library(dplyr)

##   
## Caricamento pacchetto: 'dplyr'

## I seguenti oggetti sono mascherati da 'package:stats':  
##   
## filter, lag

## I seguenti oggetti sono mascherati da 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(moments)  
options(scipen = 999)

houses = read.csv("house\_price.csv", stringsAsFactors = TRUE)  
dim(houses)

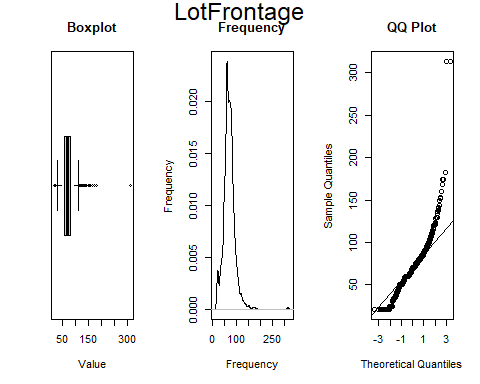
## [1] 1460 81

# Operazioni preliminari:  
# 1. Verifico quali righe e colonne hanno troppi valori mancanti  
# 2. Rimuovo righe e colonne con troppi valori mancanti  
# 3. Divido il dataset in due parti: una con le variabili numeriche e una con le variabili categoriche  
  
quant\_cont\_cols = c("LotFrontage", "LotArea", "MasVnrArea", "BsmtFinSF1", "BsmtFinSF2", "BsmtUnfSF", "TotalBsmtSF", "X1stFlrSF", "X2ndFlrSF", "LowQualFinSF", "GrLivArea", "GarageArea", "WoodDeckSF", "OpenPorchSF", "EnclosedPorch", "X3SsnPorch", "ScreenPorch")  
qual\_cols = c("MSSubClass", "MSZoning", "Street", "LotShape", "LandContour", "Utilities", "LotConfig", "LandSlope", "Neighborhood", "Condition1", "Condition2", "BldgType", "HouseStyle", "RoofStyle", "RoofMatl", "Exterior1st", "Exterior2nd", "MasVnrType", "ExterQual", "ExterCond", "Foundation", "BsmtQual", "BsmtCond", "BsmtExposure", "BsmtFinType1", "BsmtFinType2", "Heating", "HeatingQC", "CentralAir", "Electrical", "KitchenQual", "Functional", "FireplaceQu", "GarageType", "GarageFinish", "GarageQual", "GarageCond", "PavedDrive", "SaleType", "SaleCondition")  
quant\_disc\_cols = c("BsmtFullBath", "BsmtHalfBath", "FullBath", "HalfBath", "BedroomAbvGr", "KitchenAbvGr", "TotRmsAbvGrd", "Fireplaces", "GarageCars")  
year\_cols = c("YearBuilt", "YearRemodAdd", "GarageYrBlt")  
mark\_cols = c("OverallQual", "OverallCond")  
  
houses[, qual\_cols[1]] = as.factor(houses[, qual\_cols[1]])  
  
quant\_cont\_vars = houses[, quant\_cont\_cols]  
qual\_vars = houses[, qual\_cols]  
quant\_discr\_vars = houses[, quant\_disc\_cols]  
year\_vars = houses[, year\_cols]  
mark\_vars = houses[, mark\_cols]

