Evaluación asigunaturas Visualización de datos y Visualización de la información

Gráficas con ggplot

Inmaculada Perea Fern?ndez

Agosto 2017

El conjunto de datos elegido para este trabajo es uno de los datasets de aprendizaje de kaggle, en el que el objetivo es predecir el precio de venta de viviendas (variable objetivo *SalePrice*) a partir de sus características.

Para más información ir a:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

1. Carga de librerías y opciones por defecto

```
if (!require('Rcpp')) install.packages('Rcpp'); library('Rcpp')
if (!require('naniar')) install.packages('naniar'); library('naniar')
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('scales')) install.packages('scales'); library('scales')
if (!require('forcats')) install.packages('forcats'); library('forcats')
if (!require('GGally')) install.packages('GGally'); library('GGally')
if (!require('mi')) install.packages('mi'); library('mi')
if (!require('extracat')) install.packages('extracat'); library('extracat')
if (!require('data.table')) install.packages('data.table'); library('data.table')
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('maps')) install.packages('maps'); library('maps')
if (!require('ggalt')) install.packages('ggalt'); library('ggalt')
if (!require('ggExtra')) install.packages('ggExtra'); library('ggExtra')
```

```
# deshabilita la notación científica
options(scipen=999)

# Establece tema por defecto
theme_set(theme_bw())
```

2. Carga y resumen de los datos

```
data <- read.csv("data/train.csv", header=T, dec=".", sep=",")
dim(data)
```

```
## [1] 1460   81
```

```
summary(data)
```

```
##        Id           MSSubClass       MSZoning     LotFrontage
##  Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   RH     :  16   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                   Max.   :313.00
##                                                   NA's   :259
##     LotArea        Street      Alley      LotShape  LandContour
##  Min.   :  1300   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63
##  1st Qu.:  7554   Pave:1454   Pave:  41   IR2: 41   HLS:  50
##  Median :  9478               NA's:1369   IR3: 10   Low:  36
##  Mean   : 10517                           Reg:925   Lvl:1311
##  3rd Qu.: 11602
##  Max.   :215245
##
##   Utilities      LotConfig    LandSlope   Neighborhood   Condition1
##  AllPub:1459   Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260
##  NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81
##                FR2    :  47   Sev:  13   OldTown:113   Artery :  48
##                FR3    :   4              Edwards:100   RRAn   :  26
##                Inside :1052              Somerst: 86   PosN   :  19
##                                          Gilbert: 79   RRAe   :  11
##                                          (Other):707   (Other):  15
##   Condition2     BldgType      HouseStyle   OverallQual
##  Norm   :1445   1Fam  :1220   1Story :726   Min.   : 1.000
##  Feedr  :   6   2fmCon:  31   2Story :445   1st Qu.: 5.000
##  Artery :   2   Duplex:  52   1.5Fin :154   Median : 6.000
##  PosN   :   2   Twnhs :  43   SLvl   : 65   Mean   : 6.099
##  RRNn   :   2   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000
##  PosA   :   1                 1.5Unf : 14   Max.   :10.000
##  (Other):   2                 (Other): 19
##   OverallCond      YearBuilt     YearRemodAdd    RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141
##  Median :5.000   Median :1973   Median :1994   Gambrel:  11
##  Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2
##
##     RoofMatl     Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
##  CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.0
##  Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.0
##  WdShngl:   6   MetalSd:220   HdBoard:207   None   :864   Median :   0.0
##  WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 103.7
##  ClyTile:   1   Plywood:108   Plywood:142   NA's   :  8   3rd Qu.: 166.0
##  Membran:   1   CemntBd: 61   CmentBd: 60                 Max.   :1600.0
##  (Other):   2   (Other):128   (Other):136                 NA's   :8
##  ExterQual ExterCond  Foundation  BsmtQual   BsmtCond    BsmtExposure
##  Ex: 52    Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221
##  Fa: 14    Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134
##  Gd:488    Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114
##  TA:906    Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953
##            TA:1282   Stone :  6   NA's: 37   NA's:  37   NA's: 38
##                      Wood  :  3
##
##  BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
##  ALQ :220     Min.   :   0.0  ALQ :  19    Min.   :   0.00
##  BLQ :148     1st Qu.:   0.0  BLQ :  33    1st Qu.:   0.00
##  GLQ :418     Median : 383.5  GLQ :  14    Median :   0.00
##  LwQ : 74     Mean   : 443.6  LwQ :  46    Mean   :  46.55
##  Rec :133     3rd Qu.: 712.2  Rec :  54    3rd Qu.:   0.00
##  Unf :430     Max.   :5644.0  Unf :1256    Max.   :1474.00
##  NA's: 37                     NA's:  38
##    BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
##  Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
##  1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
##  Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  FuseP:   3   Median :1087   Median :   0   Median :  0.000
##  Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
##  SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  NA's :   1   Max.   :4692   Max.   :2065   Max.   :572.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath       FullBath
##  Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :1464   Median :0.0000   Median :0.00000   Median :2.000
##  Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
##  3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000
##
##     HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000
##
##   TotRmsAbvGrd    Functional    Fireplaces    FireplaceQu   GarageType
##  Min.   : 2.000   Maj1:  14   Min.   :0.000   Ex  : 24    2Types :  6
##  1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   Fa  : 33    Attchd :870
##  Median : 6.000   Min1:  31   Median :1.000   Gd  :380    Basment: 19
##  Mean   : 6.518   Min2:  34   Mean   :0.613   Po  : 20    BuiltIn: 88
##  3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   TA  :313    CarPort:  9
##  Max.   :14.000   Sev :   1   Max.   :3.000   NA's:690    Detchd :387
##                   Typ :1360                               NA's   : 81
##   GarageYrBlt   GarageFinish  GarageCars      GarageArea     GarageQual
##  Min.   :1900   Fin :352     Min.   :0.000   Min.   :   0.0   Ex  :   3
##  1st Qu.:1961   RFn :422     1st Qu.:1.000   1st Qu.: 334.5   Fa  :  48
##  Median :1980   Unf :605     Median :2.000   Median : 480.0   Gd  :  14
##  Mean   :1979   NA's: 81     Mean   :1.767   Mean   : 473.0   Po  :   3
##  3rd Qu.:2002                3rd Qu.:2.000   3rd Qu.: 576.0   TA  :1311
##  Max.   :2010                Max.   :4.000   Max.   :1418.0   NA's:  81
##  NA's   :81
##  GarageCond  PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
##  Ex  :   2   N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  Fa  :  35   P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Gd  :   9   Y:1340     Median :  0.00   Median : 25.00   Median :  0.00
##  Po  :   7              Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
##  TA  :1326              3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
##  NA's:  81              Max.   :857.00   Max.   :547.00   Max.   :552.00
##
##    X3SsnPorch      ScreenPorch        PoolArea        PoolQC
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Ex  :   2
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Fa  :   2
##  Median :  0.00   Median :  0.00   Median :  0.000   Gd  :   3
##  Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   NA's:1453
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##    Fence      MiscFeature    MiscVal             MoSold
##  GdPrv:  59   Gar2:   2   Min.   :    0.00   Min.   : 1.000
##  GdWo :  54   Othr:   2   1st Qu.:    0.00   1st Qu.: 5.000
##  MnPrv: 157   Shed:  49   Median :    0.00   Median : 6.000
##  MnWw :  11   TenC:   1   Mean   :   43.49   Mean   : 6.322
##  NA's :1179   NA's:1406   3rd Qu.:    0.00   3rd Qu.: 8.000
##                           Max.   :15500.00   Max.   :12.000
##
##      YrSold        SaleType    SaleCondition    SalePrice
##  Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
##  1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
##  Median :2008   COD    :  43   Alloca :  12   Median :163000
##  Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
##  3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
##  Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
##                 (Other):   9
```

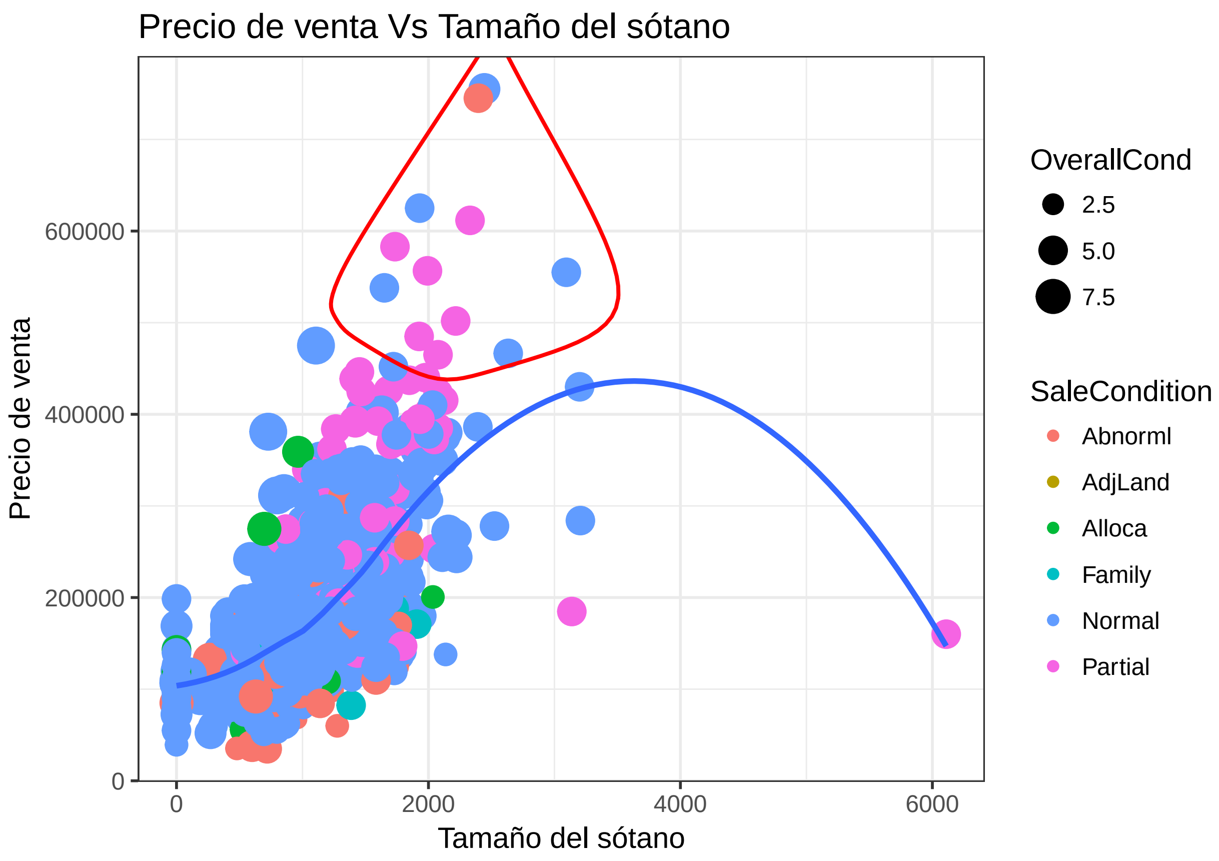
```r
str(data)
```

```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

3. Análisis Gráfico de datos

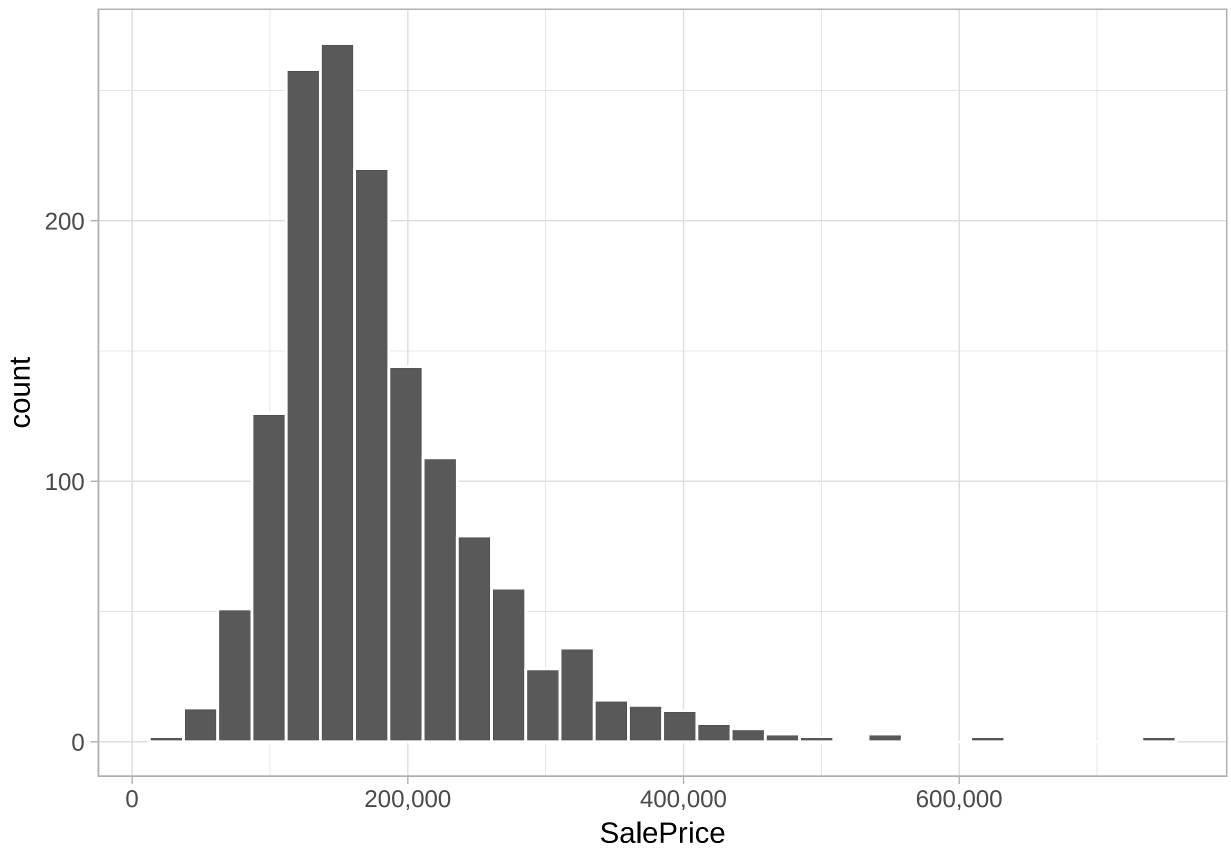
```
luxury <- data %>% filter(SalePrice > 500000)

ggplot(data, aes( x=TotalBsmtSF, y=SalePrice)) +
  geom_point(aes(col=SaleCondition, size=OverallCond)) +
  geom_smooth(method="loess", se=F) +
  geom_encircle(aes(x=TotalBsmtSF, y=SalePrice),
                data=luxury,
                color="red",
                size=2,
                expand=0.08) +
  labs(y="Precio de venta",
       x="Tamaño del sótano",
       title="Precio de venta Vs Tamaño del sótano")
```

```
## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0xf

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0xd

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0xf

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0xd
```

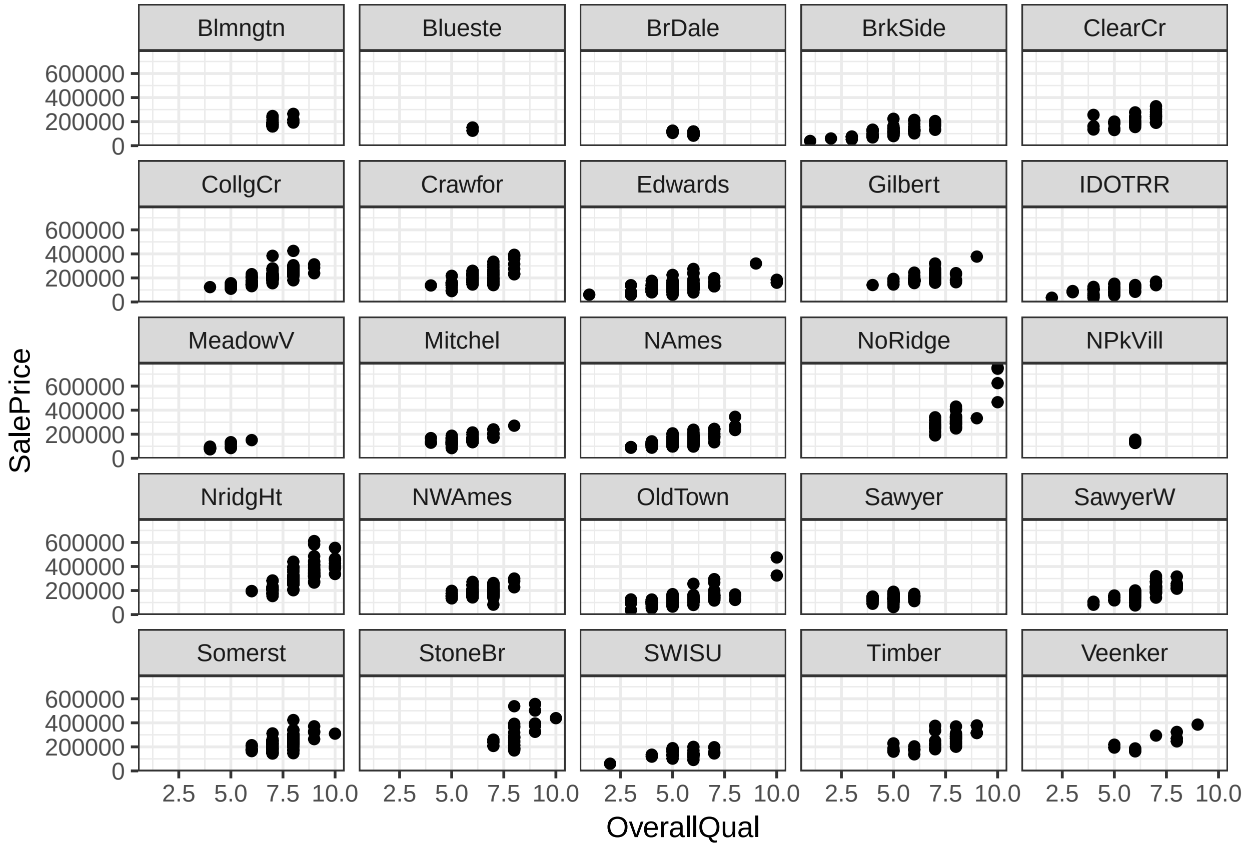


