Housing

Yuqing Yang

7/20/2020

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
list.files(path = "../input")
```

```
## character(0)
```

```
train = read.csv("~/Desktop/HTrainW19Final.csv")
test = read.csv("~/Desktop/HTestW19Final No Y values.csv")
head(train)
```

```
##   Ob MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         50       RL          82   12375   Pave  <NA>      Reg         Lvl
## 2  2         30       RM          60   10800   Pave  Grvl      Reg         Lvl
## 3  3         45       RM          58    6380   Pave  <NA>      Reg         Lvl
## 4  4         20       RL          60    7200   Pave  <NA>      Reg         Lvl
## 5  5         60       FV         100   13162   Pave  <NA>      Reg         Lvl
## 6  6         80       RL          61    9734   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl       Sawyer      Feedr       Norm     1Fam
## 2    AllPub    Inside       Gtl      OldTown       Norm       Norm     1Fam
## 3    AllPub    Inside       Gtl      BrkSide       Norm       Norm     1Fam
## 4    AllPub    Inside       Gtl        NAmes       Norm       Norm     1Fam
## 5    AllPub    Corner       Gtl      Somerst      Feedr       Norm     1Fam
## 6    AllPub    Inside       Gtl      Gilbert       RRAn       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     1.5Fin           5           5      1951         1951     Gable  CompShg
## 2     1Story           4           7      1885         1995   Mansard  CompShg
## 3     1.5Unf           5           6      1922         1950     Gable  CompShg
## 4     1Story           5           8      1950         2002     Gable  CompShg
## 5     2Story           9           5      2006         2006     Gable  CompShg
## 6       SLvl           7           5      2004         2004     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     HdBoard     HdBoard      Stone         41        TA        Fa     CBlock
## 2     VinylSd     VinylSd       None          0        TA        TA     BrkTil
## 3     MetalSd     MetalSd       None          0        TA        TA     BrkTil
## 4     VinylSd     VinylSd       None          0        TA        TA     CBlock
## 5     VinylSd     VinylSd       None          0        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        Gd        TA      PConc
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
```

```
## 1       TA       TA           No          BLQ        329          Unf
## 2       Fa       TA           No          Unf          0          Unf
## 3       TA       Fa           No          Unf          0          Unf
## 4       TA       TA           No          ALQ        398          BLQ
## 5       Ex       TA           No          GLQ       1836          Unf
## 6       Gd       TA           Mn          GLQ        241          Rec
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       477         806    GasA        TA          Y      SBrkr
## 2          0       641         641    GasA        Gd          Y      SBrkr
## 3          0       993         993    GasA        TA          Y      FuseA
## 4        149       317         864    GasA        Gd          Y      SBrkr
## 5          0       200        2036    GasA        Ex          Y      SBrkr
## 6        113        30         384    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1      1081       341            0      1422            1            0        1
## 2      1047         0            0      1047            0            0        1
## 3      1048         0            0      1048            0            0        1
## 4       864         0            0       864            1            0        1
## 5      2036       604            0      2640            1            0        3
## 6       744       630            0      1374            0            0        2
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        0            3            1          TA            7        Typ
## 2        0            2            1          TA            6        Typ
## 3        0            2            1          TA            5        Typ
## 4        0            3            1          Gd            5        Typ
## 5        1            3            1          Ex           11        Typ
## 6        1            3            1          Gd            7        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          1          TA     Detchd        1951          Unf          1
## 2          0        <NA>     Detchd        1954          Unf          1
## 3          1          Gd     Detchd        1922          Unf          1
## 4          0        <NA>     Detchd        1980          RFn          2
## 5          1          Gd     Attchd        2006          RFn          3
## 6          0        <NA>    BuiltIn        2004          Fin          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        288         TA         TA          Y          0           0
## 2        273         Fa         Fa          N          0           0
## 3        280         TA         TA          Y          0           0
## 4        720         TA         TA          Y        194           0
## 5        792         TA         TA          Y          0         265
## 6        400         TA         TA          Y          0           0
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  GdWo        <NA>
## 2             0          0           0        0   <NA>  <NA>        Shed
## 3           116          0           0        0   <NA>  <NA>        <NA>
## 4             0          0           0        0   <NA>  <NA>        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0          0           0        0   <NA>  <NA>        <NA>
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      6   2006       WD        Normal  131878.1
## 2     450      8   2007       WD        Normal  104069.3
## 3       0      8   2006       WD        Normal  116843.7
## 4       0      7   2007       WD        Normal  132932.2
## 5       0     11   2006      New       Partial  429077.4
```

```
## 6       0      5   2009       WD        Normal  174735.6
```

```
head(test)
```

```
##   Ob MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         20       RL          53    3710   Pave  <NA>      Reg         Lvl
## 2  2         60       RL          NA   10304   Pave  <NA>      IR1         Lvl
## 3  3         75       RM          90    8100   Pave  <NA>      Reg         Lvl
## 4  4         20       RL          80   14680   Pave  Grvl      IR1         HLS
## 5  5         90       RM         110    8472   Grvl  <NA>      IR2         Bnk
## 6  6         20       FV          72    8640   Pave  <NA>      Reg         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      Blmngtn       Norm       Norm     1Fam
## 2    AllPub   CulDSac       Gtl       NWAmes       PosN       Norm     1Fam
## 3    AllPub    Corner       Gtl      OldTown       Norm       Norm     1Fam
## 4    AllPub    Inside       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub    Corner       Mod       IDOTRR       RRNn       Norm   Duplex
## 6    AllPub    Inside       Gtl      Somerst       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     1Story           7           5      2007         2008     Gable  CompShg
## 2     2Story           5           7      1976         1976     Gable  CompShg
## 3     2.5Unf           5           5      1898         1965       Hip  CompShg
## 4     1Story           5           4      1960         1960     Gable  CompShg
## 5     1Story           5           5      1963         1963     Gable  CompShg
## 6     1Story           8           5      2007         2008     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     WdShing     Wd Shng    BrkFace         20        Gd        TA      PConc
## 2     Plywood     Plywood    BrkFace         44        TA        Gd     CBlock
## 3     AsbShng     AsbShng       None          0        TA        TA      PConc
## 4     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 5     Wd Sdng     Wd Sdng       None          0        Fa        TA     CBlock
## 6     VinylSd     VinylSd       None          0        Gd        TA      PConc
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           Gd          Unf          0          Unf
## 2       TA       TA           No          ALQ        381          Unf
## 3       TA       TA           No          Unf          0          Unf
## 4       TA       TA           No          Rec        793          Unf
## 5       Gd       TA           Gd          LwQ        104          GLQ
## 6       Gd       TA           No          GLQ         24          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0      1146        1146    GasA        Ex          Y      SBrkr
## 2          0       399         780    GasA        Ex          Y      SBrkr
## 3          0       849         849    GasA        TA          N      FuseA
## 4          0       480        1273    GasA        Ex          Y      SBrkr
## 5        712         0         816    GasA        TA          N      SBrkr
## 6          0      1339        1363    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1      1246         0            0      1246            0            0        2
## 2      1088       780            0      1868            1            0        2
## 3      1075      1063            0      2138            0            0        2
## 4      1273         0            0      1273            0            0        1
## 5       816         0            0       816            1            0        1
## 6      1372         0            0      1372            0            0        2
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        0            2            1          Gd            5        Typ
```

```
## 2        1            4            1          Gd            9        Typ
## 3        0            2            3          TA           11        Typ
## 4        0            2            1          TA            5        Typ
## 5        0            2            1          TA            5        Typ
## 6        0            3            1          Gd            6        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          1          Gd     Attchd        2007          Fin          2
## 2          1          TA     Attchd        1976          Unf          2
## 3          0        <NA>     Detchd        1910          Unf          2
## 4          0        <NA>     Attchd        1960          Unf          1
## 5          0        <NA>    CarPort        1963          Unf          2
## 6          0        <NA>     Attchd        2008          RFn          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        428         TA         TA          Y        100          24
## 2        484         TA         TA          Y        448          96
## 3        360         Fa         Po          N         40         156
## 4        307         TA         TA          Y        483           0
## 5        516         TA         TA          Y        106           0
## 6        588         TA         TA          Y        192         113
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA> MnPrv        <NA>
## 4             0          0         115        0   <NA> MnPrv        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0          0           0        0   <NA>  <NA>        <NA>
##   MiscVal MoSold YrSold SaleType SaleCondition
## 1       0      3   2008      New       Partial
## 2       0     10   2009       WD        Normal
## 3       0     11   2009       WD        Normal
## 4       0      6   2009       WD        Normal
## 5       0      5   2010       WD        Normal
## 6       0      7   2008      New       Partial
```