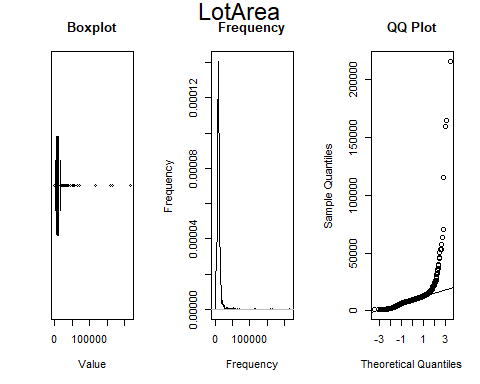
cont\_info = function (x, i) {  
 if (colnames(quant\_cont\_vars[i]) != "LotFrontage" & colnames(quant\_cont\_vars[i]) != "LotArea") {  
 x = x[x != 0]  
 }  
 print(colnames(quant\_cont\_vars[i]))  
 print(summary(x))  
 print("Skewness")  
 print(skewness(x, na.rm = TRUE))  
 print("Curtosi")  
 print(kurtosis(x, na.rm = TRUE))  
 par(mfrow = (c(1, 3)))  
 boxplot(x, main = "Boxplot", xlab = "Value", horizontal = TRUE)  
 plot(density(x, na.rm = T), main="Frequency", xlab = "Frequency", ylab = "Frequency")  
 qqnorm(x, main = "QQ Plot")  
 qqline(x)  
 mtext(text=colnames(quant\_cont\_vars[i]), line = -1.75, outer = T, cex = 1.5)  
}  
  
disc\_info = function(x, i) {  
 print(colnames(quant\_discr\_vars[i]))  
 print(summary(x, na.rm = T))  
 par(mfrow = (c(1, 1)))  
 barplot(prop.table(table(x)), main = colnames(quant\_discr\_vars[i]), xlab = "Value", ylab = "Frequency")  
}  
  
year\_info = function(x, i) {  
 par(mfrow = (c(1, 1)))  
 print("Minimo")  
 print(min(x, na.rm=T))  
 print("Massimo")  
 print(max(x, na.rm=T))  
 print("Quantili")  
 print(quantile(x, na.rm=T))  
 hist(x, main = colnames(year\_vars[i]), xlab = "Value", ylab = "Frequency")  
}  
  
  
mark\_info = function(x, i) {  
 par(mfrow = (c(1, 1)))  
 print(colnames(mark\_vars[i]))  
 print("Minimo")  
 print(min(x, na.rm=T))  
 print("Massimo")  
 print(max(x, na.rm=T))  
 print("Quantili")  
 print(quantile(x, na.rm=T))  
 barplot(prop.table(table(x)), main = colnames(mark\_vars[i]), xlab = "Value", ylab = "Frequency")  
}  
  
  
qual\_info = function (x, i) {  
 print(colnames(quant\_cont\_vars[i]))  
 print(table(x))  
 print(prop.table(table(x)))  
 par(mfrow = (c(1, 2)))  
 barplot(table(x), main = "Frequenze assolute", xlab = "Value", ylab = "Frequenze")  
 barplot(prop.table(table(x)), main = "Frequenze relative", xlab = "Value", ylab = "Frequenze")  
 mtext(text=colnames(quant\_cont\_vars[i]), line = -1.75, outer = T, cex = 1.5)  
}

for (i in seq\_along(quant\_cont\_vars)) {  
 cont\_info(quant\_cont\_vars[, i], i)  
}

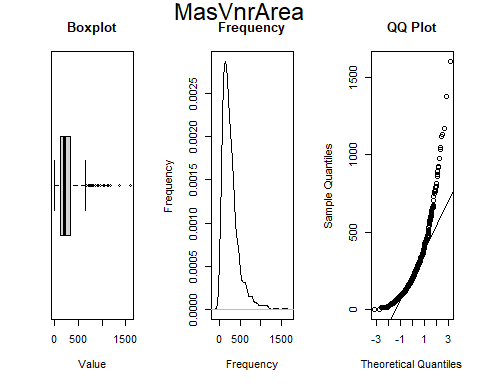
## [1] "LotFrontage"  
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 21.00 59.00 69.00 70.05 80.00 313.00 259   
## [1] "Skewness"  
## [1] 2.160866  
## [1] "Curtosi"  
## [1] 20.3753



