Filtragem de variáveis baseadas no coeficiente de correlação/contingência

Valor absoluto do coeficiente de correlação/contingência utilizado nos resultados abaixo:

0.5

Variáveis numéricas:

Quantidade: 29

Variáveis que NÃO possuem uma correlação forte com alguma variável:

MSSubClass, LotFrontage, LotArea, OverallCond, MasVnrArea, BsmtFinSF2, LowQualFinSF, WoodDeckSF, OpenPorchSF, EnclosedPorch, X3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold

Quantidade: 16

Variáveis que possuem uma correlação forte com alguma variável:

OverallQual, YearBuilt, YearRemodAdd, BsmtFinSF1, BsmtUnfSF, TotalBsmtSF, X1stFlrSF, X2ndFlrSF, GrLivArea, TotRmsAbvGrd, GarageYrBlt, GarageCars, GarageArea

Quantidade: 13

Variáveis que conseguem estar correlacionadas com todas as outras variáveis:

OverallQual, TotalBsmtSF, GrLivArea, BsmtFinSF1

Quantidade: 4

Variáveis categóricas:

Quantidade: 46

Variáveis que NÃO possuem uma associação forte com alguma variável:

Street, LotShape, Utilities, LotConfig, BsmtExposure, BsmtFinType2, Heating, CentralAir, Functional, PavedDrive, Fireplaces

Quantidade: 11

Variáveis que possuem uma associação forte com alguma variável:

MSZoning, LandContour, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtFinType1, HeatingQC, Electrical, KitchenQual, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, SaleType, SaleCondition, BsmtHalfBath, BsmtFullBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr

Quantidade: 35

Variáveis que conseguem estar associadas com todas as outras variáveis:

Neighborhood, Condition2, BldgType, HouseStyle, BsmtFinType1, GarageQual, LandContour, RoofStyle, SaleType, KitchenAbvGr

Quantidade: 10

Resultado da extração de variáveis:

Variáveis que foram escolhidas:

Street, LotShape, Utilities, LotConfig, BsmtExposure, BsmtFinType2, Heating, CentralAir, Functional, PavedDrive, Fireplaces, MSSubClass, LotFrontage, LotArea, OverallCond, MasVnrArea, BsmtFinSF2, LowQualFinSF, WoodDeckSF, OpenPorchSF, EnclosedPorch, X3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold, Neighborhood, Condition2, BldgType, HouseStyle, BsmtFinType1, GarageQual, LandContour, RoofStyle, SaleType, KitchenAbvGr, OverallQual, TotalBsmtSF, GrLivArea, BsmtFinSF1

