','BsmtUnfSF','TotalBsmtSF','1stFlr
SF','2ndFlrSF','LowQualFinSF','GrLivArea','BsmtFullBath','BsmtHalfBath','FullBa
th','HalfBath','BedroomAbvGr','KitchenAbvGr','TotRmsAbvGrd','Fireplaces','Garag
eYrBlt','GarageCars','GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','3
SsnPorch','ScreenPorch','PoolArea','MiscVal','MoSold','YrSold','SalePrice')])
#es converteix a data frame
train.scaled <- as.data.frame(scale(train.scaled))
```

Spearman:

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
for (i in 1:(ncol(train.scaled) - 1)) {
 if (is.integer(train.scaled[,i]) | is.numeric(train.scaled[,i])) {
   spearman_test = cor.test(train.scaled[,i],
   train.scaled[,length(train.scaled)], method = "spearman")
   corr_coef = spearman_test$estimate
   p_val = spearman_test$p.value
# Add row to matrix
   pair = matrix(ncol = 2, nrow = 1)
   pair[1][1] = corr_coef
   pair[2][1] = p_val
   corr_matrix <- rbind(corr_matrix, pair)
   rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(train.scaled)[i]
  }
}
print(corr_matrix)
```

```
##                  estimate       p-value
## OverallQual    0.80982859  0.000000e+00
## OverallCond   -0.12932495  7.118552e-07
## YearBuilt      0.65268155 5.693841e-178
## YearRemodAdd   0.57115898 3.557233e-127
## MasVnrArea     0.42130950  1.472613e-63
## BsmtFinSF1     0.30187120  3.857909e-32
## BsmtFinSF2    -0.03880613  1.383221e-01
## BsmtUnfSF      0.18519663  9.886861e-13
## TotalBsmtSF    0.60272544 4.157300e-145
## 1stFlrSF       0.57540784 1.780246e-129
## 2ndFlrSF       0.29359799  2.040344e-30
## LowQualFinSF  -0.06771915  9.645078e-03
## GrLivArea      0.73130958 1.431015e-244
## BsmtFullBath   0.22512487  3.130150e-18
## BsmtHalfBath  -0.01218888  6.416775e-01
## FullBath       0.63595706 2.729574e-166
## HalfBath       0.34300755  1.422950e-41
## BedroomAbvGr   0.23490672  9.402132e-20
## KitchenAbvGr  -0.16482575  2.358516e-10
## TotRmsAbvGrd   0.53258594 9.553211e-108
## Fireplaces     0.51924745 1.354698e-101
## GarageYrBlt    0.59378833 3.611689e-132
## GarageCars     0.69071097 1.654517e-207
## GarageArea     0.64937853 1.320918e-175
## WoodDeckSF     0.35380161  2.688484e-44
## OpenPorchSF    0.47756066  4.860098e-84
## EnclosedPorch -0.21839362  3.180474e-17
## 3SsnPorch      0.06544022  1.238409e-02
## ScreenPorch    0.10006972  1.281429e-04
## PoolArea       0.05845300  2.551713e-02
## MiscVal       -0.06272700  1.652526e-02
## MoSold         0.06943224  7.955957e-03
## YrSold        -0.02989913  2.535700e-01
```

Les variables més fortament correlacionades, observades amb valors estimats més alts són:

- OverallQual: Rates the overall material and finish of the house
- GrLivArea: Above grade (ground) living area square feet
- YearBuilt: Original construction date
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet

Les variables menys interessants són les que tenen una estimació més a prop de 0. Per tal de reduir la dimensionalitat es pot prescindir de les variables entre -0.2 i 0.2 que no tindrien massa correlació amb el Sale Price:

- OverallCond
- BsmtFinSF2
- BsmtUnfSF
- LowQualFinSF
- BsmtHalfBath

- KitchenAbvGr
- 3SsnPorch
- ScreenPorch
- PoolArea
- MiscVal
- MoSold
- YrSold

```
train = subset(train, select = -c(OverallCond,BsmtFinSF2,BsmtUnfSF,LowQualFinS
F,
BsmtHalfBath,BsmtHalfBath,KitchenAbvGr,`3SsnPorch`,ScreenPorch,PoolArea,
MiscVal,MoSold,YrSold) )
```

4.4 Tractament d'outliers

4.4.1 Identificació

1.Variable de classe-dependent: Sale Price

```
summary(train$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

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

```
##  [1] 345000 385000 438780 383970 372402 412500 501837 475000 386250 403000
## [11] 415298 360000 375000 342643 354000 377426 437154 394432 426000 555000
## [21] 440000 380000 374000 430000 402861 446261 369900 451950 359100 345000
## [31] 370878 350000 402000 423000 372500 392000 755000 361919 341000 538000
## [41] 395000 485000 582933 385000 350000 611657 395192 348000 556581 424870
## [51] 625000 392500 745000 367294 465000 378500 381000 410000 466500 377500
## [61] 394617
```

Observem com la majoria d'outliers són preus alts comparats amb la mitja que és de 163000. Es donen per valors vàlids, donat que s'observen característiques d'alt standing en aquestes cases, el que fa suposar que els valors són vàlids

```
train.data = train[c("OverallQual","GrLivArea","YearBuilt","GarageCars","Garage
Area","SalePrice")]


filter(train.data,SalePrice %in% (boxplot.stats(train$SalePrice)$out))
```

```
##    OverallQual GrLivArea YearBuilt GarageCars GarageArea SalePrice
## 1            9      2324      2005          3        736    345000
## 2            9      1842      1981          3        894    385000
## 3           10      2945      2006          3        641    438780
## 4            7      2696      2007          3        792    383970
## 5            8      1710      2007          3        866    372402
## 6            9      2668      2003          3        726    412500
## 7            9      2234      2008          3       1166    501837
## 8           10      3608      1892          3        840    475000
## 9           10      2392      2003          3        968    386250
## 10           8      2794      1995          3        810    403000
## 11           9      2121      2006          3        732    415298
## 12           9      1944      2003          3        708    360000
## 13           7      2036      1965          2        513    375000
## 14           9      2596      2006          3        840    342643
## 15           8      2468      2004          3        872    354000
## 16           9      1922      2005          3        676    377426
## 17           9      2728      2005          3        706    437154
## 18           9      1856      2010          3        834    394432
## 19          10      2332      2007          3        846    426000
## 20          10      2402      2008          3        672    555000
## 21           8      1976      2006          3        908    440000
## 22           9      2643      2006          3        694    380000
## 23           9      1792      2003          3        874    374000
## 24           8      3228      1992          2        546    430000
## 25          10      2020      2009          3        900    402861
## 26           9      2713      2008          3        858    446261
## 27           8      2028      2005          3        880    369900
## 28          10      2296      2008          3        842    451950
## 29           8      3194      1934          2        380    359100
## 30           8      2704      1972          2        538    345000
## 31           9      1766      2009          3        478    370878
## 32           8      2113      1995          3        839    350000
## 33           8      2448      1994          3        711    402000
## 34           8      2097      2005          3       1134    423000
## 35           8      2046      2008          3        834    372500
## 36           8      1419      2007          2        567    392000
## 37          10      4316      1994          3        832    755000
## 38           8      2576      2006          3        666    361919
## 39           7      2418      1993          3        983    341000
## 40           8      3279      2003          3        841    538000
## 41           8      1973      2006          3        895    395000
## 42           9      3140      2008          3        820    485000
## 43           9      2822      2008          3       1020    582933
## 44          10      2084      2007          3       1220    385000
## 45           9      2224      2004          3        738    350000
## 46           9      2364      2009          3        820    611657
## 47           9      1940      2009          3        606    395192
## 48           8      2392      1997          3        870    348000
## 49           9      2868      2005          3        716    556581
## 50           8      2828      2006          3       1052    424870
```

```
## 51          10       3627       1995          3        807    625000
## 52           8       1652       2008          2        482    392500
## 53          10       4476       1996          3        813    745000
## 54           9       1702       2008          3       1052    367294
## 55          10       2076       2006          3        850    465000
## 56           9       2018       2008          3        746    378500
## 57           8       3447       1935          3       1014    381000
## 58           8       3238       1995          3        666    410000
## 59          10       2633       2001          3        804    466500
## 60           9       1746       2006          3        758    377500
## 61           8       1932       2008          3        774    394617
```

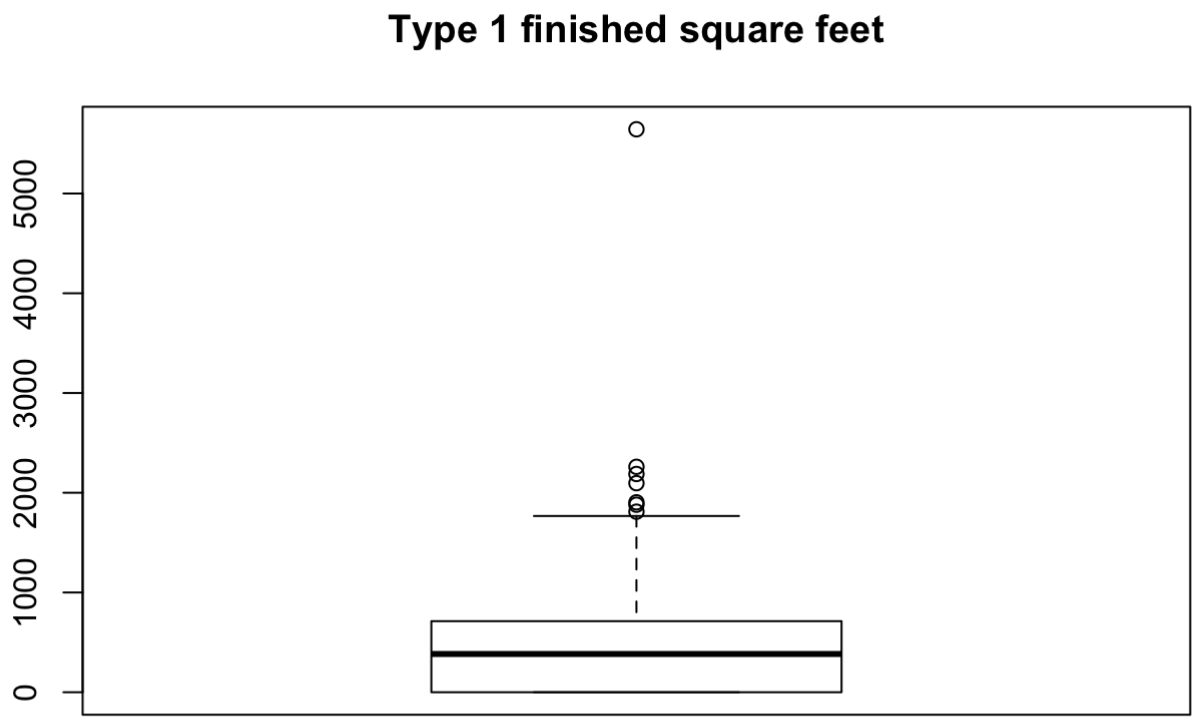
2.Variables independents:

Probablement són cases de finals del segle XIX.

```
boxplot.stats(train$YearBuilt)$out
```

```
## [1] 1880 1880 1880 1882 1880 1875 1872
```

```
bp <-boxplot(train$BsmtFinSF1,main="Type 1 finished square feet")
```




```
bp$out
```

```
## [1] 1810 1880 1904 2260 2188 2096 5644
```

```
#boxplot.stats(train$BsmtFinSF1)$out
```

```
boxplot.stats(train$TotalBsmtSF)$out
```

```
##  [1]    0    0 2223    0    0    0 2216    0 2392    0 2121 2136 3206    0
## [15]    0    0    0 3094 2153 3200    0 3138    0    0    0    0 2109 2077
## [29] 2444    0    0    0    0 2078    0 2217    0    0 2330    0    0    0
## [43]    0 2524    0    0    0    0    0 2396 2158    0    0 2136    0 2076
## [57] 2110 6110    0 2633    0
```

Hi ha o un total de metres quadrats de la basement area (sotan) molt gran > 2000 o bé no hi ha sòtan i és 0

```
boxplot.stats(train$`1stFlrSF`)$out
```

```
##  [1] 2207 2223 2259 2158 2234 2392 2402 3228 3138 2515 2444 2217 2364 2898
## [15] 2524 2411 2196 4692 2156 2633
```

```
boxplot.stats(train$`2ndFlrSF`)$out
```

```
## [1] 1872 2065
```

```
boxplot.stats(train$GrLivArea)$out
```

```
##  [1] 2945 3222 3608 3112 2794 3493 2978 3228 4676 2775 3194 3395 4316 3279
## [15] 3140 2822 2872 2898 3082 2868 2828 3627 3086 2872 4476 3447 5642 2810
## [29] 2792 3238 2784
```

```
boxplot.stats(train$TotRmsAbvGrd)$out
```

```
##  [1] 11 11 12 11 11 11 11 14 11 12 11 12 11 11 12 11 12 11 12 11 11 12 12
## [24] 11 11 12 12 12 11 11
```

```
boxplot.stats(train$Bedroom)$out
```

```
##  [1] 0 5 5 6 0 5 6 5 5 6 5 6 5 0 8 5 6 5 5 6 5 5 5 5 5 5 5 0 0 5 0 5 6 5 5
```

```
boxplot.stats(train$Fireplaces)$out
```

```
## [1] 3 3 3 3 3
```

```
boxplot.stats(train$GarageYrBlt)$out
```

```
## numeric(0)
```

quan no hi ha garatge l'any de construcció s'emplena amb un 0 que seria un outlier.

