RealEstateModelling

2023-09-10

# Loading libraries

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.1 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(e1071)

# Loading the dataset

df <- read.csv("C:/Users/user/Desktop/RealEstateModelling/HousingValuationTest-V2.csv")

# Creating a data frame for numerical variables

numerical\_vars <- select(df, "YearBuilt", "FullBath", "HalfBath", "BedroomAbvGr", "KitchenAbvGr", "TotalRmsAbvGrd", "Fireplaces", "GarageCars", "MoSold", "YrSold", "LotArea", "TotalBSF", "LowQualFinSF", "LivingArea", "PoolArea", "OpenPorchSF", "SalePrice")

# Numerical variables summary statistics

summary\_stats <- summary(numerical\_vars, na.rm = TRUE)  
print(summary\_stats)

## YearBuilt FullBath HalfBath BedroomAbvGr   
## Min. :1872 Min. :0.000 Min. :0.0000 Min. :0.000   
## 1st Qu.:1954 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:2.000   
## Median :1973 Median :2.000 Median :0.0000 Median :3.000   
## Mean :1971 Mean :1.566 Mean :0.3831 Mean :2.869   
## 3rd Qu.:2001 3rd Qu.:2.000 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :2010 Max. :3.000 Max. :2.0000 Max. :8.000   
##   
## KitchenAbvGr TotalRmsAbvGrd Fireplaces GarageCars   
## Min. :0.000 Min. : 2.00 Min. :0.0000 Min. :0.000   
## 1st Qu.:1.000 1st Qu.: 5.00 1st Qu.:0.0000 1st Qu.:1.000   
## Median :1.000 Median : 6.00 Median :1.0000 Median :2.000   
## Mean :1.047 Mean : 6.52 Mean :0.6142 Mean :1.771   
## 3rd Qu.:1.000 3rd Qu.: 7.00 3rd Qu.:1.0000 3rd Qu.:2.000   
## Max. :3.000 Max. :14.00 Max. :3.0000 Max. :4.000   
##   
## MoSold YrSold LotArea TotalBSF   
## Min. : 1.000 Min. :2006 Min. : 1300 Min. : 0   
## 1st Qu.: 5.000 1st Qu.:2007 1st Qu.: 7544 1st Qu.: 799   
## Median : 6.000 Median :2008 Median : 9478 Median : 999   
## Mean : 6.319 Mean :2008 Mean : 10521 Mean :1065   
## 3rd Qu.: 8.000 3rd Qu.:2009 3rd Qu.: 11604 3rd Qu.:1304   
## Max. :12.000 Max. :2010 Max. :215245 Max. :6110   
## NA's :49   
## LowQualFinSF LivingArea PoolArea OpenPorchSF   
## Min. : 0.000 Min. : 334 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.000 1st Qu.:1131 1st Qu.: 0.00 1st Qu.: 0.00   
## Median : 0.000 Median :1467 Median : 0.00 Median : 25.00   
## Mean : 5.869 Mean :1517 Mean : 2.77 Mean : 46.37   
## 3rd Qu.: 0.000 3rd Qu.:1780 3rd Qu.: 0.00 3rd Qu.: 68.00   
## Max. :572.000 Max. :5642 Max. :738.00 Max. :547.00   
## NA's :22   
## SalePrice   
## Min. : 34900   
## 1st Qu.:130000   
## Median :163250   
## Mean :181112   
## 3rd Qu.:214000   
## Max. :755000   
##

# Numerical variables standard deviation

numerical\_sd <- round(sapply(numerical\_vars, sd), 4)  
print(numerical\_sd)

## YearBuilt FullBath HalfBath BedroomAbvGr KitchenAbvGr   
## 30.0979 0.5485 0.5030 0.8153 0.2208   
## TotalRmsAbvGrd Fireplaces GarageCars MoSold YrSold   
## 1.6246 0.6450 0.7442 2.7030 1.3287   
## LotArea TotalBSF LowQualFinSF LivingArea PoolArea   
## 10000.4637 NA 48.7219 NA 40.2598   
## OpenPorchSF SalePrice   
## 65.1386 79331.6932

# Calculating the standard deviation for TotalBSF with NA removal

std\_dev\_TotalBSF <- sd(numerical\_vars$TotalBSF, na.rm = TRUE)  
print(std\_dev\_TotalBSF)

## [1] 440.4783

# Calculating the standard deviation for LivingArea with NA removal

std\_dev\_LivingArea <- sd(numerical\_vars$LivingArea, na.rm = TRUE)  
print(std\_dev\_LivingArea)

## [1] 525.7061

# Numerical variables skewness

numerical\_skewness <- round(sapply(numerical\_vars, skewness), 4)  
print(numerical\_skewness)

## YearBuilt FullBath HalfBath BedroomAbvGr KitchenAbvGr   
## -0.6185 0.0375 0.6743 0.2144 4.4686   
## TotalRmsAbvGrd Fireplaces GarageCars MoSold YrSold   
## 0.6775 0.6465 -0.3348 0.2118 0.0988   
## LotArea TotalBSF LowQualFinSF LivingArea PoolArea   
## 12.1614 NA 8.9736 NA 14.7670   
## OpenPorchSF SalePrice   
## 2.2637 1.8897

# Calculating the skewness for TotalBSF with NA removal

skewness\_TotalBSF <- round(skewness(numerical\_vars$TotalBSF, na.rm = TRUE), 4)  
print(skewness\_TotalBSF)

## [1] 1.5414

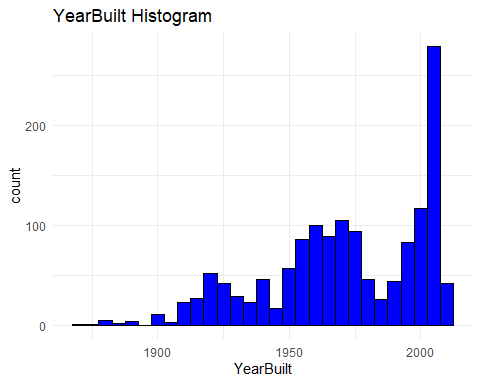
# Calculating the skewness for LivingArea with NA removal

skewness\_LivingArea <- round(skewness(numerical\_vars$LivingArea, na.rm = TRUE), 4)  
print(skewness\_LivingArea)

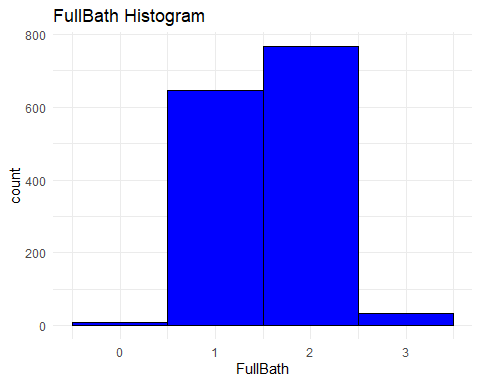
## [1] 1.3725

## Histograms of numerical variables

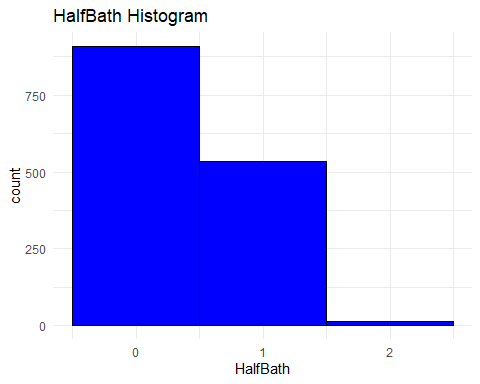
hist\_YearBuilt <- ggplot(numerical\_vars, aes(x = YearBuilt)) +  
 geom\_histogram(binwidth = 5, fill = "blue", color = "black") +  
 labs(title = "YearBuilt Histogram", x = "YearBuilt") +  
 theme\_minimal()  
print(hist\_YearBuilt)



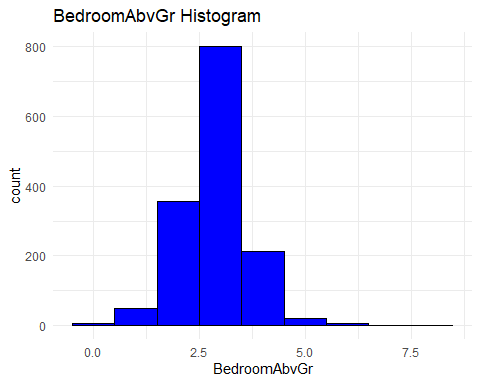
hist\_FullBath <- ggplot(numerical\_vars, aes(x = FullBath)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "FullBath Histogram", x = "FullBath") +  
 theme\_minimal()  
print(hist\_FullBath)