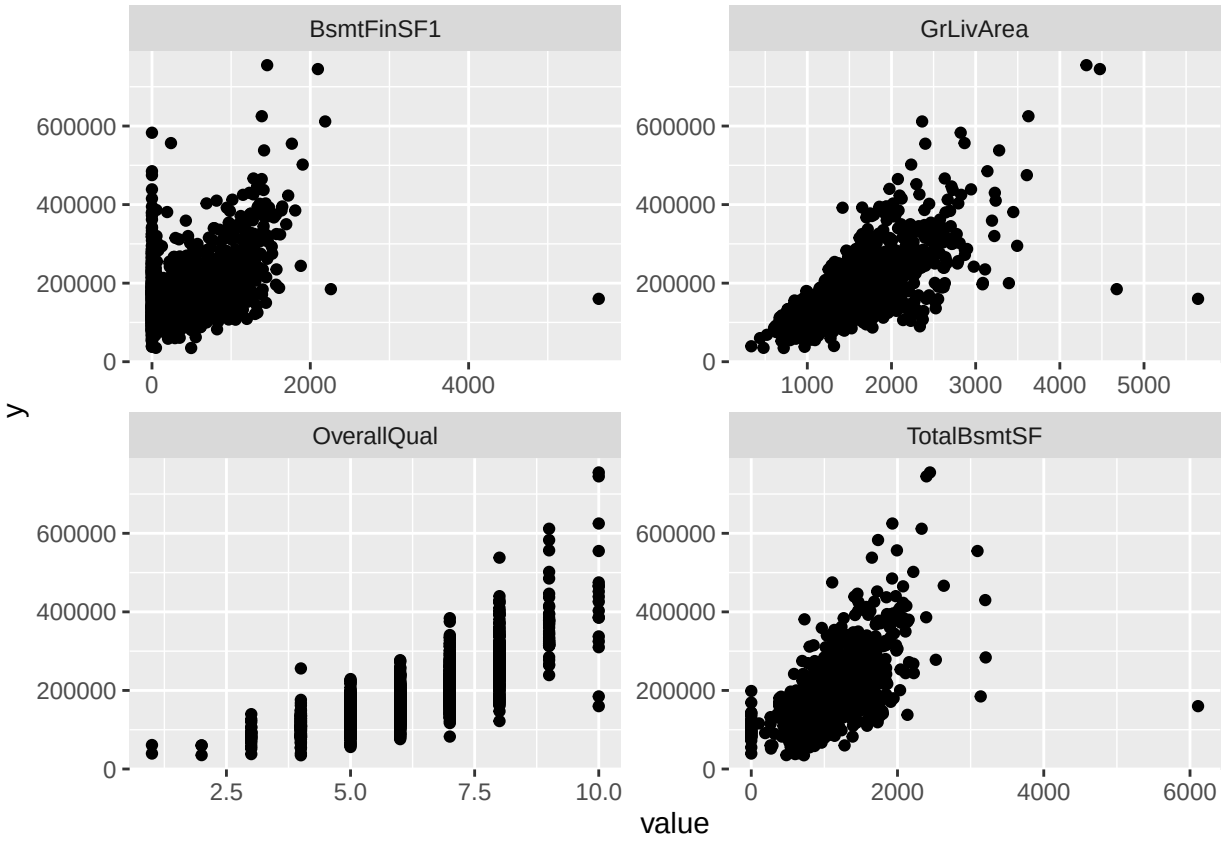
Quantidade de numéricas: 20

Quantidade de categóricas: 21

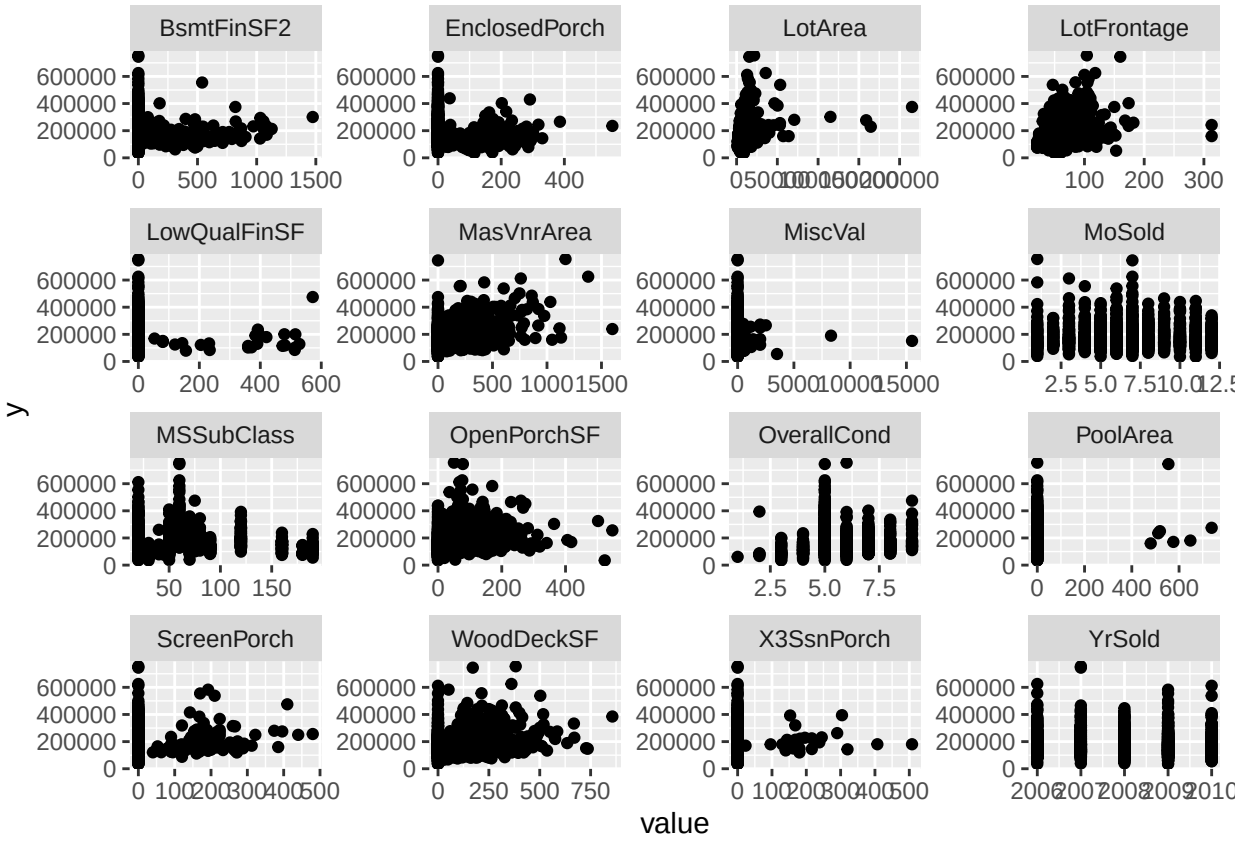
Quantidade total: 41

Quantidade anterior: 76

Análise das variáveis numéricas pela variável resposta




```
## Warning: Removed 267 rows containing missing values (geom_point).
```



Coeficiente de correlação

Resultado do p-valor

```
##   OverallQual   TotalBsmtSF     GrLivArea    BsmtFinSF1    MSSubClass
##       0.00000       0.00000       0.00000       0.00000       0.00127