```
dim(train)
```

```
## [1] 2500   81
```

```
dim(test)
```

```
## [1] 1500   80
```

```
colSums(is.na(train))
```

```
##            Ob    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             4           394             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0          2323             0             0             2
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##            19            16             0             0             0
```

```
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            71            70            70            68             1
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            69             1             1             1             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             3             3             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             1             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             1             0          1225           134           134
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##           134             0             0           134           134
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          2491          2018          2421
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             1             0
##     SalePrice
##             0
```

```
#remove missing values for train data
#I chose to delete the following columns: Alley, PoolQC, Fence, MiscFeature because there are too many
grep(c("Alley"), colnames(train))
```

```
## [1] 7
```

```
grep(c("PoolQC"), colnames(train))
```

```
## [1] 73
```

```
grep(c("Fence"), colnames(train))
```

```
## [1] 74
```

```
grep(c("MiscFeature"), colnames(train))
```

```
## [1] 75
```

```
train2 = train[, c(-7,-73,-74,-75)]
#LotFrontage is a numerical column with 394 NAs, I choose to replace all NAs by the mean
train2$LotFrontage[which(is.na(train2$LotFrontage))] <- mean(na.omit(train2$LotFrontage))
#MasVnrArea is a numerical column with 16 NAs, I choose to replace all NAs by the mean
train2$MasVnrArea[which(is.na(train2$MasVnrArea))] <- mean(na.omit(train2$MasVnrArea))
#GaraageYrBlt is a numerical column witnh 134 NAs, I choose to replace all NAs by the median
train2$GarageYrBlt[which(is.na(train2$GarageYrBlt))] <- median(na.omit(train2$GarageYrBlt))
#BsmtFinSF1 is a numerical column witnh 1 NAs, I choose to replace all NAs by the mean
train2$BsmtFinSF1[which(is.na(train2$BsmtFinSF1))] <- median(na.omit(train2$BsmtFinSF1))
#BsmtFinSF2 is a numerical column witnh 69 NAs, I choose to replace all NAs by the mean
train2$BsmtFinSF2[which(is.na(train2$BsmtFinSF2))] <- median(na.omit(train2$BsmtFinSF2))
#BsmtUnfSF is a numerical column witnh 1 NAs, I choose to replace all NAs by the mean
train2$BsmtUnfSF[which(is.na(train2$BsmtUnfSF))] <- median(na.omit(train2$BsmtUnfSF))
#TotalBsmtSF is a numerical column witnh 1 NAs, I choose to replace all NAs by the mean
train2$TotalBsmtSF[which(is.na(train2$TotalBsmtSF))] <- median(na.omit(train2$TotalBsmtSF))
#BsmtFullBath is a numerical column witnh 1 NAs, I choose to replace all NAs by the mean
```

```
train2$BsmtFullBath[which(is.na(train2$BsmtFullBath))] <- median(na.omit(train2$BsmtFullBath))
#BsmtHalfBath is a numerical column witnh 1 NAs, I choose to replace all NAs by the mean
train2$BsmtHalfBath[which(is.na(train2$BsmtHalfBath))] <- median(na.omit(train2$BsmtHalfBath))

#remove missing values for categorical variables
#MSZoing
train2$MSZoning[which(is.na(train2$MSZoning))] <- as.character(train2$MSZoning[which.max(table(train2$MS
#Utilities
train2$Utilities[which(is.na(train2$Utilities))] <- as.character(train2$Utilities[which.max(table(train1
#MasVnrTyoe
train2$MasVnrType[which(is.na(train2$MasVnrType))] <- as.character(train2$MasVnrType[which.max(table(tra
#BsmtQual
train2$BsmtQual[which(is.na(train2$BsmtQual))] <- as.character(train2$BsmtQual[which.max(table(train2$Bs
#BsmtCond
train2$BsmtCond[which(is.na(train2$BsmtCond))] <- as.character(train2$BsmtCond[which.max(table(train2$Bs
#BsmtExposure
train2$BsmtExposure[which(is.na(train2$BsmtExposure))] <- as.character(train2$BsmtExposure[which.max(tab
#BsmtFinType1
train2$BsmtFinType1[which(is.na(train2$BsmtFinType1))] <- as.character(train2$BsmtFinType1[which.max(tab
#BsmtFinType2
train2$BsmtFinType2[which(is.na(train2$BsmtFinType2))] <- as.character(train2$BsmtFinType2[which.max(tab
#kitchenQual
train2$KitchenQual[which(is.na(train2$KitchenQual))] <- as.character(train2$KitchenQual[which.max(table
#GarageCond
train2$GarageCond[which(is.na(train2$GarageCond))] <- as.character(train2$GarageCond[which.max(table(tra
#GaraQual
train2$GarageQual[which(is.na(train2$GarageQual))] <- as.character(train2$GarageQual[which.max(table(tra
#BsmtQual
train2$SaleType[which(is.na(train2$SaleType))] <- as.character(train2$SaleType[which.max(table(train2$Sa
#Functional
train2$Functional[which(is.na(train2$Functional))] <- as.character(train2$Functional[which.max(table(tra
#FireplaceQu
train2$FireplaceQu[which(is.na(train2$FireplaceQu))] <- as.character(train2$FireplaceQu[which.max(table
#GarageType
train2$GarageType[which(is.na(train2$GarageType))] <- as.character(train2$GarageType[which.max(table(tra
#GarageFinish
train2$GarageFinish[which(is.na(train2$GarageFinish))] <- as.character(train2$GarageFinish[which.max(tab
colSums(is.na(train2))
```

```
##            Ob    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             0             0             0
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##             0             0             0             0             0
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##             0             0             0             0             0
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
```

```
##             0             0             0             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             0             0             0
##    Fireplaces   FireplaceQu    GarageType   GarageYrBlt  GarageFinish
##             0             0             0             0             0
##    GarageCars    GarageArea    GarageQual    GarageCond    PavedDrive
##             0             0             0             0             0
##    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch
##             0             0             0             0             0
##      PoolArea       MiscVal        MoSold        YrSold      SaleType
##             0             0             0             0             0
## SaleCondition     SalePrice
##             0             0
```

```
sum(is.na(train2))
```

```
## [1] 0
```

```
#remove all the NAs for the test data
grep(c("Alley"), colnames(test))
```

```
## [1] 7
```

```
grep(c("PoolQC"), colnames(test))
```

```
## [1] 73
```

```
grep(c("Fence"), colnames(test))
```

```
## [1] 74
```

```
grep(c("MiscFeature"), colnames(test))
```

```
## [1] 75
```

```
grep(c("FireplaceQu"), colnames(test))
```

```
## [1] 58
```

```
test2 = test[, c(-7,-73,-74,-75,-58)]

#MSZoing
test2$MSZoning[which(is.na(test2$MSZoning))] <- as.character(test2$MSZoning[which.max(table(test2$MSZoni
#LotFrontage
test2$LotFrontage[which(is.na(test2$LotFrontage))] <- mean(na.omit(test2$LotFrontage))
#Exterior1st
test2$Exterior1st[which(is.na(test2$Exterior1st))] <- as.character(test2$Exterior1st[which.max(table(tes
#Exterior12nd
test2$Exterior2nd[which(is.na(test2$Exterior2nd))] <- as.character(test2$Exterior2nd[which.max(table(tes
#MasVnrType
test2$MasVnrType[which(is.na(test2$MasVnrArea))] <- as.character(test2$MasVnrType[which.max(table(test2$
#MasVnrArea
test2$MasVnrArea[which(is.na(test2$MasVnrArea))] <- mean(na.omit(test2$MasVnrArea))
#BsmtQual
test2$BsmtQual[which(is.na(test2$BsmtQual))] <- as.character(test2$BsmtQual[which.max(table(test2$BsmtQu
```