```
boxplot.stats(train$GarageCars)$out
```

```
## [1] 4 4 4 4 4
```

```
boxplot.stats(train$GarageArea)$out
```

```
##  [1] 1166  968 1053 1025  947 1390 1134  983 1020 1220 1248 1043 1052  995
## [15] 1356 1052  954 1014 1418  968 1069
```

```
boxplot.stats(train$WoodDeckSF)$out
```

```
##  [1] 857 576 476 574 441 468 670 495 536 519 466 517 426 503 486 486 511
## [18] 421 550 509 474 728 436 431 448 439 635 500 668 586 431 736
```

(Wood deck area in square feet outliers)

```
boxplot.stats(train$OpenPorchSF)$out
```

```
##  [1] 204 213 258 199 234 184 205 228 238 260 198 172 208 228 184 250 175
## [18] 195 214 231 192 187 176 523 285 406 182 502 274 172 243 235 312 267
## [35] 265 288 341 204 174 247 291 312 418 240 364 188 207 234 192 191 252
## [52] 189 282 224 319 244 185 200 180 263 304 234 240 192 229 211 198 287
## [69] 292 207 241 547 211 184 262 210 236
```

(Open porch area in square feet outliers)

```
boxplot.stats(train$EnclosedPorch)$out
```

```
##   [1] 272 228 205 176 205  87 172 102  37 144  64 114 202 128 156  44  77
##  [18] 144 192 144 140 180 228 128 183  39 184  40 552  30 126  96  60 150
##  [35] 120 202  77 112 252  52 224 234 144 244 268 137  24 108 294 177 218
##  [52] 242  91 112 160 130 184 126 169 105  34  96 248 236 120  32  80 115
##  [69] 291 184 116 158 112 210  36 156 144 200  84 148 116 120 136 102 240
##  [86]  54 112  39 100  36 189 293 164  40 216 239 112 252 240 180  67  90
## [103] 120  56 112 129  40  98 143 216 234 112 112  70 386 154 185 156 156
## [120] 134 196 264 185 275  96 120 112 116 230 254  68 194 192  34 150 164
## [137] 112 224  32 318 244  48  94 138 108 112 226 192 174 228  19 170 220
## [154] 128  80 115 137 192 252 112  96 176 216 176 214 280  96 116 102 190
## [171] 236 192  84 330 208 145 259 126 264  81 164  42 123 162 100 286 190
## [188] 168  20 301 198  96 221 112 212  50 150 168 112 160 114 216 154  99
## [205] 158 216 252 112
```

Enclosed porch area in square feet outliers: La mitja són 21.95 i el màxim 552 no havent-hi ni primer ni tercer quartil. Per tant podem dir que en la majoria de casos no hi ha enclosed porch, però quan hi ha és de molts metres quadrats.

4.4.2 Anàlisi dels outliers:

Tots el outliers analitzats tenen sentint dintre del domini de cada atribut, per tant es decideix no aplicar cap mètode corrector.

4.5 Creació de noves variables numèriques a partir de les categòriques

Abans hem trobat la correlació de les variables numèriques amb SalePrice fent servir la funció cor() i la correlació d'Spearman. Ara, intentarem fer el mateix però amb les variables categòriques.

Per tal d'aconseguir-ho crearem nous atributs numèrics a partir dels atributs categòrics existents. Això ho podrem fer revisant cadascuna de les variables i definint uns valors numèrics que descriguin de forma equivalent les categories representades.

La nomenclatura que farem servir per assignar aquests nous atributs serà **vn** al principi del nom, per tal d'indicar que es tracta d'un valor numèric.

Començarem per la variable Street, que té dos possibles valors: Grvl o Pave; és a dir el camí d'accés a la vivenda pot ser de grava o pavimentat.

```
#Comprovem els possibles valors de la variable Street
table(train$Street)
```

```
##
## Grvl Pave
##    6 1454
```

```
#Creem la nova variable numèrica assignant 1 o 0 en funció de si el carrer està
pavimentat (1) o no (0)
train$vnStreet[train$Street == "Pave"] <- 1
train$vnStreet[train$Street != "Pave"] <- 0
```

L'atribut LotShape ens descriu la forma de la vivenda en 4 nivells desde regular (millor) fins a irregular (pitjor).

```
#Comprovem els possibles valors de la variable LotShape
table(train$LotShape)
```

```
##
## IR1 IR2 IR3 Reg
## 484  41  10 925
```

```
#Creem la nova variable numèrica que pot agafar 4 possibles valors: regular(4), una mica irregular (3), bastant irregular (2), irregular (1)
train$vnLotShape[train$LotShape == "Reg"] <- 4
train$vnLotShape[train$LotShape == "IR1"] <- 3
train$vnLotShape[train$LotShape == "IR2"] <- 2
train$vnLotShape[train$LotShape == "IR3"] <- 1
```

LandContour indica el desnivell de la vivenda.

```
#Comprovem els possibles valors de la variable LandContour
table(train$LandContour)
```

```
##
##  Bnk  HLS  Low  Lvl
##   63   50   36 1311
```

```
#Creem la nova variable numèrica en la que considerarem únicament dos possibles valors: a nivell (1) o amb desnivell (0)
train$vnLandContour[train$LandContour == "Lvl"] <- 1
train$vnLandContour[train$LandContour != "Lvl"] <- 0
```

Utilities descriu els serveis de que consta la vivenda.

```
#Comprovem els possibles valors de la variable Utilities
table(train$Utilities)
```

```
##
## AllPub NoSeWa
##   1459      1
```

```
#La nova variable numèrica prendrà 4 possibles valors: AllPub(4), NoSewr(3), No SeWa(2), ELO(1)
train$vnUtilities[train$Utilities == "AllPub"] <- 4
train$vnUtilities[train$Utilities == "NoSewr"] <- 3
train$vnUtilities[train$Utilities == "NoSeWa"] <- 2
train$vnUtilities[train$Utilities == "ELO"] <- 1
```

LandSlope indica el pendent del terreny.

```
#Comprovem els possibles valors de la variable LandSlope
table(train$LandSlope)
```

```
##
##  Gtl  Mod  Sev
## 1382   65   13
```

```
#La nova variable numèrica prendrà 3 possibles valors: Gtl(3), Mod(2), Sev(1)
train$vnLandSlope[train$LandSlope == "Gtl"] <- 3
train$vnLandSlope[train$LandSlope == "Mod"] <- 2
train$vnLandSlope[train$LandSlope == "Sev"] <- 1
```

Amb LotConfig per tal de donar un valor numèric que tingui sentit, tenint en compte que no es pot establir cap ordre de millor o pitjor per a la configuració de la propietat, el que farem es fixar-nos en el preu mig per a cada tipus de configuració. D'aquesta manera podrem ordenar de la configuració més cara a la més barata les vivendes.

```
#Comprovem quin és el SalePrice mig per a cadascuna de les configuracions
summarize(group_by(train, LotConfig), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 5 x 2
##   LotConfig `mean(SalePrice, na.rm = T)`
##   <chr>                            <dbl>
## 1 Corner                         181623.
## 2 CulDSac                        223855.
## 3 FR2                            177935.
## 4 FR3                            208475
## 5 Inside                         176938.
```

```
#Ara assignem 5 possibles valors numèrics a la nova variable, seguint un ordre
 de configuració més valuosa a menys: CulDSac(5), FR3(4), Corner(3), FR2(2), In
side(1)
train$vnLotConfig[train$LotConfig == "CulDSac"] <- 5
train$vnLotConfig[train$LotConfig == "FR3"] <- 4
train$vnLotConfig[train$LotConfig == "Corner"] <- 3
train$vnLotConfig[train$LotConfig == "FR2"] <- 2
train$vnLotConfig[train$LotConfig == "Inside"] <- 1
```

Per assignar un valor numèric a l'atribut Neighborhood farem el mateix. Definirem el preu mig de venta de les vivendes de cada veïnat, i a continuació classificarem aquests barris en 3 grups en funció del preu mig de les vivendes entre barats(1), preu mitjà(2) i cars(3).

```
#Comprovem quin és el preu de venta mig per a les vivendes de cadascun dels bar
ris
nbhdprice <- summarize(group_by(train, Neighborhood),mean(SalePrice, na.rm=T))
#Definim com a barri barat (nbhdprice_lo) els barris amb un preu mitjà inferior
a 140.000, barri mitjà (nbhdprice_med) els que tenen un preu entre 140.000 i 20
0.000 i barri car (nbhdprice_hi) els que les propietats tenen un preu superior
 a 200.000
nbhdprice_lo <- filter(nbhdprice, nbhdprice$`mean(SalePrice, na.rm = T)` < 1400
00)
nbhdprice_med <- filter(nbhdprice, nbhdprice$`mean(SalePrice, na.rm = T)` < 200
000 & nbhdprice$`mean(SalePrice, na.rm = T)` >= 140000 )
nbhdprice_hi <- filter(nbhdprice, nbhdprice$`mean(SalePrice, na.rm = T)` >= 200
000)
#Finalment assignem 3 possibles valors depenent del tipus de veinat: nbhdprice_
hi(3), nbhdprice_med(2), nbhdprice_lo(1)
train$vnNeighborhood[train$Neighborhood %in% nbhdprice_lo$Neighborhood] <- 1
train$vnNeighborhood[train$Neighborhood %in% nbhdprice_med$Neighborhood] <- 2
train$vnNeighborhood[train$Neighborhood %in% nbhdprice_hi$Neighborhood] <- 3
```

Els atributs Condition1 i Condition2 els convertirem a valors numèrics d'igual forma. I farem servir el mateix mètode que en els anteriors casos, comprovant el preu de venta mig per a cadascuna de les categories i establint un ordre per a que els valors numèrics tinguin sentit.

```
#Comprovem SalePrice per cada opció
summarize(group_by(train, Condition1),mean(SalePrice, na.rm=T))
```

```
## # A tibble: 9 x 2
##   Condition1 `mean(SalePrice, na.rm = T)`
##   <chr>                             <dbl>
## 1 Artery                          135092.
## 2 Feedr                           142475.
## 3 Norm                            184495.
## 4 PosA                            225875
## 5 PosN                            215184.
## 6 RRAe                            138400
## 7 RRAn                            184397.
## 8 RRNe                            190750
## 9 RRNn                            212400
```

```
summarize(group_by(train, Condition2),mean(SalePrice, na.rm=T))
```

```
## # A tibble: 8 x 2
##   Condition2 `mean(SalePrice, na.rm = T)`
##   <chr>                             <dbl>
## 1 Artery                          106500
## 2 Feedr                           121167.
## 3 Norm                            181169.
## 4 PosA                            325000
## 5 PosN                            284875
## 6 RRAe                            190000
## 7 RRAn                            136905
## 8 RRNn                             96750
```

```
#A partir del resultat anterior veiem que PosA i PosN tenen un preu més elevat
 i es diferencien de la resta tant en Condition1 com en Condition2. Definim dos
valors numèrics 1 i 0, en funció de si l'atribut té valor PosA o PosN o no.
train$vnCondition1[train$Condition1 %in% c("PosA", "PosN")] <- 1
train$vnCondition1[!train$Condition1 %in% c("PosA", "PosN")] <- 0
train$vnCondition2[train$Condition2 %in% c("PosA", "PosN")] <- 1
train$vnCondition2[!train$Condition2 %in% c("PosA", "PosN")] <- 0
```

BldgType indica el tipus de construcció

```
#Comprovem quin és el SalePrice mig per a cada BldgType
summarize(group_by(train, BldgType), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 5 x 2
##   BldgType `mean(SalePrice, na.rm = T)`
##   <chr>                             <dbl>
## 1 1Fam                            185764.
## 2 2fmCon                          128432.
## 3 Duplex                          133541.
## 4 Twnhs                           135912.
## 5 TwnhsE                          181959.
```

```
#Ara assignem valors numèrics a la nova variable, seguint un ordre de tipus d'e
dificació de més valuosa a menys: 1Fam(5), TwnhsE(4), Twnhs(3), Duplex(2), 2fmC
on(1)
train$vnBldgType[train$BldgType == "1Fam"] <- 5
train$vnBldgType[train$BldgType == "TwnhsE"] <- 4
train$vnBldgType[train$BldgType == "Twnhsr"] <- 3
train$vnBldgType[train$BldgType == "Duplex"] <- 2
train$vnBldgType[train$BldgType == "2fmCon"] <- 1
```

HouseStyle descriu el tipus de vivenda

```
#Comprovem quin és el SalePrice mig per a cada HouseStyle
housestyle_price <- summarize(group_by(train, HouseStyle), mean(SalePrice, na.r
m=T))
#Definim 3 categories en funció si el preu és inferior a 140.000 (housestyle_l
o), es troba entre 140.000 i 200.000 (housestyle_med), o és superior a 200.000
 (housestyle_hi)
housestyle_lo <- filter(housestyle_price, housestyle_price$`mean(SalePrice, na.
rm = T)` < 140000)
housestyle_med <- filter(housestyle_price, housestyle_price$`mean(SalePrice, n
a.rm = T)` < 200000 & housestyle_price$`mean(SalePrice, na.rm = T)` >= 140000 )
housestyle_hi <- filter(housestyle_price, housestyle_price$`mean(SalePrice, na.
rm = T)` >= 200000)
#Finalment assignem 3 possibles valors depenent del tipus d'habitatge: housesty
le_hi(3), housestyle_med(2), housestyle_lo(1)
train$vnHouseStyle[train$HouseStyle %in% housestyle_lo$HouseStyle] <- 1
train$vnHouseStyle[train$HouseStyle %in% housestyle_med$HouseStyle] <- 2
train$vnHouseStyle[train$HouseStyle %in% housestyle_hi$HouseStyle] <- 3
```

RoofStyle descriu el tipus de teulada.

```
#Comprovem quin és el SalePrice mig per a cada RoofStyle
summarize(group_by(train, RoofStyle), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 6 x 2
##   RoofStyle `mean(SalePrice, na.rm = T)`
##   <chr>                           <dbl>
## 1 Flat                           194690
## 2 Gable                          171484.
## 3 Gambrel                        148909.
## 4 Hip                            218877.
## 5 Mansard                        180568.
## 6 Shed                           225000
```

```
#Ara assignem valors numèrics a la nova variable, seguint un ordre de tipus de
 teulada de més valuosa a menys: Shed(6), Hip(5), Flat(4), Mansard(3), Gable
(2), Gambrel(1)
train$vnRoofStyle[train$RoofStyle == "Shed"] <- 6
train$vnRoofStyle[train$RoofStyle == "Hip"] <- 5
train$vnRoofStyle[train$RoofStyle == "Flat"] <- 4
train$vnRoofStyle[train$RoofStyle == "Mansard"] <- 3
train$vnRoofStyle[train$RoofStyle == "Gable"] <- 2
train$vnRoofStyle[train$RoofStyle == "Gambrel"] <- 1
```

RoofMatl descriu el material amb que està feta la teulada.

```
#Comprovem quina és la mitja del SalePrice per cada RoofMatl
summarize(group_by(train, RoofMatl), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 8 x 2
##   RoofMatl `mean(SalePrice, na.rm = T)`
##   <chr>                           <dbl>
## 1 ClyTile                        160000
## 2 CompShg                        179804.
## 3 Membran                        241500
## 4 Metal                          180000
## 5 Roll                           137000
## 6 Tar&Grv                        185406.
## 7 WdShake                        241400
## 8 WdShngl                        390250
```

```
#Assignem valor 1 o 0 a cada material en funció de si la mitja del preu de vent
a de les vivendes amb teulades construides amb aquell material és superior a 20
0.000 (1) o inferior (0).
train$vnRoofMatl[train$RoofMatl %in% c("Membran", "WdShake", "WdShngl")] <- 1
train$vnRoofMatl[!train$RoofMatl %in% c("Membran", "WdShake", "WdShngl")] <- 0
```

Exterior1st i Exterior2nd descriuen el tipus d'acabats del recobriment exterior de les parets de la vivenda.