##   LotFrontage       LotArea   OverallCond    MasVnrArea    BsmtFinSF2
##       0.00000       0.00000       0.00291       0.00000       0.66400
##  LowQualFinSF    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##       0.32821       0.00000       0.00000       0.00000       0.08858
##   ScreenPorch      PoolArea       MiscVal        MoSold        YrSold
##       0.00002       0.00041       0.41849       0.07613       0.26941
```

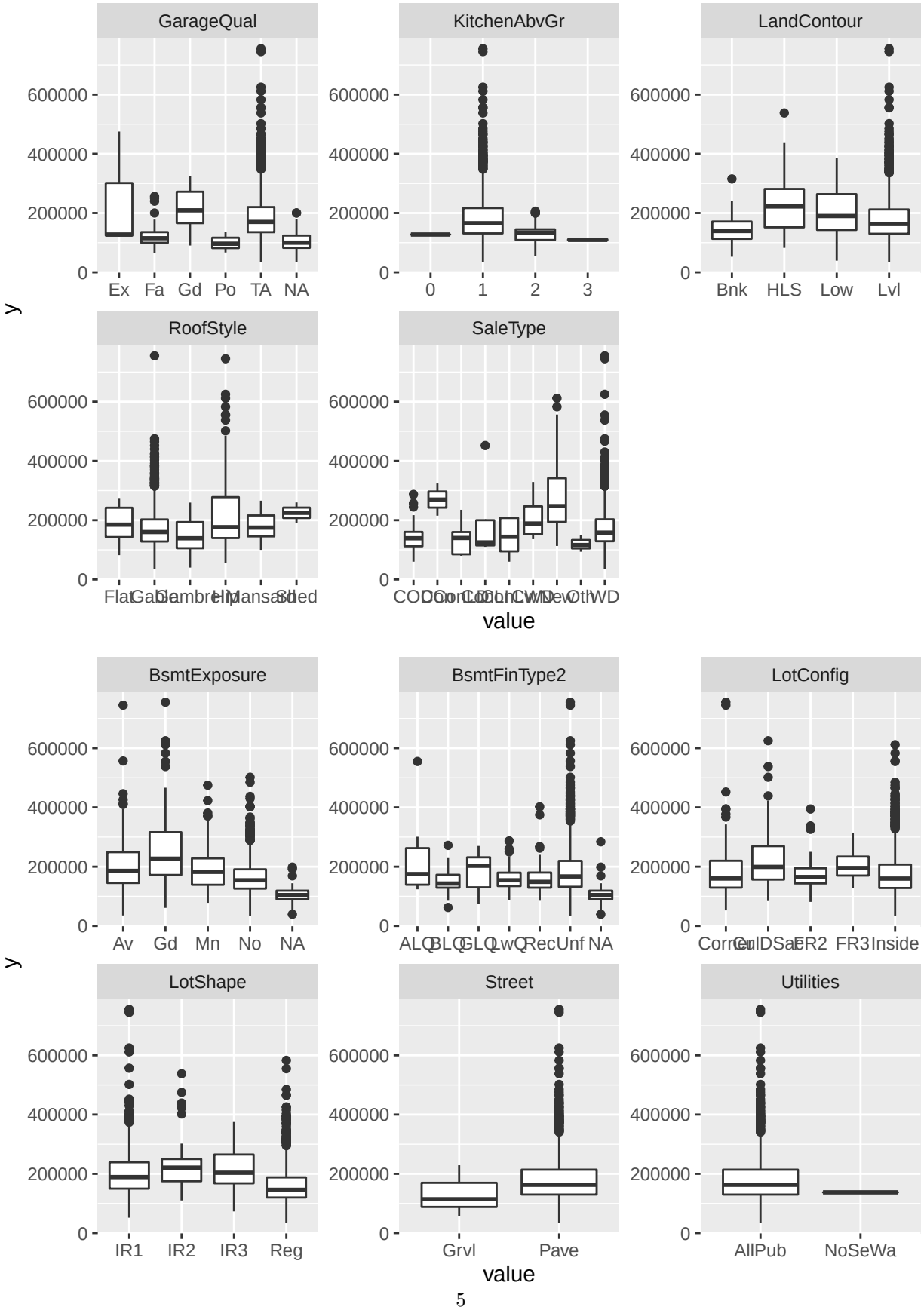
Variáveis filtradas pelo p-valor(0,05)