```
#BsmtCond
test2$BsmtCond[which(is.na(test2$BsmtCond))] <- as.character(test2$BsmtCond[which.max(table(test2$BsmtC
#BsmtExposure
test2$BsmtExposure[which(is.na(test2$BsmtExposure))] <- as.character(test2$BsmtExposure[which.max(table
#BsmtFinType1
test2$BsmtFinType1[which(is.na(test2$BsmtFinType1))] <- as.character(test2$BsmtFinType1[which.max(table
#BsmtFinType2
test2$BsmtFinType2[which(is.na(test2$BsmtFinType2))] <- as.character(test2$BsmtFinType2[which.max(table
#BsmtFinSF1
test2$BsmtFinSF1[which(is.na(test2$BsmtFinSF1))] <- median(na.omit(test2$BsmtFinSF1))
#BsmtFinSF2
test2$BsmtFinSF2[which(is.na(test2$BsmtFinSF2))] <- median(na.omit(test2$BsmtFinSF2))
#BsmtUnfSF
test2$BsmtUnfSF[which(is.na(test2$BsmtUnfSF))] <- median(na.omit(test2$BsmtUnfSF))
#TotalBsmtSF
test2$TotalBsmtSF[which(is.na(test2$TotalBsmtSF))] <- median(na.omit(test2$TotalBsmtSF))
#Electrical
test2$Electrical[which(is.na(test2$Electrical))] <- as.character(test2$Electrical[which.max(table(test2$
#Functional
test2$Functional[which(is.na(test2$Functional))] <- as.character(test2$Functional[which.max(table(test2$
#FireplaceQu
test2$FireplaceQu[which(is.na(test2$FireplaceQu))] <- as.character(test2$FireplaceQu[which.max(table(tes
#GarageType
test2$GarageType[which(is.na(test2$GarageType))] <- as.character(test2$GarageType[which.max(table(test2$
#GarageFinish
test2$GarageFinish[which(is.na(test2$GarageFinish))] <- as.character(test2$GarageFinish[which.max(table
#GarageYrBlt
test2$GarageYrBlt[which(is.na(test2$GarageYrBlt))] <- median(na.omit(test2$GarageYrBlt))
#GarageCars
test2$GarageCars[which(is.na(test2$GarageCars))] <- mean(na.omit(test2$GarageCars))
#kitchenQual
test2$KitchenQual[which(is.na(test2$KitchenQual))] <- as.character(test2$KitchenQual[which.max(table(te
#BsmtFullBath
test2$BsmtFullBath[which(is.na(test2$BsmtFullBath))] <- median(na.omit(test2$BsmtFullBath))
#BsmtHalfBath
test2$BsmtHalfBath[which(is.na(test2$BsmtHalfBath))] <- median(na.omit(test2$BsmtHalfBath))
#GarageCond
test2$GarageCond[which(is.na(test2$GarageCond))] <- as.character(test2$GarageCond[which.max(table(test2$
#GaraQual
test2$GarageQual[which(is.na(test2$GarageQual))] <- as.character(test2$GarageQual[which.max(table(test2$
#GarageArea
test2$GarageArea[which(is.na(test2$GarageArea))] <- mean(na.omit(test2$GarageArea))

colSums(is.na(test2))
```

```
##            Ob    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street     LotShape   LandContour     Utilities     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
```

```
##             0             0             0             0             0
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##             0             0             0             0             0
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##             0             0             0             0             0
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             0             0             0
##    Fireplaces    GarageType   GarageYrBlt  GarageFinish    GarageCars
##             0             0             0             0             0
##    GarageArea    GarageQual    GarageCond    PavedDrive    WoodDeckSF
##             0             0             0             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
```

```
remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}
train2$MSSubClass = remove_outliers(train2$MSSubClass)
train2$LotFrontage = remove_outliers(train2$LotFrontage)
train2$LotArea = remove_outliers(train2$LotArea)
train2$OverallQual = remove_outliers(train2$OverallQual)
train2$OverallCond = remove_outliers(train2$OverallCond)
train2$YearBuilt = remove_outliers(train2$YearBuilt)
train2$YearRemodAdd = remove_outliers(train2$YearRemodAdd)
train2$MasVnrArea = remove_outliers(train2$MasVnrArea)
train2$BsmtFinSF1 = remove_outliers(train2$BsmtFinSF1)
train2$BsmtFinSF2 = remove_outliers(train2$BsmtFinSF2)
train2$BsmtUnfSF = remove_outliers(train2$BsmtUnfSF)
train2$TotalBsmtSF = remove_outliers(train2$TotalBsmtSF)
train2$X1stFlrSF = remove_outliers(train2$X1stFlrSF)
train2$X2ndFlrSF = remove_outliers(train2$X2ndFlrSF)
train2$LowQualFinSF = remove_outliers(train2$LowQualFinSF)
train2$GrLivArea = remove_outliers(train2$GrLivArea)
train2$BsmtFullBath = remove_outliers(train2$BsmtFullBath)
train2$BsmtHalfBath = remove_outliers(train2$BsmtHalfBath)
train2$FullBath = remove_outliers(train2$FullBath)
train2$BedroomAbvGr = remove_outliers(train2$BedroomAbvGr)
train2$KitchenAbvGr = remove_outliers(train2$KitchenAbvGr)
train2$TotRmsAbvGrd = remove_outliers(train2$TotRmsAbvGrd)
train2$Fireplaces = remove_outliers(train2$Fireplaces)
train2$GarageYrBlt = remove_outliers(train2$GarageYrBlt)
```

```
train2$GarageCars = remove_outliers(train2$GarageCars)
train2$GarageArea = remove_outliers(train2$GarageArea)
train2$WoodDeckSF = remove_outliers(train2$WoodDeckSF)
train2$OpenPorchSF = remove_outliers(train2$OpenPorchSF)
train2$EnclosedPorch = remove_outliers(train2$EnclosedPorch)
train2$X3SsnPorch = remove_outliers(train2$X3SsnPorch)
train2$ScreenPorch = remove_outliers(train2$ScreenPorch)
train2$PoolArea = remove_outliers(train2$PoolArea)
train2$MiscVal = remove_outliers(train2$MiscVal)

