MAP 535 - Data Analysis Project

Adrien Toulouse & Paul-Antoine Girard

Introduction

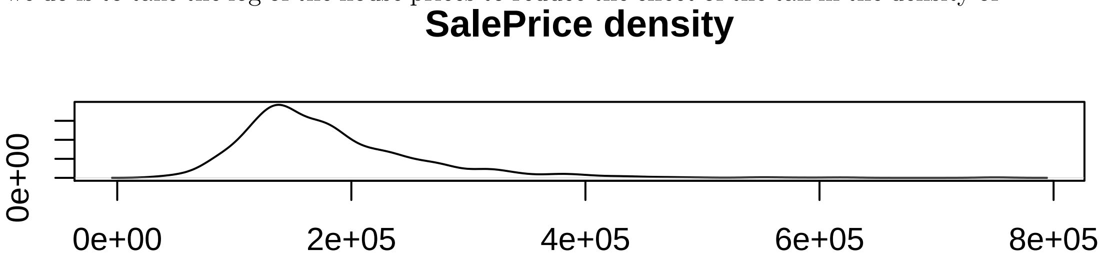
Our task is to analyse a dataset, named **House Prices: Advanced Regression Techniques**. It contains our response variable, the sale price of about 1500 residential homes located in Ames, Iowa, along with 79 explanatory variables describing (almost) every aspect of the houses. The dataset has already been preprocessed to deal with missing values, so we will work on a reduced dataset containing 68 variables.

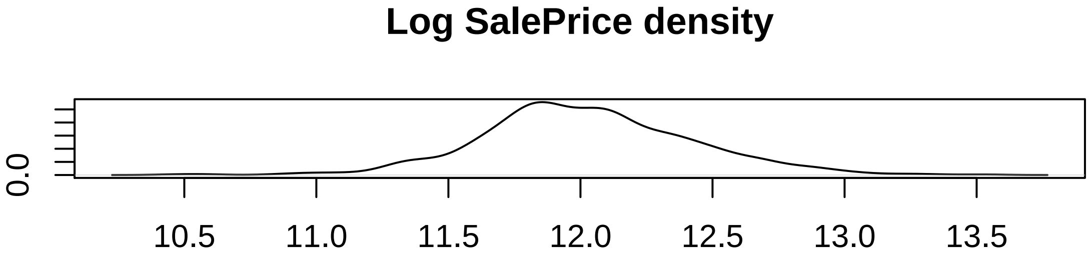
Our work will be focused on finding the best linear prediction model using a minimum number of variables. We can therefore state our research hypothesis as follows: **Can we construct a performant linear regression model by selecting only the 10 most appropriate variables? How does it compare to larger models?**

To test our hypothesis, we will start by describing the data and apply descriptive statistics to better apprehend it and preprocess the variables if necessary. We will then try to select the most important variables and after checking if the linear model assumptions are verified, we will build multiple linear regression models and compare them.

Exploratory Data Analysis

A first transformation we do is to take the log of the house prices to reduce the effect of the tail in the density of our response variable.

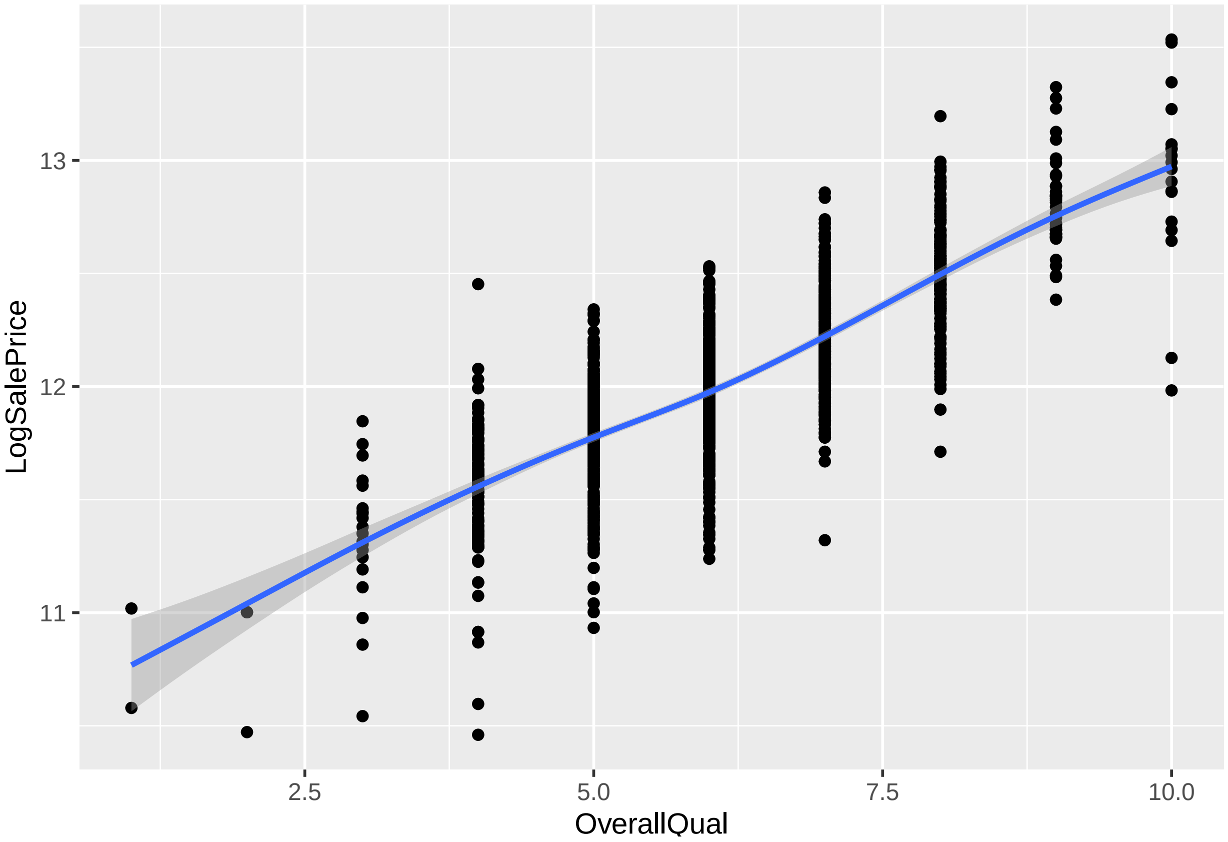




By ploting scatter plots, we find that some variables are strongly correlated with the Sale Price. Such variables are interesting to keep in a linear model as they explain well the price.

