Assignment 1

Alicia Chimeno & Silvia Ferrer

2023-11-19

# Clear space and Libraries

# Load dataset

#setwd("C:/Users/Silvia/OneDrive - Universitat Politècnica de Catalunya/Escritorio/UPC/MASTER DS/1A/SIM/SIM\_assignment\_1")  
setwd("/Users/ali/Desktop/MASTER/SIM/PROJECT 1")  
train<-read.delim("train.csv", sep=',')

# Explore and Preproces the dataset

We start by looking at an overview of our dataset.

dim(train)

## [1] 1460 81

#names(train)  
str(train)

## '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 : chr "RL" "RL" "RL" "RL" ...  
## $ 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 : 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 : 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 : 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 : int 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 : int 706 978 486 216 655 732 1369 859 0 851 ...  
## $ BsmtFinType2 : chr "Unf" "Unf" "Unf" "Unf" ...  
## $ 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 : chr "GasA" "GasA" "GasA" "GasA" ...  
## $ HeatingQC : chr "Ex" "Ex" "Ex" "Gd" ...  
## $ CentralAir : chr "Y" "Y" "Y" "Y" ...  
## $ Electrical : chr "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...  
## $ 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 : chr "Gd" "TA" "Gd" "Gd" ...  
## $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...  
## $ Functional : chr "Typ" "Typ" "Typ" "Typ" ...  
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...  
## $ FireplaceQu : chr NA "TA" "TA" "Gd" ...  
## $ GarageType : chr "Attchd" "Attchd" "Attchd" "Detchd" ...  
## $ GarageYrBlt : int 2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...  
## $ GarageFinish : chr "RFn" "RFn" "RFn" "Unf" ...  
## $ 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 : chr "TA" "TA" "TA" "TA" ...  
## $ GarageCond : chr "TA" "TA" "TA" "TA" ...  
## $ PavedDrive : chr "Y" "Y" "Y" "Y" ...  
## $ 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 : chr NA NA NA NA ...  
## $ Fence : chr NA NA NA NA ...  
## $ MiscFeature : chr NA NA 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 : chr "WD" "WD" "WD" "WD" ...  
## $ SaleCondition: chr "Normal" "Normal" "Normal" "Abnorml" ...  
## $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

#skim(train)

## Preprocess and select variables

### Record incorrect NA’s

train$Alley[which(is.na(train$Alley))] <- "No alley access"  
train$BsmtQual[which(is.na(train$BsmtQual))] <- "No Basement"  
train$BsmtCond[which(is.na(train$BsmtCond))] <- "No Basement"  
train$BsmtExposure[which(is.na(train$BsmtExposure))] <- "No Basement"  
train$BsmtFinType1[which(is.na(train$BsmtFinType1))] <- "No Basement"  
train$BsmtFinType2[which(is.na(train$BsmtFinType2))] <- "No Basement"  
train$FireplaceQu[which(is.na(train$FireplaceQu))] <- "No Fireplace"  
train$GarageType[which(is.na(train$GarageType))] <- "No Garage"  
train$GarageFinish[which(is.na(train$GarageFinish))] <- "No Garage"  
train$GarageCond[which(is.na(train$GarageCond))] <- "No Garage"  
train$PoolQC[which(is.na(train$PoolQC))] <- "No Pool"  
train$Fence[which(is.na(train$Fence))] <- "No Fence"  
train$MiscFeature[which(is.na(train$MiscFeature))] <- "None"

## Recode variables to factor

# Recode: character to factor  
char\_cols <- which(sapply(train, is.character))  
train[, char\_cols] <- lapply(train[, char\_cols], as.factor)   
  
# Recode: numeric to factor  
train$MSSubClass <- factor(train$MSSubClass)  
train$MoSold <- factor(train$MoSold) # month

## Target variable exploration “SalePrice”

We asses a normality test to the target variable. We reject null hypothesis (H0: variable is normal), therefore we shall not consider that the target variable is normally distributed.

res.con <- condes(train,81)  
res.con$quali

## R2 p.value  
## Neighborhood 0.545574991 1.558600e-225  
## ExterQual 0.477387778 1.439551e-204  
## BsmtQual 0.464993778 8.158548e-196  
## KitchenQual 0.456598624 3.032213e-192  
## GarageFinish 0.305873737 6.228747e-115  
## FireplaceQu 0.293960754 2.971217e-107  
## Foundation 0.256368402 5.791895e-91  
## GarageType 0.249204231 6.117026e-87  
## MSSubClass 0.246315973 8.662166e-79  
## BsmtFinType1 0.210810410 2.386358e-71  
## HeatingQC 0.195500486 2.667062e-67  
## MasVnrType 0.188822500 1.054025e-64  
## BsmtExposure 0.149500332 7.557758e-50  
## SaleCondition 0.135497477 7.988268e-44  
## Exterior1st 0.152773123 2.586089e-43  
## Exterior2nd 0.153829860 4.842186e-43  
## SaleType 0.137287487 5.039767e-42  
## MSZoning 0.107559683 8.817634e-35  
## HouseStyle 0.086312627 3.376777e-25  
## GarageQual 0.081421478 5.388762e-25  
## GarageCond 0.081346203 5.711746e-25  
## LotShape 0.076375708 6.447524e-25  
## CentralAir 0.063165846 1.809506e-22  
## PavedDrive 0.054539728 1.803569e-18  
## Electrical 0.059691460 8.226925e-18  
## RoofStyle 0.057696630 3.653523e-17  
## BsmtCond 0.051395803 8.195794e-16  
## Fence 0.035614717 9.379977e-11  
## BldgType 0.034534027 2.056736e-10  
## LandContour 0.025794085 2.742217e-08  
## BsmtFinType2 0.030293984 5.225649e-08  
## RoofMatl 0.031413123 7.231445e-08  
## Condition1 0.032630640 8.904549e-08  
## Alley 0.020407545 2.996380e-07  
## ExterCond 0.023617621 5.106681e-07  
## PoolQC 0.021195920 7.700989e-07  
## LotConfig 0.021019365 3.163167e-06  
## Functional 0.016480386 4.841697e-04  
## Heating 0.014437135 7.534721e-04  
## MiscFeature 0.007079753 3.500367e-02  
## Condition2 0.009899161 4.342566e-02

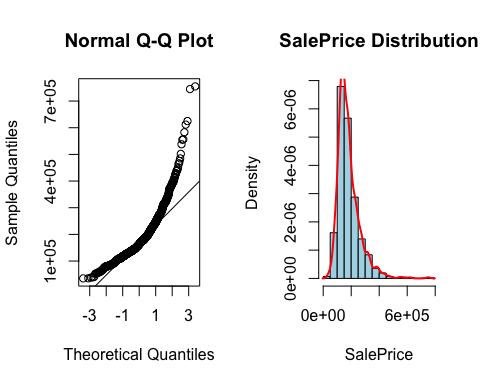
shapiro.test(train$SalePrice)

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

par(mfrow=c(1,2))  
shapiro.test(log(train$SalePrice)) # log-normality

##   
## Shapiro-Wilk normality test  
##   
## data: log(train$SalePrice)  
## W = 0.99121, p-value = 1.149e-07

qqnorm(train$SalePrice)  
qqline(train$SalePrice)  
hist(train$SalePrice, prob = TRUE, col = 'lightblue', main = 'SalePrice Distribution', xlab = 'SalePrice')  
lines(density(train$SalePrice), col = 'red', lwd = 2)



par(mfrow=c(1,1))

## Selection of 10 factor variables and numerical variables

We select the first 10 categorical variables that are more associated with the target variable.

vect<-row.names(res.con$quali)  
vect[1:10]

## [1] "Neighborhood" "ExterQual" "BsmtQual" "KitchenQual" "GarageFinish"  
## [6] "FireplaceQu" "Foundation" "GarageType" "MSSubClass" "BsmtFinType1"

factor\_train<-train[,c(vect[1:10])]  
num\_cols <- names(train)[sapply(train, is.numeric)]  
train2 <- train[,c(num\_cols, vect[1:10])]  
dim(train2)

## [1] 1460 46

# Variable analysis (EDA and univariate outliers)

## Analysis functions

To proceed with the predictor analysis, we do create the following functions for analyzing our numerical features:

## numeric variable description function  
numeric\_description <- function(variable, n\_breaks) {  
 cat("Summary:\n")  
 print(summary(variable))  
   
 cat("\nCount of missing values:",sum(is.na(variable)),"\n")  
 column\_name <- sub(".+\\$", "", deparse(substitute(variable)))  
 hist(variable, breaks = n\_breaks, freq = F, main = paste("Histograma de", column\_name), xlab = column\_name)  
 curve(dnorm(x, mean(variable), sd(variable)), add = T)  
   
 # Normality test with Shapiro-Wilk  
 print(shapiro.test(variable))  
}  
  
## analyze outliers functions  
analyze\_outliers <- function(data, column\_name) {  
 sm <- summary(data[[column\_name]])  
 iqr <- sm["3rd Qu."] - sm["1st Qu."]  
   
 # Mild Outliers  
 mild\_ub <- sm["3rd Qu."] + 1.5 \* iqr  
 mild\_lb <- sm["1st Qu."] - 1.5 \* iqr  
   
 mild\_outliers <- length(which(data[[column\_name]] > mild\_ub | data[[column\_name]] < mild\_lb))  
 cat("Number of mild outliers:", mild\_outliers, "\n")  
   
 # Plotting mild outliers  
 Boxplot(data[[column\_name]], main = paste("Outlier Analysis for", column\_name),  
 ylab = column\_name, outline = TRUE)  
 abline(h = mild\_ub, col = "orange", lwd = 2)  
 abline(h = mild\_lb, col = "orange", lwd = 2)  
   
 # Severe Outliers  
 severe\_ub <- sm["3rd Qu."] + 3 \* iqr  
 severe\_lb <- sm["1st Qu."] - 3 \* iqr  
 severe\_outliers <- length(which(data[[column\_name]] > severe\_ub | data[[column\_name]] < severe\_lb))  
 cat("Number of severe outliers:", severe\_outliers, "\n")  
 # Plotting severe outliers  
 abline(h = severe\_ub, col = "red", lwd = 2.5)  
 abline(h = severe\_lb, col = "red", lwd = 2.5)  
}  
# count outliers (new column)  
  
train2$univ\_outl\_count <- 0 # initialize  
  
# count outliers  
update\_outliers\_count <- function(dataframe, column\_name) {  
 df <- dataframe  
 sm <- summary(df[[column\_name]])  
 iqr <- sm["3rd Qu."] - sm["1st Qu."]  
 # Severe Outliers  
 severe\_ub <- sm["3rd Qu."] + 3 \* iqr  
 severe\_lb <- sm["1st Qu."] - 3 \* iqr  
 severe\_outliers\_id <- which(df[[column\_name]] > severe\_ub | df[[column\_name]] < severe\_lb)  
 df$univ\_outl\_count[severe\_outliers\_id] <- df$univ\_outl\_count[severe\_outliers\_id] + 1  
 return(df)  
}

We continue performing a descriptive analysis of each variable. Therefore, we are using the function update\_outliers\_count where we are updating a new variable named univ\_outl\_count, which is a count of the total univariate outliers an instance has, in order to track the data quality and associate having univariate outliers to some features with a correlation matrix. The order of this descriptive analysis follows the criteria of splitting into numerical and then categorical variables.

skim(train2) # overall exploration

Data summary

|  |  |
| --- | --- |
| Name | train2 |
| Number of rows | 1460 |
| Number of columns | 47 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 10 |
| numeric | 37 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Neighborhood | 0 | 1 | FALSE | 25 | NAm: 225, Col: 150, Old: 113, Edw: 100 |
| ExterQual | 0 | 1 | FALSE | 4 | TA: 906, Gd: 488, Ex: 52, Fa: 14 |
| BsmtQual | 0 | 1 | FALSE | 5 | TA: 649, Gd: 618, Ex: 121, No : 37 |
| KitchenQual | 0 | 1 | FALSE | 4 | TA: 735, Gd: 586, Ex: 100, Fa: 39 |
| GarageFinish | 0 | 1 | FALSE | 4 | Unf: 605, RFn: 422, Fin: 352, No : 81 |
| FireplaceQu | 0 | 1 | FALSE | 6 | No : 690, Gd: 380, TA: 313, Fa: 33 |
| Foundation | 0 | 1 | FALSE | 6 | PCo: 647, CBl: 634, Brk: 146, Sla: 24 |
| GarageType | 0 | 1 | FALSE | 7 | Att: 870, Det: 387, Bui: 88, No : 81 |
| MSSubClass | 0 | 1 | FALSE | 15 | 20: 536, 60: 299, 50: 144, 120: 87 |
| BsmtFinType1 | 0 | 1 | FALSE | 7 | Unf: 430, GLQ: 418, ALQ: 220, BLQ: 148 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1.00 | 730.50 | 421.61 | 1 | 365.75 | 730.5 | 1095.25 | 1460 | ▇▇▇▇▇ |
| LotFrontage | 259 | 0.82 | 70.05 | 24.28 | 21 | 59.00 | 69.0 | 80.00 | 313 | ▇▃▁▁▁ |
| LotArea | 0 | 1.00 | 10516.83 | 9981.26 | 1300 | 7553.50 | 9478.5 | 11601.50 | 215245 | ▇▁▁▁▁ |
| OverallQual | 0 | 1.00 | 6.10 | 1.38 | 1 | 5.00 | 6.0 | 7.00 | 10 | ▁▂▇▅▁ |
| OverallCond | 0 | 1.00 | 5.58 | 1.11 | 1 | 5.00 | 5.0 | 6.00 | 9 | ▁▁▇▅▁ |
| YearBuilt | 0 | 1.00 | 1971.27 | 30.20 | 1872 | 1954.00 | 1973.0 | 2000.00 | 2010 | ▁▂▃▆▇ |
| YearRemodAdd | 0 | 1.00 | 1984.87 | 20.65 | 1950 | 1967.00 | 1994.0 | 2004.00 | 2010 | ▅▂▂▃▇ |
| MasVnrArea | 8 | 0.99 | 103.69 | 181.07 | 0 | 0.00 | 0.0 | 166.00 | 1600 | ▇▁▁▁▁ |
| BsmtFinSF1 | 0 | 1.00 | 443.64 | 456.10 | 0 | 0.00 | 383.5 | 712.25 | 5644 | ▇▁▁▁▁ |
| BsmtFinSF2 | 0 | 1.00 | 46.55 | 161.32 | 0 | 0.00 | 0.0 | 0.00 | 1474 | ▇▁▁▁▁ |
| BsmtUnfSF | 0 | 1.00 | 567.24 | 441.87 | 0 | 223.00 | 477.5 | 808.00 | 2336 | ▇▅▂▁▁ |
| TotalBsmtSF | 0 | 1.00 | 1057.43 | 438.71 | 0 | 795.75 | 991.5 | 1298.25 | 6110 | ▇▃▁▁▁ |
| X1stFlrSF | 0 | 1.00 | 1162.63 | 386.59 | 334 | 882.00 | 1087.0 | 1391.25 | 4692 | ▇▅▁▁▁ |
| X2ndFlrSF | 0 | 1.00 | 346.99 | 436.53 | 0 | 0.00 | 0.0 | 728.00 | 2065 | ▇▃▂▁▁ |
| LowQualFinSF | 0 | 1.00 | 5.84 | 48.62 | 0 | 0.00 | 0.0 | 0.00 | 572 | ▇▁▁▁▁ |
| GrLivArea | 0 | 1.00 | 1515.46 | 525.48 | 334 | 1129.50 | 1464.0 | 1776.75 | 5642 | ▇▇▁▁▁ |
| BsmtFullBath | 0 | 1.00 | 0.43 | 0.52 | 0 | 0.00 | 0.0 | 1.00 | 3 | ▇▆▁▁▁ |
| BsmtHalfBath | 0 | 1.00 | 0.06 | 0.24 | 0 | 0.00 | 0.0 | 0.00 | 2 | ▇▁▁▁▁ |
| FullBath | 0 | 1.00 | 1.57 | 0.55 | 0 | 1.00 | 2.0 | 2.00 | 3 | ▁▇▁▇▁ |
| HalfBath | 0 | 1.00 | 0.38 | 0.50 | 0 | 0.00 | 0.0 | 1.00 | 2 | ▇▁▅▁▁ |
| BedroomAbvGr | 0 | 1.00 | 2.87 | 0.82 | 0 | 2.00 | 3.0 | 3.00 | 8 | ▁▇▂▁▁ |
| KitchenAbvGr | 0 | 1.00 | 1.05 | 0.22 | 0 | 1.00 | 1.0 | 1.00 | 3 | ▁▇▁▁▁ |
| TotRmsAbvGrd | 0 | 1.00 | 6.52 | 1.63 | 2 | 5.00 | 6.0 | 7.00 | 14 | ▂▇▇▁▁ |
| Fireplaces | 0 | 1.00 | 0.61 | 0.64 | 0 | 0.00 | 1.0 | 1.00 | 3 | ▇▇▁▁▁ |
| GarageYrBlt | 81 | 0.94 | 1978.51 | 24.69 | 1900 | 1961.00 | 1980.0 | 2002.00 | 2010 | ▁▁▅▅▇ |
| GarageCars | 0 | 1.00 | 1.77 | 0.75 | 0 | 1.00 | 2.0 | 2.00 | 4 | ▁▃▇▂▁ |
| GarageArea | 0 | 1.00 | 472.98 | 213.80 | 0 | 334.50 | 480.0 | 576.00 | 1418 | ▂▇▃▁▁ |
| WoodDeckSF | 0 | 1.00 | 94.24 | 125.34 | 0 | 0.00 | 0.0 | 168.00 | 857 | ▇▂▁▁▁ |
| OpenPorchSF | 0 | 1.00 | 46.66 | 66.26 | 0 | 0.00 | 25.0 | 68.00 | 547 | ▇▁▁▁▁ |
| EnclosedPorch | 0 | 1.00 | 21.95 | 61.12 | 0 | 0.00 | 0.0 | 0.00 | 552 | ▇▁▁▁▁ |
| X3SsnPorch | 0 | 1.00 | 3.41 | 29.32 | 0 | 0.00 | 0.0 | 0.00 | 508 | ▇▁▁▁▁ |
| ScreenPorch | 0 | 1.00 | 15.06 | 55.76 | 0 | 0.00 | 0.0 | 0.00 | 480 | ▇▁▁▁▁ |
| PoolArea | 0 | 1.00 | 2.76 | 40.18 | 0 | 0.00 | 0.0 | 0.00 | 738 | ▇▁▁▁▁ |
| MiscVal | 0 | 1.00 | 43.49 | 496.12 | 0 | 0.00 | 0.0 | 0.00 | 15500 | ▇▁▁▁▁ |
| YrSold | 0 | 1.00 | 2007.82 | 1.33 | 2006 | 2007.00 | 2008.0 | 2009.00 | 2010 | ▇▇▇▇▅ |
| SalePrice | 0 | 1.00 | 180921.20 | 79442.50 | 34900 | 129975.00 | 163000.0 | 214000.00 | 755000 | ▇▅▁▁▁ |
| univ\_outl\_count | 0 | 1.00 | 0.00 | 0.00 | 0 | 0.00 | 0.0 | 0.00 | 0 | ▁▁▇▁▁ |

names(train2)

## [1] "Id" "LotFrontage" "LotArea" "OverallQual"   
## [5] "OverallCond" "YearBuilt" "YearRemodAdd" "MasVnrArea"   
## [9] "BsmtFinSF1" "BsmtFinSF2" "BsmtUnfSF" "TotalBsmtSF"   
## [13] "X1stFlrSF" "X2ndFlrSF" "LowQualFinSF" "GrLivArea"   
## [17] "BsmtFullBath" "BsmtHalfBath" "FullBath" "HalfBath"   
## [21] "BedroomAbvGr" "KitchenAbvGr" "TotRmsAbvGrd" "Fireplaces"   
## [25] "GarageYrBlt" "GarageCars" "GarageArea" "WoodDeckSF"   
## [29] "OpenPorchSF" "EnclosedPorch" "X3SsnPorch" "ScreenPorch"   
## [33] "PoolArea" "MiscVal" "YrSold" "SalePrice"   
## [37] "Neighborhood" "ExterQual" "BsmtQual" "KitchenQual"   
## [41] "GarageFinish" "FireplaceQu" "Foundation" "GarageType"   
## [45] "MSSubClass" "BsmtFinType1" "univ\_outl\_count"

## Numerical variables

Firstly, we are analyzing the numerical variables, excluding the variable ‘Id’, which is not considered for the analysis. We distinguish between continuous variables (areas,) and discrete variables (nummber of garages, rooms…)

### “LotFrontage” Linear feet of street connected to property (cont)

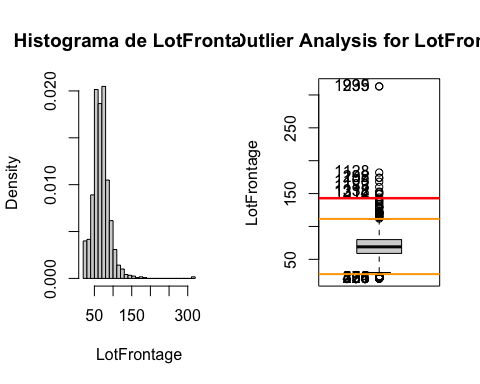
par(mfrow=c(1,2))  
numeric\_description(train2$LotFrontage, 30) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 21.00 59.00 69.00 70.05 80.00 313.00 259   
##   
## Count of missing values: 259

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.8804, p-value < 2.2e-16

analyze\_outliers(train2, "LotFrontage")

## Number of mild outliers: 88



## Number of severe outliers: 12

train2 <- update\_outliers\_count(train2, "LotFrontage")

### “LotArea” Lot size in square feet (cont)

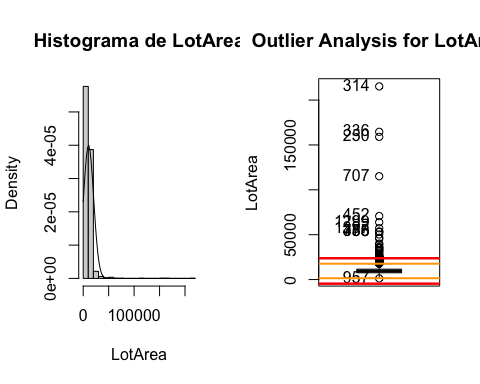
par(mfrow=c(1,2))  
numeric\_description(train2$LotArea, 20) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1300 7554 9478 10517 11602 215245   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.35106, p-value < 2.2e-16

analyze\_outliers(train2, "LotArea")