#remove NAs for train2
#MSSubclass
train2$MSSubClass[which(is.na(train2$MSSubClass))] <- median(na.omit(train2$MSSubClass))
#LotFrontage
train2$LotFrontage[which(is.na(train2$LotFrontage))] <- median(na.omit(train2$LotFrontage))
#MSZoing
train2$MSZoning[which(is.na(train2$MSZoning))] <- as.character(train2$MSZoning[which.max(table(train2$MS
train2$LotArea[which(is.na(train2$LotArea))] <- median(na.omit(train2$LotArea))
train2$OverallCond[which(is.na(train2$OverallCond))] <- median(na.omit(train2$OverallCond))
train2$OverallQual[which(is.na(train2$OverallQual))] <- median(na.omit(train2$OverallQual))
train2$YearBuilt[which(is.na(train2$YearBuilt))] <- median(na.omit(train2$YearBuilt))
train2$MasVnrArea[which(is.na(train2$MasVnrArea))] <- median(na.omit(train2$MasVnrArea))
train2$BsmtFinSF1[which(is.na(train2$BsmtFinSF1))] <- median(na.omit(train2$BsmtFinSF1))
train2$BsmtFinSF2[which(is.na(train2$BsmtFinSF2))] <- median(na.omit(train2$BsmtFinSF2))
train2$BsmtUnfSF[which(is.na(train2$BsmtUnfSF))] <- median(na.omit(train2$BsmtUnfSF))
train2$TotalBsmtSF[which(is.na(train2$TotalBsmtSF))] <- median(na.omit(train2$TotalBsmtSF))
train2$X1stFlrSF[which(is.na(train2$X1stFlrSF))] <- median(na.omit(train2$X1stFlrSF))
train2$X2ndFlrSF[which(is.na(train2$X2ndFlrSF))] <- median(na.omit(train2$X2ndFlrSF))
train2$LowQualFinSF[which(is.na(train2$LowQualFinSF))] <- median(na.omit(train2$LowQualFinSF))
train2$GrLivArea[which(is.na(train2$GrLivArea))] <- median(na.omit(train2$GrLivArea))
train2$BsmtFullBath[which(is.na(train2$BsmtFullBath))] <- median(na.omit(train2$BsmtFullBath))
train2$BsmtHalfBath[which(is.na(train2$BsmtHalfBath))] <- median(na.omit(train2$BsmtHalfBath))
train2$FullBath[which(is.na(train2$FullBath))] <- median(na.omit(train2$FullBath))
train2$BedroomAbvGr[which(is.na(train2$BedroomAbvGr))] <- median(na.omit(train2$BedroomAbvGr))
train2$KitchenQual[which(is.na(train2$KitchenQual))] <- as.character(train2$KitchenQual[which.max(table
train2$TotRmsAbvGrd[which(is.na(train2$TotRmsAbvGrd))] <- median(na.omit(train2$TotRmsAbvGrd))
train2$Fireplaces[which(is.na(train2$Fireplaces))] <- median(na.omit(train2$Fireplaces))
train2$GarageYrBlt[which(is.na(train2$GarageYrBlt))] <- median(na.omit(train2$GarageYrBlt))
train2$GarageCars[which(is.na(train2$GarageCars))] <- median(na.omit(train2$GarageCars))
train2$GarageArea[which(is.na(train2$GarageArea))] <- median(na.omit(train2$GarageArea))
train2$WoodDeckSF[which(is.na(train2$WoodDeckSF))] <- median(na.omit(train2$WoodDeckSF))
train2$OpenPorchSF[which(is.na(train2$OpenPorchSF))] <- median(na.omit(train2$OpenPorchSF))
train2$EnclosedPorch[which(is.na(train2$EnclosedPorch))] <- median(na.omit(train2$EnclosedPorch))
train2$X3SsnPorch[which(is.na(train2$X3SsnPorch))] <- median(na.omit(train2$X3SsnPorch))
train2$ScreenPorch[which(is.na(train2$ScreenPorch))] <- median(na.omit(train2$ScreenPorch))
train2$PoolArea[which(is.na(train2$PoolArea))] <- median(na.omit(train2$PoolArea))
train2$MiscVal[which(is.na(train2$MiscVal))] <- median(na.omit(train2$MiscVal))
colSums(is.na(train2))
```

```
##            Ob    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street     LotShape   LandContour     Utilities     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
```

```
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             0             0             0
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##             0             0             0             0             0
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##             0             0             0             0             0
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0           126             0             0             0
##    Fireplaces   FireplaceQu    GarageType   GarageYrBlt  GarageFinish
##             0             0             0             0             0
##    GarageCars    GarageArea    GarageQual    GarageCond    PavedDrive
##             0             0             0             0             0
##    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch
##             0             0             0             0             0
##      PoolArea       MiscVal        MoSold        YrSold      SaleType
##             0             0             0             0             0
## SaleCondition     SalePrice
##             0             0
```

```r
remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}
test2$MSSubClass = remove_outliers(test2$MSSubClass)
test2$LotFrontage = remove_outliers(test2$LotFrontage)
test2$LotArea = remove_outliers(test2$LotArea)
test2$OverallQual = remove_outliers(test2$OverallQual)
test2$OverallCond = remove_outliers(test2$OverallCond)
test2$YearBuilt = remove_outliers(test2$YearBuilt)
test2$YearRemodAdd = remove_outliers(test2$YearRemodAdd)
test2$MasVnrArea = remove_outliers(test2$MasVnrArea)
test2$BsmtFinSF1 = remove_outliers(test2$BsmtFinSF1)
test2$BsmtFinSF2 = remove_outliers(test2$BsmtFinSF2)
test2$BsmtUnfSF = remove_outliers(test2$BsmtUnfSF)
test2$TotalBsmtSF = remove_outliers(test2$TotalBsmtSF)
test2$X1stFlrSF = remove_outliers(test2$X1stFlrSF)
test2$X2ndFlrSF = remove_outliers(test2$X2ndFlrSF)
test2$LowQualFinSF = remove_outliers(test2$LowQualFinSF)
test2$GrLivArea = remove_outliers(test2$GrLivArea)
test2$BsmtFullBath = remove_outliers(test2$BsmtFullBath)
test2$BsmtHalfBath = remove_outliers(test2$BsmtHalfBath)
```

```
test2$FullBath = remove_outliers(test2$FullBath)
test2$BedroomAbvGr = remove_outliers(test2$BedroomAbvGr)
test2$KitchenAbvGr = remove_outliers(test2$KitchenAbvGr)
test2$TotRmsAbvGrd = remove_outliers(test2$TotRmsAbvGrd)
test2$Fireplaces = remove_outliers(test2$Fireplaces)
test2$GarageYrBlt = remove_outliers(test2$GarageYrBlt)
test2$GarageCars = remove_outliers(test2$GarageCars)
test2$GarageArea = remove_outliers(test2$GarageArea)
test2$WoodDeckSF = remove_outliers(test2$WoodDeckSF)
test2$OpenPorchSF = remove_outliers(test2$OpenPorchSF)
test2$EnclosedPorch = remove_outliers(test2$EnclosedPorch)
test2$X3SsnPorch = remove_outliers(test2$X3SsnPorch)
test2$ScreenPorch = remove_outliers(test2$ScreenPorch)
test2$PoolArea = remove_outliers(test2$PoolArea)
test2$MiscVal = remove_outliers(test2$MiscVal)

test2$MSSubClass[which(is.na(test2$MSSubClass))] <- median(na.omit(test2$MSSubClass))
test2$LotFrontage[which(is.na(test2$LotFrontage))] <- mean(na.omit(test2$LotFrontage))

test2$LotArea[which(is.na(test2$LotArea))] <- median(na.omit(test2$LotArea))

test2$OverallCond[which(is.na(test2$OverallCond))] <- median(na.omit(test2$OverallCond))
test2$OverallQual[which(is.na(test2$OverallQual))] <- median(na.omit(test2$OverallQual))
test2$YearBuilt[which(is.na(test2$YearBuilt))] <- median(na.omit(test2$YearBuilt))

test2$MasVnrArea[which(is.na(test2$MasVnrArea))] <- median(na.omit(test2$MasVnrArea))
test2$BsmtFinSF1[which(is.na(test2$BsmtFinSF1))] <- median(na.omit(test2$BsmtFinSF1))
test2$BsmtFinSF2[which(is.na(test2$BsmtFinSF2))] <- median(na.omit(test2$BsmtFinSF2))

test2$BsmtUnfSF[which(is.na(test2$BsmtUnfSF))] <- median(na.omit(test2$BsmtUnfSF))
test2$TotalBsmtSF[which(is.na(test2$TotalBsmtSF))] <- median(na.omit(test2$TotalBsmtSF))
test2$X1stFlrSF[which(is.na(test2$X1stFlrSF))] <- median(na.omit(test2$X1stFlrSF))
test2$X2ndFlrSF[which(is.na(test2$X2ndFlrSF))] <- median(na.omit(test2$X2ndFlrSF))
test2$LowQualFinSF[which(is.na(test2$LowQualFinSF))] <- median(na.omit(test2$LowQualFinSF))
test2$GrLivArea[which(is.na(test2$GrLivArea))] <- median(na.omit(test2$GrLivArea))

test2$BsmtHalfBath[which(is.na(test2$BsmtHalfBath))] <- median(na.omit(test2$BsmtHalfBath))

test2$FullBath[which(is.na(test2$FullBath))] <- median(na.omit(test2$FullBath))
test2$BedroomAbvGr[which(is.na(test2$BedroomAbvGr))] <- median(na.omit(test2$BedroomAbvGr))

test2$TotRmsAbvGrd[which(is.na(test2$TotRmsAbvGrd))] <- median(na.omit(test2$TotRmsAbvGrd))
test2$Fireplaces[which(is.na(test2$Fireplaces))] <- median(na.omit(test2$Fireplaces))
test2$GarageYrBlt[which(is.na(test2$GarageYrBlt))] <- median(na.omit(test2$GarageYrBlt))
test2$GarageCars[which(is.na(test2$GarageCars))] <- median(na.omit(test2$GarageCars))

test2$GarageArea[which(is.na(test2$GarageArea))] <- median(na.omit(test2$GarageArea))
test2$WoodDeckSF[which(is.na(test2$WoodDeckSF))] <- median(na.omit(test2$WoodDeckSF))
test2$OpenPorchSF[which(is.na(test2$OpenPorchSF))] <- median(na.omit(test2$OpenPorchSF))
test2$EnclosedPorch[which(is.na(test2$EnclosedPorch))] <- median(na.omit(test2$EnclosedPorch))
test2$X3SsnPorch[which(is.na(test2$X3SsnPorch))] <- median(na.omit(test2$X3SsnPorch))
test2$ScreenPorch[which(is.na(test2$ScreenPorch))] <- median(na.omit(test2$ScreenPorch))
test2$PoolArea[which(is.na(test2$PoolArea))] <- median(na.omit(test2$PoolArea))
```