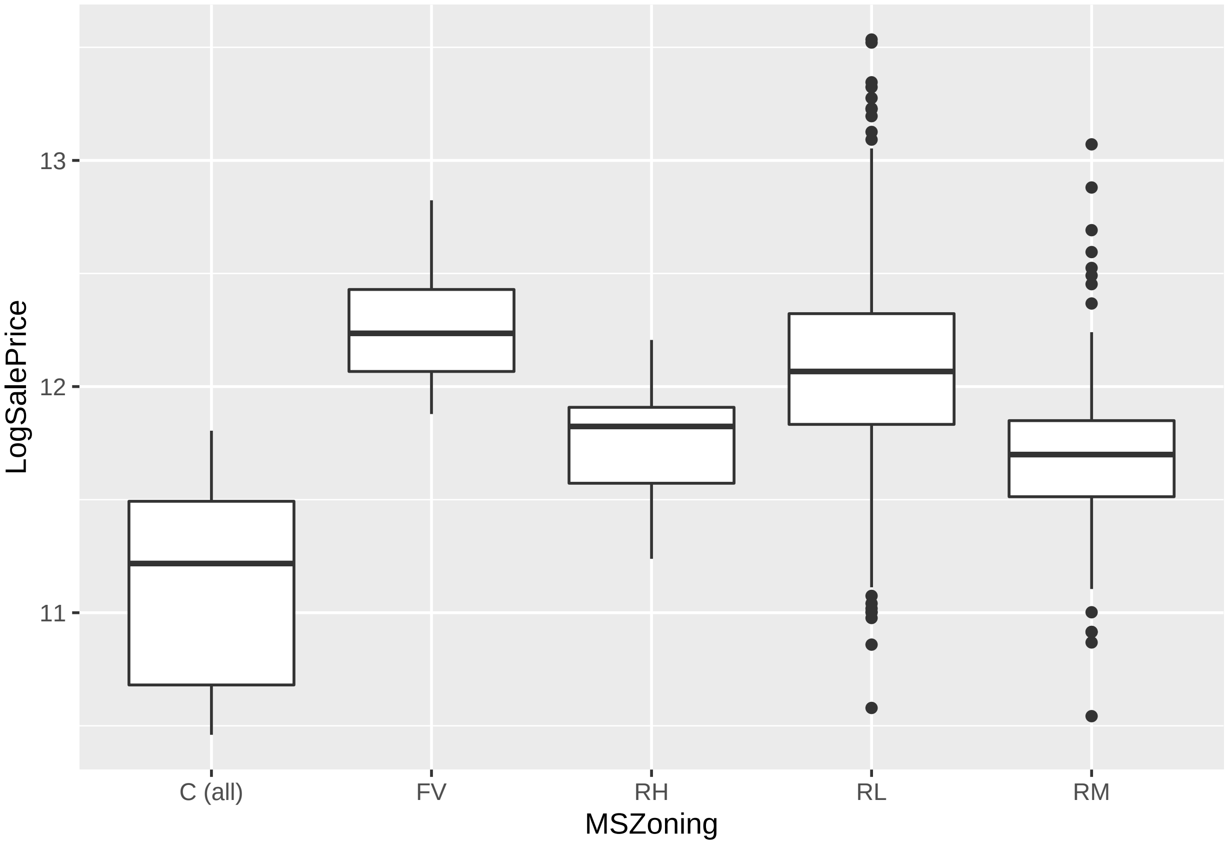
```
ggplot(trainImputed, aes(OverallQual, LogSalePrice)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



We can also plot boxplots for categorical variables. We are looking for variables that have different boxplots in each category when considering the Sale Price as this will indicate a clear dependency between the two variables.

```
ggplot(data = trainImputed) +
  geom_boxplot(aes(y=LogSalePrice, x = MSZoning))
```



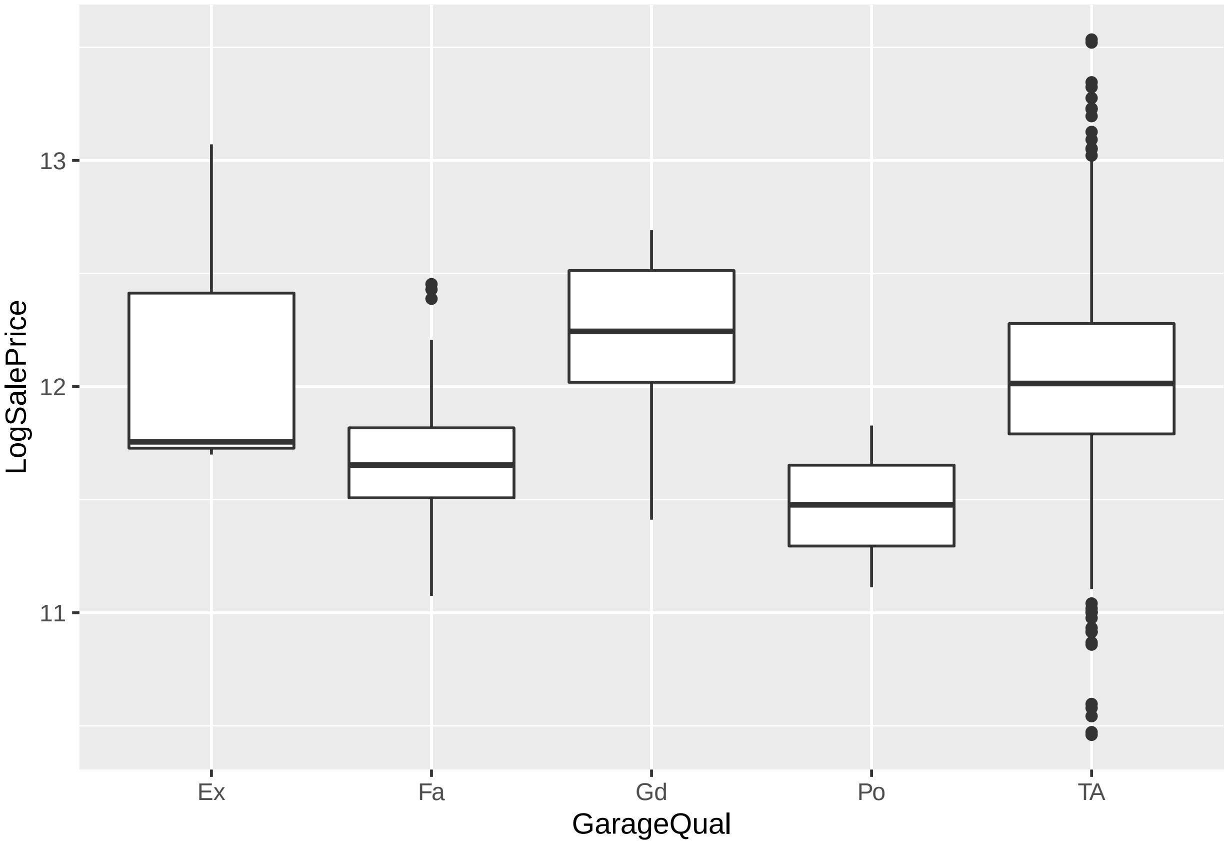
This first analysis visually indicates clearly that MSZoning is an important variable to explain the Sale Price.