```r
ggplot(data, aes(x=SalePrice)) +
  geom_histogram(col = 'white') +
  theme_light() +
  scale_x_continuous(labels = comma)
```

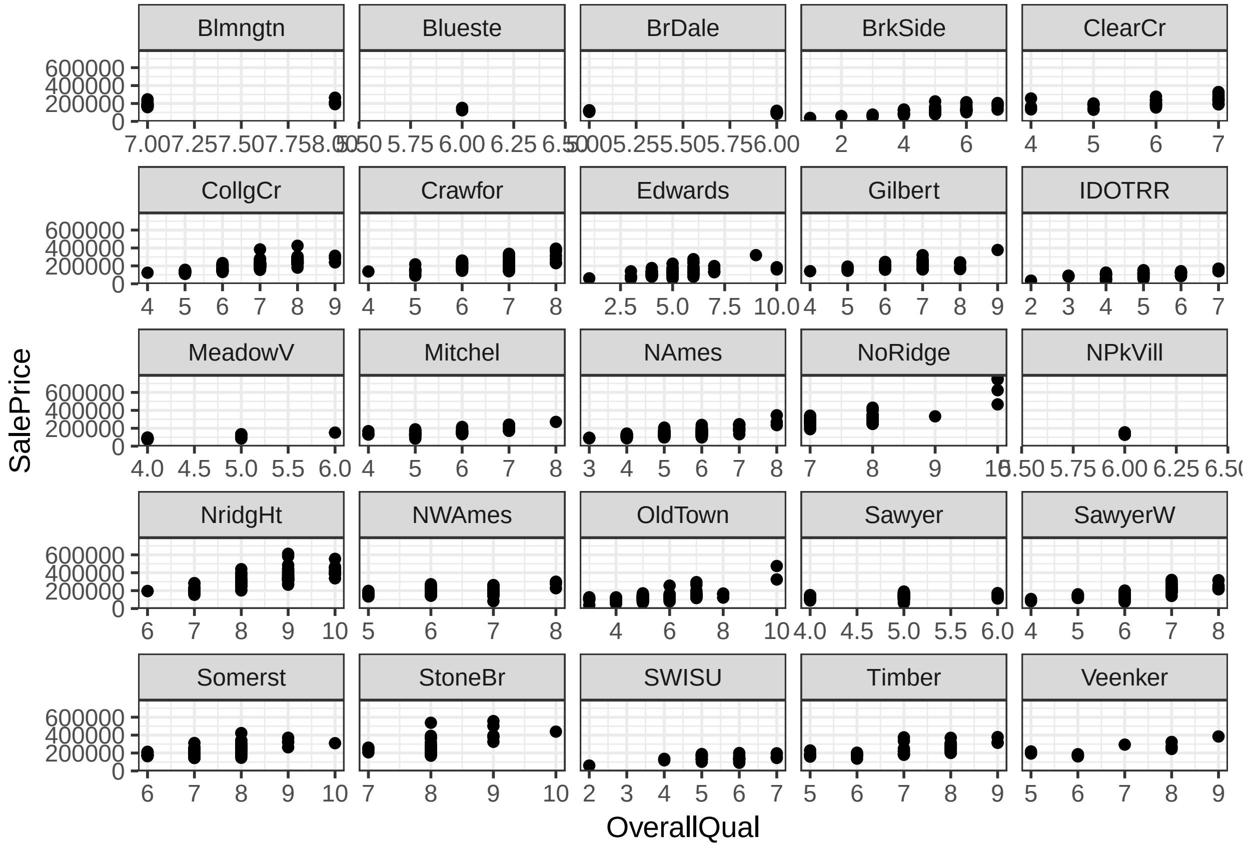
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



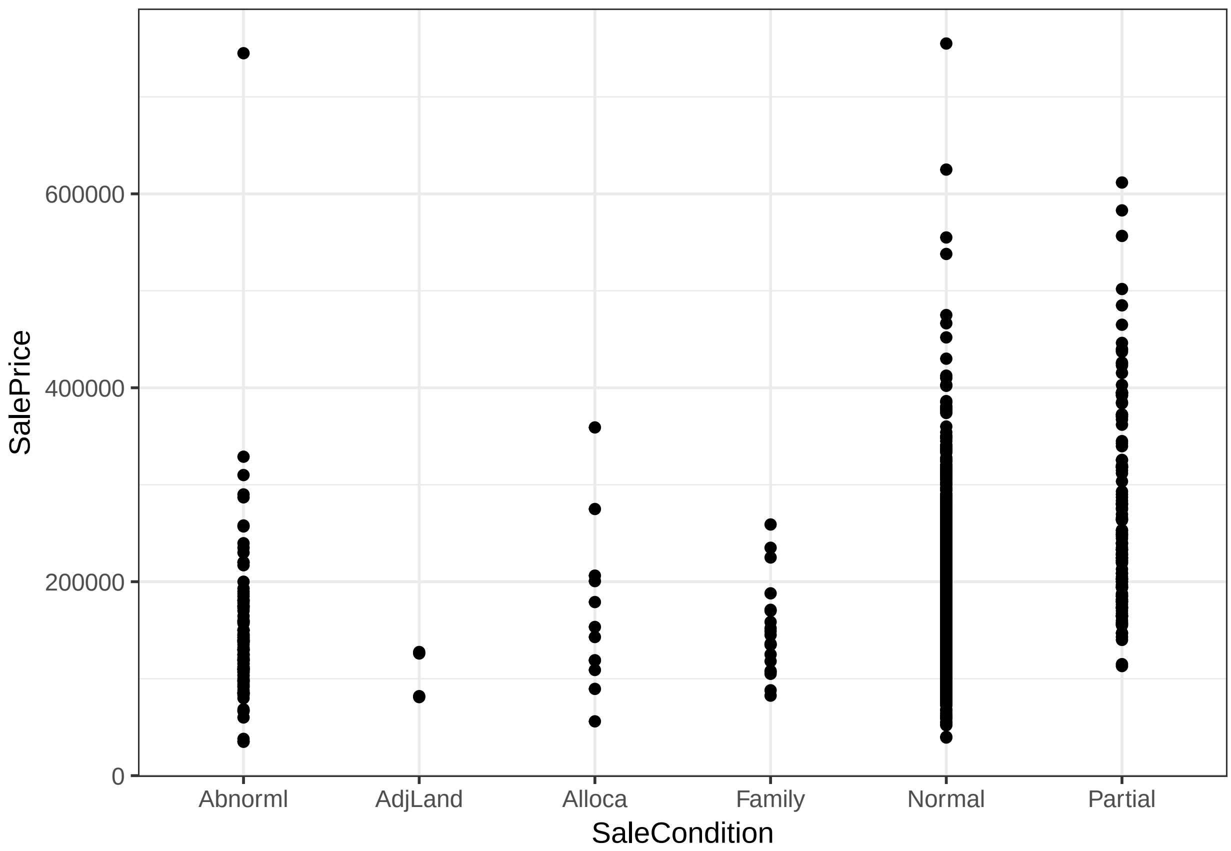
```
ggplot(data, aes(x=OverallQual, y=SalePrice)) +
geom_point() +
facet_wrap('Neighborhood')
```



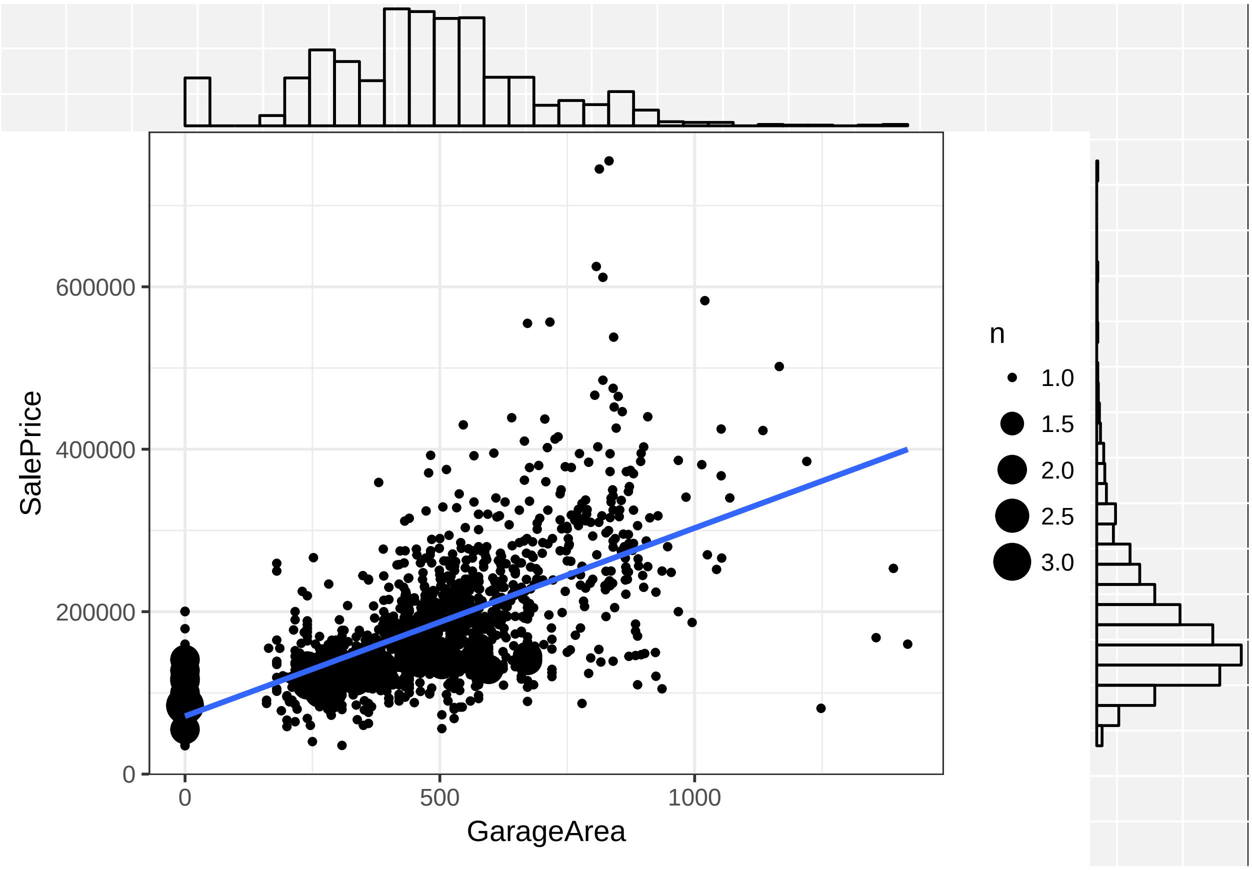
```r
ggplot(data, aes(x=OverallQual, y=SalePrice)) +
geom_point() +
facet_wrap('Neighborhood', scales='free_x')
```



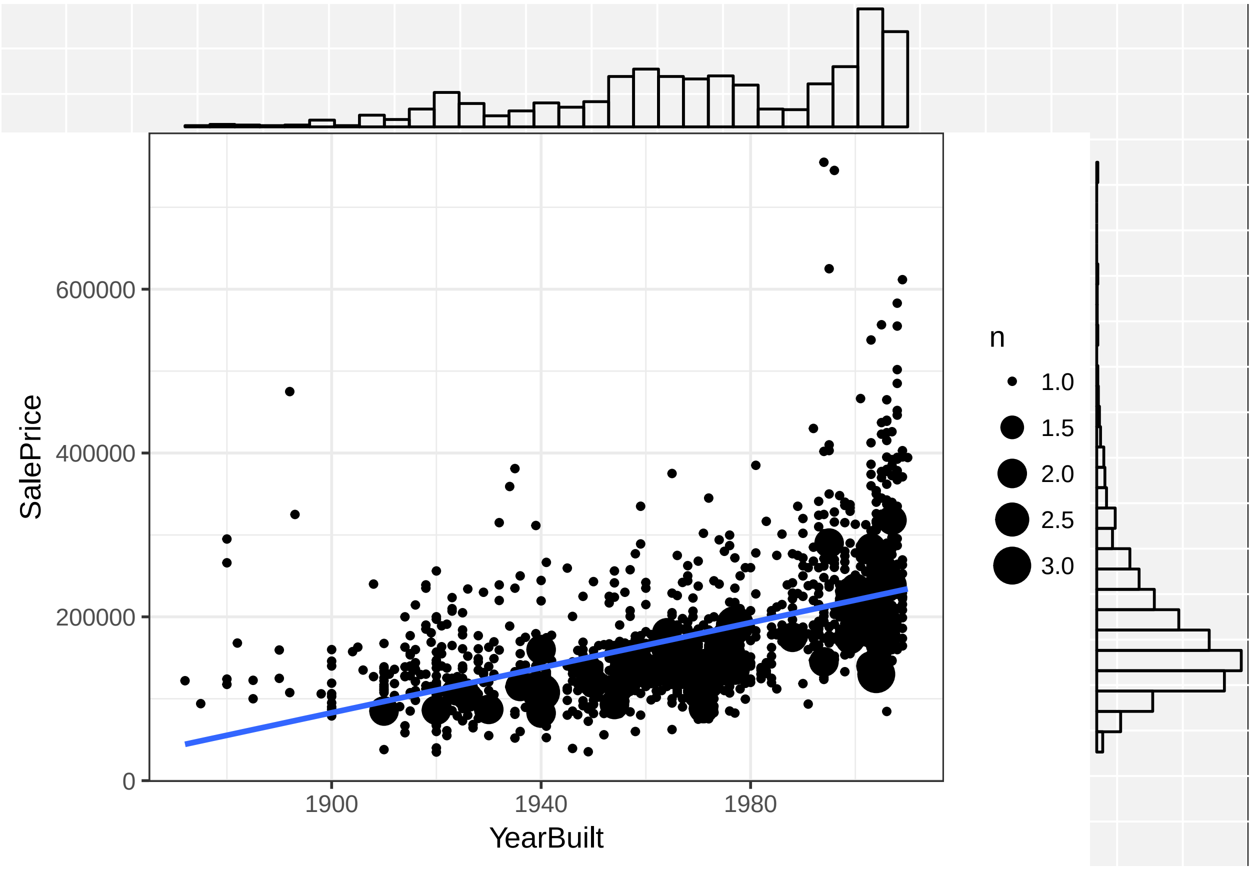
```
ggplot(data, aes(x=SaleCondition, y=SalePrice)) +
geom_point()
```



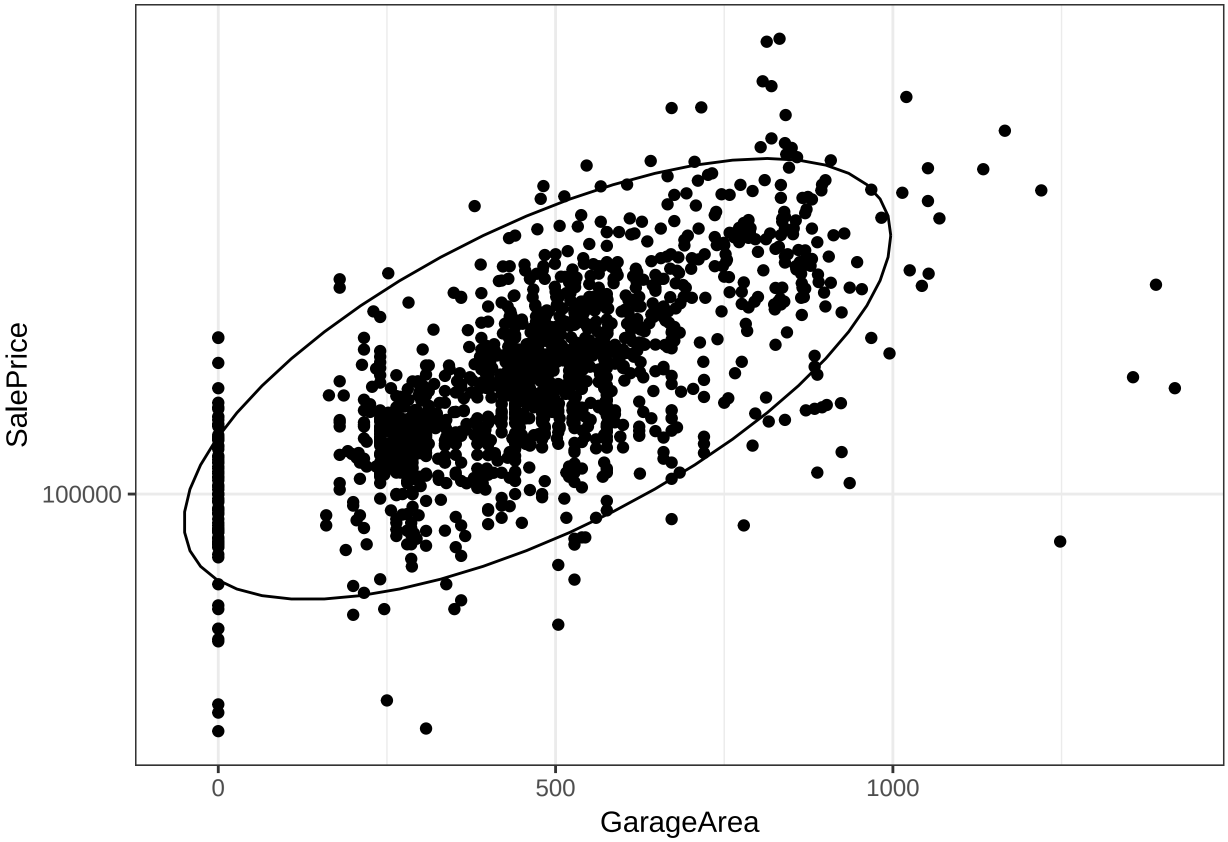
```
g <- ggplot(data, aes(GarageArea, SalePrice)) +
  geom_count() +
  geom_smooth(method="lm", se=F)

g2<-ggMarginal(g, type = "histogram", fill="transparent")
plot(g2)
```



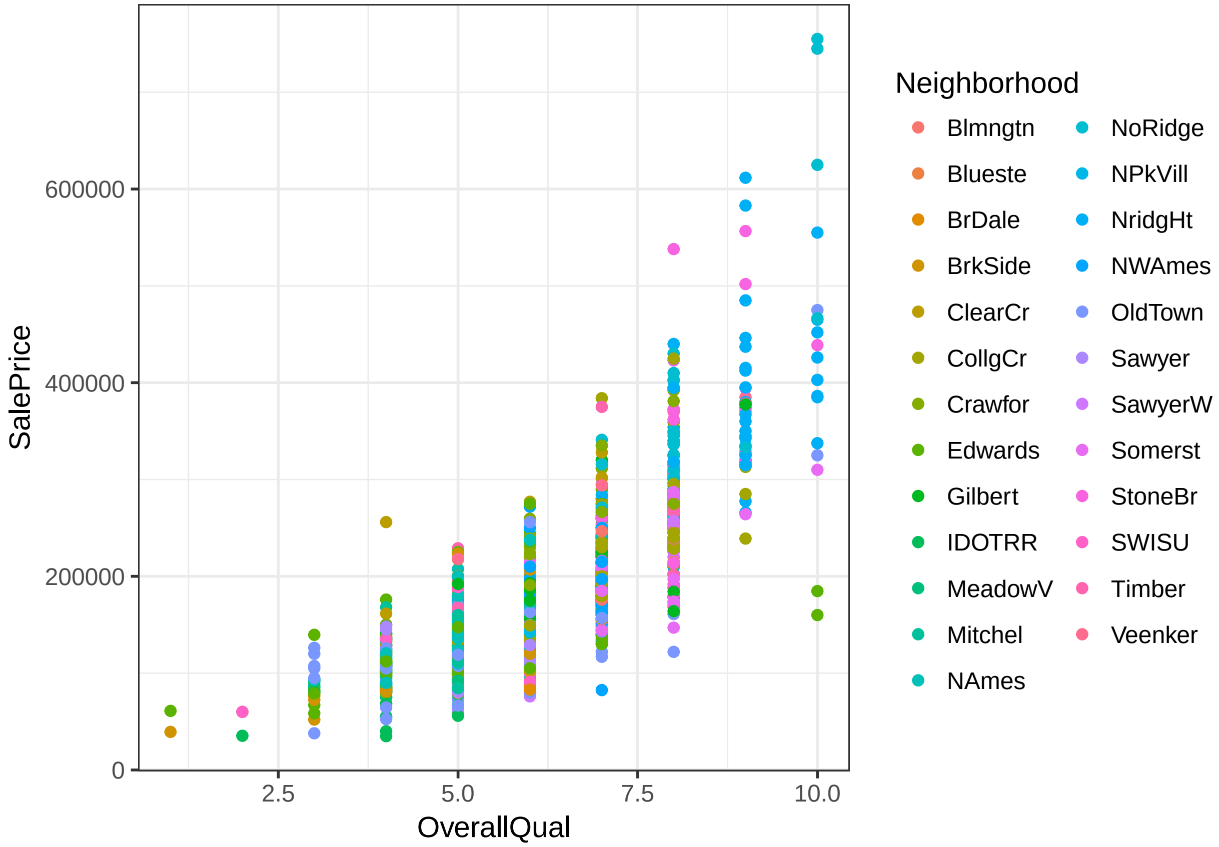
```r
g <- ggplot(data, aes(YearBuilt, SalePrice)) +
  geom_count() +
  geom_smooth(method="lm", se=F)