```
#Comprovem quina és la mitja del SalePrice per a cada tipus d'Exterior1st
ext1_price <- summarize(group_by(train, Exterior1st), mean(SalePrice, na.rm=T))
#Definim 3 categories en funció si el preu és inferior a 140.000 (ext1_lo), es
 troba entre 140.000 i 200.000 (ext1_med), o és superior a 200.000 (ext1_hi)
ext1_lo <- filter(ext1_price, ext1_price$`mean(SalePrice, na.rm = T)` < 140000)
ext1_med <- filter(ext1_price, ext1_price$`mean(SalePrice, na.rm = T)` < 200000
& ext1_price$`mean(SalePrice, na.rm = T)` >= 140000 )
ext1_hi <- filter(ext1_price, ext1_price$`mean(SalePrice, na.rm = T)` >= 200000
)
#Finalment assignem 3 possibles valors depenent del tipus d'exterior: ext1_hi
(3), ext1_med(2), ext1_lo(1)
train$vnExterior1st[train$Exterior1st %in% ext1_lo$Exterior1st] <- 1
train$vnExterior1st[train$Exterior1st %in% ext1_med$Exterior1st] <- 2
train$vnExterior1st[train$Exterior1st %in% ext1_hi$Exterior1st] <- 3
#Comprovem quina és la mitja del SalePrice per a cada tipus d'Exterior2nd
ext2_price <- summarize(group_by(train, Exterior2nd), mean(SalePrice, na.rm=T))
#Definim 3 categories en funció si el preu és inferior a 140.000 (ext2_lo), es
 troba entre 140.000 i 200.000 (ext2_med), o és superior a 200.000 (ext2_hi)
ext2_lo <- filter(ext2_price, ext2_price$`mean(SalePrice, na.rm = T)` < 140000)
ext2_med <- filter(ext2_price, ext2_price$`mean(SalePrice, na.rm = T)` < 200000
& ext2_price$`mean(SalePrice, na.rm = T)` >= 140000 )
ext2_hi <- filter(ext2_price, ext2_price$`mean(SalePrice, na.rm = T)` >= 200000
)
#Finalment assignem 3 possibles valors depenent del tipus d'exterior: ext1_hi
(3), ext1_med(2), ext1_lo(1)
train$vnExterior2nd[train$Exterior2nd %in% ext2_lo$Exterior2nd] <- 1
train$vnExterior2nd[train$Exterior2nd %in% ext2_med$Exterior2nd] <- 2
train$vnExterior2nd[train$Exterior2nd %in% ext2_hi$Exterior2nd] <- 3
```

MasVnrType indica el tipus de mamposteria.

```
#Comprovem quina és la mitja del SalePrice per cada MasVnrType
summarize(group_by(train, MasVnrType),mean(SalePrice, na.rm=T))
```

```
## # A tibble: 5 x 2
##   MasVnrType `mean(SalePrice, na.rm = T)`
##   <chr>                            <dbl>
## 1 BrkCmn                         146318.
## 2 BrkFace                        204692.
## 3 None                           156222.
## 4 Stone                          265584.
## 5 <NA>                           236484.
```

Observem que hi han valors NA.

```
#Assignem valor 1 o 0 a cada material en funció de si la mitja del preu de vent
a de les vivendes amb mamposteria feta amb aquell material és superior a 200.00
0 (1) o inferior (0).
train$vnMasVnrType[train$MasVnrType %in% c("Stone", "BrkFace") | is.na(train$Ma
sVnrType)] <- 1
train$vnMasVnrType[!train$MasVnrType %in% c("Stone", "BrkFace") & !is.na(train
$MasVnrType)] <- 0
```

ExterQual, ens diu la qualitat dels materials emprats en els exteriors.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (5), Good (4), Ave
rage/Typical (3), Fair(2), Poor (1)
train$vnExterQual[train$ExterQual == "Ex"] <- 5
train$vnExterQual[train$ExterQual == "Gd"] <- 4
train$vnExterQual[train$ExterQual == "TA"] <- 3
train$vnExterQual[train$ExterQual == "Fa"] <- 2
train$vnExterQual[train$ExterQual == "Po"] <- 1
```

ExterCond indica l'estat actual dels materials emprats en els exteriors.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (5), Good (4), Ave
rage/Typical (3), Fair(2), Poor (1)
train$vnExterCond[train$ExterCond == "Ex"] <- 5
train$vnExterCond[train$ExterCond == "Gd"] <- 4
train$vnExterCond[train$ExterCond == "TA"] <- 3
train$vnExterCond[train$ExterCond == "Fa"] <- 2
train$vnExterCond[train$ExterCond == "Po"] <- 1
```

Foundation indica el tipus de fonaments de la vivenda.

```
#Comprovem quin és el SalePrice mig per a cada tipus de fonaments.
summarize(group_by(train, Foundation), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 6 x 2
##   Foundation `mean(SalePrice, na.rm = T)`
##   <chr>                            <dbl>
## 1 BrkTil                         132291.
## 2 CBlock                         149806.
## 3 PConc                          225230.
## 4 Slab                           107366.
## 5 Stone                          165959.
## 6 Wood                           185667.
```

```
#Ara assignem valors numèrics a la nova variable, seguint un ordre dels fonamen
ts més valorats als que menys: PConc(6), Wood(5), Stone(4), CBlock(3), BrkTil
(2), Slab(1)
train$vnFoundation[train$Foundation == "PConc"] <- 6
train$vnFoundation[train$Foundation == "Wood"] <- 5
train$vnFoundation[train$Foundation == "Stone"] <- 4
train$vnFoundation[train$Foundation == "CBlock"] <- 3
train$vnFoundation[train$Foundation == "BrkTil"] <- 2
train$vnFoundation[train$Foundation == "Slab"] <- 1
```

BsmtQual, indica la qualitat del soterrani en funció de la seva alçada.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (6), Good (5), Typ
ical (4), Fair (3), Poor (2), No Basement (1)
train$vnBsmtQual[train$BsmtQual == "Ex"] <- 6
train$vnBsmtQual[train$BsmtQual == "Gd"] <- 5
train$vnBsmtQual[train$BsmtQual == "TA"] <- 4
train$vnBsmtQual[train$BsmtQual == "Fa"] <- 3
train$vnBsmtQual[train$BsmtQual == "Po"] <- 2
train$vnBsmtQual[is.na(train$BsmtQual)] <- 1
```

BsmtCond, fa una evaluació general de les condicions actuals del soterrani.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (6), Good (5), Typ
ical (4), Fair (3), Poor (2), No Basement (1)
train$vnBsmtCond[train$BsmtCond == "Ex"] <- 6
train$vnBsmtCond[train$BsmtCond == "Gd"] <- 5
train$vnBsmtCond[train$BsmtCond == "TA"] <- 4
train$vnBsmtCond[train$BsmtCond == "Fa"] <- 3
train$vnBsmtCond[train$BsmtCond == "Po"] <- 2
train$vnBsmtCond[is.na(train$BsmtCond)] <- 1
```

BsmtExposure fa referència a la visibilitat del soterrani.

```
#Assignem valors numèrics de més qualitat a menys: Good Exposure (5), Average E
xposure (4), Minimum Exposure (3), No Exposure (2), No Basement (1)
train$vnBsmtExposure[train$BsmtExposure == "Gd"] <- 5
train$vnBsmtExposure[train$BsmtExposure == "Av"] <- 4
train$vnBsmtExposure[train$BsmtExposure == "Mn"] <- 3
train$vnBsmtExposure[train$BsmtExposure == "No"] <- 2
train$vnBsmtExposure[is.na(train$BsmtExposure)] <- 1
```

BsmtFinType1 i BsmtFinType2 fa referència a la qualitat del soterrani i la seva habitabilitat.

```
#Assignem valors numèrics de més qualitat a menys: Good Living Quarters (7), Av
erage Living Quarters (6), Below Average Living Quarters (5), Average Rec Room
 (4), Low Quality (3), Unfinished (2), No Basement (1)
train$vnBsmtFinType1[train$BsmtFinType1 == "GLQ"] <- 7
train$vnBsmtFinType1[train$BsmtFinType1 == "ALQ"] <- 6
train$vnBsmtFinType1[train$BsmtFinType1 == "BLQ"] <- 5
train$vnBsmtFinType1[train$BsmtFinType1 == "Rec"] <- 4
train$vnBsmtFinType1[train$BsmtFinType1 == "LwQ"] <- 3
train$vnBsmtFinType1[train$BsmtFinType1 == "Unf"] <- 2
train$vnBsmtFinType1[is.na(train$BsmtFinType1)] <- 1
train$vnBsmtFinType2[train$BsmtFinType2 == "GLQ"] <- 7
train$vnBsmtFinType2[train$BsmtFinType2 == "ALQ"] <- 6
train$vnBsmtFinType2[train$BsmtFinType2 == "BLQ"] <- 5
train$vnBsmtFinType2[train$BsmtFinType2 == "Rec"] <- 4
train$vnBsmtFinType2[train$BsmtFinType2 == "LwQ"] <- 3
train$vnBsmtFinType2[train$BsmtFinType2 == "Unf"] <- 2
train$vnBsmtFinType2[is.na(train$BsmtFinType2)] <- 1
```

Heating descriu el tipus de calefacció.

```
#Comprovem quin és el SalePrice mig per a cada tipus de calefacció
summarize(group_by(train, Heating), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 6 x 2
##   Heating `mean(SalePrice, na.rm = T)`
##   <chr>                          <dbl>
## 1 Floor                          72500
## 2 GasA                          182021.
## 3 GasW                          166632.
## 4 Grav                           75271.
## 5 OthW                          125750
## 6 Wall                           92100
```

```
#Ara assignem valors numèrics a la nova variable, seguint un ordre del tipus de calefacció que dóna més valor a la vivenda al que menys: GasA(6), GasW(5), OthW (4), Wall(3), Grav(2), Floor(1)
train$vnHeating[train$Heating == "GasA"] <- 6
train$vnHeating[train$Heating == "GasW"] <- 5
train$vnHeating[train$Heating == "OthW"] <- 4
train$vnHeating[train$Heating == "Wall"] <- 3
train$vnHeating[train$Heating == "Grav"] <- 2
train$vnHeating[train$Heating == "Floor"] <- 1
```

HeatingQC ens diu la qualitat i l'estat de la instal·lació de la calefacció.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (5), Good (4), Average/Typical (3), Fair (2), Poor (1)
train$vnHeatingQC[train$HeatingQC == "Ex"] <- 5
train$vnHeatingQC[train$HeatingQC == "Gd"] <- 4
train$vnHeatingQC[train$HeatingQC == "TA"] <- 3
train$vnHeatingQC[train$HeatingQC == "Fa"] <- 2
train$vnHeatingQC[train$HeatingQC == "Po"] <- 1
```

CentralAir indica si l'habitatge té aire acondicionat central.

```
#Assignem el valor 1 si té AC i 0 si no en té.
train$vnCentralAir[train$CentralAir == "Y"] <- 1
train$vnCentralAir[train$CentralAir == "N"] <- 0
```

Electrical descriu el tipus d'instal·lació elèctrica de la vivenda.

```
#Comprovem quin és el SalePrice mig de les vivendes per a cada tipus d'instal·lació
summarize(group_by(train, Electrical), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 6 x 2
##   Electrical `mean(SalePrice, na.rm = T)`
##   <chr>                             <dbl>
## 1 FuseA                           122197.
## 2 FuseF                           107675.
## 3 FuseP                            97333.
## 4 Mix                              67000
## 5 SBrkr                           186825.
## 6 <NA>                            167500
```

```
#Ara assignem valors numèrics a la nova variable, seguint un ordre del tipus
 d'instal·lació que dóna més valor a la vivenda al que menys: SBrkr(6), NA(5),
 FuseA(4), FuseF(3), FuseP(2), Mix(1)
train$vnElectrical[train$Electrical == "SBrkr"] <- 6
train$vnElectrical[is.na(train$Electrical)] <- 5
train$vnElectrical[train$Electrical == "FuseA"] <- 4
train$vnElectrical[train$Electrical == "FuseF"] <- 3
train$vnElectrical[train$Electrical == "FuseP"] <- 2
train$vnElectrical[train$Electrical == "Mix"] <- 1
```

KitchenQual, valora la qualitat general de la cuina.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (5), Good (4), Ave
rage/Typical (3), Fair (2), Poor (1)
train$vnKitchenQual[train$KitchenQual == "Ex"] <- 5
train$vnKitchenQual[train$KitchenQual == "Gd"] <- 4
train$vnKitchenQual[train$KitchenQual == "TA"] <- 3
train$vnKitchenQual[train$KitchenQual == "Fa"] <- 2
train$vnKitchenQual[train$KitchenQual == "Po"] <- 1
```

FireplaceQu, indica la qualitat del foc a terra.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (6), Good (5), Ave
rage (4), Fair (3), Poor (2), No Fireplace (1)
train$vnFireplaceQu[train$FireplaceQu == "Ex"] <- 6
train$vnFireplaceQu[train$FireplaceQu == "Gd"] <- 5
train$vnFireplaceQu[train$FireplaceQu == "TA"] <- 4
train$vnFireplaceQu[train$FireplaceQu == "Fa"] <- 3
train$vnFireplaceQu[train$FireplaceQu == "Po"] <- 2
train$vnFireplaceQu[is.na(train$FireplaceQu)] <- 1
```

GarageType, tipus de garatge.

```
#Comprovem quin és el SalePrice mig de les vivendes segons el tipus de garatge
summarize(group_by(train, GarageType), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 7 x 2
##   GarageType `mean(SalePrice, na.rm = T)`
##   <chr>                             <dbl>
## 1 2Types                          151283.
## 2 Attchd                          202893.
## 3 Basment                         160571.
## 4 BuiltIn                         254752.
## 5 CarPort                         109962.
## 6 Detchd                          134091.
## 7 <NA>                            103317.
```

```
#Ara assignem valors numèrics a la nova variable, seguint un ordre del tipus de
garatge que dóna més valor a la vivenda al que menys: BuiltIn(7), Attached(6),
 Basement(5), More than one type(4), Detached(3), Car Port(2), No Garage(1)
train$vnGarageType[train$GarageType == "BuiltIn"] <- 7
train$vnGarageType[train$GarageType == "Attchd"] <- 6
train$vnGarageType[train$GarageType == "Basment"] <- 5
train$vnGarageType[train$GarageType == "2Types"] <- 4
train$vnGarageType[train$GarageType == "Detchd"] <- 3
train$vnGarageType[train$GarageType == "CarPort"] <- 2
train$vnGarageType[is.na(train$GarageType)] <- 1
```

GarageFinish, ens diu si l'interior del garatge està acabat o no.

```
#Assignem valors numèrics en funció del grau dels acabats del garatge: Finished
(4), Rough Finished (3), Unfinished (2), No Garage (1)
train$vnGarageFinish[train$GarageFinish == "Fin"] <- 4
train$vnGarageFinish[train$GarageFinish == "RFn"] <- 3
train$vnGarageFinish[train$GarageFinish == "Unf"] <- 2
train$vnGarageFinish[is.na(train$GarageFinish)] <- 1
```

GarageQual, indica la qualitat general del garatge.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (6), Good (5), Ave
rage (4), Fair (3), Poor (2), No Garage (1)
train$vnGarageQual[train$GarageQual == "Ex"] <- 6
train$vnGarageQual[train$GarageQual == "Gd"] <- 5
train$vnGarageQual[train$GarageQual == "TA"] <- 4
train$vnGarageQual[train$GarageQual == "Fa"] <- 3
train$vnGarageQual[train$GarageQual == "Po"] <- 2
train$vnGarageQual[is.na(train$GarageQual)] <- 1
```

GarageCond, indica l'estat actual del garatge.