```
test2$MiscVal[which(is.na(test2$MiscVal))] <- median(na.omit(test2$MiscVal))
colSums(is.na(train2))
```

```
##            Ob    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             0             0             0
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##             0             0             0             0             0
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##             0             0             0             0             0
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0           126             0             0             0
##    Fireplaces   FireplaceQu    GarageType   GarageYrBlt  GarageFinish
##             0             0             0             0             0
##    GarageCars    GarageArea    GarageQual    GarageCond    PavedDrive
##             0             0             0             0             0
##    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch
##             0             0             0             0             0
##      PoolArea       MiscVal        MoSold        YrSold      SaleType
##             0             0             0             0             0
## SaleCondition     SalePrice
##             0             0
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
#First selecting numerical variables
model <- lm(SalePrice ~ MSZoning + LotFrontage + LotArea + OverallQual + OverallCond + YearBuilt + Yearl
summary(model)
```

```
##
## Call:
```

```
## lm(formula = SalePrice ~ MSZoning + LotFrontage + LotArea + OverallQual +
##     OverallCond + YearBuilt + YearRemodAdd + MasVnrArea + BsmtFinSF1 +
##     BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + X1stFlrSF + X2ndFlrSF +
##     KitchenAbvGr + LowQualFinSF + GrLivArea + BsmtFullBath +
##     BsmtHalfBath + FullBath + HalfBath + BedroomAbvGr + TotRmsAbvGrd +
##     Fireplaces + GarageYrBlt + GarageArea + WoodDeckSF + OpenPorchSF +
##     EnclosedPorch + X3SsnPorch + ScreenPorch + PoolArea + MiscVal +
##     MoSold + YrSold, data = train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140209  -16660   -2997   11119  386127
##
## Coefficients: (9 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.923e+05  9.858e+05   0.499 0.617540
## MSZoningFV    -1.530e+04  9.310e+03  -1.643 0.100496
## MSZoningRH    -2.797e+04  1.243e+04  -2.250 0.024528 *
## MSZoningRL    -1.607e+04  8.734e+03  -1.839 0.065974 .
## MSZoningRM    -1.229e+04  8.732e+03  -1.407 0.159557
## LotFrontage    1.792e+02  5.220e+01   3.432 0.000609 ***
## LotArea        1.523e+00  2.632e-01   5.788 8.08e-09 ***
## OverallQual    2.320e+04  7.748e+02  29.942  < 2e-16 ***
## OverallCond    1.690e+03  9.223e+02   1.832 0.067090 .
## YearBuilt      1.488e+02  4.705e+01   3.162 0.001585 **
## YearRemodAdd   1.020e+02  4.561e+01   2.237 0.025392 *
## MasVnrArea    -2.853e+01  6.561e+00  -4.348 1.43e-05 ***
## BsmtFinSF1     3.421e+01  2.716e+00  12.595  < 2e-16 ***
## BsmtFinSF2            NA         NA      NA       NA
## BsmtUnfSF      6.420e+00  2.629e+00   2.442 0.014700 *
## TotalBsmtSF    1.196e+01  4.060e+00   2.947 0.003244 **
## X1stFlrSF      3.014e+01  4.511e+00   6.681 2.96e-11 ***
## X2ndFlrSF      1.645e+01  3.498e+00   4.704 2.70e-06 ***
## KitchenAbvGr          NA         NA      NA       NA
## LowQualFinSF          NA         NA      NA       NA
## GrLivArea      6.984e+00  3.334e+00   2.095 0.036297 *
## BsmtFullBath   3.464e+03  1.683e+03   2.059 0.039631 *
## BsmtHalfBath          NA         NA      NA       NA
## FullBath       4.293e+03  1.932e+03   2.221 0.026423 *
## HalfBath       5.375e+03  1.899e+03   2.831 0.004681 **
## BedroomAbvGr  -4.121e+03  1.284e+03  -3.211 0.001342 **
## TotRmsAbvGrd   6.959e+03  8.105e+02   8.585  < 2e-16 ***
## Fireplaces     5.688e+03  1.257e+03   4.523 6.39e-06 ***
## GarageYrBlt    5.326e+01  4.942e+01   1.078 0.281285
## GarageArea     3.093e+01  4.666e+00   6.629 4.19e-11 ***
## WoodDeckSF     1.733e+01  6.390e+00   2.711 0.006748 **
## OpenPorchSF    3.599e+01  1.561e+01   2.306 0.021215 *
## EnclosedPorch         NA         NA      NA       NA
## X3SsnPorch            NA         NA      NA       NA
## ScreenPorch           NA         NA      NA       NA
## PoolArea              NA         NA      NA       NA
## MiscVal               NA         NA      NA       NA
## MoSold         7.022e+01  2.465e+02   0.285 0.775771
## YrSold        -6.042e+02  4.889e+02  -1.236 0.216703
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30770 on 2344 degrees of freedom
##   (126 observations deleted due to missingness)
## Multiple R-squared:  0.8503, Adjusted R-squared:  0.8484
## F-statistic:   459 on 29 and 2344 DF,  p-value: < 2.2e-16
```

```
# keep variables with siginicant p-values
model.1 <- lm(SalePrice ~ LotArea + OverallQual + YearBuilt + YearRemodAdd + MasVnrArea + BsmtFinSF1 + B
summary(model.1)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + YearBuilt +
##     YearRemodAdd + MasVnrArea + BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF +
##     TotalBsmtSF + X1stFlrSF + X2ndFlrSF + BsmtFullBath + BedroomAbvGr +
##     TotRmsAbvGrd + Fireplaces + GarageArea + ScreenPorch + PoolArea +
##     MiscVal, data = train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148337  -16198   -2534   11506  408221
##
## Coefficients: (4 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.745e+05  7.886e+04 -11.089  < 2e-16 ***
## LotArea       1.897e+00  2.278e-01   8.329  < 2e-16 ***
## OverallQual   2.380e+04  7.421e+02  32.079  < 2e-16 ***
## YearBuilt     2.443e+02  3.197e+01   7.643 3.00e-14 ***
## YearRemodAdd  1.451e+02  4.205e+01   3.452 0.000567 ***
## MasVnrArea   -2.560e+01  6.369e+00  -4.020 5.99e-05 ***
## BsmtFinSF1    3.433e+01  2.554e+00  13.446  < 2e-16 ***
## BsmtFinSF2           NA         NA      NA       NA
## BsmtUnfSF     6.537e+00  2.384e+00   2.742 0.006152 **
## TotalBsmtSF   1.135e+01  3.822e+00   2.970 0.003002 **
## X1stFlrSF     3.466e+01  3.893e+00   8.903  < 2e-16 ***
## X2ndFlrSF     2.861e+01  2.376e+00  12.040  < 2e-16 ***
## BsmtFullBath  2.670e+03  1.584e+03   1.686 0.092005 .
## BedroomAbvGr -4.589e+03  1.203e+03  -3.816 0.000139 ***
## TotRmsAbvGrd  6.621e+03  6.954e+02   9.522  < 2e-16 ***
## Fireplaces    6.697e+03  1.183e+03   5.663 1.66e-08 ***
## GarageArea    3.129e+01  4.240e+00   7.381 2.13e-13 ***
## ScreenPorch          NA         NA      NA       NA
## PoolArea             NA         NA      NA       NA
## MiscVal              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31210 on 2484 degrees of freedom
## Multiple R-squared:  0.842,  Adjusted R-squared:  0.841
## F-statistic: 882.4 on 15 and 2484 DF,  p-value: < 2.2e-16
```