g2<-ggMarginal(g, type = "histogram", fill="transparent")
plot(g2)
```



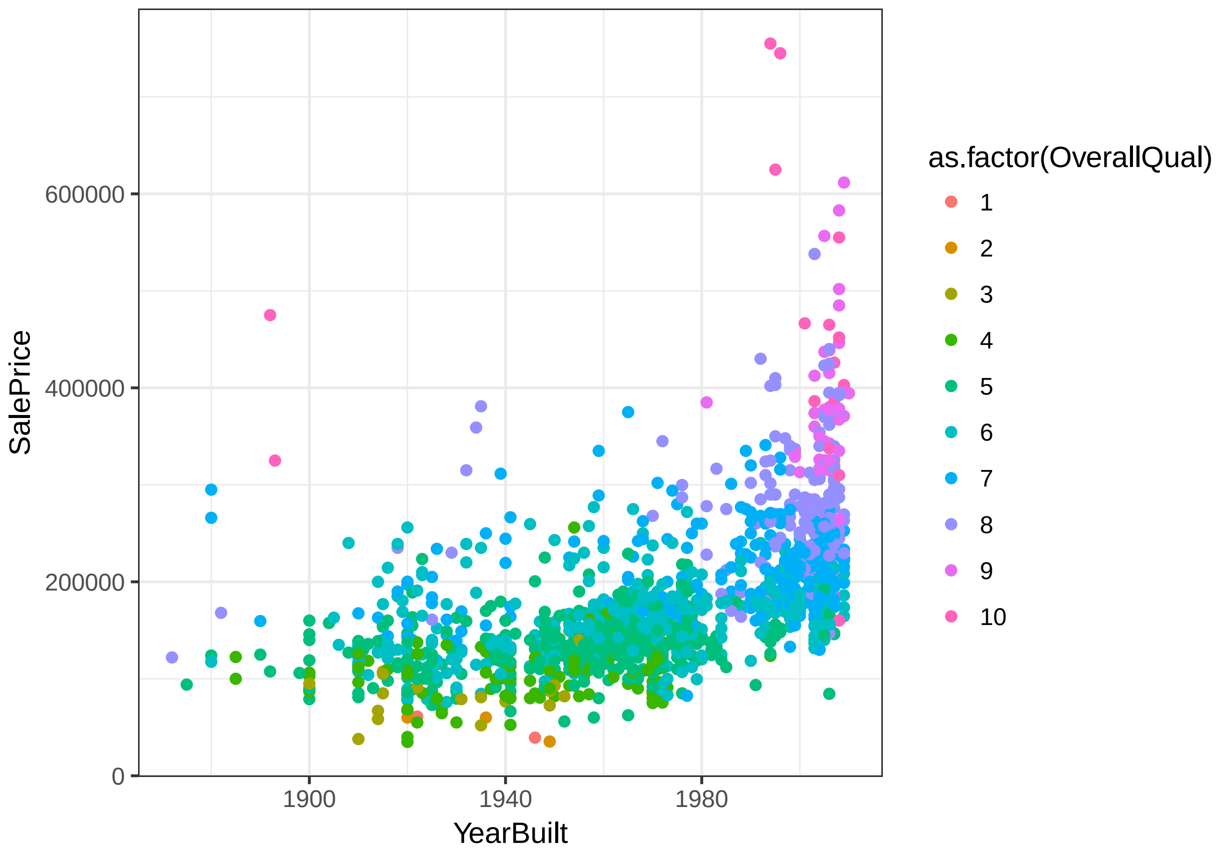
```r
ggplot(data, aes(x=GarageArea, y=SalePrice)) +
geom_point() +
scale_y_log10() +
stat_ellipse(type='norm')
```



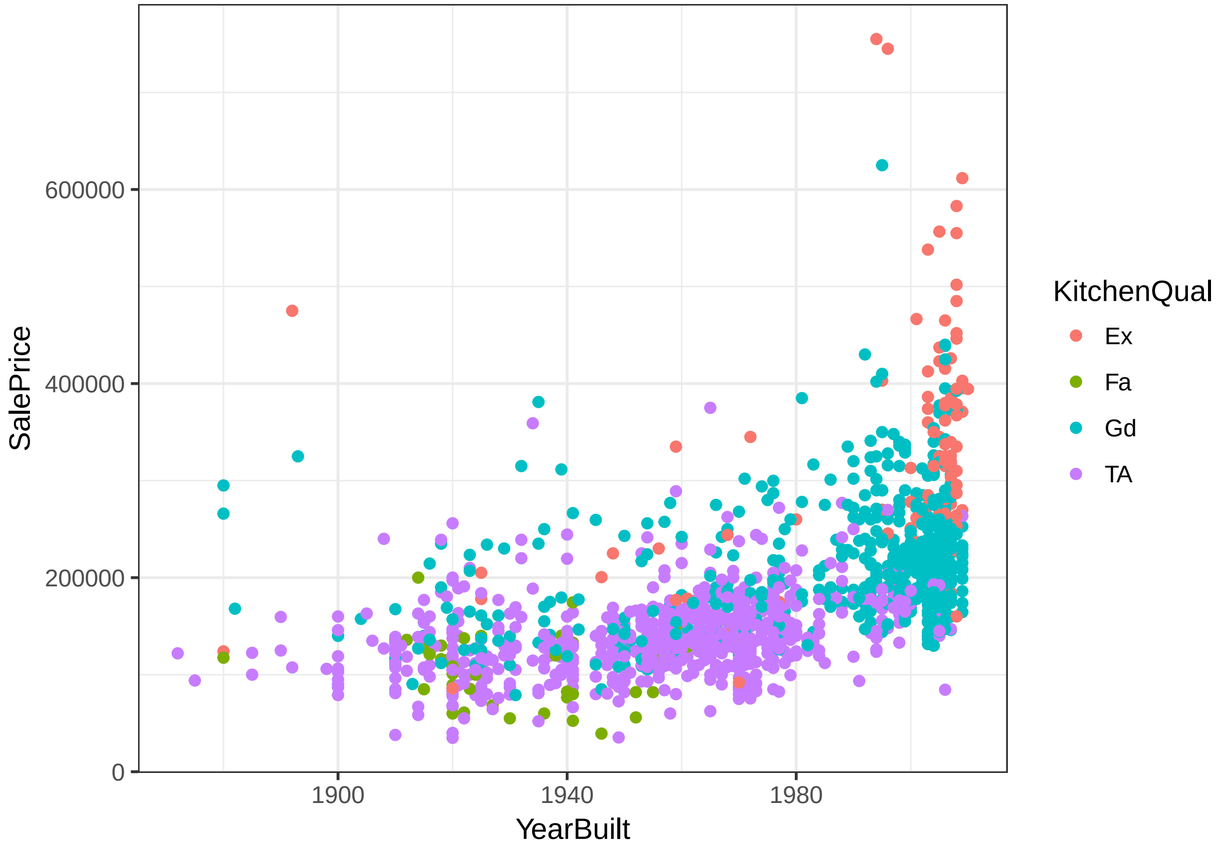
```
ggplot(data, aes(x=OverallQual, y=SalePrice, colour=Neighborhood)) +
  geom_point()
```



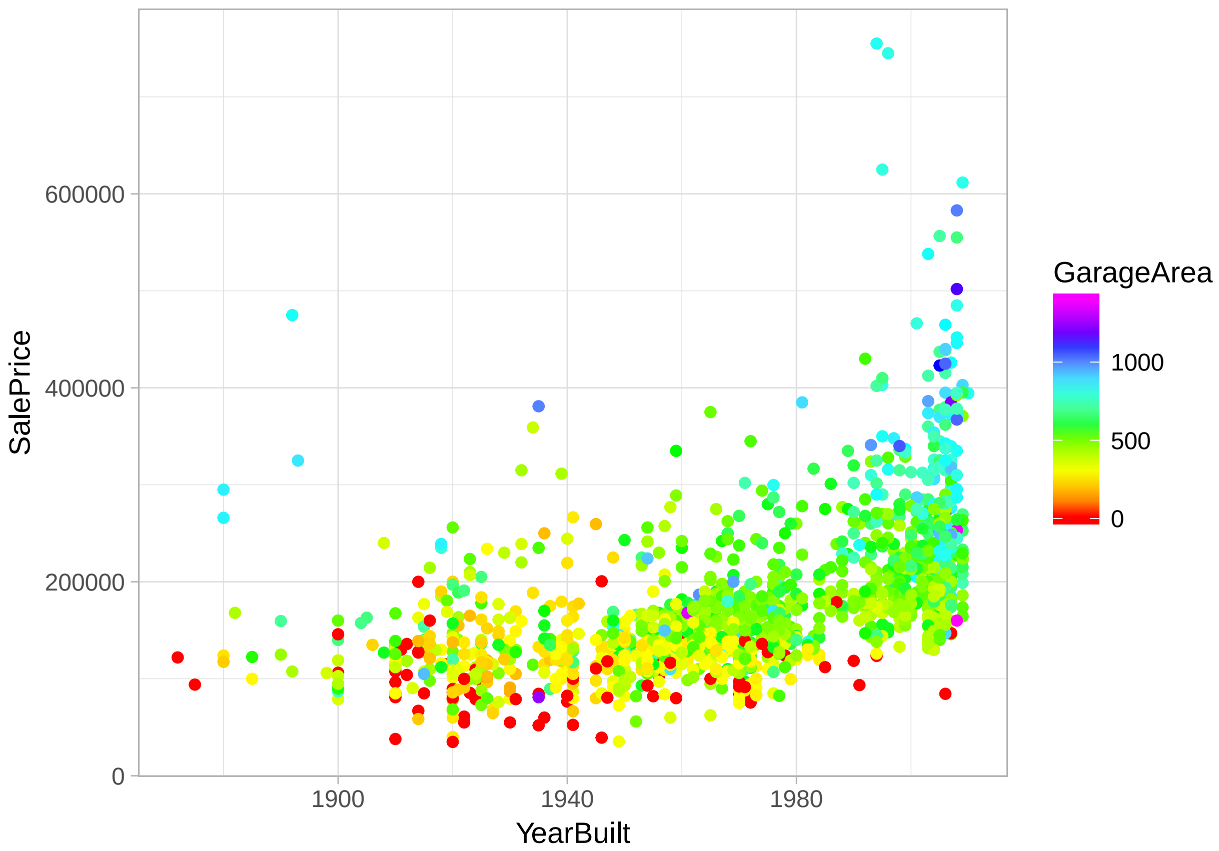
```
ggplot(data, aes(x=YearBuilt, y=SalePrice, colour=as.factor(OverallQual))) +
  geom_point()
```



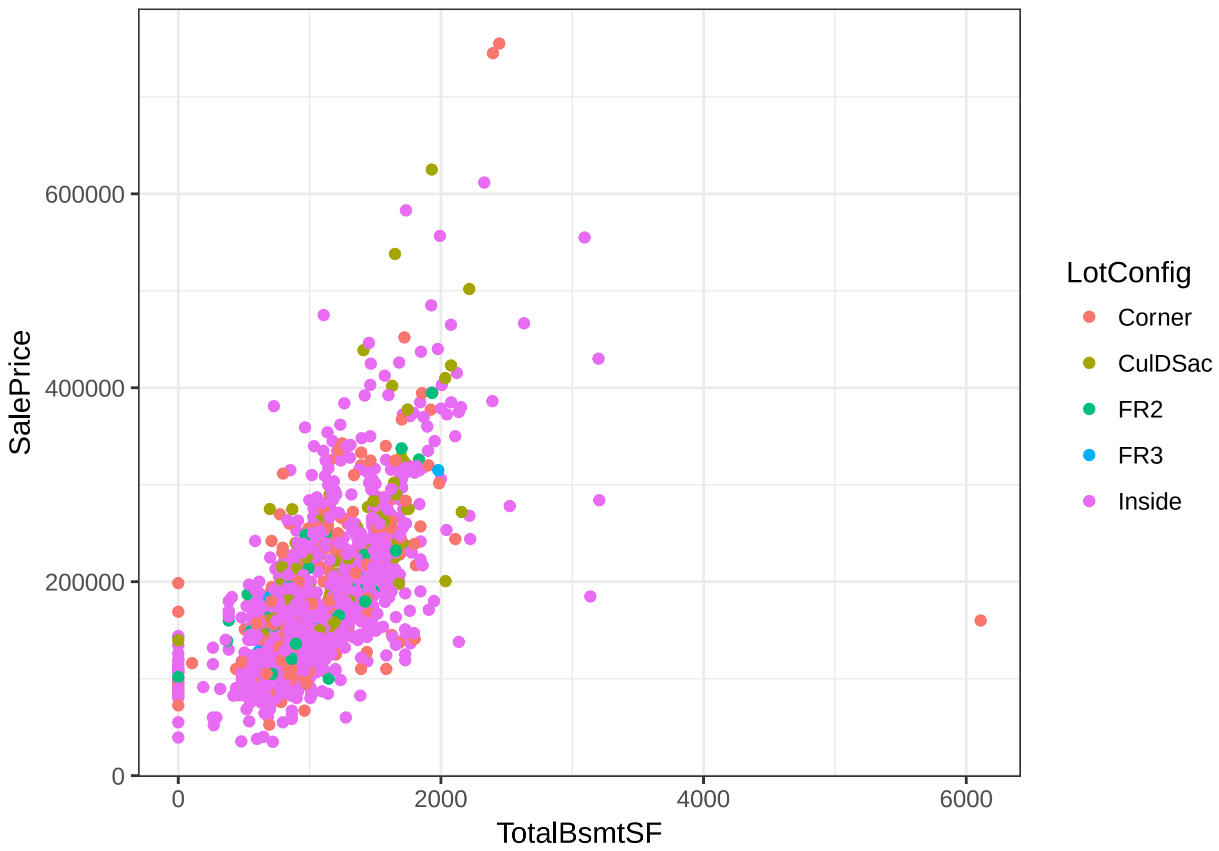
```
ggplot(data, aes(x=YearBuilt, y=SalePrice, colour=KitchenQual)) + geom_point()
```



```r
ggplot(data, aes(x=YearBuilt, y=SalePrice, colour=GarageArea)) +
  geom_point() +
  theme_light() +
  scale_colour_gradientn(colours=rainbow(6))
```

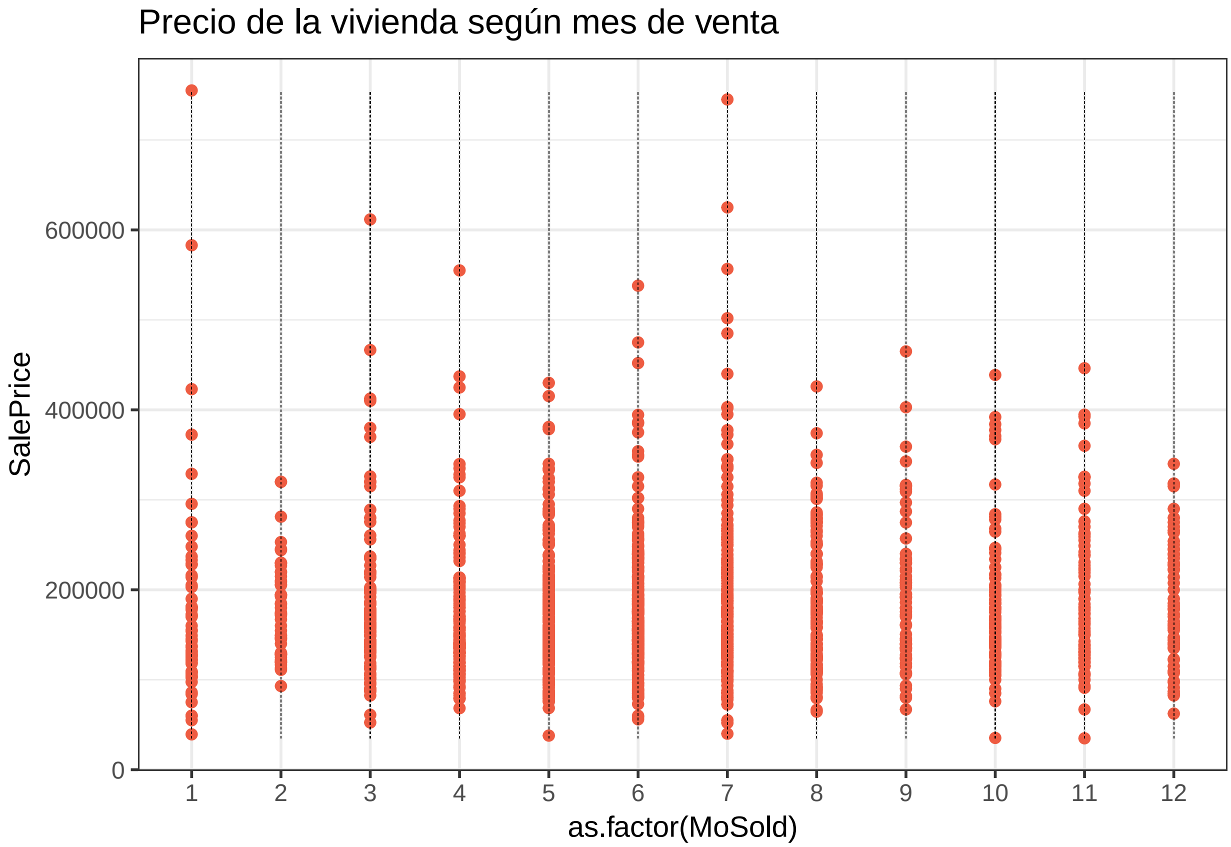


```
ggplot(data, aes(x=TotalBsmtSF, y=SalePrice, color=LotConfig)) +
  geom_point()
```



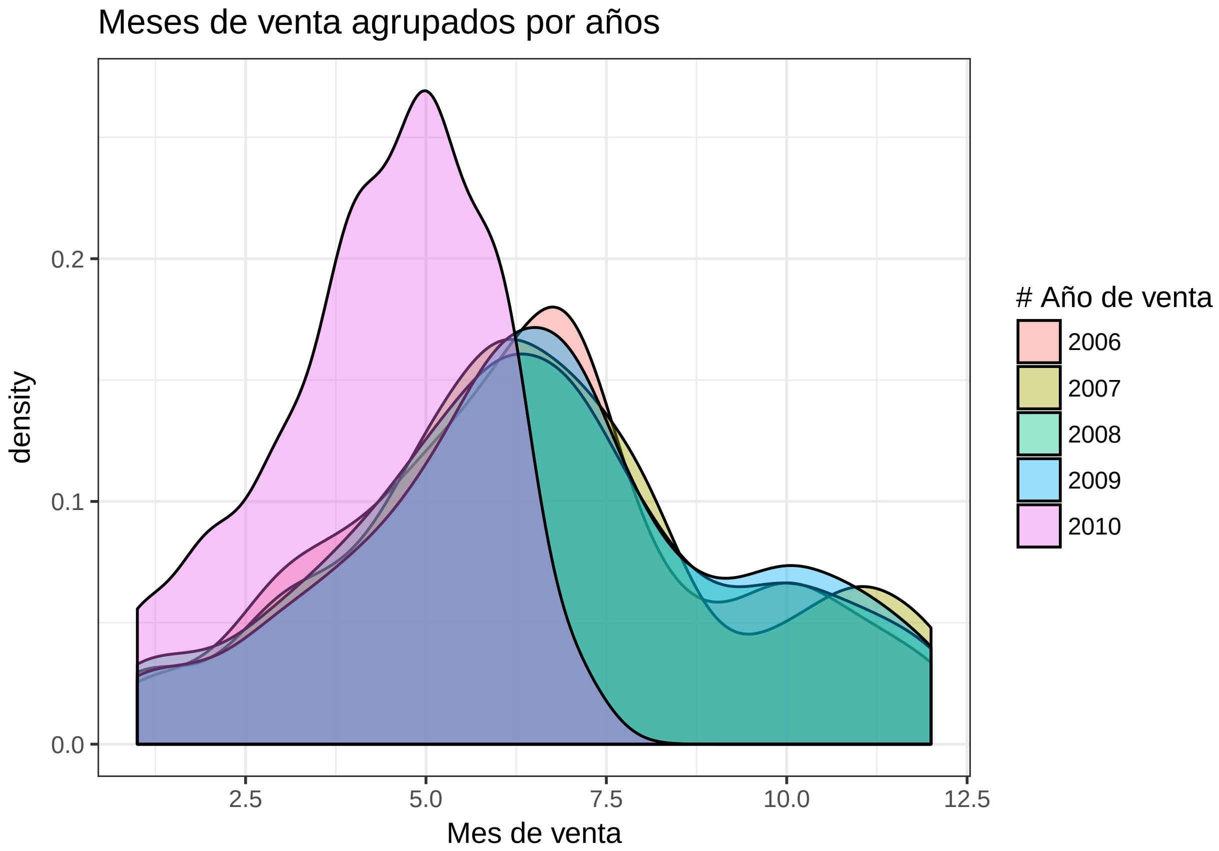
```
ggplot(data, aes(x=as.factor(MoSold), y=SalePrice)) +
  geom_point(col="tomato2", size=1.5) +
  geom_segment(aes(x=MoSold,
                   xend=MoSold,
                   y=min(SalePrice),
                   yend=max(SalePrice)),
               linetype="dashed",
               size=0.1) +
  labs(title="Precio de la vivienda según mes de venta")
```

```
## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0x6
```

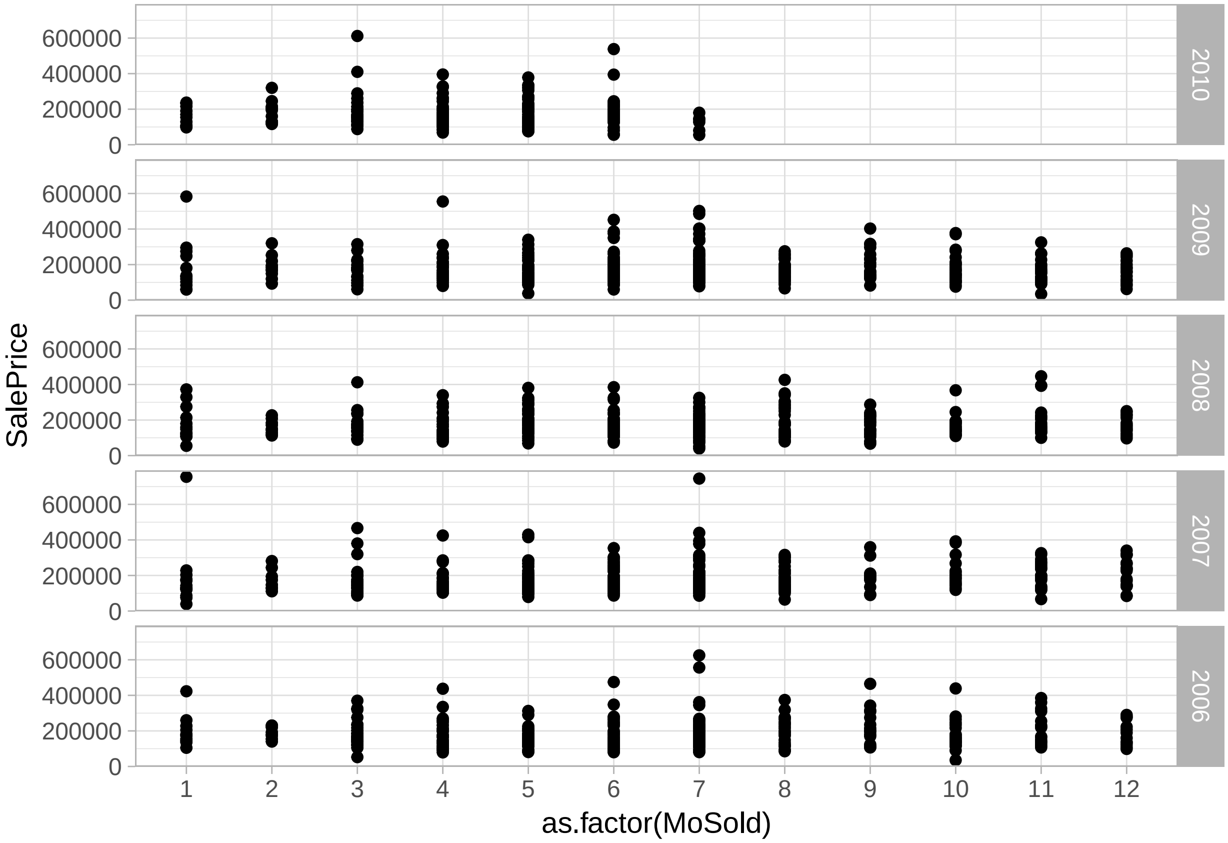


```
g <- ggplot(data, aes(MoSold))
g + geom_density(aes(fill=factor(YrSold)), alpha=0.4) +
    labs(title="Meses de venta agrupados por años",
         x="Mes de venta",
         fill="# Año de venta")
```

```
## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0xf

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0xf
```

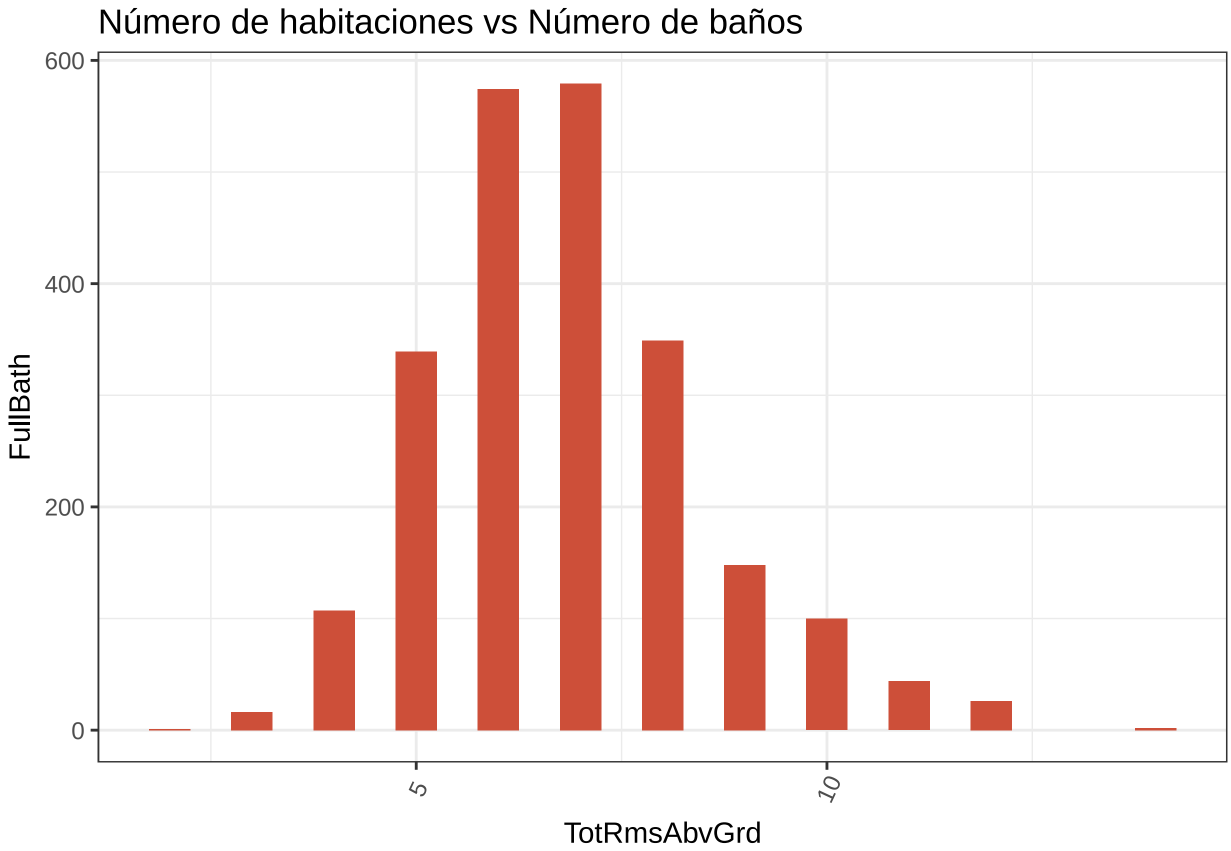