```
#Assignem valors numèrics de més qualitat a menys: Excellent (6), Good (5), Ave
rage (4), Fair (3), Poor (2), No Garage (1)
train$vnGarageCond[train$GarageCond == "Ex"] <- 6
train$vnGarageCond[train$GarageCond == "Gd"] <- 5
train$vnGarageCond[train$GarageCond == "TA"] <- 4
train$vnGarageCond[train$GarageCond == "Fa"] <- 3
train$vnGarageCond[train$GarageCond == "Po"] <- 2
train$vnGarageCond[is.na(train$GarageCond)] <- 1
```

PavedDrive indica si la calçada està asfaltada.

```
#Hi han 3 possibles valors: Paved (3), Partial Pavement (2), Dirt/Gravel (1)
train$vnPavedDrive[train$PavedDrive == "Y"] <- 3
train$vnPavedDrive[train$PavedDrive == "P"] <- 2
train$vnPavedDrive[train$PavedDrive == "N"] <- 1
```

Functional, indica la funcionalitat de l'habitatge (assumeix la opció Typical si no hi ha cap deducció garantida).

```
train$vnFunctional[train$Functional == "Typ"] <- 1
train$vnFunctional[train$Functional != "Typ"] <- 0
```

SaleType fa referència al tipus de venta.

```
#Comprovem quin és el preu mitjà de venta per a les vivendes en funció del tipus de venta
summarize(group_by(train, SaleType), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 9 x 2
##   SaleType `mean(SalePrice, na.rm = T)`
##   <chr>                           <dbl>
## 1 COD                           143973.
## 2 Con                           269600
## 3 ConLD                         138781.
## 4 ConLI                         200390
## 5 ConLw                         143700
## 6 CWD                           210600
## 7 New                           274945.
## 8 Oth                           119850
## 9 WD                            173402.
```

```
#Ara assignem valors numèrics i agrupem les variables que observem que tenen preus mitjans de venta semblants.
test$vnSaleType[test$SaleType %in% c("New", "Con")] <- 5
test$vnSaleType[test$SaleType %in% c("CWD", "ConLI")] <- 4
test$vnSaleType[test$SaleType %in% c("WD")] <- 3
test$vnSaleType[test$SaleType %in% c("COD", "ConLw", "ConLD")] <- 2
test$vnSaleType[test$SaleType %in% c("Oth")] <- 1
```

SaleCondition descriu les condicions en que s'ha realitzat la venta.

```
#Comprovem quin és el preu mitjà de venta per a les vivendes en funció de la condició de venta
summarize(group_by(train, SaleCondition), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 6 x 2
##   SaleCondition `mean(SalePrice, na.rm = T)`
##   <chr>                                <dbl>
## 1 Abnorml                            146527.
## 2 AdjLand                            104125
## 3 Alloca                             167377.
## 4 Family                             149600
## 5 Normal                             175202.
## 6 Partial                            272292.
```

```
#Ara assignem valors numèrics a les variables
train$vnSaleCondition[train$SaleCondition == "Partial"] <- 6
train$vnSaleCondition[train$SaleCondition == "Normal"] <- 5
train$vnSaleCondition[train$SaleCondition == "Alloca"] <- 4
train$vnSaleCondition[train$SaleCondition == "Family"] <- 3
train$vnSaleCondition[train$SaleCondition == "Abnorml"] <- 2
train$vnSaleCondition[train$SaleCondition == "Adjland"] <- 1
```

MSZoning identifica el tipus de zona on es realitza la venta.

```
#Comprovem quin és el preu mitjà de venta per a les vivendes en funció de la zo
na de venta
summarize(group_by(train, MSZoning), mean(SalePrice, na.rm=T))
```

```
## # A tibble: 5 x 2
##   MSZoning `mean(SalePrice, na.rm = T)`
##   <chr>                           <dbl>
## 1 C (all)                         74528
## 2 FV                             214014.
## 3 RH                             131558.
## 4 RL                             191005.
## 5 RM                             126317.
```

```
#Ara assignem valors numèrics a les variables
test$vnMSZoning[test$MSZoning %in% c("FV")] <- 5
test$vnMSZoning[test$MSZoning %in% c("RL")] <- 4
test$vnMSZoning[test$MSZoning %in% c("RH")] <- 3
test$vnMSZoning[test$MSZoning %in% c("RM")] <- 2
test$vnMSZoning[test$MSZoning %in% c("C (all)")] <- 1
```

MSSubClass identifica el tipus de vivenda venuda.

```
#Comprovem quin és el preu de venta mig per a les vivendes de cada tipus
subclassprice <- summarize(group_by(train, MSSubClass),mean(SalePrice, na.rm=T
))
#Definim com a tipus de vivenda barata (subclass_lo) les propietats amb un preu
mitjà inferior a 140.000, vivenda assequible (subclass_med) les que tenen un pr
eu entre 140.000 i 200.000 i propietat cara (subclass_hi) els tipus de propieta
ts que tenen un preu superior a 200.000
subclass_lo <- filter(subclassprice, subclassprice$`mean(SalePrice, na.rm = T)`
< 140000)
subclass_med <- filter(subclassprice, subclassprice$`mean(SalePrice, na.rm = T)
` < 200000 & subclassprice$`mean(SalePrice, na.rm = T)` >= 140000 )
subclass_hi <- filter(subclassprice, subclassprice$`mean(SalePrice, na.rm = T)`
>= 200000)
#Finalment assignem 3 possibles valors: subclass_hi(3), subclass_med(2), subcla
ss_lo(1)
train$vnMSSubClass[train$MSSubClass %in% subclass_lo$MSSubClass] <- 1
train$vnMSSubClass[train$MSSubClass %in% subclass_med$MSSubClass] <- 2
train$vnMSSubClass[train$MSSubClass %in% subclass_hi$MSSubClass] <- 3
```

A continuació s'observala correlació, donat que a les variables numèriques no seguien una distribució normal, s'aplica també Spearman sobre les variables transformades vn, afegint SalePrice al final com a variable dependent

```
train.vn <- train[,c("vnStreet","vnLotShape","vnLandContour","vnUtilities",
"vnLandSlope","vnLotConfig","vnNeighborhood","vnCondition1","vnCondition2","vnB
ldgType",
"vnHouseStyle","vnRoofStyle","vnRoofMatl","vnExterior1st","vnExterior2nd",
"vnMasVnrType","vnExterQual","vnExterCond","vnFoundation","vnBsmtQual","vnBsmtC
ond",
"vnBsmtExposure" ,"vnBsmtFinType1","vnBsmtFinType2","vnHeating","vnHeatingQC",
"vnCentralAir","vnElectrical","vnKitchenQual" ,"vnFireplaceQu","vnGarageType",
"vnGarageFinish","vnGarageQual","vnGarageCond","vnPavedDrive","vnFunctional",
"vnSaleCondition","SalePrice")]
```

Spearman en les variables vn

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")
for (i in 1:(ncol(train.vn) - 1)) {
# if( substr(colnames(train.vn[i]), start=1, stop=2)=="vn")
 # {
   spearman_test = cor.test(train.vn[,i],
   train.vn[,length(train.vn)], method = "spearman")
   corr_coef = spearman_test$estimate
   p_val = spearman_test$p.value
# Add row to matrix
   pair = matrix(ncol = 2, nrow = 1)
   pair[1][1] = corr_coef
   pair[2][1] = p_val
   corr_matrix <- rbind(corr_matrix, pair)
   rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(train.vn)[i]
  #}
}
print(corr_matrix)
```

```
##                    estimate       p-value
## vnStreet          0.04581419  8.012225e-02
## vnLotShape       -0.32105533  2.341724e-36
## vnLandContour    -0.01789297  4.945066e-01
## vnUtilities       0.01670961  5.234926e-01
## vnLandSlope      -0.05031026  5.461545e-02
## vnLotConfig       0.10502919  5.798914e-05
## vnNeighborhood    0.68525784 5.403916e-203
## vnCondition1      0.09586658  2.441008e-04
## vnCondition2      0.05517806  3.501794e-02
## vnBldgType        0.11570952  1.262589e-05
## vnHouseStyle      0.31338851  1.236615e-34
## vnRoofStyle       0.16378164  3.068415e-10
## vnRoofMatl        0.11053654  2.306410e-05
## vnExterior1st     0.41592760  3.653377e-62
## vnExterior2nd     0.41265735  3.997079e-61
## vnMasVnrType      0.41061097  1.762322e-60
## vnExterQual       0.68401380 5.605572e-202
## vnExterCond       0.01168189  6.555992e-01
## vnFoundation      0.57358006 1.755079e-128
## vnBsmtQual        0.67802625 3.696559e-197
## vnBsmtCond        0.26937252  1.088293e-25
## vnBsmtExposure    0.34420665  7.175863e-42
## vnBsmtFinType1    0.36162475  2.443878e-46
## vnBsmtFinType2    0.03981255  1.283765e-01
## vnHeating         0.12194854  2.966772e-06
## vnHeatingQC       0.49139191  1.347224e-89
## vnCentralAir      0.31328617  1.302833e-34
## vnElectrical      0.29757768  3.074431e-31
## vnKitchenQual     0.67284855 4.400509e-193
## vnFireplaceQu     0.53760183 3.933062e-110
## vnGarageType      0.59881437 8.906579e-143
## vnGarageFinish    0.63397362 5.937952e-165
## vnGarageQual      0.35108157  1.336417e-43
## vnGarageCond      0.33901490  1.360807e-40
## vnPavedDrive      0.28060152  8.018188e-28
## vnFunctional      0.13498257  2.255201e-07
## vnSaleCondition   0.31498304  6.752142e-35
```

Les variables categòriques més correlacionades amb el Preu són:

- vnNeighborhood: Barri. (Agrupació de barris en 3 grups segons el preu mitjà de les cases.)
- vnExterQual: Qualtitat del material al exterior
- vnKitchenQual: Qualitat de la cuina
- vnBsmtQual: Evalua el gruix del pis que està directament relacionat amb la qualitat.
- vnGarageFinish: Mesura la qualitat del garatge

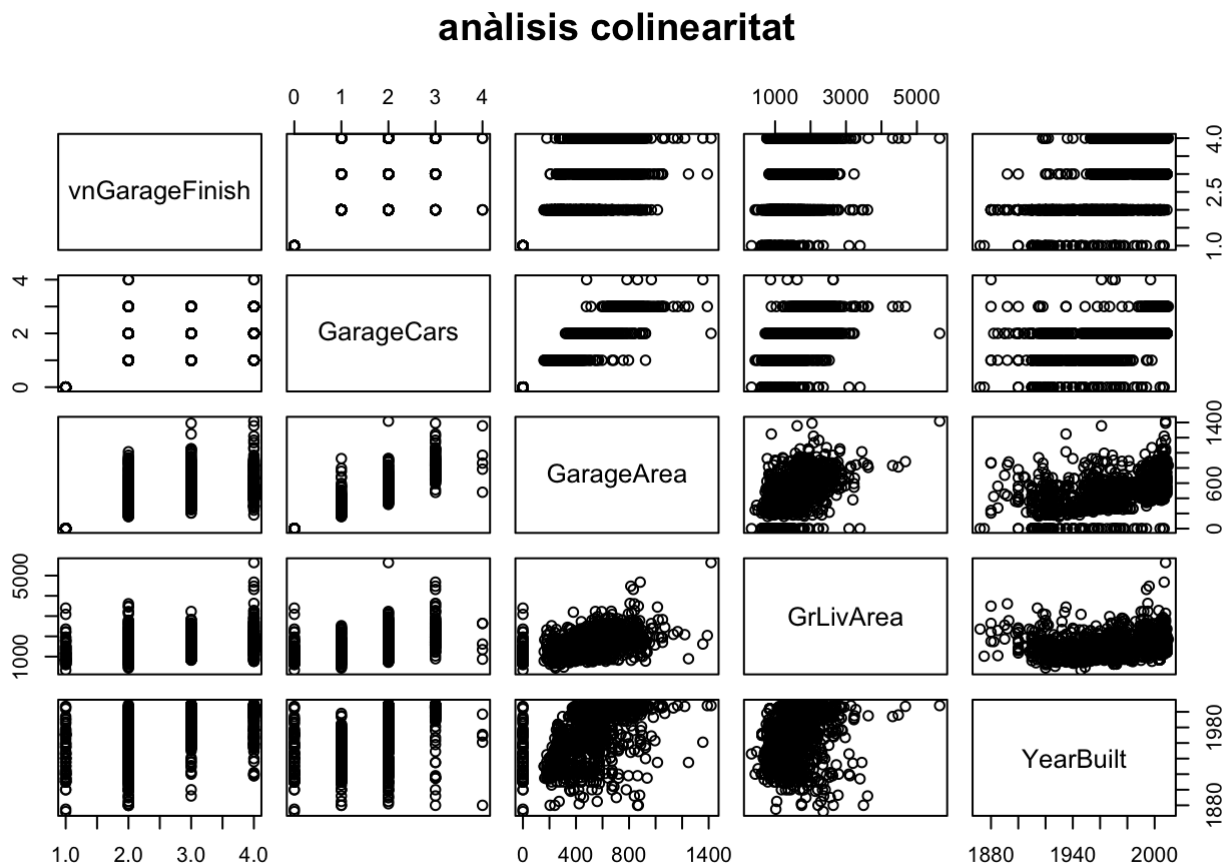
Recordem també les variables numèriques més correlacionades eren:

- OverallQual: Rates the overall material and finish of the house
- GrLivArea: Above grade (ground) living area square feet
- YearBuilt: Original construction date

- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet

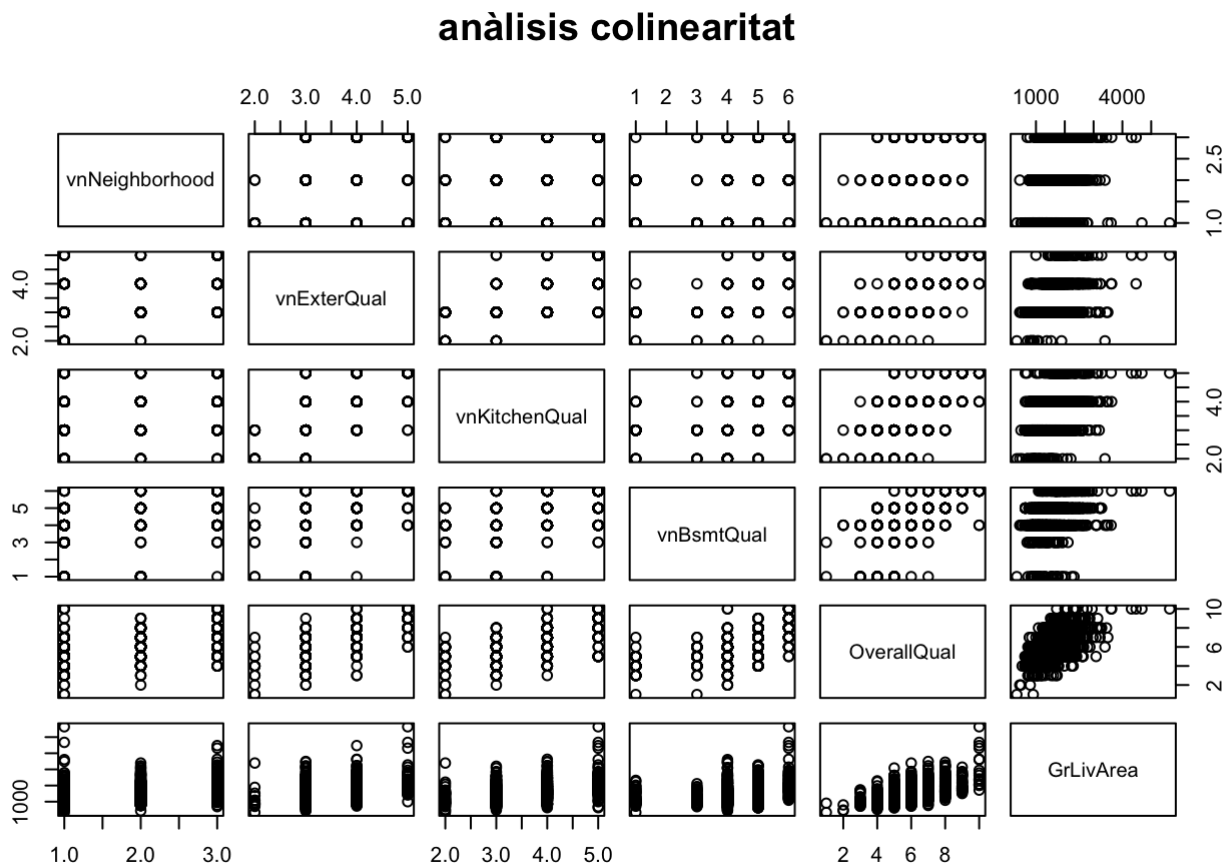
A continuació s'analitza la colinealitat entre aquestes variables en dos cops:

```
pairs(~vnGarageFinish+GarageCars+GarageArea+GrLivArea+YearBuilt,data=train,
   main="anàlisis colinearitat")
```



S'observa certa correlació entre Garage Area i YearBuilt.

```
pairs(~vnNeighborhood+vnExterQual+vnKitchenQual+vnBsmtQual+
        OverallQual+GrLivArea,data=train,
   main="anàlisis colinearitat")
```



s'observa una colinearitat molt més clara entre GrLivArea i OverallQual.

4.6 Comparació de barris en relació al preu de venda (Comparació de grups)

A la fase de creació de variables numèriques a partir de qualitatives es va veure que es podia establir un ordre en els barris d'acord amb el seu preu mitjà. A continuació es fa una prova d'hipòtesi per a comparar aquests barris com grups i determinar si són iguals en quant a la seva mitjana i homoscedasticitat. Com ja vam veure, la variable SalePrice no seguexi una distribució normal, a continuació es fa la prova d'homoscedasticitat amb el test de Fligner-Killeen.

Establim com a Hipòtesi 0 que les variances són iguals amb un nivell de confiança d'un 95% (nivell de significació per defecte 0.05)

```
fligner.test(SalePrice ~ Neighborhood, data = train)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SalePrice by Neighborhood
## Fligner-Killeen:med chi-squared = 282.78, df = 24, p-value <
## 2.2e-16
```

Com que el p-value <0.05 es rebutja la hipòtesi nul·la i per tant podem dir que el preu de venda presenta variàncies estadísticament diferents segons el barri.

A continuació es compara d'una forma no paramètrica les distribucions dels barris que van ser categoritzats com grup 1 i grup 2 aplicant el test de Wilcox.

```
  wilcox.test(SalePrice ~ vnNeighborhood, data = train, subset = vnNeighborhood
<3)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  SalePrice by vnNeighborhood
## W = 52771, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Per tant es comprova d'una manera més formal, que es refusa la hipòtesi nul·la perquè p-vale <0.05 que els grups de barris 1 i 2 són estadísticament diferents en quan al preu de venda de les cases.

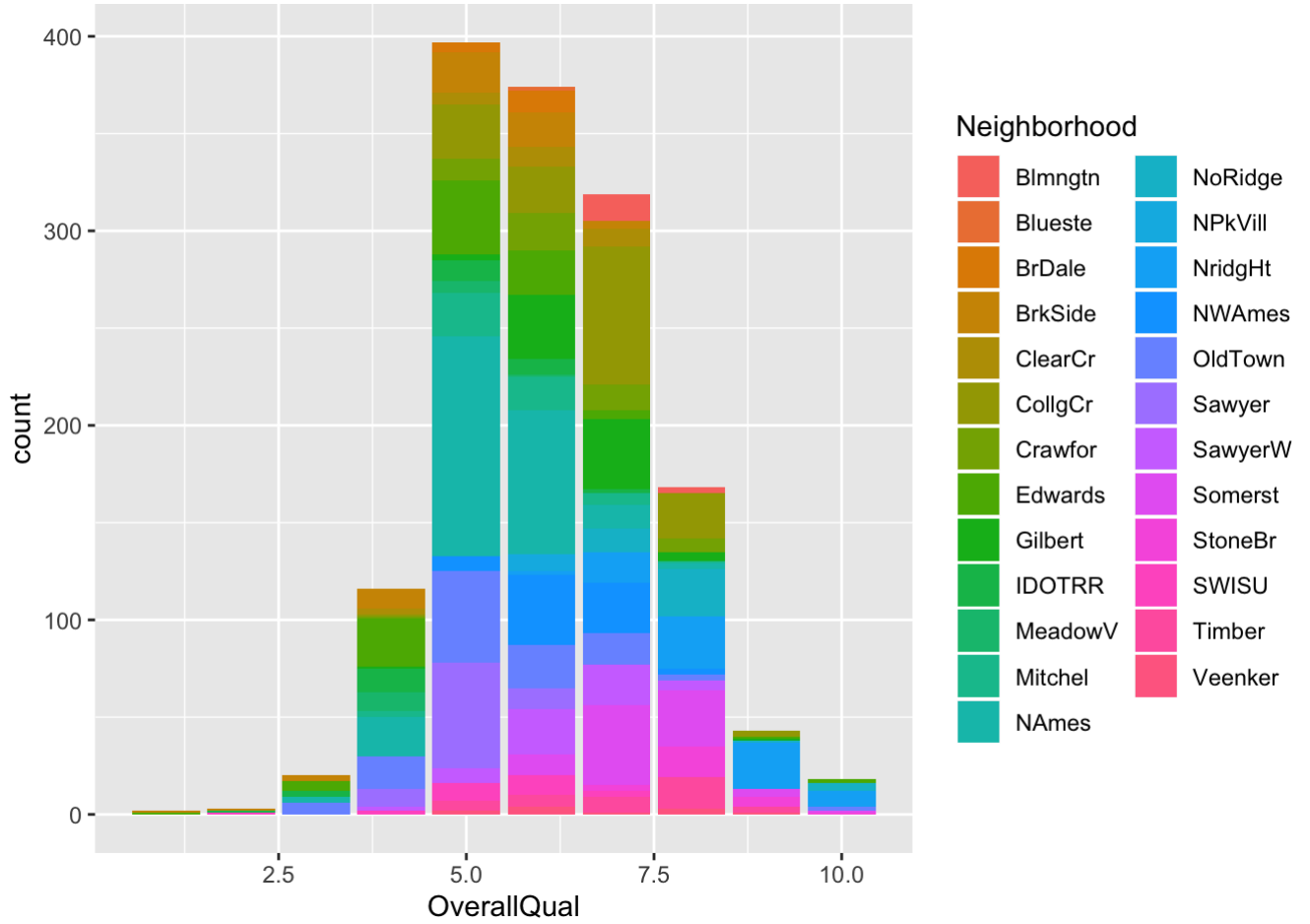
Cada barri està en un grup (segons la mitja de preus de venda. Grup 1 amb els barris que tenen les cases més barates, grup 3: barris amb les cases més cares)

```
unique ((train[c("vnNeighborhood","Neighborhood")]))
```

```
##     vnNeighborhood Neighborhood
## 1                2      CollgCr
## 2                3      Veenker
## 4                3      Crawfor
## 5                3      NoRidge
## 6                2      Mitchel
## 7                3      Somerst
## 8                2       NWAmes
## 9                1      OldTown
## 10               1      BrkSide
## 11               1       Sawyer
## 12               3      NridgHt
## 15               2        NAmes
## 19               2      SawyerW
## 22               1       IDOTRR
## 24               1      MeadowV
## 40               1      Edwards
## 42               3       Timber
## 51               2      Gilbert
## 59               3      StoneBr
## 70               3      ClearCr
## 127              2      NPkVill
## 220              2      Blmngtn
## 226              1       BrDale
## 268              2        SWISU
## 600              1      Blueste
```

s'analitza visualment per OverallQual (ratis de qualitat en quant acabats de la casa) els barris relacionats.

```
filas=dim(train)[1] #1460 rows
ggplot(data=train[1:filas,],aes(x=OverallQual,fill=Neighborhood))+geom_bar()
```



S'observa per exemple que la relació de qualitat de la casa en els valors de 9-10 està present en els barris del grup 3.

4.7 Comparació de preus de cases per decada de construcció de la casa

Hi ha una diferència entre els preus de venda dels anys 80 als anys 90?

Creació dels subsets:

```
train.SalePrice80s <-train$SalePrice[train$YearBuilt >1979 & train$YearBuilt <=
1989]
train.SalePrice90s <-train$SalePrice[train$YearBuilt >1989 & train$YearBuilt <=
2000]
```

Es comprova la normalitat, donat que serien subsets de la variable SalePrice. S'usa el mètode de Kolmogorov-Smirnov

```
ks.test(train.SalePrice80s, pnorm, mean(train.SalePrice80s), sd(train.SalePrice
80s))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  train.SalePrice80s
## D = 0.13989, p-value = 0.1836
## alternative hypothesis: two-sided
```

Assumint com a hipòtesi nul·la que la població està distribuïda normalment, Al ser el p-valor més gran que el nivell de significació,(per defecte α=0,05) la hipòtesi nul·la no es podria rebutjar i amb un 95% de confiança podríem dir que la distribució és normal.

Ara bé, passant el test de shapiro, la hipòtesi nul·la sí s'hauria de rebutjar:

```
shapiro.test(train.SalePrice80s)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train.SalePrice80s
## W = 0.91085, p-value = 0.000298
```

Es comprova si en la decada dels 90 els preus segueixen una distribució normal

```
ks.test(train.SalePrice90s, pnorm, mean(train.SalePrice90s), sd(train.SalePrice
90s))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  train.SalePrice90s
## D = 0.15275, p-value = 0.0003244
## alternative hypothesis: two-sided
```

Segons Kolmogorov-Smirnov, la distribució dels 90 no seria normal. Es comprova com tampoc ho és seguint Shapiro-Wilk.

```
shapiro.test(train.SalePrice90s)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train.SalePrice90s
## W = 0.72836, p-value < 2.2e-16
```

Es fa el test d'homoscedasticitat per a comprovar si les variances són iguals.

Es comprova l'anàlisi de variances amb el test no paramètric de Fligner. Com que el nombre d'elements en cada sample és diferent es fa un stack previ en un nou dataframe anomenat stacked:

```
stacked <- stack(list(train.SalePrice80s=train.SalePrice80s,train.SalePrice90s=
train.SalePrice90s))
 fligner.test(values ~ ind, data = stacked)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  values by ind
## Fligner-Killeen:med chi-squared = 0.37698, df = 1, p-value =
## 0.5392
```

Segons el test de Fligner-Killeen les variances seguirien una distribució similar entre SalesPrice 80 i 90.

Es comparen els preus dels 80s amb els 90s seguint el test no paramètric de Wilcoxon i Mann-Whitney. La hipòtesi nul·la seria la igualtat de les distribucions :

```
stacked <- stack(list(train.SalePrice80s=train.SalePrice80s,train.SalePrice90s=
train.SalePrice90s))
wilcox.test(values ~ ind, data = stacked)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  values by ind
## W = 3909.5, p-value = 0.0002273
## alternative hypothesis: true location shift is not equal to 0
```