```
res.aov <- aov(LogSalePrice ~ MSZoning, data = trainImputed)
summary(res.aov)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)    
## MSZoning        4  40.94  10.234   77.61 <2e-16 ***
## Residuals    1455 191.87   0.132                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is confirmed by the one way Anova test as the p-value is less than the significance level 0.05. We can conclude that there are significant differences between the MSZoning caregories when considering the Sale Price and make us say that we should include MSZoning in our model.

```
ggplot(data = trainImputed) +
  geom_boxplot(aes(y=LogSalePrice, x = GarageQual))
```



```
res.aov <- aov(LogSalePrice ~ GarageQual, data = trainImputed)
summary(res.aov)
```

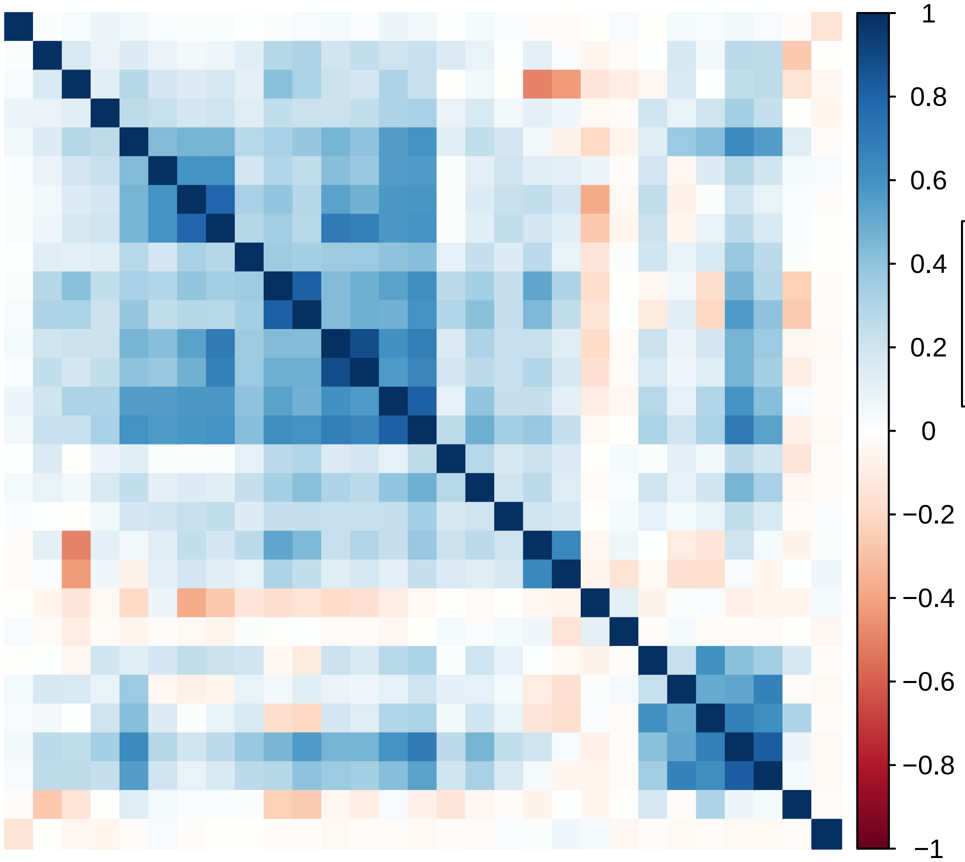
```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## GarageQual     4   7.54  1.8848   12.17 9.75e-10 ***
## Residuals   1455 225.26  0.1548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

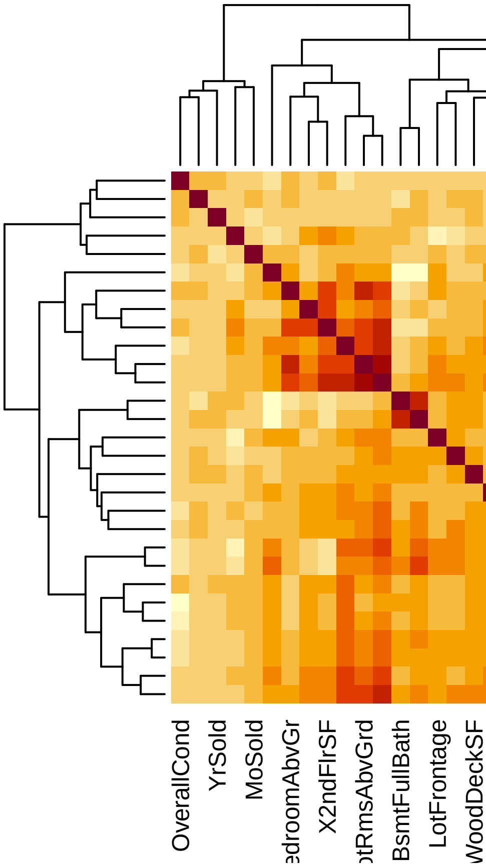
We find similar results when looking at the GarageQual variable. However, the boxplots do not completely follow intuition. Indeed the boxplots indicate that on average houses with garages in excellent quality have a lower sale price that garages in good quality. This seems to indicate that garage quality is not a key variable when trying to explain the sale price of a house.

To select regressors, it is important to also analyze the correlation between numerical variables and to create clusters of variables based on correlations. Indeed we know that having correlated response variables is not efficient in linear models. This will help us know which variables contain redundant information and which variables to keep.

```
var.numeric <- colnames(trainImputed)[sapply(trainImputed, is.numeric)]

trainImputed %>%
  select(var.numeric) %>%
  cor() %>%
  corrplot(method = 'color', order = "hclust", tl.pos = 'n') %>%
  heatmap (symm=T)
```





```
## Don't know how to print only the second graph
```

Now that we have more information on our data, let's try to build multiple linear models. We will use what we found in this exploratory analysis part to select variables and build these different models.

Modeling and Diagnostics

In this part, we are going to build different linear regression models and analyse their differences to select the one that we think fits the best our research hypothesis.

First, we start by doing a linear regression with all the variables of the dataset.