```
ggplot(data, aes(SalePrice, as.factor(MoSold))) +
geom_point() + coord_flip() +
facet_grid(YrSold ~ ., as.table=FALSE) +
theme_light() + scale_colour_gradientn(colours=rainbow(6))
```

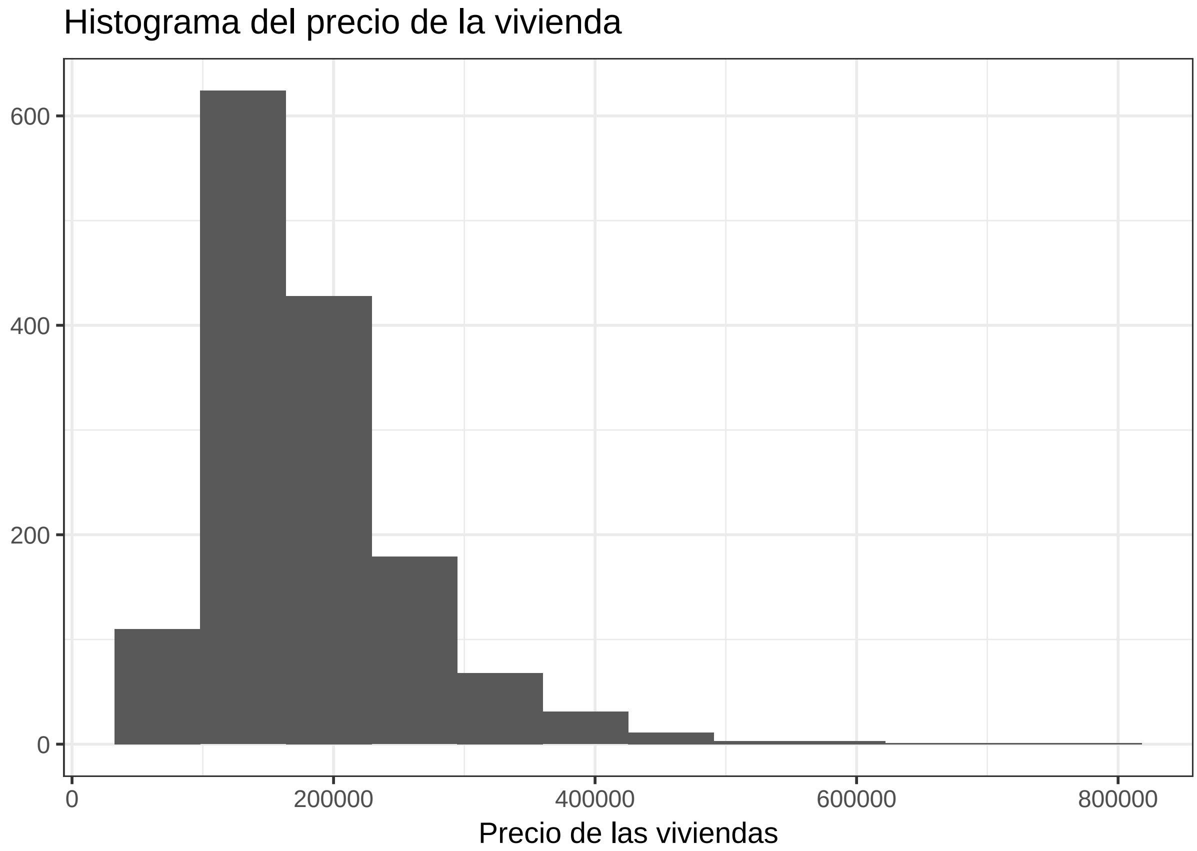


```
ggplot(data, aes(x=TotRmsAbvGrd, y=FullBath)) +
  geom_bar(stat="identity", width=.5, fill="tomato3") +
  labs(title="Número de habitaciones vs Número de baños") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```

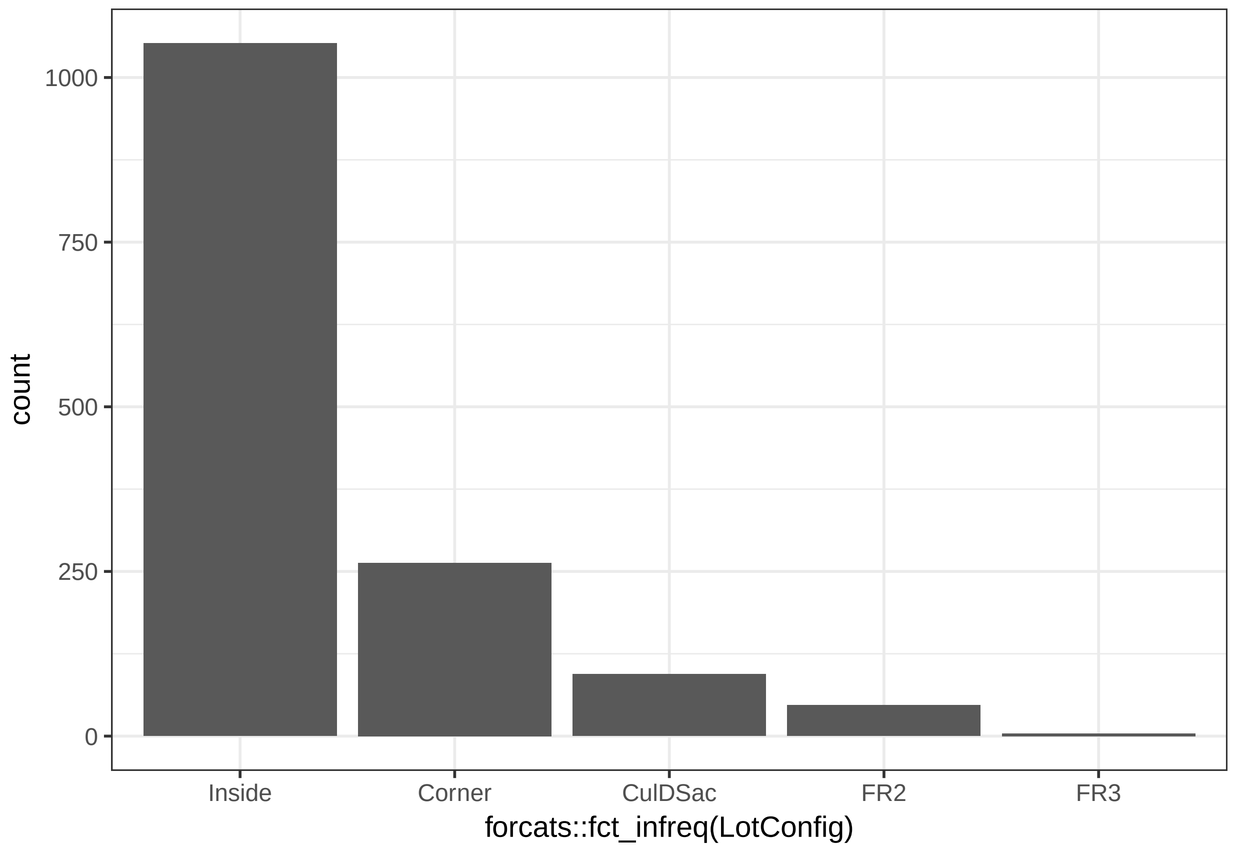
```
## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0x6

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0x6

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0xf
```



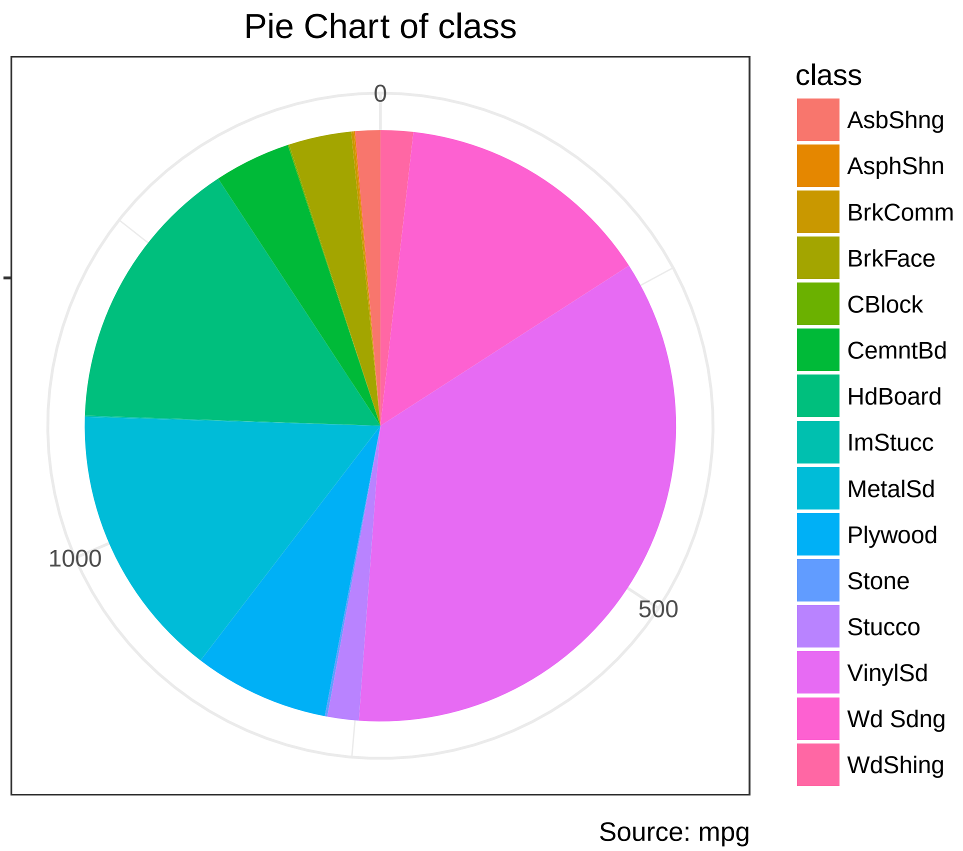
```
ggplot(data, aes(SalePrice)) +
geom_histogram(bins=nclass.Sturges(data$SalePrice)) +
xlab('Precio de las viviendas') +
ylab('') +
ggtitle('Histograma del precio de la vivienda')
```



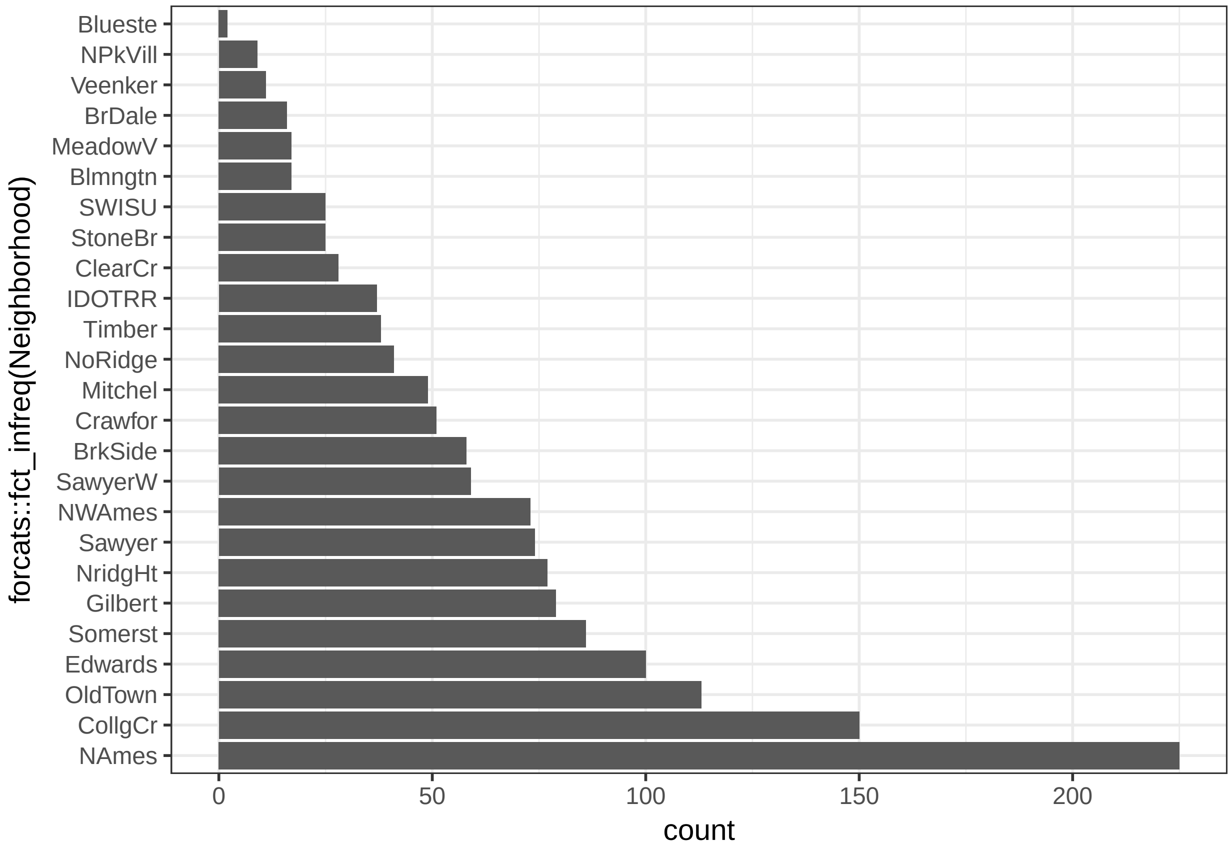
```r
ggplot(data, aes(forcats::fct_infreq(LotConfig))) +
  geom_bar()
```



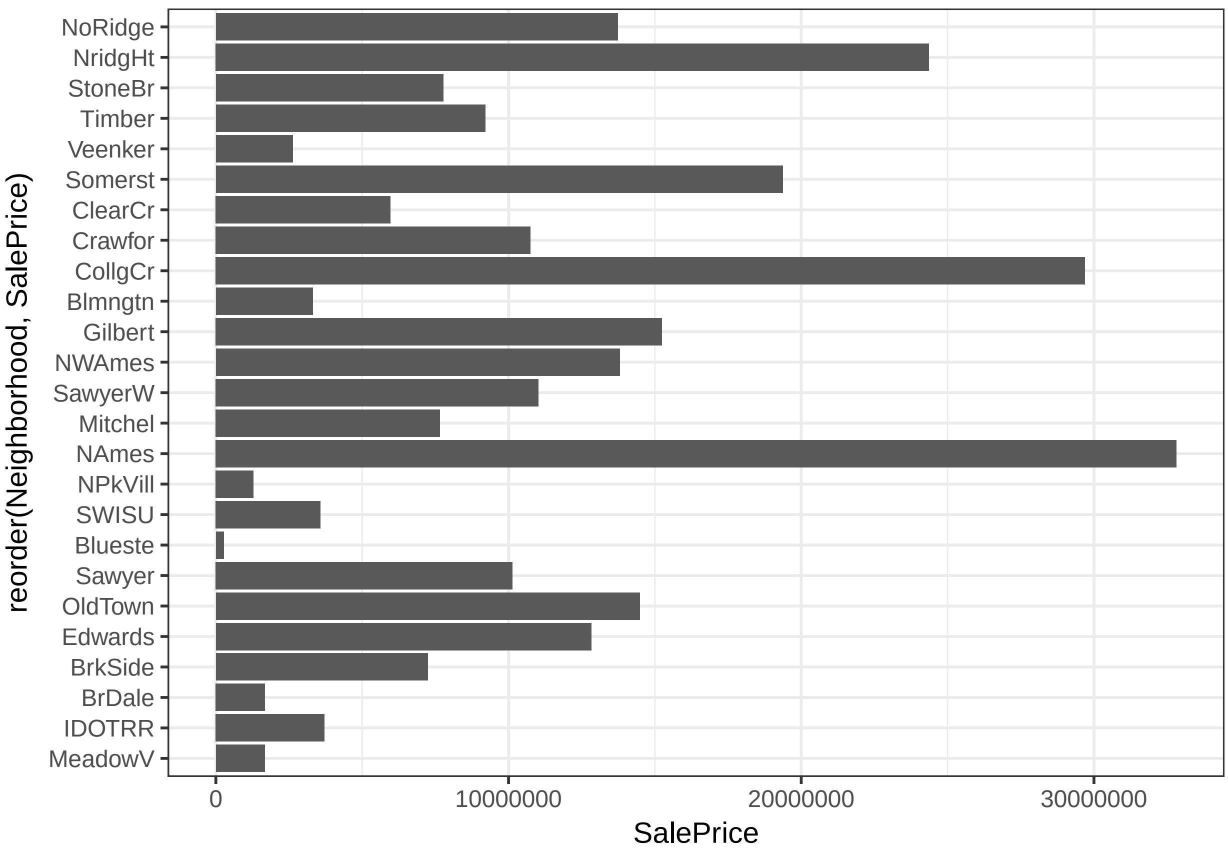
```
pie <- ggplot(data, aes(x = "", fill =Exterior1st )) +
  geom_bar(width = 1) +
  theme(axis.line = element_blank(),
        plot.title = element_text(hjust=0.5)) +
  labs(fill="class",
       x=NULL,
       y=NULL,
       title="Pie Chart of class",
       caption="Source: mpg")

pie + coord_polar(theta = "y", start=0)
```



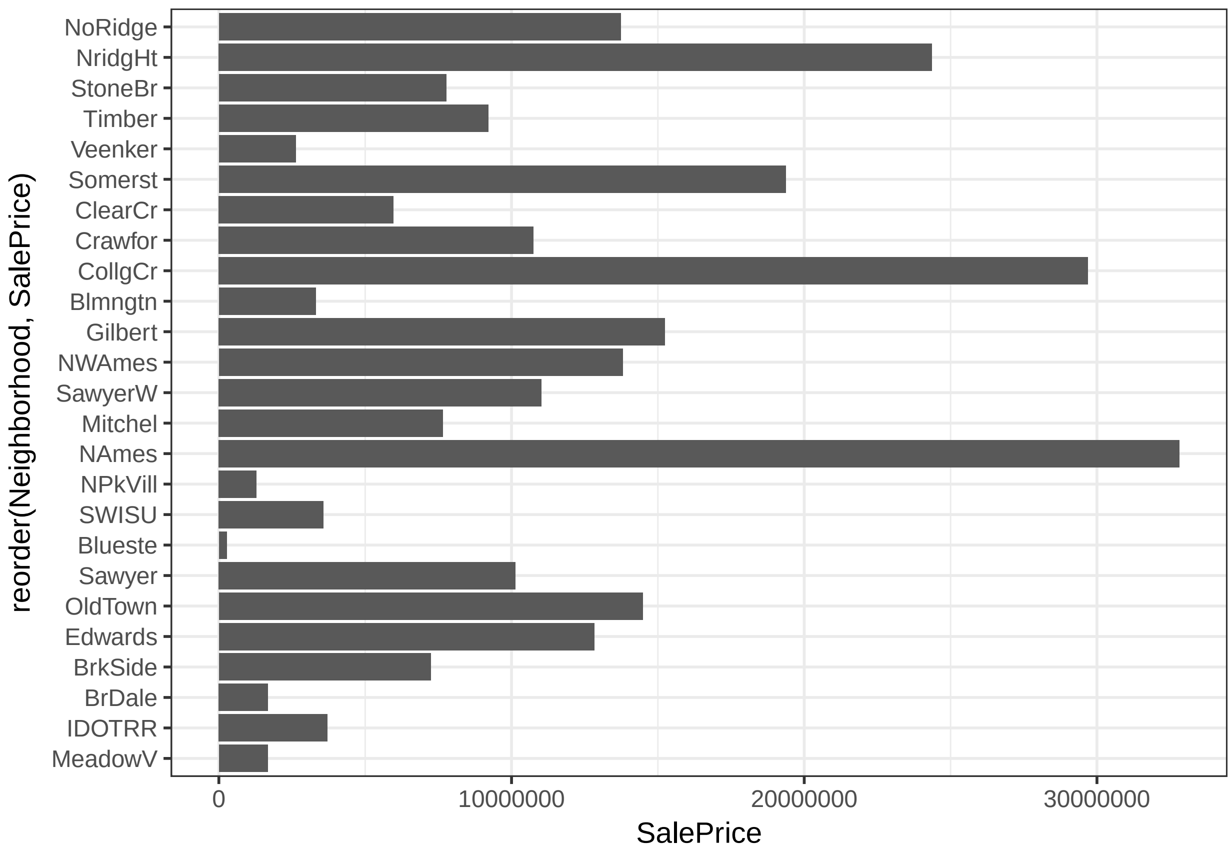
```
ggplot(data, aes(forcats::fct_infreq(Neighborhood))) +
  geom_bar() +
  coord_flip()
```



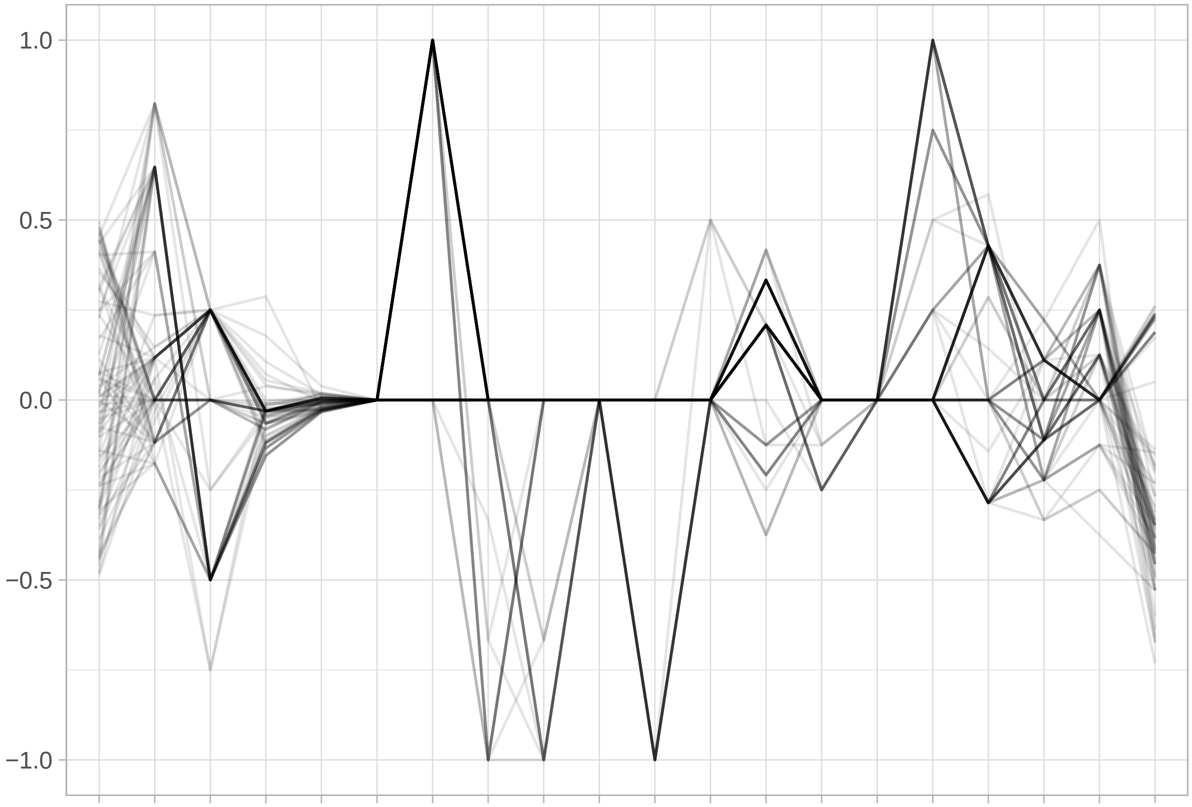
```
ggplot(data,
aes(reorder(Neighborhood, SalePrice), SalePrice)) +
geom_bar(stat='identity') + coord_flip()
```



