Assignment 1

Alicia Chimeno & Silvia Ferrer

2023-11-18

## 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 dataset and Preproces 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)

## 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), therfore we shall not consider that the target variable is normally distributed.

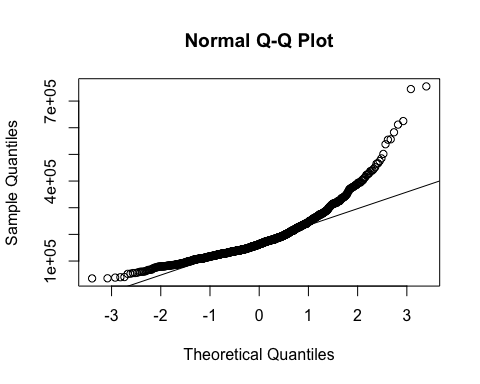
res.con <- condes(train,81)  
head(res.con$quali)

## R2 p.value  
## Neighborhood 0.5455750 1.558600e-225  
## ExterQual 0.4773878 1.439551e-204  
## BsmtQual 0.4649938 8.158548e-196  
## KitchenQual 0.4565986 3.032213e-192  
## GarageFinish 0.3058737 6.228747e-115  
## FireplaceQu 0.2939608 2.971217e-107

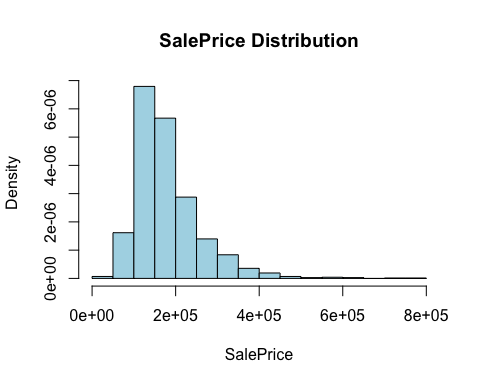
shapiro.test(train$SalePrice)

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

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)

## Selection of 10 factor variables and numerical variables

We select the first 10th 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

# Analysis functions

To proceede with the predictor analysis, we shall create the following functions:

## 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)  
}

# EDA and univariate outliers of each variable

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)

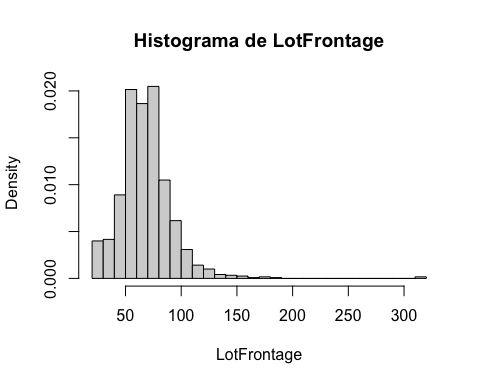
### Numerical variables

## Id (not considered for the analysis)

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

numeric\_description(train2$LotFrontage, 30) # not normal distribution

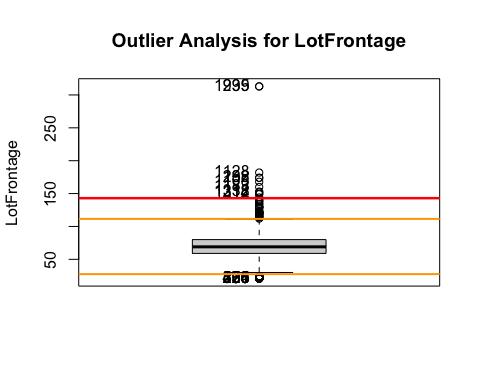
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 21.00 59.00 69.00 70.05 80.00 313.00 259   
##   
## Count of missing values: 259



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

analyze\_outliers(train2, "LotFrontage")

## Number of mild outliers: 88



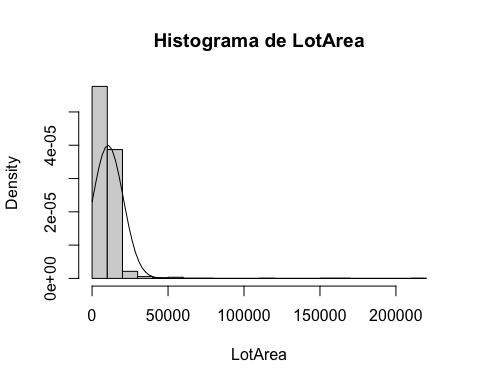
## Number of severe outliers: 12

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

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

numeric\_description(train2$LotArea, 20) # not normal distribution

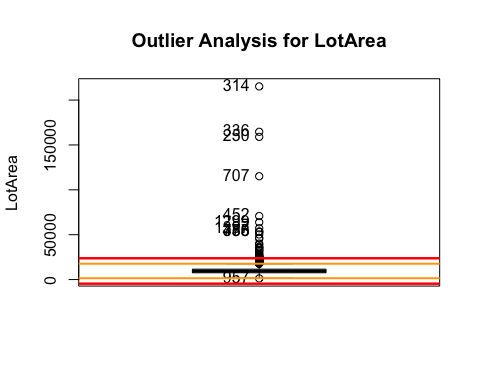
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1300 7554 9478 10517 11602 215245   
##   
## Count of missing values: 0



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

analyze\_outliers(train2, "LotArea")

## Number of mild outliers: 69



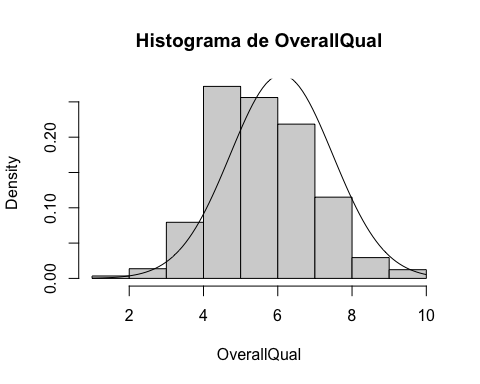
## Number of severe outliers: 34

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

## “OverallQual” (cont)

numeric\_description(train2$OverallQual, 10) # not normal distribution

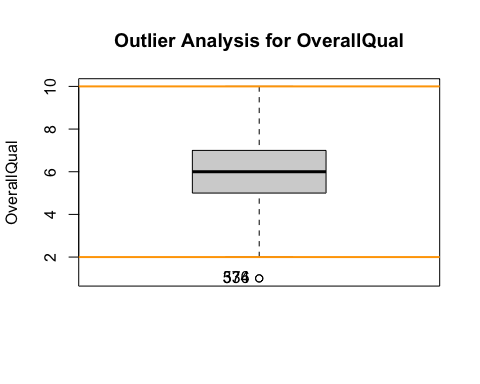
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 5.000 6.000 6.099 7.000 10.000   
##   
## Count of missing values: 0



##   
## 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

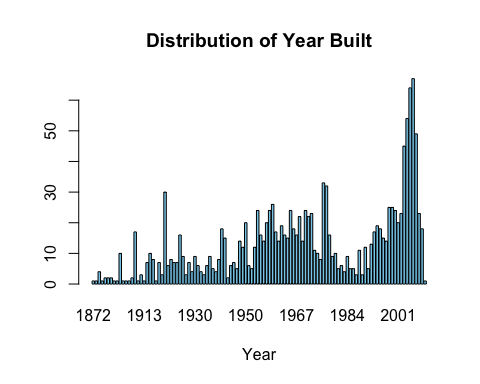
train2 <- update\_outliers\_count(train2, "OverallQual")

## “YearBuilt” (int)

summary(train2$YearBuilt)

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

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

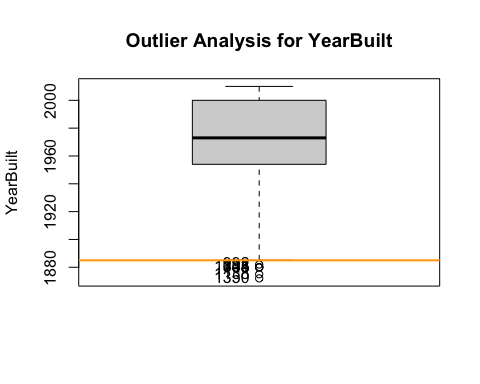


sum(is.na(train2$YearBuilt))

## [1] 0

analyze\_outliers(train2, "YearBuilt")

## Number of mild outliers: 7



## Number of severe outliers: 0

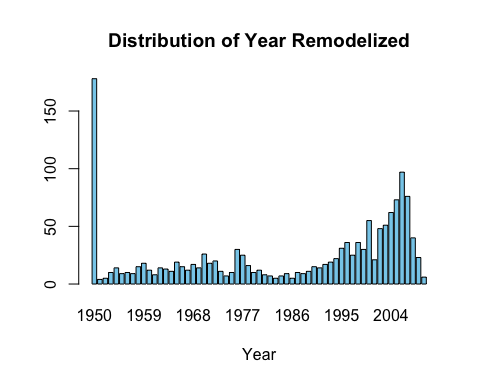
train2 <- update\_outliers\_count(train2, "YearBuilt")

## “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")

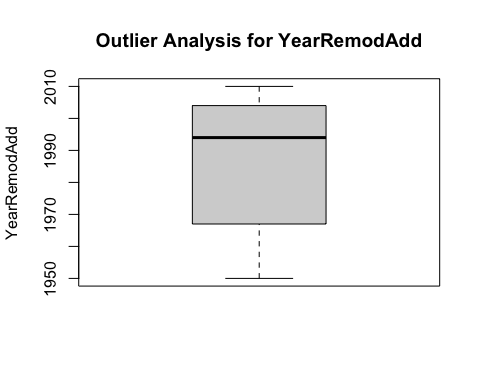


sum(is.na(train2$YearRemodAdd))

## [1] 0

analyze\_outliers(train2, "YearRemodAdd")

## Number of mild outliers: 0



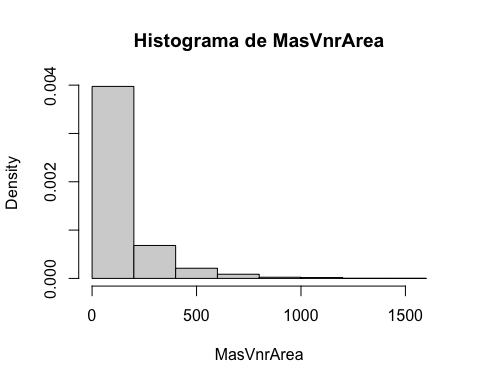
## Number of severe outliers: 0

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

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

numeric\_description(train2$MasVnrArea, 10)# not normal distribution

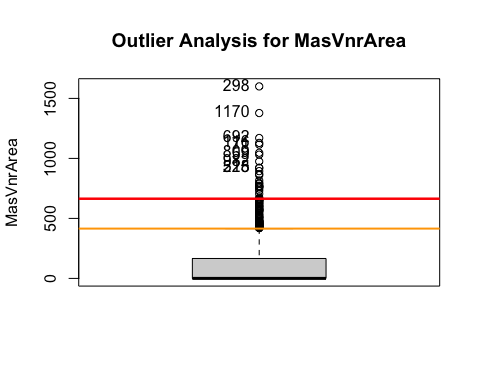
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 0.0 0.0 103.7 166.0 1600.0 8   
##   
## Count of missing values: 8



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

analyze\_outliers(train2,"MasVnrArea")

## Number of mild outliers: 96



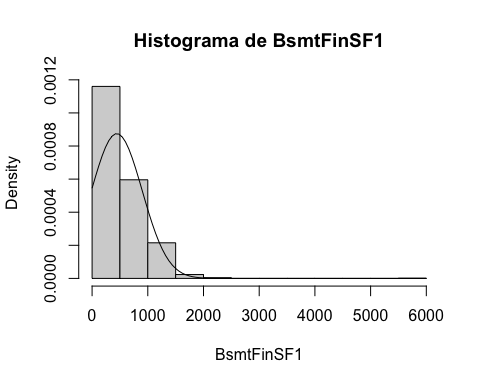
## Number of severe outliers: 25

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

numeric\_description(train2$BsmtFinSF1, 10)# not normal distribution

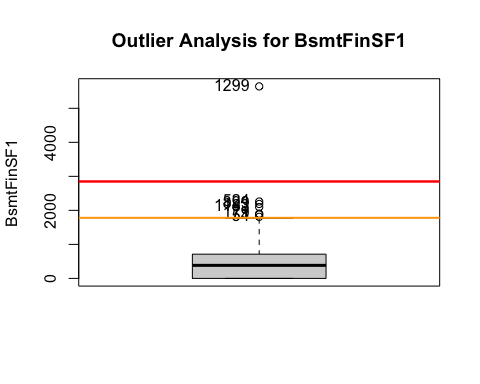
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 383.5 443.6 712.2 5644.0   
##   
## Count of missing values: 0



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

analyze\_outliers(train2,"BsmtFinSF1")

## Number of mild outliers: 7



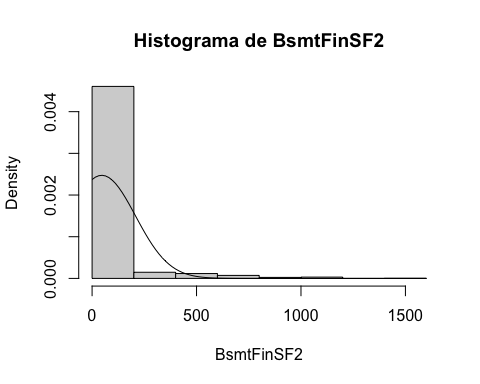
## Number of severe outliers: 1

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

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

numeric\_description(train2$BsmtFinSF2, 10) # not normal distribution

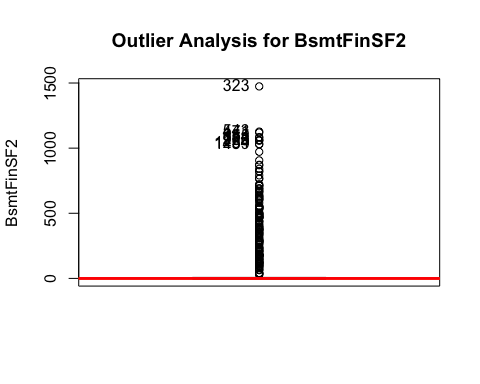
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 46.55 0.00 1474.00   
##   
## Count of missing values: 0



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

analyze\_outliers(train2,"BsmtFinSF2")

## Number of mild outliers: 167



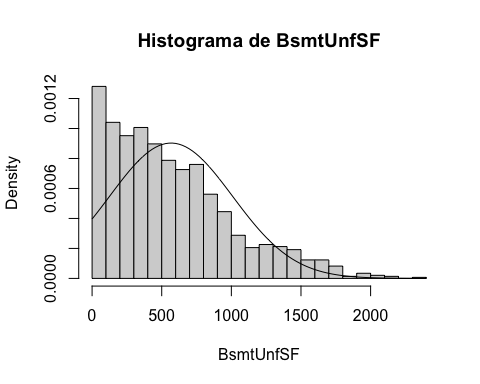
## Number of severe outliers: 167

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

# “BsmtUnfSF” (cont)

numeric\_description(train2$BsmtUnfSF, 20) # not normal distribution

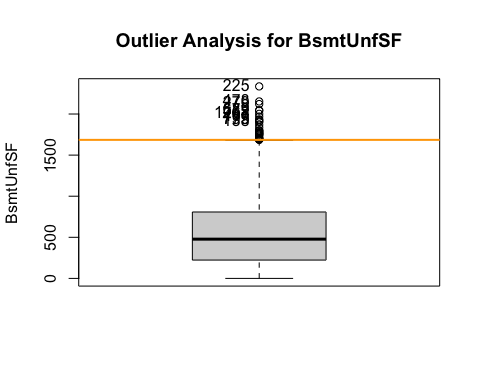
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 223.0 477.5 567.2 808.0 2336.0   
##   
## Count of missing values: 0



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

analyze\_outliers(train2,"BsmtUnfSF")