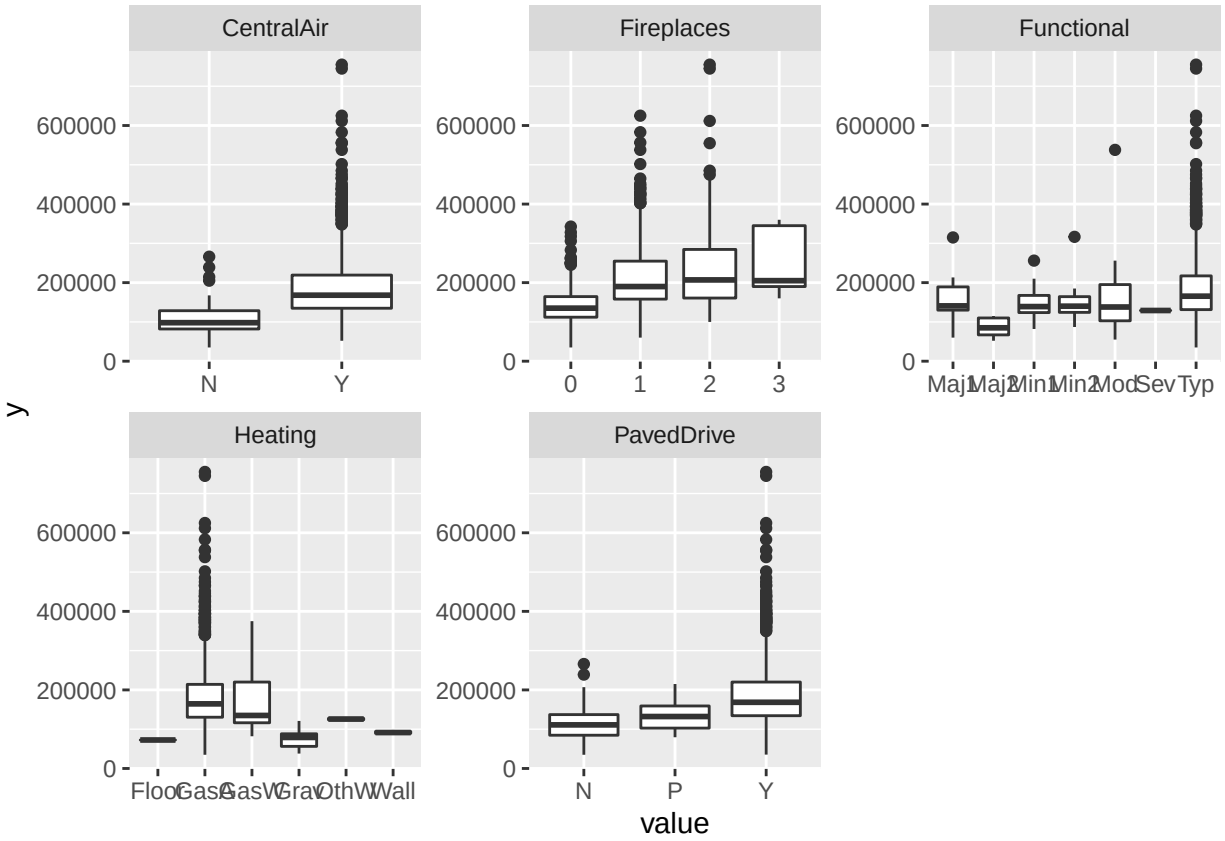
```
##  [1] "OverallQual"   "TotalBsmtSF"   "GrLivArea"     "BsmtFinSF1"
##  [5] "MSSubClass"    "LotFrontage"   "LotArea"       "OverallCond"
##  [9] "MasVnrArea"    "WoodDeckSF"    "OpenPorchSF"   "EnclosedPorch"
## [13] "ScreenPorch"   "PoolArea"
```