```
price_lm = lm(LogSalePrice ~ ., data = trainPP)
summary(price_lm)
```

```
##
## Call:
## lm(formula = LogSalePrice ~ ., data = trainPP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68809 -0.04386  0.00091  0.05102  0.55334
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.4793140  0.2673474  35.457  < 2e-16 ***
## MSSubClass             -0.0001861  0.0003715  -0.501 0.616456
## MSZoningFV              0.4526922  0.0535759   8.450  < 2e-16 ***
## MSZoningRH           0.4226000  0.0533679   7.919 5.31e-15 ***
## MSZoningRL           0.4025102  0.0456157   8.824  < 2e-16 ***
## MSZoningRM           0.3729896  0.0427194   8.731  < 2e-16 ***
## LotFrontage          0.0008833  0.0055650   0.159 0.873906    
## LotArea              0.0489112  0.0063277   7.730 2.22e-14 ***
## StreetPave           0.0767609  0.0528265   1.453 0.146456    
## LotShapeIR2          0.0126210  0.0190877   0.661 0.508601    
## LotShapeIR3          0.0353481  0.0390426   0.905 0.365444    
## LotShapeReg          0.0110053  0.0072078   1.527 0.127051    
## LandContourHLS       0.0243042  0.0229205   1.060 0.289184    
## LandContourLow      -0.0256108  0.0282328  -0.907 0.364515    
## LandContourLvl       0.0219546  0.0165077   1.330 0.183775    
## UtilitiesNoSeWa     -0.1430956  0.1171010  -1.222 0.221947    
## LotConfigCulDSac     0.0157507  0.0149397   1.054 0.291958    
## LotConfigFR2        -0.0544385  0.0180258  -3.020 0.002579 ** 
## LotConfigFR3        -0.1152056  0.0563323  -2.045 0.041056 *  
## LotConfigInside     -0.0172591  0.0080059  -2.156 0.031292 *  
## LandSlopeMod         0.0308997  0.0177563   1.740 0.082070 .  
## LandSlopeSev        -0.1087672  0.0431244  -2.522 0.011788 *  
## NeighborhoodBlueste -0.0245805  0.0857111  -0.287 0.774326    
## NeighborhoodBrDale  -0.0269889  0.0500587  -0.539 0.589883    
## NeighborhoodBrkSide  0.0041497  0.0431036   0.096 0.923319    
## NeighborhoodClearCr -0.0049221  0.0427146  -0.115 0.908281    
## NeighborhoodCollgCr -0.0393713  0.0341471  -1.153 0.249137    
## NeighborhoodCrawfor  0.0839128  0.0395629   2.121 0.034120 *  
## NeighborhoodEdwards -0.1055271  0.0371069  -2.844 0.004530 ** 
## NeighborhoodGilbert -0.0445283  0.0360396  -1.236 0.216865    
## NeighborhoodIDOTRR  -0.0573013  0.0492140  -1.164 0.244515    
## NeighborhoodMeadowV -0.1243754  0.0506757  -2.454 0.014252 *  
## NeighborhoodMitchel -0.0835179  0.0380199  -2.197 0.028228 *  
## NeighborhoodNAmes   -0.0642935  0.0363386  -1.769 0.077092 .  
## NeighborhoodNoRidge  0.0588854  0.0381866   1.542 0.123319    
## NeighborhoodNPkVill  0.0121784  0.0633095   0.192 0.847489    
## NeighborhoodNridgHt  0.0465634  0.0348998   1.334 0.182384    
## NeighborhoodNWAmes  -0.0750364  0.0376248  -1.994 0.046335 *  
## NeighborhoodOldTown -0.0674804  0.0439033  -1.537 0.124544    
## NeighborhoodSawyer  -0.0532408  0.0375885  -1.416 0.156907    
## NeighborhoodSawyerW -0.0371805  0.0362650  -1.025 0.305448    
## NeighborhoodSomerst -0.0103365  0.0411644  -0.251 0.801776    
## NeighborhoodStoneBr  0.1197527  0.0379097   3.159 0.001622 ** 
## NeighborhoodSWISU   -0.0090610  0.0443895  -0.204 0.838289    
## NeighborhoodTimber  -0.0326540  0.0380736  -0.858 0.391249    
## NeighborhoodVeenker -0.0016063  0.0483351  -0.033 0.973494    
## Condition1Feedr      0.0438623  0.0223752   1.960 0.050184 .  
## Condition1Norm       0.0919456  0.0185824   4.948 8.53e-07 ***
## Condition1PosA       0.0418415  0.0444308   0.942 0.346518    
## Condition1PosN       0.1058951  0.0330455   3.205 0.001387 ** 
## Condition1RRAe      -0.0557096  0.0407691  -1.366 0.172042    
## Condition1RRAn       0.0446338  0.0305283   1.462 0.143982    
## Condition1RRNe       0.0482936  0.0780764   0.619 0.536331    
## Condition1RRNn       0.1096593  0.0573166   1.913 0.055950 .  
## Condition2Feedr      0.1412271  0.1033490   1.367 0.172028    
## Condition2Norm       0.1018588  0.0893881   1.140 0.254710    
## Condition2PosA       0.2958184  0.1638619   1.805 0.071272 .  
## Condition2PosN       -0.6299007  0.1221136  -5.158 2.90e-07 ***
## Condition2RRAe       -0.1964957  0.1967216  -0.999 0.318062    
## Condition2RRAn       -0.0225677  0.1409978  -0.160 0.872862    
## Condition2RRNn        0.1509159  0.1208199   1.249 0.211866    
## BldgType2fmCon       -0.0162692  0.0552960  -0.294 0.768640    
## BldgTypeDuplex       -0.0718467  0.0275282  -2.610 0.009165 ** 
## BldgTypeTwnhs        -0.0117428  0.0448850  -0.262 0.793659    
## BldgTypeTwnhsE        0.0074103  0.0402828   0.184 0.854077    
## HouseStyle1.5Unf      0.0703525  0.0410542   1.714 0.086844 .  