## Number of mild outliers: 69

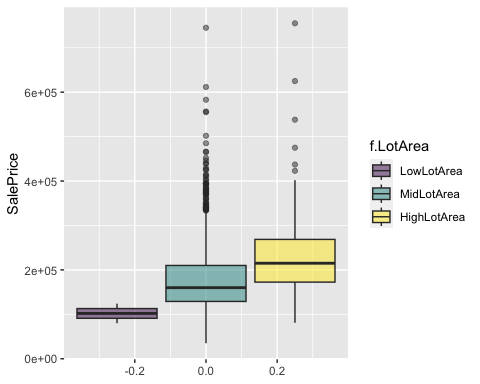


## Number of severe outliers: 34

train2 <- update\_outliers\_count(train2, "LotArea")  
  
# new variable:  
sm<-summary(train2$LotArea)  
plot(train2$LotArea,train2$SalePrice)  
  
iqr <- sm["3rd Qu."] - sm["1st Qu."]  
mild\_ub <- sm["3rd Qu."] + 1.5 \* iqr  
mild\_lb <- sm["1st Qu."] - 1.5 \* iqr  
  
train2$f.LotArea <- ifelse(  
 train2$LotArea <= mild\_lb,   
 1,   
 ifelse(  
 train2$LotArea > mild\_lb & train2$LotArea < mild\_ub,   
 2,  
 ifelse(  
 train2$LotArea >= mild\_ub,   
 3,  
   
 NA   
 )  
 )  
)  
  
train2$f.LotArea<- factor(train2$f.LotArea, labels=c("LowLotArea","MidLotArea","HighLotArea"), order = T, levels=c(1,2,3))  
table(train2$f.LotArea)

##   
## LowLotArea MidLotArea HighLotArea   
## 2 1391 67

ggplot(train2, aes(y=SalePrice, fill=f.LotArea)) + geom\_boxplot(alpha=0.5)



### “OverallQual” (cont)

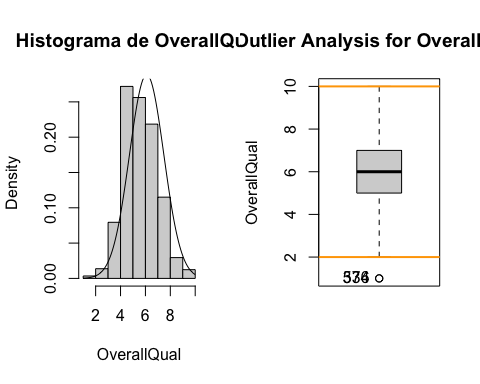
par(mfrow=c(1,2))  
numeric\_description(train2$OverallQual, 10) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 5.000 6.000 6.099 7.000 10.000   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.94801, p-value < 2.2e-16

analyze\_outliers(train2, "OverallQual")

## Number of mild outliers: 2



## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "OverallQual")

### “YearBuilt” (int)

summary(train2$YearBuilt)

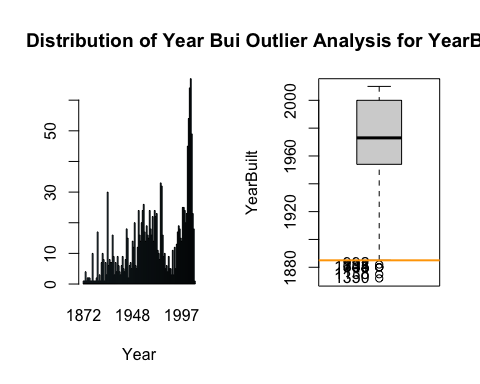
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1872 1954 1973 1971 2000 2010

par(mfrow=c(1,2))  
barplot(table(train2$YearBuilt), main = "Distribution of Year Built",xlab = "Year",col = "skyblue")  
sum(is.na(train2$YearBuilt))

## [1] 0

analyze\_outliers(train2, "YearBuilt")

## Number of mild outliers: 7



## Number of severe outliers: 0

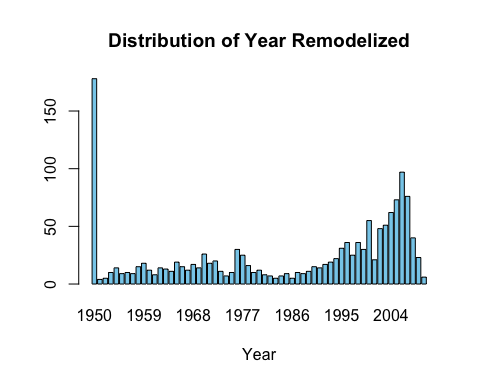
train2 <- update\_outliers\_count(train2, "YearBuilt")

### “YearRemodAdd” (int) - Remodel date (same as construction date if no remodeling or additions)

summary(train2$YearRemodAdd)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1950 1967 1994 1985 2004 2010

barplot(table(train2$YearRemodAdd), main = "Distribution of Year Remodelized",xlab = "Year",col = "skyblue")

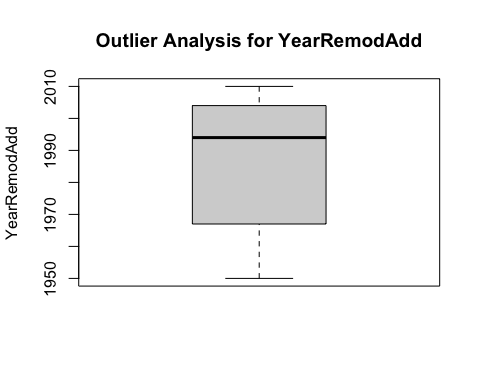


sum(is.na(train2$YearRemodAdd))

## [1] 0

analyze\_outliers(train2, "YearRemodAdd")

## Number of mild outliers: 0



## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "YearRemodAdd")

### MasVnrArea - Masonry veneer area in square feet (cont)

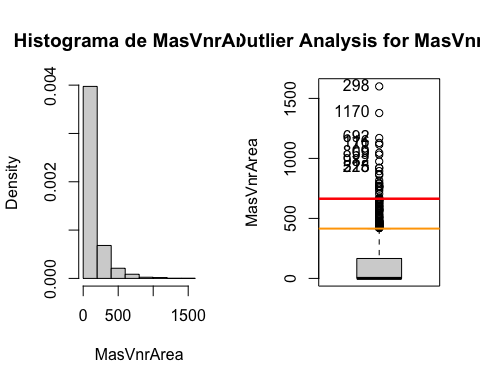
par(mfrow=c(1,2))  
numeric\_description(train2$MasVnrArea, 10)# not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 0.0 0.0 103.7 166.0 1600.0 8   
##   
## Count of missing values: 8

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.63929, p-value < 2.2e-16

analyze\_outliers(train2,"MasVnrArea")

## Number of mild outliers: 96



## Number of severe outliers: 25

train2 <- update\_outliers\_count(train2, "MasVnrArea")  
# find values that are NA instead of 0  
# if it has a material but a 0 in masvnrarea is a NA  
train2$MasVnrArea[which(train2$MasVnrArea == 0)] <- NA  
none\_idx <- which(train$MasVnrType == 'None')  
train2$MasVnrArea[none\_idx] <- 0

### “BsmtFinSF1” - Type 1 finished square feet (cont)

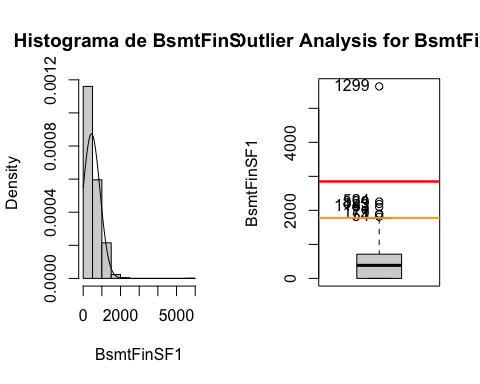
par(mfrow=c(1,2))  
numeric\_description(train2$BsmtFinSF1, 10)# not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 383.5 443.6 712.2 5644.0   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.84796, p-value < 2.2e-16

analyze\_outliers(train2,"BsmtFinSF1")

## Number of mild outliers: 7



## Number of severe outliers: 1

train2 <- update\_outliers\_count(train2, "BsmtFinSF1")

### BsmtFinSF2 - Rating of basement finished area (if multiple types) (cont)

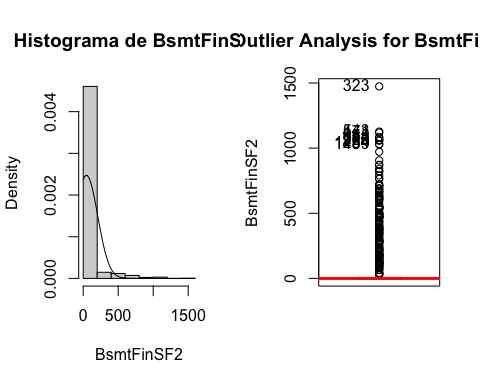
par(mfrow=c(1,2))  
numeric\_description(train2$BsmtFinSF2, 10) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 46.55 0.00 1474.00   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.32728, p-value < 2.2e-16

analyze\_outliers(train2,"BsmtFinSF2")

## Number of mild outliers: 167



## Number of severe outliers: 167

train2 <- update\_outliers\_count(train2, "BsmtFinSF2")

### “BsmtUnfSF” (cont)

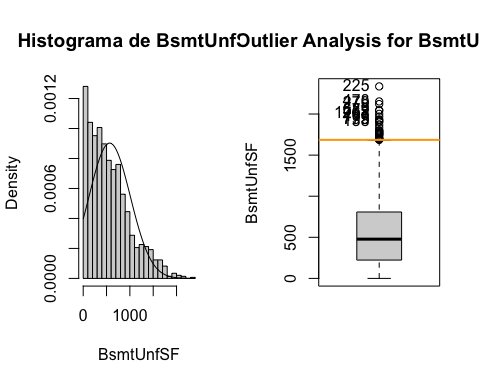
par(mfrow=c(1,2))  
numeric\_description(train2$BsmtUnfSF, 20) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 223.0 477.5 567.2 808.0 2336.0   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.93042, p-value < 2.2e-16

analyze\_outliers(train2,"BsmtUnfSF")

## Number of mild outliers: 29



## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "BsmtUnfSF")

### “TotalBsmtSF” (cont)

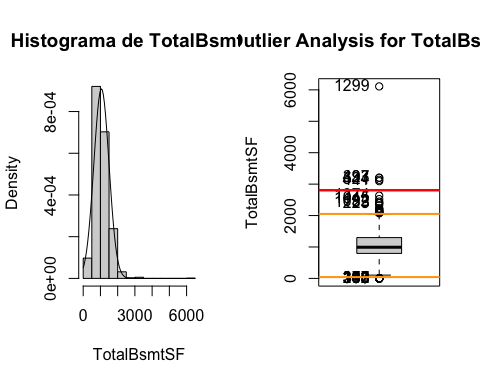
par(mfrow=c(1,2))  
numeric\_description(train2$TotalBsmtSF, 10)# not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 795.8 991.5 1057.4 1298.2 6110.0   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.91735, p-value < 2.2e-16

analyze\_outliers(train2,"TotalBsmtSF")

## Number of mild outliers: 61



## Number of severe outliers: 5

train2 <- update\_outliers\_count(train2, "TotalBsmtSF")

### X1stFlrSF - First Floor square feet (cont)

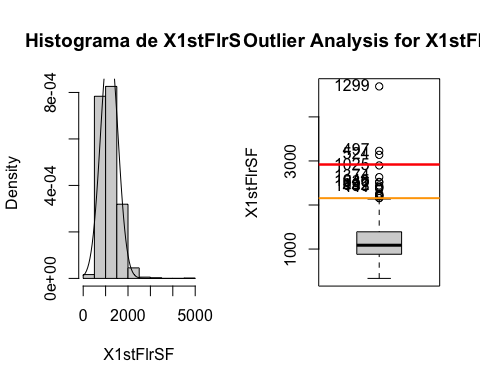
par(mfrow=c(1,2))  
numeric\_description(train2$X1stFlrSF, 10) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 882 1087 1163 1391 4692   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.92695, p-value < 2.2e-16

analyze\_outliers(train2,"X1stFlrSF")

## Number of mild outliers: 20



## Number of severe outliers: 3

train2 <- update\_outliers\_count(train2, "X1stFlrSF")

### X2ndFlrSF - Second floor square feet (cont)

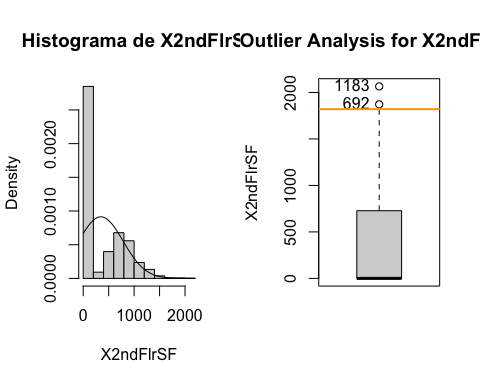
par(mfrow=c(1,2))  
numeric\_description(train2$X2ndFlrSF, 10)# not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 347 728 2065   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.7668, p-value < 2.2e-16

analyze\_outliers(train2,"X2ndFlrSF")

## Number of mild outliers: 2



## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "X2ndFlrSF")  
  
# create a new variable if has second floor = 1 / no second floor = 0.   
train2$secondfloor<-0  
train2$secondfloor[which(train2$X2ndFlrSF!=0)]<-1  
train2$secondfloor<-factor(train2$secondfloor)  
table(train2$secondfloor)

##   
## 0 1   
## 829 631

### LowQualFinSF - Low quality finished square feet (all floors) (cont)

par(mfrow=c(1,2))  
numeric\_description(train2$LowQualFinSF, 10) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 5.845 0.000 572.000   
##   
## Count of missing values: 0

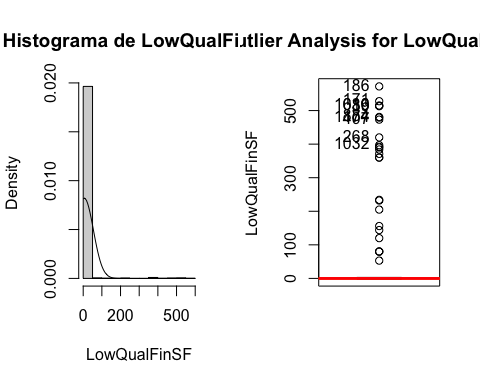
##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.09799, p-value < 2.2e-16

table(train2$LowQualFinSF)# very centered in 0

##   
## 0 53 80 120 144 156 205 232 234 360 371 384 390 392 397 420   
## 1434 1 3 1 1 1 1 1 1 2 1 1 1 1 1 1   
## 473 479 481 513 514 515 528 572   
## 1 1 1 1 1 1 1 1

analyze\_outliers(train2,"LowQualFinSF")

## Number of mild outliers: 26



## Number of severe outliers: 26

nearZeroVar(train2$LowQualFinSF, saveMetrics = TRUE) # Variable with near zero variance

## freqRatio percentUnique zeroVar nzv  
## 1 478 1.643836 FALSE TRUE

train2 <- update\_outliers\_count(train2, "LowQualFinSF")

### GrLivArea - Above grade (ground) living area square feet (cont)

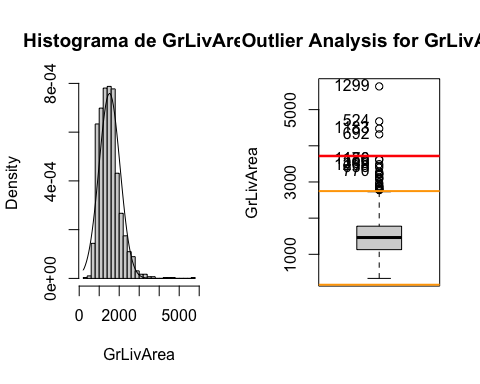
par(mfrow=c(1,2))  
numeric\_description(train2$GrLivArea, 20) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1130 1464 1515 1777 5642   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.92798, p-value < 2.2e-16

analyze\_outliers(train2,"GrLivArea")

## Number of mild outliers: 31



## Number of severe outliers: 4

train2 <- update\_outliers\_count(train2, "GrLivArea")

### BsmtFullBath - # full baths in the basement (int)

summary(train2$BsmtFullBath)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.4253 1.0000 3.0000

par(mfrow=c(1,2))  
analyze\_outliers(train2,"BsmtFullBath")

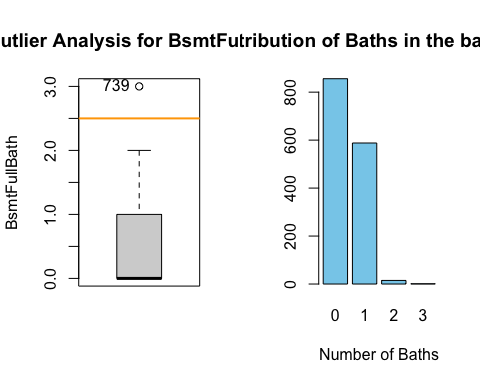
## Number of mild outliers: 1

## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "BsmtFullBath")  
#table(train2$BsmtFullBath)  
sum(is.na(train2$BsmtFullBath))

## [1] 0

barplot(table(train2$BsmtFullBath), main = "Distribution of Baths in the basement",xlab = "Number of Baths",col = "skyblue")



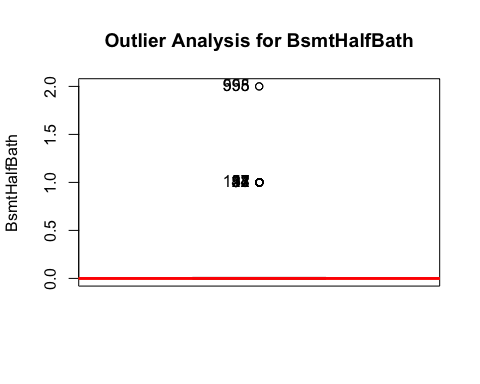
### BsmtHalfBath - # half baths in the basement (int)

summary(train2$BsmtHalfBath)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.05753 0.00000 2.00000

analyze\_outliers(train2,"BsmtHalfBath")

## Number of mild outliers: 82



## Number of severe outliers: 82

train2 <- update\_outliers\_count(train2, "BsmtHalfBath")  
table(train2$BsmtHalfBath) # very centered in 0

##   
## 0 1 2   
## 1378 80 2

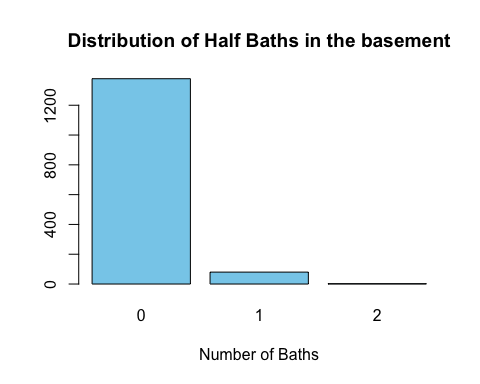
nearZeroVar(train2$BsmtHalfBath, saveMetrics = TRUE) # Variable not with near zero variance

## freqRatio percentUnique zeroVar nzv  
## 1 17.225 0.2054795 FALSE FALSE

sum(is.na(train2$BsmtHalfBath))

## [1] 0

barplot(table(train2$BsmtHalfBath), main = "Distribution of Half Baths in the basement",xlab = "Number of Baths",col = "skyblue")



### FullBath - full baths above grade (int)

summary(train2$FullBath)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 1.565 2.000 3.000

par(mfrow=c(1,2))  
analyze\_outliers(train2,"FullBath")

## Number of mild outliers: 0

## Number of severe outliers: 0

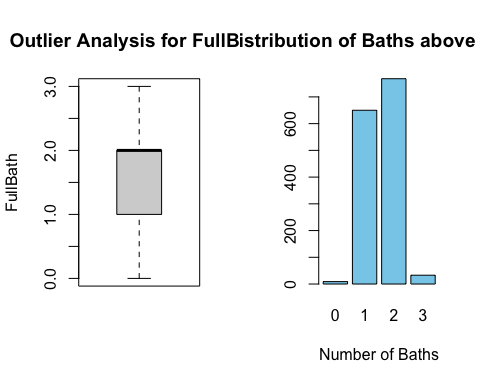
train2 <- update\_outliers\_count(train2, "FullBath")  
table(train2$FullBath)

##   
## 0 1 2 3   
## 9 650 768 33

sum(is.na(train2$FullBath))

## [1] 0

barplot(table(train2$FullBath), main = "Distribution of Baths above grade",xlab = "Number of Baths",col = "skyblue")



### HalfBath - Half baths above grade (int)

summary(train2$HalfBath)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.3829 1.0000 2.0000

par(mfrow=c(1,2))  
analyze\_outliers(train2,"HalfBath")

## Number of mild outliers: 0

## Number of severe outliers: 0

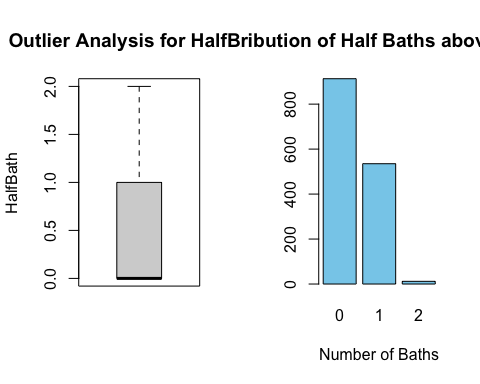
train2 <- update\_outliers\_count(train2, "HalfBath")  
table(train2$HalfBath)

##   
## 0 1 2   
## 913 535 12

sum(is.na(train2$HalfBath))

## [1] 0

barplot(table(train2$HalfBath), main = "Distribution of Half Baths above grade",xlab = "Number of Baths",col = "skyblue")



### BedroomAbvGr - Bedrooms above grade (does NOT include basement bedrooms)

summary(train2$BedroomAbvGr)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 3.000 2.866 3.000 8.000

par(mfrow=c(1,2))  
analyze\_outliers(train2,"BedroomAbvGr")

## Number of mild outliers: 35

## Number of severe outliers: 1

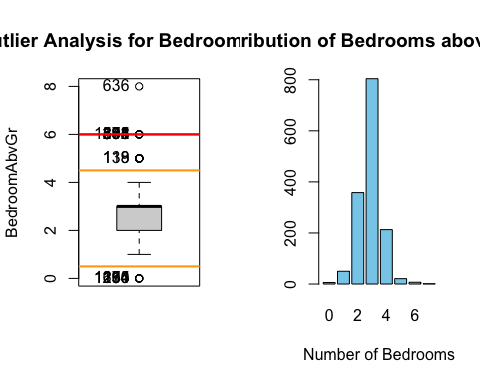
train2 <- update\_outliers\_count(train2, "BedroomAbvGr")  
table(train2$BedroomAbvGr)