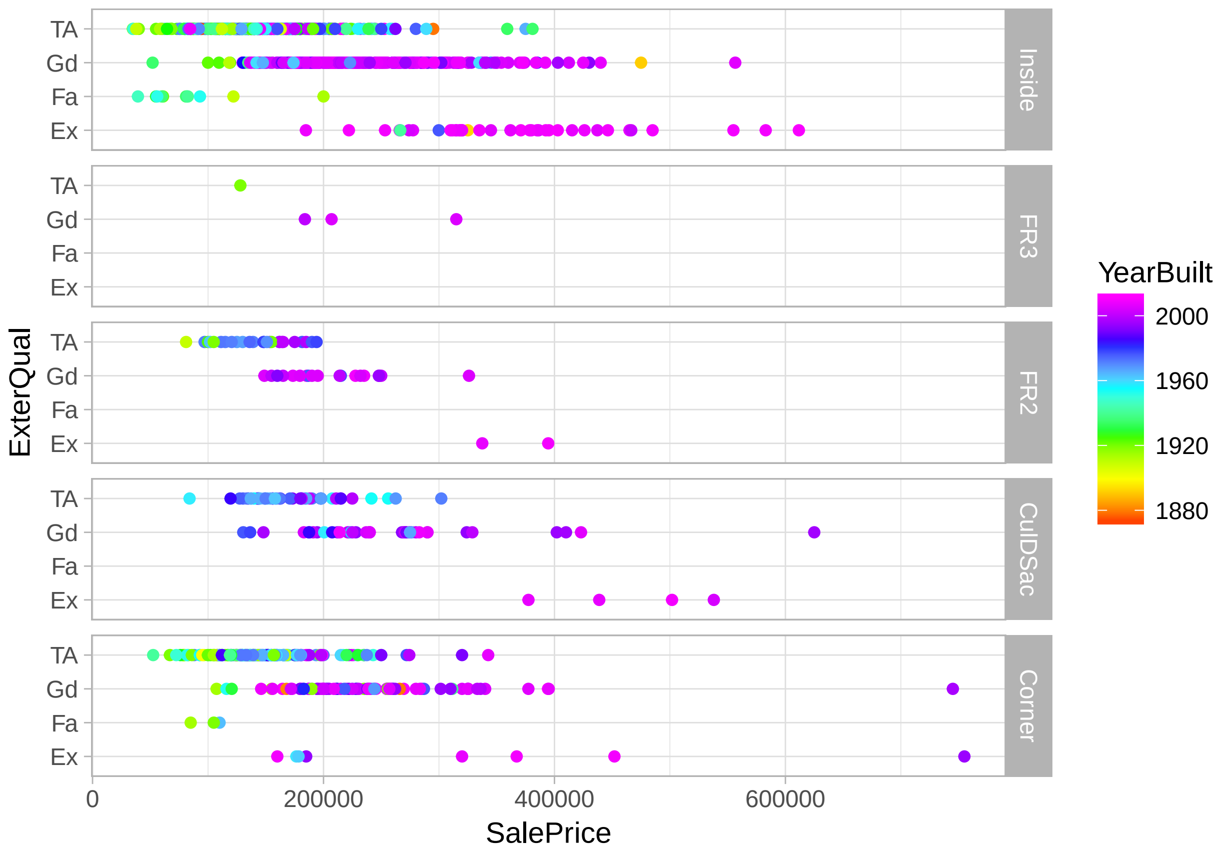
```
ggplot(data,
aes(reorder(Neighborhood, SalePrice), SalePrice)) +
geom_bar(stat='identity') + coord_flip()
```



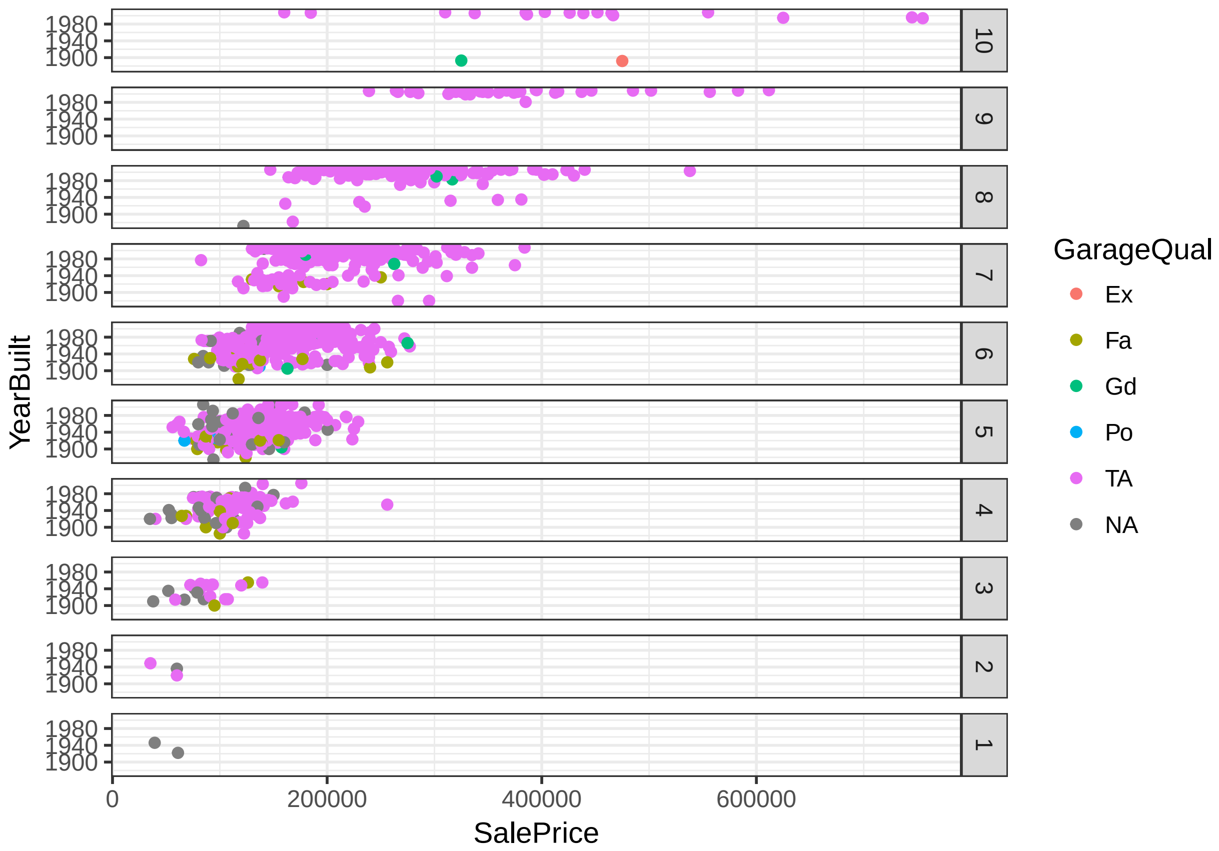
```
ggparcoord(data, columns=1:20, alphaLines=0.1,
scale='center', scaleSummary='median') +
xlab('') + ylab('') +
scale_x_discrete(labels=NULL) + theme_light()
```



```
ggplot(data, aes(ExterQual, SalePrice, colour=YearBuilt)) +
  geom_point() +
  facet_grid(LotConfig ~ ., as.table=FALSE) +
  theme_light() +
  coord_flip() +
  scale_colour_gradientn(colours=rainbow(6))
```

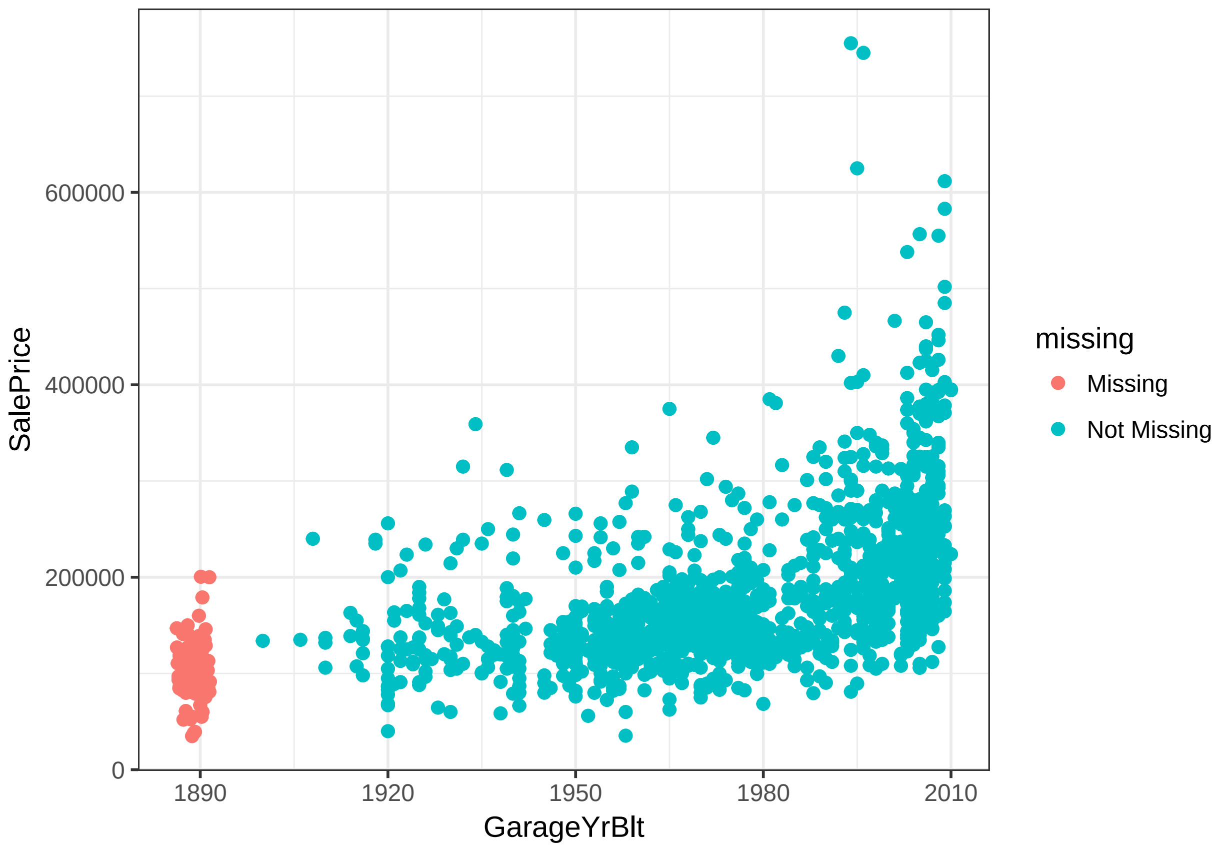


```r
ggplot(data, aes(YearBuilt, SalePrice, colour=GarageQual)) +
  geom_point() +
  facet_grid(as.factor(OverallQual) ~ ., as.table=FALSE) +
  coord_flip() +
  theme_bw()
```

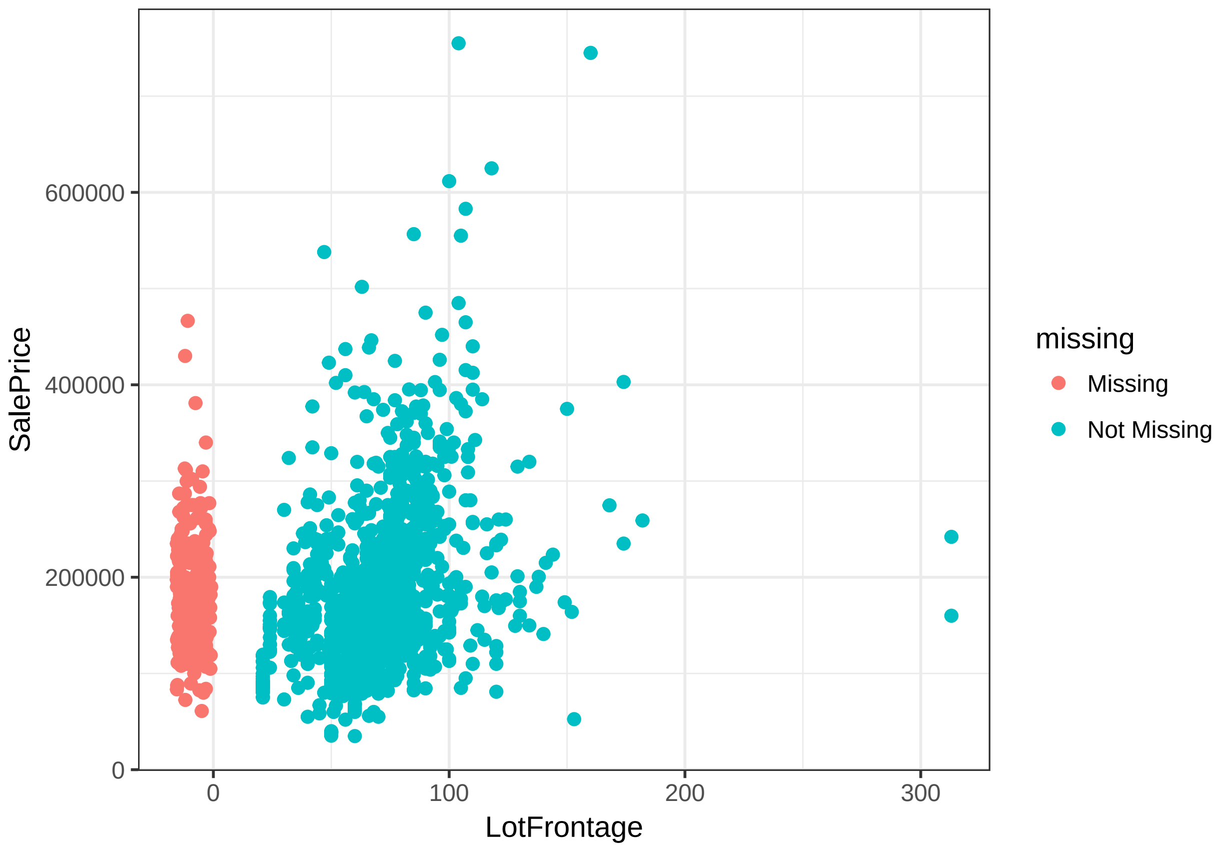


4. Análisis gráfico de valores perdidos

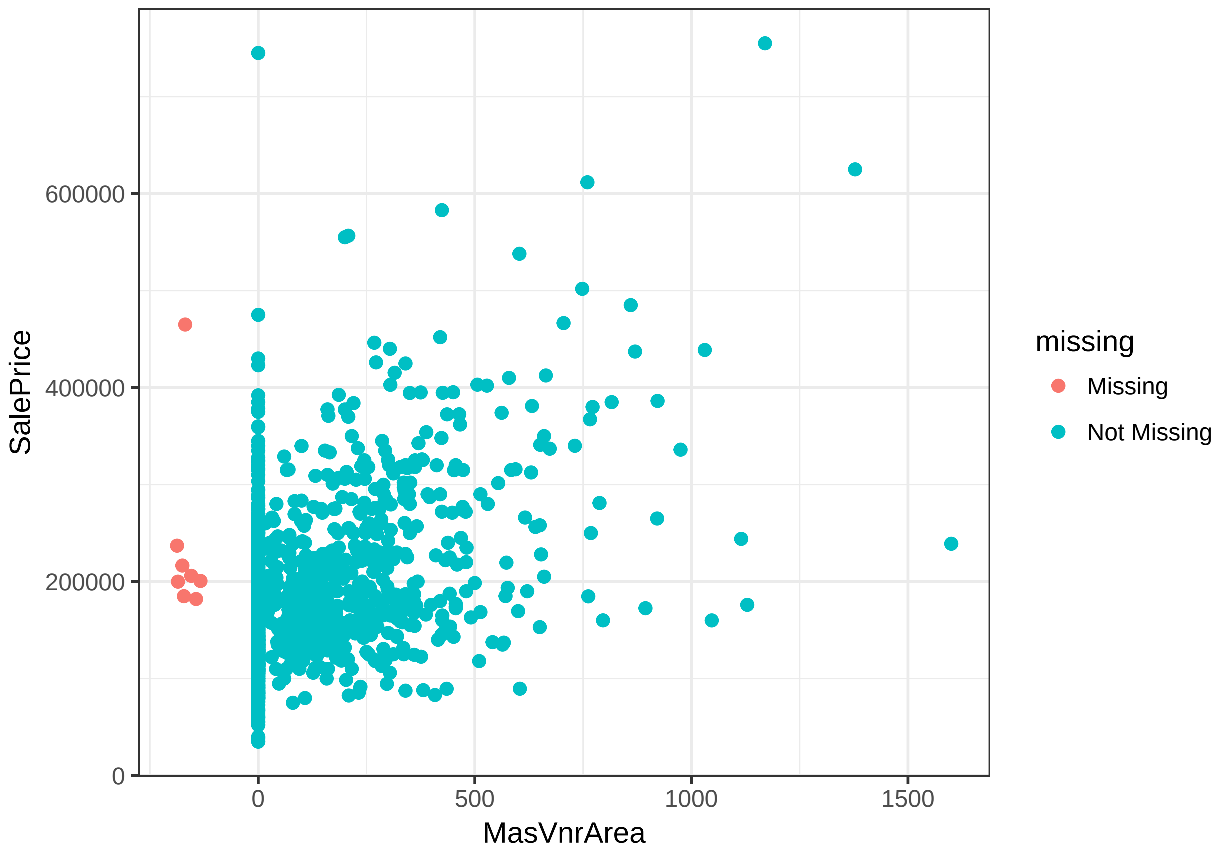
```
ggplot(data = data, aes(x=GarageYrBlt, y=SalePrice)) +
  geom_missing_point()
```



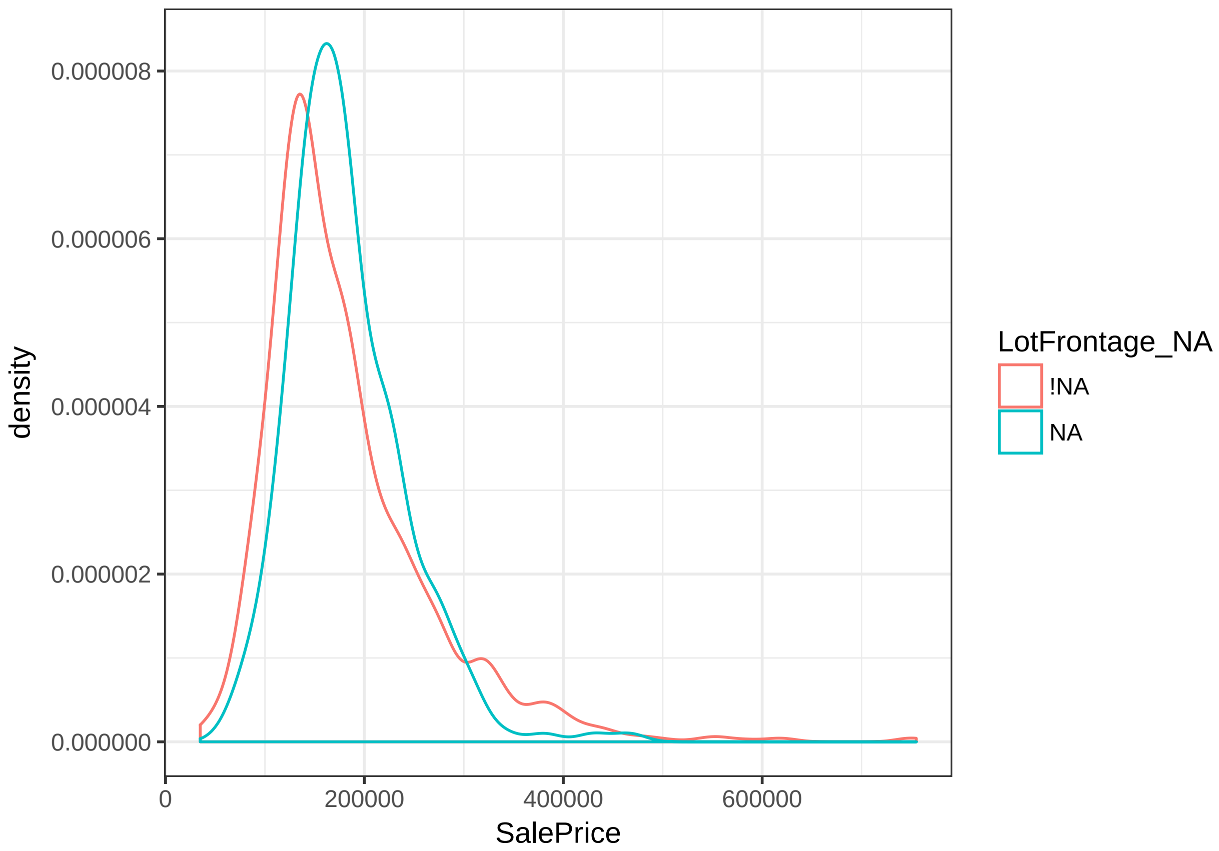
```
ggplot(data = data, aes(x=LotFrontage, y=SalePrice)) + geom_missing_point()
```