```
#wilcox.test(values ~ ind, data = stacked, subset = Month %in% c(5, 8))
```

Donat que el test dona un p-value < 0.05, s'hauria de rebutjar la hipòtesi nul·la i per tant concloure que els preus són estadísticament diferents entre els 80 i els 90s

Donat que alguns tests són positius i altres negatius, es fa una representació visual també. Es pot concloure que la mitja del preu de venda, tot i diferent, no pujaria excessivament dels 80s als 90s. Possiblement un dels factors que està impactant en la comparació paramètrica dels preus, serien alguns preus molt alts (Outliers) trobats en la dècada dels 90.

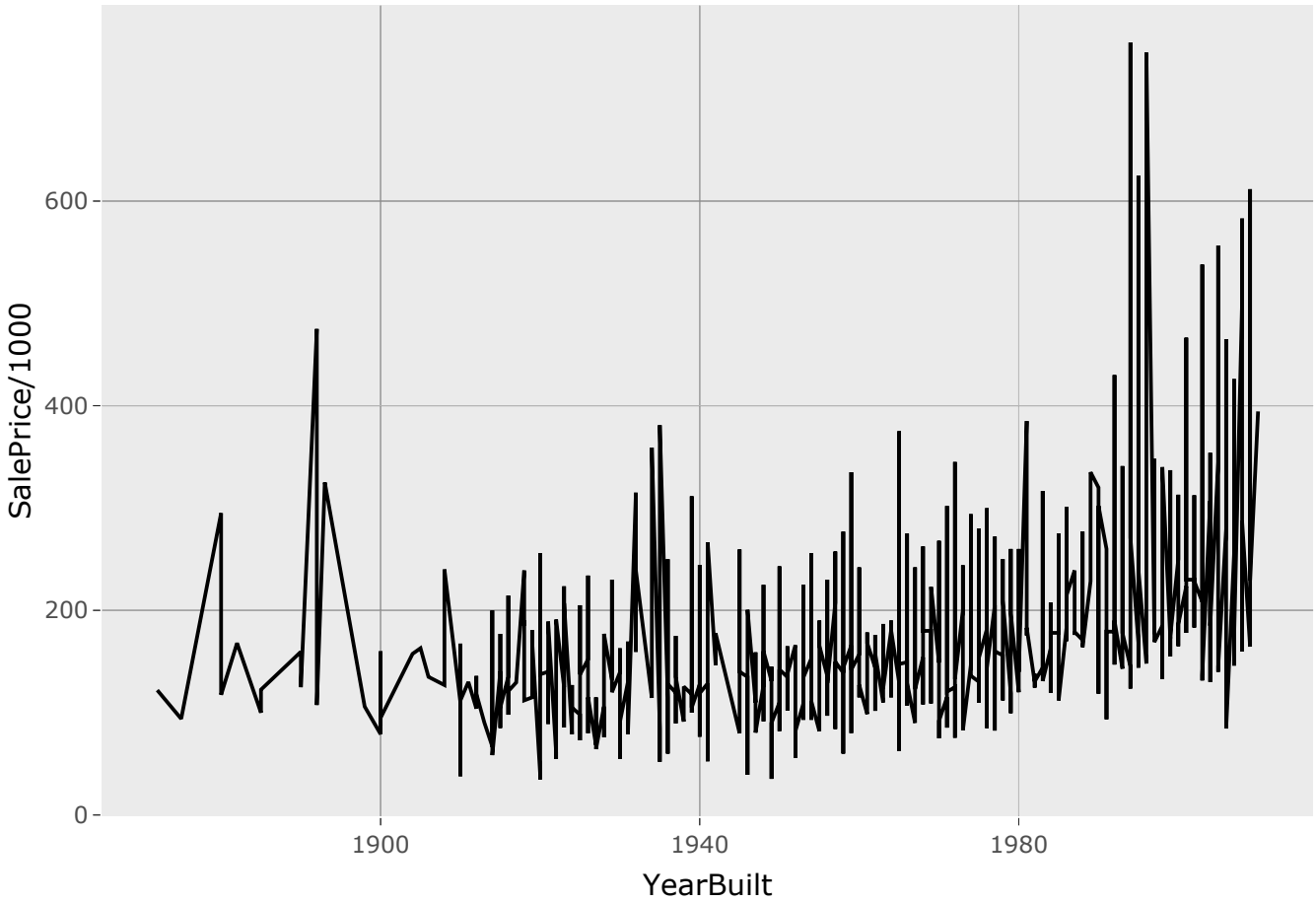
```
# Plot
summary(train.SalePrice80s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  112000  143000  178000  190080  215000  385000
```

```
summary(train.SalePrice90s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   93500  177250  204000  226141  250500  755000
```

```
# Plot
p <- ggplot(train, aes(x=YearBuilt, y=SalePrice/1000)) + geom_line()
ggplotly(p)
```



4.8 Anàlisi de l'evolució del preu de venda al llarg dels anys

En el nostre dataset tenim 4 variables de temps que ens poden servir per obtenir un anàlisi interessant sobre l'evolució dels preus de les vivendes al llarg del temps. Aquests 4 atributs són:

1. YearBuilt
2. YearRemodAdd
3. GarageYrBlt
4. YrSold

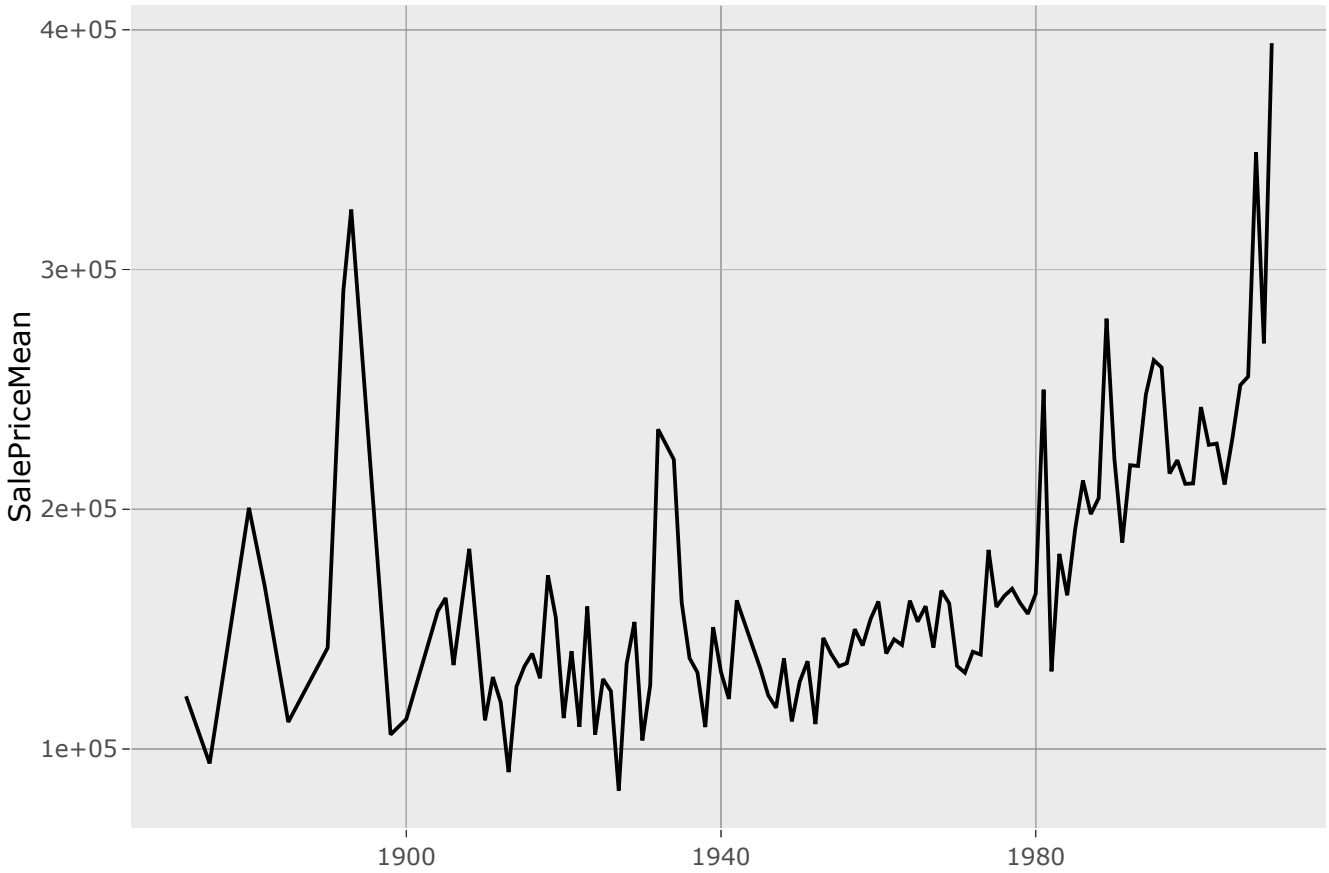
A continuació estudiarem el preu de venda en funció d'aquestes variables de temps.

```
#Tornem a llegir el dataset i agafem només les columnes que necessitem per aque
st anàlisi
trainTime <- read_csv('HousePrices/train.csv' )
trainTime = subset(trainTime, select = c(YearBuilt, YearRemodAdd, GarageYrBlt,
 YrSold, SalePrice) )
```

Primer de tot observem l'evolució del preu en funció de YearBuilt, és a dir, l'any de construcció de la vivenda.

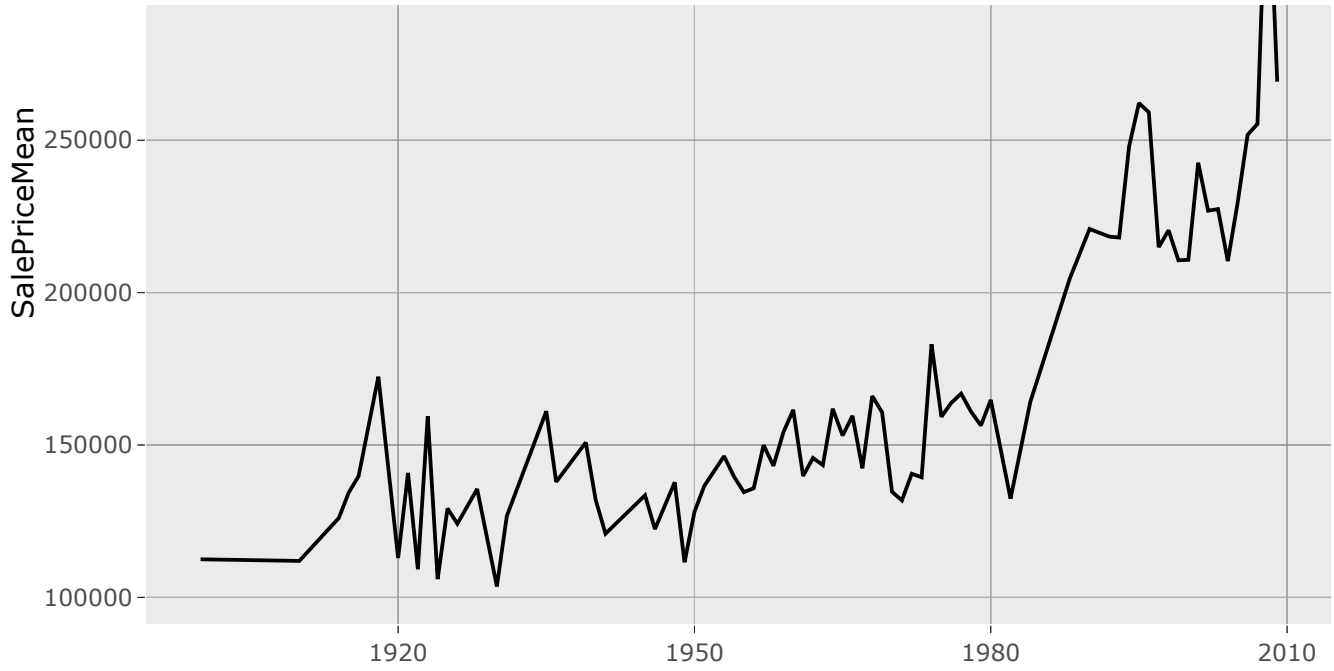
```
#Calculem la mitja del SalePrice en funció de YearBuilt
priceYearBuilt <- summarize(group_by(trainTime, YearBuilt), SalePriceMean = mea
n(SalePrice, na.rm=T), n=n())

#Mostrem el gràfic fent servir el paquet plotly que ens permet interactuar amb
 les dades mostrades
p <- ggplot(priceYearBuilt, aes(x=YearBuilt, y=SalePriceMean)) + geom_line() +
 xlab("")
ggplotly(p)
```



Veiem que en determinats anys el valor mitjà varia molt respecte als anys anteriors i posteriors. Això és degut a que en alguns casos tenim molt poques observacions per alguns anys concrets. Per a que la gràfica representi de forma més acurada l'evolució dels preus filtrem per a mostrar únicament els preus mitjans calculats tenint més de 5 observacions per aquell any.

```
priceYearBuilt5 <- filter(priceYearBuilt, n>5)
p5 <- ggplot(priceYearBuilt5, aes(x=YearBuilt, y=SalePriceMean)) + geom_line()
 + xlab("")
ggplotly(p5)
```



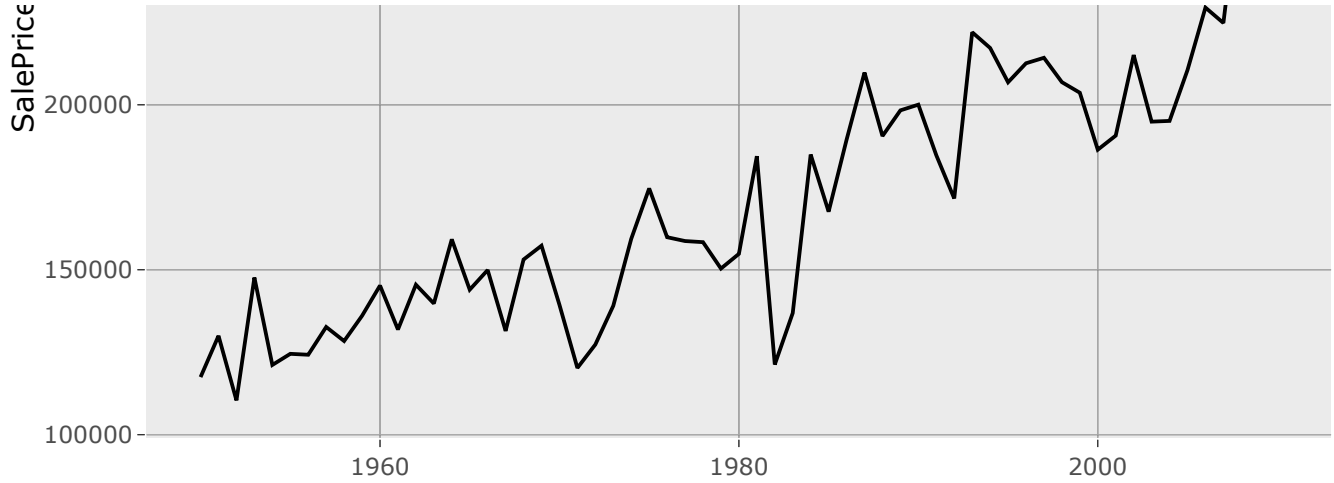
Veiem que excloent els anys amb poques observacions, en general, com més noves són les construccions més valor tenen les vivendes. Aquest fet té sentit, tot i que apreciem també que el preu no té una relació massa directa amb això, sinó que més aviat té a veure amb altres factors a nivell econòmic. Arribem a aquesta conclusió observant com a l'any 2009 hi ha una baixada molt pronunciada del preu de les vivendes respecte de l'any anterior, tot i ser més noves. Aquest fet és degut a l'efecte de la crisi econòmica que es va iniciar a partir del 2008 i que va afectar al preu de venda de les vivendes de nova construcció els anys següents.

Si observem el primer gràfic amb totes les dades, veiem que s'han venut algunes vivendes antigues (en concret una vivenda l'any 1893 i dues l'any 1892) amb un preu molt elevat. Analitzant amb detall aquestes observacions veiem que es tracta en tots els casos de vivendes que han estat remodelades posteriorment i de gran tamany i qualitat.

Ara observarem l'evolució dels preus per la variable YearRemodAdd.