```
# And I still choose GrLivArea over X1stFlrSF + X2ndFlrSF
# I keep FUll bath (need transformation) instead of BsmtFullBath
```

```r
#i only keep TotRmsAbvGrd between BedroomAbvGr and TotRmsAbvGrd
```

```r
dim(train2)
```

```
## [1] 2500   77
```

```r
dim(test2)
```

```
## [1] 1500   75
```

```r
# I combined YearBuit and YearRemodAdd into one predictor: Age
train2[78] <- data.frame("Age" = train2$YearRemodAdd - train2$YearBuilt)
test2[76] <- data.frame("Age" = test2$YearRemodAdd - test2$YearBuilt)
```

```r
model.2 <- lm(SalePrice ~ LotArea + OverallQual  + Age + BsmtFinSF1 + + BsmtUnfSF +TotalBsmtSF  + GrLiv
summary(model.2)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + Age + BsmtFinSF1 +
##     +BsmtUnfSF + TotalBsmtSF + GrLivArea + FullBath + TotRmsAbvGrd +
##     Fireplaces + GarageArea, data = train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -154461  -17852   -2071   11827  378416
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.335e+05  3.967e+03 -33.663  < 2e-16 ***
## LotArea       1.774e+00  2.298e-01   7.720 1.67e-14 ***
## OverallQual   2.743e+04  6.705e+02  40.912  < 2e-16 ***
## Age          -5.688e+01  3.023e+01  -1.882    0.060 .
## BsmtFinSF1    3.413e+01  2.440e+00  13.990  < 2e-16 ***
## BsmtUnfSF     1.827e+00  2.413e+00   0.757    0.449
## TotalBsmtSF   2.073e+01  2.767e+00   7.492 9.33e-14 ***
## GrLivArea     1.946e+01  2.594e+00   7.503 8.64e-14 ***
## FullBath      7.804e+03  1.638e+03   4.764 2.01e-06 ***
## TotRmsAbvGrd  5.421e+03  6.818e+02   7.951 2.78e-15 ***
## Fireplaces    6.551e+03  1.198e+03   5.467 5.04e-08 ***
## GarageArea    3.923e+01  4.286e+00   9.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32180 on 2488 degrees of freedom
## Multiple R-squared:  0.8317, Adjusted R-squared:  0.8309
## F-statistic:  1117 on 11 and 2488 DF,  p-value: < 2.2e-16
```

```r
vif(model.2)
```

```
##      LotArea  OverallQual          Age   BsmtFinSF1    BsmtUnfSF  TotalBsmtSF
##     1.321707     2.194310     1.202336     2.715034     2.446185     2.467341
##    GrLivArea     FullBath TotRmsAbvGrd   Fireplaces   GarageArea
##     3.145987     1.891017     2.345328     1.362708     1.760802
```

```r
#delete predictors (vif greater than 5)
model.3 <- lm(SalePrice ~ LotArea + OverallQual  + Age + GrLivArea + FullBath + TotRmsAbvGrd + Fireplac
```

```r
summary(model.3)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + Age + GrLivArea +
##     FullBath + TotRmsAbvGrd + Fireplaces + GarageArea, data = train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152071  -21373   -3368   16388  392619
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.254e+05  4.333e+03 -28.932  < 2e-16 ***
## LotArea       2.763e+00  2.497e-01  11.065  < 2e-16 ***
## OverallQual   3.017e+04  7.141e+02  42.248  < 2e-16 ***
## Age          -1.821e+02  3.305e+01  -5.510 3.96e-08 ***
## GrLivArea     2.138e+01  2.891e+00   7.398 1.89e-13 ***
## FullBath      6.734e+03  1.808e+03   3.724 0.000201 ***
## TotRmsAbvGrd  4.277e+03  7.571e+02   5.650 1.79e-08 ***
## Fireplaces    1.169e+04  1.310e+03   8.917  < 2e-16 ***
## GarageArea    5.755e+01  4.687e+00  12.279  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35880 on 2491 degrees of freedom
## Multiple R-squared:  0.7905, Adjusted R-squared:  0.7898
## F-statistic:  1175 on 8 and 2491 DF,  p-value: < 2.2e-16
```

```r
vif(model.3)
```

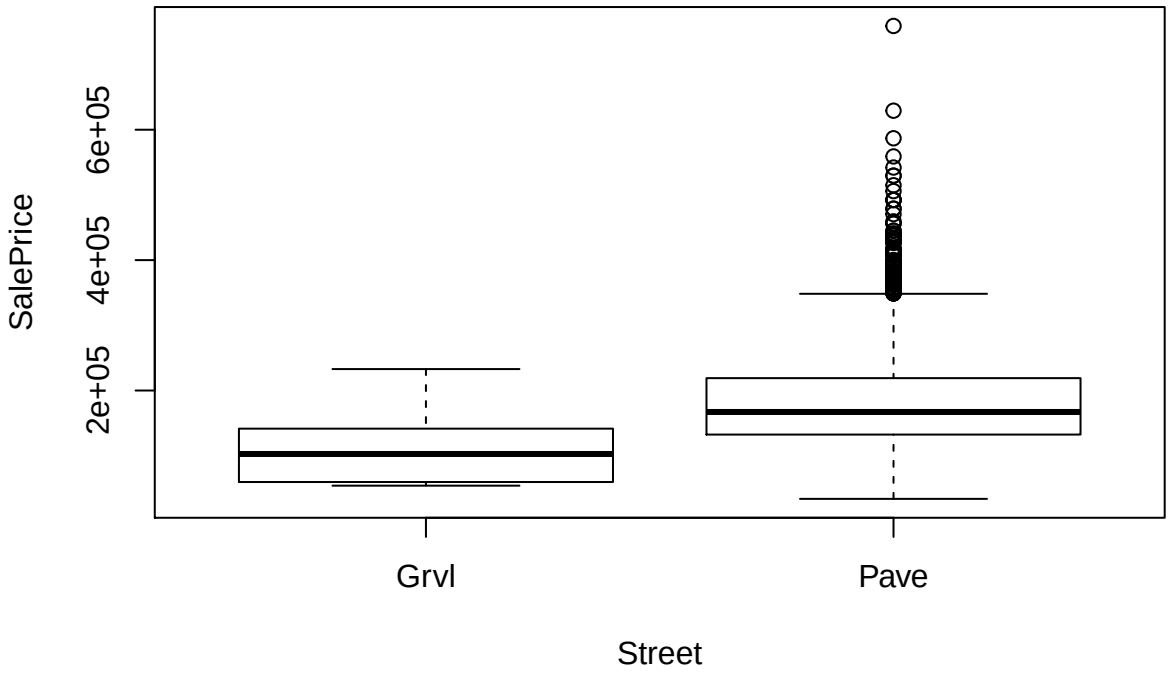
```
##      LotArea  OverallQual          Age    GrLivArea     FullBath TotRmsAbvGrd
##     1.256462     2.002431     1.156760     3.142608     1.853543     2.326601
##   Fireplaces   GarageArea
##     1.311142     1.693859
```