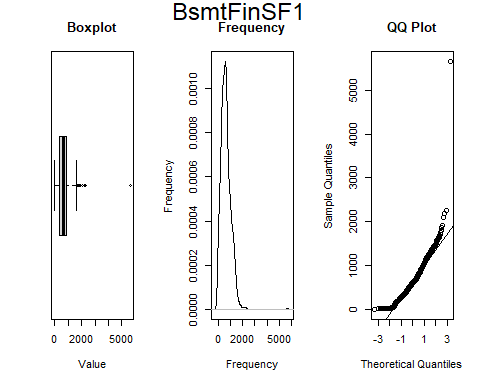
## [1] "LotArea"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1300 7554 9478 10517 11602 215245   
## [1] "Skewness"  
## [1] 12.19514  
## [1] "Curtosi"  
## [1] 205.5438



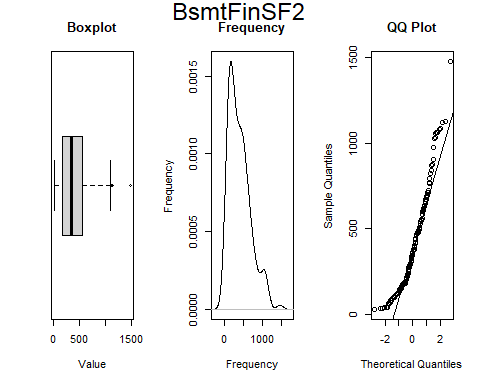
## [1] "MasVnrArea"  
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.0 113.0 203.0 254.7 330.5 1600.0 8   
## [1] "Skewness"  
## [1] 2.088559  
## [1] "Curtosi"  
## [1] 9.682093



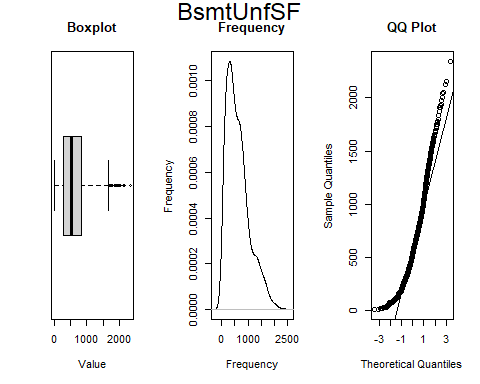
## [1] "BsmtFinSF1"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 371.0 604.0 652.3 867.0 5644.0   
## [1] "Skewness"  
## [1] 2.298795  
## [1] "Curtosi"  
## [1] 24.21043



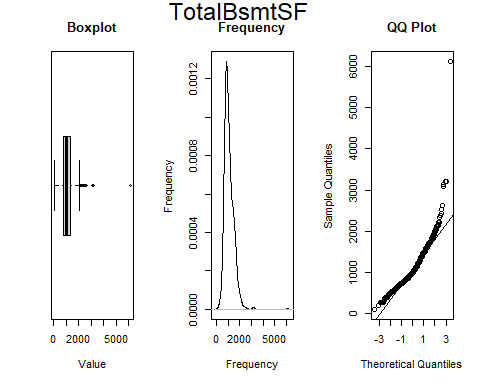
## [1] "BsmtFinSF2"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 28.0 178.5 354.0 407.0 551.0 1474.0   
## [1] "Skewness"  
## [1] 0.9852846  
## [1] "Curtosi"  
## [1] 3.668218



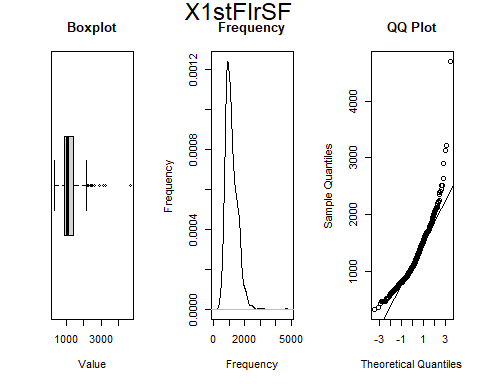
## [1] "BsmtUnfSF"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 14.0 288.0 536.0 617.1 843.2 2336.0   
## [1] "Skewness"  
## [1] 0.9695924  
## [1] "Curtosi"  
## [1] 3.549353