##   
## 0 1 2 3 4 5 6 8   
## 6 50 358 804 213 21 7 1

sum(is.na(train2$BedroomAbvGr))

## [1] 0

barplot(table(train2$BedroomAbvGr), main = "Distribution of Bedrooms above grade",xlab = "Number of Bedrooms",col = "skyblue")



### KitchenAbvGr - Kitchens above grade (int)

summary(train2$KitchenAbvGr)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 1.000 1.047 1.000 3.000

par(mfrow=c(1,2))  
analyze\_outliers(train2,"KitchenAbvGr")

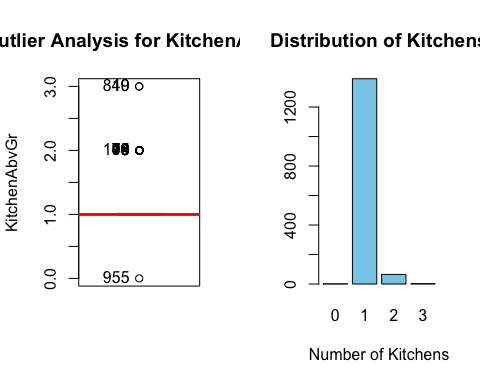
## Number of mild outliers: 68

## Number of severe outliers: 68

train2 <- update\_outliers\_count(train2, "KitchenAbvGr")  
sum(is.na(train2$KitchenAbvGr))

## [1] 0

barplot(table(train2$KitchenAbvGr), main = "Distribution of Kitchens",xlab = "Number of Kitchens",col = "skyblue")



table(train2$KitchenAbvGr) # variable very centered in 1

##   
## 0 1 2 3   
## 1 1392 65 2

nearZeroVar(train2$KitchenAbvGr,saveMetrics = TRUE) # near zero variance

## freqRatio percentUnique zeroVar nzv  
## 1 21.41538 0.2739726 FALSE TRUE

### “TotRmsAbvGrd” - Total rooms above grade (does not include bathrooms) - continuous ratio variable (int)

par(mfrow=c(1,2))  
summary(train2$TotRmsAbvGrd)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 5.000 6.000 6.518 7.000 14.000

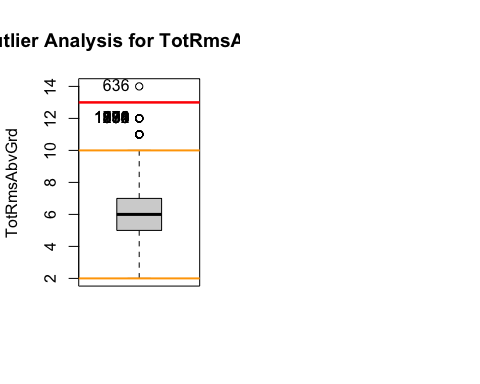
analyze\_outliers(train2,"TotRmsAbvGrd")

## Number of mild outliers: 30

## Number of severe outliers: 1

train2 <- update\_outliers\_count(train2, "TotRmsAbvGrd")  
#table(train2$TotRmsAbvGrd)  
sum(is.na(train2$TotRmsAbvGrd))

## [1] 0



### “Fireplaces” - Number of fireplaces (int)

summary(train2$Fireplaces)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 1.000 0.613 1.000 3.000

par(mfrow=c(1,2))  
#table(train2$Fireplaces)  
sum(is.na(train2$Fireplaces))

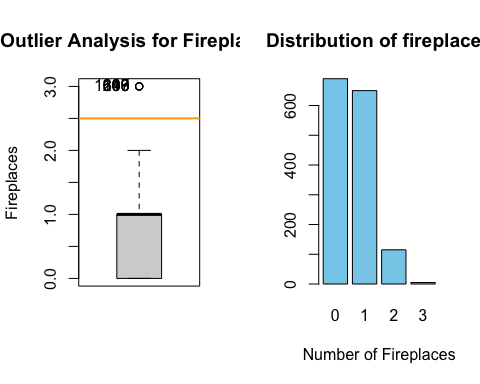
## [1] 0

analyze\_outliers(train2, "Fireplaces")

## Number of mild outliers: 5

## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "Fireplaces")  
barplot(table(train2$Fireplaces), main = "Distribution of fireplaces",xlab = "Number of Fireplaces",col = "skyblue")



### “GarageCars” - Size of garage in car capacity (int)

summary(train2$GarageCars)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 1.767 2.000 4.000

#table(train2$GarageCars)  
sum(is.na(train2$GarageCars)) #0

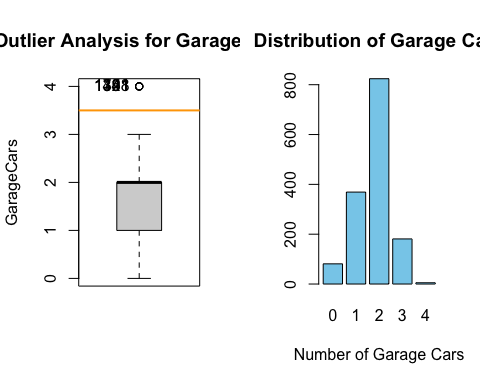
## [1] 0

par(mfrow=c(1,2))  
analyze\_outliers(train2, "GarageCars")

## Number of mild outliers: 5

## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "GarageCars")  
barplot(table(train2$GarageCars), main = "Distribution of Garage Cars",xlab = "Number of Garage Cars",col = "skyblue")



### “GarageArea” - Size of garage in square feet (cont)

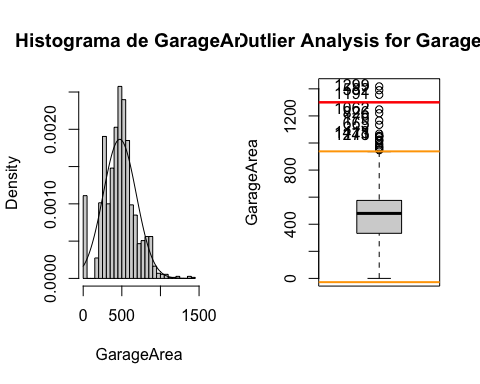
par(mfrow=c(1,2))  
numeric\_description(train2$GarageArea, 30) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 334.5 480.0 473.0 576.0 1418.0   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.97533, p-value = 4.017e-15

analyze\_outliers(train2, "GarageArea")

## Number of mild outliers: 21



## Number of severe outliers: 3

train2 <- update\_outliers\_count(train2, "GarageArea")  
sm<-summary(train2$GarageArea)  
  
# checking if 0 mean that they don't have garage  
which(train$GarageCars == 0)

## [1] 40 49 79 89 90 100 109 126 128 141 149 156 164 166 199  
## [16] 211 242 251 288 292 308 376 387 394 432 435 442 465 496 521  
## [31] 529 534 536 563 583 614 615 621 636 637 639 650 706 711 739  
## [46] 751 785 827 844 922 943 955 961 969 971 977 1010 1012 1031 1039  
## [61] 1097 1124 1132 1138 1144 1174 1180 1219 1220 1235 1258 1284 1324 1326 1327  
## [76] 1338 1350 1408 1450 1451 1454

which(train$GarageArea == 0)

## [1] 40 49 79 89 90 100 109 126 128 141 149 156 164 166 199  
## [16] 211 242 251 288 292 308 376 387 394 432 435 442 465 496 521  
## [31] 529 534 536 563 583 614 615 621 636 637 639 650 706 711 739  
## [46] 751 785 827 844 922 943 955 961 969 971 977 1010 1012 1031 1039  
## [61] 1097 1124 1132 1138 1144 1174 1180 1219 1220 1235 1258 1284 1324 1326 1327  
## [76] 1338 1350 1408 1450 1451 1454

which(train$GarageFinish=='No Garage')

## [1] 40 49 79 89 90 100 109 126 128 141 149 156 164 166 199  
## [16] 211 242 251 288 292 308 376 387 394 432 435 442 465 496 521  
## [31] 529 534 536 563 583 614 615 621 636 637 639 650 706 711 739  
## [46] 751 785 827 844 922 943 955 961 969 971 977 1010 1012 1031 1039  
## [61] 1097 1124 1132 1138 1144 1174 1180 1219 1220 1235 1258 1284 1324 1326 1327  
## [76] 1338 1350 1408 1450 1451 1454

### “WoodDeckSF” - Wood deck area in square feet - continuous ratio variable

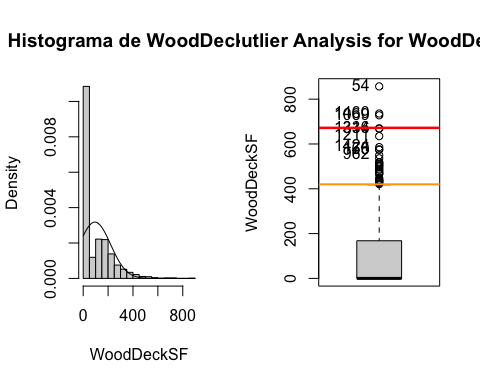
par(mfrow=c(1,2))  
numeric\_description(train2$WoodDeckSF, 30) # not normal distribution

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 94.24 168.00 857.00   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.76852, p-value < 2.2e-16

#table(train2$WoodDeckSF) # centered data in 0  
analyze\_outliers(train2, "WoodDeckSF")

## Number of mild outliers: 32



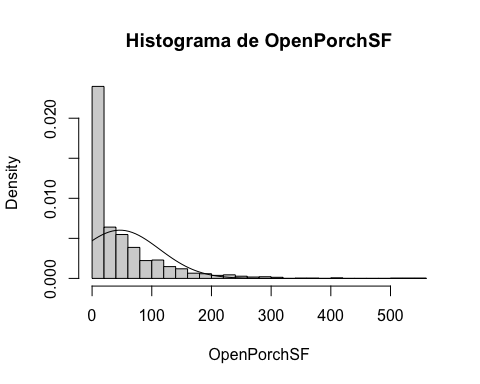
## Number of severe outliers: 3

train2 <- update\_outliers\_count(train2, "WoodDeckSF")

### “OpenPorchSF” - Open porch area in square feet - continuous ratio variable (continua)

numeric\_description(train2$OpenPorchSF, 30)

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 25.00 46.66 68.00 547.00   
##   
## Count of missing values: 0



##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.72717, p-value < 2.2e-16

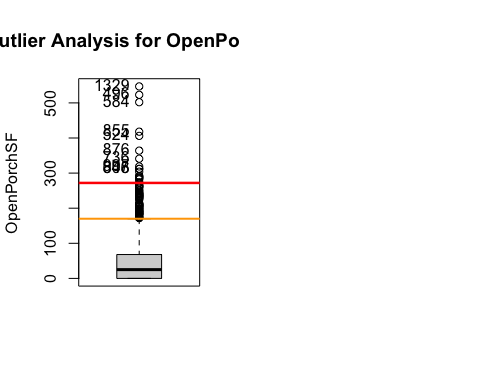
par(mfrow=c(1,2))  
analyze\_outliers(train2,"OpenPorchSF")

## Number of mild outliers: 77

## Number of severe outliers: 18

train2 <- update\_outliers\_count(train2, "OpenPorchSF")  
#table(train2$OpenPorchSF) # very centered data in 0  
train2$OpenPorch\_binary<-0  
train2$OpenPorch\_binary[which(train2$OpenPorchSF!=0)]<-1  
train2$OpenPorch\_binary<-as.factor(train2$OpenPorch\_binary)  
table(train2$OpenPorch\_binary)

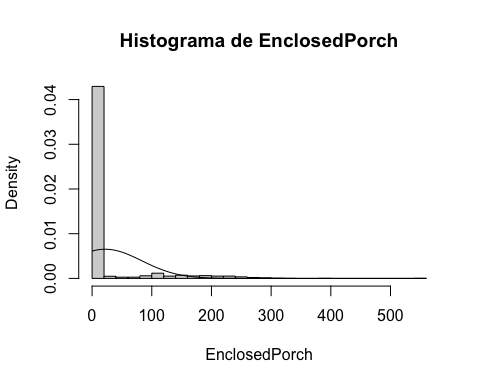
##   
## 0 1   
## 656 804



### EnclosedPorch - Enclosed porch area in square feet (cont)

numeric\_description(train2$EnclosedPorch, 30)

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 21.95 0.00 552.00   
##   
## Count of missing values: 0



##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.41444, p-value < 2.2e-16

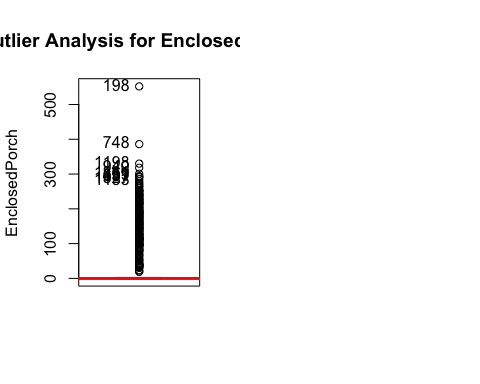
par(mfrow=c(1,2))  
analyze\_outliers(train2,"EnclosedPorch")

## Number of mild outliers: 208

## Number of severe outliers: 208

train2 <- update\_outliers\_count(train2, "EnclosedPorch")  
#table(train2$EnclosedPorch) # very centered data in 0  
  
# therefore we will create a factor 0 if they dont have enclosed porch area,1 if they do have.  
train2$EnclosedPorch\_binary<-0  
train2$EnclosedPorch\_binary[which(train2$EnclosedPorch!=0)]<-1  
train2$EnclosedPorch\_binary<-as.factor(train2$EnclosedPorch\_binary)  
#table(train2$EnclosedPorch\_binary)  
  
# create new variable - has porch or not - mixing both binary variables  
train2$HasPorch\_binary<-0  
train2$HasPorch\_binary[which(train2$EnclosedPorch!=0 | train2$OpenPorchSF!=0)]<-1  
train2$HasPorch\_binary<-as.factor(train2$HasPorch\_binary)  
table(train2$HasPorch\_binary)

##   
## 0 1   
## 513 947



### X3SsnPorch - Three season porch area in square feet (cont)

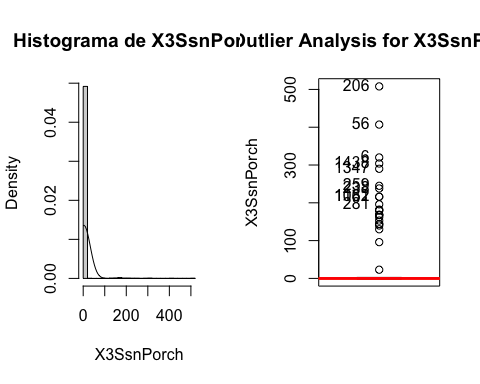
par(mfrow=c(1,2))  
numeric\_description(train2$X3SsnPorch, 30)

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 3.41 0.00 508.00   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.094934, p-value < 2.2e-16

#table(train2$X3SsnPorch) # very centered in 0 -> dont have three season porch  
analyze\_outliers(train2,"X3SsnPorch")

## Number of mild outliers: 24



## Number of severe outliers: 24

train2 <- update\_outliers\_count(train2, "X3SsnPorch")

### ScreenPorch - Screen porch area in square feet (cont)

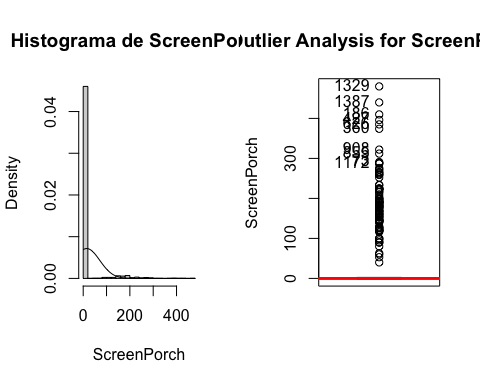
par(mfrow=c(1,2))  
numeric\_description(train2$ScreenPorch, 30) # not normally distributed

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 15.06 0.00 480.00   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.29821, p-value < 2.2e-16

#table(train2$ScreenPorch) # very centered in 0 -> dont have screen porch  
analyze\_outliers(train2,"ScreenPorch")

## Number of mild outliers: 116



## Number of severe outliers: 116

train2 <- update\_outliers\_count(train2, "ScreenPorch")

### PoolArea - Pool area in square feet (cont)

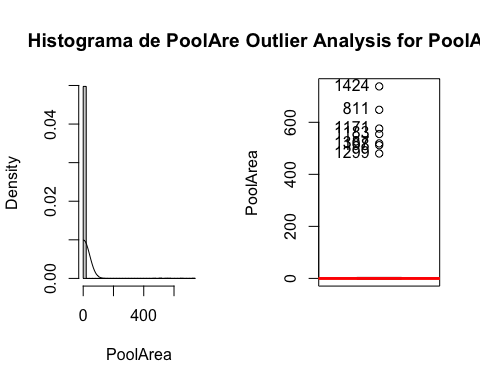
par(mfrow=c(1,2))  
numeric\_description(train2$PoolArea, 30)# not normally distributed

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 2.759 0.000 738.000   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.041202, p-value < 2.2e-16

#table(train2$PoolArea) # very centered in 0 -> dont have pool  
analyze\_outliers(train2,"PoolArea")

## Number of mild outliers: 7



## Number of severe outliers: 7

train2 <- update\_outliers\_count(train2, "PoolArea")  
  
# create a new variable : 0=no pool/1= pool  
train2$Pool\_binary<-0  
train2$Pool\_binary[which(train2$PoolArea!=0)]<-1  
train2$Pool\_binary<-as.factor(train2$Pool\_binary)  
table(train2$Pool\_binary) # not very usefull , the majority dont have pools, very centered

##   
## 0 1   
## 1453 7

#train2$Pool\_binary<-NULL

### MiscVal - $Value of miscellaneous feature (cont)

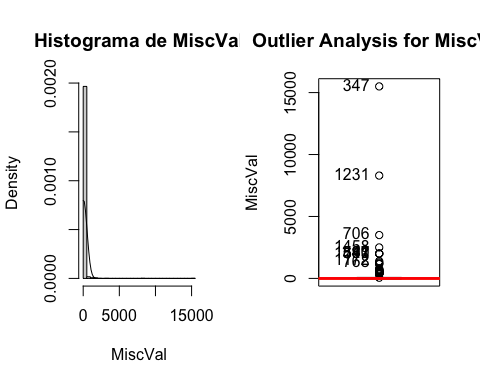
par(mfrow=c(1,2))  
numeric\_description(train2$MiscVal, 30) # not normally distributed

## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 43.49 0.00 15500.00   
##   
## Count of missing values: 0

##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.058233, p-value < 2.2e-16

#table(train2$MiscVal) # very centered in 0   
analyze\_outliers(train2,"MiscVal")

## Number of mild outliers: 52



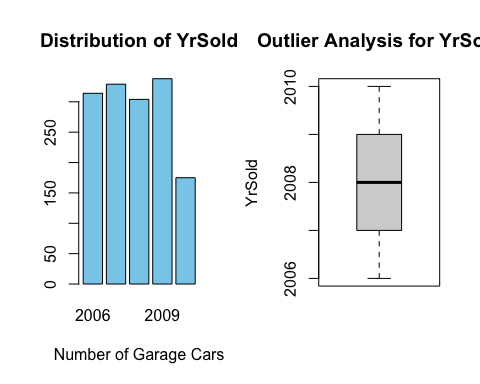
## Number of severe outliers: 52

train2 <- update\_outliers\_count(train2, "MiscVal")

### YrSold - (int)

par(mfrow=c(1,2))  
barplot(table(train2$YrSold), main = "Distribution of YrSold",xlab = "Number of Garage Cars",col = "skyblue")  
analyze\_outliers(train2,"YrSold")

## Number of mild outliers: 0



## Number of severe outliers: 0

train2 <- update\_outliers\_count(train2, "YrSold")

## Categorical Variables

Categorical variables are analyzed.