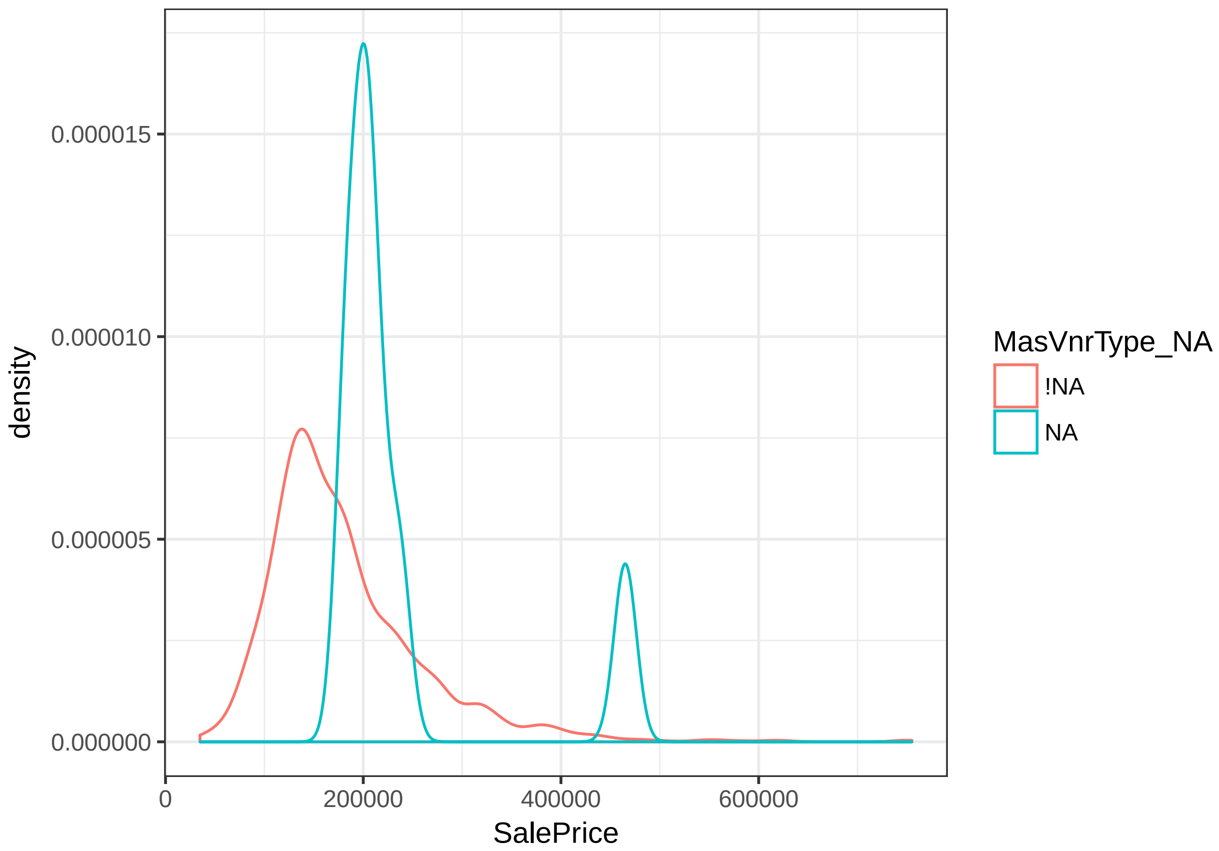
```
ggplot(data = data, aes(x=MasVnrArea, y=SalePrice)) + geom_missing_point()
```



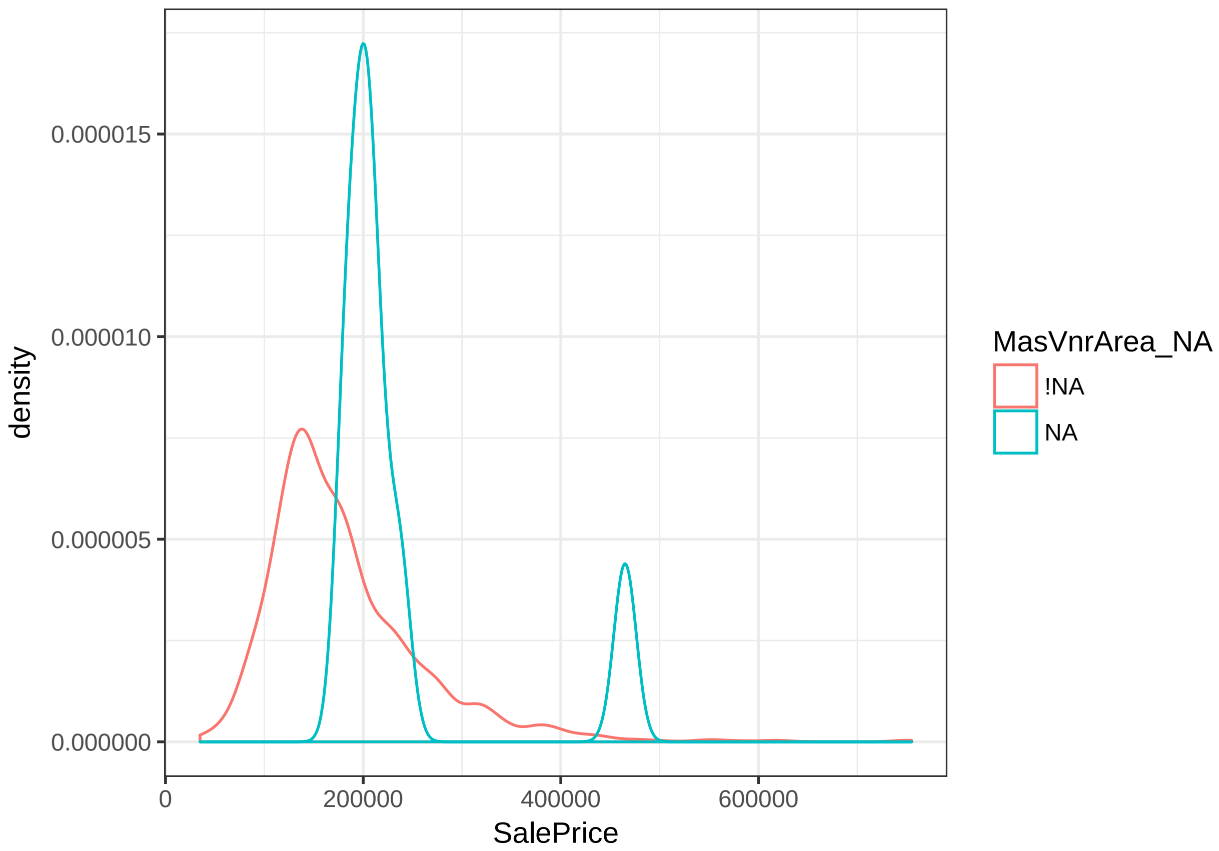
```
ggplot(data = bind_shadow(data),
aes(x = SalePrice, color = LotFrontage_NA)) +
geom_density()
```



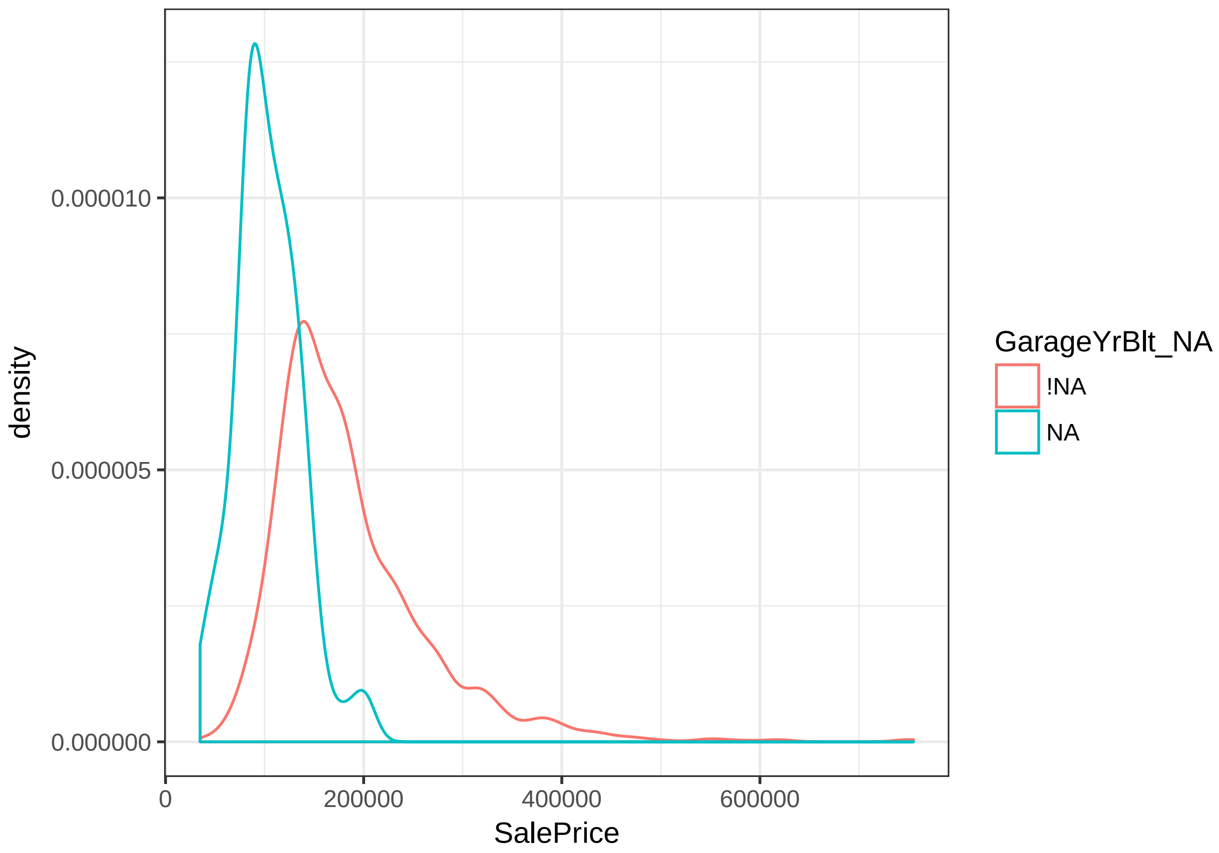
```
ggplot(data = bind_shadow(data),
aes(x = SalePrice, color = MasVnrType_NA)) +
geom_density()
```



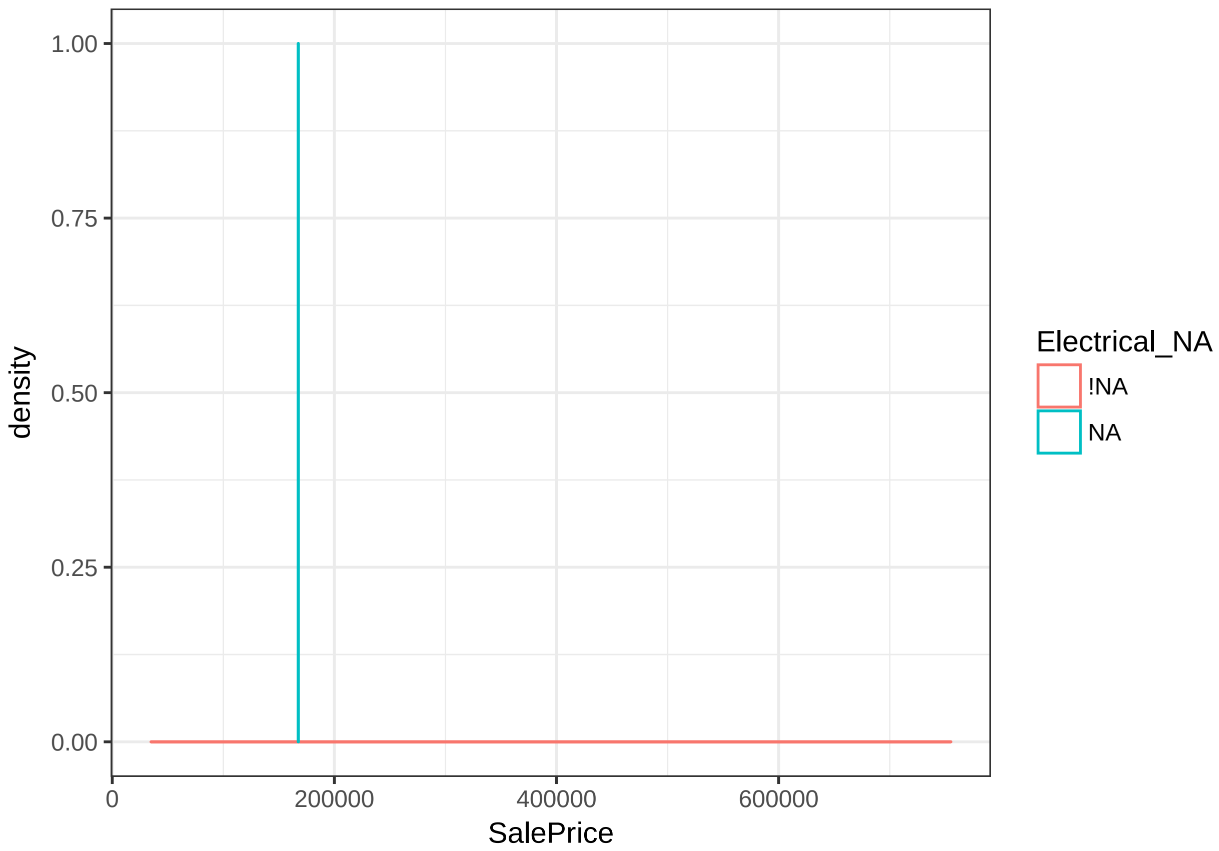
```
ggplot(data = bind_shadow(data),
aes(x = SalePrice, color = MasVnrArea_NA)) +
geom_density()
```



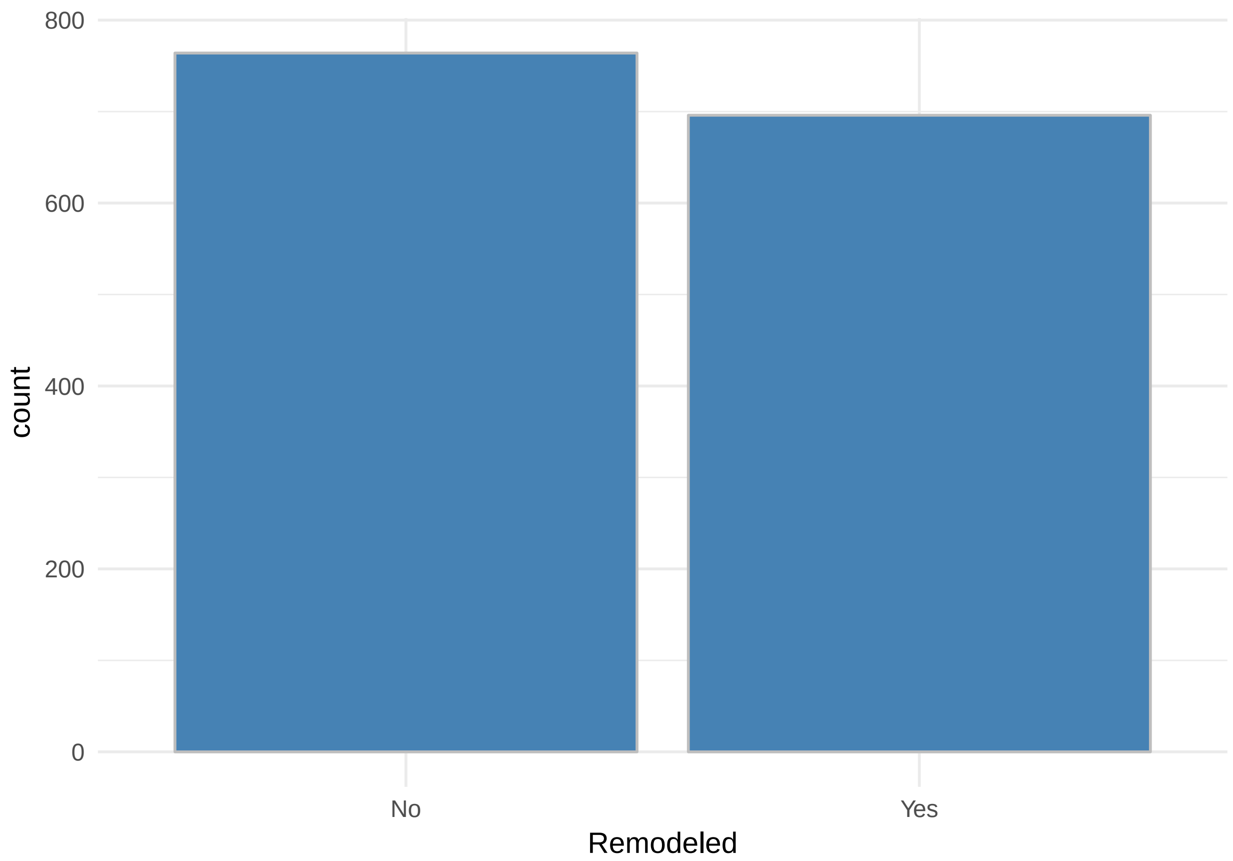
```
ggplot(data = bind_shadow(data),
aes(x = SalePrice, color = GarageYrBlt_NA)) +
geom_density()
```



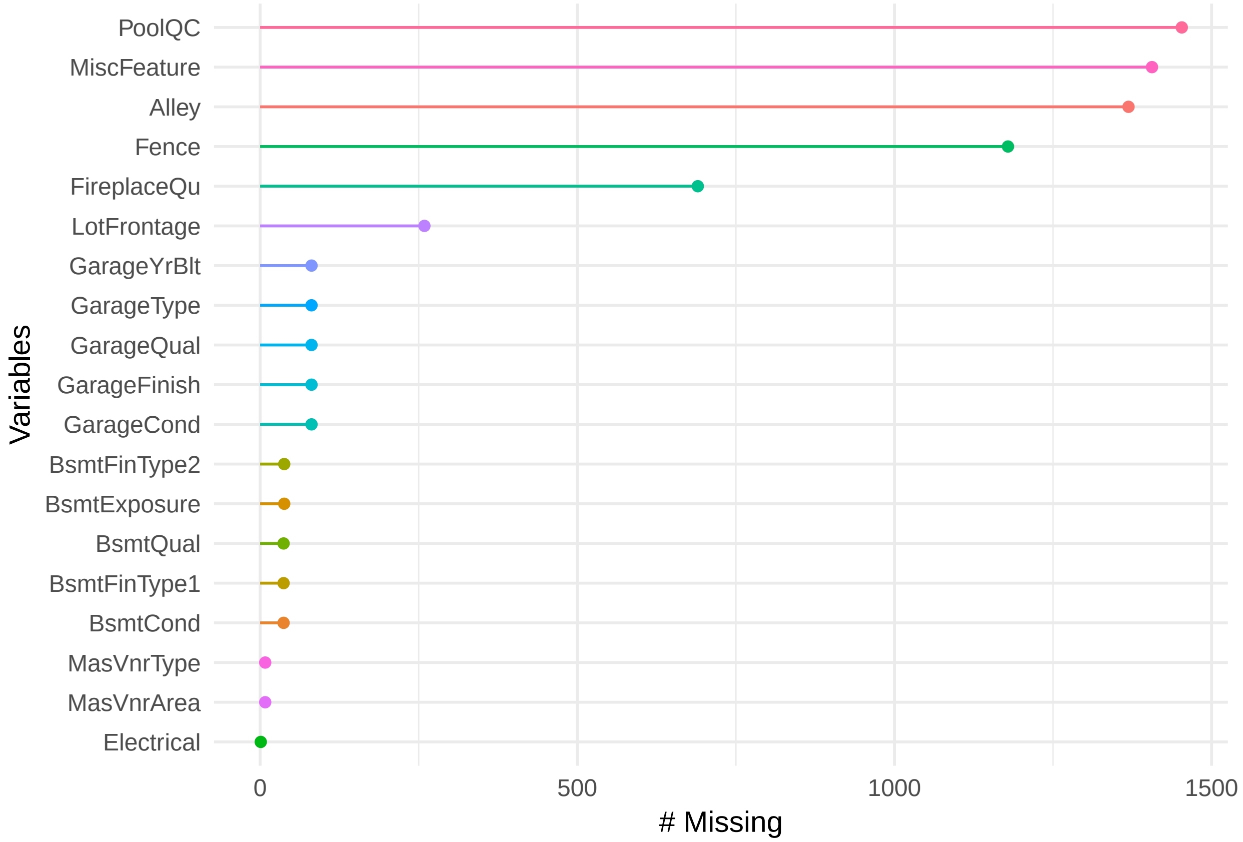
```
ggplot(data = bind_shadow(data),
aes(x = SalePrice, color = Electrical_NA)) +
geom_density()
```



```
data %>% select(YearBuilt, YearRemodAdd) %>%    mutate(Remodeled = as.integer(YearBuilt != YearRemodAdd
```

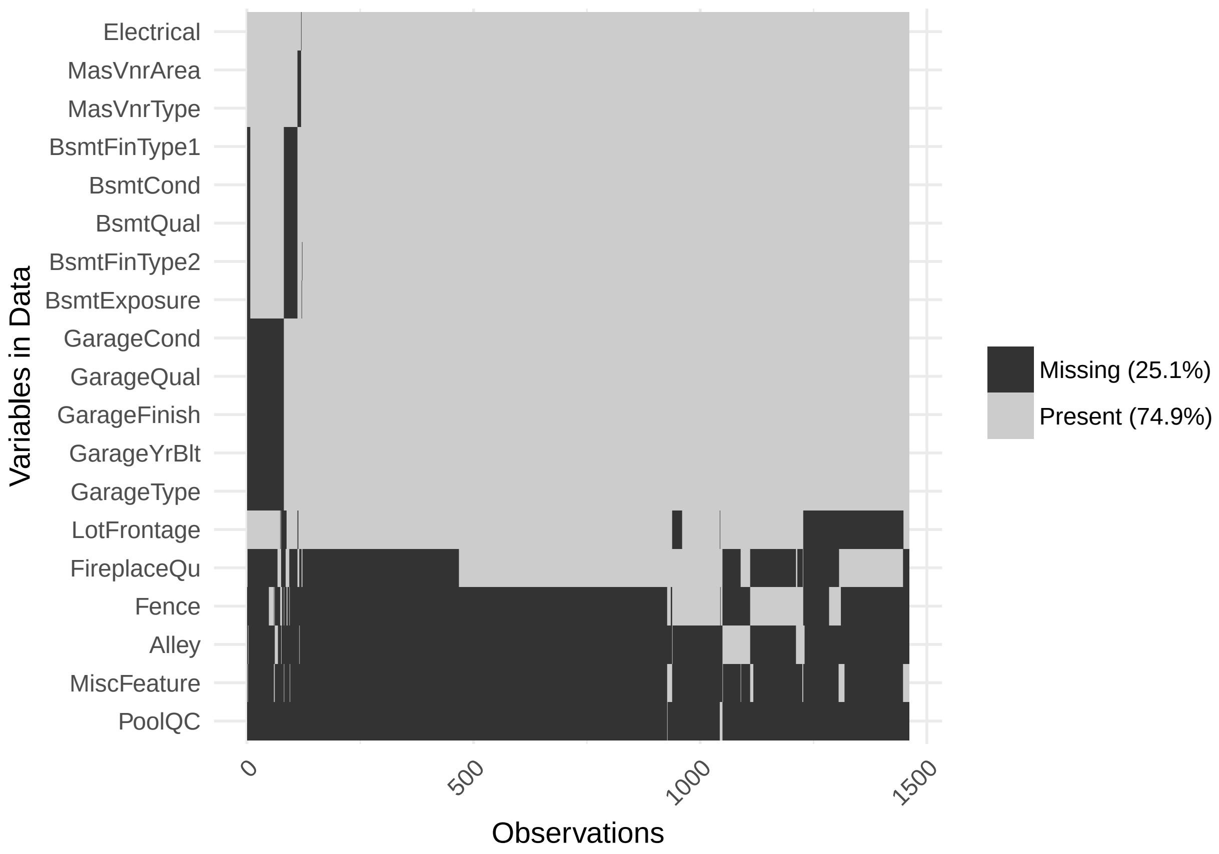


```
gg_missing_var(data[,colSums(is.na(data)) > 0])
```