```r
# delete meaningless predictors(in my opinion) and we can see that R^2 does not get affected much
#the predictors inside model.4 is all the numerical predictors that we pick
model.4 <- lm(SalePrice ~ LotArea + OverallQual + Age + GrLivArea + FullBath + TotRmsAbvGrd + GarageArea
summary(model.4)
```

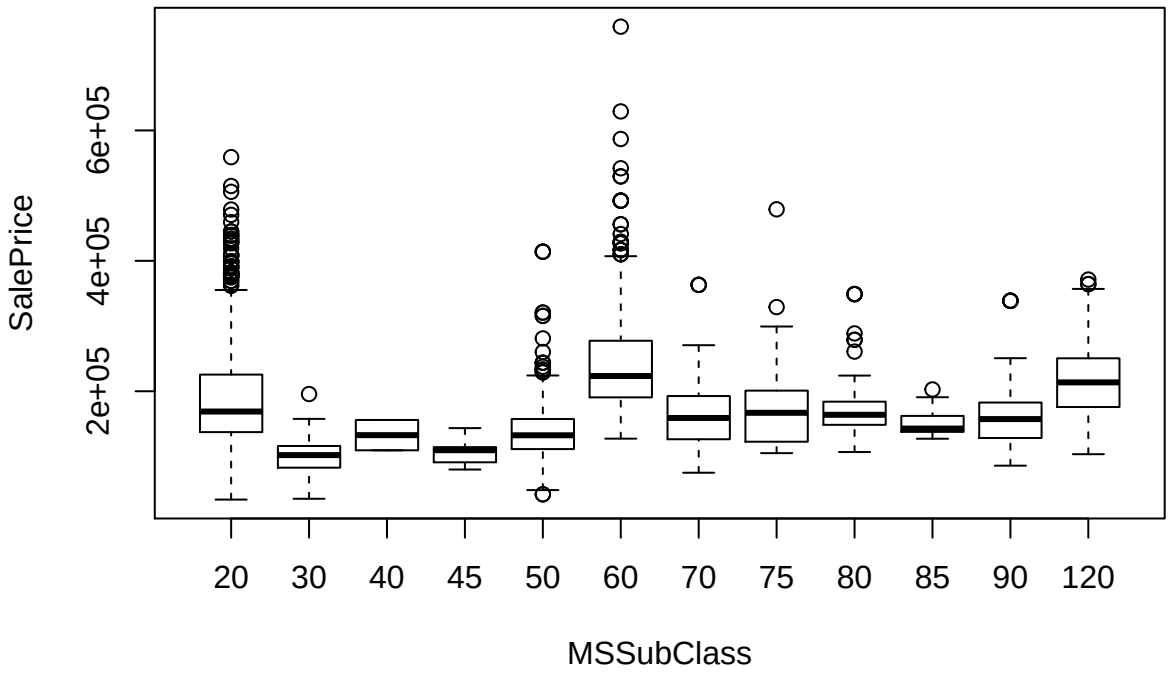
```
##
## Call:
## lm(formula = SalePrice ~ LotArea + OverallQual + Age + GrLivArea +
##     FullBath + TotRmsAbvGrd + GarageArea, data = train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -149754  -21270   -3409   16881  406322
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.341e+05  4.287e+03 -31.286  < 2e-16 ***
## LotArea       3.029e+00  2.518e-01  12.029  < 2e-16 ***
## OverallQual   3.154e+04  7.084e+02  44.521  < 2e-16 ***
```

```
## Age          -1.967e+02  3.353e+01  -5.866 5.05e-09 ***
## GrLivArea     2.648e+01  2.878e+00   9.199  < 2e-16 ***
## FullBath      4.983e+03  1.826e+03   2.730  0.00639 **
## TotRmsAbvGrd  4.346e+03  7.689e+02   5.652 1.76e-08 ***
## GarageArea    5.706e+01  4.760e+00  11.988  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36440 on 2492 degrees of freedom
## Multiple R-squared:  0.7838, Adjusted R-squared:  0.7832
## F-statistic:  1291 on 7 and 2492 DF,  p-value: < 2.2e-16
```

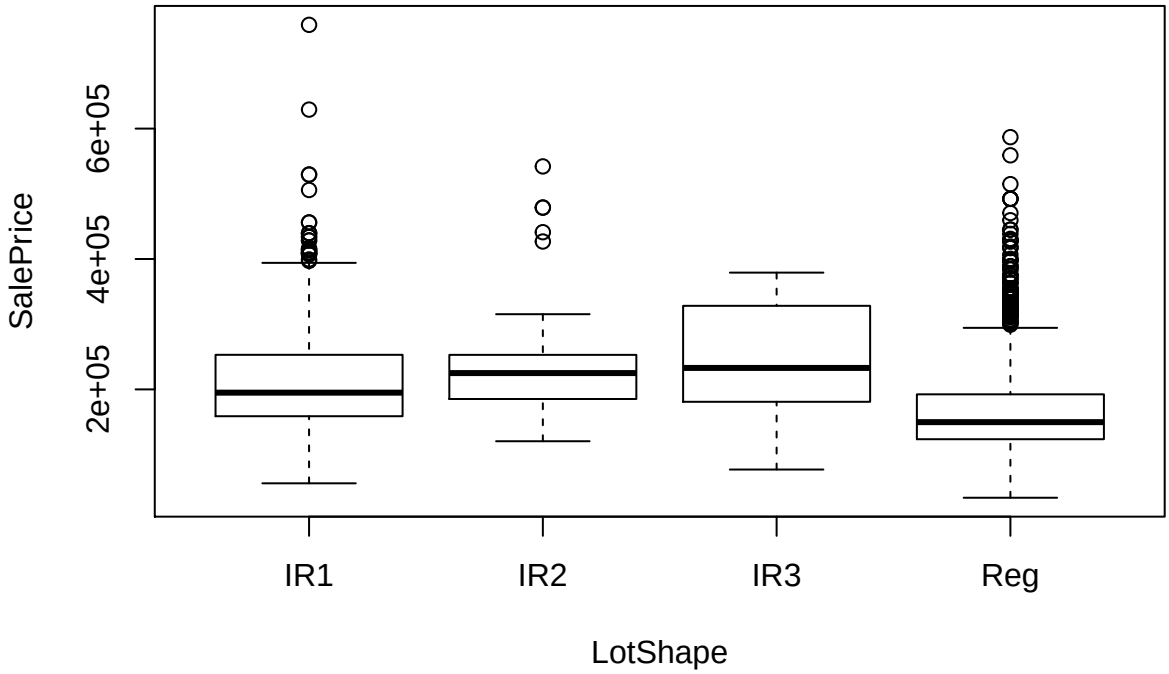
```
#pick categorial variables
attach(train2)
boxplot(SalePrice ~ Street)
```



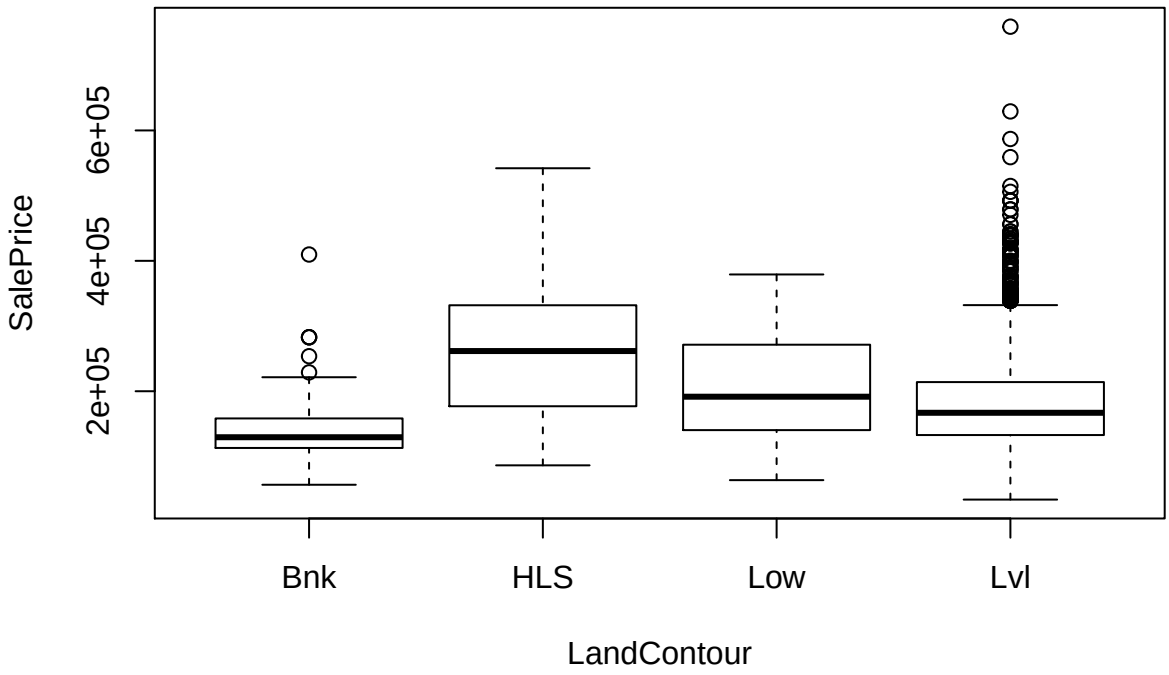