### Neighborhood

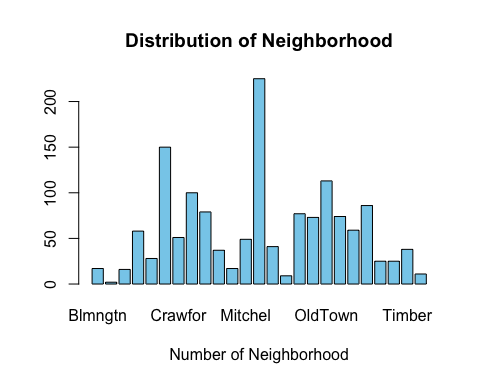
sum(is.na(train2$Neighborhood)) #0

## [1] 0

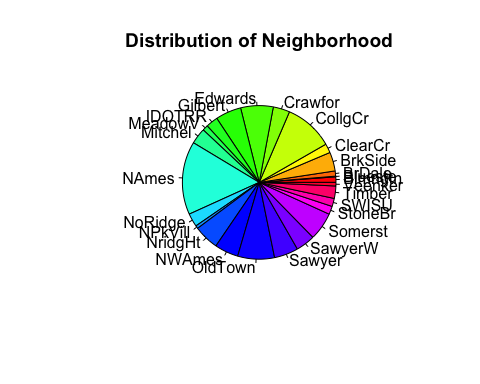
table(train2$Neighborhood)

##   
## Blmngtn Blueste BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert IDOTRR   
## 17 2 16 58 28 150 51 100 79 37   
## MeadowV Mitchel NAmes NoRidge NPkVill NridgHt NWAmes OldTown Sawyer SawyerW   
## 17 49 225 41 9 77 73 113 74 59   
## Somerst StoneBr SWISU Timber Veenker   
## 86 25 25 38 11

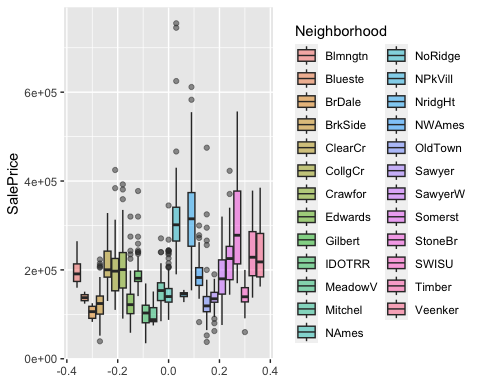
barplot(table(train2$Neighborhood), main = "Distribution of Neighborhood",xlab = "Number of Neighborhood",col = "skyblue")



pie(table(train2$Neighborhood), main = "Distribution of Neighborhood", col = rainbow(length(levels(train2$Neighborhood))))



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=Neighborhood)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=Neighborhood)) +  
 geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$Neighborhood)

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

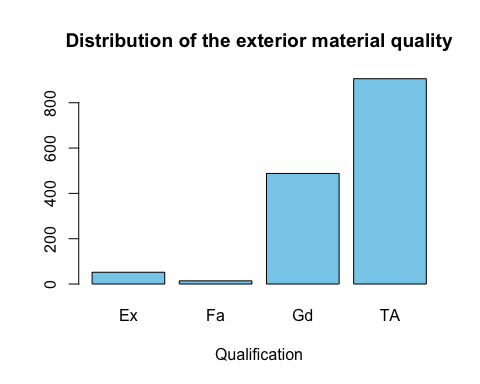
sum(is.na(train2$ExterQual)) #0

## [1] 0

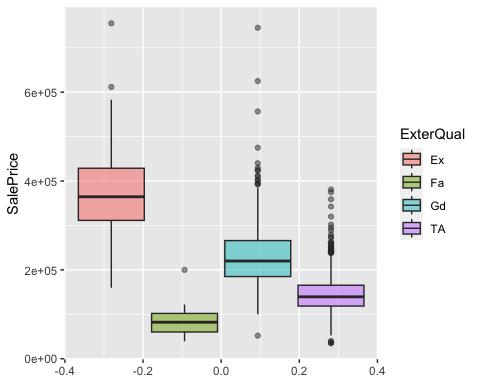
table(train2$ExterQual)

##   
## Ex Fa Gd TA   
## 52 14 488 906

barplot(table(train2$ExterQual), main = "Distribution of the exterior material quality",xlab = "Qualification",col = "skyblue")



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=ExterQual)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=ExterQual)) +geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$ExterQual)

### BsmtQual - Evaluates the height of the basement

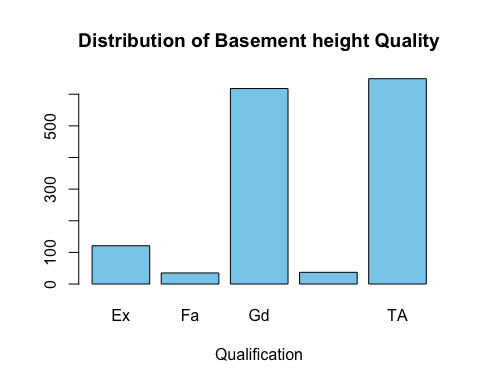
sum(is.na(train2$BsmtQual)) #0

## [1] 0

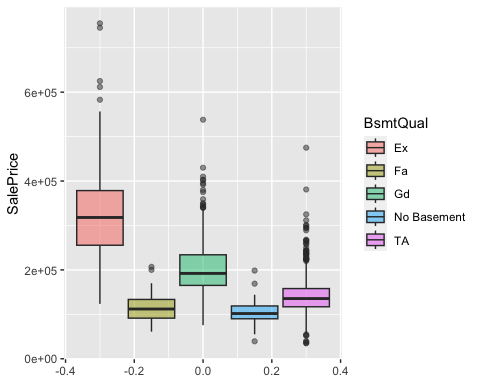
table(train2$BsmtQual)

##   
## Ex Fa Gd No Basement TA   
## 121 35 618 37 649

barplot(table(train2$BsmtQual), main = "Distribution of Basement height Quality",xlab = "Qualification",col = "skyblue")



#pie(table(train2$BsmtQual), main = "Distribution of BsmtQual", col = rainbow(length(levels(train2$BsmtQual))))  
  
# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=BsmtQual)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=BsmtQual)) + geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$BsmtQual)

### KitchenQual - Kitchen quality

sum(is.na(train2$KitchenQual)) #0

## [1] 0

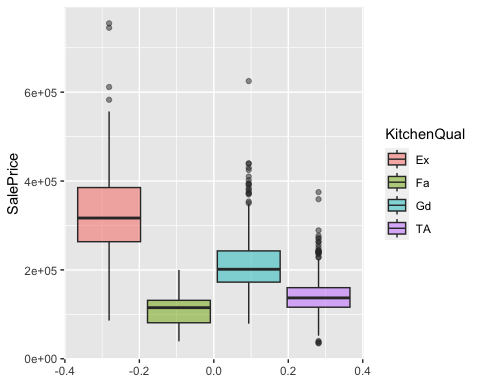
table(train2$KitchenQual)

##   
## Ex Fa Gd TA   
## 100 39 586 735

barplot(table(train2$KitchenQual), main = "Distribution of Kitchen Quality",xlab = "Qualification",col = "skyblue")



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=KitchenQual)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=KitchenQual)) + geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$KitchenQual)  
#ks.test(train2$SalePrice[train2$KitchenQual=="TA"],train2$SalePrice[train2$KitchenQual=="Fa"]) # test if we can join certain categoris together to make a easier analysis ->rejected

### GarageFinish - Interior finish of the garage

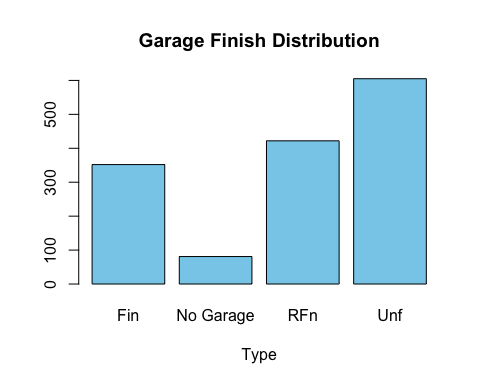
sum(is.na(train2$GarageFinish)) #0

## [1] 0

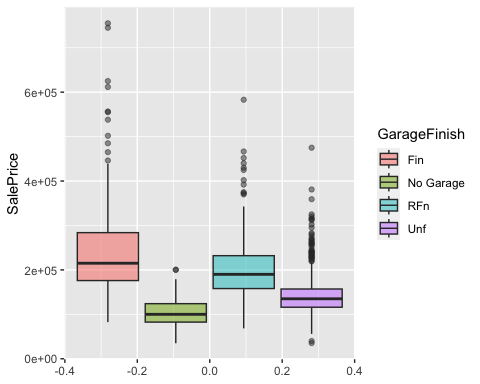
table(train2$GarageFinish)

##   
## Fin No Garage RFn Unf   
## 352 81 422 605

barplot(table(train2$GarageFinish), main = "Garage Finish Distribution",xlab = "Type",col = "skyblue")



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=GarageFinish)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=GarageFinish)) +  
 geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$GarageFinish)

### FireplaceQu - Fireplace quality

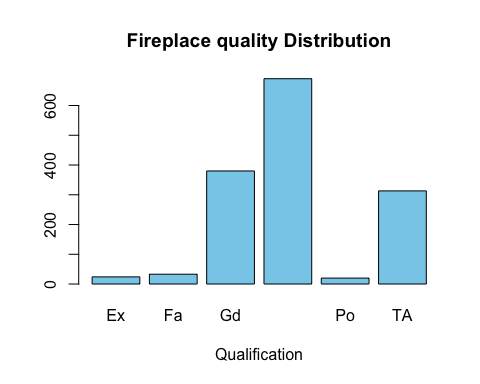
sum(is.na(train2$FireplaceQu)) #0

## [1] 0

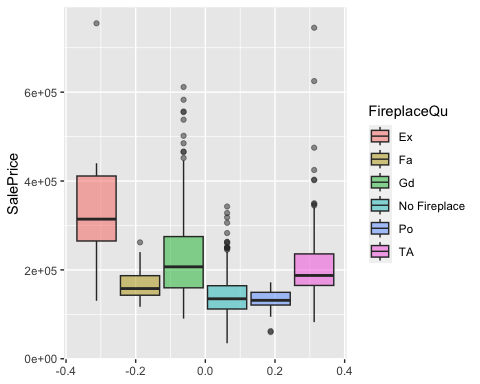
table(train2$FireplaceQu)

##   
## Ex Fa Gd No Fireplace Po TA   
## 24 33 380 690 20 313

barplot(table(train2$FireplaceQu), main = "Fireplace quality Distribution",xlab = "Qualification",col = "skyblue")



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=FireplaceQu)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=FireplaceQu)) +  
 geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$FireplaceQu)

### Foundation - Type of foundation

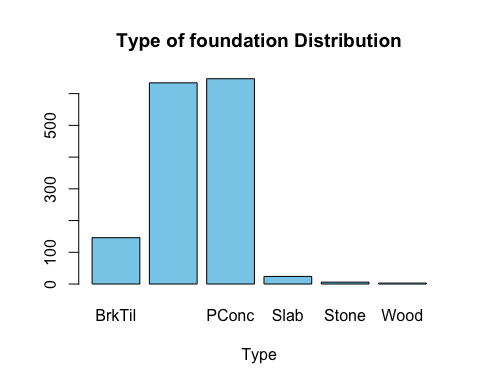
sum(is.na(train2$Foundation)) #0

## [1] 0

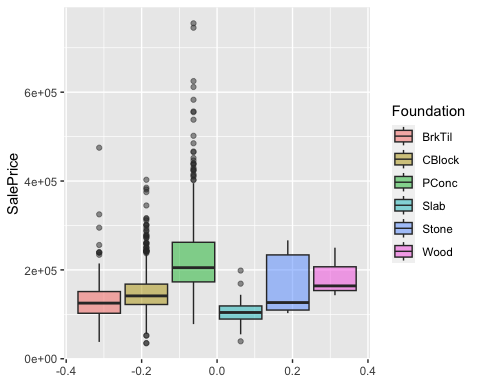
table(train2$Foundation)

##   
## BrkTil CBlock PConc Slab Stone Wood   
## 146 634 647 24 6 3

barplot(table(train2$Foundation), main = "Type of foundation Distribution",xlab = "Type",col = "skyblue")



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=Foundation)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=Foundation)) +  
 geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$Foundation)

### GarageType - Garage location

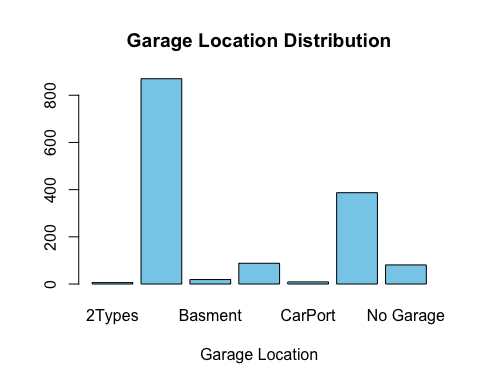
sum(is.na(train2$GarageType)) #0

## [1] 0

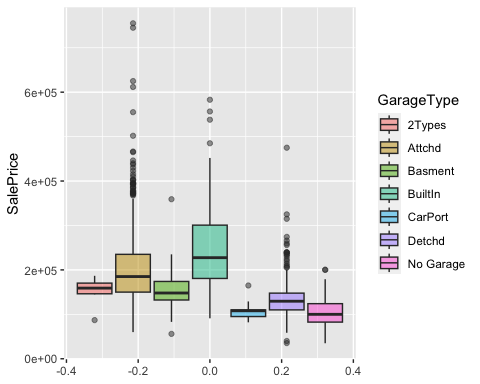
table(train2$GarageType)

##   
## 2Types Attchd Basment BuiltIn CarPort Detchd No Garage   
## 6 870 19 88 9 387 81

barplot(table(train2$GarageType), main = "Garage Location Distribution",xlab = "Garage Location",col = "skyblue")



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=GarageType)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=GarageType)) +  
 geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$GarageType)

### MSSubClass

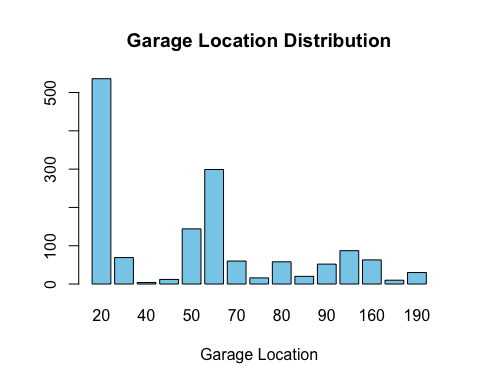
sum(is.na(train2$MSSubClass)) #0

## [1] 0

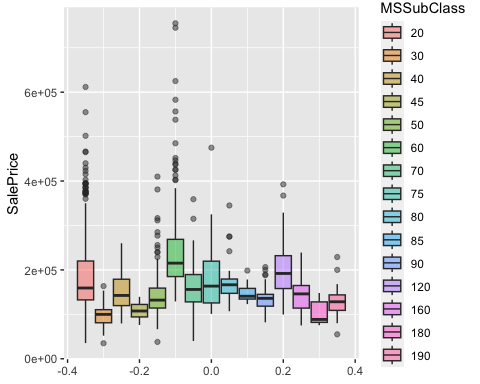
table(train2$MSSubClass) # +classes

##   
## 20 30 40 45 50 60 70 75 80 85 90 120 160 180 190   
## 536 69 4 12 144 299 60 16 58 20 52 87 63 10 30

barplot(table(train2$MSSubClass), main = "Garage Location Distribution",xlab = "Garage Location",col = "skyblue")



# analysis with the target variable  
#ggplot(train2, aes(x=SalePrice, fill=MSSubClass)) +geom\_density(alpha=.5)  
ggplot(train2, aes(y=SalePrice, fill=MSSubClass)) +  
 geom\_boxplot(alpha=0.5)



#pairwise.wilcox.test(train2$SalePrice, train2$MSSubClass)

## Missing values and imputation

Checking which variables have NA’s we can see that LotFrontage, MasVnrArea and GarageYrBlt are the ones with missing values.

#skim(train2)  
sum(is.na(train2$LotFrontage))

## [1] 259

sum(is.na(train2$MasVnrArea))

## [1] 10

sum(is.na(train2$GarageYrBlt))

## [1] 81

We decided to exclude the imputation of the GarageYrBlt feature because it has no sense to impute this variable, as it is a year.

mice\_imp<-mice(train2[, !names(train2) %in% "GarageYrBlt"],method = "cart")

##   
## iter imp variable  
## 1 1 LotFrontage MasVnrArea  
## 1 2 LotFrontage MasVnrArea  
## 1 3 LotFrontage MasVnrArea  
## 1 4 LotFrontage MasVnrArea  
## 1 5 LotFrontage MasVnrArea  
## 2 1 LotFrontage MasVnrArea  
## 2 2 LotFrontage MasVnrArea  
## 2 3 LotFrontage MasVnrArea  
## 2 4 LotFrontage MasVnrArea  
## 2 5 LotFrontage MasVnrArea  
## 3 1 LotFrontage MasVnrArea  
## 3 2 LotFrontage MasVnrArea  
## 3 3 LotFrontage MasVnrArea  
## 3 4 LotFrontage MasVnrArea  
## 3 5 LotFrontage MasVnrArea  
## 4 1 LotFrontage MasVnrArea  
## 4 2 LotFrontage MasVnrArea  
## 4 3 LotFrontage MasVnrArea  
## 4 4 LotFrontage MasVnrArea  
## 4 5 LotFrontage MasVnrArea  
## 5 1 LotFrontage MasVnrArea  
## 5 2 LotFrontage MasVnrArea  
## 5 3 LotFrontage MasVnrArea  
## 5 4 LotFrontage MasVnrArea  
## 5 5 LotFrontage MasVnrArea

## Warning: Number of logged events: 50

imputed\_data<-complete(mice\_imp)

We proceed to validate the imputation.

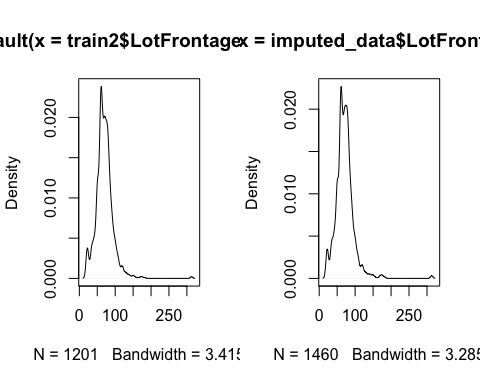
# LotFrontage validation  
summary(imputed\_data$LotFrontage)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.00 60.00 70.00 71.14 81.00 313.00

summary(train2$LotFrontage)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 21.00 59.00 69.00 70.05 80.00 313.00 259

par(mfrow=c(1,2))  
plot(density(train2$LotFrontage,na.rm=TRUE))  
plot(density(imputed\_data$LotFrontage,na.rm=TRUE))



par(mfrow=c(1,1))

Checking the results we can conclude that the imputation doesn’t have a significant change in the density nor the summary.

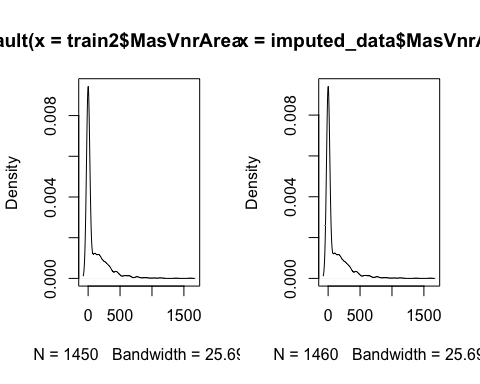
# MasVnrArea validation  
summary(imputed\_data$MasVnrArea)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 103.7 164.2 1600.0

summary(train2$MasVnrArea)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 0.0 0.0 103.2 164.0 1600.0 10

par(mfrow=c(1,2))  
plot(density(train2$MasVnrArea,na.rm=TRUE))  
plot(density(imputed\_data$MasVnrArea,na.rm=TRUE))



par(mfrow=c(1,1))

# Data Quality Exploration

To check the quality of the dataset we have a count on each individual of the univariate outliers. We can see that there are 7 instances that have 4 or more univariate outliers, therefore are candidates to delete from the analysis.

max(imputed\_data$univ\_outl\_count)

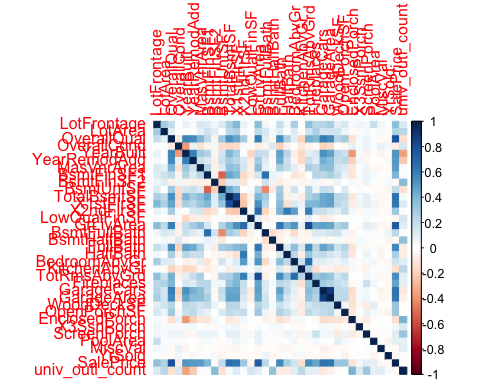
## [1] 10

ind\_delete<-imputed\_data[which(imputed\_data$univ\_outl\_count > 3),]  
imputed\_data<-imputed\_data[-c(ind\_delete)$Id,]

## Correlation Matrix

The correlation matrix shows a moderate negative correlation to the BsmtUnfSF (then the less Unfinished square feet of basement area, the more univ outliers) and a significant positive correlation to both BsmtFinSF2, Type 2 finished square feet, and ScreenPorch.

cols <- num\_cols[! num\_cols %in% c("Id","GarageYrBlt")]  
df\_of\_interest <- imputed\_data[,c(cols,"univ\_outl\_count")]  
cor\_outl = cor(df\_of\_interest)  
par(mfrow=c(1,1))  
corrplot(cor\_outl, method = 'color')



cor\_outl[34, ]

## LotFrontage LotArea OverallQual OverallCond YearBuilt   
## 0.35276847 0.26760836 0.79609375 -0.07666732 0.52597552   
## YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF   
## 0.50833651 0.49232756 0.40868965 -0.01296088 0.21579373   
## TotalBsmtSF X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea   
## 0.65096994 0.63122613 0.31988123 -0.03308338 0.74067275   
## BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr   
## 0.22762274 -0.01687411 0.56184539 0.28530765 0.16741336   
## KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea   
## -0.13877110 0.54284615 0.46954556 0.64249602 0.63038954   
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch   
## 0.32675949 0.32458454 -0.13663365 0.04469333 0.10961720   
## PoolArea MiscVal YrSold SalePrice univ\_outl\_count   
## 0.09827141 -0.02725834 -0.02770387 1.00000000 -0.01560585

colnames(cor\_outl)[which.min(cor\_outl[34, ])];min(cor\_outl[34, ])

## [1] "KitchenAbvGr"

## [1] -0.1387711

colnames(cor\_outl)[which.max(cor\_outl[34,1:33])];max(cor\_outl[34,1:33])

## [1] "OverallQual"

## [1] 0.7960937

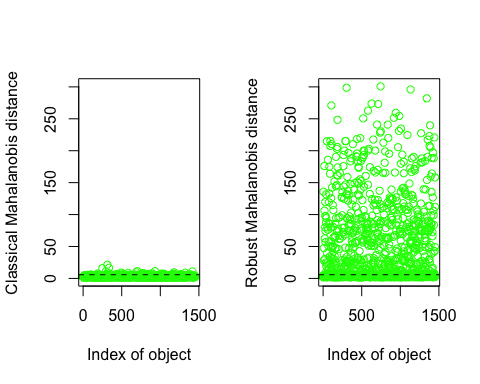
## Multivariate outliers

We filter variables in order to have only those with similar distribution of values. For that reason, we have filtered the values that had more unique values than 40. Then, we have excluded the variables that were centered in the ‘0’ value. We are not taking off these outliers because the length is 44 and taking into account the dimensions of our dataset, 1460, it would be a reduction of the 3%. Also, we have performed other Moutliers in order to see different scenarios that we consider rare data.