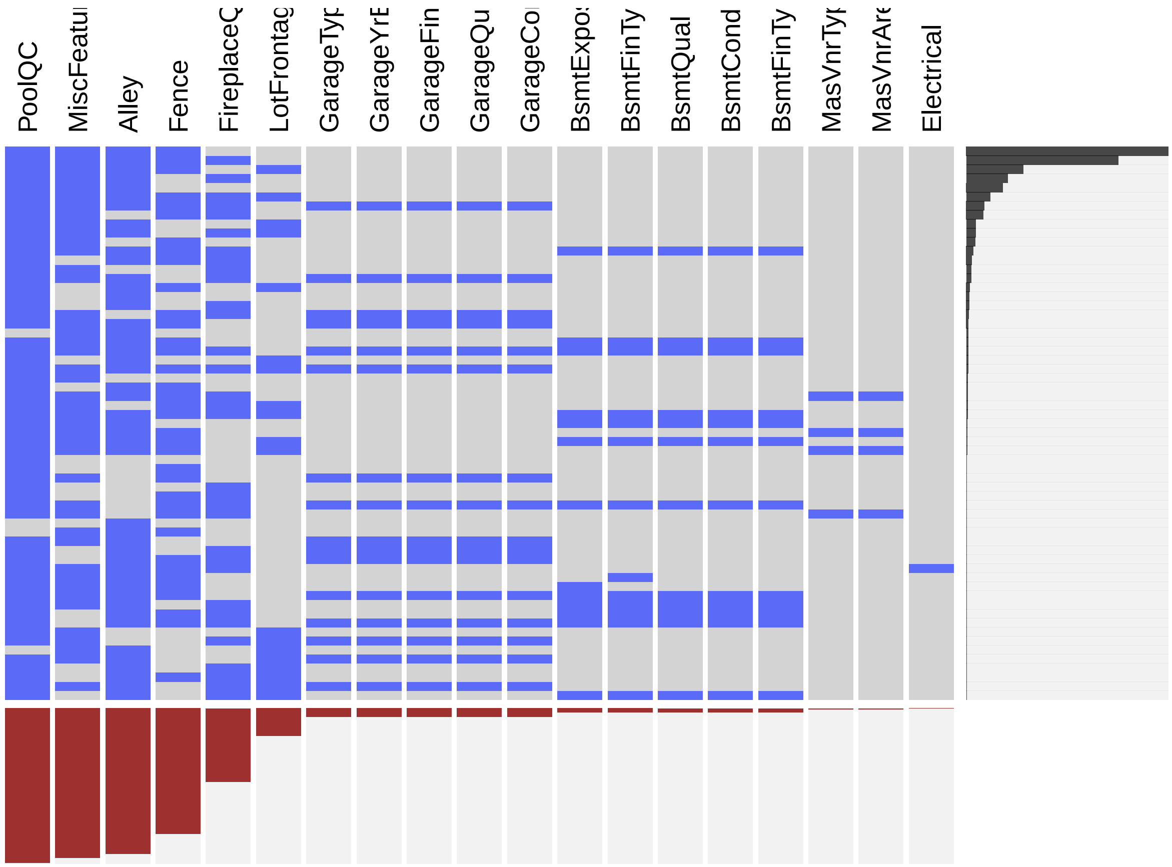


```
visdat::vis_miss(data[,colSums(is.na(data)) > 0], cluster = TRUE, sort_miss = TRUE) + coord_flip()
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

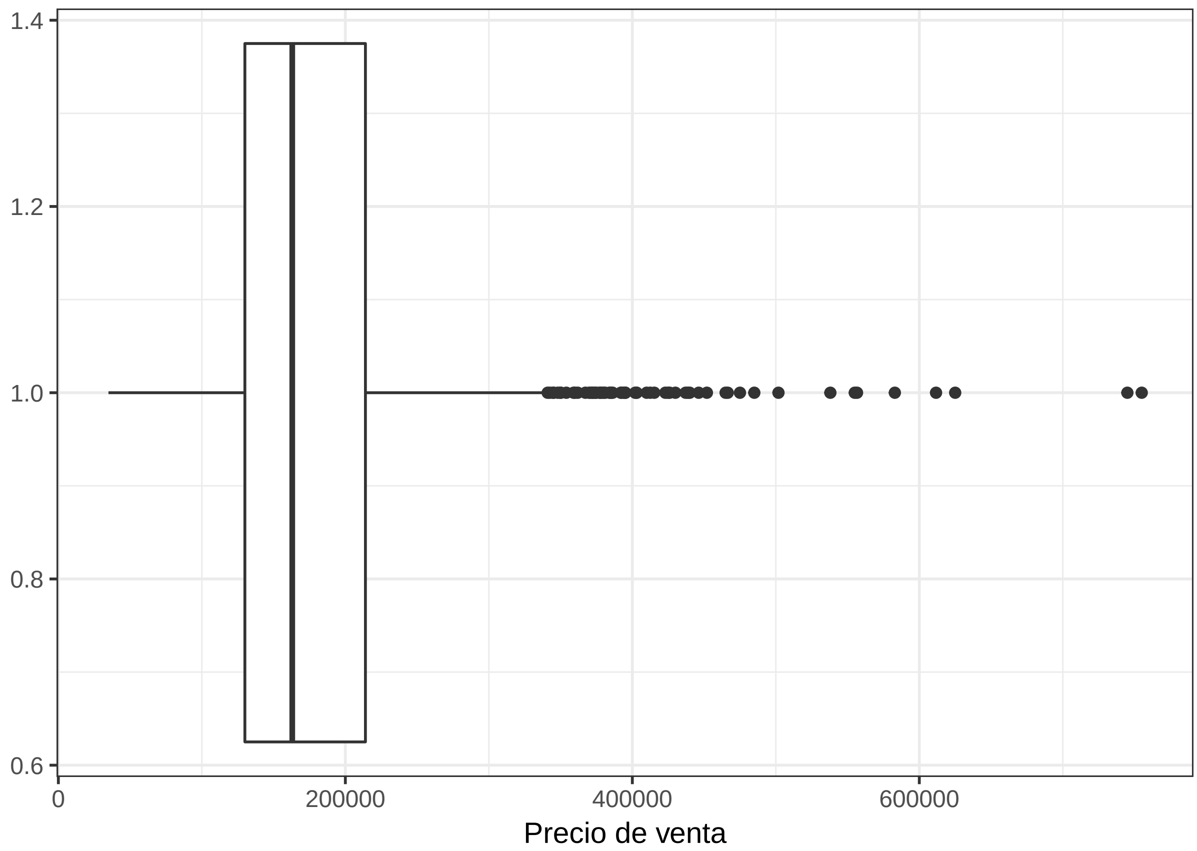


```r
extracat::visna(data[,colSums(is.na(data)) > 0], sort = "b")
```

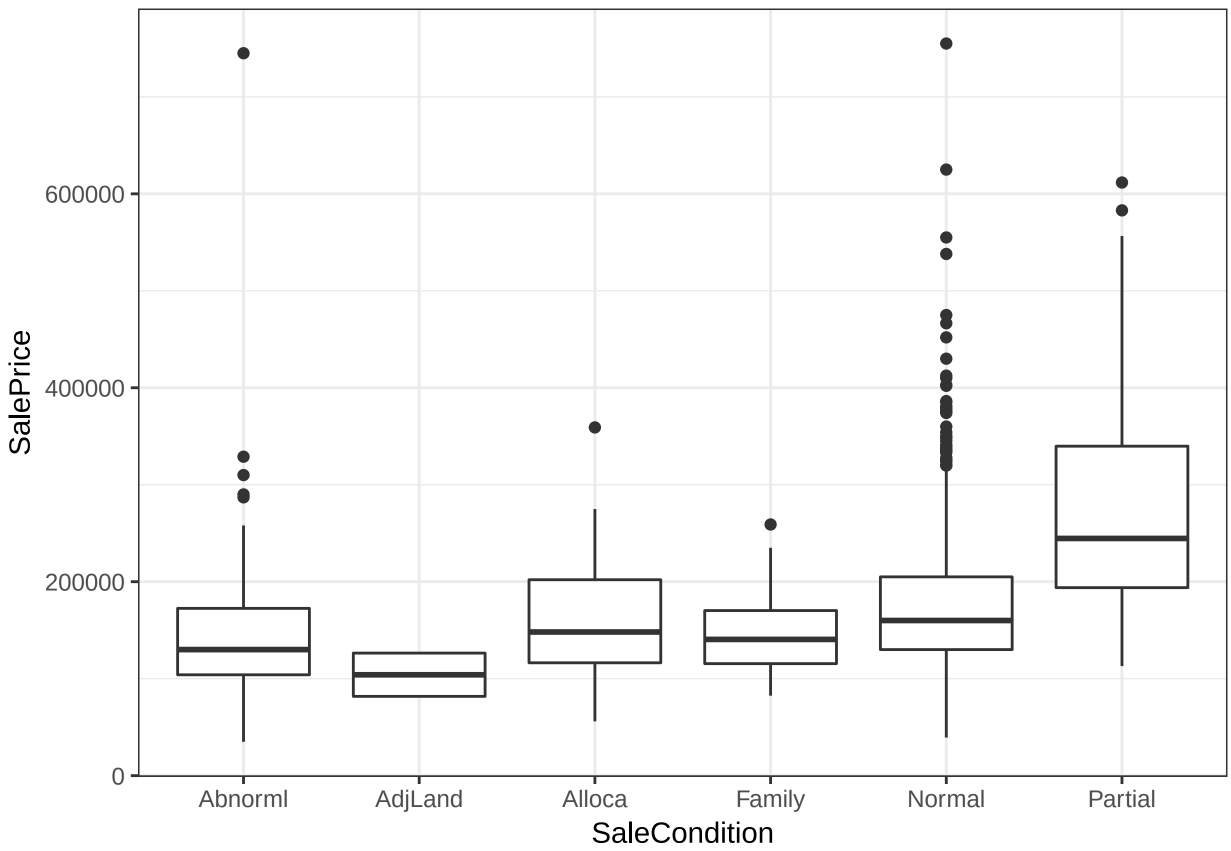


5. Análisis gráfico de valores atípicos

```
ggplot(data, aes(1, SalePrice)) +
geom_boxplot() + coord_flip() +
xlab('') +
ylab('Precio de venta')
```



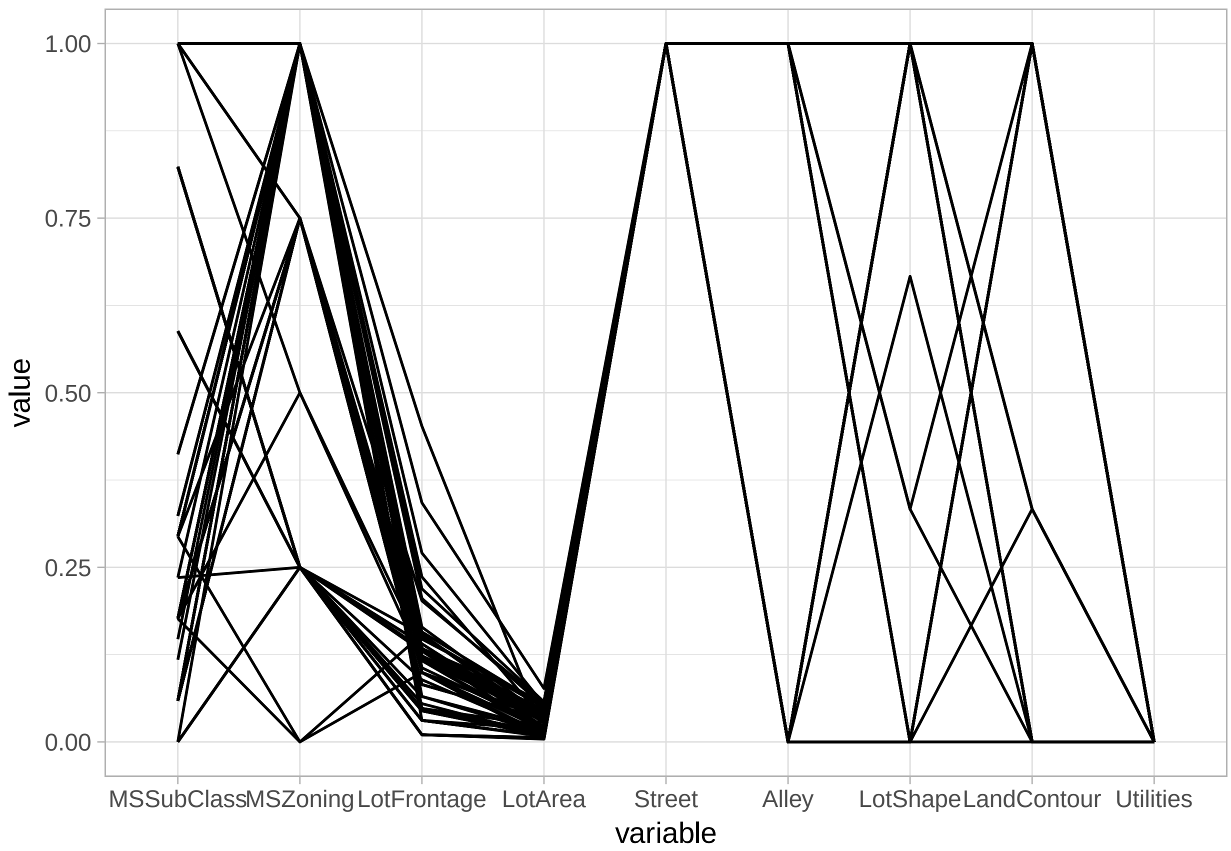
```
ggplot(data, aes(SaleCondition, SalePrice)) +
geom_boxplot()
```



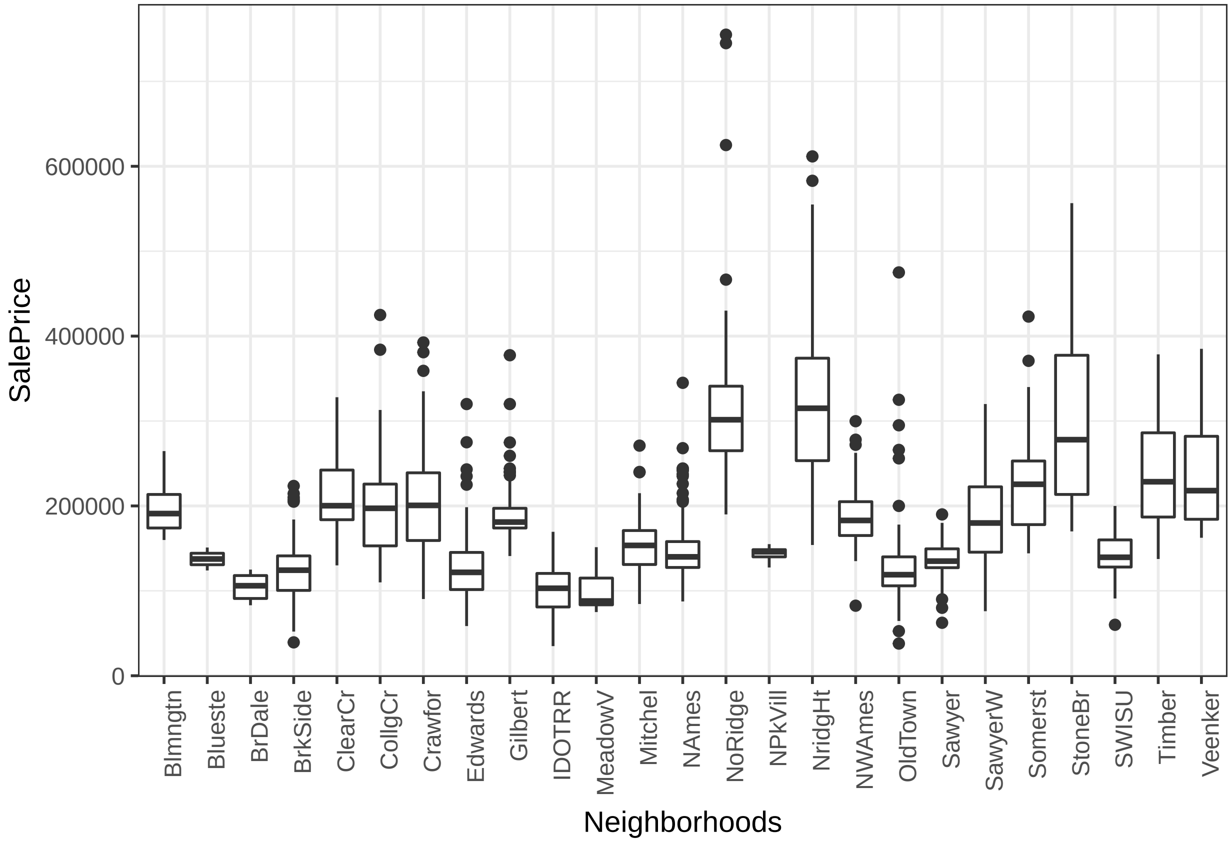
```
data %>%
select(SalePrice) %>%
filter(SalePrice > 350000)
```

```
##    SalePrice
## 1     385000
## 2     438780
## 3     383970
## 4     372402
## 5     412500
## 6     501837
## 7     475000
## 8     386250
## 9     403000
## 10    415298
## 11    360000
## 12    375000
## 13    354000
## 14    377426
## 15    437154
## 16    394432
## 17    426000
## 18    555000
## 19    440000
## 20    380000
## 21    374000
## 22    430000
## 23    402861
## 24    446261
## 25    369900
## 26    451950
## 27    359100
## 28    370878
## 29    402000
## 30    423000
## 31    372500
## 32    392000
## 33    755000
## 34    361919
## 35    538000
## 36    395000
## 37    485000
## 38    582933
## 39    385000
## 40    611657
## 41    395192
## 42    556581
## 43    424870
## 44    625000
## 45    392500
## 46    745000
## 47    367294
## 48    465000
## 49    378500
## 50    381000
## 51    410000
## 52    466500
## 53    377500
## 54    394617
```

```
ggparcoord(data, columns = 2:10,
scale = "uniminmax") + theme_light()
```

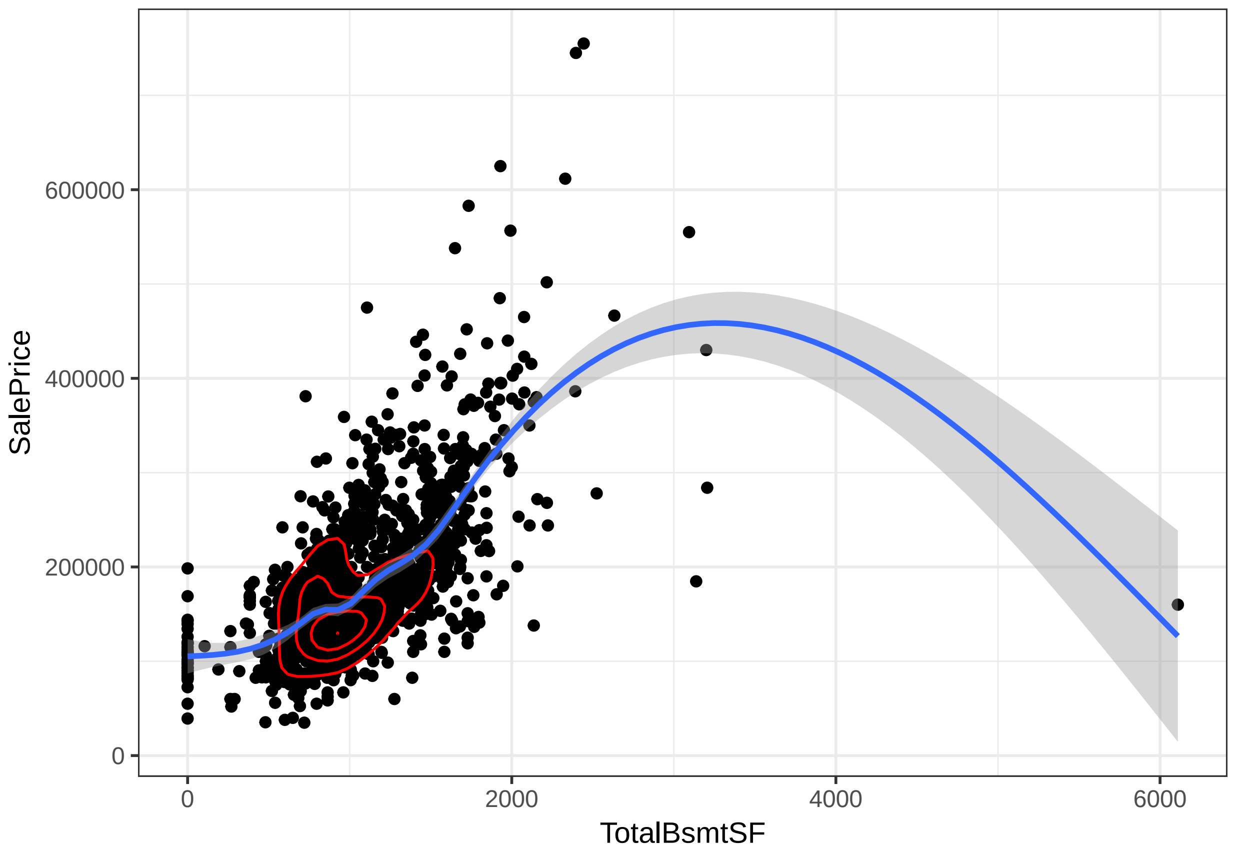