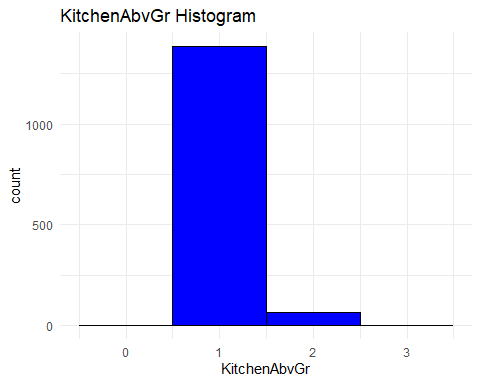
hist\_HalfBath <- ggplot(numerical\_vars, aes(x = HalfBath)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "HalfBath Histogram", x = "HalfBath") +  
 theme\_minimal()  
print(hist\_HalfBath)



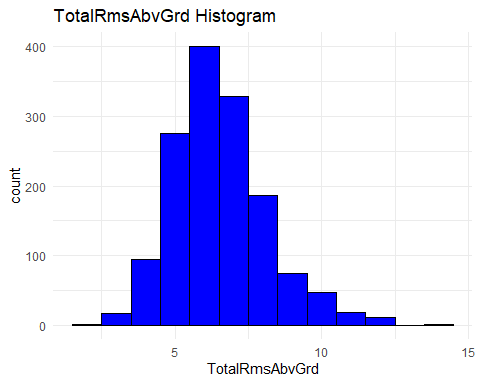
hist\_BedroomAbvGr <- ggplot(numerical\_vars, aes(x = BedroomAbvGr)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "BedroomAbvGr Histogram", x = "BedroomAbvGr") +  
 theme\_minimal()  
print(hist\_BedroomAbvGr)



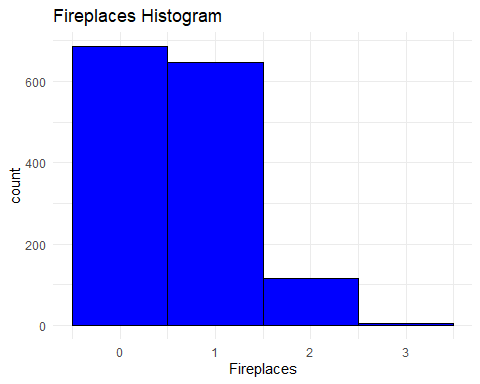
hist\_KitchenAbvGr <- ggplot(numerical\_vars, aes(x = KitchenAbvGr)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "KitchenAbvGr Histogram", x = "KitchenAbvGr") +  
 theme\_minimal()  
print(hist\_KitchenAbvGr)



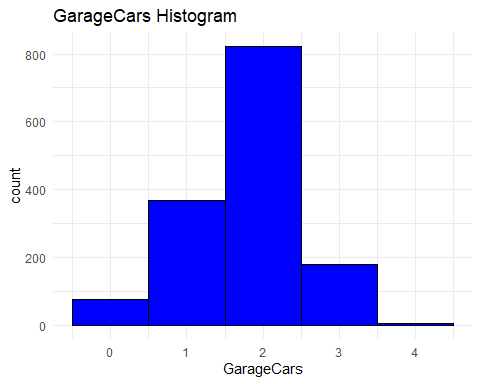
hist\_TotalRmsAbvGrd <- ggplot(numerical\_vars, aes(x = TotalRmsAbvGrd)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "TotalRmsAbvGrd Histogram", x = "TotalRmsAbvGrd") +  
 theme\_minimal()  
print(hist\_TotalRmsAbvGrd)



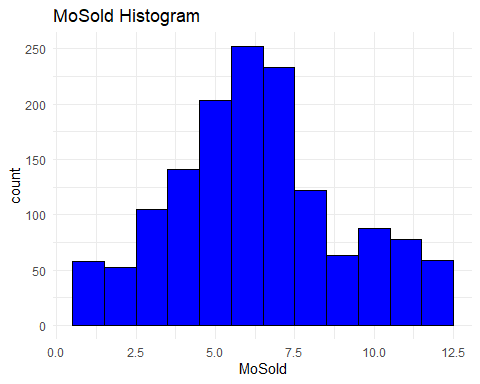
hist\_Fireplaces <- ggplot(numerical\_vars, aes(x = Fireplaces)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "Fireplaces Histogram", x = "Fireplaces") +  
 theme\_minimal()  
print(hist\_Fireplaces)



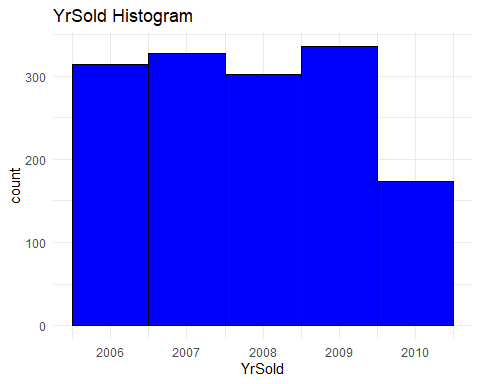
hist\_GarageCars <- ggplot(numerical\_vars, aes(x = GarageCars)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "GarageCars Histogram", x = "GarageCars") +  
 theme\_minimal()  
print(hist\_GarageCars)



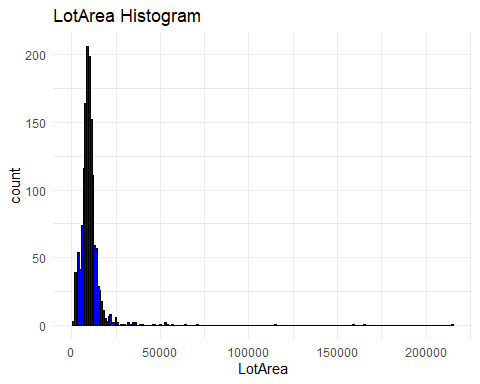
hist\_MoSold <- ggplot(numerical\_vars, aes(x = MoSold)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "MoSold Histogram", x = "MoSold") +  
 theme\_minimal()  
print(hist\_MoSold)



hist\_YrSold <- ggplot(numerical\_vars, aes(x = YrSold)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 labs(title = "YrSold Histogram", x = "YrSold") +  
 theme\_minimal()  
print(hist\_YrSold)

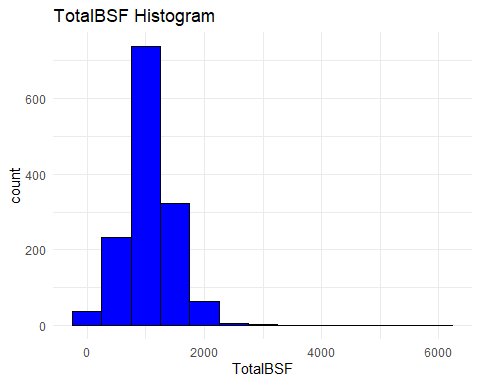


hist\_LotArea <- ggplot(numerical\_vars, aes(x = LotArea)) +  
 geom\_histogram(binwidth = 1000, fill = "blue", color = "black") +  
 labs(title = "LotArea Histogram", x = "LotArea") +  
 theme\_minimal()  
print(hist\_LotArea)

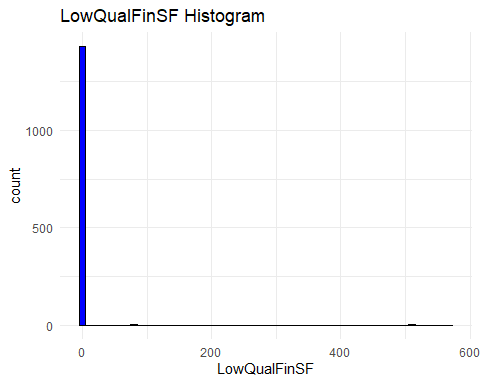


hist\_TotalBSF <- ggplot(numerical\_vars, aes(x = TotalBSF)) +  
 geom\_histogram(binwidth = 500, fill = "blue", color = "black") +  
 labs(title = "TotalBSF Histogram", x = "TotalBSF") +  
 theme\_minimal()  
print(hist\_TotalBSF)