# 1. All variable with similar high unique values  
n\_cols <- names(imputed\_data)[sapply(imputed\_data, is.numeric)]  
n\_cols <- n\_cols[n\_cols != "Id"]  
df\_of\_interest <- imputed\_data[,c(n\_cols)]  
  
threshold <- 40  
df\_cols <- sapply(df\_of\_interest, function(x) length(unique(x)) > threshold)  
df\_of\_interest <- df\_of\_interest[, df\_cols]  
  
names(df\_of\_interest)

## [1] "LotFrontage" "LotArea" "YearBuilt" "YearRemodAdd"   
## [5] "MasVnrArea" "BsmtFinSF1" "BsmtFinSF2" "BsmtUnfSF"   
## [9] "TotalBsmtSF" "X1stFlrSF" "X2ndFlrSF" "GrLivArea"   
## [13] "GarageArea" "WoodDeckSF" "OpenPorchSF" "EnclosedPorch"  
## [17] "ScreenPorch" "SalePrice"

res.out = Moutlier(df\_of\_interest[, !(names(df\_of\_interest) %in% c("MasVnrArea","BsmtFinSF1", "BsmtFinSF2", "X2ndFlrSF", "WoodDeckSF", "EnclosedPorch", "ScreenPorch"))], quantile = 0.9995, col="green")



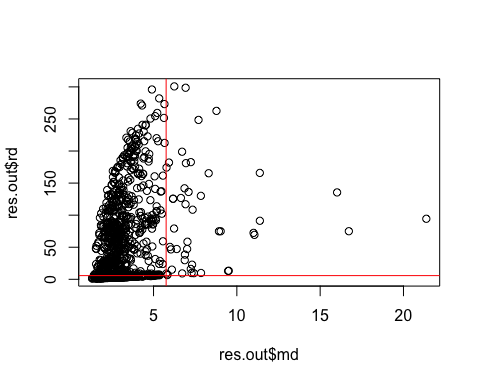
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 103 179 186 250 294 305 314 333 336 385 441 452 496 497 582 584   
## 103 179 186 249 293 304 313 332 335 384 440 451 495 496 579 581   
## 609 665 667 692 707 748 770 779 804 808 855 869 895 899 935 1012   
## 606 661 663 688 703 744 766 775 800 804 850 864 890 894 930 1007   
## 1025 1046 1047 1049 1062 1170 1174 1183 1185 1191 1269 1293 1329 1397 1424   
## 1020 1041 1042 1044 1057 1165 1169 1178 1180 1186 1263 1287 1322 1389 1416

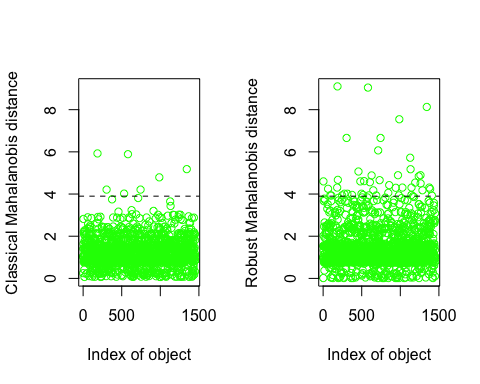
length(which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))) #44

## [1] 47

par(mfrow=c(1,1))  
plot( res.out$md, res.out$rd )  
abline(h=res.out$cutoff, col="red")  
abline(v=res.out$cutoff, col="red")



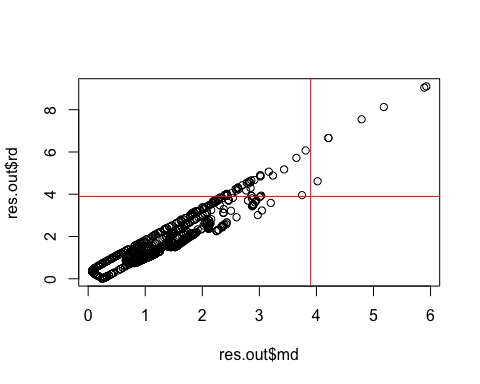
# take off m outliers  
#imputed\_data <- imputed\_data[-which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff)),]  
  
# 2. Generally the houses built latetly are the ones with most quality  
res.out <- Moutlier(imputed\_data[,c("YearBuilt","OverallQual")], quantile = 0.9995, col="green")



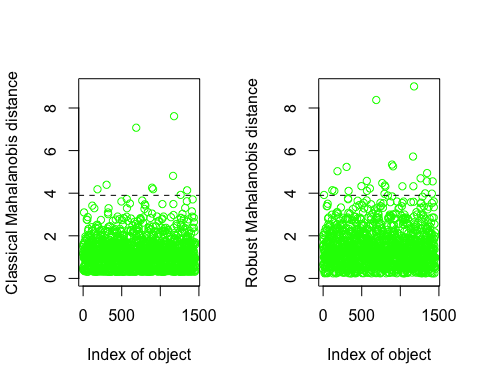
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 186 305 534 584 748 992 1350   
## 186 304 531 581 744 987 1343

par(mfrow=c(1,1))  
plot( res.out$md, res.out$rd )  
abline(h=res.out$cutoff, col="red")  
abline(v=res.out$cutoff, col="red")



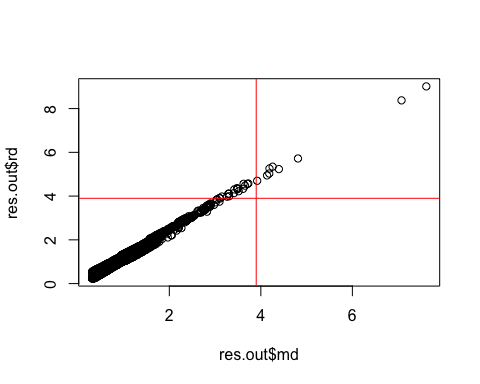
# 3. Generally the houses with more area are the ones with more rooms  
res.out <- Moutlier(imputed\_data[,c("GrLivArea","TotRmsAbvGrd")], quantile = 0.9995, col="green")



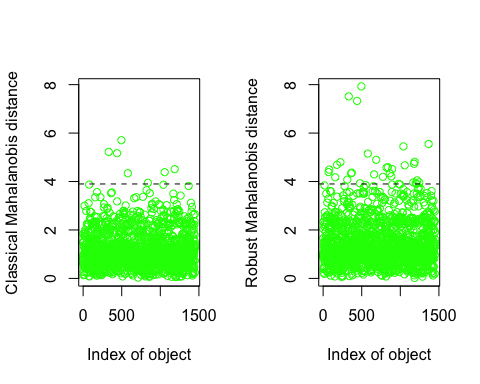
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 186 305 692 898 911 1170 1183 1269 1354   
## 186 304 688 893 906 1165 1178 1263 1347

par(mfrow=c(1,1))  
plot( res.out$md, res.out$rd )  
abline(h=res.out$cutoff, col="red")  
abline(v=res.out$cutoff, col="red")



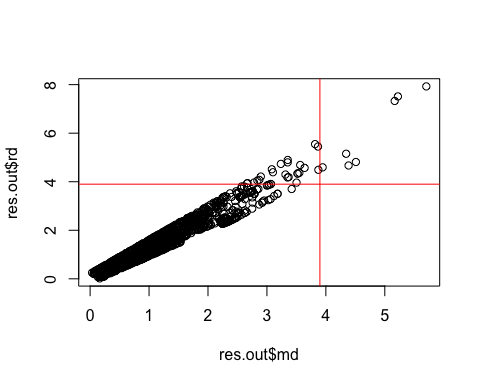
# 4. Generally the houses with more total area are the ones with more garage area also  
res.out <- Moutlier(imputed\_data[,c("GarageArea","TotalBsmtSF")], quantile = 0.9995, col="green")



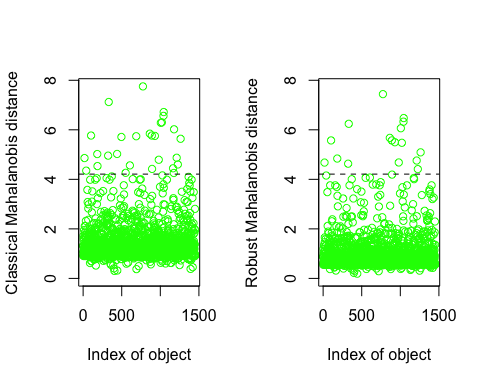
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 333 441 497 582 844 1062 1191   
## 332 440 496 579 839 1057 1186

par(mfrow=c(1,1))  
plot( res.out$md, res.out$rd )  
abline(h=res.out$cutoff, col="red")  
abline(v=res.out$cutoff, col="red")



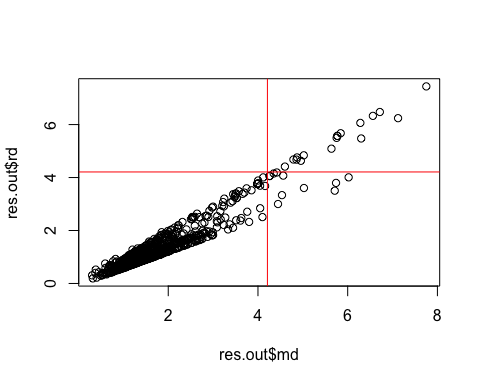
# 5. Moutliers about total areas zones  
res.out <- Moutlier(imputed\_data[,c("X1stFlrSF","TotalBsmtSF","GrLivArea")], quantile = 0.9995, col="green") #GarageArea, si se lo sumo hay menos outliers



which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 18 103 183 325 333 779 869 895 935 985 1012 1025 1046 1049 1217 1233   
## 18 103 183 324 332 775 864 890 930 980 1007 1020 1041 1044 1212 1227   
## 1269   
## 1263

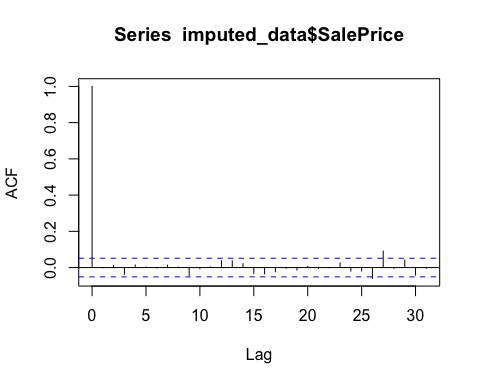
par(mfrow=c(1,1))  
plot( res.out$md, res.out$rd )  
abline(h=res.out$cutoff, col="red")  
abline(v=res.out$cutoff, col="red")



# Profiling

As we have declared before in the target variable exploration, the distribution is not normal. To test for autocorrelation the acf() function is used.

acf(imputed\_data$SalePrice)



#duplicate of the section target exploration  
shapiro.test(imputed\_data$SalePrice) ## Normality test of the target variable ->>> # We test the null hypothesis H0: variable is normal, H1: variable not normal. # The p-value is less than 0.05, therefore we reject H0. # For that we will not consider the target variable as normally distributed

##   
## Shapiro-Wilk normality test  
##   
## data: imputed\_data$SalePrice  
## W = 0.86922, p-value < 2.2e-16

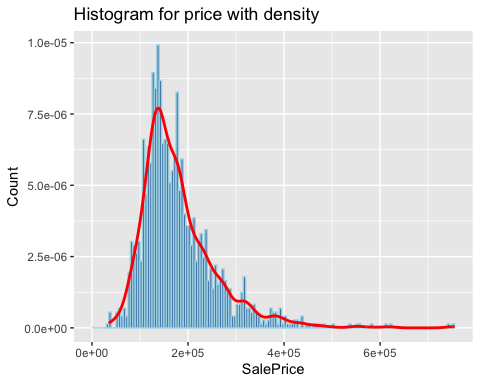
ks.test(imputed\_data$SalePrice, 'pnorm', mean(imputed\_data$SalePrice), sd(imputed\_data$SalePrice))

## Warning in ks.test.default(imputed\_data$SalePrice, "pnorm",  
## mean(imputed\_data$SalePrice), : ties should not be present for the  
## Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: imputed\_data$SalePrice  
## D = 0.12376, p-value < 2.2e-16  
## alternative hypothesis: two-sided

ggplot(data=imputed\_data, aes(SalePrice, y = after\_stat(density))) +  
 geom\_histogram(breaks=seq(0, max(imputed\_data$SalePrice), by=5000),  
 col='lightblue',  
 fill='steelblue') +  
 geom\_density(lwd=1,  
 col='red') +  
 labs(title="Histogram for price with density", x="SalePrice", y="Count")

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



## Variables with most association with the response variable

For the quantitative variables there are some variable that are highly significant positively correlated (r > 0.50, p ~ 0) to the target variable, SalesPrice: OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, X1stFlrSF, FullBath, TotRmsAbvGrd, YearBuilt and YearRemodAdd. This seems logical as the higher the overall condition rating of the house is and the bigger the area above grade living, the more it would cost. There isn’t any variable which is highly significant negatively correlated to the SalePrice (r < -0.50 p = 0).

For the qualitative variables it is clear that the Neighborhood explains the most variance in the SalePrice variable (R^2 = 0.546, p ~ 0).  
The influence of the other qualitative variables (with R^2 > 0.3) are in order: Neighborhood, ExterQual, BsmtQual, KitchenQual, f.GarageArea, GarageFinish. EnclosedPorch\_binary, Secondfloor and Pool\_binary are poorly associated as they have a R^2-value under 4%.

res.con <- condes(imputed\_data,35) #names(imputed\_data)[35] - [1] "SalePrice"  
res.con$quanti

## correlation p.value  
## OverallQual 0.79609375 1.398670e-318  
## GrLivArea 0.74067275 1.025988e-252  
## TotalBsmtSF 0.65096994 8.783474e-176  
## GarageCars 0.64249602 7.623680e-170  
## X1stFlrSF 0.63122613 3.121909e-162  
## GarageArea 0.63038954 1.114177e-161  
## FullBath 0.56184539 1.375148e-121  
## TotRmsAbvGrd 0.54284615 4.622921e-112  
## YearBuilt 0.52597552 4.236239e-104  
## YearRemodAdd 0.50833651 3.052623e-96  
## MasVnrArea 0.49232756 1.688940e-89  
## Fireplaces 0.46954556 1.682802e-80  
## BsmtFinSF1 0.40868965 1.464343e-59  
## LotFrontage 0.35276847 8.450612e-44  
## WoodDeckSF 0.32675949 1.787547e-37  
## OpenPorchSF 0.32458454 5.676640e-37  
## X2ndFlrSF 0.31988123 6.688459e-36  
## HalfBath 0.28530765 1.346076e-28  
## LotArea 0.26760836 3.111931e-25  
## BsmtFullBath 0.22762274 1.611002e-18  
## BsmtUnfSF 0.21579373 9.260246e-17  
## BedroomAbvGr 0.16741336 1.370234e-10  
## ScreenPorch 0.10961720 2.839149e-05  
## PoolArea 0.09827141 1.764824e-04  
## OverallCond -0.07666732 3.464488e-03  
## EnclosedPorch -0.13663365 1.727619e-07  
## KitchenAbvGr -0.13877110 1.102034e-07

#done before for selecting  
res.con$quali

## R2 p.value  
## Neighborhood 0.54755148 1.595761e-225  
## ExterQual 0.48847936 3.501182e-210  
## BsmtQual 0.47147373 1.488792e-198  
## KitchenQual 0.46306298 6.028665e-195  
## GarageFinish 0.31013396 3.123779e-116  
## FireplaceQu 0.29575431 1.902525e-107  
## Foundation 0.25784442 4.504809e-91  
## GarageType 0.25257359 7.585307e-88  
## MSSubClass 0.24826775 4.225305e-79  
## BsmtFinType1 0.21246055 1.362078e-71  
## OpenPorch\_binary 0.17113708 4.009012e-61  
## HasPorch\_binary 0.09173377 3.479772e-32  
## EnclosedPorch\_binary 0.03445322 1.012072e-12  
## f.LotArea 0.03035151 2.004842e-10  
## secondfloor 0.01860096 1.819444e-07  
## Pool\_binary 0.01146507 4.344490e-05

# Modeling

To create a model, first we split the train data (imputed\_data) into train\_new and test\_new. We need the test\_new to validate the accuracy of the model with data that has the target variable and the model has not used, to check if the model has overfitting.

set.seed(123)  
rows <- sample(nrow(imputed\_data), .7 \* nrow(imputed\_data))  
train\_new <- imputed\_data[rows, ]  
test\_new <- imputed\_data[-rows, ]

Before building the model, we should first study the multicolinearity and check if there are any predictors that are highly correlated. We have seen that the previous variables are very correlated, GarageArea with GarageCars with 88% correlation, TotRmsAbvGrd and GrLivArea with 83% correlation, X1stFlrSF and TotalBsmtSF with 81% correlation and SalePrice with OverallQual with 79% what is indeed our target variable. We will not, for now, in the modeling process consider excluding some of the variables that are very correlated.

num\_cols <- names(train\_new)[sapply(train\_new, is.numeric)]  
num\_data<-train\_new[,c(num\_cols)]  
descrCor <- cor(num\_data)  
highCorr <- sum(abs(descrCor[upper.tri(descrCor)]) > .99)   
  
max\_cor\_value <-max(abs(descrCor[upper.tri(descrCor)])) # maximum correlated  
which(descrCor == max\_cor\_value, arr.ind = TRUE) # garageArea and garagecars with 88% correlation

## row col  
## GarageArea 26 25  
## GarageCars 25 26

vec <- as.vector(descrCor)  
#sort(unique(vec), decreasing = TRUE)  
sort(unique(vec), decreasing = TRUE)[3];which(descrCor == sort(unique(vec), decreasing = TRUE)[3], arr.ind = TRUE)

## [1] 0.8341302

## row col  
## TotRmsAbvGrd 23 16  
## GrLivArea 16 23

sort(unique(vec), decreasing = TRUE)[4];which(descrCor == sort(unique(vec), decreasing = TRUE)[4], arr.ind = TRUE)

## [1] 0.8080929

## row col  
## SalePrice 35 4  
## OverallQual 4 35

sort(unique(vec), decreasing = TRUE)[5];which(descrCor == sort(unique(vec), decreasing = TRUE)[5], arr.ind = TRUE)

## [1] 0.7947344

## row col  
## X1stFlrSF 13 12  
## TotalBsmtSF 12 13

## Build different models with their metrics

We start by building the model with a manual stepwise technique, we will check the R^2, RSS and AIC metrics to select the better model. We can start by adding the more correlated numerical variables to the target variable.

display(m0 <- lm(SalePrice ~ 1, data = train\_new))  
display(m0 <- lm(log(SalePrice) ~ 1, data = train\_new))  
display(m1 <- lm(log(SalePrice) ~ OverallQual , data = train\_new))  
#plot(m1)  
anova(m0,m1) # rss  
as.matrix(AIC(m0, m1))   
  
display(m2 <- lm(log(SalePrice) ~ OverallQual\*log(GrLivArea), data = train\_new)) # with interaction  
display(m2 <- lm(log(SalePrice) ~ OverallQual+log(GrLivArea), data = train\_new)) # without interaction  
#plot(m2)  
anova(m1,m2)  
as.matrix(AIC(m1, m2))  
  
display(m3 <- lm(log(SalePrice) ~OverallQual+log(GrLivArea)\*GarageCars , data = train\_new))  
display(m3 <- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars , data = train\_new))  
anova(m2,m3)   
as.matrix(AIC(m2, m3))   
summary(m3)  
  
display(m4 <- lm(log(SalePrice)~OverallQual+log(GrLivArea)+GarageCars+GarageArea , data = train\_new))  
anova(m3,m4)  
as.matrix(AIC(m3, m4))  
summary(m4)  
  
display(m5<- lm(log(SalePrice) ~OverallQual +log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF , data = train\_new))  
anova(m4,m5)  
summary(m5)   
as.matrix(AIC(m4, m5))   
  
display(m6<- lm(log(SalePrice) ~OverallQual +log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+ X1stFlrSF, data = train\_new))  
anova(m5,m6)  
summary(m6)   
as.matrix(AIC(m5, m6)) #does not improve  
  
display(m7<- lm(log(SalePrice) ~OverallQual +log(GrLivArea)+GarageCars+TotalBsmtSF+FullBath, data = train\_new))  
anova(m5,m7) # does not improve much  
summary(m7)   
as.matrix(AIC(m5, m7))   
  
display(m8<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+TotRmsAbvGrd, data = train\_new))  
anova(m5,m8) # not improve  
as.matrix(AIC(m7, m8))   
  
display(m9<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt, data = train\_new))  
anova(m5,m9)   
summary(m8)  
as.matrix(AIC(m8, m9)) # improves  
  
display(m10<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt+YearRemodAdd, data = train\_new))  
anova(m9,m10)   
summary(m10)  
as.matrix(AIC(m9, m10))   
  
display(m11<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt+YearRemodAdd+MasVnrArea, data = train\_new))  
anova(m10,m11)   
summary(m11)  
as.matrix(AIC(m10, m11)) # not much better  
  
  
display(m12<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces, data = train\_new))  
anova(m10,m12)   
summary(m12) # r^2 better  
as.matrix(AIC(m10, m12))   
  
  
display(m13<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1, data = train\_new))  
anova(m12,m13)   
summary(m13) # r^2 NOT better , rss and AIC yes -> keep  
as.matrix(AIC(m10, m13))  
  
display(m14<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1+LotFrontage, data = train\_new))  
anova(m13,m14) # R^2 bbetter -> keep  
summary(m14)  
as.matrix(AIC(m13, m14))  
  
display(m15<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1+LotFrontage+WoodDeckSF, data = train\_new))  
anova(m14,m15) # better  
summary(m15)   
as.matrix(AIC(m14, m15))  
  
# delete garagearea bv very correlated to garage  
display(m16<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1+LotFrontage+WoodDeckSF+OpenPorchSF, data = train\_new))  
anova(m15,m16) # not better  
summary(m16)   
as.matrix(AIC(m14, m16))  
  
display(m17<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1+LotFrontage+WoodDeckSF+X2ndFlrSF, data = train\_new))  
anova(m15,m17) # not better   
summary(m17)   
as.matrix(AIC(m15, m17))  
vif(m17)  
  
display(m18<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1+LotFrontage+WoodDeckSF+HalfBath, data = train\_new))  
anova(m15,m18) # not better   
summary(m18)   
as.matrix(AIC(m15, m18))  
vif(m18)

We stop here, the model we end up with is :

display(m15<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+GarageArea+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1+LotFrontage+WoodDeckSF, data = train\_new))

## lm(formula = log(SalePrice) ~ OverallQual + log(GrLivArea) +   
## GarageCars + GarageArea + TotalBsmtSF + YearBuilt + YearRemodAdd +   
## Fireplaces + BsmtFinSF1 + LotFrontage + WoodDeckSF, data = train\_new)  
## coef.est coef.se  
## (Intercept) 0.12 0.55   
## OverallQual 0.08 0.01   
## log(GrLivArea) 0.36 0.02   
## GarageCars 0.03 0.01   
## GarageArea 0.00 0.00   
## TotalBsmtSF 0.00 0.00   
## YearBuilt 0.00 0.00   
## YearRemodAdd 0.00 0.00   
## Fireplaces 0.05 0.01   
## BsmtFinSF1 0.00 0.00   
## LotFrontage 0.00 0.00   
## WoodDeckSF 0.00 0.00   
## ---  
## n = 1016, k = 12  
## residual sd = 0.14, R-Squared = 0.89

as.matrix(AIC(m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11, m12, m13, m14, m15))

## df AIC  
## m1 3 -177.4676  
## m2 4 -462.3334  
## m3 5 -594.1448  
## m4 6 -617.5179  
## m5 7 -785.4945  
## m6 8 -785.6116  
## m7 7 -783.1480  
## m8 8 -785.1551  
## m9 8 -906.5419  
## m10 9 -960.7879  
## m11 10 -965.0052  
## m12 10 -1022.0880  
## m13 11 -1118.0725  
## m14 12 -1159.0330  
## m15 13 -1170.0386

summary(m15)

##   
## Call:  
## lm(formula = log(SalePrice) ~ OverallQual + log(GrLivArea) +   
## GarageCars + GarageArea + TotalBsmtSF + YearBuilt + YearRemodAdd +   
## Fireplaces + BsmtFinSF1 + LotFrontage + WoodDeckSF, data = train\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.84277 -0.06602 0.00726 0.07612 0.46559   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.160e-01 5.475e-01 0.212 0.832205   
## OverallQual 8.255e-02 5.190e-03 15.906 < 2e-16 \*\*\*  
## log(GrLivArea) 3.573e-01 1.903e-02 18.777 < 2e-16 \*\*\*  
## GarageCars 2.594e-02 1.304e-02 1.990 0.046886 \*   
## GarageArea 1.218e-04 4.479e-05 2.718 0.006679 \*\*   
## TotalBsmtSF 1.144e-04 1.424e-05 8.032 2.68e-15 \*\*\*  
## YearBuilt 1.561e-03 1.990e-04 7.846 1.10e-14 \*\*\*  
## YearRemodAdd 2.686e-03 2.724e-04 9.863 < 2e-16 \*\*\*  
## Fireplaces 4.560e-02 7.863e-03 5.800 8.90e-09 \*\*\*  
## BsmtFinSF1 1.142e-04 1.152e-05 9.914 < 2e-16 \*\*\*  
## LotFrontage 1.189e-03 1.807e-04 6.585 7.35e-11 \*\*\*  
## WoodDeckSF 1.301e-04 3.618e-05 3.596 0.000338 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1351 on 1004 degrees of freedom  
## Multiple R-squared: 0.8857, Adjusted R-squared: 0.8845   
## F-statistic: 707.6 on 11 and 1004 DF, p-value: < 2.2e-16

summary(m15)$r.squared

## [1] 0.885748

# check if there are any correlated predictors  
treshold\_vif<-function(rsq){  
 vif<- 1/(1-rsq)  
 return(vif)  
}  
treshold\_vif(summary(m15)$r.squared)

## [1] 8.752584

vif(m15) # there are not

## OverallQual log(GrLivArea) GarageCars GarageArea TotalBsmtSF   
## 2.936543 2.138552 5.507931 5.191456 1.846333   
## YearBuilt YearRemodAdd Fireplaces BsmtFinSF1 LotFrontage   
## 2.009515 1.745579 1.451028 1.378614 1.198423   
## WoodDeckSF   
## 1.137807

We can also use the automatic forward stepwise step() and see which are the variables that will create a model with higher accuracy.