## Number of mild outliers: 29



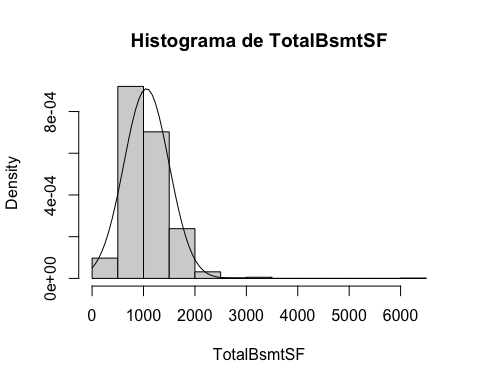
## Number of severe outliers: 0

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

# “TotalBsmtSF” (cont)

numeric\_description(train2$TotalBsmtSF, 10)# not normal distribution

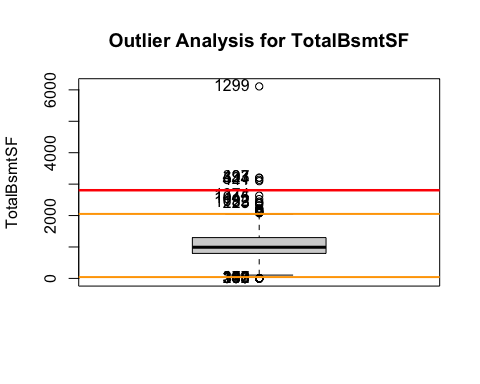
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 795.8 991.5 1057.4 1298.2 6110.0   
##   
## Count of missing values: 0



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

analyze\_outliers(train2,"TotalBsmtSF")

## Number of mild outliers: 61



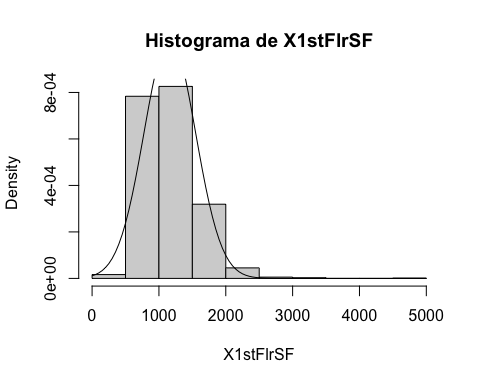
## Number of severe outliers: 5

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

# X1stFlrSF - First Floor square feet (cont)

numeric\_description(train2$X1stFlrSF, 10)

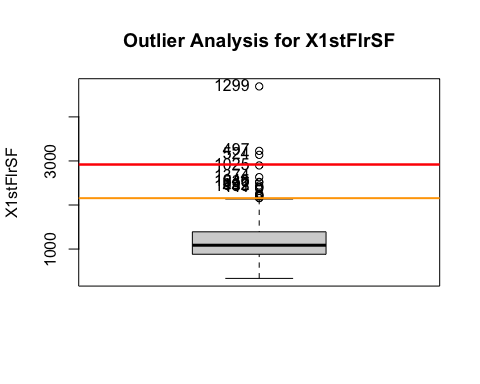
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 882 1087 1163 1391 4692   
##   
## Count of missing values: 0



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

# not normal distribution  
analyze\_outliers(train2,"X1stFlrSF")

## Number of mild outliers: 20



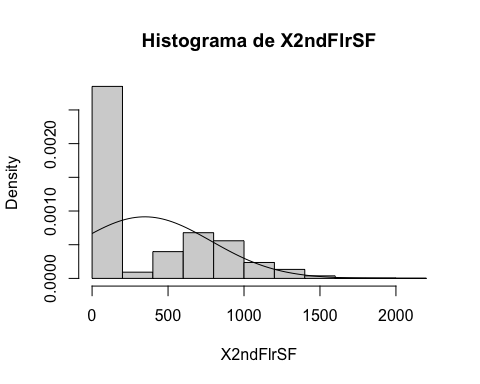
## Number of severe outliers: 3

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

## X2ndFlrSF - Second floor square feet (cont)

numeric\_description(train2$X2ndFlrSF, 10)# not normal distribution

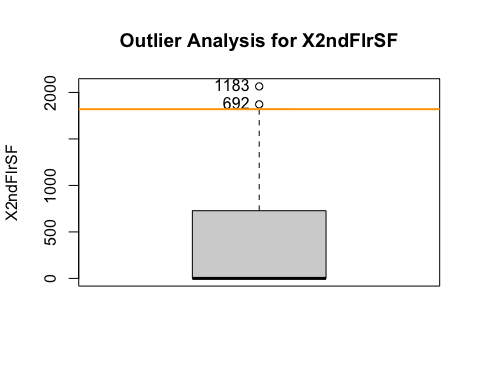
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 347 728 2065   
##   
## Count of missing values: 0



##   
## 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

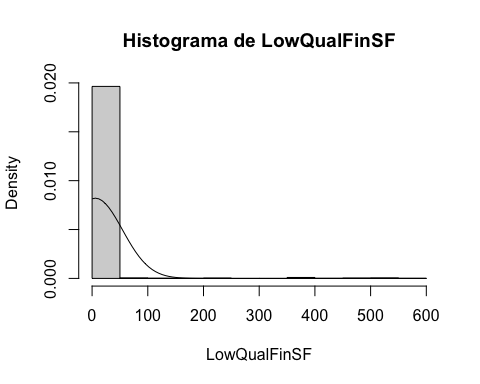
train2 <- update\_outliers\_count(train2, "X2ndFlrSF")  
  
# create a new variable if has second floor = 1 / no second floor = 0.   
train2$secondfloor<-0  
train2$secondfloor[which(train2$X2ndFlrSF!=0)]<-1  
train2$secondfloor<-factor(train2$secondfloor)  
table(train2$secondfloor)

##   
## 0 1   
## 829 631

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

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

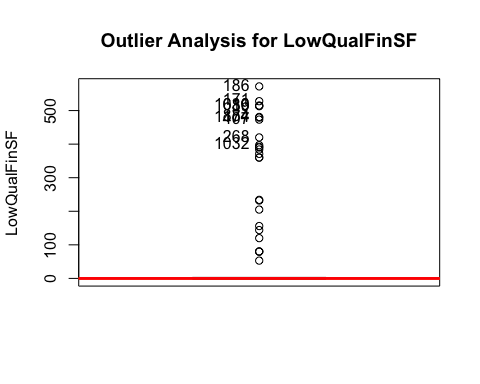
##   
## 0 53 80 120 144 156 205 232 234 360 371 384 390 392 397 420   
## 1434 1 3 1 1 1 1 1 1 2 1 1 1 1 1 1   
## 473 479 481 513 514 515 528 572   
## 1 1 1 1 1 1 1 1

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

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

analyze\_outliers(train2,"LowQualFinSF")

## Number of mild outliers: 26



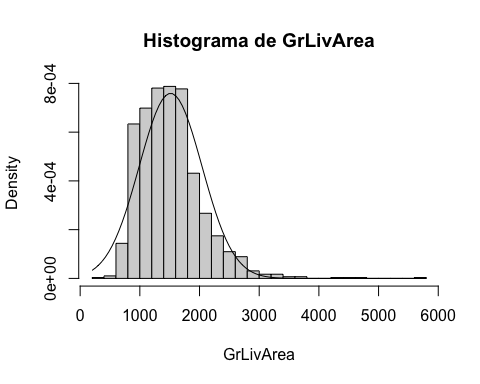
## Number of severe outliers: 26

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

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

numeric\_description(train2$GrLivArea, 20) # not normal distribution

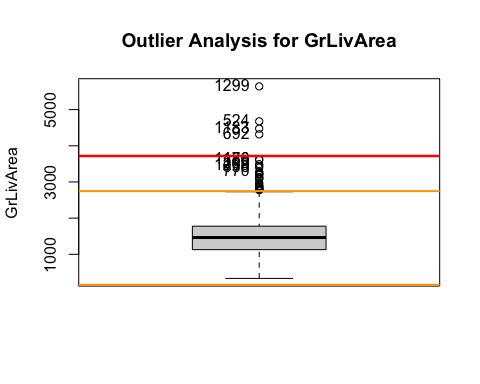
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1130 1464 1515 1777 5642   
##   
## Count of missing values: 0



##   
## 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")

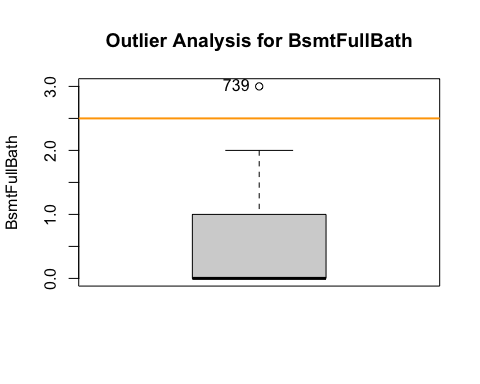
## BsmtFullBath - # full baths in the basement (int)

summary(train2$BsmtFullBath)

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

analyze\_outliers(train2,"BsmtFullBath")

## Number of mild outliers: 1



## Number of severe outliers: 0

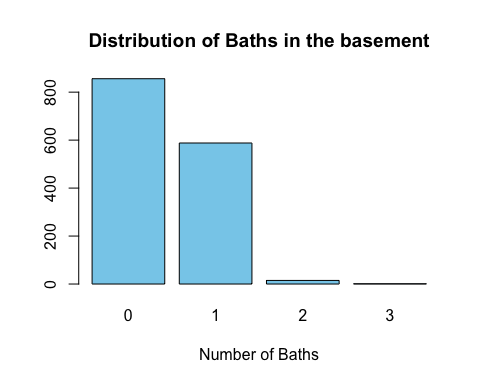
train2 <- update\_outliers\_count(train2, "BsmtFullBath")  
table(train2$BsmtFullBath)

##   
## 0 1 2 3   
## 856 588 15 1

sum(is.na(train2$BsmtFullBath))

## [1] 0

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



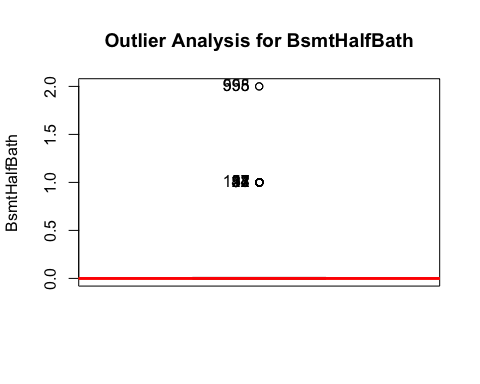
## BsmtHalfBath - # half baths in the basement (int)

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

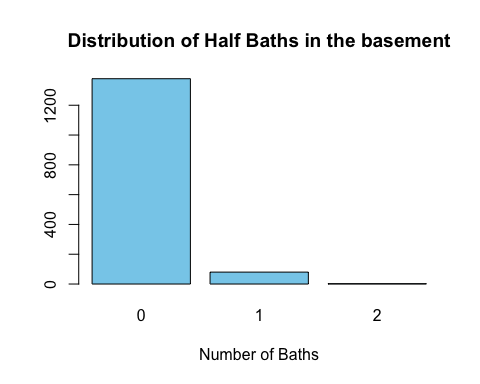
nearZeroVar(train2$BsmtHalfBath, saveMetrics = TRUE) # Variable not with near zero variance

## 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")



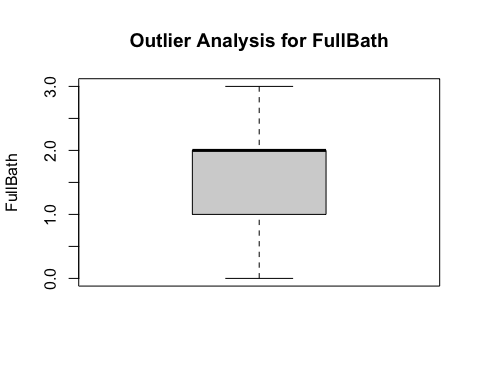
## FullBath - full baths above grade (int)

summary(train2$FullBath)

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

analyze\_outliers(train2,"FullBath")

## Number of mild outliers: 0



## Number of severe outliers: 0

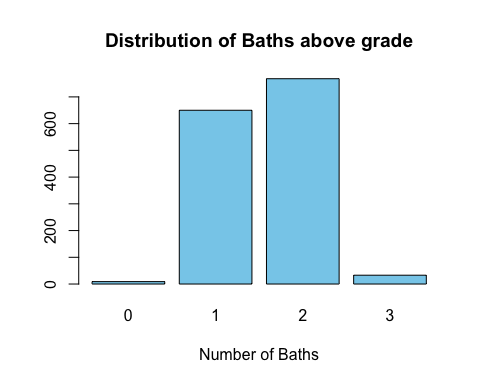
train2 <- update\_outliers\_count(train2, "FullBath")  
table(train2$FullBath)

##   
## 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")



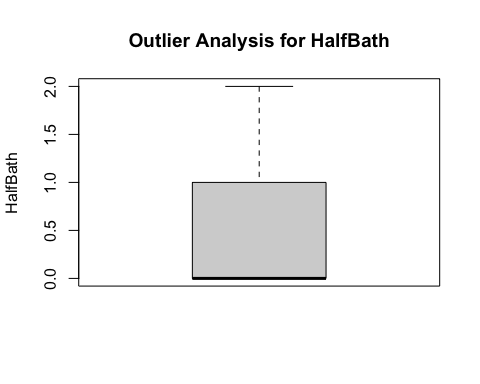
## HalfBath - Half baths above grade (int)

summary(train2$HalfBath)

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

analyze\_outliers(train2,"HalfBath")

## Number of mild outliers: 0



## Number of severe outliers: 0

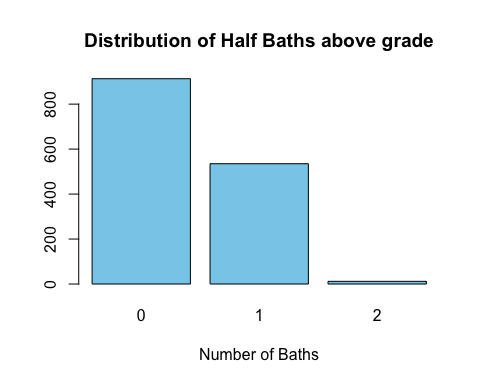
train2 <- update\_outliers\_count(train2, "HalfBath")  
table(train2$HalfBath)

##   
## 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")



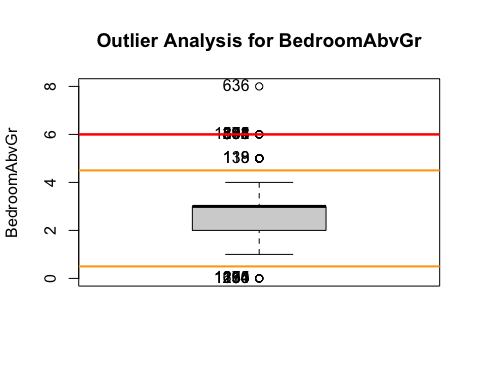
## BedroomAbvGr - Bedrooms above grade (does NOT include basement bedrooms)

summary(train2$BedroomAbvGr)

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

analyze\_outliers(train2,"BedroomAbvGr")

## Number of mild outliers: 35



## Number of severe outliers: 1

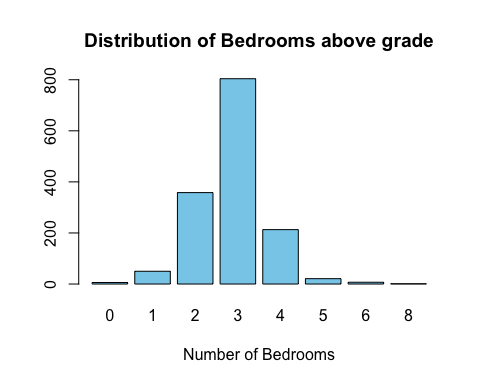
train2 <- update\_outliers\_count(train2, "BedroomAbvGr")  
table(train2$BedroomAbvGr)

##   
## 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")



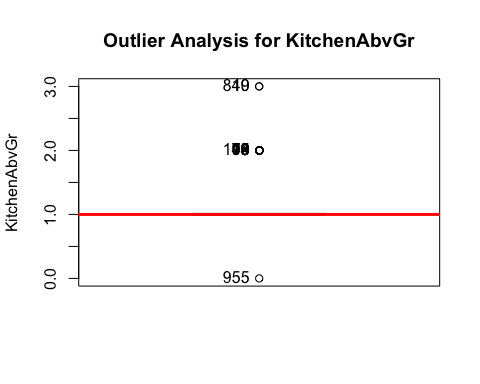
## KitchenAbvGr - Kitchens above grade (int)

summary(train2$KitchenAbvGr)