```
#Calculem la mitja del SalePrice en funció de YearRemodAdd
priceYearRemodAdd <- summarize(group_by(trainTime, YearRemodAdd), SalePriceMean
= mean(SalePrice, na.rm=T), n=n())

#Mostrem el gràfic
p <- ggplot(priceYearRemodAdd, aes(x=YearRemodAdd, y=SalePriceMean)) + geom_lin
e() + xlab("")
ggplotly(p)
```



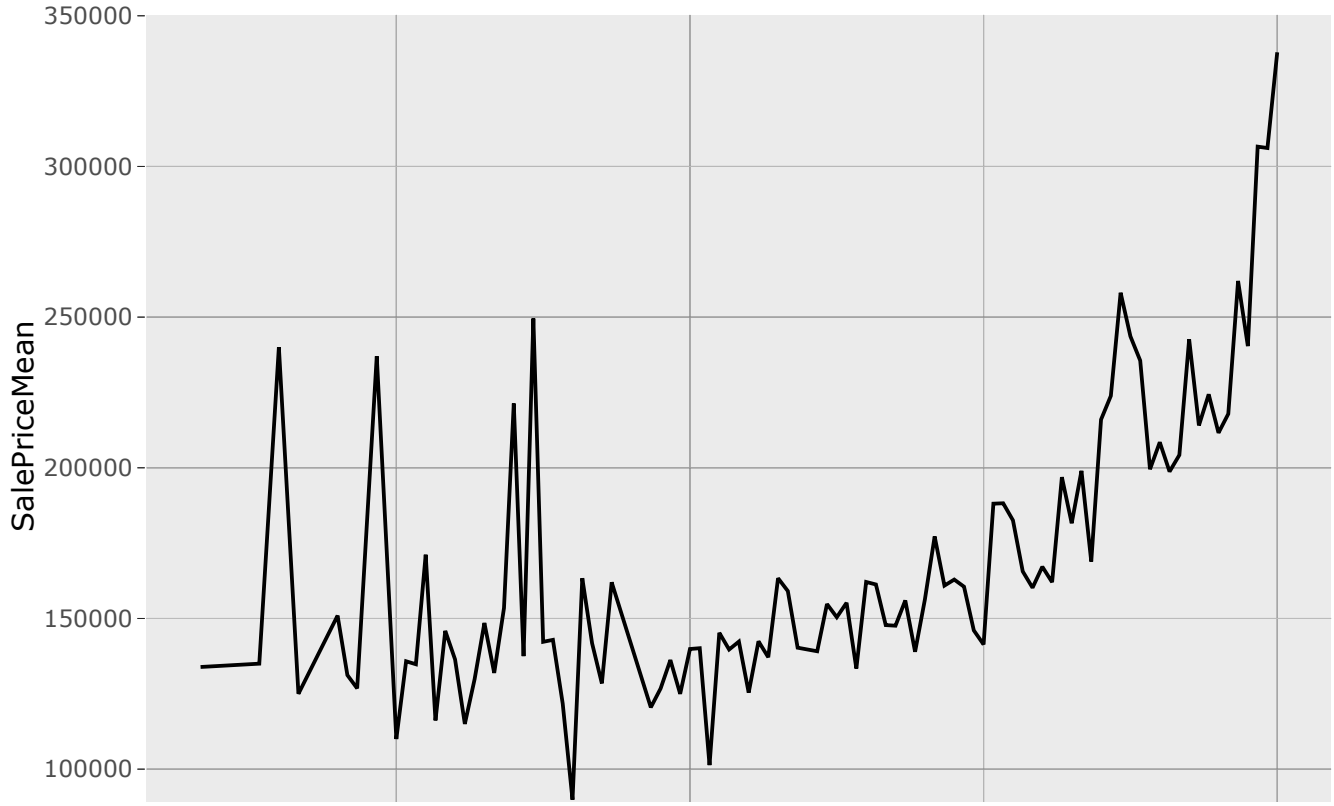
En aquest cas, totes les observacions són posteriors a l'any 1950 i no hi ha cap any amb menys de 4 observacions, amb la qual cosa no cal filtrar els anys amb poques observacions.

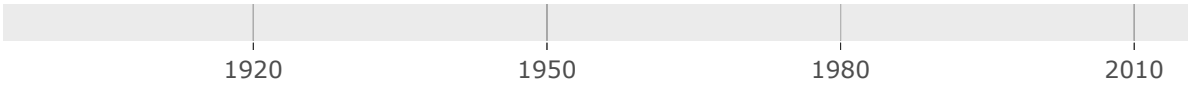
Pel que fa al gràfic observem com el preu de venda és clarament més elevat com més recent és la remodelació de la vivenda. Aquest increment del preu és especialment pronunciat si la reforma s'ha efectuat dintre dels 3 últims anys (en el cas de la mostra del 2007 al 2010).

Comprovem ara GarageYrBlt

```
#Calculem la mitja del SalePrice en funció de GarageYrBlt
priceGarageYrBlt <- summarize(group_by(trainTime, GarageYrBlt), SalePriceMean =
mean(SalePrice, na.rm=T), n=n())

#Mostrem el gràfic
p <- ggplot(priceGarageYrBlt, aes(x=GarageYrBlt, y=SalePriceMean)) + geom_line
() + xlab("")
ggplotly(p)
```



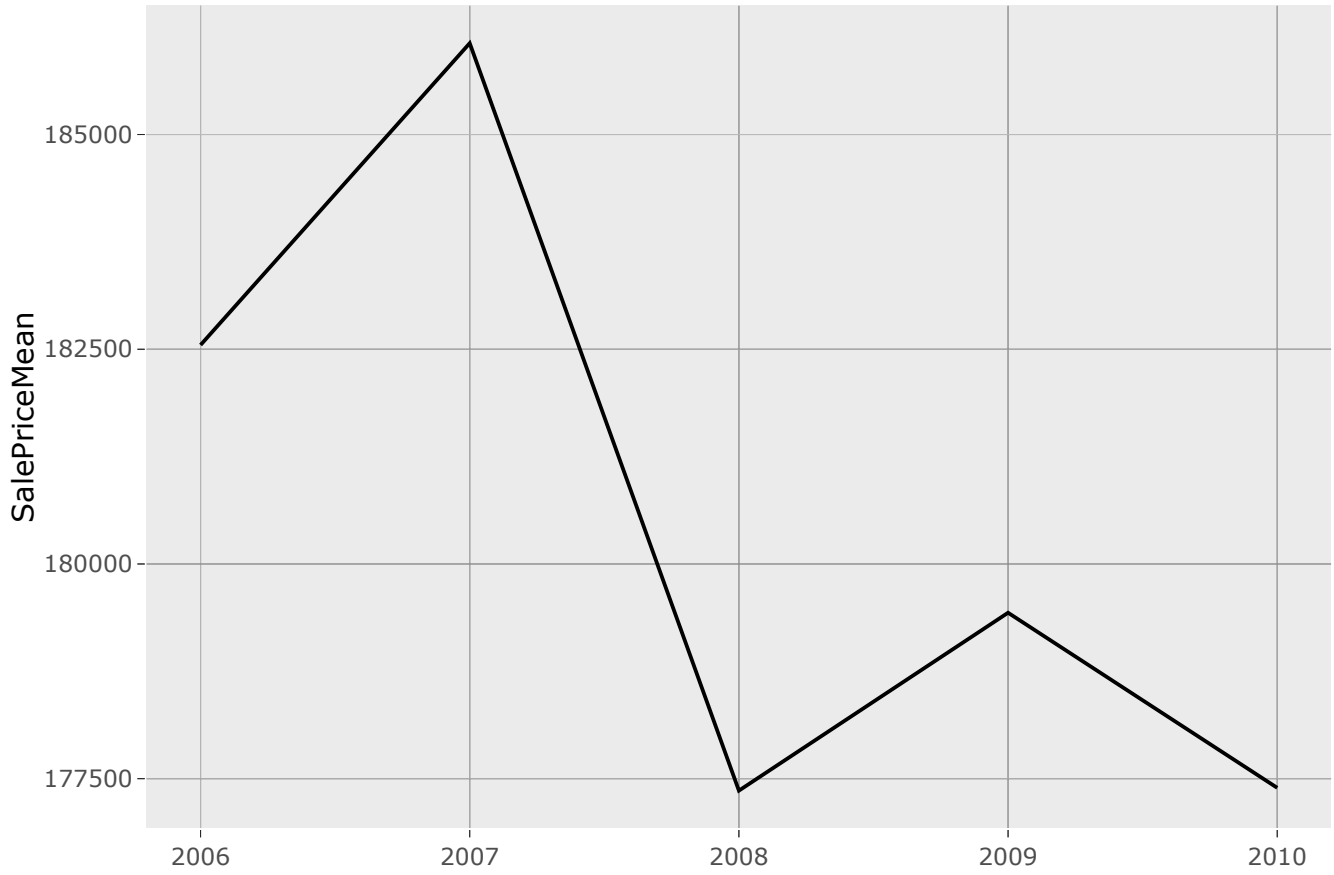


En aquest cas veiem que els resultats s'assemblen bastant als obtinguts per la variable YearBuilt. Aquesta sebmlança té sentit, ja que en la majoria dels casos, s'acostuma a construir el garatge al mateix temps que la resta de la casa.

Finalment, mirem la variable YrSold.

```
#Calculem la mitja del SalePrice en funció de YrSold
priceYrSold <- summarize(group_by(trainTime, YrSold), SalePriceMean = mean(Sale
Price, na.rm=T), n=n())

#Mostrem el gràfic
p <- ggplot(priceYrSold, aes(x=YrSold, y=SalePriceMean)) + geom_line() + xlab(
"")
ggplotly(p)
```



En aquest cas, les observacions que tenim mostren el preu de venda mig de les vivendes venudes entre els anys 2006 i 2010. Aquesta franja de temps és interessant, ja que com podem veure a la gràfica, es veu clarament la incidència de la crisi econòmica del 2008 en la baixada importantíssima dels preus de venda a partir d'aquell any.

5 Model: Regressió Lineal

Es fa una aproximació de la relació de dependència lineal entre la variable a predir: Sale Price i un subset de variables abans identificades com més correlacionades amb SalesPrice.

```
lmSalePrice1  = lm(SalePrice~OverallQual+YearBuilt+vnNeighborhood+vnKitchenQual
                  +FullBath+GarageCars+GarageArea+GrLivArea,data=train)
#summary(lmSalePrice)
```

Veiem com usant les variables originals el R-Squared millora, sent de 8.1.

```
lmSalePrice2  = lm(SalePrice~OverallQual+YearBuilt+Neighborhood+KitchenQual
                  +FullBath+GarageCars+GarageArea+GrLivArea,data=train)
#summary(lmSalePrice)
```

Amb altres variacions provades el R-Square seria menor.

```
lmSalePrice3  = lm(SalePrice~FullBath+YearBuilt+vnNeighborhood+vnKitchenQual
                  +OverallQual+GarageFinish+GrLivArea,data=train)
#summary(lmSalePrice)$r.squared
```

```
lmSalePrice4  = lm(SalePrice~FullBath+YearBuilt+Neighborhood+KitchenQual
                  +OverallQual+GarageFinish+GrLivArea+ExterQual+BsmtQual+TotalB
smtSF,data=train)
```

Es comprova que no hi ha un guany signiifcatiu afegint moltes variables predictores.

```
lmSalePrice5  = lm(SalePrice~MSSubClass+MSZoning+Neighborhood+KitchenQual
                  +OverallQual+GarageFinish+GrLivArea+ExterQual+BsmtQual+TotalB
smtSF+
                    LotArea+LotShape+LandContour+BldgType+HouseStyle+
                    YearRemodAdd+Exterior1st+ExterQual+Foundation+HalfBath+
                    BedroomAbvGr+Fireplaces+FireplaceQu+
                    GarageCars+GarageArea+GarageQual+PavedDrive+WoodDeckSF+Open
PorchSF
                   ,data=train)
```

```
print (paste("Model 1 -->", summary(lmSalePrice1)$r.squared))
```

```
## [1] "Model 1 --> 0.782895088977925"
```

```
print (paste("Model 2 -->", summary(lmSalePrice2)$r.squared))
```

```
## [1] "Model 2 --> 0.819081072671528"
```

```
print (paste("Model 3 -->", summary(lmSalePrice3)$r.squared))
```

```
## [1] "Model 3 --> 0.770873063098381"
```

```
print (paste("Model 4 -->", summary(lmSalePrice4)$r.squared))
```

```
## [1] "Model 4 --> 0.824611474027129"
```

```
print (paste("Model 5 -->", summary(lmSalePrice4)$r.squared))
```

```
## [1] "Model 5 --> 0.824611474027129"
```

```
# no hi ha guany significatiu afegint més variables print (paste("Model 5 -->", summary(lmSalePrice5)$r.squared))
```

El model amb un major coeficient de determinació és el model 5, però no s'obté una guany significatiu respecte el model 4 afegint un gran nombre de variables predictores. Per tant el model triat és el model 4, capaç d'explicar la variabilitat de les dades en un 81,9% (Adjusted R-Squared).