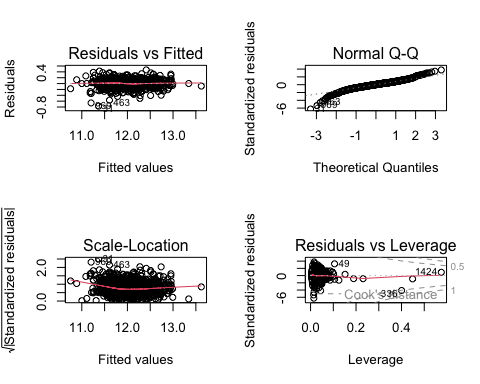
mod.fow <- stats::step(lm(SalePrice~. , data = train\_new), trace = FALSE,  
 direction = "forward")  
summary(mod.fow)

##   
## Call:  
## lm(formula = SalePrice ~ Id + LotFrontage + LotArea + OverallQual +   
## OverallCond + YearBuilt + YearRemodAdd + MasVnrArea + BsmtFinSF1 +   
## BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + X1stFlrSF + X2ndFlrSF +   
## LowQualFinSF + GrLivArea + BsmtFullBath + BsmtHalfBath +   
## FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd +   
## Fireplaces + GarageCars + GarageArea + WoodDeckSF + OpenPorchSF +   
## EnclosedPorch + X3SsnPorch + ScreenPorch + PoolArea + MiscVal +   
## YrSold + Neighborhood + ExterQual + BsmtQual + KitchenQual +   
## GarageFinish + FireplaceQu + Foundation + GarageType + MSSubClass +   
## BsmtFinType1 + univ\_outl\_count + f.LotArea + secondfloor +   
## OpenPorch\_binary + EnclosedPorch\_binary + HasPorch\_binary +   
## Pool\_binary, data = train\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -90295 -10985 587 10404 167892   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.150e+06 1.233e+06 1.743 0.081655 .   
## Id 1.393e+00 1.875e+00 0.743 0.457470   
## LotFrontage 8.663e+01 4.297e+01 2.016 0.044072 \*   
## LotArea 5.710e-01 1.435e-01 3.978 7.49e-05 \*\*\*  
## OverallQual 8.527e+03 1.182e+03 7.216 1.13e-12 \*\*\*  
## OverallCond 6.507e+03 9.297e+02 6.999 5.01e-12 \*\*\*  
## YearBuilt 4.399e+02 9.682e+01 4.543 6.29e-06 \*\*\*  
## YearRemodAdd 9.810e+01 6.133e+01 1.600 0.110028   
## MasVnrArea 2.475e+01 5.601e+00 4.418 1.11e-05 \*\*\*  
## BsmtFinSF1 4.410e+01 5.904e+00 7.468 1.91e-13 \*\*\*  
## BsmtFinSF2 3.795e+01 8.444e+00 4.494 7.89e-06 \*\*\*  
## BsmtUnfSF 2.815e+01 5.403e+00 5.209 2.35e-07 \*\*\*  
## TotalBsmtSF NA NA NA NA   
## X1stFlrSF 4.181e+01 6.246e+00 6.693 3.82e-11 \*\*\*  
## X2ndFlrSF 7.664e+01 7.191e+00 10.658 < 2e-16 \*\*\*  
## LowQualFinSF 1.607e+01 1.994e+01 0.806 0.420660   
## GrLivArea NA NA NA NA   
## BsmtFullBath 1.462e+03 2.358e+03 0.620 0.535244   
## BsmtHalfBath 6.264e+03 4.405e+03 1.422 0.155314   
## FullBath 5.851e+03 2.637e+03 2.219 0.026761 \*   
## HalfBath 3.405e+03 2.486e+03 1.370 0.171151   
## BedroomAbvGr -5.327e+03 1.641e+03 -3.246 0.001214 \*\*   
## KitchenAbvGr -7.138e+03 6.655e+03 -1.072 0.283803   
## TotRmsAbvGrd 3.378e+02 1.164e+03 0.290 0.771769   
## Fireplaces 5.585e+03 3.034e+03 1.841 0.066002 .   
## GarageCars 6.529e+03 2.588e+03 2.523 0.011818 \*   
## GarageArea 6.965e+00 8.881e+00 0.784 0.433056   
## WoodDeckSF 1.793e+01 6.976e+00 2.570 0.010330 \*   
## OpenPorchSF 3.012e+01 1.797e+01 1.676 0.094010 .   
## EnclosedPorch -1.803e+01 3.072e+01 -0.587 0.557359   
## X3SsnPorch 4.383e+01 2.476e+01 1.771 0.076979 .   
## ScreenPorch 6.795e+01 1.897e+01 3.582 0.000360 \*\*\*  
## PoolArea 2.272e+02 4.145e+02 0.548 0.583802   
## MiscVal 6.187e+00 6.154e+00 1.005 0.314975   
## YrSold -1.578e+03 6.065e+02 -2.601 0.009443 \*\*   
## NeighborhoodBlueste -1.321e+03 2.641e+04 -0.050 0.960125   
## NeighborhoodBrDale -1.435e+04 1.268e+04 -1.132 0.258120   
## NeighborhoodBrkSide -1.410e+04 1.066e+04 -1.323 0.186234   
## NeighborhoodClearCr -3.238e+04 1.106e+04 -2.927 0.003508 \*\*   
## NeighborhoodCollgCr -2.362e+04 9.228e+03 -2.560 0.010629 \*   
## NeighborhoodCrawfor 3.302e+03 1.030e+04 0.321 0.748539   
## NeighborhoodEdwards -2.767e+04 9.890e+03 -2.798 0.005252 \*\*   
## NeighborhoodGilbert -2.438e+04 9.528e+03 -2.559 0.010650 \*   
## NeighborhoodIDOTRR -2.596e+04 1.130e+04 -2.296 0.021897 \*   
## NeighborhoodMeadowV -7.851e+03 1.283e+04 -0.612 0.540675   
## NeighborhoodMitchel -3.529e+04 1.018e+04 -3.466 0.000553 \*\*\*  
## NeighborhoodNAmes -2.510e+04 9.612e+03 -2.612 0.009164 \*\*   
## NeighborhoodNoRidge 2.063e+03 1.046e+04 0.197 0.843660   
## NeighborhoodNPkVill 1.035e+03 1.318e+04 0.079 0.937439   
## NeighborhoodNridgHt -6.858e+03 9.533e+03 -0.719 0.472068   
## NeighborhoodNWAmes -3.300e+04 9.843e+03 -3.352 0.000834 \*\*\*  
## NeighborhoodOldTown -2.933e+04 1.025e+04 -2.860 0.004331 \*\*   
## NeighborhoodSawyer -2.501e+04 9.957e+03 -2.512 0.012180 \*   
## NeighborhoodSawyerW -2.248e+04 9.713e+03 -2.314 0.020881 \*   
## NeighborhoodSomerst -9.199e+03 9.600e+03 -0.958 0.338219   
## NeighborhoodStoneBr 2.254e+04 1.031e+04 2.187 0.028975 \*   
## NeighborhoodSWISU -2.530e+04 1.159e+04 -2.183 0.029298 \*   
## NeighborhoodTimber -2.455e+04 1.007e+04 -2.437 0.015009 \*   
## NeighborhoodVeenker 3.020e+03 1.302e+04 0.232 0.816615   
## ExterQualFa -1.321e+04 1.115e+04 -1.184 0.236556   
## ExterQualGd -2.078e+04 5.430e+03 -3.827 0.000139 \*\*\*  
## ExterQualTA -2.156e+04 6.065e+03 -3.555 0.000397 \*\*\*  
## BsmtQualFa -2.785e+04 7.279e+03 -3.826 0.000139 \*\*\*  
## BsmtQualGd -2.855e+04 4.204e+03 -6.792 2.00e-11 \*\*\*  
## BsmtQualNo Basement 9.985e+02 1.195e+04 0.084 0.933452   
## BsmtQualTA -2.605e+04 5.118e+03 -5.090 4.35e-07 \*\*\*  
## KitchenQualFa -1.782e+04 7.519e+03 -2.370 0.018000 \*   
## KitchenQualGd -1.433e+04 4.086e+03 -3.508 0.000473 \*\*\*  
## KitchenQualTA -1.588e+04 4.605e+03 -3.448 0.000591 \*\*\*  
## GarageFinishNo Garage 2.569e+04 1.432e+04 1.794 0.073159 .   
## GarageFinishRFn -2.321e+03 2.435e+03 -0.953 0.340801   
## GarageFinishUnf 1.736e+02 2.954e+03 0.059 0.953145   
## FireplaceQuFa -1.683e+04 8.140e+03 -2.068 0.038915 \*   
## FireplaceQuGd -1.074e+04 6.306e+03 -1.703 0.088930 .   
## FireplaceQuNo Fireplace -4.854e+03 7.647e+03 -0.635 0.525734   
## FireplaceQuPo -1.039e+04 9.048e+03 -1.148 0.251286   
## FireplaceQuTA -1.344e+04 6.562e+03 -2.048 0.040859 \*   
## FoundationCBlock 4.954e+02 3.724e+03 0.133 0.894192   
## FoundationPConc 3.753e+03 4.234e+03 0.886 0.375624   
## FoundationSlab -1.548e+03 1.132e+04 -0.137 0.891306   
## FoundationStone 1.746e+04 1.480e+04 1.180 0.238344   
## FoundationWood -2.858e+04 1.563e+04 -1.829 0.067762 .   
## GarageTypeAttchd 1.333e+04 1.311e+04 1.016 0.309681   
## GarageTypeBasment 2.315e+04 1.513e+04 1.530 0.126382   
## GarageTypeBuiltIn 1.934e+04 1.371e+04 1.411 0.158627   
## GarageTypeCarPort -5.510e+03 1.606e+04 -0.343 0.731552   
## GarageTypeDetchd 1.527e+04 1.310e+04 1.166 0.243970   
## GarageTypeNo Garage NA NA NA NA   
## MSSubClass30 3.392e+03 5.562e+03 0.610 0.542087   
## MSSubClass40 1.673e+04 2.541e+04 0.658 0.510489   
## MSSubClass45 2.724e+02 9.758e+03 0.028 0.977739   
## MSSubClass50 1.199e+04 6.800e+03 1.763 0.078265 .   
## MSSubClass60 4.850e+02 6.219e+03 0.078 0.937855   
## MSSubClass70 4.770e+03 7.899e+03 0.604 0.546090   
## MSSubClass75 1.440e+04 1.130e+04 1.274 0.202940   
## MSSubClass80 3.258e+03 4.828e+03 0.675 0.499914   
## MSSubClass85 3.165e+03 7.252e+03 0.436 0.662636   
## MSSubClass90 -1.158e+04 6.799e+03 -1.704 0.088800 .   
## MSSubClass120 -2.204e+04 4.771e+03 -4.620 4.38e-06 \*\*\*  
## MSSubClass160 -2.402e+04 8.392e+03 -2.863 0.004299 \*\*   
## MSSubClass180 -2.170e+04 1.240e+04 -1.749 0.080551 .   
## MSSubClass190 2.522e+02 7.600e+03 0.033 0.973536   
## BsmtFinType1BLQ 3.935e+03 3.223e+03 1.221 0.222387   
## BsmtFinType1GLQ 3.420e+03 3.003e+03 1.139 0.255166   
## BsmtFinType1LwQ -4.070e+03 4.430e+03 -0.919 0.358447   
## BsmtFinType1No Basement NA NA NA NA   
## BsmtFinType1Rec -3.399e+03 3.417e+03 -0.995 0.320108   
## BsmtFinType1Unf -5.938e+01 3.342e+03 -0.018 0.985826   
## univ\_outl\_count -4.555e+03 2.563e+03 -1.777 0.075889 .   
## f.LotArea.L 9.738e+03 3.854e+03 2.527 0.011676 \*   
## secondfloor1 -2.577e+04 7.212e+03 -3.574 0.000371 \*\*\*  
## OpenPorch\_binary1 1.460e+03 5.189e+03 0.281 0.778437   
## EnclosedPorch\_binary1 1.360e+04 7.266e+03 1.871 0.061636 .   
## HasPorch\_binary1 -4.190e+03 5.288e+03 -0.792 0.428291   
## Pool\_binary1 -1.201e+05 2.861e+05 -0.420 0.674844   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23480 on 905 degrees of freedom  
## Multiple R-squared: 0.9202, Adjusted R-squared: 0.9105   
## F-statistic: 94.89 on 110 and 905 DF, p-value: < 2.2e-16

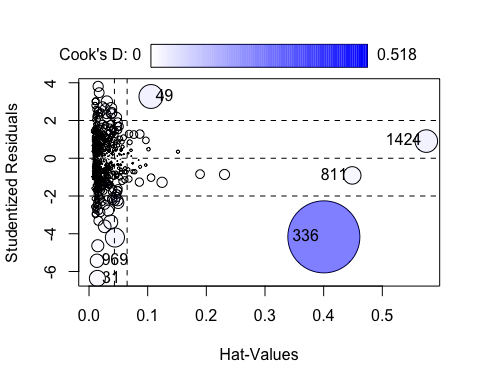
# Model:   
  
n00<- lm(log(SalePrice) ~ LotArea + OverallQual + OverallCond + YearBuilt +  
 MasVnrArea + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +  
 X1stFlrSF + X2ndFlrSF +  
 BsmtFullBath + FullBath + HalfBath + BedroomAbvGr +  
 KitchenAbvGr + Fireplaces + GarageCars +  
 OpenPorchSF + X3SsnPorch + ScreenPorch +  
 PoolArea, data = train\_new)  
#summary(n00)  
summary(n00)$r.squared

## [1] 0.9042354

par(mfrow=c(2,2))  
plot(n00)

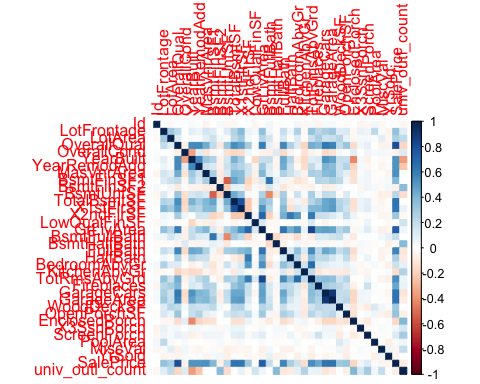


par(mfrow=c(1,1))  
influencePlot(n00)



## StudRes Hat CookD  
## 811 -0.9089625 0.44873948 0.03057613  
## 31 -6.3657646 0.01427203 0.02564917  
## 49 3.2717008 0.10554825 0.05685901  
## 336 -4.1646691 0.40032649 0.51779124  
## 969 -5.4309935 0.01346120 0.01778403  
## 1424 0.9089625 0.57499487 0.05081770

corrplot(descrCor, method = 'color')



# check if there are any correlated predictors  
treshold\_vif(summary(n00)$r.squared)

## [1] 10.44227

vif(n00) # there are not

## LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1   
## 1.203208 3.300425 1.252460 2.836137 1.430048 5.355830   
## BsmtFinSF2 BsmtUnfSF X1stFlrSF X2ndFlrSF BsmtFullBath FullBath   
## 1.532306 4.333847 4.479553 4.574664 1.980008 2.891431   
## HalfBath BedroomAbvGr KitchenAbvGr Fireplaces GarageCars OpenPorchSF   
## 2.246190 1.670376 1.286951 1.529900 1.967153 1.189330   
## X3SsnPorch ScreenPorch PoolArea   
## 1.014946 1.063377 1.019173

which(as.vector(vif(n00))>treshold\_vif(summary(n00)$r.squared))# there are not highly correlated variables in the model

## integer(0)

# delete with a backwards stepwise : MasVnrArea, FullBath, HalfBath, BedroomAbvGr,OpenPorchSF, X3SsnPorch  
n0<- lm(log(SalePrice) ~ LotArea + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+ X1stFlrSF+ X2ndFlrSF +  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea, data = train\_new)

Now we have 2 potential models, m15 created with mechanic stepwise and n0 with automatic stepwise, we shall check which is better.

as.matrix(AIC(m15,n0))

## df AIC  
## m15 13 -1170.039  
## n0 16 -1289.761

anova(m15,n0)

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ OverallQual + log(GrLivArea) + GarageCars +   
## GarageArea + TotalBsmtSF + YearBuilt + YearRemodAdd + Fireplaces +   
## BsmtFinSF1 + LotFrontage + WoodDeckSF  
## Model 2: log(SalePrice) ~ LotArea + OverallQual + OverallCond + YearBuilt +   
## BsmtFinSF1 + BsmtFinSF2 + X1stFlrSF + X2ndFlrSF + BsmtFullBath +   
## KitchenAbvGr + Fireplaces + GarageCars + ScreenPorch + PoolArea  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1004 18.330   
## 2 1001 16.197 3 2.1335 43.952 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(n0)$r.squared

## [1] 0.8990461

summary(m15)$r.squared

## [1] 0.885748

We will proceed with the n0 model.

## Plot to show which transformations we shall apply

# Check BoxTidwell tests:   
# We guess that the variables with a higher range will have a logaritmic transformation (the area variables)  
  
n0<- lm(log(SalePrice) ~ LotArea + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+ X1stFlrSF+ X2ndFlrSF +  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea, data = train\_new)  
  
# we do BoxTidwall test to see which transformations we shall apply in the predictors to make a better model.   
# LotArea  
boxTidwell(log(SalePrice) ~ LotArea, ~OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+ X1stFlrSF+ X2ndFlrSF +  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea,data=train\_new, max.iter = 70 )

## MLE of lambda Score Statistic (z) Pr(>|z|)   
## -0.3036 -6.7678 1.307e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 5

v0<- lm(log(SalePrice) ~ log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+ X1stFlrSF+ X2ndFlrSF +  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea, data = train\_new)  
# we shall convert the LotArea variable into logaritmic  
  
# X1stFlrSF  
boxTidwell(log(SalePrice) ~ X1stFlrSF,~ log(LotArea)+ OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+ X2ndFlrSF +  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea,data=train\_new, max.iter = 70 )

## MLE of lambda Score Statistic (z) Pr(>|z|)   
## 0.50049 -2.8813 0.00396 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 3

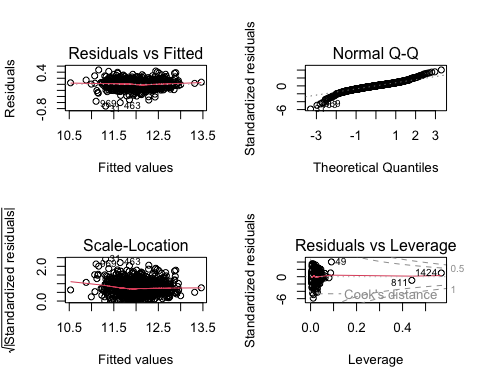
v1<- lm(log(SalePrice) ~ log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+ log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea, data = train\_new)  
  
  
# X2stFlrSF  
  
boxTidwell(log(SalePrice) ~ I(X2ndFlrSF+0.01),~ log(LotArea)+ OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+X1stFlrSF +  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea,data=train\_new, max.iter = 70 ) # we dont need

## MLE of lambda Score Statistic (z) Pr(>|z|)  
## 1.1024 1.0891 0.2761  
##   
## iterations = 2

#summary(v1)  
summary(v1)$r.squared

## [1] 0.904978

par(mfrow=c(2,2))  
plot(v1)



par(mfrow=c(1,1))  
#influencePlot(v1)  
which(as.vector(vif(v1))>treshold\_vif(summary(v1)$r.squared))# there are not highly correlated variables in the model

## integer(0)

## Select the best model

We have been looking on the accuracy of the model, currently on 90% (will depend on the seed), but we need to check if there is overfitting using the test sample. Looking at these results, seems that the model has no overfitting. No significant difference on the measures of the model between the train sample and the test.

rmse <- function(fitted, actual){  
 sqrt(mean((fitted - actual)^2))  
}  
  
RSQUARE <- function(predictions, actual\_values) {  
 ss\_residual <- sum((actual\_values - predictions)^2)  
 ss\_total <- sum((actual\_values - mean(actual\_values))^2)  
 rsquare <- 1 - (ss\_residual / ss\_total)  
 return(rsquare)  
}  
  
  
results <- data.frame(Model = c("Model with train\_new sample",  
 "Model with test\_new sample"),  
 RSQRT = round(c(RSQUARE(exp(fitted(v1)), train\_new$SalePrice),  
 RSQUARE(exp(predict(v1, newdata = test\_new)), test\_new$SalePrice)),2),  
 RMSE = round(c(rmse(exp(fitted(v1)), train\_new$SalePrice),  
 rmse(exp(predict(v1, newdata = test\_new)), test\_new$SalePrice)),2))  
results

## Model RSQRT RMSE  
## 1 Model with train\_new sample 0.91 23661.06  
## 2 Model with test\_new sample 0.90 25469.78

## Add categorical variables to the model

Now we shall proceed to adding the categorical variables for our model creation. We also will consider interactions if needed.