## HouseStyle1Story      0.0171870  0.0268421   0.640 0.522098    
## HouseStyle2.5Fin     -0.0723367  0.0474322  -1.525 0.127501    
## HouseStyle2.5Unf      0.0093626  0.0408196   0.229 0.818624    
## HouseStyle2Story     -0.0113909  0.0151707  -0.751 0.452887    
## HouseStyleSFoyer      0.0423097  0.0338245   1.251 0.211222    
## HouseStyleSLvl        0.0207094  0.0279937   0.740 0.459570    
## OverallQual           0.0522045  0.0062060   8.412  < 2e-16 ***
## OverallCond           0.0416889  0.0042835   9.732  < 2e-16 ***
## YearBuilt             0.0503115  0.0102513   4.908 1.04e-06 ***
## YearRemodAdd          0.0184994  0.0050632   3.654 0.000269 ***
## RoofStyleGable        0.0035441  0.0821375   0.043 0.965590    
## RoofStyleGambrel     -0.0223886  0.0895496  -0.250 0.802619    
## RoofStyleHip          0.0058891  0.0823858   0.071 0.943026    
## RoofStyleMansard      0.0515132  0.0955092   0.539 0.589740    
## RoofStyleShed         0.2425767  0.1535285   1.580 0.114360    
## RoofMatlCompShg       1.8656676  0.1336128  13.963  < 2e-16 ***
## RoofMatlMembran       2.1125993  0.2011721  10.501  < 2e-16 ***
## RoofMatlMetal         2.0333440  0.1954852  10.402  < 2e-16 ***
## RoofMatlRoll          1.8673307  0.1737182  10.749  < 2e-16 ***
## RoofMatlTar&Grv       1.8963447  0.1587917  11.942  < 2e-16 ***
## RoofMatlWdShake       1.8472738  0.1497425  12.336  < 2e-16 ***
## RoofMatlWdShngl       1.9540877  0.1414543  13.814  < 2e-16 ***
## Exterior1stAsphShn   -0.0332386  0.1480146  -0.225 0.822356    
## Exterior1stBrkComm   -0.2337965  0.1240439  -1.885 0.059693 .  
## Exterior1stBrkFace    0.0511861  0.0565454   0.905 0.365525    
## Exterior1stCBlock    -0.0877420  0.1217986  -0.720 0.471424    
## Exterior1stCemntBd   -0.0712749  0.0849345  -0.839 0.401533    
## Exterior1stHdBoard   -0.0461894  0.0573227  -0.806 0.420526    
## Exterior1stImStucc   -0.1403510  0.1228985  -1.142 0.253672    
## Exterior1stMetalSd    0.0067508  0.0649358   0.104 0.917218    
## Exterior1stPlywood   -0.0417231  0.0564859  -0.739 0.460262    
## Exterior1stStone     -0.0355814  0.1063063  -0.335 0.737904    
## Exterior1stStucco    -0.0038388  0.0624662  -0.061 0.951008    
## Exterior1stVinylSd   -0.0490367  0.0589590  -0.832 0.405734    
## Exterior1stWd Sdng   -0.0731175  0.0548011  -1.334 0.182373    
## Exterior1stWdShing   -0.0332531  0.0589695  -0.564 0.572921    
## Exterior2ndAsphShn    0.0348981  0.0994882   0.351 0.725816    
## Exterior2ndBrk Cmn    0.0400107  0.0901047   0.444 0.657086    
## Exterior2ndBrkFace    0.0046903  0.0585509   0.080 0.936165    
## Exterior2ndCBlock            NA         NA      NA       NA    
## Exterior2ndCmentBd    0.0952132  0.0835622   1.139 0.254745    
## Exterior2ndHdBoard    0.0365577  0.0552897   0.661 0.508606    
## Exterior2ndImStucc    0.0831774  0.0634955   1.310 0.190448    
## Exterior2ndMetalSd    0.0121037  0.0633527   0.191 0.848515    
## Exterior2ndOther     -0.0289568  0.1216635  -0.238 0.811915    
## Exterior2ndPlywood    0.0368938  0.0536398   0.688 0.491704
## Exterior2ndStone     -0.0182731  0.0757093  -0.241 0.809317
## Exterior2ndStucco     0.0456493  0.0604769   0.755 0.450500
## Exterior2ndVinylSd    0.0634491  0.0572628   1.108 0.268063
## Exterior2ndWd Sdng    0.0707056  0.0531755   1.330 0.183874
## Exterior2ndWd Shng    0.0340150  0.0554639   0.613 0.539803
## MasVnrTypeBrkFace     0.0394097  0.0302228   1.304 0.192486
## MasVnrTypeNone        0.0013344  0.0440521   0.030 0.975839
## MasVnrTypeStone       0.0534869  0.0319356   1.675 0.094219 .
## MasVnrArea           -0.0157376  0.0165057  -0.953 0.340542
## ExterQualFa           0.0344387  0.0488905   0.704 0.481313
## ExterQualGd          -0.0189693  0.0212684  -0.892 0.372620
## ExterQualTA          -0.0157401  0.0236101  -0.667 0.505110
## ExterCondFa          -0.1249827  0.0809343  -1.544 0.122784
## ExterCondGd          -0.1073552  0.0773628  -1.388 0.165483
## ExterCondPo          -0.1067362  0.1417264  -0.753 0.451525
## ExterCondTA          -0.0927265  0.0771494  -1.202 0.229629
## FoundationCBlock      0.0305762  0.0141762   2.157 0.031208 *
## FoundationPConc       0.0376012  0.0153055   2.457 0.014158 *
## FoundationSlab        0.0261223  0.0360560   0.724 0.468901
## FoundationStone       0.1130169  0.0491217   2.301 0.021571 *
## FoundationWood       -0.1073348  0.0651562  -1.647 0.099741 .
## BsmtQualFa           -0.0395926  0.0280977  -1.409 0.159055
## BsmtQualGd           -0.0444706  0.0146726  -3.031 0.002489 **
## BsmtQualTA           -0.0522682  0.0182640  -2.862 0.004283 **
## BsmtCondGd            0.0150156  0.0236313   0.635 0.525277
## BsmtCondPo            0.2564510  0.1317258   1.947 0.051778 .
## BsmtCondTA            0.0224583  0.0189972   1.182 0.237357
## BsmtExposureGd        0.0354376  0.0133365   2.657 0.007981 **
## BsmtExposureMn       -0.0062465  0.0136124  -0.459 0.646400
## BsmtExposureNo       -0.0146021  0.0099376  -1.469 0.141985
## BsmtFinType1BLQ      -0.0096525  0.0123218  -0.783 0.433559
## BsmtFinType1GLQ       0.0097934  0.0112555   0.870 0.384416
## BsmtFinType1LwQ      -0.0348869  0.0166775  -2.092 0.036655 *
## BsmtFinType1Rec      -0.0152538  0.0133032  -1.147 0.251760
## BsmtFinType1Unf      -0.0036978  0.0234284  -0.158 0.874614
## BsmtFinSF1            0.0155088  0.0137979   1.124 0.261234
## BsmtFinType2BLQ      -0.0621791  0.0330824  -1.880 0.060408 .