## Warning: Removed 49 rows containing non-finite values (`stat\_bin()`).

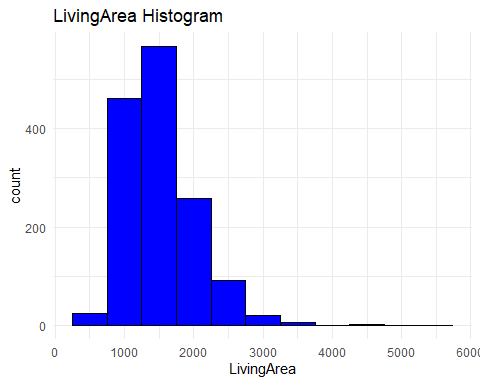


hist\_LowQualFinSF <- ggplot(numerical\_vars, aes(x = LowQualFinSF)) +  
 geom\_histogram(binwidth = 10, fill = "blue", color = "black") +  
 labs(title = "LowQualFinSF Histogram", x = "LowQualFinSF") +  
 theme\_minimal()  
print(hist\_LowQualFinSF)

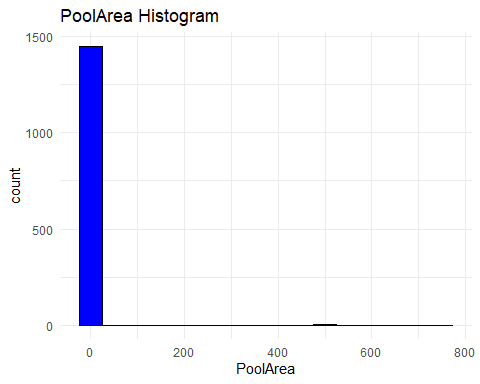


hist\_LivingArea <- ggplot(numerical\_vars, aes(x = LivingArea)) +  
 geom\_histogram(binwidth = 500, fill = "blue", color = "black") +  
 labs(title = "LivingArea Histogram", x = "LivingArea") +  
 theme\_minimal()  
print(hist\_LivingArea)

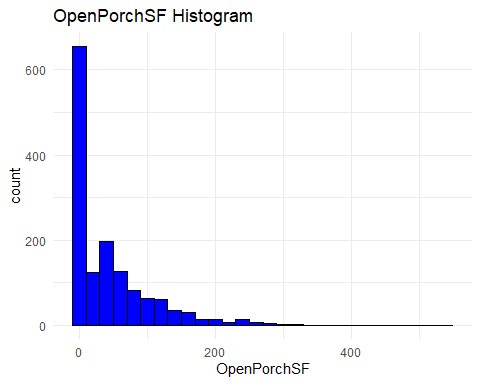
## Warning: Removed 22 rows containing non-finite values (`stat\_bin()`).



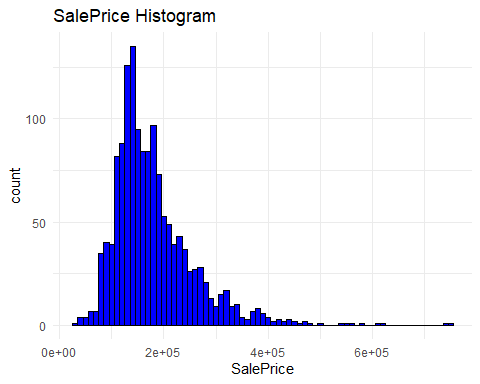
hist\_PoolArea <- ggplot(numerical\_vars, aes(x = PoolArea)) +  
 geom\_histogram(binwidth = 50, fill = "blue", color = "black") +  
 labs(title = "PoolArea Histogram", x = "PoolArea") +  
 theme\_minimal()  
print(hist\_PoolArea)



hist\_OpenPorchSF <- ggplot(numerical\_vars, aes(x = OpenPorchSF)) +  
 geom\_histogram(binwidth = 20, fill = "blue", color = "black") +  
 labs(title = "OpenPorchSF Histogram", x = "OpenPorchSF") +  
 theme\_minimal()  
print(hist\_OpenPorchSF)

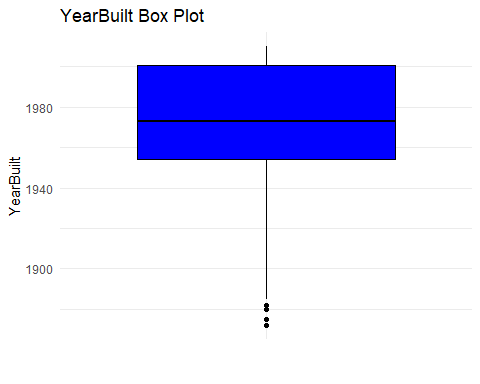


hist\_SalePrice <- ggplot(numerical\_vars, aes(x = SalePrice)) +  
 geom\_histogram(binwidth = 10000, fill = "blue", color = "black") +  
 labs(title = "SalePrice Histogram", x = "SalePrice") +  
 theme\_minimal()  
print(hist\_SalePrice)

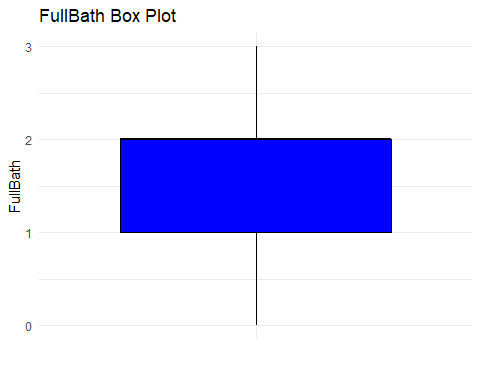


## Creating box plots for each numerical variable

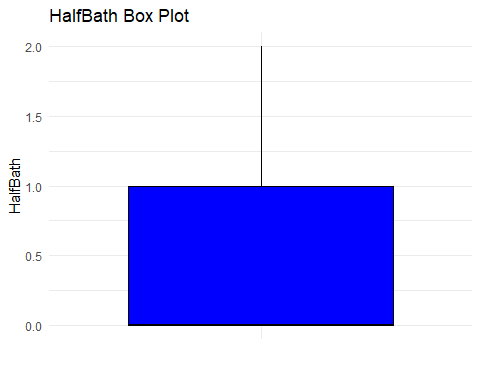
boxplot\_YearBuilt <- ggplot(numerical\_vars, aes(x = "", y = YearBuilt)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "YearBuilt Box Plot", x = "", y = "YearBuilt") +  
 theme\_minimal()  
print(boxplot\_YearBuilt)



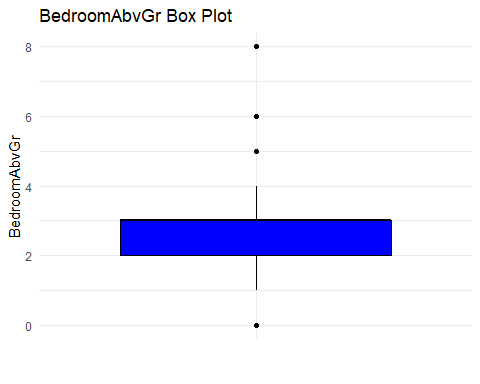
boxplot\_FullBath <- ggplot(numerical\_vars, aes(x = "", y = FullBath)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "FullBath Box Plot", x = "", y = "FullBath") +  
 theme\_minimal()  
print(boxplot\_FullBath)



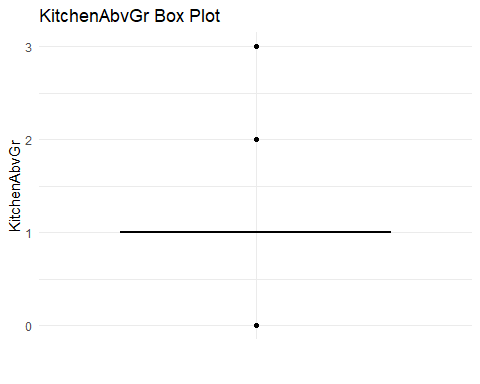
boxplot\_HalfBath <- ggplot(numerical\_vars, aes(x = "", y = HalfBath)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "HalfBath Box Plot", x = "", y = "HalfBath") +  
 theme\_minimal()  
print(boxplot\_HalfBath)