#names(train\_new)  
b0<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea + Neighborhood  
 , data = train\_new)  
#summary(b0)  
summary(b0)$r.squared

## [1] 0.9207796

anova(v1,b0)

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ log(LotArea) + OverallQual + OverallCond + YearBuilt +   
## BsmtFinSF1 + BsmtFinSF2 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## KitchenAbvGr + Fireplaces + GarageCars + ScreenPorch + PoolArea  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + Neighborhood  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1001 15.245   
## 2 975 12.710 26 2.5351 7.4799 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(v1,b0)

## df AIC  
## v1 16 -1351.285  
## b0 42 -1484.070

treshold\_vif(summary(b0)$r.squared)

## [1] 12.62301

vif(b0) # High vif-> Highly correlated predictor -> wont add to the model

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.851458 1 1.360683  
## log(LotArea) 2.895842 1 1.701717  
## OverallQual 3.645192 1 1.909239  
## OverallCond 1.346682 1 1.160466  
## YearBuilt 6.068686 1 2.463470  
## BsmtFinSF1 2.189619 1 1.479736  
## log(X1stFlrSF) 4.531546 1 2.128743  
## X2ndFlrSF 4.604495 1 2.145809  
## BsmtFullBath 1.897426 1 1.377471  
## FullBath 2.595548 1 1.611070  
## TotRmsAbvGrd 3.964616 1 1.991134  
## KitchenAbvGr 1.393114 1 1.180303  
## Fireplaces 1.683962 1 1.297676  
## GarageCars 2.134017 1 1.460828  
## ScreenPorch 1.091643 1 1.044817  
## PoolArea 1.099617 1 1.048626  
## Neighborhood 44.103779 24 1.082081

b1<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual  
 , data = train\_new)  
  
#summary(b1)  
summary(b1)$r.squared

## [1] 0.9091793

anova(n0,b1)

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotArea + OverallQual + OverallCond + YearBuilt +   
## BsmtFinSF1 + BsmtFinSF2 + X1stFlrSF + X2ndFlrSF + BsmtFullBath +   
## KitchenAbvGr + Fireplaces + GarageCars + ScreenPorch + PoolArea  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1001 16.197   
## 2 996 14.571 5 1.6257 22.226 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(n0,b1)

## df AIC  
## n0 16 -1289.761  
## b1 21 -1387.230

treshold\_vif(summary(b1)$r.squared)

## [1] 11.01071

vif(b1)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.777089 1 1.333075  
## log(LotArea) 1.863758 1 1.365195  
## OverallQual 3.559949 1 1.886783  
## OverallCond 1.292052 1 1.136685  
## YearBuilt 2.583860 1 1.607439  
## BsmtFinSF1 2.066596 1 1.437566  
## log(X1stFlrSF) 4.054252 1 2.013517  
## X2ndFlrSF 4.041876 1 2.010442  
## BsmtFullBath 1.824231 1 1.350641  
## FullBath 2.353554 1 1.534130  
## TotRmsAbvGrd 3.767768 1 1.941074  
## KitchenAbvGr 1.320928 1 1.149317  
## Fireplaces 1.489943 1 1.220632  
## GarageCars 1.989657 1 1.410552  
## ScreenPorch 1.061064 1 1.030079  
## PoolArea 1.081599 1 1.040000  
## ExterQual 2.945450 3 1.197270

# check interactions  
b11<- lm(log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea)\*ExterQual  
 , data = train\_new)  
#summary(b11)  
summary(b11)$r.squared

## [1] 0.9164772

anova(b1,b11) # the model is better with interactions but for explainability reasons we will not consider the interactions

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual  
## Model 2: log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea) \* ExterQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 996 14.571   
## 2 958 13.400 38 1.1708 2.2028 4.881e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b1,b11)

## df AIC  
## b1 21 -1387.230  
## b11 59 -1396.337

b2<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual  
 , data = train\_new)  
  
#summary(b2)  
summary(b2)$r.squared

## [1] 0.9137682

anova(b1,b2)

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 996 14.571   
## 2 992 13.835 4 0.73622 13.197 1.768e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b1,b2)

## df AIC  
## b1 21 -1387.230  
## b2 25 -1431.907

treshold\_vif(summary(b2)$r.squared)

## [1] 11.59665

vif(b2)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.788374 1 1.337301  
## log(LotArea) 1.874197 1 1.369013  
## OverallQual 3.827846 1 1.956488  
## OverallCond 1.312793 1 1.145772  
## YearBuilt 3.314812 1 1.820662  
## BsmtFinSF1 2.121379 1 1.456495  
## log(X1stFlrSF) 4.059837 1 2.014904  
## X2ndFlrSF 4.114908 1 2.028524  
## BsmtFullBath 1.838431 1 1.355888  
## FullBath 2.462803 1 1.569332  
## TotRmsAbvGrd 3.913099 1 1.978155  
## KitchenAbvGr 1.421952 1 1.192456  
## Fireplaces 1.494298 1 1.222415  
## GarageCars 2.020539 1 1.421457  
## ScreenPorch 1.064253 1 1.031627  
## PoolArea 1.082560 1 1.040461  
## ExterQual 3.825780 3 1.250605  
## BsmtQual 5.701989 4 1.243092

# interactions  
b21<- lm(log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual) \*BsmtQual  
 , data = train\_new)  
#summary(b21)  
summary(b21)$r.squared

## [1] 0.9208455

anova(b2,b21) # better with interaction

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual  
## Model 2: log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea + ExterQual) \* BsmtQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 992 13.835   
## 2 929 12.699 63 1.1355 1.3185 0.05309 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b2,b21)

## df AIC  
## b2 25 -1431.907  
## b21 88 -1392.916

anova(b11,b21) # but worse than b11

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea) \* ExterQual  
## Model 2: log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea + ExterQual) \* BsmtQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 958 13.400   
## 2 929 12.699 29 0.70085 1.7679 0.007728 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

b3<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual  
 , data = train\_new)  
  
#summary(b3)  
summary(b3)$r.squared

## [1] 0.9149776

anova(b2,b3)

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 992 13.835   
## 2 989 13.641 3 0.19403 4.6892 0.002938 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b2,b3)

## df AIC  
## b2 25 -1431.907  
## b3 28 -1440.257

treshold\_vif(summary(b3)$r.squared)

## [1] 11.7616

vif(b3)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.800847 1 1.341957  
## log(LotArea) 1.876842 1 1.369979  
## OverallQual 3.912830 1 1.978087  
## OverallCond 1.382920 1 1.175976  
## YearBuilt 3.397393 1 1.843202  
## BsmtFinSF1 2.130314 1 1.459559  
## log(X1stFlrSF) 4.109969 1 2.027306  
## X2ndFlrSF 4.151696 1 2.037571  
## BsmtFullBath 1.846843 1 1.358986  
## FullBath 2.499105 1 1.580856  
## TotRmsAbvGrd 3.935877 1 1.983904  
## KitchenAbvGr 1.441983 1 1.200826  
## Fireplaces 1.495654 1 1.222969  
## GarageCars 2.061623 1 1.435835  
## ScreenPorch 1.067600 1 1.033247  
## PoolArea 1.086731 1 1.042464  
## ExterQual 5.804914 3 1.340600  
## BsmtQual 6.297702 4 1.258629  
## KitchenQual 4.882951 3 1.302508

#with interactions  
b31<- lm(log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual)\* KitchenQual  
 , data = train\_new)  
  
#summary(b31)  
summary(b31)$r.squared

## [1] 0.923933

anova(b3,b31) # better model with interactions but we will not consider them for explainability purposes

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual  
## Model 2: log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea + ExterQual + BsmtQual) \*   
## KitchenQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 989 13.641   
## 2 931 12.204 58 1.4368 1.8898 0.0001032 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b3,b31)

## df AIC  
## b3 28 -1440.257  
## b31 86 -1437.339

b4<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+ GarageFinish  
 , data = train\_new)  
  
#summary(b4)  
summary(b4)$r.squared

## [1] 0.9152487

anova(b3,b4) # not better model -> we dont use this variable

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageFinish  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 989 13.641   
## 2 986 13.597 3 0.043506 1.0516 0.3689

AIC(b3,b4)

## df AIC  
## b3 28 -1440.257  
## b4 31 -1437.503

treshold\_vif(summary(b4)$r.squared)

## [1] 11.79923

vif(b4)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.816399 1 1.347738  
## log(LotArea) 1.880659 1 1.371371  
## OverallQual 3.925994 1 1.981412  
## OverallCond 1.386838 1 1.177641  
## YearBuilt 3.625750 1 1.904140  
## BsmtFinSF1 2.140705 1 1.463115  
## log(X1stFlrSF) 4.124205 1 2.030814  
## X2ndFlrSF 4.160748 1 2.039791  
## BsmtFullBath 1.851517 1 1.360704  
## FullBath 2.568238 1 1.602572  
## TotRmsAbvGrd 3.971066 1 1.992753  
## KitchenAbvGr 1.463308 1 1.209673  
## Fireplaces 1.524221 1 1.234593  
## GarageCars 3.031648 1 1.741163  
## ScreenPorch 1.069944 1 1.034381  
## PoolArea 1.088391 1 1.043260  
## ExterQual 6.006899 3 1.348264  
## BsmtQual 6.613522 4 1.266351  
## KitchenQual 4.974500 3 1.306547  
## GarageFinish 3.818648 3 1.250216

b5<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual + FireplaceQu  
 , data = train\_new)  
  
#summary(b5)  
summary(b5)$r.squared

## [1] 0.91581

anova(b3,b5) # not better model

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## FireplaceQu  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 989 13.641   
## 2 984 13.507 5 0.13355 1.9458 0.08432 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b3,b5)

## df AIC  
## b3 28 -1440.257  
## b5 33 -1440.253

treshold\_vif(summary(b5)$r.squared)

## [1] 11.87789

vif(b5)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.810956 1 1.345718  
## log(LotArea) 1.890971 1 1.375126  
## OverallQual 3.969698 1 1.992410  
## OverallCond 1.386168 1 1.177357  
## YearBuilt 3.460650 1 1.860282  
## BsmtFinSF1 2.161423 1 1.470178  
## log(X1stFlrSF) 4.154726 1 2.038314  
## X2ndFlrSF 4.245075 1 2.060358  
## BsmtFullBath 1.859216 1 1.363531  
## FullBath 2.517610 1 1.586698  
## TotRmsAbvGrd 3.992518 1 1.998129  
## KitchenAbvGr 1.457067 1 1.207090  
## Fireplaces 5.982984 1 2.446014  
## GarageCars 2.067517 1 1.437886  
## ScreenPorch 1.077556 1 1.038054  
## PoolArea 1.106797 1 1.052044  
## ExterQual 5.889637 3 1.343842  
## BsmtQual 6.682232 4 1.267988  
## KitchenQual 5.073361 3 1.310839  
## FireplaceQu 8.897276 5 1.244302

b6<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+Foundation  
 , data = train\_new)  
  
#summary(b6)  
summary(b6)$r.squared

## [1] 0.9165545

anova(b3,b6)

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## Foundation  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 989 13.641   
## 2 984 13.388 5 0.253 3.7191 0.002432 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b3,b6)

## df AIC  
## b3 28 -1440.257  
## b6 33 -1449.278

treshold\_vif(summary(b6)$r.squared)

## [1] 11.98387

vif(b6) # we dont add it because highly correlated

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.872340 1 1.368335  
## log(LotArea) 1.883598 1 1.372442  
## OverallQual 4.032696 1 2.008157  
## OverallCond 1.435297 1 1.198039  
## YearBuilt 5.501804 1 2.345592  
## BsmtFinSF1 2.168408 1 1.472551  
## log(X1stFlrSF) 4.171936 1 2.042532  
## X2ndFlrSF 4.213454 1 2.052670  
## BsmtFullBath 1.852599 1 1.361102  
## FullBath 2.533519 1 1.591703  
## TotRmsAbvGrd 3.957372 1 1.989314  
## KitchenAbvGr 1.444076 1 1.201697  
## Fireplaces 1.527015 1 1.235724  
## GarageCars 2.106745 1 1.451463  
## ScreenPorch 1.071350 1 1.035060  
## PoolArea 1.090082 1 1.044070  
## ExterQual 6.086642 3 1.351231  
## BsmtQual 19.243536 4 1.447224  
## KitchenQual 5.156556 3 1.314397  
## Foundation 17.400460 5 1.330626

b7<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+GarageType  
 , data = train\_new)  
  
#summary(b7)  
summary(b7)$r.squared

## [1] 0.9161309

anova(b3,b7)

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 989 13.641   
## 2 983 13.456 6 0.18503 2.2529 0.03638 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b3,b7)

## df AIC  
## b3 28 -1440.257  
## b7 34 -1442.134

treshold\_vif(summary(b7)$r.squared)

## [1] 11.92334

vif(b7)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.842124 1 1.357249  
## log(LotArea) 1.908711 1 1.381561  
## OverallQual 3.961637 1 1.990386  
## OverallCond 1.394138 1 1.180736  
## YearBuilt 3.989315 1 1.997327  
## BsmtFinSF1 2.147832 1 1.465548  
## log(X1stFlrSF) 4.212123 1 2.052346  
## X2ndFlrSF 4.363389 1 2.088873  
## BsmtFullBath 1.851576 1 1.360726  
## FullBath 2.543429 1 1.594813  
## TotRmsAbvGrd 3.972219 1 1.993043  
## KitchenAbvGr 1.508272 1 1.228117  
## Fireplaces 1.530400 1 1.237094  
## GarageCars 3.212864 1 1.792446  
## ScreenPorch 1.073345 1 1.036024  
## PoolArea 1.088088 1 1.043115  
## ExterQual 6.021147 3 1.348797  
## BsmtQual 6.669671 4 1.267690  
## KitchenQual 5.088363 3 1.311484  
## GarageType 4.825743 6 1.140154

# interactions   
b71<- lm(log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea)\*GarageType +ExterQual + BsmtQual + KitchenQual  
 , data = train\_new)  
  
#summary(b71)  
summary(b71)$r.squared

## [1] 0.9271972

anova(b7,b71) # better model

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType  
## Model 2: log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea) \* GarageType + ExterQual +   
## BsmtQual + KitchenQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 983 13.456   
## 2 919 11.680 64 1.7754 2.1827 6.518e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b7,b71)

## df AIC  
## b7 34 -1442.134  
## b71 98 -1457.900

treshold\_vif(summary(b71)$r.squared)

## [1] 13.73573

b8<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+GarageType+MSSubClass  
 , data = train\_new)  
  
#summary(b8)  
summary(b8)$r.squared

## [1] 0.9176482

anova(b7,b8) # not better model

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType + MSSubClass  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 983 13.456   
## 2 969 13.212 14 0.24344 1.2753 0.216

AIC(b7,b8)

## df AIC  
## b7 34 -1442.134  
## b8 48 -1432.683

treshold\_vif(summary(b8)$r.squared)

## [1] 12.14303

vif(b8) # highly correlated

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.946073 1 1.395017  
## log(LotArea) 2.814618 1 1.677682  
## OverallQual 4.210427 1 2.051933  
## OverallCond 1.452258 1 1.205097  
## YearBuilt 8.525699 1 2.919880  
## BsmtFinSF1 2.193624 1 1.481089  
## log(X1stFlrSF) 4.867109 1 2.206152  
## X2ndFlrSF 9.267584 1 3.044271  
## BsmtFullBath 1.924930 1 1.387418  
## FullBath 2.600169 1 1.612504  
## TotRmsAbvGrd 4.221822 1 2.054707  
## KitchenAbvGr 2.963685 1 1.721536  
## Fireplaces 1.607585 1 1.267906  
## GarageCars 3.291834 1 1.814341  
## ScreenPorch 1.083741 1 1.041029  
## PoolArea 1.106702 1 1.051999  
## ExterQual 6.402642 3 1.362678  
## BsmtQual 8.276830 4 1.302366  
## KitchenQual 5.439541 3 1.326153  
## GarageType 6.960219 6 1.175489  
## MSSubClass 402.965275 14 1.238926

b9<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+GarageType+BsmtFinType1  
 , data = train\_new)  
  
#summary(b9)  
summary(b9)$r.squared

## [1] 0.9165275

anova(b7,b9) # not better model

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType + BsmtFinType1  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 983 13.456   
## 2 978 13.392 5 0.063631 0.9294 0.461

AIC(b7,b9)

## df AIC  
## b7 34 -1442.134  
## b9 39 -1436.950

treshold\_vif(summary(b9)$r.squared)

## [1] 11.97999

#f.LotArea  
b10<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+GarageType+f.LotArea  
 , data = train\_new)  
  
#summary(b10)  
summary(b10)$r.squared

## [1] 0.9161417

anova(b7,b10) # not better model

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType  
## Model 2: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType + f.LotArea  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 983 13.456   
## 2 982 13.454 1 0.0017376 0.1268 0.7218

AIC(b7,b10)

## df AIC  
## b7 34 -1442.134  
## b10 35 -1440.265

treshold\_vif(summary(b10)$r.squared)

## [1] 11.92488

vif(b10)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.846397 1 1.358822  
## log(LotArea) 2.280272 1 1.510057  
## OverallQual 3.978483 1 1.994614  
## OverallCond 1.394490 1 1.180885  
## YearBuilt 3.990041 1 1.997509  
## BsmtFinSF1 2.156750 1 1.468588  
## log(X1stFlrSF) 4.218010 1 2.053779  
## X2ndFlrSF 4.388750 1 2.094934  
## BsmtFullBath 1.855848 1 1.362295  
## FullBath 2.558967 1 1.599677  
## TotRmsAbvGrd 4.004391 1 2.001098  
## KitchenAbvGr 1.508963 1 1.228399  
## Fireplaces 1.537156 1 1.239821  
## GarageCars 3.214239 1 1.792830  
## ScreenPorch 1.073376 1 1.036039  
## PoolArea 1.101554 1 1.049550  
## ExterQual 6.038510 3 1.349444  
## BsmtQual 6.681981 4 1.267982  
## KitchenQual 5.111433 3 1.312473  
## GarageType 4.852209 6 1.140674  
## f.LotArea 1.384129 1 1.176490

b12<- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+GarageType+HasPorch\_binary  
 , data = train\_new)  
  
#summary(b11)  
summary(b11)$r.squared

## [1] 0.9164772

anova(b7,b11) # a little better model, but not much, we better keep simpler model

## Analysis of Variance Table  
##   
## Model 1: log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType  
## Model 2: log(SalePrice) ~ (LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea) \* ExterQual  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 983 13.456   
## 2 958 13.400 25 0.055559 0.1589 1

AIC(b7,b11)

## df AIC  
## b7 34 -1442.134  
## b11 59 -1396.337

AIC(b7,b71,b3,b31,b2,b21,b1,b11)

## df AIC  
## b7 34 -1442.134  
## b71 98 -1457.900  
## b3 28 -1440.257  
## b31 86 -1437.339  
## b2 25 -1431.907  
## b21 88 -1392.916  
## b1 21 -1387.230  
## b11 59 -1396.337

final\_model<-b3

## Final model metrics

results <- data.frame(Model = c("Model with train\_new sample",  
 "Model with test\_new sample"),  
 RSQRT = round(c(RSQUARE(exp(fitted(final\_model)), train\_new$SalePrice),  
 RSQUARE(exp(predict(final\_model, newdata = test\_new)), test\_new$SalePrice)),2),  
 RMSE = round(c(rmse(exp(fitted(final\_model)), train\_new$SalePrice),  
 rmse(exp(predict(final\_model, newdata = test\_new)), test\_new$SalePrice)),2))  
results

## Model RSQRT RMSE  
## 1 Model with train\_new sample 0.92 21684.9  
## 2 Model with test\_new sample 0.92 23867.8

## Study the presence of a priori influential data observations

In this case, 26 a priori values where found.

mean\_hat <- mean(hatvalues(final\_model));mean\_hat

## [1] 0.0265748

priori <- which(hatvalues(final\_model)>4\*mean\_hat)  
length(priori)

## [1] 14

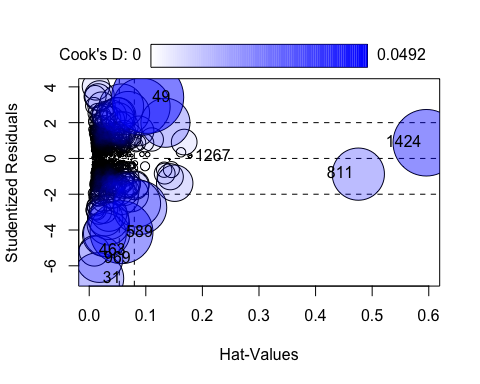
## Study the presence of a posteriori influential values

For studying the posteriori influential values, we have looked to cook’s distance outliers, where 3 influent data are found and are removed from the sample. The best model is then reconstructed on this new data set and compared with its original to see the difference. Comparing the previous coefficients and the current coefficients, we can say that the change has no relevant changes in the coefficients. The R-squared of the model has increased up to 0.94.

betas <- as.data.frame(dfbetas(final\_model))  
betas\_cutoff <- 2 / sqrt(dim(train\_new)[1]);betas\_cutoff

## [1] 0.06274558

influencePlot(final\_model, id=list(n=3, method="noteworthy"))



## StudRes Hat CookD  
## 811 -0.8773371 0.475563773 2.585750e-02  
## 31 -6.7047610 0.015501905 2.510085e-02  
## 49 3.4153958 0.103308543 4.924400e-02  
## 589 -4.1233678 0.057087322 3.751784e-02  
## 1267 0.1014985 0.178502316 8.299044e-05  
## 969 -5.6220666 0.017964124 2.077161e-02  
## 1424 0.8773371 0.595675580 4.200976e-02  
## 463 -5.1664335 0.009040735 8.790790e-03

llcoo <-c("1424","811","1183")  
# residual outliers  
llres <- which(abs(rstudent(final\_model))>qnorm(0.995));length(llres)

## [1] 26

llrem <- unique(c(rownames(train\_new)[llres],llcoo)); length(llrem)

## [1] 29

llremreg<-which(rownames(train\_new)%in%llrem);llremreg

## [1] 32 55 116 228 241 266 305 331 351 383 558 595 608 621 673  
## [16] 708 723 732 740 753 857 868 914 923 938 950 979 1000