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

analyze\_outliers(train2,"KitchenAbvGr")

## Number of mild outliers: 68



## Number of severe outliers: 68

train2 <- update\_outliers\_count(train2, "KitchenAbvGr")  
table(train2$KitchenAbvGr) # variable very centered in 1

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

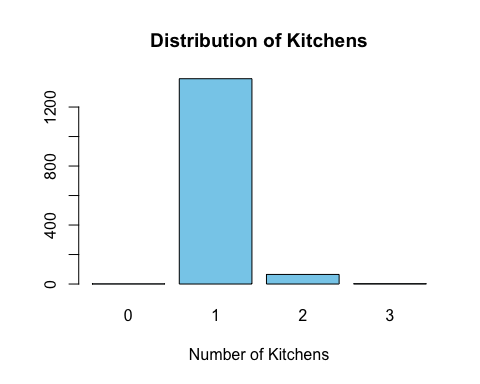
nearZeroVar(train2$KitchenAbvGr,saveMetrics = TRUE) # near zero variance

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

sum(is.na(train2$KitchenAbvGr))

## [1] 0

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



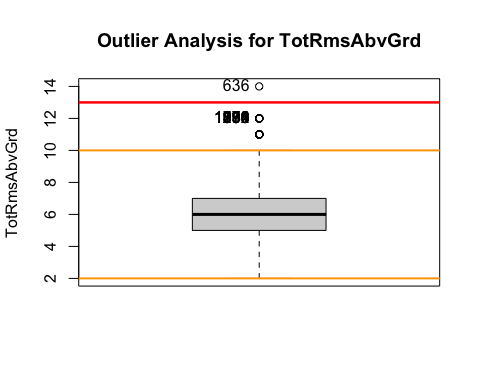
## “TotRmsAbvGrd” - Total rooms above grade (does not include bathrooms) - continuous ratio variable (int)

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

table(train2$Fireplaces)

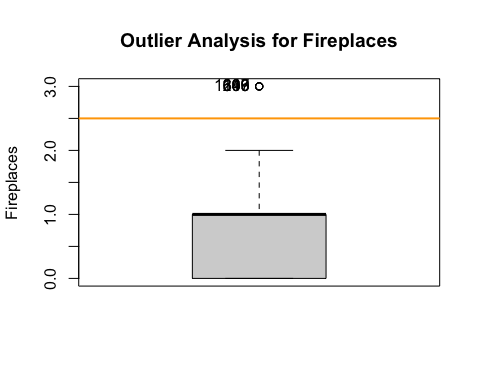
##   
## 0 1 2 3   
## 690 650 115 5

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

## [1] 0

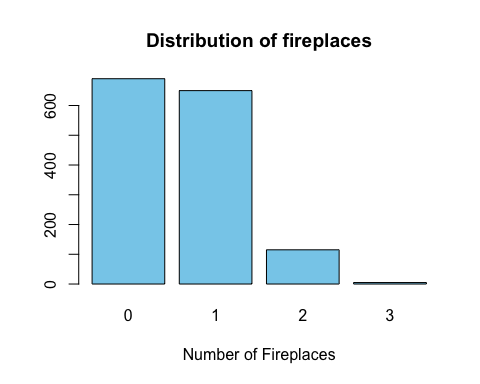
analyze\_outliers(train2, "Fireplaces")

## 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)

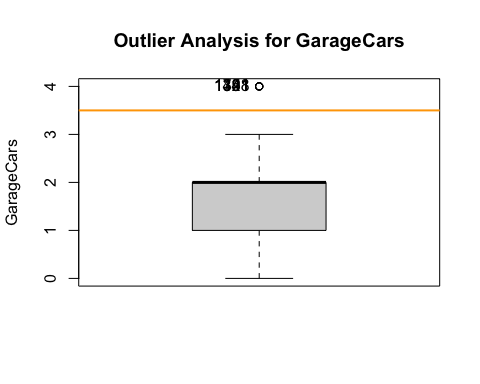
##   
## 0 1 2 3 4   
## 81 369 824 181 5

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

## [1] 0

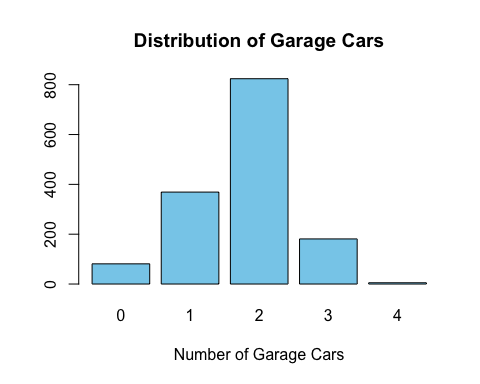
analyze\_outliers(train2, "GarageCars")

## Number of mild outliers: 5



## Number of severe outliers: 0

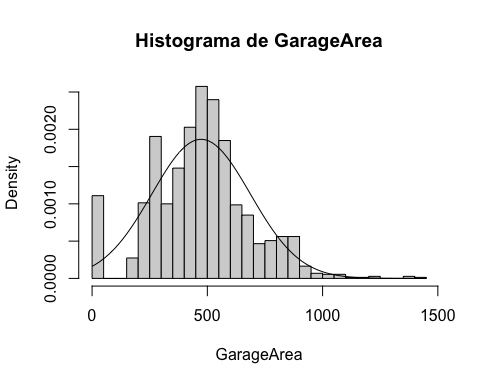
train2 <- update\_outliers\_count(train2, "GarageCars")  
barplot(table(train2$GarageCars), main = "Distribution of Garage Cars",xlab = "Number of Garage Cars",col = "skyblue")



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

numeric\_description(train2$GarageArea, 30)

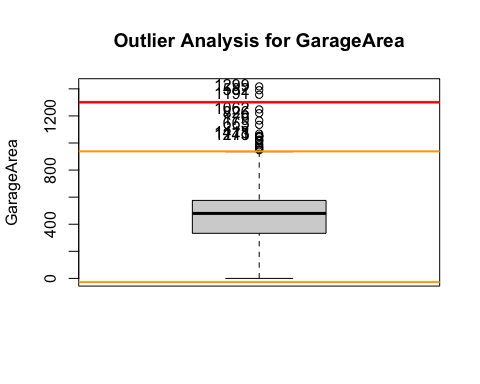
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 334.5 480.0 473.0 576.0 1418.0   
##   
## Count of missing values: 0



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

# not normal distribution  
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

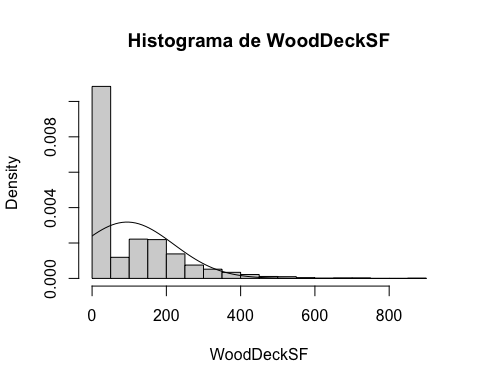
which(train$GarageFinish=='No Garage')

## [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

numeric\_description(train2$WoodDeckSF, 30) # not normal distribution

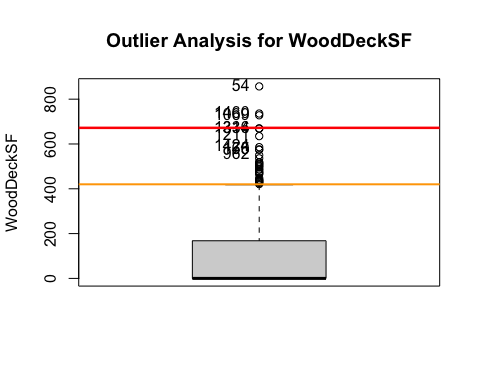
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 94.24 168.00 857.00   
##   
## Count of missing values: 0



##   
## 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



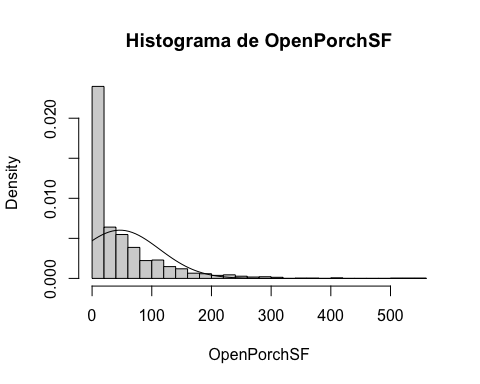
## Number of severe outliers: 3

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

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

numeric\_description(train2$OpenPorchSF, 30)

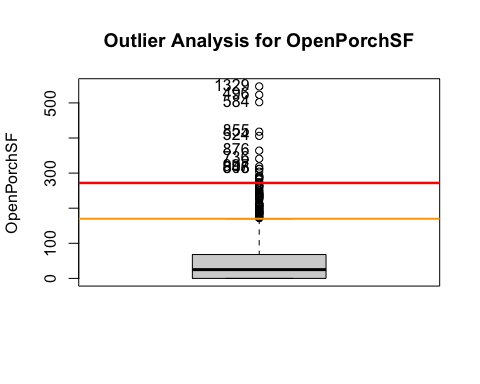
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 25.00 46.66 68.00 547.00   
##   
## Count of missing values: 0



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

analyze\_outliers(train2,"OpenPorchSF")

## Number of mild outliers: 77



## Number of severe outliers: 18

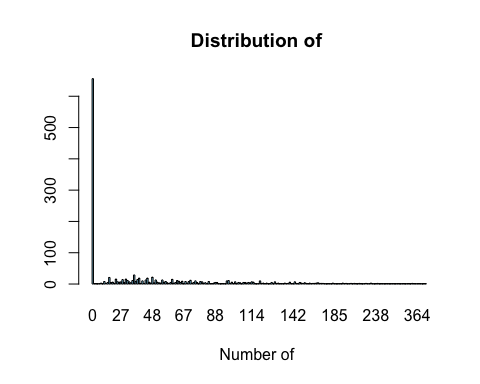
train2 <- update\_outliers\_count(train2, "OpenPorchSF")  
summary(train2$OpenPorchSF)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 25.00 46.66 68.00 547.00

#table(train2$OpenPorchSF) # very centered data in 0  
sum(is.na(train2$OpenPorchSF))

## [1] 0

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



sm<-summary(train2$OpenPorchSF)  
train2$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)

summary(train2$EnclosedPorch)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 21.95 0.00 552.00

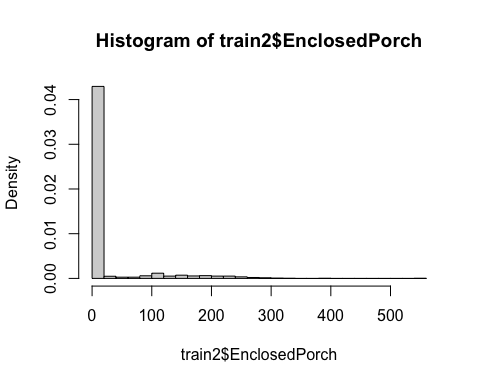
table(train2$EnclosedPorch) # very centered data in 0

##   
## 0 19 20 24 30 32 34 36 37 39 40 42 44 48 50 52   
## 1252 1 1 1 1 2 2 2 1 2 3 1 1 1 1 1   
## 54 56 60 64 67 68 70 77 80 81 84 87 90 91 94 96   
## 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 6   
## 98 99 100 102 105 108 112 114 115 116 120 123 126 128 129 130   
## 1 1 2 3 1 2 15 2 2 4 5 1 3 3 1 1   
## 134 136 137 138 140 143 144 145 148 150 154 156 158 160 162 164   
## 1 1 2 1 1 1 5 1 1 3 2 4 2 2 1 3   
## 168 169 170 172 174 176 177 180 183 184 185 189 190 192 194 196   
## 2 1 1 1 1 3 1 2 1 3 2 1 2 5 1 1   
## 198 200 202 205 208 210 212 214 216 218 220 221 224 226 228 230   
## 1 1 2 2 1 1 1 1 5 1 1 1 2 1 3 1   
## 234 236 239 240 242 244 248 252 254 259 264 268 272 275 280 286   
## 2 2 1 2 1 2 1 4 1 1 2 1 1 1 1 1   
## 291 293 294 301 318 330 386 552   
## 1 1 1 1 1 1 1 1

sum(is.na(train2$EnclosedPorch))

## [1] 0

hist(train2$EnclosedPorch, breaks = 30, freq = F)



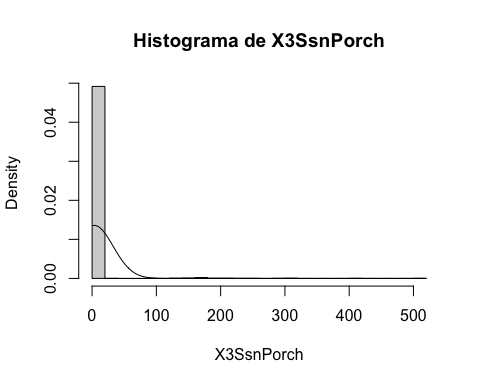
# therefore we will create a factor 0 if they dont have enclosed porch area,1 if they do have.  
train2$EnclosedPorch\_binary<-0  
train2$EnclosedPorch\_binary[which(train2$EnclosedPorch!=0)]<-1  
train2$EnclosedPorch\_binary<-as.factor(train2$EnclosedPorch\_binary)  
#table(train2$EnclosedPorch\_binary)  
  
# 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)

numeric\_description(train2$X3SsnPorch, 30)

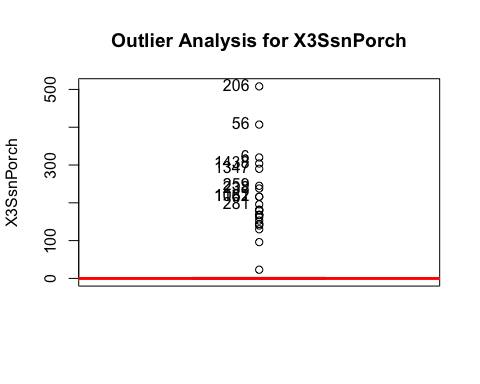
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 3.41 0.00 508.00   
##   
## Count of missing values: 0



##   
## 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



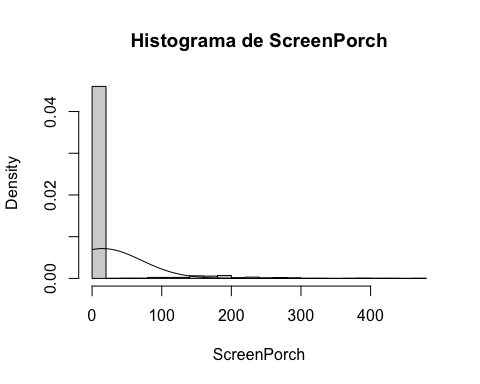
## Number of severe outliers: 24

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

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

numeric\_description(train2$ScreenPorch, 30) # not normally distributed

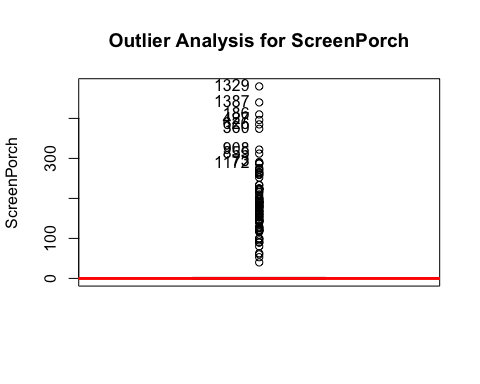
## Summary:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 15.06 0.00 480.00   
##   
## Count of missing values: 0