Variáveis quem tem uma correlação absoluta maior que 0.5 com a variável resposta

```
## OverallQual TotalBsmtSF   GrLivArea
##     0.79098     0.61358     0.70862
```

Análise das variáveis categóricas pela variável resposta





Resultado Anova

```
##                Df        Sum Sq      Mean Sq F value                Pr(>F)
## Neighborhood   24 4434287052729 184761960530 107.358 < 0.0000000000000002 ***
## Condition2      7   54057554379   7722507768   4.487  0.00005999920284937 ***
## BldgType        4  331972175205  82993043801  48.224 < 0.0000000000000002 ***
## HouseStyle      7  145640915208  20805845030  12.089  0.00000000000000527 ***
## BsmtFinType1    5  233122198135  46624439627  27.092 < 0.0000000000000002 ***
## GarageQual      4   49566836666  12391709167   7.200  0.00000989107111308 ***
## LandContour     3   25844222369   8614740790   5.006              0.001874 **
## RoofStyle       5  161428153213  32285630643  18.760 < 0.0000000000000002 ***
## SaleType        8  129349209290  16168651161   9.395  0.00000000000115245 ***
## KitchenAbvGr    1    1175985963   1175985963   0.683              0.408605
## Street          1    1726108435   1726108435   1.003              0.316787
## LotShape        3   49188988800  16396329600   9.527  0.00000318669834792 ***
## Utilities       1    8020109814   8020109814   4.660              0.031061 *
## LotConfig       4   38538417677   9634604419   5.598              0.000182 ***
## BsmtExposure    3  287037945553  95679315184  55.595 < 0.0000000000000002 ***
## BsmtFinType2    5   12473317264   2494663453   1.450              0.203711
## Heating         3    8886911646   2962303882   1.721              0.160777
## CentralAir      1   39241001246  39241001246  22.801  0.00000200891223276 ***
## Functional      6   30269723263   5044953877   2.931              0.007630 **
## PavedDrive      2    7172809952   3586404976   2.084              0.124875
## Fireplaces      1  211073105168 211073105168 122.646 < 0.0000000000000002 ***
## Residuals    1248 2147803343660   1720996269
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 113 observations deleted due to missingness
```

Variáveis filtradas pelo p-valor(0,05)

```
##  [1] "Neighborhood" "Condition2"   "BldgType"     "HouseStyle"   "BsmtFinType1"
##  [6] "GarageQual"   "LandContour"  "RoofStyle"    "SaleType"     "LotShape"
## [11] "Utilities"    "LotConfig"    "BsmtExposure" "CentralAir"   "Functional"
## [16] "Fireplaces"   ""
```

Resultado da extração de variáveis

Dado a relação das variáveis explicativas pela variável resposta

```
##  [1] "OverallQual"  "TotalBsmtSF"  "GrLivArea"    "Neighborhood" "Condition2"
##  [6] "BldgType"     "HouseStyle"   "BsmtFinType1" "GarageQual"   "LandContour"
## [11] "RoofStyle"    "SaleType"     "LotShape"     "Utilities"    "LotConfig"
## [16] "BsmtExposure" "CentralAir"   "Functional"   "Fireplaces"
```

Número de variáveis anteriormente: 76

Número de variáveis agora: 19

tirar dúvidas:

- qual a melhor medida para se usar quando eu tenho uma variável contínua e outra categórica? já que o banco de dados é grande
- como tratar variáveis que tem muitas categorias? como é o caso do neighborhood