```
summary(lmSalePrice4)
```

```
##
## Call:
## lm(formula = SalePrice ~ FullBath + YearBuilt + Neighborhood +
##     KitchenQual + OverallQual + GarageFinish + GrLivArea + ExterQual +
##     BsmtQual + TotalBsmtSF, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -426134  -14303     -41   13595  215198
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.766e+05  1.653e+05  -1.069 0.285452
## FullBath             5.053e+02  2.694e+03   0.188 0.851232
## YearBuilt            1.348e+02  8.231e+01   1.637 0.101828
## NeighborhoodBlueste -1.365e+04  2.542e+04  -0.537 0.591416
## NeighborhoodBrDale  -1.909e+04  1.273e+04  -1.499 0.134033
## NeighborhoodBrkSide  6.293e+03  1.104e+04   0.570 0.568674
## NeighborhoodClearCr  3.715e+04  1.104e+04   3.365 0.000788 ***
## NeighborhoodCollgCr  1.979e+04  8.838e+03   2.239 0.025310 *
## NeighborhoodCrawfor  3.427e+04  1.055e+04   3.247 0.001197 **
## NeighborhoodEdwards -1.184e+04  9.934e+03  -1.192 0.233637
## NeighborhoodGilbert  1.171e+04  9.258e+03   1.264 0.206281
## NeighborhoodIDOTRR  -5.328e+03  1.183e+04  -0.450 0.652458
## NeighborhoodMeadowV -1.411e+04  1.333e+04  -1.058 0.290336
## NeighborhoodMitchel  9.424e+03  1.020e+04   0.924 0.355696
## NeighborhoodNAmes    6.517e+03  9.479e+03   0.688 0.491869
## NeighborhoodNoRidge  7.711e+04  1.013e+04   7.609 5.25e-14 ***
## NeighborhoodNPkVill  2.507e+02  1.438e+04   0.017 0.986089
## NeighborhoodNridgHt  4.474e+04  9.541e+03   4.689 3.03e-06 ***
## NeighborhoodNWAmes   9.735e+03  9.666e+03   1.007 0.314024
## NeighborhoodOldTown -7.763e+03  1.068e+04  -0.727 0.467255
## NeighborhoodSawyer   1.064e+04  9.997e+03   1.064 0.287582
## NeighborhoodSawyerW  1.685e+04  9.647e+03   1.747 0.080937 .
## NeighborhoodSomerst  2.393e+04  9.124e+03   2.622 0.008834 **
## NeighborhoodStoneBr  6.081e+04  1.076e+04   5.649 1.98e-08 ***
## NeighborhoodSWISU   -9.832e+03  1.252e+04  -0.785 0.432495
## NeighborhoodTimber   2.816e+04  9.991e+03   2.818 0.004899 **
## NeighborhoodVeenker  5.142e+04  1.326e+04   3.879 0.000110 ***
## KitchenQualFa       -4.564e+04  8.952e+03  -5.098 3.94e-07 ***
## KitchenQualGd       -2.901e+04  4.823e+03  -6.016 2.32e-09 ***
## KitchenQualTA       -3.870e+04  5.347e+03  -7.238 7.72e-13 ***
## OverallQual          1.201e+04  1.327e+03   9.052  < 2e-16 ***
## GarageFinishRFn     -6.353e+03  2.710e+03  -2.345 0.019199 *
## GarageFinishUnf     -9.457e+03  3.128e+03  -3.023 0.002553 **
## GrLivArea            4.894e+01  2.963e+00  16.520  < 2e-16 ***
## ExterQualFa         -3.302e+04  1.499e+04  -2.203 0.027789 *
## ExterQualGd         -2.224e+04  6.343e+03  -3.506 0.000470 ***
## ExterQualTA         -1.963e+04  7.081e+03  -2.773 0.005639 **
## BsmtQualFa          -3.596e+04  8.515e+03  -4.223 2.57e-05 ***
## BsmtQualGd          -3.287e+04  4.451e+03  -7.384 2.72e-13 ***
```

```
## BsmtQualTA          -3.194e+04  5.448e+03  -5.864 5.72e-09 ***
## TotalBsmtSF          1.746e+01  3.024e+00   5.776 9.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33600 on 1308 degrees of freedom
##   (111 observations deleted due to missingness)
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.8192
## F-statistic: 153.7 on 40 and 1308 DF,  p-value: < 2.2e-16
```

```
#table(train$Street)
```

Predicció de preus de cases

```
#newHousePrice <- data.frame()

newHouse <- data.frame( FullBath = 2, YearBuilt = 2003, Neighborhood = "CollgC
r",
                        KitchenQual = "Gd", OverallQual = 7, GarageFinish="RFn",
                        GrLivArea=1710, ExterQual ="Gd",BsmtQual="Gd", TotalBsmt
SF=856)
predict(lmSalePrice4, newHouse)
```

```
##      1
## 206370
```

Aquesta new house passada al model existeix en el sistema amb el valor:208500

```
head(subset(train, select = c(FullBath,YearBuilt,Neighborhood,KitchenQual
                          ,OverallQual,GarageFinish,GrLivArea,ExterQual,BsmtQua
l,TotalBsmtSF,SalePrice) ),1)
```

```
##   FullBath YearBuilt Neighborhood KitchenQual OverallQual GarageFinish
## 1        2      2003      CollgCr          Gd           7          RFn
##   GrLivArea ExterQual BsmtQual TotalBsmtSF SalePrice
## 1      1710        Gd       Gd         856    208500
```

Predicció del dataset de test usant el model de regressió lineal. Per Id es prediu el preu de venda. Això és també el repte demanat a Kaggle.

```
predictHousePrice<- predict(lmSalePrice4, test)
#output <- cbind(testdata, prediction)
output <- cbind(test, predictHousePrice)
#output
```

S'exporta a un fitxer CSV el Id del data source de test i el preu de venda predit pel model. Aquest fitxer és el demanat al challenge de Kaggle.

```
kaggleChallenge = subset(output, select = c(Id,predictHousePrice ) )

write.csv(kaggleChallenge,".\\PRA2DamEus_Submission.csv", row.names = FALSE)
```

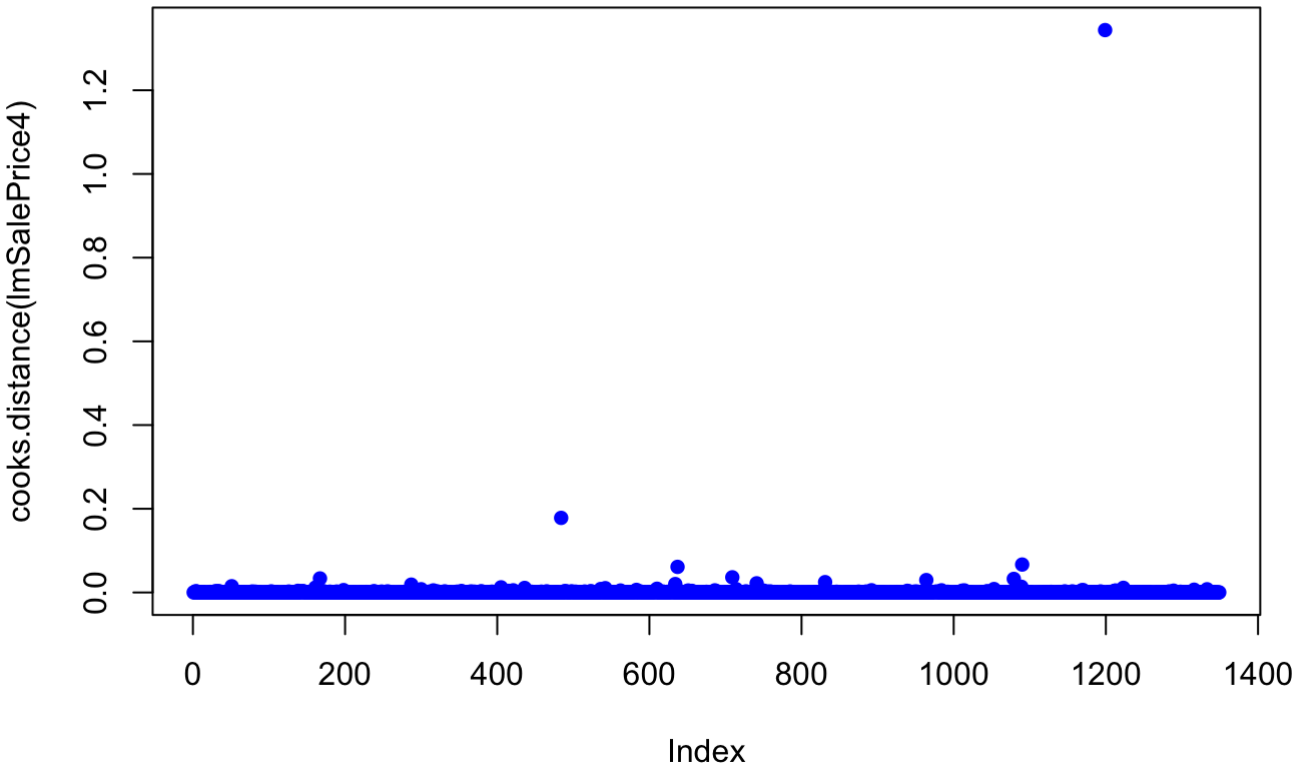
De les 1459 observacions del datasource de test, s'observa que 116 de les prediccions han estat na i per tant el model no ha estat capaç de predir.

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

```
##               Id       MSSubClass         MSZoning      LotFrontage
##                0                0                4              227
##          LotArea           Street            Alley         LotShape
##                0                0             1352                0
##      LandContour        Utilities        LotConfig        LandSlope
##                0                2                0                0
##     Neighborhood       Condition1       Condition2         BldgType
##                0                0                0                0
##       HouseStyle      OverallQual      OverallCond        YearBuilt
##                0                0                0                0
##     YearRemodAdd        RoofStyle          RoofMatl      Exterior1st
##                0                0                0                1
##      Exterior2nd       MasVnrType       MasVnrArea        ExterQual
##                1               16               15                0
##       ExterCond       Foundation         BsmtQual         BsmtCond
##                0                0               44               45
##     BsmtExposure     BsmtFinType1       BsmtFinSF1     BsmtFinType2
##               44               42                1               42
##       BsmtFinSF2       BsmtUnfSF     TotalBsmtSF          Heating
##                1                1                1                0
##        HeatingQC       CentralAir       Electrical         1stFlrSF
##                0                0                0                0
##         2ndFlrSF     LowQualFinSF        GrLivArea     BsmtFullBath
##                0                0                0                2
##     BsmtHalfBath         FullBath         HalfBath     BedroomAbvGr
##                2                0                0                0
##     KitchenAbvGr      KitchenQual     TotRmsAbvGrd       Functional
##                0                1                0                2
##       Fireplaces      FireplaceQu       GarageType      GarageYrBlt
##                0              730               76               78
##     GarageFinish       GarageCars       GarageArea       GarageQual
##               78                1                1               78
##       GarageCond       PavedDrive       WoodDeckSF      OpenPorchSF
##               78                0                0                0
##    EnclosedPorch        3SsnPorch      ScreenPorch         PoolArea
##                0                0                0                0
##           PoolQC            Fence      MiscFeature          MiscVal
##             1456             1169             1408                0
##           MoSold           YrSold         SaleType    SaleCondition
##                0                0                1                0
##       vnSaleType       vnMSZoning predictHousePrice
##                1                4              116
```

En la Distància de cook del model, s'observa un punt molt allunyat i altres amb una certa distància.

```
plot(cooks.distance(lmSalePrice4), pch = 16, col = "blue")
```



6 Conclusions

Tras analitzar les característiques d'aquest data set sobre les cases de Iowa, s'observa que els valors NA són propis del domini, o almenys així es registra en la definició del data set. S'aprecien valors outliers que impacten la potència estadística, però es decideix no treurel's donat que semblen valors correctes analitzat el contexte d'altres variables que indiquen que el preu podria estar justificat, per la qualitat dels materials, any de construcció, metres del garatge o altres característiques que fan pensar que el valor és correcte. En aquest sentit hem tingut un dilema en quant a l'eliminació o tractament dels outliers per a obtenir una major potència estadística vs mantenir els valors al comprovar que tot indica que són correctes. Finalment s'ha decidit la segona opció.

Es fa una reducció de la dimensionalitat: eliminant aquells atributs que tenen el valor NA en la majoria dels seus registres i aquells atributs on no s'observa cap correlació amb la variable SalePrice. Per a analitzar la correlació dels atributs categòrics es creen variables numèriques ordinals a partir de les variables qualitatives. En els casos on no es pot establir un ordre es fa una mitja del preu per cadascú de les categories de la variable i s'estableix així l'ordre. Donat que hi ha una manca de normalitat en les variables, s'aplica el test de correlació no paramètric de Spearman.

Els atributs amb una més alta correlació són els incorporats al model de regressió lineal múltiple, que és el que s'usa per predir el preu en el challenge de Kaggle. El model obtingut té un coeficient de determinació d'un 82%. S'observa que en 16 casos dels 1451 observacions del dataset de test no és capaç de predir i el valor és NA. En general amb les proves fetes s'ha observat que el preu del model s'aproxima bastant al preu real.

En quant a la comparació de grups, tant en la comparació del preu per barris com comparant els preus dels anys 80 i 90, s'observen diferencies estadístiques en el preu de venda, més clares en la comparació per barris i no tan evidents en la comparació de preus entre els 80 i els 90 on hi ha un lleuger augment de la mitjana dels preus i una forta presència de preus molt elevats en els anys 90 (outliers).

Finalment, pel que fa a l'evolució dels preus al llarg del temps, amb l'anàlisi realitzat sobre les variables temporals del dataset hem pogut observar resultats interessants com, per exemple, el fet de que les cases reformades en els últims 2-3 anys experimenten un augment molt significatiu del seu preu de venda. O també, que amb la crisi econòmica del 2008 els preus de venda van baixar dràsticament per a tots els casos.

7 Taula de contribucions

Contribucions	**Firma**
Investigació prèvia	Eusebio Garcia i Damián Martínez
Redacció de les respostes	Eusebio Garcia i Damián Martínez
Desenvolupament codi	Eusebio Garcia i Damián Martínez

8 WEBGRAPHY

HousePrices - kaggle https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview)

HousePrices - kaggle - Tutorials https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview/tutorials (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview/tutorials)

HousePrices - Kaggle - Start Here - Visualization and Model Stacking https://www.kaggle.com/rp1611/start-here-visualization-and-model-stacking (https://www.kaggle.com/rp1611/start-here-visualization-and-model-stacking)

Trabajo de Fin de Grado - ESTUDIO DE TÉCNICAS SUPERVISADAS DE REDUCCIÓN DE DIMENSIONALIDAD PARA PROBLEMAS DE CLASIFICACIÓN - Álvaro Soriano Maganto 2016-2017 https://e-archivo.uc3m.es/bitstream/handle/10016/26504/TFG_Alvaro_Soriano_Maganto.pdf (https://e-archivo.uc3m.es/bitstream/handle/10016/26504/TFG_Alvaro_Soriano_Maganto.pdf)

The most common dimension techniques - RPubs https://rpubs.com/Saskia/520216 (https://rpubs.com/Saskia/520216)

Análisis de Regresión - Alfonso Novales - Departamente de Economía Cuantitativa 2010 https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf (https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf)

StackOverflow - R - Perform a levene test with two samples with different sizes https://stackoverflow.com/questions/43749166/r-perform-a-levene-test-with-two-samples-with-different-sizes (https://stackoverflow.com/questions/43749166/r-perform-a-levene-test-with-two-samples-with-

different-sizes)