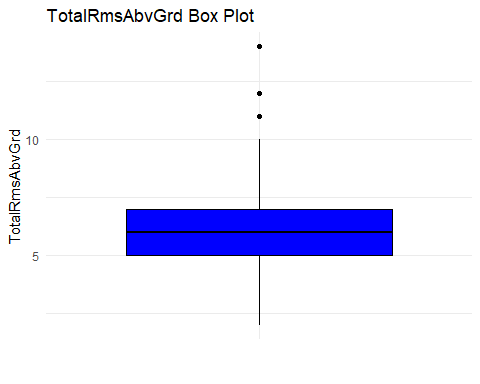
boxplot\_BedroomAbvGr <- ggplot(numerical\_vars, aes(x = "", y = BedroomAbvGr)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "BedroomAbvGr Box Plot", x = "", y = "BedroomAbvGr") +  
 theme\_minimal()  
print(boxplot\_BedroomAbvGr)



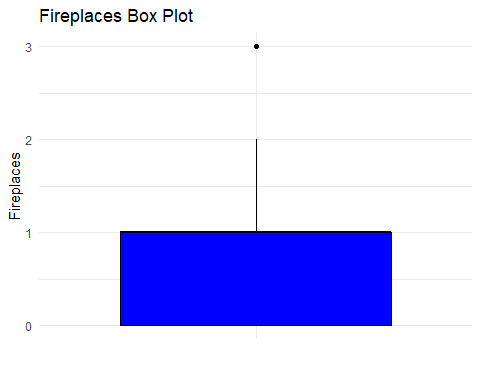
boxplot\_KitchenAbvGr <- ggplot(numerical\_vars, aes(x = "", y = KitchenAbvGr)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "KitchenAbvGr Box Plot", x = "", y = "KitchenAbvGr") +  
 theme\_minimal()  
print(boxplot\_KitchenAbvGr)



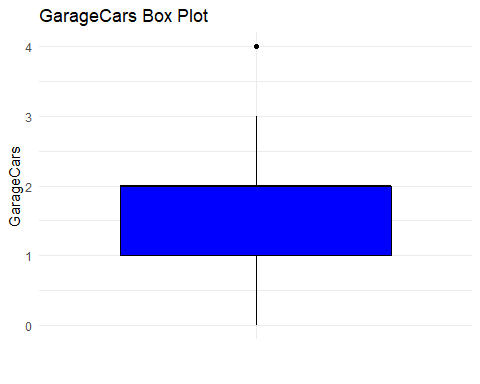
boxplot\_TotalRmsAbvGrd <- ggplot(numerical\_vars, aes(x = "", y = TotalRmsAbvGrd)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "TotalRmsAbvGrd Box Plot", x = "", y = "TotalRmsAbvGrd") +  
 theme\_minimal()  
print(boxplot\_TotalRmsAbvGrd)



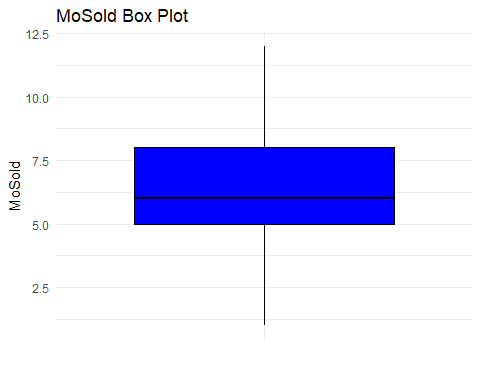
boxplot\_Fireplaces <- ggplot(numerical\_vars, aes(x = "", y = Fireplaces)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "Fireplaces Box Plot", x = "", y = "Fireplaces") +  
 theme\_minimal()  
print(boxplot\_Fireplaces)



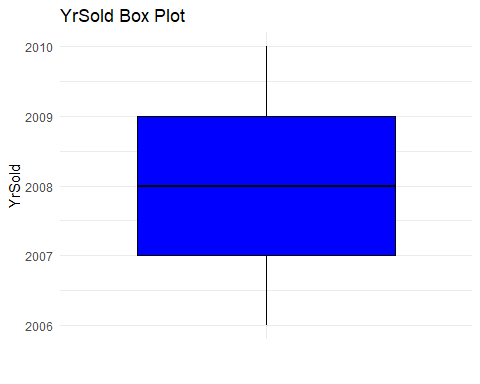
boxplot\_GarageCars <- ggplot(numerical\_vars, aes(x = "", y = GarageCars)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "GarageCars Box Plot", x = "", y = "GarageCars") +  
 theme\_minimal()  
print(boxplot\_GarageCars)



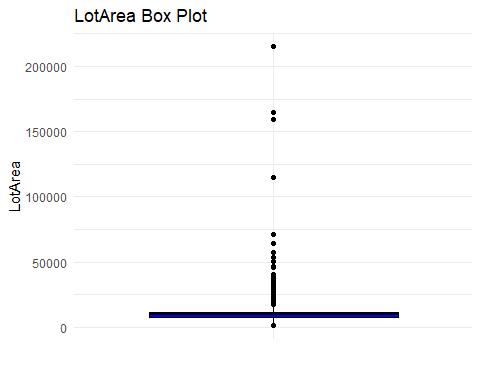
boxplot\_MoSold <- ggplot(numerical\_vars, aes(x = "", y = MoSold)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "MoSold Box Plot", x = "", y = "MoSold") +  
 theme\_minimal()  
print(boxplot\_MoSold)



boxplot\_YrSold <- ggplot(numerical\_vars, aes(x = "", y = YrSold)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "YrSold Box Plot", x = "", y = "YrSold") +  
 theme\_minimal()  
print(boxplot\_YrSold)

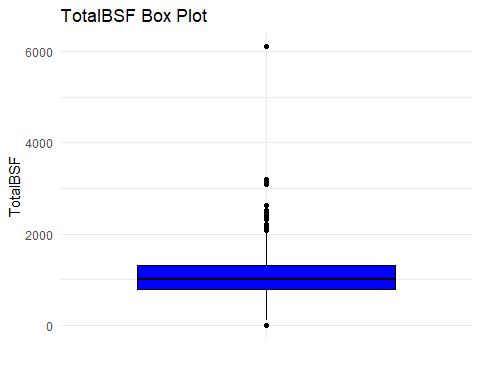


boxplot\_LotArea <- ggplot(numerical\_vars, aes(x = "", y = LotArea)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "LotArea Box Plot", x = "", y = "LotArea") +  
 theme\_minimal()  
print(boxplot\_LotArea)

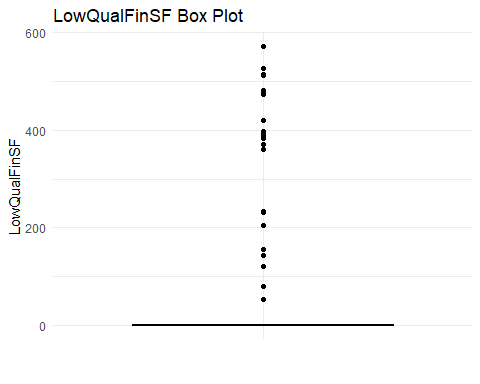


boxplot\_TotalBSF <- ggplot(numerical\_vars, aes(x = "", y = TotalBSF)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "TotalBSF Box Plot", x = "", y = "TotalBSF") +  
 theme\_minimal()  
print(boxplot\_TotalBSF)

## Warning: Removed 49 rows containing non-finite values (`stat\_boxplot()`).

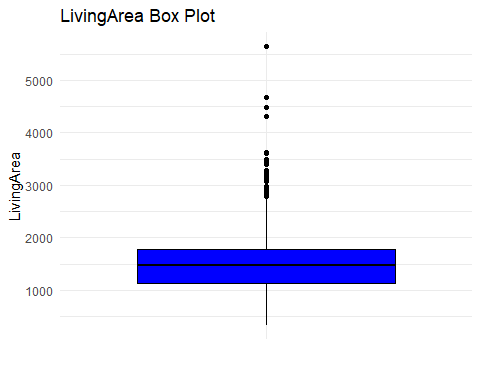


boxplot\_LowQualFinSF <- ggplot(numerical\_vars, aes(x = "", y = LowQualFinSF)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "LowQualFinSF Box Plot", x = "", y = "LowQualFinSF") +  
 theme\_minimal()  
print(boxplot\_LowQualFinSF)