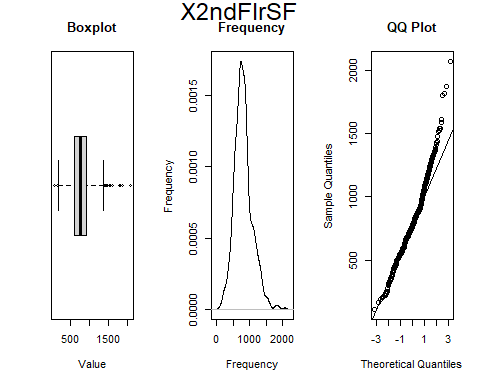
## [1] "TotalBsmtSF"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 105.0 810.5 1004.0 1084.9 1309.5 6110.0   
## [1] "Skewness"  
## [1] 2.168831  
## [1] "Curtosi"  
## [1] 20.14677



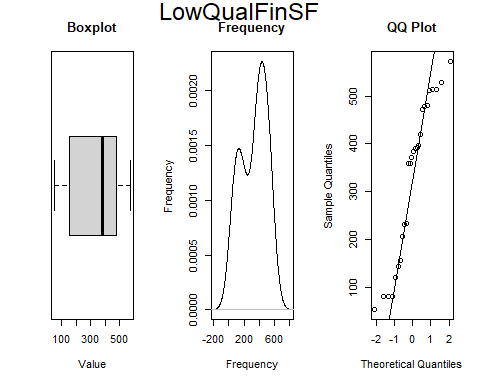
## [1] "X1stFlrSF"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 882 1087 1163 1391 4692   
## [1] "Skewness"  
## [1] 1.375342  
## [1] "Curtosi"  
## [1] 8.722076



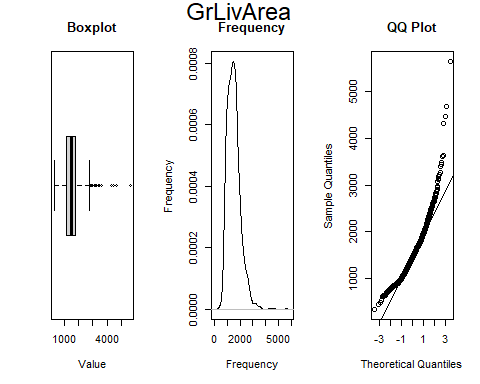
## [1] "X2ndFlrSF"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 110.0 625.0 776.0 802.9 926.5 2065.0   
## [1] "Skewness"  
## [1] 0.7011031  
## [1] "Curtosi"  
## [1] 4.273049



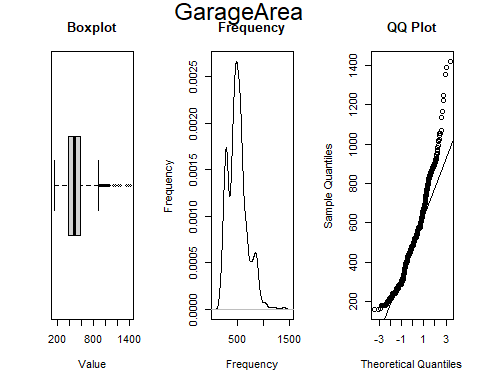
## [1] "LowQualFinSF"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 53.0 168.2 377.5 328.2 477.5 572.0   
## [1] "Skewness"  
## [1] -0.3231395  
## [1] "Curtosi"  
## [1] 1.691515