##   
## 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



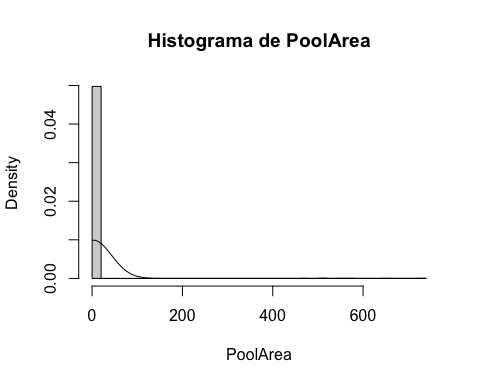
## Number of severe outliers: 116

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

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

numeric\_description(train2$PoolArea, 30)

## 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



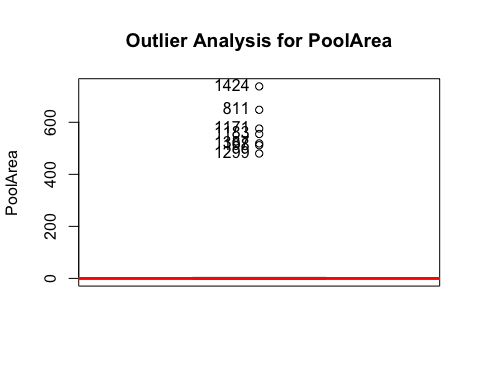
##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.041202, p-value < 2.2e-16

# not normally distributed  
table(train2$PoolArea) # very centered in 0 -> dont have pool

##   
## 0 480 512 519 555 576 648 738   
## 1453 1 1 1 1 1 1 1

analyze\_outliers(train2,"PoolArea")

## Number of mild outliers: 7



## Number of severe outliers: 7

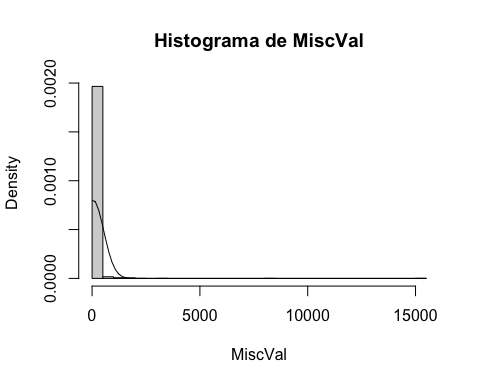
train2 <- update\_outliers\_count(train2, "PoolArea")  
# create a new variable : 0=no pool/1= pool  
train2$Pool\_binary<-0  
train2$Pool\_binary[which(train2$PoolArea!=0)]<-1  
train2$Pool\_binary<-as.factor(train2$Pool\_binary)  
table(train2$Pool\_binary) # not very usefull , the majority dont have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

##   
## 0 1   
## 1453 7

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

numeric\_description(train2$MiscVal, 30)

## 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



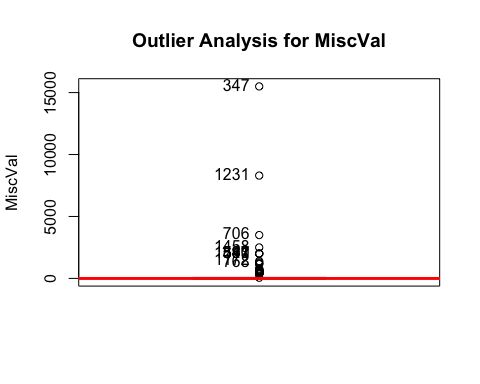
##   
## Shapiro-Wilk normality test  
##   
## data: variable  
## W = 0.058233, p-value < 2.2e-16

# not normally distributed  
table(train2$MiscVal) # very centered in 0

##   
## 0 54 350 400 450 480 500 560 600 620 700 800 1150   
## 1408 1 1 11 4 2 8 1 4 1 5 1 1   
## 1200 1300 1400 2000 2500 3500 8300 15500   
## 2 1 1 4 1 1 1 1

analyze\_outliers(train2,"MiscVal")

## Number of mild outliers: 52

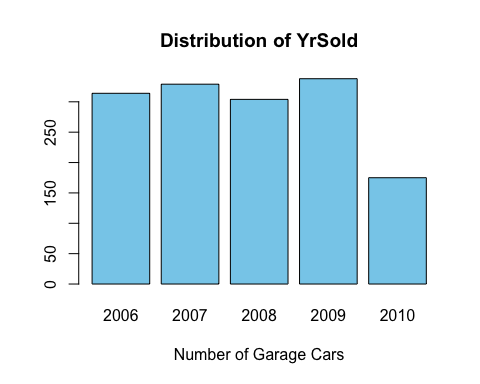


## Number of severe outliers: 52

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

## YrSold - (int)

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

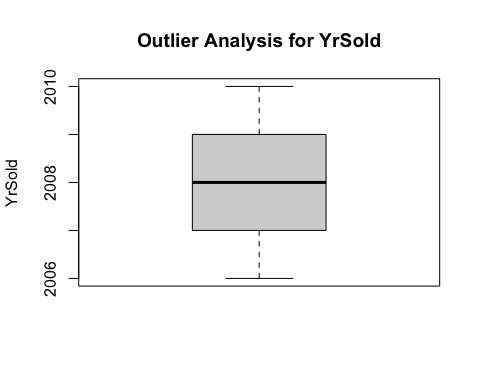


table(train2$YrSold)

##   
## 2006 2007 2008 2009 2010   
## 314 329 304 338 175

analyze\_outliers(train2,"YrSold")

## Number of mild outliers: 0



## Number of severe outliers: 0

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

### Categorical Variables

## Neighborhood

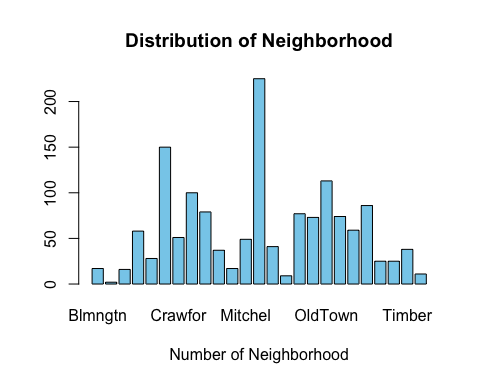
sum(is.na(train2$Neighborhood)) #0

## [1] 0

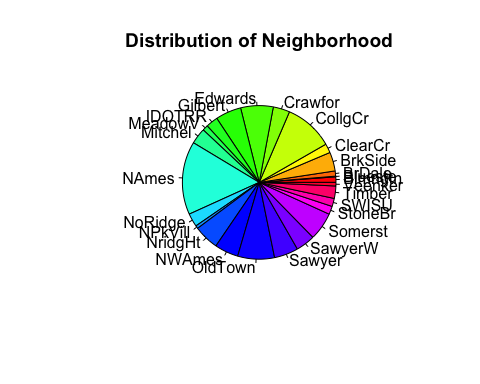
table(train2$Neighborhood)

##   
## 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

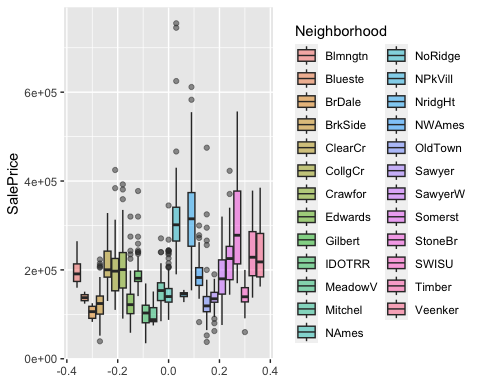
barplot(table(train2$Neighborhood), main = "Distribution of Neighborhood",xlab = "Number of Neighborhood",col = "skyblue")



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

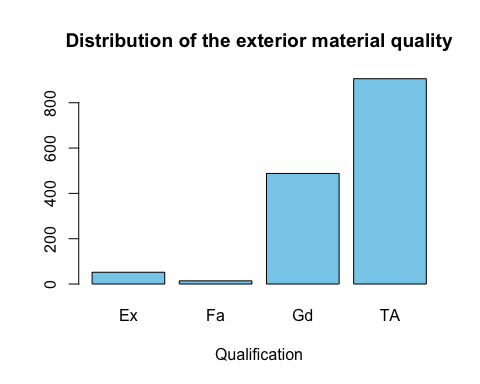
sum(is.na(train2$ExterQual)) #0

## [1] 0

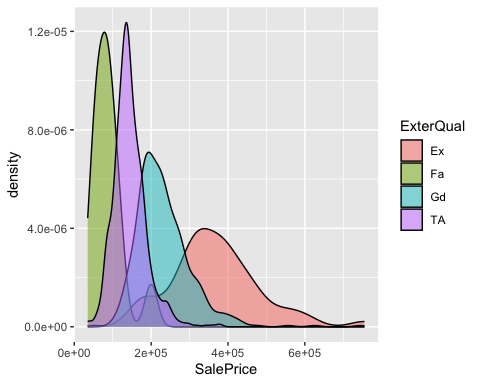
table(train2$ExterQual)

##   
## 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

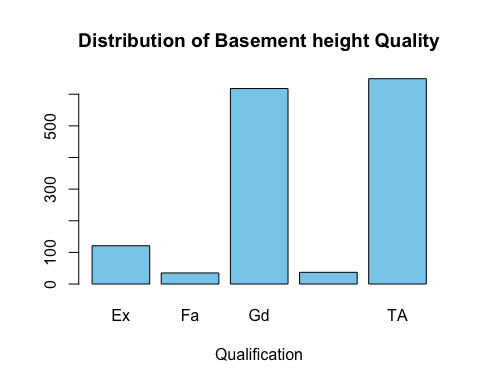
sum(is.na(train2$BsmtQual)) #0

## [1] 0

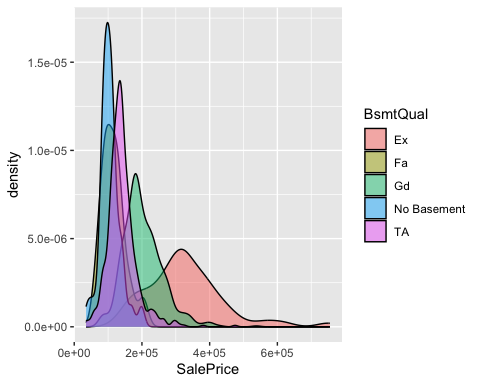
table(train2$BsmtQual)

##   
## 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

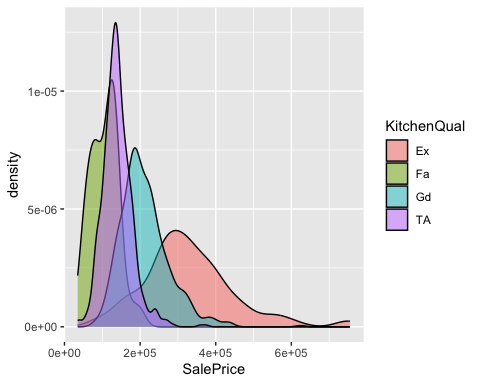
table(train2$KitchenQual)

##   
## 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

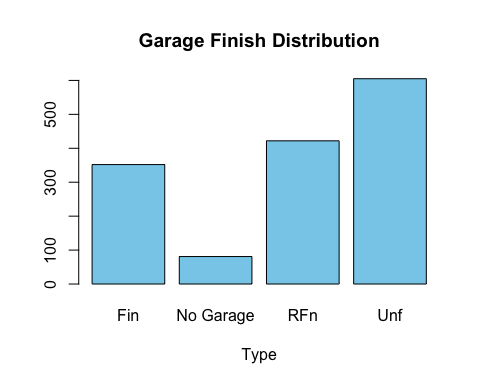
sum(is.na(train2$GarageFinish)) #0

## [1] 0

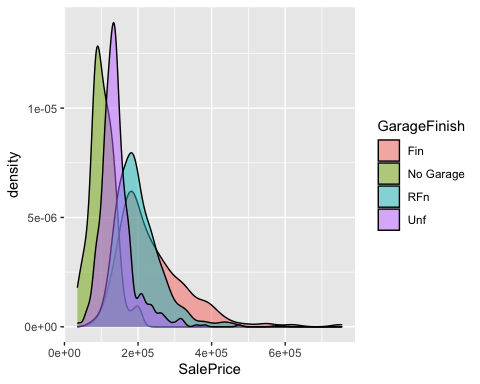
table(train2$GarageFinish)

##   
## 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

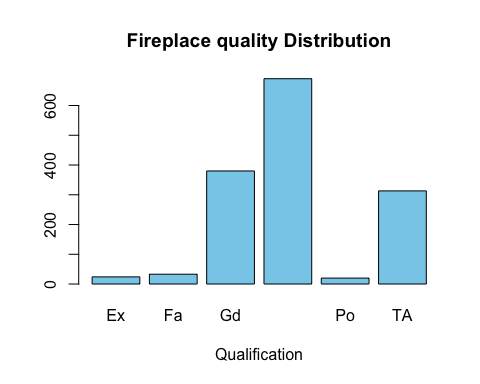
sum(is.na(train2$FireplaceQu)) #0

## [1] 0

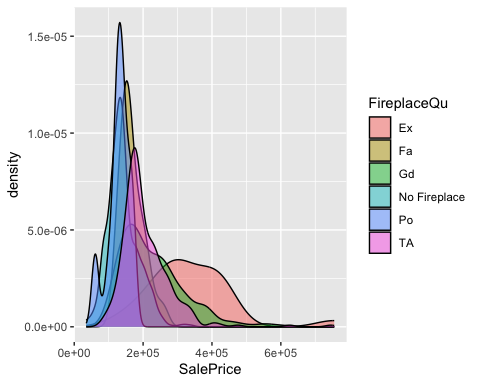
table(train2$FireplaceQu)

##   
## 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

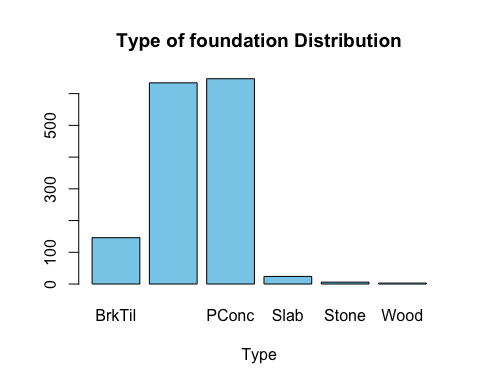
sum(is.na(train2$Foundation)) #0

## [1] 0

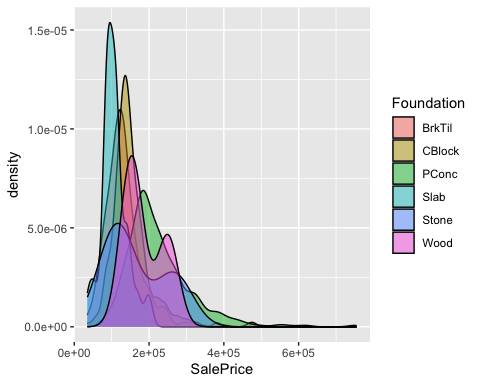
table(train2$Foundation)

##   
## 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

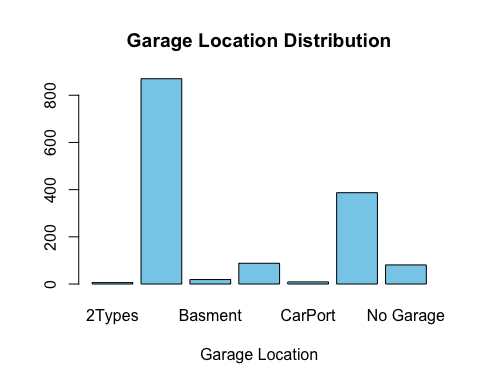
sum(is.na(train2$GarageType)) #0

## [1] 0

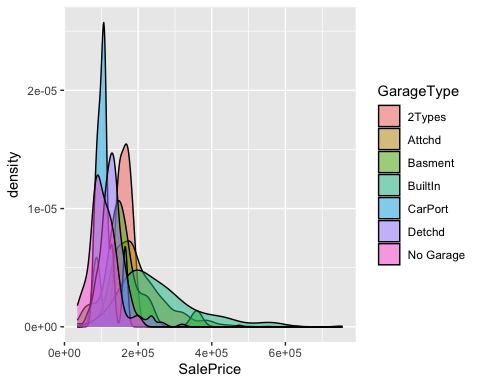
table(train2$GarageType)

##   
## 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

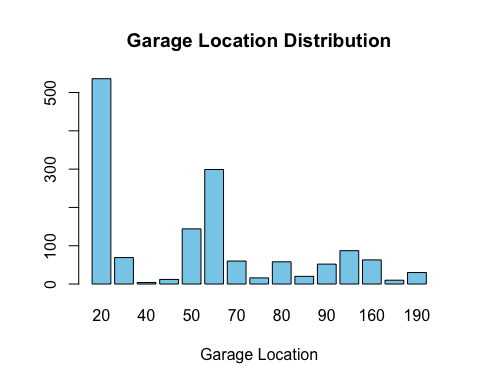
sum(is.na(train2$MSSubClass)) #0

## [1] 0

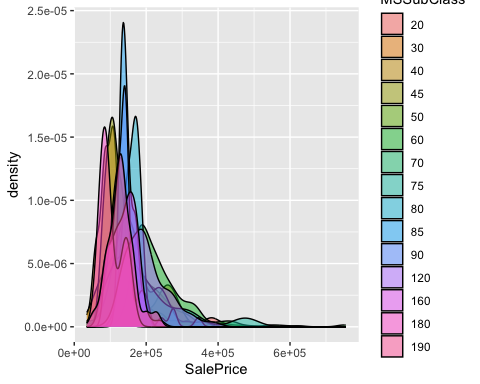
table(train2$MSSubClass) # +classes

##   
## 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

First, we check which variables have NA’s and then we decide whether we impute or not.