## BsmtFinType2GLQ       0.0019766  0.0416895   0.047 0.962192
## BsmtFinType2LwQ      -0.0409004  0.0319671  -1.279 0.200978
## BsmtFinType2Rec      -0.0386072  0.0313209  -1.233 0.217947
## BsmtFinType2Unf      -0.0270602  0.0293032  -0.923 0.355950
## BsmtUnfSF            -0.0201018  0.0077869  -2.581 0.009952 **
## TotalBsmtSF           0.0517243  0.0094569   5.469 5.46e-08 ***
## HeatingGasA           0.0393773  0.1111540   0.354 0.723205
## HeatingGasW           0.1153364  0.1148193   1.005 0.315333
## HeatingGrav          -0.1092006  0.1218449  -0.896 0.370307
## HeatingOthW           0.0601762  0.1374512   0.438 0.661608
## HeatingWall           0.1007952  0.1287508   0.783 0.433853
## HeatingQCFa          -0.0203172  0.0206942  -0.982 0.326399
## HeatingQCGd          -0.0207670  0.0092107  -2.255 0.024329 *
## HeatingQCPo          -0.0298812  0.1192983  -0.250 0.802262
## HeatingQCTA          -0.0338199  0.0092023  -3.675 0.000248 ***
## CentralAirY           0.0575063  0.0173919   3.306 0.000972 ***
## ElectricalFuseF       0.0056348  0.0257383   0.219 0.826743    
## ElectricalFuseP      -0.0843031  0.0828254  -1.018 0.308953    
## ElectricalMix        -0.1754861  0.1978121  -0.887 0.375179    
## ElectricalSBrkr      -0.0146923  0.0132104  -1.112 0.266276    
## X1stFlrSF             0.0055497  0.0125758   0.441 0.659072    
## X2ndFlrSF             0.0115132  0.0151348   0.761 0.446973    
## GrLivArea             0.1312134  0.0157687   8.321 2.28e-16 ***
## BsmtFullBath          0.0089077  0.0044877   1.985 0.047373 *  
## BsmtHalfBath          0.0009091  0.0032186   0.282 0.777652    
## FullBath              0.0098929  0.0053360   1.854 0.063979 .  
## HalfBath              0.0106476  0.0047936   2.221 0.026517 *  
## BedroomAbvGr          0.0025636  0.0049548   0.517 0.604968    
## KitchenQualFa        -0.0698525  0.0276978  -2.522 0.011796 *  
## KitchenQualGd        -0.0736513  0.0154615  -4.764 2.13e-06 ***
## KitchenQualTA        -0.0756274  0.0173776  -4.352 1.46e-05 ***
## TotRmsAbvGrd         -0.0010948  0.0066514  -0.165 0.869288    
## FunctionalMaj2       -0.2252490  0.0644472  -3.495 0.000491 ***
## FunctionalMin1        0.0226186  0.0381157   0.593 0.553009    
## FunctionalMin2        0.0172495  0.0381444   0.452 0.651192    
## FunctionalMod        -0.0474732  0.0460992  -1.030 0.303303    
## FunctionalSev        -0.2924472  0.1291855  -2.264 0.023761 *  
## FunctionalTyp         0.0786944  0.0330331   2.382 0.017356 *  
## Fireplaces            0.0137099  0.0038787   3.535 0.000423 ***
## GarageTypeAttchd      0.1076317  0.0487099   2.210 0.027312 *  
## GarageTypeBasment     0.0963315  0.0562625   1.712 0.087115 .  
## GarageTypeBuiltIn     0.1056783  0.0504013   2.097 0.036220 *  
## GarageTypeCarPort     0.1141287  0.0650800   1.754 0.079736 .  
## GarageTypeDetchd      0.1210922  0.0484907   2.497 0.012646 *  
## GarageYrBlt          -0.0094430  0.0068048  -1.388 0.165475    
## GarageFinishRFn      -0.0059494  0.0087421  -0.681 0.496287    
## GarageFinishUnf      -0.0087645  0.0107081  -0.818 0.413233    
## GarageCars            0.0172577  0.0073839   2.337 0.019587 *  
## GarageArea            0.0224287  0.0076751   2.922 0.003538 ** 
## GarageQualFa         -0.5153833  0.1302044  -3.958 7.98e-05 ***
## GarageQualGd         -0.4408788  0.1328944  -3.318 0.000935 ***
## GarageQualPo         -0.5252749  0.1687697  -3.112 0.001898 ** 
## GarageQualTA         -0.4789013  0.1284619  -3.728 0.000202 ***
## GarageCondFa          0.3676978  0.1508577   2.437 0.014934 *  
## GarageCondGd          0.4322617  0.1557497   2.775 0.005597 ** 
## GarageCondPo          0.4617344  0.1624901   2.842 0.004562 ** 
## GarageCondTA          0.4056741  0.1491294   2.720 0.006614 ** 
## PavedDriveP          -0.0182488  0.0244559  -0.746 0.455693    
## PavedDriveY           0.0118909  0.0152851   0.778 0.436752    
## WoodDeckSF            0.0065961  0.0033204   1.987 0.047193 *  
## OpenPorchSF           0.0017742  0.0035948   0.494 0.621715    
## MoSold               -0.0022753  0.0029470  -0.772 0.440222    
## YrSold               -0.0039091  0.0030406  -1.286 0.198816    
## SaleTypeCon           0.0740254  0.0790112   0.937 0.348994    
## SaleTypeConLD         0.1267997  0.0433014   2.928 0.003471 ** 
## SaleTypeConLI        -0.0155355  0.0514124  -0.302 0.762570    
## SaleTypeConLw         0.0064634  0.0544615   0.119 0.905549    
## SaleTypeCWD           0.0368273  0.0577148   0.638 0.523532    
## SaleTypeNew           0.1140157  0.0687657   1.658 0.097565 .  
## SaleTypeOth           0.0732226  0.0644938   1.135 0.256452    
## SaleTypeWD           -0.0135076  0.0187226  -0.721 0.470762
## SaleConditionAdjLand  0.1294335  0.0639010   2.026 0.043027 *
## SaleConditionAlloca   0.0674897  0.0377495   1.788 0.074048 .
## SaleConditionFamily  -0.0059943  0.0271633  -0.221 0.825380
## SaleConditionNormal   0.0583650  0.0128586   4.539 6.20e-06 ***
## SaleConditionPartial -0.0270972  0.0661918  -0.409 0.682336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.102 on 1236 degrees of freedom
## Multiple R-squared:  0.9447, Adjusted R-squared:  0.9347
## F-statistic: 94.71 on 223 and 1236 DF,  p-value: < 2.2e-16
```

The results of this model show us that some of the variables are irrelevant if we consider the p-value of the Student test statistic realized for each variable. At a significance level of 0.001, this technique suggests us to only keep the following variables: MSZoningFV, MSZoningRH, MSZoningRL, MSZoningRM, LotArea, Condition1Norm, Condition2PosN, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofMatlCompShg, RoofMatlMembran, RoofMatlMetal, RoofMatlRoll, RoofMatlTar&Grv, RoofMatlWdShake, RoofMatlWdShngl, BldgTypeDuplex, KitchenQualGd, KitchenQualTA, FunctionalMaj2, Fireplaces, GarageQualGd, GarageQualFa.

We then use a method based on minimizing the AIC to automatically select a reduced number of variables for our model.