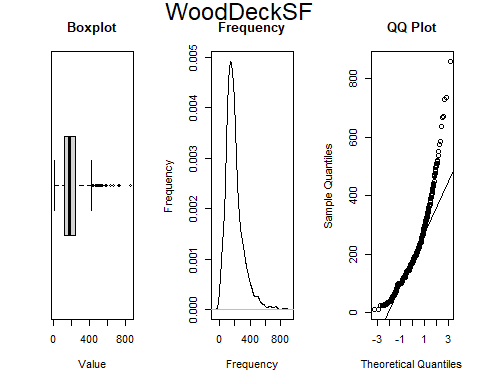
## [1] "GrLivArea"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1130 1464 1515 1777 5642   
## [1] "Skewness"  
## [1] 1.365156  
## [1] "Curtosi"  
## [1] 7.874266



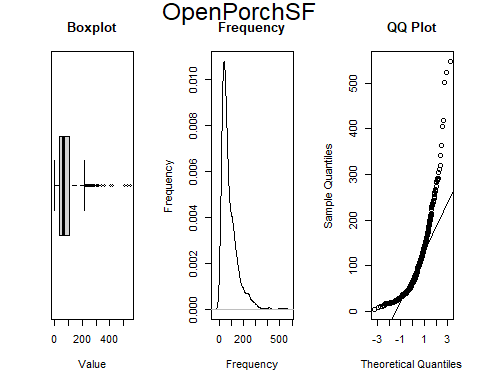
## [1] "GarageArea"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 160.0 380.0 484.0 500.8 580.0 1418.0   
## [1] "Skewness"  
## [1] 0.8101544  
## [1] "Curtosi"  
## [1] 4.18098



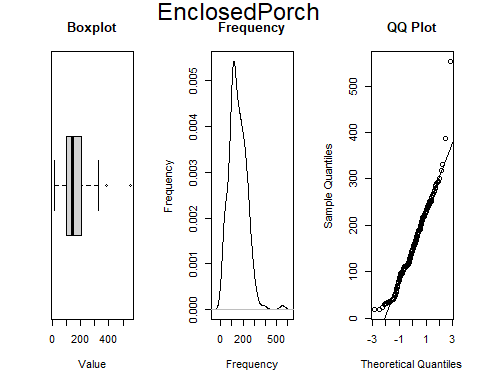
## [1] "WoodDeckSF"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.0 120.0 171.0 196.8 240.0 857.0   
## [1] "Skewness"  
## [1] 1.614144  
## [1] "Curtosi"  
## [1] 7.247074



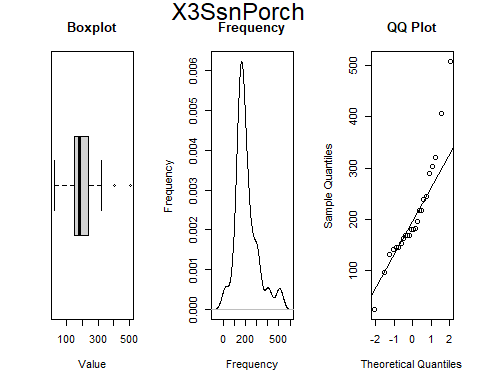
## [1] "OpenPorchSF"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 39.00 63.00 84.73 112.00 547.00   
## [1] "Skewness"  
## [1] 2.244353  
## [1] "Curtosi"  
## [1] 10.75368



## [1] "EnclosedPorch"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.0 104.2 144.5 154.1 205.0 552.0   
## [1] "Skewness"  
## [1] 0.8582936  
## [1] "Curtosi"  
## [1] 5.552907



## [1] "X3SsnPorch"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 23.0 150.8 180.0 207.4 239.8 508.0   
## [1] "Skewness"  
## [1] 1.205196  
## [1] "Curtosi"  
## [1] 4.839964



## [1] "ScreenPorch"  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 40.0 143.8 180.0 189.6 224.0 480.0   
## [1] "Skewness"  
## [1] 1.171071  
## [1] "Curtosi"  
## [1] 5.116482