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

## [1] 259

sum(is.na(train2$MasVnrArea))

## [1] 10

sum(is.na(train2$GarageYrBlt))

## [1] 81

We exclude the imputation of the GarageYrBlt feature because it has no sense to impute this variable.

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

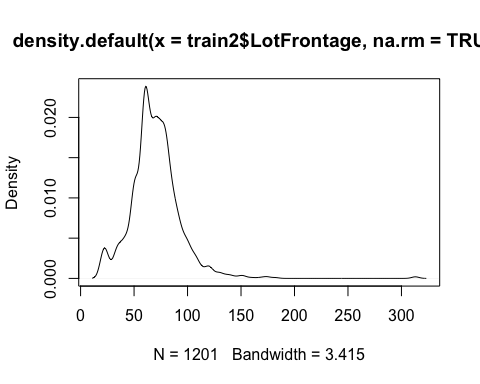
# LotFrontage validation  
summary(imputed\_data$LotFrontage)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.00 60.00 70.00 70.75 80.00 313.00

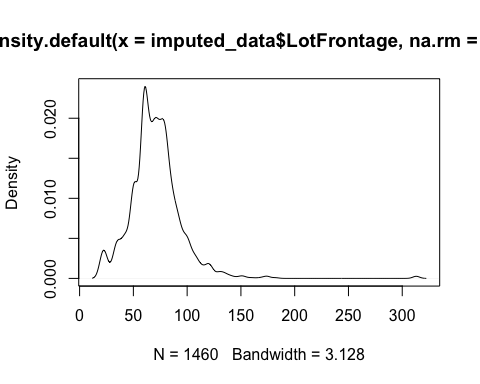
summary(train2$LotFrontage)

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

plot(density(train2$LotFrontage,na.rm=TRUE))



plot(density(imputed\_data$LotFrontage,na.rm=TRUE))

 The imputation doesn’t change much the density nor summary.

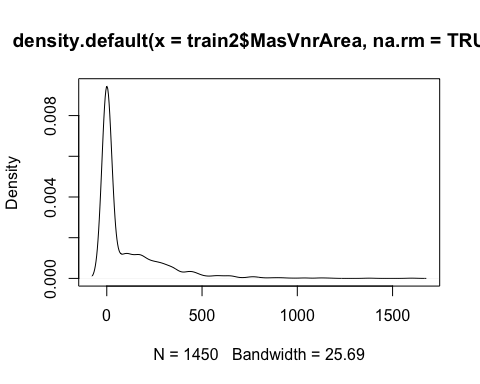
# MasVnrArea validation  
summary(imputed\_data$MasVnrArea)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 103.4 165.2 1600.0

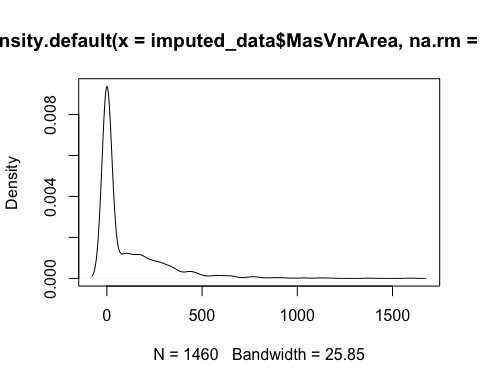
summary(train2$MasVnrArea)

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

plot(density(train2$MasVnrArea,na.rm=TRUE))



plot(density(imputed\_data$MasVnrArea,na.rm=TRUE))

 The imputation doesnt change much the density nor summary

## 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)

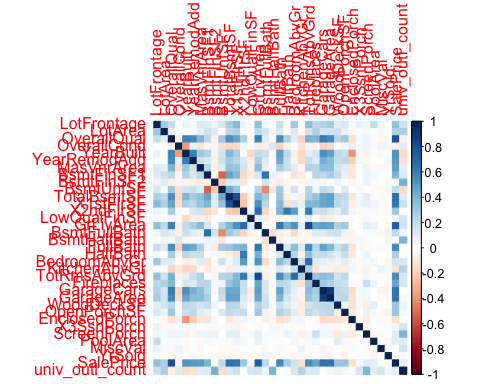
## [1] 10

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   
## 0.34017783 0.26760706 0.79585433 -0.07670080 0.52598160   
## YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF   
## 0.50827902 0.48834456 0.40869598 -0.01287439 0.21580512   
## TotalBsmtSF X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea   
## 0.65090160 0.63120195 0.31955392 -0.03309214 0.73959382   
## BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr   
## 0.22750432 -0.01689242 0.56178277 0.28454057 0.16740870   
## KitchenAbvGr TotRmsAbvGrd Fireplaces GarageCars GarageArea   
## -0.13752026 0.54089962 0.46953187 0.64249623 0.63035535   
## WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch   
## 0.32676978 0.32454345 -0.13626825 0.04468370 0.10959267   
## PoolArea MiscVal YrSold SalePrice univ\_outl\_count   
## 0.09826675 -0.02317955 -0.02774883 1.00000000 0.08206935

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

## [1] "KitchenAbvGr"

## [1] -0.1375203

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

## [1] "OverallQual"

## [1] 0.7958543

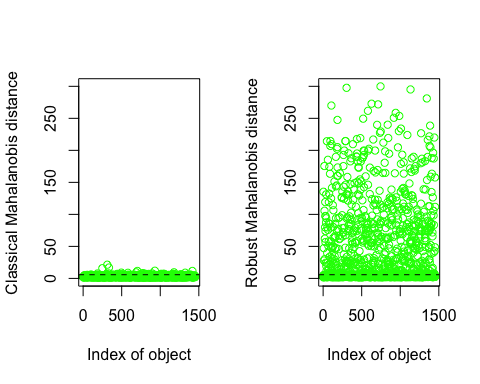
# Multivariate outliers

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.

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")



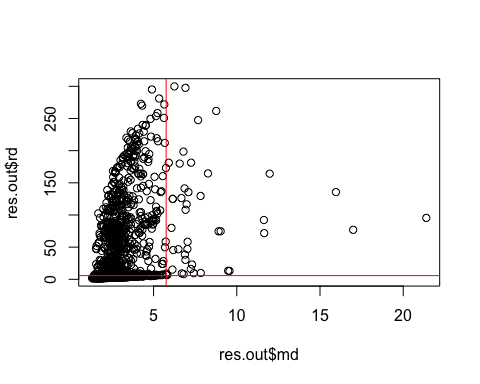
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 103 179 186 250 305 314 333 336 385 441 452 496 497 582 584 692   
## 103 179 186 249 304 313 332 335 384 440 451 495 496 579 581 688   
## 707 748 770 779 804 808 855 869 895 899 935 1012 1025 1046 1047 1049   
## 703 744 766 775 800 804 850 864 890 894 930 1007 1020 1041 1042 1044   
## 1062 1170 1174 1183 1185 1191 1269 1293 1329 1397 1424   
## 1057 1165 1169 1178 1180 1186 1264 1288 1323 1390 1417

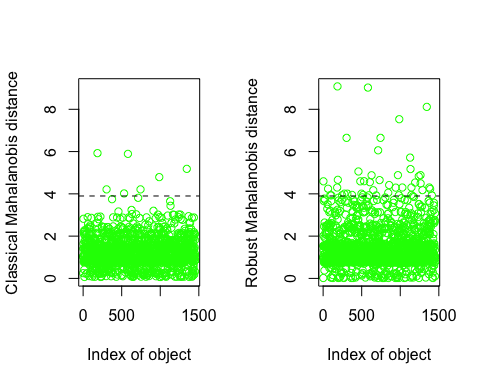
length(which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))) #44

## [1] 43

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")



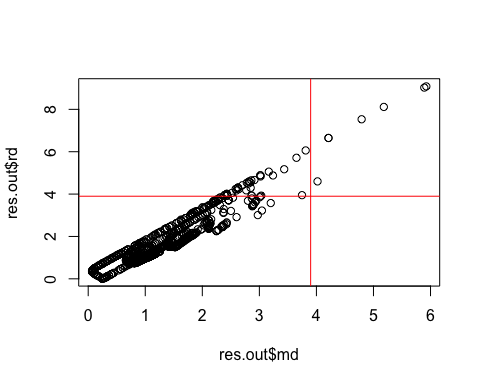
# take off m outliers  
#df = df[-which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff)),]  
  
# 2. Generally the houses built latetly are the ones with most quality  
res.out <- Moutlier(imputed\_data[,c("YearBuilt","OverallQual")], quantile = 0.9995, col="green")



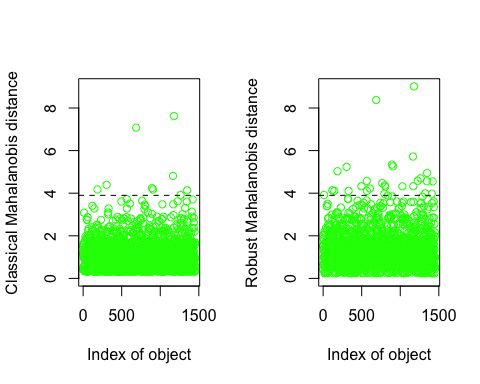
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 186 305 534 584 748 992 1350   
## 186 304 531 581 744 987 1344

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")



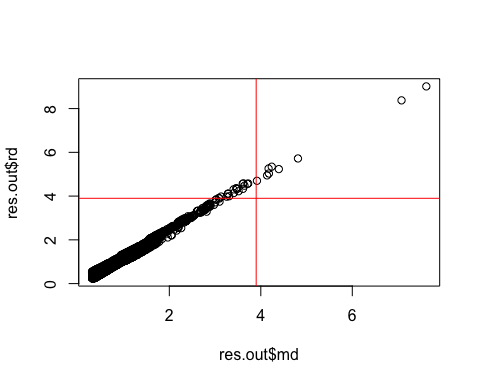
# 3. Generally the houses with more area are the ones with more rooms  
res.out <- Moutlier(imputed\_data[,c("GrLivArea","TotRmsAbvGrd")], quantile = 0.9995, col="green")



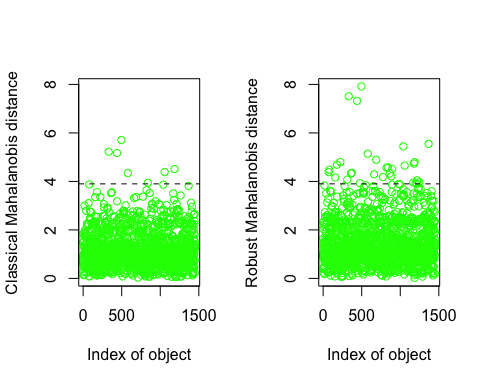
which((res.out$md > res.out$cutoff)&(res.out$rd > res.out$cutoff))

## 186 305 692 898 911 1170 1183 1269 1354   
## 186 304 688 893 906 1165 1178 1264 1348

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")



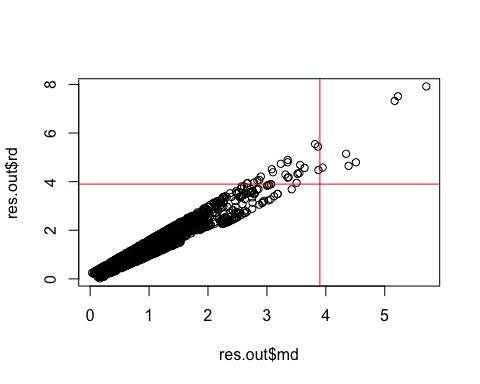
# 4. Generally the houses with more total area are the ones with more garage area also  
res.out <- Moutlier(imputed\_data[,c("GarageArea","TotalBsmtSF")], quantile = 0.9995, col="green")



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

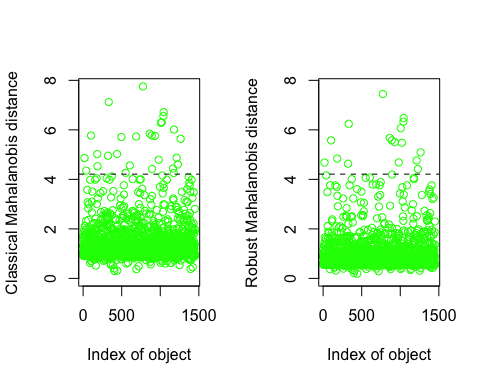
## 333 441 497 582 844 1062 1191   
## 332 440 496 579 839 1057 1186

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

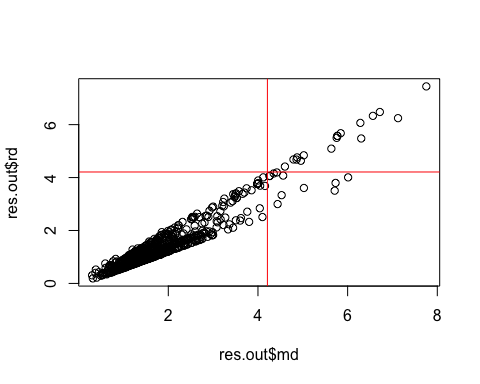
## Warning in covMcd(X): The covariance matrix has become singular during  
## the iterations of the MCD algorithm.  
## There are 572 observations (in the entire dataset of 1453 obs.) lying  
## on the plane with equation 0.70711 (x\_i1-m\_1) + -0.70711 (x\_i2-m\_2) +  
## -1.1578e-15 (x\_i3-m\_3) = 0 with (m\_1,m\_2) the mean of these  
## observations.