```r
select.variables.backward = step(price_lm,scope= ~1,direction="backward",trace=FALSE)
summary(select.variables.backward)
```

```
##
## Call:
## lm(formula = LogSalePrice ~ MSZoning + LotArea + Street + LotConfig +
##     LandSlope + Neighborhood + Condition1 + Condition2 + BldgType +
##     OverallQual + OverallCond + YearBuilt + YearRemodAdd + RoofMatl +
##     Exterior1st + Foundation + BsmtQual + BsmtExposure + BsmtFinSF1 +
##     BsmtUnfSF + TotalBsmtSF + Heating + HeatingQC + CentralAir +
##     X1stFlrSF + GrLivArea + BsmtFullBath + FullBath + HalfBath +
##     KitchenQual + Functional + Fireplaces + GarageYrBlt + GarageCars +
##     GarageArea + GarageQual + GarageCond + WoodDeckSF + SaleType +
##     SaleCondition, data = trainPP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69217 -0.04441  0.00199  0.05184  0.56486
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.624657   0.216082  44.542  < 2e-16 ***
## MSZoningFV             0.431602   0.051224   8.426  < 2e-16 ***
## MSZoningRH             0.407910   0.051116   7.980 3.16e-15 ***
## MSZoningRL             0.394814   0.043478   9.081  < 2e-16 ***
## MSZoningRM             0.366275   0.040626   9.016  < 2e-16 ***
## LotArea                0.045087   0.005423   8.315 2.27e-16 ***
## StreetPave             0.075663   0.047792   1.583 0.113621
## LotConfigCulDSac       0.007124   0.013520   0.527 0.598306
## LotConfigFR2          -0.051007   0.017286  -2.951 0.003225 **
## LotConfigFR3          -0.101803   0.055488  -1.835 0.066777 .
## LotConfigInside       -0.016990   0.007527  -2.257 0.024162 *
## LandSlopeMod          0.016047   0.014970   1.072 0.283950    
## LandSlopeSev         -0.084716   0.037469  -2.261 0.023924 *  
## NeighborhoodBlueste  -0.044172   0.081767  -0.540 0.589132    
## NeighborhoodBrDale   -0.027848   0.046734  -0.596 0.551356    
## NeighborhoodBrkSide   0.015134   0.039012   0.388 0.698137    
## NeighborhoodClearCr  -0.016458   0.039033  -0.422 0.673346    
## NeighborhoodCollgCr  -0.046948   0.031076  -1.511 0.131093    
## NeighborhoodCrawfor   0.078255   0.035937   2.178 0.029615 *  
## NeighborhoodEdwards  -0.096386   0.033867  -2.846 0.004495 ** 
## NeighborhoodGilbert  -0.053563   0.033063  -1.620 0.105469    
## NeighborhoodIDOTRR   -0.047432   0.045342  -1.046 0.295711    
## NeighborhoodMeadowV  -0.134498   0.046850  -2.871 0.004160 ** 
## NeighborhoodMitchel  -0.090673   0.034774  -2.607 0.009224 ** 
## NeighborhoodNAmes    -0.059208   0.033066  -1.791 0.073586 .  
## NeighborhoodNoRidge   0.055500   0.034932   1.589 0.112347    
## NeighborhoodNPkVill  -0.000660   0.046842  -0.014 0.988760    
## NeighborhoodNridgHt   0.040928   0.031259   1.309 0.190651    
## NeighborhoodNWAmes   -0.070988   0.034288  -2.070 0.038612 *  
## NeighborhoodOldTown  -0.058886   0.040201  -1.465 0.143218    
## NeighborhoodSawyer   -0.054358   0.034648  -1.569 0.116917    
## NeighborhoodSawyerW  -0.043977   0.033554  -1.311 0.190202    
## NeighborhoodSomerst  -0.006645   0.037810  -0.176 0.860520    
## NeighborhoodStoneBr   0.097474   0.034888   2.794 0.005283 ** 
## NeighborhoodSWISU    -0.018429   0.040098  -0.460 0.645880    
## NeighborhoodTimber   -0.051120   0.034868  -1.466 0.142862    
## NeighborhoodVeenker  -0.025772   0.045273  -0.569 0.569275    
## Condition1Feedr       0.050333   0.021309   2.362 0.018317 *  
## Condition1Norm        0.092418   0.017625   5.244 1.83e-07 ***
## Condition1PosA        0.050176   0.042435   1.182 0.237255    
## Condition1PosN        0.100490   0.031329   3.208 0.001371 ** 
## Condition1RRAe       -0.043770   0.039535  -1.107 0.268444    
## Condition1RRAn        0.053370   0.029128   1.832 0.067142 .  
## Condition1RRNe        0.036732   0.077114   0.476 0.633911    
## Condition1RRNn        0.110867   0.053995   2.053 0.040243 *  
## Condition2Feedr       0.098789   0.096372   1.025 0.305514    
## Condition2Norm        0.078202   0.082453   0.948 0.343083    
## Condition2PosA        0.426718   0.134601   3.170 0.001558 ** 
## Condition2PosN       -0.656386   0.115753  -5.671 1.75e-08 ***
## Condition2RRAe        0.006090   0.134558   0.045 0.963906    
## Condition2RRAn       -0.026403   0.135259  -0.195 0.845261    
## Condition2RRNn        0.084949   0.114096   0.745 0.456686    
## BldgType2fmCon       -0.049230   0.021984  -2.239 0.025301 *  
## BldgTypeDuplex       -0.075497   0.019396  -3.892 0.000104 ***
## BldgTypeTwnhs        -0.035860   0.024391  -1.470 0.141746    
## BldgTypeTwnhsE       -0.018482   0.015946  -1.159 0.246652    
## OverallQual           0.056467   0.005757   9.808  < 2e-16 ***
## OverallCond           0.042226   0.003834  11.012  < 2e-16 ***
## YearBuilt             0.060143   0.009274   6.485 1.25e-10 ***
## YearRemodAdd          0.016751   0.004714   3.553 0.000394 ***
## RoofMatlCompShg       1.806358   0.120522  14.988  < 2e-16 ***
## RoofMatlMembran       1.987869   0.166788  11.919  < 2e-16 ***
## RoofMatlMetal         1.900982   0.165011  11.520  < 2e-16 ***
## RoofMatlRoll          1.869122   0.162873  11.476  < 2e-16 ***
## RoofMatlTar&Grv       1.832440   0.126292  14.510  < 2e-16 ***
## RoofMatlWdShake       1.860148   0.131236  14.174  < 2e-16 ***
## RoofMatlWdShngl       1.885866   0.127488  14.793  < 2e-16 ***
## Exterior1stAsphShn    0.002666   0.110667   0.024 0.980786    
## Exterior1stBrkComm   -0.175463   0.084829  -2.068 0.038794 *  
## Exterior1stBrkFace    0.093364   0.030584   3.053 0.002313 ** 
## Exterior1stCBlock    -0.006139   0.109472  -0.056 0.955289    
## Exterior1stCemntBd    0.044695   0.031950   1.399 