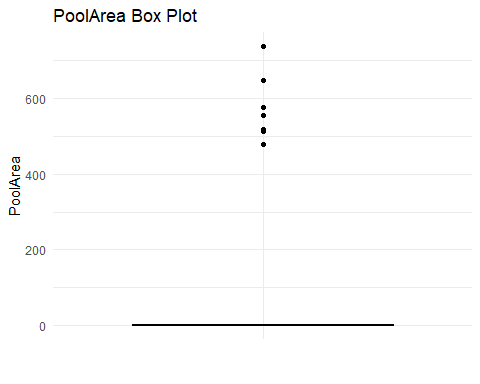


boxplot\_LivingArea <- ggplot(numerical\_vars, aes(x = "", y = LivingArea)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "LivingArea Box Plot", x = "", y = "LivingArea") +  
 theme\_minimal()  
print(boxplot\_LivingArea)

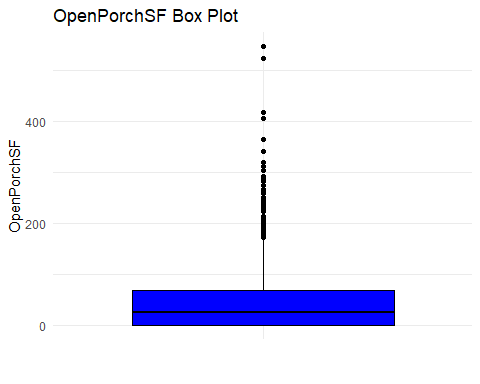
## Warning: Removed 22 rows containing non-finite values (`stat\_boxplot()`).



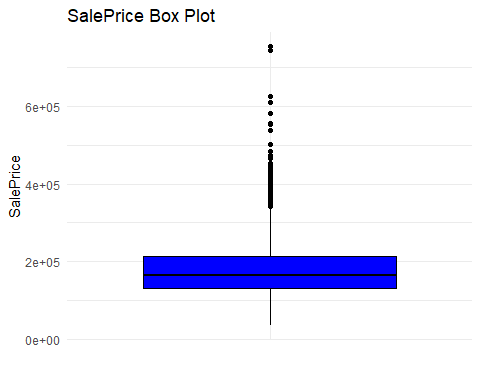
boxplot\_PoolArea <- ggplot(numerical\_vars, aes(x = "", y = PoolArea)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "PoolArea Box Plot", x = "", y = "PoolArea") +  
 theme\_minimal()  
print(boxplot\_PoolArea)



boxplot\_OpenPorchSF <- ggplot(numerical\_vars, aes(x = "", y = OpenPorchSF)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "OpenPorchSF Box Plot", x = "", y = "OpenPorchSF") +  
 theme\_minimal()  
print(boxplot\_OpenPorchSF)



boxplot\_SalePrice <- ggplot(numerical\_vars, aes(x = "", y = SalePrice)) +  
 geom\_boxplot(fill = "blue", color = "black") +  
 labs(title = "SalePrice Box Plot", x = "", y = "SalePrice") +  
 theme\_minimal()  
print(boxplot\_SalePrice)



# Identifying missing values

cols <- colnames(df)  
count\_missing\_values <- sapply(df, function(x) sum(is.na(x)))  
missing\_values <- data.frame(Variable = cols,   
 NumberOfMissingValues = count\_missing\_values)  
print(missing\_values)

## Variable NumberOfMissingValues  
## Id Id 0  
## LotArea LotArea 0  
## LotShape LotShape 0  
## LandContour LandContour 0  
## Utilities Utilities 0  
## LotConfig LotConfig 0  
## Slope Slope 0  
## DwellClass DwellClass 0  
## OverallQuality OverallQuality 0  
## OverallCondition OverallCondition 0  
## YearBuilt YearBuilt 0  
## ExteriorCondition ExteriorCondition 0  
## BasementCondition BasementCondition 0  
## TotalBSF TotalBSF 49  
## CentralAir CentralAir 0  
## LowQualFinSF LowQualFinSF 0  
## LivingArea LivingArea 22  
## FullBath FullBath 0  
## HalfBath HalfBath 0  
## BedroomAbvGr BedroomAbvGr 0  
## KitchenQuality KitchenQuality 0  
## KitchenAbvGr KitchenAbvGr 0  
## TotalRmsAbvGrd TotalRmsAbvGrd 0  
## Fireplaces Fireplaces 0  
## GarageType GarageType 78  
## GarageCars GarageCars 0  
## PavedDrive PavedDrive 0  
## PoolArea PoolArea 0  
## OpenPorchSF OpenPorchSF 0  
## MoSold MoSold 0  
## YrSold YrSold 0  
## SalePrice SalePrice 0  
## X X 1454  
## X.1 X.1 1454

# Handling missing values with Method 1: Imputation

df\_method1 <- df  
# Impute missing values in TotalBSF with the mean of the non-missing values  
mean\_TotalBSF <- mean(df\_method1$TotalBSF, na.rm = TRUE)  
print(mean\_TotalBSF)

## [1] 1064.512

df\_method1$TotalBSF[is.na(df\_method1$TotalBSF)] <- mean\_TotalBSF  
# Impute missing values in TotalBSF with the mean of the non-missing values  
mean\_LivingArea <- mean(df\_method1$LivingArea, na.rm = TRUE)  
print(mean\_LivingArea)

## [1] 1517.222

df\_method1$LivingArea[is.na(df\_method1$LivingArea)] <- mean\_LivingArea  
# Impute missing values in GarageType with the mode of the non-missing values  
garage\_table <- c("2Types" = 6, "Attchd" = 870, "Basment" = 19, "BuiltIn" = 88,   
 "CarPort" = 9, "Detchd" = 384)  
garage\_mode <- names(which.max(garage\_table))  
print(garage\_mode)

## [1] "Attchd"

df\_method1$GarageType[is.na(df\_method1$GarageType)] <- garage\_mode  
# Identifying missing values after method 1  
cols1 <- colnames(df\_method1)  
count\_missing\_values\_method1 <- sapply(df\_method1, function(x) sum(is.na(x)))  
missing\_values\_method1 <- data.frame(Variable = cols1,   
 NumberOfMissingValues = count\_missing\_values\_method1)  
print(missing\_values\_method1)

## Variable NumberOfMissingValues  
## Id Id 0  
## LotArea LotArea 0  
## LotShape LotShape 0  
## LandContour LandContour 0  
## Utilities Utilities 0  
## LotConfig LotConfig 0  
## Slope Slope 0  
## DwellClass DwellClass 0  
## OverallQuality OverallQuality 0  
## OverallCondition OverallCondition 0  
## YearBuilt YearBuilt 0  
## ExteriorCondition ExteriorCondition 0  
## BasementCondition BasementCondition 0  
## TotalBSF TotalBSF 0  
## CentralAir CentralAir 0  
## LowQualFinSF LowQualFinSF 0  
## LivingArea LivingArea 0  
## FullBath FullBath 0  
## HalfBath HalfBath 0  
## BedroomAbvGr BedroomAbvGr 0  
## KitchenQuality KitchenQuality 0  
## KitchenAbvGr KitchenAbvGr 0  
## TotalRmsAbvGrd TotalRmsAbvGrd 0  
## Fireplaces Fireplaces 0  
## GarageType GarageType 0  
## GarageCars GarageCars 0  
## PavedDrive PavedDrive 0  
## PoolArea PoolArea 0  
## OpenPorchSF OpenPorchSF 0  
## MoSold MoSold 0  
## YrSold YrSold 0  
## SalePrice SalePrice 0  
## X X 1454  
## X.1 X.1 1454

# Handling missing values with Method 2: Deleting missing values

df\_method2 <- df  
rows\_with\_na <- rowSums(is.na(df\_method2[c("TotalBSF", "LivingArea", "GarageType")])) > 0  
df\_method2 <- df\_method2[!rows\_with\_na, ]  
  
# Identifying missing values after method 2  
cols2 <- colnames(df\_method2)  
count\_missing\_values\_method2 <- sapply(df\_method2, function(x) sum(is.na(x)))  
missing\_values\_method2 <- data.frame(Variable = cols2,   
 NumberOfMissingValues = count\_missing\_values\_method2)  
print(missing\_values\_method2)