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 1228   
## 1269   
## 1264

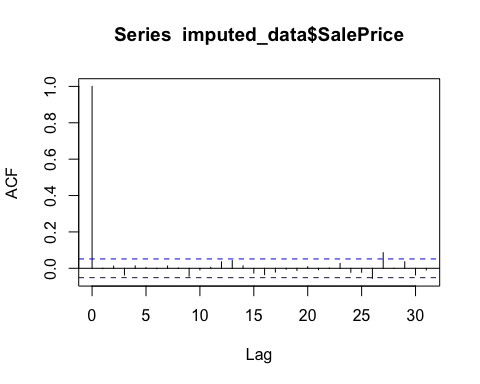
par(mfrow=c(1,1))  
plot( res.out$md, res.out$rd )  
abline(h=res.out$cutoff, col="red")  
abline(v=res.out$cutoff, col="red")



# Profiling

To test for autocorrelation the acf() function is used, of which the result can be seen below: target is not normally distributed and autocorrelation is found, so a randomized dataframe is reset

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 = 0.86925, p-value < 2.2e-16

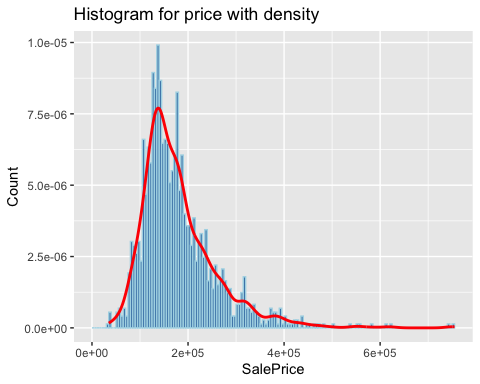
ks.test(imputed\_data$SalePrice, 'pnorm', mean(imputed\_data$SalePrice), sd(imputed\_data$SalePrice))

## 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 = 0.124, p-value < 2.2e-16  
## alternative hypothesis: two-sided

#qqnorm(imputed\_data$SalePrice) + qqline(imputed\_data$SalePrice)  
#hist(imputed\_data$SalePrice, prob = TRUE, col = 'lightblue', main = 'SalePrice Distribution', xlab = 'SalePrice') + lines(density(imputed\_data$SalePrice), col = 'red', lwd = 2)  
  
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 0.79585433 1.800118e-318  
## GrLivArea 0.73959382 8.947769e-252  
## TotalBsmtSF 0.65090160 7.455967e-176  
## GarageCars 0.64249623 5.837898e-170  
## X1stFlrSF 0.63120195 2.511377e-162  
## GarageArea 0.63035535 9.107151e-162  
## FullBath 0.56178277 1.225399e-121  
## TotRmsAbvGrd 0.54089962 3.400072e-111  
## YearBuilt 0.52598160 3.578620e-104  
## YearRemodAdd 0.50827902 2.782822e-96  
## MasVnrArea 0.48834456 6.179887e-88  
## Fireplaces 0.46953187 1.503171e-80  
## BsmtFinSF1 0.40869598 1.329980e-59  
## LotFrontage 0.34017783 1.090517e-40  
## WoodDeckSF 0.32676978 1.679594e-37  
## OpenPorchSF 0.32454345 5.485499e-37  
## X2ndFlrSF 0.31955392 7.509846e-36  
## HalfBath 0.28454057 1.826130e-28  
## LotArea 0.26760706 2.999057e-25  
## BsmtFullBath 0.22750432 1.635036e-18  
## BsmtUnfSF 0.21580512 9.004815e-17  
## BedroomAbvGr 0.16740870 1.352061e-10  
## ScreenPorch 0.10959267 2.832892e-05  
## PoolArea 0.09826675 1.756995e-04  
## univ\_outl\_count 0.08206935 1.742542e-03  
## OverallCond -0.07670080 3.439104e-03  
## EnclosedPorch -0.13626825 1.846376e-07  
## KitchenAbvGr -0.13752026 1.420791e-07

#done before for selecting  
res.con$quali

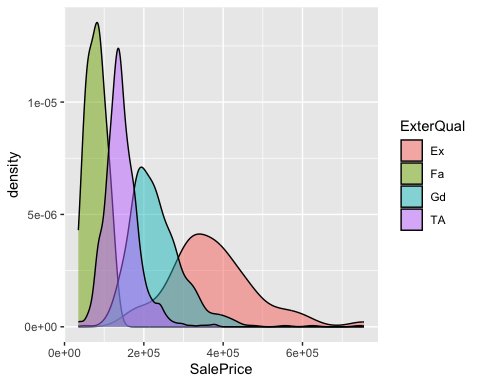
## R2 p.value  
## Neighborhood 0.54724351 1.747377e-225  
## ExterQual 0.48825569 3.437545e-210  
## BsmtQual 0.47146120 1.101824e-198  
## KitchenQual 0.46279482 6.343171e-195  
## GarageFinish 0.31012463 2.620955e-116  
## FireplaceQu 0.29573420 1.631423e-107  
## Foundation 0.25767350 4.584555e-91  
## GarageType 0.25210757 1.026767e-87  
## MSSubClass 0.24792113 5.081782e-79  
## BsmtFinType1 0.21223984 1.479192e-71  
## OpenPorch\_binary 0.17109800 3.776037e-61  
## HasPorch\_binary 0.09173386 3.314957e-32  
## EnclosedPorch\_binary 0.03423831 1.171882e-12  
## secondfloor 0.01860862 1.791339e-07  
## Pool\_binary 0.01146405 4.321554e-05

## Define a polytomic factor ExterQual, evaluates the quality of the material on the exterior, and argue if the average price depends on the level of ExterQual

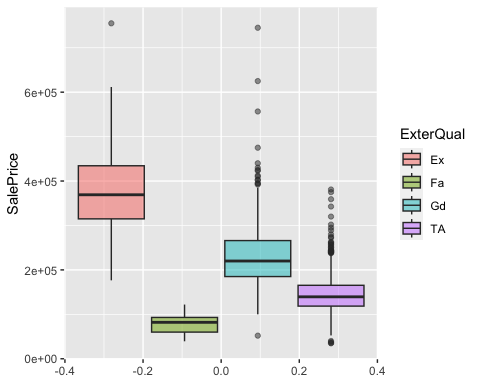
Define a polytomic factor f.age for the covariate car age according to its quartiles and argue if the average price depends on the level of age. Statistically justify the answer

Given the evidence below, we argue that the price is dependent on the level of ExterQual. Firstly, in the boxplots it can be observed that the better the quality (Q1) the lower the average of the house sale price. This is also seen clearly in the distribution plot below. —Secondly, the wilcoxon test shows that the means of these levels are not equal meaning that some relation exists between the factor and the response variable

ggplot(imputed\_data, aes(x=SalePrice, fill=ExterQual)) +  
 geom\_density(alpha=.5)



ggplot(imputed\_data, aes(y=SalePrice, fill=ExterQual)) +  
 geom\_boxplot(alpha=0.5)

 We perform pairwise wilcoxon test to check for similar means. Looking at the boxplot and distribution plot we can already expect that these are not going to be similar. The result of the wilcoxon test indicates our hypothesis was right that there is a clear difference between the means of the different quartiles

pairwise.wilcox.test(imputed\_data$SalePrice, imputed\_data$ExterQual)

##   
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction   
##   
## data: imputed\_data$SalePrice and imputed\_data$ExterQual   
##   
## Ex Fa Gd   
## Fa 7.1e-08 - -   
## Gd < 2e-16 3.0e-09 -   
## TA < 2e-16 9.9e-08 < 2e-16  
##   
## P value adjustment method: holm

## LowAge LowMidAge HighMidAge  
## LowMidAge <2e-16 - -   
## HighMidAge <2e-16 <2e-16 -   
## HighAge <2e-16 <2e-16 <2e-16  
  
#We perform pairwise wilcoxon test to check for similar means. Looking at the boxplot   
#and distribution plot we can already expect that these are not going to be similar. The   
#result of the wilcoxon test indicates our hypothesis was right that there is a clear   
#difference between the means of the different quartiles

## 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.

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] 0.8341127

## 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] 0.8000323

## 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] 0.7926493

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

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.

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)   
summary(m15)   
as.matrix(AIC(m14, m15))  
  
# delete garagearea bv very correlated to garage  
display(m15<- lm(log(SalePrice) ~OverallQual+log(GrLivArea)+GarageCars+TotalBsmtSF+YearBuilt+YearRemodAdd+Fireplaces+BsmtFinSF1+LotFrontage+WoodDeckSF, data = train\_new))  
anova(m14,m15)   
summary(m15)   
as.matrix(AIC(m14, m15))

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

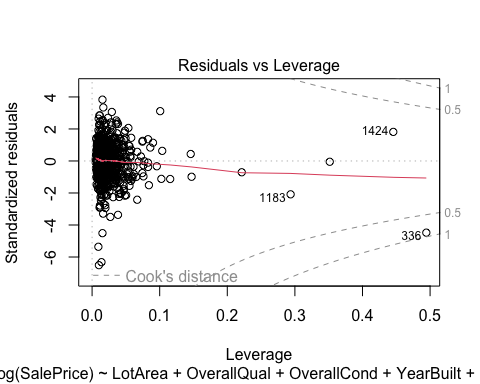
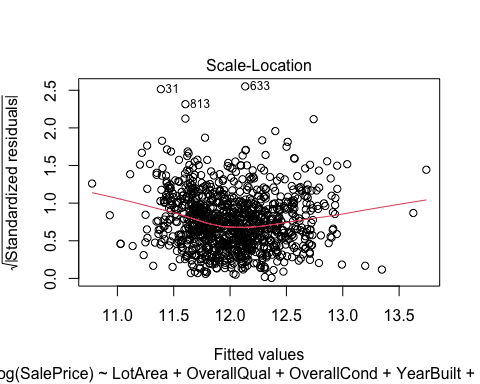
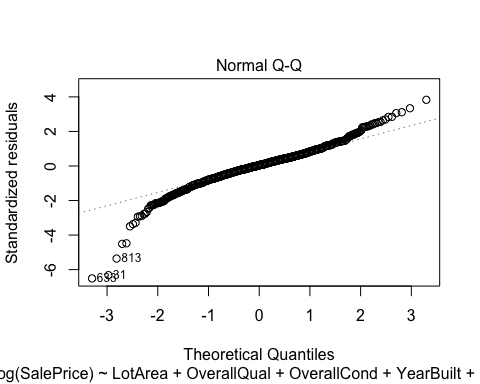
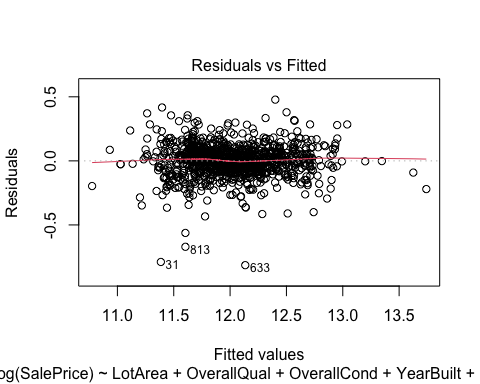
mod.fow <- stats::step(lm(SalePrice~. , data = train\_new), trace = FALSE,  
 direction = "forward")  
summary(mod.fow)

##   
## 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 + secondfloor + OpenPorch\_binary +   
## EnclosedPorch\_binary + HasPorch\_binary + Pool\_binary, data = train\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -93965 -10558 689 10598 155483   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.340e+06 1.217e+06 1.101 0.271202   
## Id 2.215e+00 1.871e+00 1.184 0.236790   
## LotFrontage 1.365e+02 4.690e+01 2.911 0.003687 \*\*   
## LotArea 7.830e-01 1.384e-01 5.658 2.06e-08 \*\*\*  
## OverallQual 8.266e+03 1.173e+03 7.045 3.66e-12 \*\*\*  
## OverallCond 5.698e+03 9.487e+02 6.006 2.75e-09 \*\*\*  
## YearBuilt 4.392e+02 9.472e+01 4.637 4.05e-06 \*\*\*  
## YearRemodAdd 1.129e+02 6.153e+01 1.834 0.066924 .   
## MasVnrArea 2.187e+01 5.723e+00 3.821 0.000142 \*\*\*  
## BsmtFinSF1 5.012e+01 5.990e+00 8.368 < 2e-16 \*\*\*  
## BsmtFinSF2 3.436e+01 8.422e+00 4.080 4.91e-05 \*\*\*  
## BsmtUnfSF 2.813e+01 5.394e+00 5.216 2.27e-07 \*\*\*  
## TotalBsmtSF NA NA NA NA   
## X1stFlrSF 4.866e+01 6.076e+00 8.009 3.53e-15 \*\*\*  
## X2ndFlrSF 7.992e+01 7.191e+00 11.114 < 2e-16 \*\*\*  
## LowQualFinSF 3.981e+01 2.147e+01 1.854 0.064024 .   
## GrLivArea NA NA NA NA   
## BsmtFullBath 1.168e+03 2.328e+03 0.502 0.615893   
## BsmtHalfBath 2.474e+03 4.579e+03 0.540 0.589135   
## FullBath 6.010e+03 2.685e+03 2.238 0.025466 \*   
## HalfBath 3.238e+03 2.442e+03 1.326 0.185255   
## BedroomAbvGr -8.257e+03 1.640e+03 -5.034 5.80e-07 \*\*\*  
## KitchenAbvGr -9.483e+03 7.069e+03 -1.342 0.180075   
## TotRmsAbvGrd 1.230e+03 1.140e+03 1.080 0.280595   
## Fireplaces 6.013e+03 3.008e+03 1.999 0.045913 \*   
## GarageCars 7.488e+03 2.607e+03 2.873 0.004166 \*\*   
## GarageArea -2.908e+00 9.019e+00 -0.322 0.747183   
## WoodDeckSF 1.165e+01 6.927e+00 1.682 0.092828 .   
## OpenPorchSF 2.853e+01 1.641e+01 1.739 0.082326 .   
## EnclosedPorch -8.400e+00 3.265e+01 -0.257 0.797018   
## X3SsnPorch 2.183e+01 2.442e+01 0.894 0.371633   
## ScreenPorch 3.325e+01 1.875e+01 1.774 0.076449 .   
## PoolArea -7.331e+02 2.066e+02 -3.549 0.000407 \*\*\*  
## MiscVal -2.942e+00 2.859e+00 -1.029 0.303711   
## YrSold -1.200e+03 5.986e+02 -2.004 0.045368 \*   
## NeighborhoodBlueste -1.563e+03 2.638e+04 -0.059 0.952749   
## NeighborhoodBrDale -1.488e+04 1.252e+04 -1.189 0.234823   
## NeighborhoodBrkSide -1.629e+04 1.065e+04 -1.529 0.126501   
## NeighborhoodClearCr -2.913e+04 1.102e+04 -2.644 0.008343 \*\*   
## NeighborhoodCollgCr -2.557e+04 9.148e+03 -2.795 0.005301 \*\*   
## NeighborhoodCrawfor -9.023e+02 1.033e+04 -0.087 0.930420   
## NeighborhoodEdwards -2.838e+04 9.929e+03 -2.859 0.004351 \*\*   
## NeighborhoodGilbert -2.704e+04 9.485e+03 -2.850 0.004467 \*\*   
## NeighborhoodIDOTRR -3.189e+04 1.129e+04 -2.824 0.004842 \*\*   
## NeighborhoodMeadowV -2.402e+03 1.337e+04 -0.180 0.857451   
## NeighborhoodMitchel -3.458e+04 1.003e+04 -3.448 0.000592 \*\*\*  
## NeighborhoodNAmes -2.639e+04 9.621e+03 -2.742 0.006220 \*\*   
## NeighborhoodNoRidge -1.728e+03 1.055e+04 -0.164 0.869972   
## NeighborhoodNPkVill 1.745e+03 1.314e+04 0.133 0.894356   
## NeighborhoodNridgHt -9.326e+03 9.473e+03 -0.984 0.325149   
## NeighborhoodNWAmes -3.832e+04 9.837e+03 -3.896 0.000105 \*\*\*  
## NeighborhoodOldTown -3.036e+04 1.027e+04 -2.955 0.003205 \*\*   
## NeighborhoodSawyer -2.759e+04 9.986e+03 -2.763 0.005838 \*\*   
## NeighborhoodSawyerW -2.659e+04 9.574e+03 -2.777 0.005593 \*\*   
## NeighborhoodSomerst -1.241e+04 9.504e+03 -1.306 0.191954   
## NeighborhoodStoneBr 2.082e+04 1.025e+04 2.032 0.042427 \*   
## NeighborhoodSWISU -2.553e+04 1.176e+04 -2.171 0.030167 \*   
## NeighborhoodTimber -2.860e+04 9.988e+03 -2.863 0.004287 \*\*   
## NeighborhoodVeenker -6.038e+03 1.247e+04 -0.484 0.628464   
## ExterQualFa -2.427e+04 1.068e+04 -2.271 0.023361 \*   
## ExterQualGd -1.985e+04 5.317e+03 -3.733 0.000201 \*\*\*  
## ExterQualTA -2.145e+04 5.951e+03 -3.604 0.000331 \*\*\*  
## BsmtQualFa -2.314e+04 7.221e+03 -3.204 0.001403 \*\*   
## BsmtQualGd -2.498e+04 4.173e+03 -5.988 3.07e-09 \*\*\*  
## BsmtQualNo Basement 7.486e+03 1.210e+04 0.619 0.536200   
## BsmtQualTA -2.310e+04 5.024e+03 -4.597 4.89e-06 \*\*\*  
## KitchenQualFa -1.719e+04 7.189e+03 -2.391 0.016991 \*   
## KitchenQualGd -1.601e+04 4.110e+03 -3.895 0.000106 \*\*\*  
## KitchenQualTA -1.702e+04 4.590e+03 -3.708 0.000222 \*\*\*  
## GarageFinishNo Garage 3.226e+04 1.312e+04 2.460 0.014097 \*   
## GarageFinishRFn -8.386e+02 2.462e+03 -0.341 0.733491   
## GarageFinishUnf 2.247e+03 2.976e+03 0.755 0.450401   
## FireplaceQuFa -1.795e+04 8.434e+03 -2.128 0.033612 \*   
## FireplaceQuGd -7.458e+03 6.397e+03 -1.166 0.244015   
## FireplaceQuNo Fireplace -4.739e+02 7.739e+03 -0.061 0.951182   
## FireplaceQuPo -4.483e+02 9.377e+03 -0.048 0.961882   
## FireplaceQuTA -1.002e+04 6.696e+03 -1.496 0.135086   
## FoundationCBlock 2.458e+03 3.644e+03 0.675 0.500128   
## FoundationPConc 6.272e+03 3.968e+03 1.581 0.114293   
## FoundationSlab 2.638e+03 1.132e+04 0.233 0.815810   
## FoundationStone 1.279e+04 1.491e+04 0.857 0.391405   
## FoundationWood -2.918e+04 1.543e+04 -1.892 0.058835 .   
## GarageTypeAttchd 1.559e+04 1.173e+04 1.329 0.184103   
## GarageTypeBasment 1.747e+04 1.437e+04 1.216 0.224260   
## GarageTypeBuiltIn 1.797e+04 1.243e+04 1.446 0.148624   
## GarageTypeCarPort 1.230e+04 1.539e+04 0.799 0.424331   
## GarageTypeDetchd 1.998e+04 1.165e+04 1.715 0.086764 .   
## GarageTypeNo Garage NA NA NA NA   
## MSSubClass30 3.771e+03 5.479e+03 0.688 0.491479   
## MSSubClass40 -1.641e+03 1.827e+04 -0.090 0.928462   
## MSSubClass45 1.845e+02 1.007e+04 0.018 0.985388   
## MSSubClass50 9.891e+03 6.758e+03 1.464 0.143656   
## MSSubClass60 -7.009e+02 6.129e+03 -0.114 0.908982   
## MSSubClass70 1.584e+03 7.721e+03 0.205 0.837532   
## MSSubClass75 7.821e+03 1.210e+04 0.646 0.518353   
## MSSubClass80 6.811e+03 4.932e+03 1.381 0.167661   
## MSSubClass85 -7.291e+02 6.970e+03 -0.105 0.916703   
## MSSubClass90 -2.173e+04 7.322e+03 -2.968 0.003073 \*\*   
## MSSubClass120 -2.130e+04 4.771e+03 -4.465 9.01e-06 \*\*\*  
## MSSubClass160 -2.360e+04 8.460e+03 -2.789 0.005391 \*\*   
## MSSubClass180 -3.496e+04 1.508e+04 -2.319 0.020626 \*   
## MSSubClass190 -4.817e+03 8.021e+03 -0.601 0.548247   
## BsmtFinType1BLQ 4.470e+03 3.287e+03 1.360 0.174256   
## BsmtFinType1GLQ 3.939e+03 2.984e+03 1.320 0.187164   
## BsmtFinType1LwQ 5.733e+02 4.438e+03 0.129 0.897245   
## BsmtFinType1No Basement NA NA NA NA   
## BsmtFinType1Rec -7.355e+02 3.427e+03 -0.215 0.830104   
## BsmtFinType1Unf 4.301e+03 3.269e+03 1.315 0.188678   
## univ\_outl\_count -3.756e+01 2.539e+03 -0.015 0.988202   
## secondfloor1 -2.195e+04 7.077e+03 -3.101 0.001985 \*\*   
## OpenPorch\_binary1 -2.867e+03 5.023e+03 -0.571 0.568270   
## EnclosedPorch\_binary1 4.541e+03 6.931e+03 0.655 0.512511   
## HasPorch\_binary1 1.345e+02 5.195e+03 0.026 0.979350   
## Pool\_binary1 5.541e+05 1.347e+05 4.115 4.22e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23460 on 907 degrees of freedom  
## Multiple R-squared: 0.9256, Adjusted R-squared: 0.9167   
## F-statistic: 103.6 on 109 and 907 DF, p-value: < 2.2e-16