0.162078    
## Exterior1stHdBoard    0.006263   0.027818   0.225 0.821899    
## Exterior1stImStucc   -0.037567   0.107679  -0.349 0.727231    
## Exterior1stMetalSd    0.034463   0.027016   1.276 0.202317    
## Exterior1stPlywood    0.011667   0.029343   0.398 0.690986    
## Exterior1stStone     -0.015702   0.083583  -0.188 0.851013    
## Exterior1stStucco     0.048210   0.033943   1.420 0.155745    
## Exterior1stVinylSd    0.025844   0.027187   0.951 0.341986    
## Exterior1stWd Sdng    0.004791   0.026962   0.178 0.858989    
## Exterior1stWdShing    0.014092   0.033624   0.419 0.675198    
## FoundationCBlock      0.022097   0.013476   1.640 0.101297    
## FoundationPConc       0.034016   0.014776   2.302 0.021482 *  
## FoundationSlab        0.028639   0.033384   0.858 0.391129    
## FoundationStone       0.111615   0.045978   2.428 0.015334 *  
## FoundationWood       -0.115626   0.062636  -1.846 0.065119 .  
## BsmtQualFa           -0.050101   0.026438  -1.895 0.058311 .  
## BsmtQualGd           -0.056417   0.013754  -4.102 4.35e-05 ***
## BsmtQualTA           -0.071624   0.017127  -4.182 3.08e-05 ***
## BsmtExposureGd        0.040232   0.012748   3.156 0.001636 ** 
## BsmtExposureMn       -0.009462   0.012723  -0.744 0.457184    
## BsmtExposureNo       -0.018217   0.008927  -2.041 0.041473 *  
## BsmtFinSF1            0.018191   0.005448   3.339 0.000865 ***
## BsmtUnfSF            -0.016835   0.005880  -2.863 0.004261 ** 
## TotalBsmtSF           0.047466   0.007672   6.187 8.20e-10 ***
## HeatingGasA           0.038773   0.109413   0.354 0.723115    
## HeatingGasW           0.121723   0.112611   1.081 0.279936    
## HeatingGrav          -0.088251   0.118375  -0.746 0.456092    
## HeatingOthW           0.026819   0.134267   0.200 0.841710    
## HeatingWall           0.104119   0.124331   0.837 0.402504    
## HeatingQCFa          -0.030901   0.019602  -1.576 0.115180    
## HeatingQCGd          -0.023566   0.008900  -2.648 0.008195 ** 
## HeatingQCPo           0.006974   0.112147   0.062 0.950420    
## HeatingQCTA          -0.035683   0.008850  -4.032 5.85e-05 ***
## CentralAirY           0.051892   0.015806   3.283 0.001054 ** 
## X1stFlrSF             0.009807   0.006561   1.495 0.135223    
## GrLivArea             0.123997   0.006882  18.018  < 2e-16 ***
## BsmtFullBath          0.010707   0.003941   2.717 0.006672 ** 
## FullBath              0.008567   0.004976   1.722 0.085385 .  
## HalfBath              0.010358   0.004461   2.322 0.020390 *  
## KitchenQualFa        -0.076077   0.025311  -3.006 0.002700 ** 
## KitchenQualGd        -0.076860   0.014188  -5.417 7.19e-08 ***
## KitchenQualTA        -0.078435   0.016125  -4.864 1.29e-06 ***
## FunctionalMaj2       -0.218182   0.058164  -3.751 0.000184 ***
## FunctionalMin1        0.016348   0.035403   0.462 0.644318    
## FunctionalMin2        0.019269   0.034754   0.554 0.579385    
## FunctionalMod        -0.049839   0.041614  -1.198 0.231273    
## FunctionalSev        -0.375096   0.114806  -3.267 0.001114 ** 
## FunctionalTyp         0.077656   0.030131   2.577 0.010068 *  
## Fireplaces             0.013668   0.003731   3.664 0.000259 ***
## GarageYrBlt           -0.008902   0.006405  -1.390 0.164824    
## GarageCars             0.020222   0.007098   2.849 0.004455 ** 
## GarageArea             0.020779   0.007134   2.913 0.003645 ** 
## GarageQualFa          -0.469928   0.119489  -3.933 8.83e-05 ***
## GarageQualGd          -0.401356   0.121742  -3.297 0.001004 ** 
## GarageQualPo          -0.488195   0.148564  -3.286 0.001043 ** 
## GarageQualTA          -0.450105   0.117761  -3.822 0.000138 ***
## GarageCondFa           0.321838   0.140905   2.284 0.022526 *  
## GarageCondGd           0.378763   0.143893   2.632 0.008581 ** 
## GarageCondPo           0.429602   0.151812   2.830 0.004728 ** 
## GarageCondTA           0.368669   0.138853   2.655 0.008024 ** 
## WoodDeckSF             0.006145   0.003200   1.920 0.055024 .  
## SaleTypeCon            0.073461   0.077573   0.947 0.343814    
## SaleTypeConLD          0.128078   0.041505   3.086 0.002072 ** 
## SaleTypeConLI         -0.017842   0.050185  -0.356 0.722260    
## SaleTypeConLw          0.014675   0.051712   0.284 0.776625    
## SaleTypeCWD            0.043652   0.056160   0.777 0.437133    
## SaleTypeNew            0.144988   0.067140   2.159 0.030993 *  
## SaleTypeOth            0.068721   0.063563   1.081 0.279833    
## SaleTypeWD            -0.010422   0.017941  -0.581 0.561421    
## SaleConditionAdjLand  0.118804   0.057145   2.079 0.037809 *  
## SaleConditionAlloca   0.057640   0.035734   1.613 0.106984    
## SaleConditionFamily   0.004743   0.026464   0.179 0.857781    
## SaleConditionNormal   0.062208   0.012327   5.046 5.13e-07 ***
## SaleConditionPartial -0.049616   0.064783  -0.766 0.443883    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.102 on 1315 degrees of freedom
## Multiple R-squared:  0.9412, Adjusted R-squared:  0.9348 
## F-statistic: 146.2 on 144 and 1315 DF,  p-value: < 2.2e-16
```

This model still selects a lot of variables. Let’s try to reduce again the number of regressors based on the p_values obtained for the different regressors and what we discovered in the previous exploratory data analysis.

The summary of the previous model indicated that the following variables seem important as they obtain the lowest p_values: OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofMatl, LotArea, MSZoning, Fireplaces, GrLivArea, BsmtQual, TotalBsmtSF, KitchenQual, GarageQual.

Let’s confirm

Final model and prediction

Conclusion