## Variable NumberOfMissingValues  
## Id Id 0  
## LotArea LotArea 0  
## LotShape LotShape 0  
## LandContour LandContour 0  
## Utilities Utilities 0  
## LotConfig LotConfig 0  
## Slope Slope 0  
## DwellClass DwellClass 0  
## OverallQuality OverallQuality 0  
## OverallCondition OverallCondition 0  
## YearBuilt YearBuilt 0  
## ExteriorCondition ExteriorCondition 0  
## BasementCondition BasementCondition 0  
## TotalBSF TotalBSF 0  
## CentralAir CentralAir 0  
## LowQualFinSF LowQualFinSF 0  
## LivingArea LivingArea 0  
## FullBath FullBath 0  
## HalfBath HalfBath 0  
## BedroomAbvGr BedroomAbvGr 0  
## KitchenQuality KitchenQuality 0  
## KitchenAbvGr KitchenAbvGr 0  
## TotalRmsAbvGrd TotalRmsAbvGrd 0  
## Fireplaces Fireplaces 0  
## GarageType GarageType 0  
## GarageCars GarageCars 0  
## PavedDrive PavedDrive 0  
## PoolArea PoolArea 0  
## OpenPorchSF OpenPorchSF 0  
## MoSold MoSold 0  
## YrSold YrSold 0  
## SalePrice SalePrice 0  
## X X 1316  
## X.1 X.1 1316

# Handling missing values with Method 3: Replacing the missing values with a specific value

df\_method3 <- df  
# Replacing missing values in TotalBSF with "0"  
df\_method3$TotalBSF[is.na(df\_method3$TotalBSF)] <- 0  
# Replacing missing values in LivingArea with "0"  
df\_method3$LivingArea[is.na(df\_method3$LivingArea)] <- 0  
# Replacing missing values in GarageType with "Unknown"  
df\_method3$GarageType[is.na(df\_method3$GarageType)] <- "Unknown"  
# Identifying missing values after method 3  
cols3 <- colnames(df\_method3)  
count\_missing\_values\_method3 <- sapply(df\_method3, function(x) sum(is.na(x)))  
missing\_values\_method3 <- data.frame(Variable = cols3,   
 NumberOfMissingValues = count\_missing\_values\_method3)  
print(missing\_values\_method3)

## Variable NumberOfMissingValues  
## Id Id 0  
## LotArea LotArea 0  
## LotShape LotShape 0  
## LandContour LandContour 0  
## Utilities Utilities 0  
## LotConfig LotConfig 0  
## Slope Slope 0  
## DwellClass DwellClass 0  
## OverallQuality OverallQuality 0  
## OverallCondition OverallCondition 0  
## YearBuilt YearBuilt 0  
## ExteriorCondition ExteriorCondition 0  
## BasementCondition BasementCondition 0  
## TotalBSF TotalBSF 0  
## CentralAir CentralAir 0  
## LowQualFinSF LowQualFinSF 0  
## LivingArea LivingArea 0  
## FullBath FullBath 0  
## HalfBath HalfBath 0  
## BedroomAbvGr BedroomAbvGr 0  
## KitchenQuality KitchenQuality 0  
## KitchenAbvGr KitchenAbvGr 0  
## TotalRmsAbvGrd TotalRmsAbvGrd 0  
## Fireplaces Fireplaces 0  
## GarageType GarageType 0  
## GarageCars GarageCars 0  
## PavedDrive PavedDrive 0  
## PoolArea PoolArea 0  
## OpenPorchSF OpenPorchSF 0  
## MoSold MoSold 0  
## YrSold YrSold 0  
## SalePrice SalePrice 0  
## X X 1454  
## X.1 X.1 1454

# Summary statistics

summary(df$TotalBSF)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0 799 999 1065 1304 6110 49

summary(df$LivingArea)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 334 1131 1467 1517 1780 5642 22

summary(df\_method1$TotalBSF)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 806.2 1017.0 1064.5 1285.5 6110.0

summary(df\_method1$LivingArea)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1136 1474 1517 1775 5642

summary(df\_method2$TotalBSF)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 815.8 1020.5 1081.8 1324.0 6110.0

summary(df\_method2$LivingArea)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 438 1155 1479 1535 1792 5642

summary(df\_method3$TotalBSF)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 780.0 982.5 1028.6 1285.5 6110.0

summary(df\_method3$LivingArea)

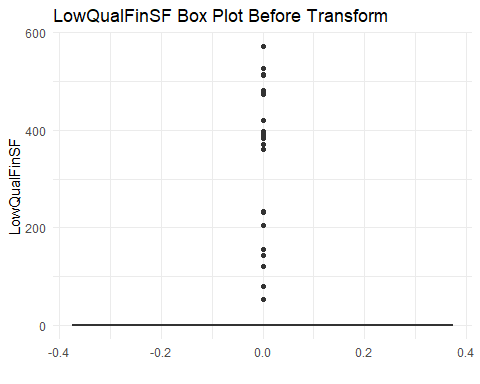
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 1120 1456 1494 1775 5642

# Transforming right skewed variables using log method

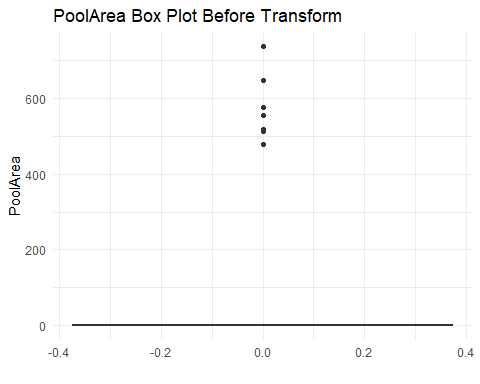
right\_skewed\_vars <- select(df, "LowQualFinSF", "PoolArea")  
right\_skewed\_vars$LowQualFinSF <- log10(right\_skewed\_vars$LowQualFinSF + 1)   
right\_skewed\_vars$PoolArea <- log10(right\_skewed\_vars$PoolArea + 1)

# Boxplots before transform

ggplot(df) +   
 geom\_boxplot(aes(y = LowQualFinSF))+  
 labs(title = "LowQualFinSF Box Plot Before Transform", x = "", y = "LowQualFinSF") +  
 theme\_minimal()

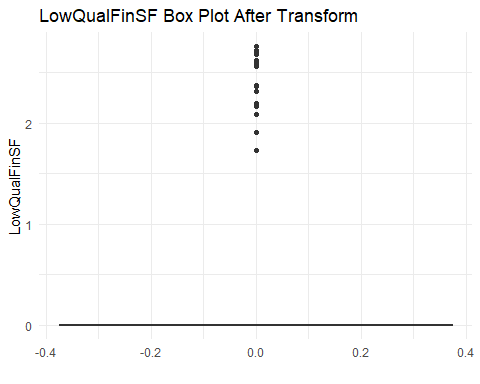


ggplot(df) +  
 geom\_boxplot(aes(y = PoolArea))+  
 labs(title = "PoolArea Box Plot Before Transform", x = "", y = "PoolArea") +  
 theme\_minimal()

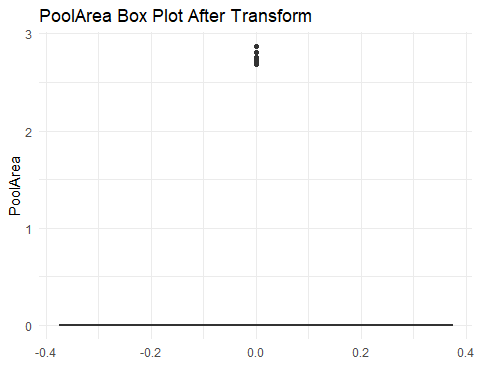


# Boxplots after transform

ggplot(right\_skewed\_vars) +  
 geom\_boxplot(aes(y = LowQualFinSF))+  
 labs(title = "LowQualFinSF Box Plot After Transform", x = "", y = "LowQualFinSF") +  
 theme\_minimal()



ggplot(right\_skewed\_vars) +  
 geom\_boxplot(aes(y = PoolArea)) +  
 labs(title = "PoolArea Box Plot After Transform", x = "", y = "PoolArea") +  
 theme\_minimal()

 # Calculating before and after skewness

skewness(df$LowQualFinSF)

## [1] 8.973624

skewness(right\_skewed\_vars$LowQualFinSF)