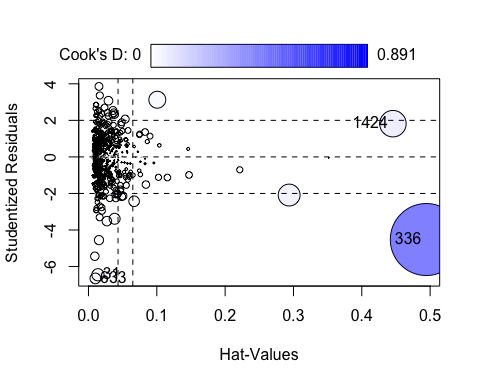
# Models:   
  
n00<- lm(log(SalePrice) ~ LotArea + OverallQual + OverallCond + YearBuilt +  
 MasVnrArea + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +  
 X1stFlrSF + X2ndFlrSF +  
 BsmtFullBath + FullBath + HalfBath + BedroomAbvGr +  
 KitchenAbvGr + Fireplaces + GarageCars +  
 OpenPorchSF + X3SsnPorch +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

## [1] 0.9043544

plot(n00)



influencePlot(n00)



## StudRes Hat CookD  
## 31 -6.449544 0.0138726 0.02555599  
## 336 -4.521449 0.4944162 0.89130698  
## 1424 1.818272 0.4455391 0.12047707  
## 633 -6.648943 0.0100976 0.01964477

#corrplot(descrCor, method = 'number')   
# check if there are any correlated predictors  
treshold\_vif<-function(rsq){  
 vif<- 1/(1-rsq)  
 return(vif)  
}  
  
treshold\_vif(summary(n00)$r.squared)

## [1] 10.45527

vif(n00) # there are not

## LotArea OverallQual OverallCond YearBuilt MasVnrArea BsmtFinSF1   
## 1.214905 3.327787 1.263210 2.879749 1.403376 5.217512   
## BsmtFinSF2 BsmtUnfSF X1stFlrSF X2ndFlrSF BsmtFullBath FullBath   
## 1.463428 4.023406 4.356839 4.455454 1.972188 2.854595   
## HalfBath BedroomAbvGr KitchenAbvGr Fireplaces GarageCars OpenPorchSF   
## 2.238717 1.686953 1.279118 1.509298 1.970413 1.181613   
## X3SsnPorch ScreenPorch PoolArea   
## 1.017610 1.078405 1.031469

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)

# 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|)   
## -0.56093 -5.6351 1.749e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 9

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|)   
## 0.59555 -2.6347 0.008421 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 2

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|)  
## 0.98748 -0.1429 0.8864  
##   
## iterations = 2

boxTidwell(log(SalePrice) ~ I(PoolArea+0.01),~ log(LotArea)+ OverallQual + OverallCond + YearBuilt +  
 BsmtFinSF1 + BsmtFinSF2+X1stFlrSF + X2ndFlrSF+  
 BsmtFullBath + KitchenAbvGr + Fireplaces + GarageCars +  
 ScreenPorch,data=train\_new, max.iter = 70 ) # no either

## Warning in boxTidwell.default(y, X1, X2, max.iter = max.iter, tol = tol, :  
## maximum iterations exceeded

## MLE of lambda Score Statistic (z) Pr(>|z|)   
## 20.465 1.711 0.08708 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## iterations = 71

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

## [1] 0.9028157

#plot(v1)  
#influencePlot(v1)  
which(as.vector(vif(v1))>treshold\_vif(summary(v1)$r.squared))# there are not highly correlated variables in the model

## integer(0)

Therefore we have 2 potential models. We have been looking on the accuracy of the model, currently on 90% (will deppend on the seed), but we need to check if there is overfitting using the test sample.

We shall check

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 24295.81  
## 2 Model with test\_new sample 0.89 24796.11

Seems that the model has no overfitting. No significant difference on the measures of the model between the train sample and the test.

Now we shall proceed to adding the categorical variables for our model creation. We now add to the model

#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] 0.9210496

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 1002 15.973   
## 2 976 12.976 26 2.9968 8.6696 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(v1,b0)

## df AIC  
## v1 16 -1306.228  
## b0 42 -1465.549

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

## [1] 12.66618

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

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.975298 1 1.405453  
## log(LotArea) 3.014180 1 1.736140  
## OverallQual 3.579863 1 1.892053  
## OverallCond 1.375475 1 1.172806  
## YearBuilt 6.311398 1 2.512250  
## BsmtFinSF1 2.237723 1 1.495902  
## log(X1stFlrSF) 4.424458 1 2.103440  
## X2ndFlrSF 4.678568 1 2.163000  
## BsmtFullBath 1.896918 1 1.377286  
## FullBath 2.538274 1 1.593196  
## TotRmsAbvGrd 4.135935 1 2.033700  
## KitchenAbvGr 1.452889 1 1.205358  
## Fireplaces 1.651287 1 1.285024  
## GarageCars 2.157259 1 1.468761  
## ScreenPorch 1.099950 1 1.048785  
## PoolArea 1.152381 1 1.073490  
## Neighborhood 47.091953 24 1.083560

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] 0.9081543

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 1002 16.557   
## 2 997 15.095 5 1.4621 19.313 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(n0,b1)

## df AIC  
## n0 16 -1269.667  
## b1 21 -1353.687

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

## [1] 10.88782

vif(b1)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.867932 1 1.366723  
## log(LotArea) 1.956537 1 1.398762  
## OverallQual 3.552590 1 1.884832  
## OverallCond 1.290987 1 1.136216  
## YearBuilt 2.591520 1 1.609820  
## BsmtFinSF1 2.135441 1 1.461315  
## log(X1stFlrSF) 4.001874 1 2.000469  
## X2ndFlrSF 4.085653 1 2.021300  
## BsmtFullBath 1.833923 1 1.354224  
## FullBath 2.300029 1 1.516585  
## TotRmsAbvGrd 3.911187 1 1.977672  
## KitchenAbvGr 1.359035 1 1.165776  
## Fireplaces 1.480379 1 1.216708  
## GarageCars 1.999640 1 1.414086  
## ScreenPorch 1.075744 1 1.037181  
## PoolArea 1.131648 1 1.063789  
## ExterQual 2.911112 3 1.194932

# 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] 0.9184348

anova(b1,b11) # better with 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)   
## 1 997 15.095   
## 2 958 13.406 39 1.6897 3.0961 1.48e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b1,b11)

## df AIC  
## b1 21 -1353.687  
## b11 60 -1396.413

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] 0.9131775

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)   
## 1 997 15.095   
## 2 993 14.270 4 0.82558 14.363 2.098e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b1,b2)

## df AIC  
## b1 21 -1353.687  
## b2 25 -1402.887

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

## [1] 11.51775

vif(b2)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.875693 1 1.369560  
## log(LotArea) 1.969617 1 1.403430  
## OverallQual 3.807098 1 1.951179  
## OverallCond 1.316540 1 1.147406  
## YearBuilt 3.275175 1 1.809744  
## BsmtFinSF1 2.187675 1 1.479079  
## log(X1stFlrSF) 4.013927 1 2.003479  
## X2ndFlrSF 4.136780 1 2.033908  
## BsmtFullBath 1.848409 1 1.359562  
## FullBath 2.399474 1 1.549024  
## TotRmsAbvGrd 3.994653 1 1.998663  
## KitchenAbvGr 1.414644 1 1.189388  
## Fireplaces 1.485518 1 1.218818  
## GarageCars 2.022029 1 1.421981  
## ScreenPorch 1.080429 1 1.039437  
## PoolArea 1.133058 1 1.064452  
## ExterQual 3.691654 3 1.243188  
## BsmtQual 5.039855 4 1.224059

# 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] 0.9223772

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)   
## 1 993 14.270   
## 2 929 12.758 64 1.512 1.7204 0.0005468 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b2,b21)

## df AIC  
## b2 25 -1402.887  
## b21 89 -1388.797

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 13.406   
## 2 929 12.758 29 0.64795 1.627 0.02002 \*  
## ---  
## 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] 0.9144228

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)   
## 1 993 14.270   
## 2 990 14.065 3 0.20468 4.8024 0.002511 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b2,b3)

## df AIC  
## b2 25 -1402.887  
## b3 28 -1411.581

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

## [1] 11.68536

vif(b3)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.888593 1 1.374261  
## log(LotArea) 1.970501 1 1.403745  
## OverallQual 3.909470 1 1.977238  
## OverallCond 1.393790 1 1.180589  
## YearBuilt 3.372218 1 1.836360  
## BsmtFinSF1 2.208846 1 1.486219  
## log(X1stFlrSF) 4.037036 1 2.009238  
## X2ndFlrSF 4.153888 1 2.038109  
## BsmtFullBath 1.859646 1 1.363688  
## FullBath 2.430228 1 1.558919  
## TotRmsAbvGrd 4.021209 1 2.005295  
## KitchenAbvGr 1.426154 1 1.194217  
## Fireplaces 1.487218 1 1.219516  
## GarageCars 2.046531 1 1.430570  
## ScreenPorch 1.086863 1 1.042527  
## PoolArea 1.137195 1 1.066394  
## ExterQual 5.748045 3 1.338402  
## BsmtQual 5.522987 4 1.238146  
## KitchenQual 4.819954 3 1.299692

#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] 0.9265619

anova(b3,b31) # better model with 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 + 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)   
## 1 990 14.065   
## 2 931 12.070 59 1.9951 2.6083 2.008e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b3,b31)

## df AIC  
## b3 28 -1411.581  
## b31 87 -1449.157

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] 0.914606

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)  
## 1 990 14.065   
## 2 987 14.035 3 0.030108 0.7058 0.5487

AIC(b3,b4)

## df AIC  
## b3 28 -1411.581  
## b4 31 -1407.760

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

## [1] 11.71043

vif(b4)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.903963 1 1.379842  
## log(LotArea) 1.975375 1 1.405480  
## OverallQual 3.917590 1 1.979290  
## OverallCond 1.395857 1 1.181464  
## YearBuilt 3.628547 1 1.904874  
## BsmtFinSF1 2.217642 1 1.489175  
## log(X1stFlrSF) 4.065513 1 2.016312  
## X2ndFlrSF 4.169498 1 2.041935  
## BsmtFullBath 1.863068 1 1.364943  
## FullBath 2.483751 1 1.575992  
## TotRmsAbvGrd 4.077258 1 2.019222  
## KitchenAbvGr 1.457353 1 1.207209  
## Fireplaces 1.517010 1 1.231670  
## GarageCars 3.054954 1 1.747843  
## ScreenPorch 1.088798 1 1.043455  
## PoolArea 1.139799 1 1.067614  
## ExterQual 5.923042 3 1.345109  
## BsmtQual 5.791210 4 1.245507  
## KitchenQual 4.908840 3 1.303656  
## GarageFinish 3.933006 3 1.256379

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] 0.9151732

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)  
## 1 990 14.065   
## 2 985 13.942 5 0.12333 1.7427 0.1222

AIC(b3,b5)

## df AIC  
## b3 28 -1411.581  
## b5 33 -1410.538

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

## [1] 11.78873

vif(b5)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.892490 1 1.375678  
## log(LotArea) 1.980670 1 1.407363  
## OverallQual 3.953619 1 1.988371  
## OverallCond 1.396254 1 1.181632  
## YearBuilt 3.442847 1 1.855491  
## BsmtFinSF1 2.243131 1 1.497709  
## log(X1stFlrSF) 4.114852 1 2.028510  
## X2ndFlrSF 4.273009 1 2.067126  
## BsmtFullBath 1.876237 1 1.369758  
## FullBath 2.450299 1 1.565343  
## TotRmsAbvGrd 4.091756 1 2.022809  
## KitchenAbvGr 1.442281 1 1.200950  
## Fireplaces 5.970216 1 2.443403  
## GarageCars 2.054792 1 1.433455  
## ScreenPorch 1.093622 1 1.045764  
## PoolArea 1.152598 1 1.073591  
## ExterQual 5.820066 3 1.341183  
## BsmtQual 5.941477 4 1.249502  
## KitchenQual 5.034650 3 1.309166  
## FireplaceQu 9.209674 5 1.248603