```
ggplot(data, aes(factor(Neighborhood), SalePrice)) + geom_boxplot() + theme(axis.text.x = element_text(
```



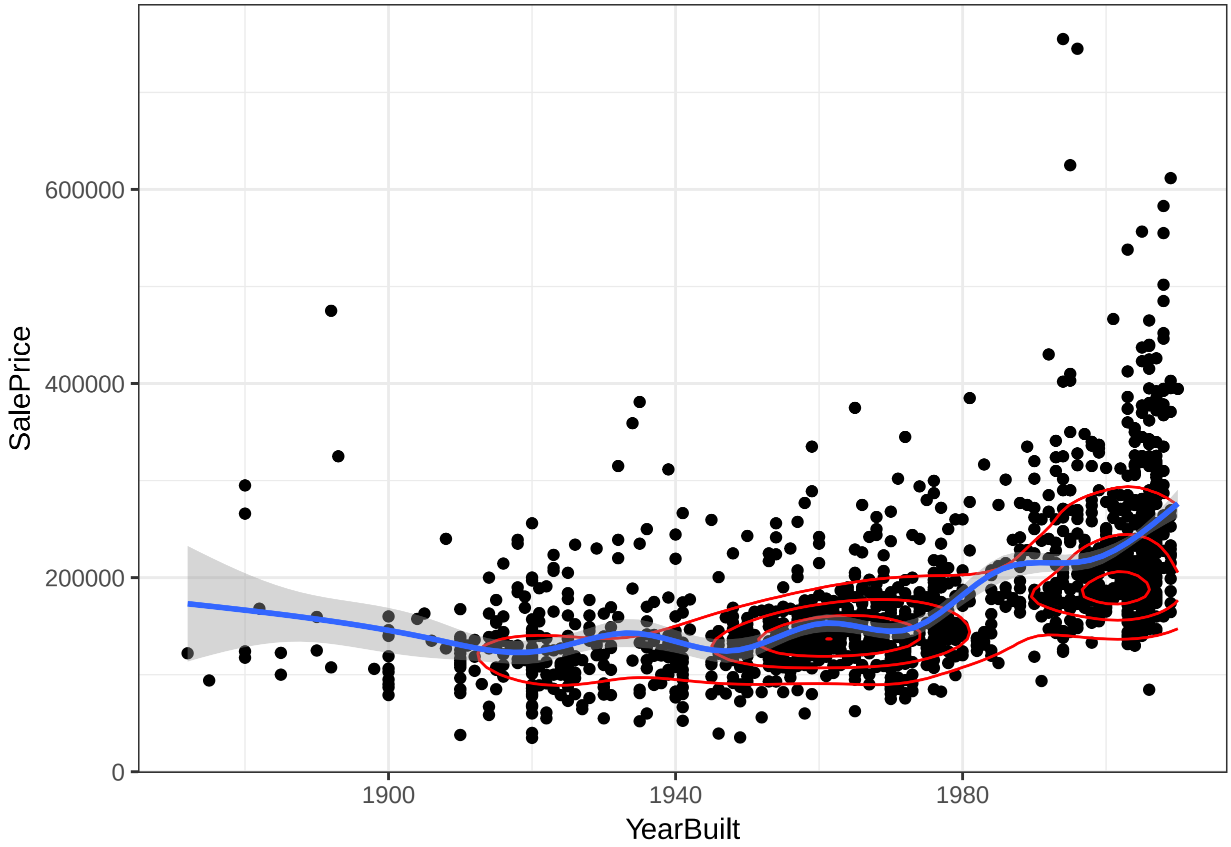
```
ggplot(data, aes(TotalBsmtSF, SalePrice)) + geom_point() +
geom_density2d(bins = 4, color = "red") +
geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```



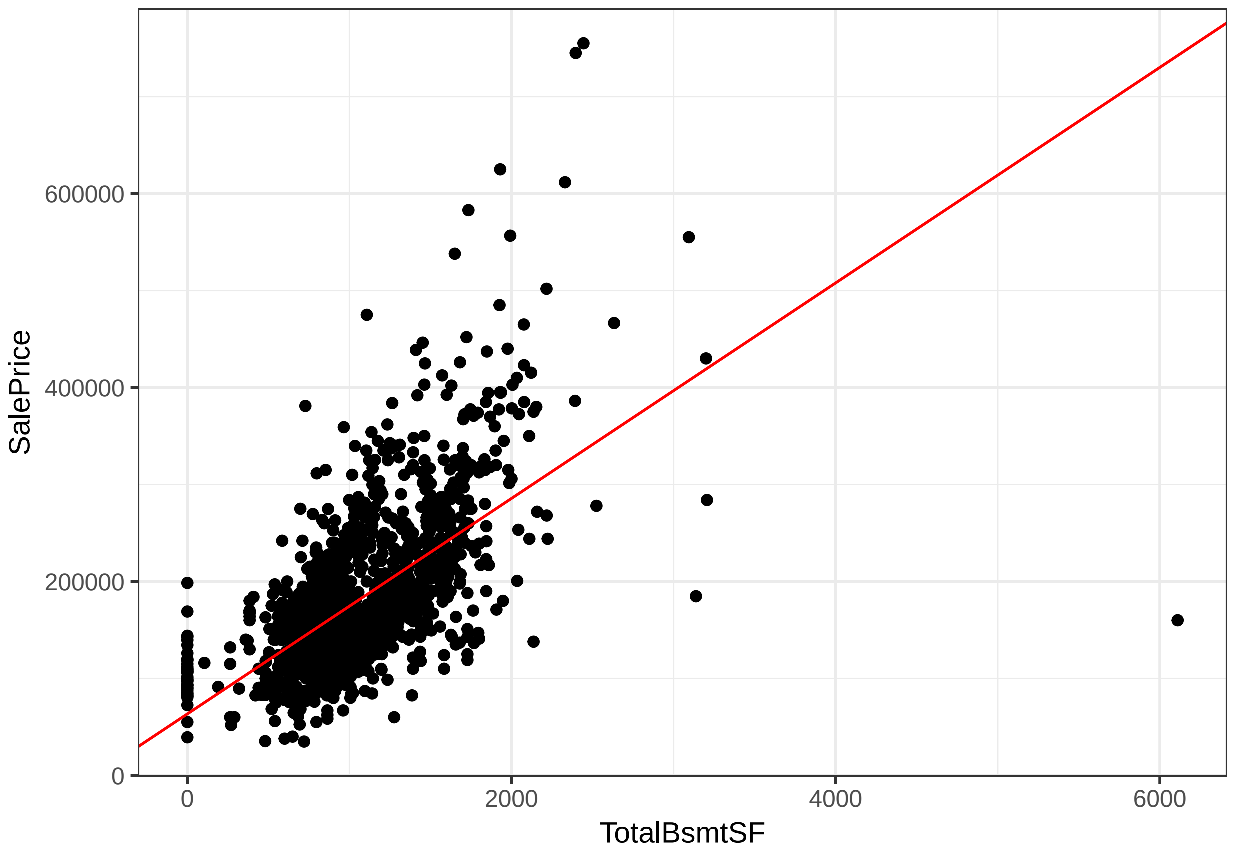
```r
ggplot(data, aes(YearBuilt, SalePrice)) + geom_point() +
geom_density2d(bins = 4, color = "red") +
geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```

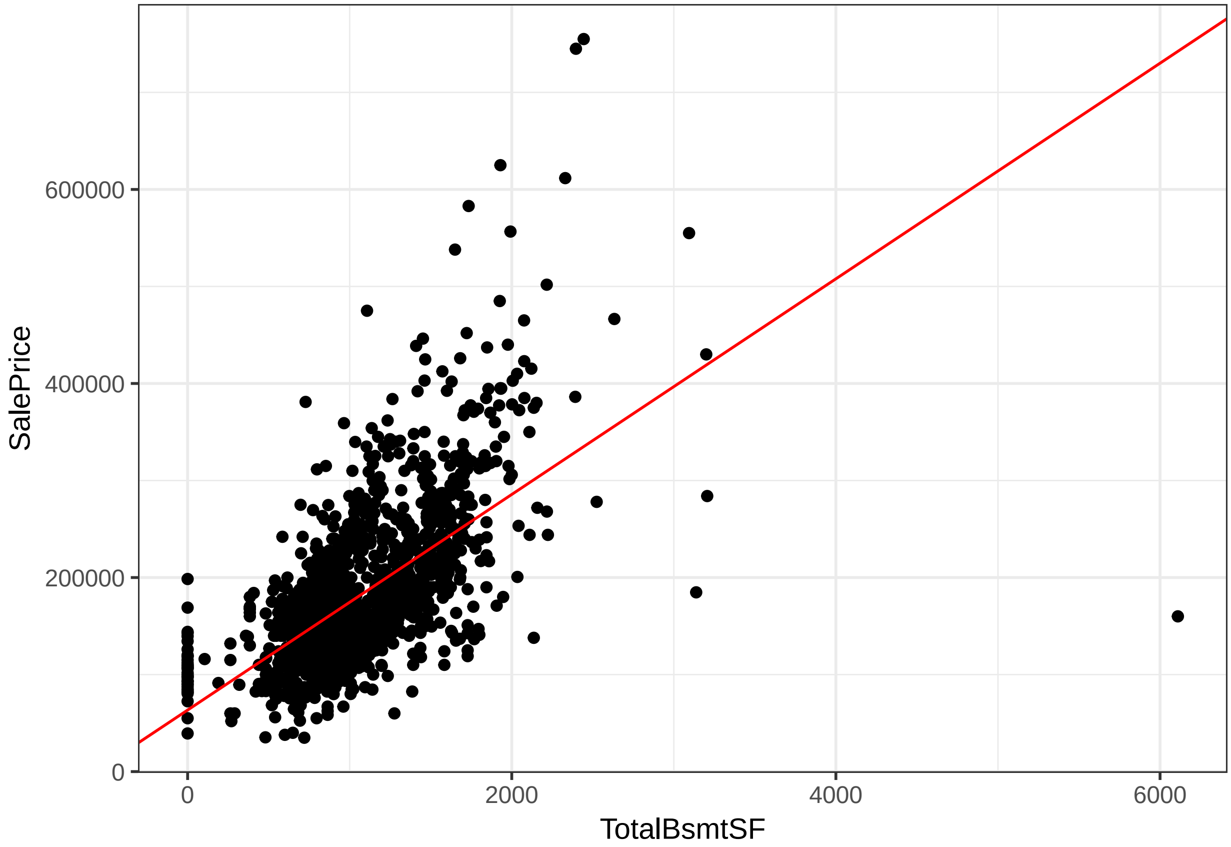


6. Análisis gráfico con modelos

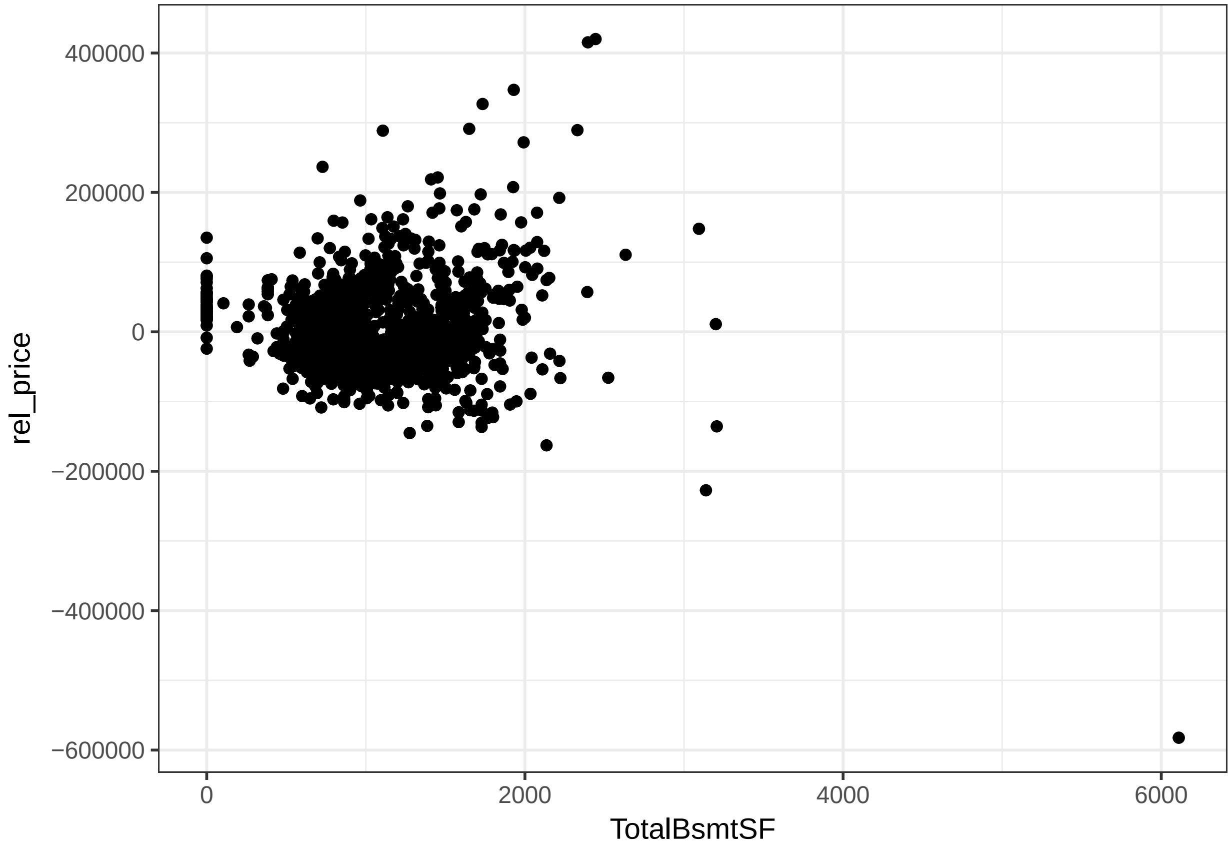
```r
modelo_lineal <- lm(SalePrice ~ TotalBsmtSF, data = data)

ggplot(data, aes(TotalBsmtSF, SalePrice)) + geom_point() +
geom_abline(intercept = coef(modelo_lineal)[1],
slope = coef(modelo_lineal)[2],
color = "red")
```



```
modelo_lineal <- lm(SalePrice ~ TotalBsmtSF, data = data)

ggplot(data, aes(TotalBsmtSF, SalePrice)) + geom_point() +
geom_abline(intercept = coef(modelo_lineal)[1],
slope = coef(modelo_lineal)[2],
color = "red")
```



```r
data2 <- data %>% mutate(rel_price = resid(modelo_lineal))
ggplot(data2, aes(TotalBsmtSF, rel_price)) +
geom_point()
```



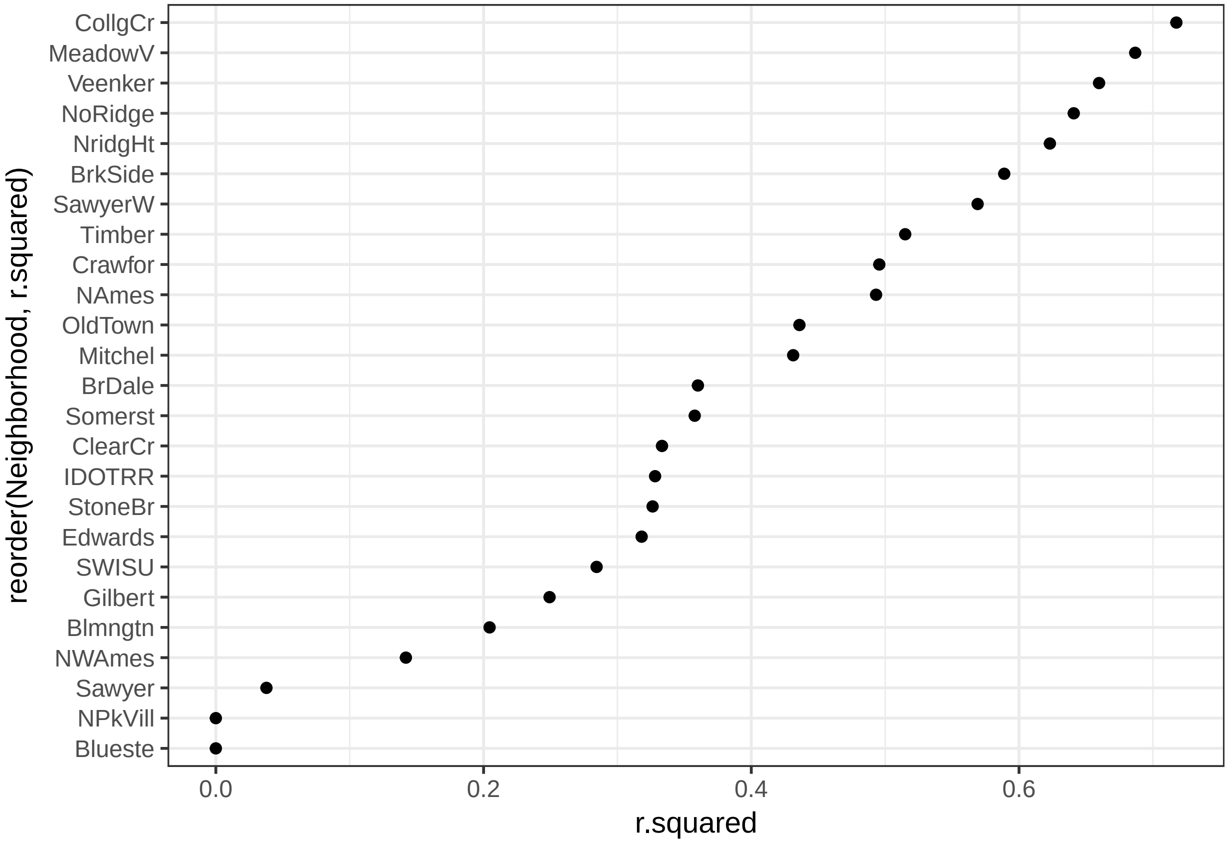
```
deseas <- function(y, x) {
resid(lm(y ~ factor(x), na.action = na.exclude))
}


data3 <- data %>%
group_by(Neighborhood) %>%
mutate(rel_sales = deseas(OverallQual, SalePrice))

models <- data3 %>%
group_by(Neighborhood) %>%
do(mod = lm(log2(SalePrice) ~ OverallQual,
data = ., na.action = na.exclude))
head(models)
```

```
## # A tibble: 6 x 2
##   Neighborhood      mod
##         <fctr>   <list>
## 1      Blmngtn <S3: lm>
## 2      Blueste <S3: lm>
## 3       BrDale <S3: lm>
## 4      BrkSide <S3: lm>
## 5      ClearCr <S3: lm>
## 6      CollgCr <S3: lm>
```

```
model_sum <- models %>% broom::glance(mod)
head(model_sum, 4)
```

```
## Source: local data frame [4 x 12]
## Groups: Neighborhood [4]
##
## # A tibble: 4 x 12
##   Neighborhood r.squared adj.r.squared     sigma statistic
##         <fctr>     <dbl>         <dbl>     <dbl>     <dbl>
## 1      Blmngtn 0.2044985     0.1514651 0.1968872   3.85603
## 2      Blueste 0.0000000     0.0000000 0.2009657        NA
## 3       BrDale 0.3601234     0.3144180 0.1666412   7.87922
## 4      BrkSide 0.5889607     0.5816207 0.3231762  80.24000
## # ... with 7 more variables: p.value <dbl>, df <int>, logLik <dbl>,
## #   AIC <dbl>, BIC <dbl>, deviance <dbl>, df.residual <int>
```

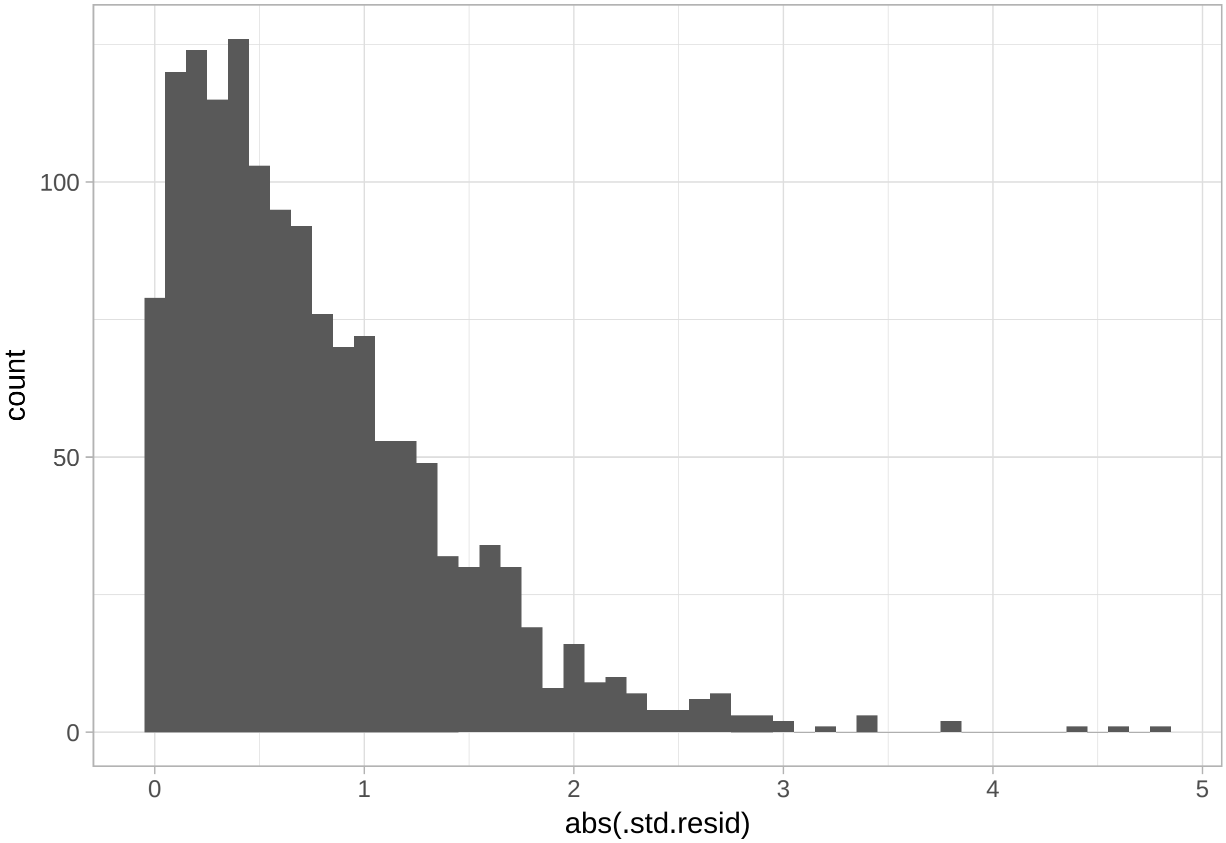
```r
ggplot(model_sum, aes(r.squared, reorder(Neighborhood, r.squared))) +
geom_point()
```



```r
obs_sum <- models %>% broom::augment(mod)
head(obs_sum, 5)
```

```
## Source: local data frame [5 x 10]
## Groups: Neighborhood [1]
##
## # A tibble: 5 x 10
##   Neighborhood log2.SalePrice. OverallQual  .fitted    .se.fit
##         <fctr>           <dbl>       <int>    <dbl>      <dbl>
## 1      Blmngtn        17.35156           7 17.51335 0.05262032
## 2      Blmngtn        17.55450           7 17.51335 0.05262032
## 3      Blmngtn        17.55075           8 17.75932 0.11367289
## 4      Blmngtn        17.39624           7 17.51335 0.05262032
## 5      Blmngtn        17.44750           7 17.51335 0.05262032
## # ... with 5 more variables: .resid <dbl>, .hat <dbl>, .sigma <dbl>,
## #   .cooksd <dbl>, .std.resid <dbl>
```

```
ggplot(obs_sum, aes(abs(.std.resid))) +
geom_histogram(binwidth = 0.1) + theme_light()
```


count
abs(.std.resid)