## [1] 7.428831

skewness(df$PoolArea)

## [1] 14.767

skewness(right\_skewed\_vars$PoolArea)

## [1] 14.30363

# Transforming categorical variables

not\_missing\_categorical\_vars <- select(df\_method1, "LotShape", "LandContour", "Utilities", "LotConfig", "Slope", "DwellClass",  
 "CentralAir", "GarageType", "PavedDrive", "OverallQuality", "OverallCondition",  
 "ExteriorCondition", "BasementCondition", "KitchenQuality")  
dummies <- not\_missing\_categorical\_vars %>%  
 mutate(  
 LotShape\_2 = if\_else(LotShape == 2, 1, 0),  
 LotShape\_3 = if\_else(LotShape == 3, 1, 0),  
 LotShape\_4 = if\_else(LotShape == 4, 1, 0),  
 LandContour\_HLS = if\_else(LandContour == "HLS", 1, 0),  
 LandContour\_Low = if\_else(LandContour == "Low", 1, 0),  
 LandContour\_Lv1 = if\_else(LandContour == "Lv1", 1, 0),  
 Utilities\_NoSewa = if\_else(Utilities == "NoSewa", 1, 0),  
 LotConfig\_CulDSac = if\_else(LotConfig == "CulDSac", 1, 0),  
 LotConfig\_FR2 = if\_else(LotConfig == "FR2", 1, 0),  
 LotConfig\_FR3 = if\_else(LotConfig == "FR3", 1, 0),  
 LotConfig\_Inside = if\_else(LotConfig == "Inside", 1, 0),  
 Slope\_Mod = if\_else(Slope == "Mod", 1, 0),  
 Slope\_Sev = if\_else(Slope == "Sev", 1, 0),  
 DwellClass\_2fmCon = if\_else(DwellClass == "2fmCon", 1, 0),  
 DwellClass\_Duplex = if\_else(DwellClass == "Duplex", 1, 0),  
 DwellClass\_Twnhs = if\_else(DwellClass == "Twnhs", 1, 0),  
 DwellClass\_TwnhsE = if\_else(DwellClass == "TwnhsE", 1, 0),  
 CentralAir\_Y = if\_else(CentralAir == "Y", 1, 0),  
 GarageType\_Attchd = if\_else(GarageType == "Attchd", 1, 0),  
 GarageType\_Basment = if\_else(GarageType == "Basment", 1, 0),  
 GarageType\_BuiltIn = if\_else(GarageType == "BuiltIn", 1, 0),  
 GarageType\_CarPort = if\_else(GarageType == "CarPort", 1, 0),  
 GarageType\_Detchd = if\_else(GarageType == "Detchd", 1, 0),  
 PavedDrive\_P = if\_else(PavedDrive == "P", 1, 0),  
 PavedDrive\_Y = if\_else(PavedDrive == "Y", 1, 0),  
 OverallQuality\_2 = if\_else(OverallQuality == 2, 1, 0),  
 OverallQuality\_3 = if\_else(OverallQuality == 3, 1, 0),  
 OverallQuality\_4 = if\_else(OverallQuality == 4, 1, 0),  
 OverallQuality\_5 = if\_else(OverallQuality == 5, 1, 0),  
 OverallQuality\_6 = if\_else(OverallQuality == 6, 1, 0),  
 OverallQuality\_7 = if\_else(OverallQuality == 7, 1, 0),  
 OverallQuality\_8 = if\_else(OverallQuality == 8, 1, 0),  
 OverallQuality\_9 = if\_else(OverallQuality == 9, 1, 0),  
 OverallQuality\_10 = if\_else(OverallQuality == 10, 1, 0),  
 OverallCondition\_3 = if\_else(OverallCondition == 3, 1, 0),  
 OverallCondition\_4 = if\_else(OverallCondition == 4, 1, 0),  
 OverallCondition\_5 = if\_else(OverallCondition == 5, 1, 0),  
 OverallCondition\_6 = if\_else(OverallCondition == 6, 1, 0),  
 OverallCondition\_7 = if\_else(OverallCondition == 7, 1, 0),  
 OverallCondition\_8 = if\_else(OverallCondition == 8, 1, 0),  
 OverallCondition\_9 = if\_else(OverallCondition == 9, 1, 0),  
 ExteriorCondition\_Gd = if\_else(ExteriorCondition == "Gd", 1, 0),  
 ExteriorCondition\_TA = if\_else(ExteriorCondition == "TA", 1, 0),  
 BasementCondition\_Fa = if\_else(BasementCondition == "Fa", 1, 0),  
 BasementCondition\_Gd = if\_else(BasementCondition == "Gd", 1, 0),  
 BasementCondition\_NB = if\_else(BasementCondition == "NB", 1, 0),  
 KitchenQuality\_Ex = if\_else(KitchenQuality == "Ex", 1, 0),  
 KitchenQuality\_Fa = if\_else(KitchenQuality == "Fa", 1, 0),  
 KitchenQuality\_Gd = if\_else(KitchenQuality == "Gd", 1, 0)  
 )

not\_missing\_numerical\_vars <- select(df\_method1, "YearBuilt", "FullBath", "HalfBath", "BedroomAbvGr", "KitchenAbvGr", "TotalRmsAbvGrd",   
 "Fireplaces", "GarageCars", "MoSold", "YrSold", "LotArea", "TotalBSF", "LowQualFinSF",   
 "LivingArea", "PoolArea", "OpenPorchSF", "SalePrice")

# Correlation analysis

correlation\_matrix <- round(cor(not\_missing\_numerical\_vars), 4)  
print(correlation\_matrix)