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] 0.9166394

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)   
## 1 990 14.065   
## 2 985 13.701 5 0.3643 5.2383 9.38e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b3,b6)

## df AIC  
## b3 28 -1411.581  
## b6 33 -1428.270

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

## [1] 11.99608

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

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.937643 1 1.391992  
## log(LotArea) 1.976554 1 1.405900  
## OverallQual 4.001029 1 2.000257  
## OverallCond 1.452379 1 1.205147  
## YearBuilt 5.031820 1 2.243172  
## BsmtFinSF1 2.251905 1 1.500635  
## log(X1stFlrSF) 4.078537 1 2.019539  
## X2ndFlrSF 4.204773 1 2.050554  
## BsmtFullBath 1.865495 1 1.365831  
## FullBath 2.464679 1 1.569930  
## TotRmsAbvGrd 4.039245 1 2.009787  
## KitchenAbvGr 1.434011 1 1.197502  
## Fireplaces 1.520984 1 1.233282  
## GarageCars 2.081777 1 1.442837  
## ScreenPorch 1.088810 1 1.043460  
## PoolArea 1.139721 1 1.067577  
## ExterQual 5.970330 3 1.346893  
## BsmtQual 20.871029 4 1.461985  
## KitchenQual 5.117119 3 1.312716  
## Foundation 17.646508 5 1.332496

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] 0.9151456

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)  
## 1 990 14.065   
## 2 984 13.946 6 0.11879 1.3969 0.2127

AIC(b3,b7)

## df AIC  
## b3 28 -1411.581  
## b7 34 -1408.207

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

## [1] 11.78489

vif(b7)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.929411 1 1.389032  
## log(LotArea) 1.994671 1 1.412328  
## OverallQual 3.969837 1 1.992445  
## OverallCond 1.402487 1 1.184266  
## YearBuilt 4.007888 1 2.001971  
## BsmtFinSF1 2.224160 1 1.491362  
## log(X1stFlrSF) 4.125141 1 2.031044  
## X2ndFlrSF 4.354338 1 2.086705  
## BsmtFullBath 1.864199 1 1.365357  
## FullBath 2.464694 1 1.569934  
## TotRmsAbvGrd 4.073226 1 2.018223  
## KitchenAbvGr 1.533635 1 1.238400  
## Fireplaces 1.521632 1 1.233544  
## GarageCars 3.203902 1 1.789945  
## ScreenPorch 1.094556 1 1.046210  
## PoolArea 1.140647 1 1.068011  
## ExterQual 5.987169 3 1.347525  
## BsmtQual 5.846841 4 1.246996  
## KitchenQual 5.055832 3 1.310083  
## GarageType 5.047246 6 1.144426

# 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] 0.9280488

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)   
## 1 984 13.946   
## 2 923 11.825 61 2.1207 2.7135 2.09e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b7,b71)

## df AIC  
## b7 34 -1408.207  
## b71 95 -1453.960

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

## [1] 13.89831

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] 0.9171122

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)   
## 1 984 13.946   
## 2 970 13.623 14 0.32322 1.6439 0.06215 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

AIC(b7,b8)

## df AIC  
## b7 34 -1408.207  
## b8 48 -1404.054

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

## [1] 12.0645

vif(b8) # highly correlated

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 2.071201 1 1.439167  
## log(LotArea) 2.888622 1 1.699595  
## OverallQual 4.253619 1 2.062430  
## OverallCond 1.480082 1 1.216586  
## YearBuilt 8.938208 1 2.989684  
## BsmtFinSF1 2.261404 1 1.503797  
## log(X1stFlrSF) 4.790107 1 2.188631  
## X2ndFlrSF 8.888887 1 2.981424  
## BsmtFullBath 1.927904 1 1.388490  
## FullBath 2.546685 1 1.595834  
## TotRmsAbvGrd 4.319798 1 2.078412  
## KitchenAbvGr 3.518620 1 1.875799  
## Fireplaces 1.604504 1 1.266690  
## GarageCars 3.266001 1 1.807208  
## ScreenPorch 1.110814 1 1.053952  
## PoolArea 1.167002 1 1.080279  
## ExterQual 6.343276 3 1.360564  
## BsmtQual 7.548953 4 1.287466  
## KitchenQual 5.496469 3 1.328456  
## GarageType 6.645431 6 1.170964  
## MSSubClass 482.254064 14 1.246899

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] 0.9153675

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)  
## 1 984 13.946   
## 2 979 13.910 5 0.036479 0.5135 0.7662

AIC(b7,b9)

## df AIC  
## b7 34 -1408.207  
## b9 39 -1400.870

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

## [1] 11.8158

AIC(b7,b71,b3,b31,b2,b21,b1,b11) # b31 better but too many df, -> we keep b3

## df AIC  
## b7 34 -1408.207  
## b71 95 -1453.960  
## b3 28 -1411.581  
## b31 87 -1449.157  
## b2 25 -1402.887  
## b21 89 -1388.797  
## b1 21 -1353.687  
## b11 60 -1396.413

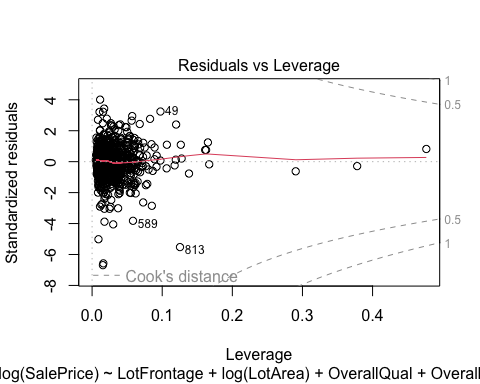
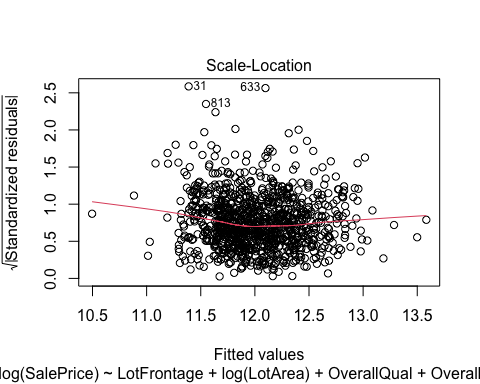
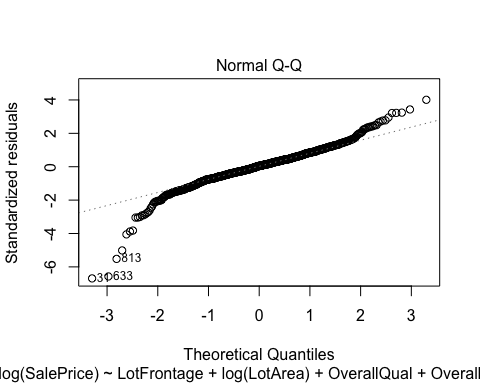
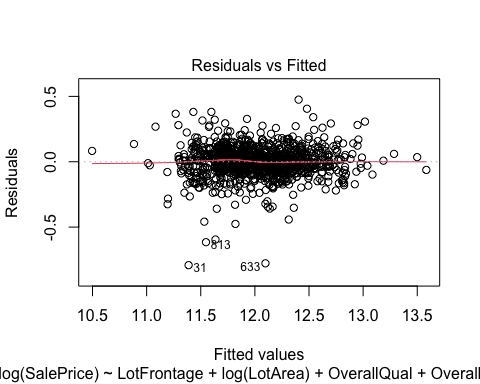
# now we have done a forward stepwise, we can see if we need all the variables doing a backwards stepwise, we can see that in the model: LotFrontage PoolArea and GarageType are less influental, for better explainability we can not consider it in our model.  
final\_model<-b3  
summary(final\_model)

##   
## 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   
## -0.79122 -0.05883 0.00545 0.06505 0.47511   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.587e+00 5.078e-01 5.094 4.20e-07 \*\*\*  
## LotFrontage -5.327e-05 2.162e-04 -0.246 0.805433   
## log(LotArea) 7.729e-02 1.055e-02 7.327 4.89e-13 \*\*\*  
## OverallQual 5.982e-02 5.264e-03 11.363 < 2e-16 \*\*\*  
## OverallCond 4.861e-02 3.954e-03 12.294 < 2e-16 \*\*\*  
## YearBuilt 2.741e-03 2.264e-04 12.109 < 2e-16 \*\*\*  
## BsmtFinSF1 7.820e-05 1.270e-05 6.156 1.08e-09 \*\*\*  
## log(X1stFlrSF) 3.857e-01 2.404e-02 16.045 < 2e-16 \*\*\*  
## X2ndFlrSF 2.322e-04 1.747e-05 13.296 < 2e-16 \*\*\*  
## BsmtFullBath 2.622e-02 9.922e-03 2.642 0.008363 \*\*   
## FullBath 1.787e-02 1.075e-02 1.663 0.096660 .   
## TotRmsAbvGrd 6.307e-03 4.630e-03 1.362 0.173436   
## KitchenAbvGr -8.063e-02 2.000e-02 -4.031 5.98e-05 \*\*\*  
## Fireplaces 2.674e-02 7.042e-03 3.797 0.000155 \*\*\*  
## GarageCars 4.811e-02 6.956e-03 6.916 8.34e-12 \*\*\*  
## ScreenPorch 2.578e-04 7.368e-05 3.498 0.000489 \*\*\*  
## PoolArea 1.167e-04 1.128e-04 1.034 0.301251   
## ExterQualFa -1.733e-01 5.128e-02 -3.380 0.000754 \*\*\*  
## ExterQualGd -2.529e-02 2.535e-02 -0.998 0.318605   
## ExterQualTA -7.255e-02 2.810e-02 -2.582 0.009971 \*\*   
## BsmtQualFa -1.636e-01 3.370e-02 -4.856 1.39e-06 \*\*\*  
## BsmtQualGd -9.679e-02 1.875e-02 -5.163 2.94e-07 \*\*\*  
## BsmtQualNo Basement -2.275e-01 3.341e-02 -6.809 1.71e-11 \*\*\*  
## BsmtQualTA -1.186e-01 2.181e-02 -5.439 6.75e-08 \*\*\*  
## KitchenQualFa -7.630e-02 3.368e-02 -2.265 0.023714 \*   
## KitchenQualGd -4.612e-02 1.922e-02 -2.400 0.016568 \*   
## KitchenQualTA -7.674e-02 2.123e-02 -3.616 0.000315 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1192 on 990 degrees of freedom  
## Multiple R-squared: 0.9144, Adjusted R-squared: 0.9122   
## F-statistic: 406.9 on 26 and 990 DF, p-value: < 2.2e-16

vif(final\_model)

## GVIF Df GVIF^(1/(2\*Df))  
## LotFrontage 1.888593 1 1.374261  
## log(LotArea) 1.970501 1 1.403745  
## OverallQual 3.909470 1 1.977238  
## OverallCond 1.393790 1 1.180589  
## YearBuilt 3.372218 1 1.836360  
## BsmtFinSF1 2.208846 1 1.486219  
## log(X1stFlrSF) 4.037036 1 2.009238  
## X2ndFlrSF 4.153888 1 2.038109  
## BsmtFullBath 1.859646 1 1.363688  
## FullBath 2.430228 1 1.558919  
## TotRmsAbvGrd 4.021209 1 2.005295  
## KitchenAbvGr 1.426154 1 1.194217  
## Fireplaces 1.487218 1 1.219516  
## GarageCars 2.046531 1 1.430570  
## ScreenPorch 1.086863 1 1.042527  
## PoolArea 1.137195 1 1.066394  
## ExterQual 5.748045 3 1.338402  
## BsmtQual 5.522987 4 1.238146  
## KitchenQual 4.819954 3 1.299692

plot(final\_model)



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 0.93 22151.87  
## 2 Model with test\_new sample 0.91 22885.14

## interpret the model

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   
## -0.79122 -0.05883 0.00545 0.06505 0.47511   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.587e+00 5.078e-01 5.094 4.20e-07 \*\*\*  
## LotFrontage -5.327e-05 2.162e-04 -0.246 0.805433   
## log(LotArea) 7.729e-02 1.055e-02 7.327 4.89e-13 \*\*\*  
## OverallQual 5.982e-02 5.264e-03 11.363 < 2e-16 \*\*\*  
## OverallCond 4.861e-02 3.954e-03 12.294 < 2e-16 \*\*\*  
## YearBuilt 2.741e-03 2.264e-04 12.109 < 2e-16 \*\*\*  
## BsmtFinSF1 7.820e-05 1.270e-05 6.156 1.08e-09 \*\*\*  
## log(X1stFlrSF) 3.857e-01 2.404e-02 16.045 < 2e-16 \*\*\*  
## X2ndFlrSF 2.322e-04 1.747e-05 13.296 < 2e-16 \*\*\*  
## BsmtFullBath 2.622e-02 9.922e-03 2.642 0.008363 \*\*   
## FullBath 1.787e-02 1.075e-02 1.663 0.096660 .   
## TotRmsAbvGrd 6.307e-03 4.630e-03 1.362 0.173436   
## KitchenAbvGr -8.063e-02 2.000e-02 -4.031 5.98e-05 \*\*\*  
## Fireplaces 2.674e-02 7.042e-03 3.797 0.000155 \*\*\*  
## GarageCars 4.811e-02 6.956e-03 6.916 8.34e-12 \*\*\*  
## ScreenPorch 2.578e-04 7.368e-05 3.498 0.000489 \*\*\*  
## PoolArea 1.167e-04 1.128e-04 1.034 0.301251   
## ExterQualFa -1.733e-01 5.128e-02 -3.380 0.000754 \*\*\*  
## ExterQualGd -2.529e-02 2.535e-02 -0.998 0.318605   
## ExterQualTA -7.255e-02 2.810e-02 -2.582 0.009971 \*\*   
## BsmtQualFa -1.636e-01 3.370e-02 -4.856 1.39e-06 \*\*\*  
## BsmtQualGd -9.679e-02 1.875e-02 -5.163 2.94e-07 \*\*\*  
## BsmtQualNo Basement -2.275e-01 3.341e-02 -6.809 1.71e-11 \*\*\*  
## BsmtQualTA -1.186e-01 2.181e-02 -5.439 6.75e-08 \*\*\*  
## KitchenQualFa -7.630e-02 3.368e-02 -2.265 0.023714 \*   
## KitchenQualGd -4.612e-02 1.922e-02 -2.400 0.016568 \*   
## KitchenQualTA -7.674e-02 2.123e-02 -3.616 0.000315 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1192 on 990 degrees of freedom  
## Multiple R-squared: 0.9144, Adjusted R-squared: 0.9122   
## F-statistic: 406.9 on 26 and 990 DF, p-value: < 2.2e-16

coefficients <- coef(final\_model)  
coefficients

## (Intercept) LotFrontage log(LotArea) OverallQual   
## 2.586628e+00 -5.327303e-05 7.728642e-02 5.982022e-02   
## OverallCond YearBuilt BsmtFinSF1 log(X1stFlrSF)   
## 4.860608e-02 2.741070e-03 7.819681e-05 3.857127e-01   
## X2ndFlrSF BsmtFullBath FullBath TotRmsAbvGrd   
## 2.322435e-04 2.621865e-02 1.786866e-02 6.306649e-03   
## KitchenAbvGr Fireplaces GarageCars ScreenPorch   
## -8.063032e-02 2.673881e-02 4.810871e-02 2.577776e-04   
## PoolArea ExterQualFa ExterQualGd ExterQualTA   
## 1.167092e-04 -1.733311e-01 -2.529165e-02 -7.254872e-02   
## BsmtQualFa BsmtQualGd BsmtQualNo Basement BsmtQualTA   
## -1.636409e-01 -9.679239e-02 -2.274507e-01 -1.186318e-01   
## KitchenQualFa KitchenQualGd KitchenQualTA   
## -7.630241e-02 -4.612220e-02 -7.674249e-02