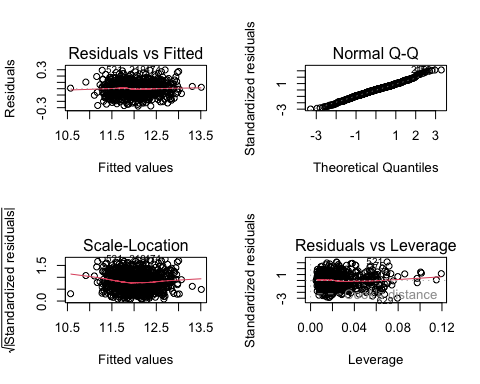
train\_new<-train\_new[-llremreg,]  
  
final\_modelp <- lm(log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + log(X1stFlrSF)+ X2ndFlrSF +  
 BsmtFullBath + FullBath + TotRmsAbvGrd+  
 KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch + PoolArea +ExterQual + BsmtQual + KitchenQual+GarageType  
 , data = train\_new)  
summary(final\_modelp)

##   
## Call:  
## lm(formula = log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea + ExterQual + BsmtQual +   
## KitchenQual + GarageType, data = train\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.272403 -0.060697 0.003073 0.058368 0.296143   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.441e+00 4.602e-01 3.131 0.001792 \*\*   
## LotFrontage -1.978e-04 1.684e-04 -1.175 0.240427   
## log(LotArea) 1.007e-01 8.444e-03 11.928 < 2e-16 \*\*\*  
## OverallQual 6.092e-02 4.333e-03 14.059 < 2e-16 \*\*\*  
## OverallCond 5.230e-02 3.193e-03 16.378 < 2e-16 \*\*\*  
## YearBuilt 3.125e-03 2.029e-04 15.397 < 2e-16 \*\*\*  
## BsmtFinSF1 7.539e-05 1.024e-05 7.360 3.96e-13 \*\*\*  
## log(X1stFlrSF) 3.990e-01 2.024e-02 19.712 < 2e-16 \*\*\*  
## X2ndFlrSF 2.437e-04 1.475e-05 16.518 < 2e-16 \*\*\*  
## BsmtFullBath 2.130e-02 7.968e-03 2.673 0.007652 \*\*   
## FullBath 2.992e-02 8.767e-03 3.413 0.000669 \*\*\*  
## TotRmsAbvGrd -1.121e-03 3.809e-03 -0.294 0.768661   
## KitchenAbvGr -7.967e-02 1.785e-02 -4.463 9.03e-06 \*\*\*  
## Fireplaces 2.121e-02 5.773e-03 3.674 0.000252 \*\*\*  
## GarageCars 4.620e-02 7.139e-03 6.471 1.55e-10 \*\*\*  
## ScreenPorch 2.872e-04 6.013e-05 4.776 2.07e-06 \*\*\*  
## PoolArea NA NA NA NA   
## ExterQualFa -5.908e-02 4.416e-02 -1.338 0.181261   
## ExterQualGd -7.465e-03 2.109e-02 -0.354 0.723432   
## ExterQualTA -4.687e-02 2.359e-02 -1.987 0.047237 \*   
## BsmtQualFa -1.702e-01 2.809e-02 -6.060 1.95e-09 \*\*\*  
## BsmtQualGd -9.258e-02 1.565e-02 -5.914 4.63e-09 \*\*\*  
## BsmtQualNo Basement -2.198e-01 2.857e-02 -7.692 3.60e-14 \*\*\*  
## BsmtQualTA -9.869e-02 1.848e-02 -5.341 1.15e-07 \*\*\*  
## KitchenQualFa -1.039e-01 3.001e-02 -3.461 0.000562 \*\*\*  
## KitchenQualGd -6.531e-02 1.574e-02 -4.150 3.61e-05 \*\*\*  
## KitchenQualTA -9.320e-02 1.765e-02 -5.281 1.59e-07 \*\*\*  
## GarageTypeAttchd 8.361e-02 5.029e-02 1.663 0.096733 .   
## GarageTypeBasment 1.171e-01 5.758e-02 2.034 0.042217 \*   
## GarageTypeBuiltIn 8.844e-02 5.219e-02 1.695 0.090448 .   
## GarageTypeCarPort 7.118e-03 6.364e-02 0.112 0.910975   
## GarageTypeDetchd 9.631e-02 5.006e-02 1.924 0.054673 .   
## GarageTypeNo Garage 9.918e-02 5.401e-02 1.836 0.066620 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.09566 on 956 degrees of freedom  
## Multiple R-squared: 0.9407, Adjusted R-squared: 0.9388   
## F-statistic: 489.5 on 31 and 956 DF, p-value: < 2.2e-16

final\_modelp <- step( final\_modelp, k=log(nrow(train\_new)))

## Start: AIC=-4449.49  
## log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + PoolArea + ExterQual + BsmtQual + KitchenQual +   
## GarageType  
##   
##   
## Step: AIC=-4449.49  
## log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + ExterQual + BsmtQual + KitchenQual + GarageType  
##   
## Df Sum of Sq RSS AIC  
## - GarageType 6 0.0962 8.8441 -4480.1  
## - ExterQual 3 0.1157 8.8636 -4457.2  
## - TotRmsAbvGrd 1 0.0008 8.7486 -4456.3  
## - LotFrontage 1 0.0126 8.7605 -4455.0  
## <none> 8.7479 -4449.5  
## - BsmtFullBath 1 0.0654 8.8132 -4449.0  
## - FullBath 1 0.1066 8.8545 -4444.4  
## - Fireplaces 1 0.1235 8.8714 -4442.5  
## - KitchenQual 3 0.2594 9.0072 -4441.3  
## - KitchenAbvGr 1 0.1823 8.9301 -4436.0  
## - ScreenPorch 1 0.2087 8.9565 -4433.1  
## - GarageCars 1 0.3832 9.1310 -4414.0  
## - BsmtQual 4 0.6619 9.4097 -4405.0  
## - BsmtFinSF1 1 0.4957 9.2435 -4401.9  
## - log(LotArea) 1 1.3019 10.0497 -4319.3  
## - OverallQual 1 1.8086 10.5565 -4270.7  
## - YearBuilt 1 2.1693 10.9172 -4237.5  
## - OverallCond 1 2.4544 11.2023 -4212.0  
## - X2ndFlrSF 1 2.4967 11.2446 -4208.3  
## - log(X1stFlrSF) 1 3.5555 12.3034 -4119.4  
##   
## Step: AIC=-4480.05  
## log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + ExterQual + BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## - ExterQual 3 0.1140 8.9581 -4488.1  
## - TotRmsAbvGrd 1 0.0002 8.8443 -4486.9  
## - LotFrontage 1 0.0136 8.8577 -4485.4  
## <none> 8.8441 -4480.1  
## - BsmtFullBath 1 0.0668 8.9109 -4479.5  
## - FullBath 1 0.1046 8.9487 -4475.3  
## - Fireplaces 1 0.1264 8.9705 -4472.9  
## - KitchenQual 3 0.2550 9.0991 -4472.7  
## - ScreenPorch 1 0.2082 9.0523 -4464.0  
## - KitchenAbvGr 1 0.2330 9.0771 -4461.3  
## - BsmtQual 4 0.6771 9.5212 -4434.8  
## - BsmtFinSF1 1 0.4842 9.3283 -4434.3  
## - GarageCars 1 0.5118 9.3559 -4431.4  
## - log(LotArea) 1 1.2562 10.1003 -4355.7  
## - OverallQual 1 1.8902 10.7343 -4295.6  
## - YearBuilt 1 2.4386 11.2827 -4246.3  
## - OverallCond 1 2.4849 11.3290 -4242.3  
## - X2ndFlrSF 1 2.5907 11.4348 -4233.1  
## - log(X1stFlrSF) 1 3.5921 12.4362 -4150.2  
##   
## Step: AIC=-4488.08  
## log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces + GarageCars +   
## ScreenPorch + BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## - TotRmsAbvGrd 1 0.0005 8.9586 -4494.9  
## - LotFrontage 1 0.0201 8.9783 -4492.8  
## <none> 8.9581 -4488.1  
## - BsmtFullBath 1 0.0731 9.0313 -4486.9  
## - FullBath 1 0.1028 9.0610 -4483.7  
## - Fireplaces 1 0.1124 9.0705 -4482.7  
## - ScreenPorch 1 0.2052 9.1634 -4472.6  
## - KitchenAbvGr 1 0.2268 9.1849 -4470.3  
## - KitchenQual 3 0.4590 9.4172 -4459.4  
## - BsmtFinSF1 1 0.4508 9.4090 -4446.5  
## - GarageCars 1 0.5151 9.4732 -4439.7  
## - BsmtQual 4 0.7216 9.6798 -4439.1  
## - log(LotArea) 1 1.2629 10.2210 -4364.7  
## - OverallQual 1 2.4288 11.3869 -4258.0  
## - OverallCond 1 2.5512 11.5094 -4247.4  
## - X2ndFlrSF 1 2.6326 11.5907 -4240.4  
## - YearBuilt 1 2.6773 11.6354 -4236.6  
## - log(X1stFlrSF) 1 3.6915 12.6497 -4154.1  
##   
## Step: AIC=-4494.93  
## log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual + OverallCond +   
## YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + KitchenAbvGr + Fireplaces + GarageCars + ScreenPorch +   
## BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## - LotFrontage 1 0.0207 8.9793 -4499.5  
## <none> 8.9586 -4494.9  
## - BsmtFullBath 1 0.0730 9.0315 -4493.8  
## - FullBath 1 0.1026 9.0612 -4490.6  
## - Fireplaces 1 0.1123 9.0709 -4489.5  
## - ScreenPorch 1 0.2064 9.1650 -4479.3  
## - KitchenAbvGr 1 0.2444 9.2030 -4475.2  
## - KitchenQual 3 0.4589 9.4174 -4466.3  
## - BsmtFinSF1 1 0.4655 9.4241 -4451.8  
## - GarageCars 1 0.5147 9.4733 -4446.6  
## - BsmtQual 4 0.7284 9.6869 -4445.3  
## - log(LotArea) 1 1.2638 10.2224 -4371.4  
## - OverallQual 1 2.4320 11.3905 -4264.5  
## - OverallCond 1 2.5517 11.5102 -4254.2  
## - YearBuilt 1 2.6780 11.6366 -4243.4  
## - X2ndFlrSF 1 4.8226 13.7812 -4076.3  
## - log(X1stFlrSF) 1 4.8432 13.8018 -4074.8  
##   
## Step: AIC=-4499.54  
## log(SalePrice) ~ log(LotArea) + OverallQual + OverallCond + YearBuilt +   
## BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF + BsmtFullBath +   
## FullBath + KitchenAbvGr + Fireplaces + GarageCars + ScreenPorch +   
## BsmtQual + KitchenQual  
##   
## Df Sum of Sq RSS AIC  
## <none> 8.9793 -4499.5  
## - BsmtFullBath 1 0.0736 9.0529 -4498.4  
## - FullBath 1 0.1030 9.0823 -4495.2  
## - Fireplaces 1 0.1162 9.0955 -4493.7  
## - ScreenPorch 1 0.2060 9.1853 -4484.0  
## - KitchenAbvGr 1 0.2392 9.2185 -4480.5  
## - KitchenQual 3 0.4681 9.4474 -4470.0  
## - BsmtFinSF1 1 0.4753 9.4546 -4455.5  
## - GarageCars 1 0.5034 9.4826 -4452.6  
## - BsmtQual 4 0.7242 9.7035 -4450.5  
## - log(LotArea) 1 1.4681 10.4474 -4356.8  
## - OverallQual 1 2.4310 11.4103 -4269.7  
## - OverallCond 1 2.5331 11.5124 -4260.9  
## - YearBuilt 1 2.6583 11.6376 -4250.2  
## - X2ndFlrSF 1 4.8079 13.7872 -4082.8  
## - log(X1stFlrSF) 1 4.8760 13.8552 -4077.9

par(mfrow=c(2,2))  
plot(final\_modelp)



coef\_orig <- coef(final\_model)  
coef\_filt <- coef(final\_modelp)  
  
#common variables  
var\_com <- intersect(names(coef\_orig), names(coef\_filt))  
  
coef\_orig <- coef\_orig[var\_com]  
coef\_filt <- coef\_filt[var\_com]  
  
coef\_comparison <- data.frame(Variable = names(coef\_orig),  
 Original\_coefficient = coef\_orig,  
 Filtered\_coefficient = coef\_filt)  
print(coef\_comparison)

## Variable Original\_coefficient  
## (Intercept) (Intercept) 1.767309e+00  
## log(LotArea) log(LotArea) 9.190035e-02  
## OverallQual OverallQual 6.673385e-02  
## OverallCond OverallCond 5.271844e-02  
## YearBuilt YearBuilt 3.043825e-03  
## BsmtFinSF1 BsmtFinSF1 7.807564e-05  
## log(X1stFlrSF) log(X1stFlrSF) 3.893113e-01  
## X2ndFlrSF X2ndFlrSF 2.329787e-04  
## BsmtFullBath BsmtFullBath 2.341133e-02  
## FullBath FullBath 2.868578e-02  
## KitchenAbvGr KitchenAbvGr -6.975039e-02  
## Fireplaces Fireplaces 2.459133e-02  
## GarageCars GarageCars 4.864742e-02  
## ScreenPorch ScreenPorch 2.955594e-04  
## BsmtQualFa BsmtQualFa -1.539022e-01  
## BsmtQualGd BsmtQualGd -9.580525e-02  
## BsmtQualNo Basement BsmtQualNo Basement -2.183703e-01  
## BsmtQualTA BsmtQualTA -1.055267e-01  
## KitchenQualFa KitchenQualFa -6.275097e-02  
## KitchenQualGd KitchenQualGd -4.250488e-02  
## KitchenQualTA KitchenQualTA -7.430136e-02  
## Filtered\_coefficient  
## (Intercept) 1.589139e+00  
## log(LotArea) 9.219227e-02  
## OverallQual 6.658310e-02  
## OverallCond 5.175633e-02  
## YearBuilt 3.126180e-03  
## BsmtFinSF1 7.241335e-05  
## log(X1stFlrSF) 3.934097e-01  
## X2ndFlrSF 2.401428e-04  
## BsmtFullBath 2.250707e-02  
## FullBath 2.895870e-02  
## KitchenAbvGr -8.620365e-02  
## Fireplaces 2.027016e-02  
## GarageCars 4.260217e-02  
## ScreenPorch 2.840820e-04  
## BsmtQualFa -1.771345e-01  
## BsmtQualGd -9.503964e-02  
## BsmtQualNo Basement -2.248469e-01  
## BsmtQualTA -1.070349e-01  
## KitchenQualFa -1.134850e-01  
## KitchenQualGd -6.608734e-02  
## KitchenQualTA -1.099259e-01

results <- data.frame(Model = c("Model with train\_new sample",  
 "Model with test\_new sample"),  
 RSQRT = round(c(RSQUARE(exp(fitted(final\_modelp)), train\_new$SalePrice),  
 RSQUARE(exp(predict(final\_modelp, newdata = test\_new)), test\_new$SalePrice)),2),  
 RMSE = round(c(rmse(exp(fitted(final\_modelp)), train\_new$SalePrice),  
 rmse(exp(predict(final\_modelp, newdata = test\_new)), test\_new$SalePrice)),2))  
results

## Model RSQRT RMSE  
## 1 Model with train\_new sample 0.94 19429.18  
## 2 Model with test\_new sample 0.92 23865.33

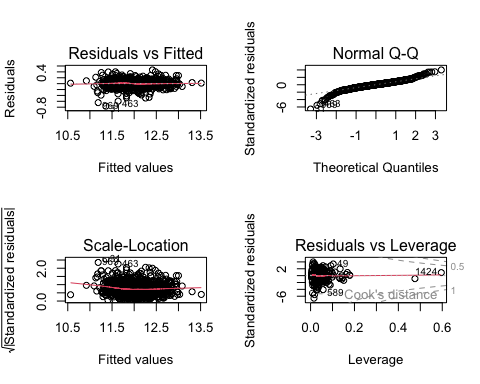
## Interpret the Model

The Residuals vs Fitted plot shows that the residuals follow a good linear pattern, which meets the regression assumptions very well. The Normal Q-Q plot shows that the standard errors are mostly normally distributed with a small amount of deviating prediction at the upper end of the tail and the beginning of the tail. The scale-location plot shows that homoscedasticity is satisfied as a nearly straight line is obtained. The residuals vs leverage plot shows that our model includes some high-leverage (so highly influential in our model) points that deviate significantly (more than 4 standardized residuals away) and asymmetrically from the prediction. An influence plot further confirms this believe.

summary(final\_model)

##   
## Call:  
## lm(formula = log(SalePrice) ~ LotFrontage + log(LotArea) + OverallQual +   
## OverallCond + YearBuilt + BsmtFinSF1 + log(X1stFlrSF) + X2ndFlrSF +   
## BsmtFullBath + FullBath + TotRmsAbvGrd + KitchenAbvGr + Fireplaces +   
## GarageCars + ScreenPorch + PoolArea + ExterQual + BsmtQual +   
## KitchenQual, data = train\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.76448 -0.06002 0.00452 0.06468 0.46643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.767e+00 5.091e-01 3.471 0.000540 \*\*\*  
## LotFrontage -1.600e-04 1.925e-04 -0.831 0.406043   
## log(LotArea) 9.190e-02 1.001e-02 9.178 < 2e-16 \*\*\*  
## OverallQual 6.673e-02 5.207e-03 12.816 < 2e-16 \*\*\*  
## OverallCond 5.272e-02 3.841e-03 13.726 < 2e-16 \*\*\*  
## YearBuilt 3.044e-03 2.249e-04 13.533 < 2e-16 \*\*\*  
## BsmtFinSF1 7.808e-05 1.245e-05 6.271 5.34e-10 \*\*\*  
## log(X1stFlrSF) 3.893e-01 2.412e-02 16.139 < 2e-16 \*\*\*  
## X2ndFlrSF 2.330e-04 1.742e-05 13.372 < 2e-16 \*\*\*  
## BsmtFullBath 2.341e-02 9.685e-03 2.417 0.015814 \*   
## FullBath 2.869e-02 1.060e-02 2.707 0.006910 \*\*   
## TotRmsAbvGrd -2.755e-04 4.592e-03 -0.060 0.952181   
## KitchenAbvGr -6.975e-02 2.018e-02 -3.456 0.000572 \*\*\*  
## Fireplaces 2.459e-02 6.938e-03 3.544 0.000412 \*\*\*  
## GarageCars 4.865e-02 6.934e-03 7.016 4.23e-12 \*\*\*  
## ScreenPorch 2.956e-04 7.351e-05 4.021 6.24e-05 \*\*\*  
## PoolArea 1.648e-04 1.248e-04 1.320 0.187017   
## ExterQualFa -6.660e-02 5.296e-02 -1.258 0.208799   
## ExterQualGd -2.317e-02 2.546e-02 -0.910 0.362928   
## ExterQualTA -5.433e-02 2.833e-02 -1.918 0.055451 .   
## BsmtQualFa -1.539e-01 3.375e-02 -4.560 5.75e-06 \*\*\*  
## BsmtQualGd -9.581e-02 1.901e-02 -5.040 5.53e-07 \*\*\*  
## BsmtQualNo Basement -2.184e-01 3.459e-02 -6.313 4.11e-10 \*\*\*  
## BsmtQualTA -1.055e-01 2.221e-02 -4.751 2.32e-06 \*\*\*  
## KitchenQualFa -6.275e-02 3.521e-02 -1.782 0.075004 .   
## KitchenQualGd -4.250e-02 1.878e-02 -2.263 0.023835 \*   
## KitchenQualTA -7.430e-02 2.103e-02 -3.533 0.000430 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1174 on 989 degrees of freedom  
## Multiple R-squared: 0.915, Adjusted R-squared: 0.9127   
## F-statistic: 409.4 on 26 and 989 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(final\_model)



coefficients <- coef(final\_model)  
coefficients

## (Intercept) LotFrontage log(LotArea) OverallQual   
## 1.767309e+00 -1.599888e-04 9.190035e-02 6.673385e-02   
## OverallCond YearBuilt BsmtFinSF1 log(X1stFlrSF)   
## 5.271844e-02 3.043825e-03 7.807564e-05 3.893113e-01   
## X2ndFlrSF BsmtFullBath FullBath TotRmsAbvGrd   
## 2.329787e-04 2.341133e-02 2.868578e-02 -2.754672e-04   
## KitchenAbvGr Fireplaces GarageCars ScreenPorch   
## -6.975039e-02 2.459133e-02 4.864742e-02 2.955594e-04   
## PoolArea ExterQualFa ExterQualGd ExterQualTA   
## 1.647560e-04 -6.660470e-02 -2.317243e-02 -5.433320e-02   
## BsmtQualFa BsmtQualGd BsmtQualNo Basement BsmtQualTA   
## -1.539022e-01 -9.580525e-02 -2.183703e-01 -1.055267e-01   
## KitchenQualFa KitchenQualGd KitchenQualTA   
## -6.275097e-02 -4.250488e-02 -7.430136e-02

We can see the coefficient with largest value (in absolute value) is log(X1stFlrSF) with positive impact, 0.385. Meaning that with all other things being equal(c.p) when the first floor square feet increases, the log-price also increases. Larger first-floor areas are often associated with more spacious living arrangements, which can contribute to higher property values. Continuing with the analysis of the coefficients, the variable “BsmtQualNo Basement” is noteworthy due to its substantial negative impact of -0.227. This suggests that houses without a basement tend to have a significant decrease in their sales price, all else being equal. Basements often contribute to the overall living space and functionality of a house, so the absence of this feature appears to have a notable effect on the property’s value. The presence of a fireplace, as indicated by the variable “Fireplaces,” has a positive coefficient of 0.0248. This suggests that houses with fireplaces tend to have higher predicted sales prices. Fireplaces are often considered desirable features, providing both functional and aesthetic benefits, which can positively impact a property’s perceived value.

Variables related to kitchen quality, such as “KitchenQualFa” (Kitchen Quality Fair) and “KitchenQualTA” (Kitchen Quality Typical), both have negative coefficients of -0.064 and -0.075, respectively. This indicates that, compared to higher-quality kitchens, houses with fair or typical kitchen quality are associated with lower predicted sales prices. Kitchens are known to be focal points for homebuyers, and a well-appointed kitchen is often considered a positive attribute.

In conclusion, the coefficients shed light on various aspects that influence the predicted sales prices of houses in the model. Factors such as the house square feet, the presence of a fireplace, and the quality of the kitchen and exterior all play roles in determining the estimated market value of a property.

## Annex