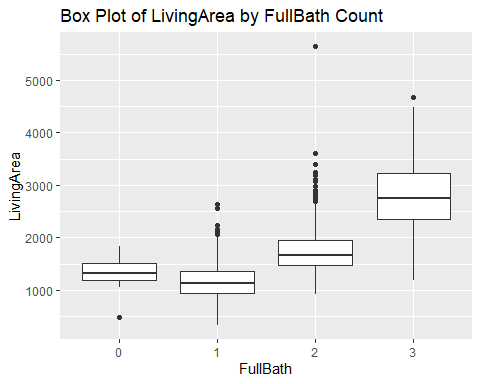
## YearBuilt FullBath HalfBath BedroomAbvGr KitchenAbvGr  
## YearBuilt 1.0000 0.4736 0.2454 -0.0751 -0.1771  
## FullBath 0.4736 1.0000 0.1365 0.3603 0.1336  
## HalfBath 0.2454 0.1365 1.0000 0.2285 -0.0685  
## BedroomAbvGr -0.0751 0.3603 0.2285 1.0000 0.1986  
## KitchenAbvGr -0.1771 0.1336 -0.0685 0.1986 1.0000  
## TotalRmsAbvGrd 0.0962 0.5523 0.3440 0.6761 0.2564  
## Fireplaces 0.1470 0.2403 0.2034 0.1051 -0.1245  
## GarageCars 0.5382 0.4662 0.2198 0.0808 -0.0521  
## MoSold 0.0145 0.0536 -0.0070 0.0459 0.0269  
## YrSold -0.0114 -0.0170 -0.0101 -0.0348 0.0322  
## LotArea 0.0141 0.1257 0.0136 0.1198 -0.0179  
## TotalBSF 0.3898 0.3200 -0.0351 0.0540 -0.0646  
## LowQualFinSF -0.1856 -0.0009 -0.0272 0.1056 0.0074  
## LivingArea 0.2007 0.6215 0.4113 0.5129 0.0992  
## PoolArea 0.0045 0.0498 0.0224 0.0707 -0.0146  
## OpenPorchSF 0.2044 0.2538 0.1988 0.0953 -0.0705  
## SalePrice 0.5258 0.5576 0.2834 0.1652 -0.1369  
## TotalRmsAbvGrd Fireplaces GarageCars MoSold YrSold LotArea  
## YearBuilt 0.0962 0.1470 0.5382 0.0145 -0.0114 0.0141  
## FullBath 0.5523 0.2403 0.4662 0.0536 -0.0170 0.1257  
## HalfBath 0.3440 0.2034 0.2198 -0.0070 -0.0101 0.0136  
## BedroomAbvGr 0.6761 0.1051 0.0808 0.0459 -0.0348 0.1198  
## KitchenAbvGr 0.2564 -0.1245 -0.0521 0.0269 0.0322 -0.0179  
## TotalRmsAbvGrd 1.0000 0.3242 0.3593 0.0354 -0.0327 0.1899  
## Fireplaces 0.3242 1.0000 0.2975 0.0458 -0.0224 0.2712  
## GarageCars 0.3593 0.2975 1.0000 0.0388 -0.0343 0.1546  
## MoSold 0.0354 0.0458 0.0388 1.0000 -0.1443 0.0012  
## YrSold -0.0327 -0.0224 -0.0343 -0.1443 1.0000 -0.0137  
## LotArea 0.1899 0.2712 0.1546 0.0012 -0.0137 1.0000  
## TotalBSF 0.2898 0.3357 0.4276 0.0018 -0.0155 0.2586  
## LowQualFinSF 0.1314 -0.0215 -0.0957 -0.0221 -0.0287 0.0047  
## LivingArea 0.8184 0.4568 0.4635 0.0466 -0.0348 0.2637  
## PoolArea 0.0839 0.0951 0.0207 -0.0337 -0.0596 0.0777  
## OpenPorchSF 0.2335 0.1705 0.2187 0.0725 -0.0599 0.0848  
## SalePrice 0.5320 0.4653 0.6394 0.0465 -0.0269 0.2637  
## TotalBSF LowQualFinSF LivingArea PoolArea OpenPorchSF SalePrice  
## YearBuilt 0.3898 -0.1856 0.2007 0.0045 0.2044 0.5258  
## FullBath 0.3200 -0.0009 0.6215 0.0498 0.2538 0.5576  
## HalfBath -0.0351 -0.0272 0.4113 0.0224 0.1988 0.2834  
## BedroomAbvGr 0.0540 0.1056 0.5129 0.0707 0.0953 0.1652  
## KitchenAbvGr -0.0646 0.0074 0.0992 -0.0146 -0.0705 -0.1369  
## TotalRmsAbvGrd 0.2898 0.1314 0.8184 0.0839 0.2335 0.5320  
## Fireplaces 0.3357 -0.0215 0.4568 0.0951 0.1705 0.4653  
## GarageCars 0.4276 -0.0957 0.4635 0.0207 0.2187 0.6394  
## MoSold 0.0018 -0.0221 0.0466 -0.0337 0.0725 0.0465  
## YrSold -0.0155 -0.0287 -0.0348 -0.0596 -0.0599 -0.0269  
## LotArea 0.2586 0.0047 0.2637 0.0777 0.0848 0.2637  
## TotalBSF 1.0000 -0.0298 0.4584 0.1269 0.2465 0.6084  
## LowQualFinSF -0.0298 1.0000 0.1355 0.0621 0.0191 -0.0260  
## LivingArea 0.4584 0.1355 1.0000 0.1716 0.3263 0.7061  
## PoolArea 0.1269 0.0621 0.1716 1.0000 0.0622 0.0926  
## OpenPorchSF 0.2465 0.0191 0.3263 0.0622 1.0000 0.3134  
## SalePrice 0.6084 -0.0260 0.7061 0.0926 0.3134 1.0000

# Filter correlations exceeding a threshold (e.g., 0.7)  
threshold <- 0.6  
high\_correlation <- which(correlation\_matrix > threshold, arr.ind = TRUE)  
print(high\_correlation)

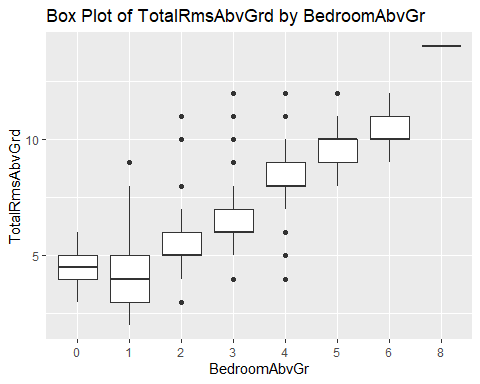
## row col  
## YearBuilt 1 1  
## FullBath 2 2  
## LivingArea 14 2  
## HalfBath 3 3  
## BedroomAbvGr 4 4  
## TotalRmsAbvGrd 6 4  
## KitchenAbvGr 5 5  
## BedroomAbvGr 4 6  
## TotalRmsAbvGrd 6 6  
## LivingArea 14 6  
## Fireplaces 7 7  
## GarageCars 8 8  
## SalePrice 17 8  
## MoSold 9 9  
## YrSold 10 10  
## LotArea 11 11  
## TotalBSF 12 12  
## SalePrice 17 12  
## LowQualFinSF 13 13  
## FullBath 2 14  
## TotalRmsAbvGrd 6 14  
## LivingArea 14 14  
## SalePrice 17 14  
## PoolArea 15 15  
## OpenPorchSF 16 16  
## GarageCars 8 17  
## TotalBSF 12 17  
## LivingArea 14 17  
## SalePrice 17 17

# Scatter plots for high correlation variables

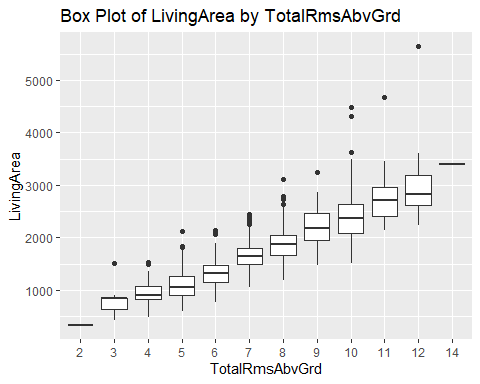
ggplot(data = not\_missing\_numerical\_vars, aes(x = factor(FullBath), y = LivingArea)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of LivingArea by FullBath Count",  
 x = "FullBath",  
 y = "LivingArea")



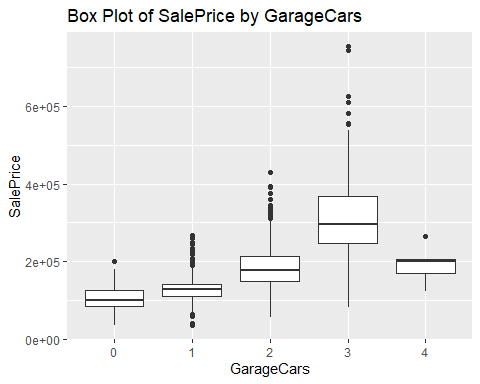
ggplot(data = not\_missing\_numerical\_vars, aes(x = factor(BedroomAbvGr), y = TotalRmsAbvGrd)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of TotalRmsAbvGrd by BedroomAbvGr",  
 x = "BedroomAbvGr",  
 y = "TotalRmsAbvGrd")



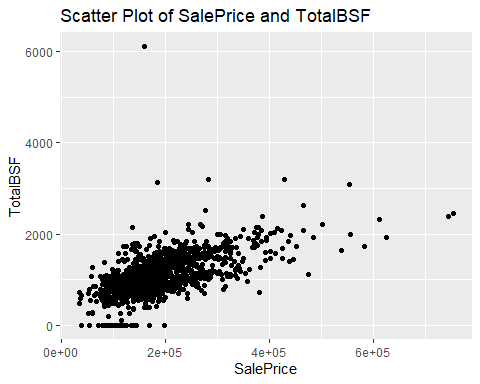
ggplot(data = not\_missing\_numerical\_vars, aes(x = factor(TotalRmsAbvGrd), y = LivingArea)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of LivingArea by TotalRmsAbvGrd",  
 x = "TotalRmsAbvGrd",  
 y = "LivingArea")



ggplot(data = not\_missing\_numerical\_vars, aes(x = factor(GarageCars), y = SalePrice)) +  
 geom\_boxplot() +  
 labs(title = "Box Plot of SalePrice by GarageCars",  
 x = "GarageCars",  
 y = "SalePrice")



ggplot(data = not\_missing\_numerical\_vars, aes(x = SalePrice, y = TotalBSF)) +  
 geom\_point() +  
 labs(title = "Scatter Plot of SalePrice and TotalBSF",  
 x = "SalePrice",  
 y = "TotalBSF")



ggplot(data = not\_missing\_numerical\_vars, aes(x = LivingArea, y = SalePrice)) +  
 geom\_point() +  
 labs(title = "Scatter Plot of LivingArea and SalePrice",  
 x = "LivingArea",  
 y = "SalePrice")