```
boxplot(SalePrice ~ MSSubClass)
```



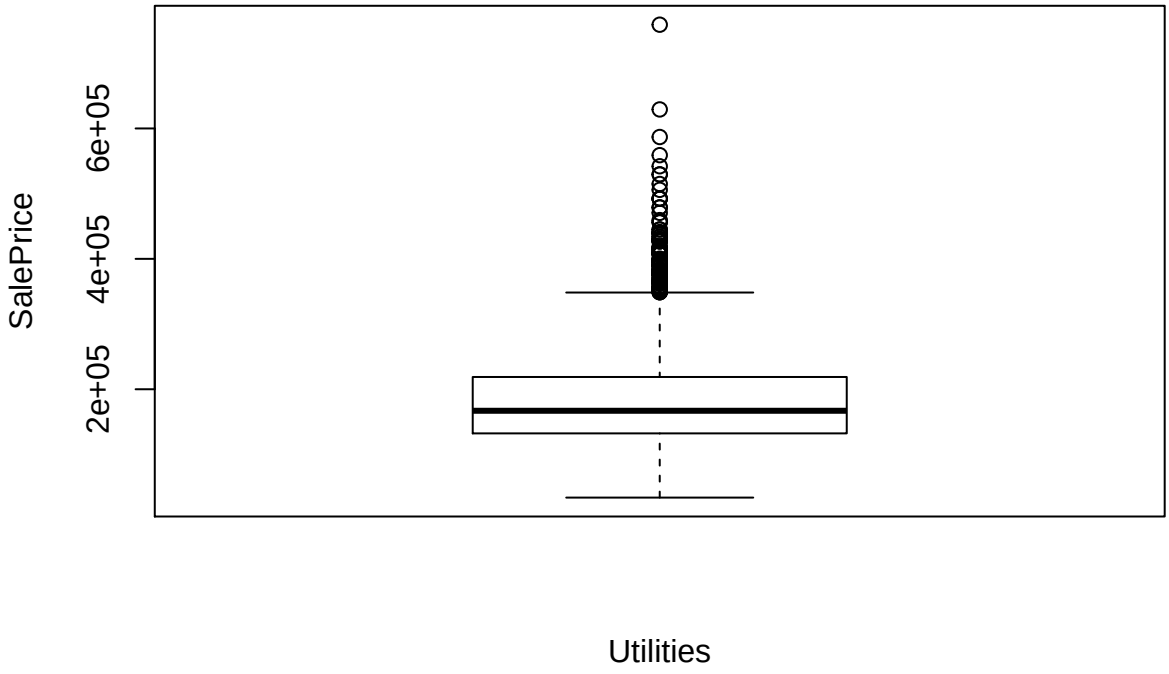

```
boxplot(SalePrice ~ LotShape)
```



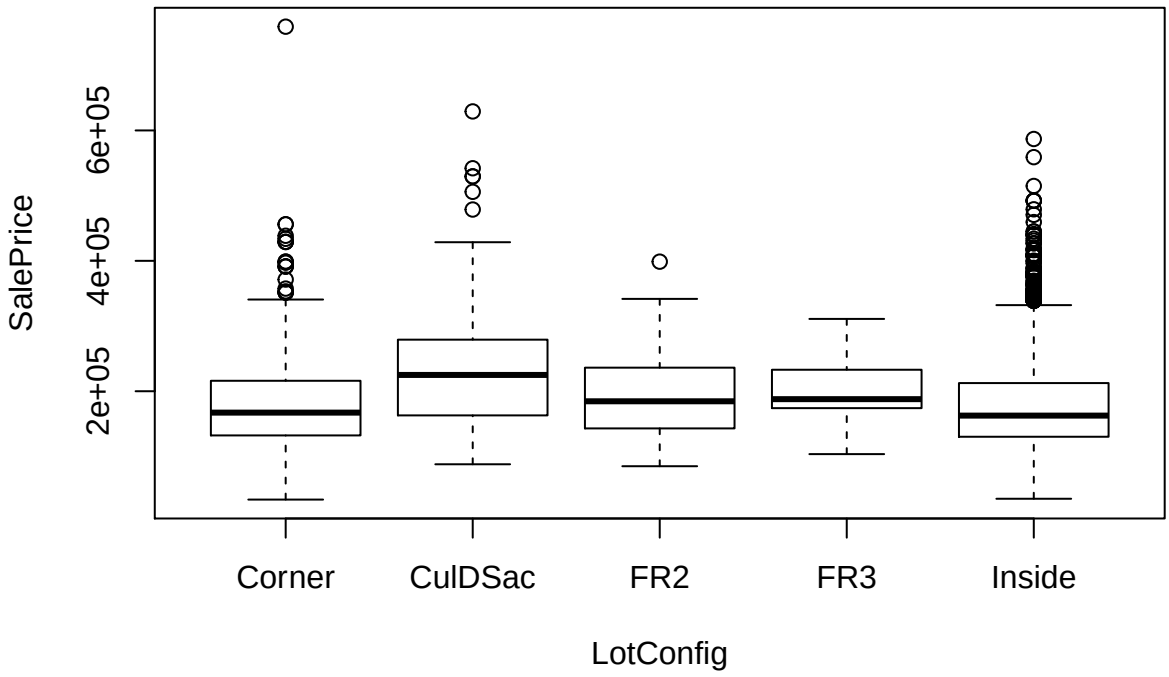

```
boxplot(SalePrice ~ LandContour)
```



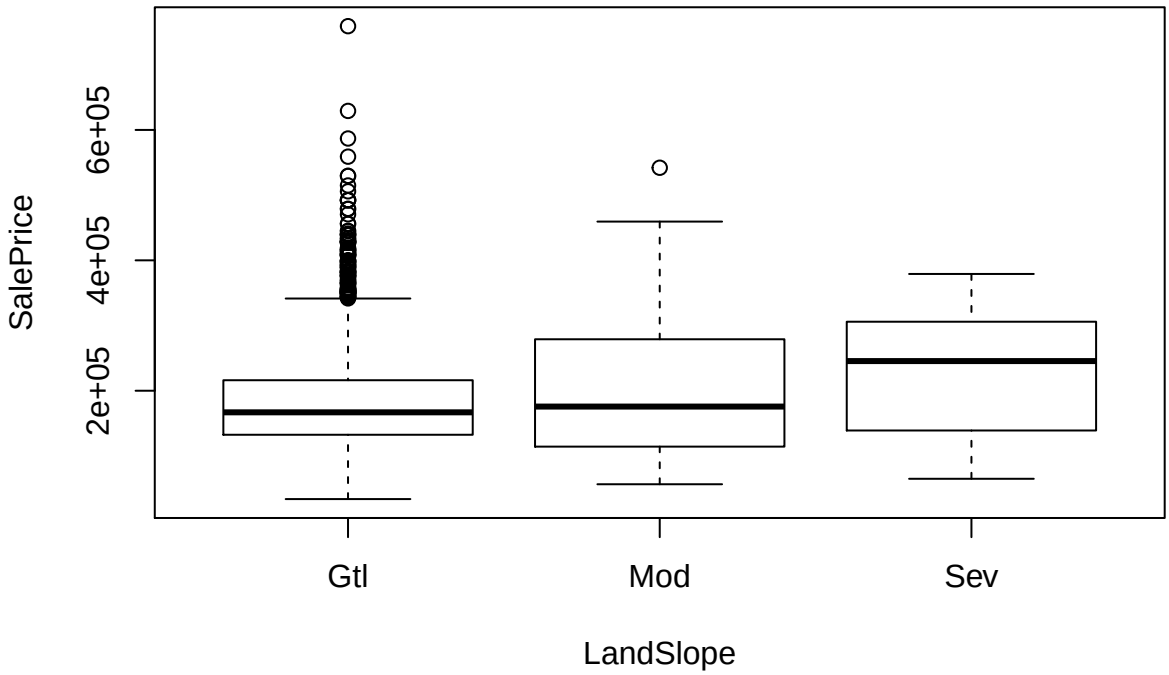

```
boxplot(SalePrice ~ Utilities)
```



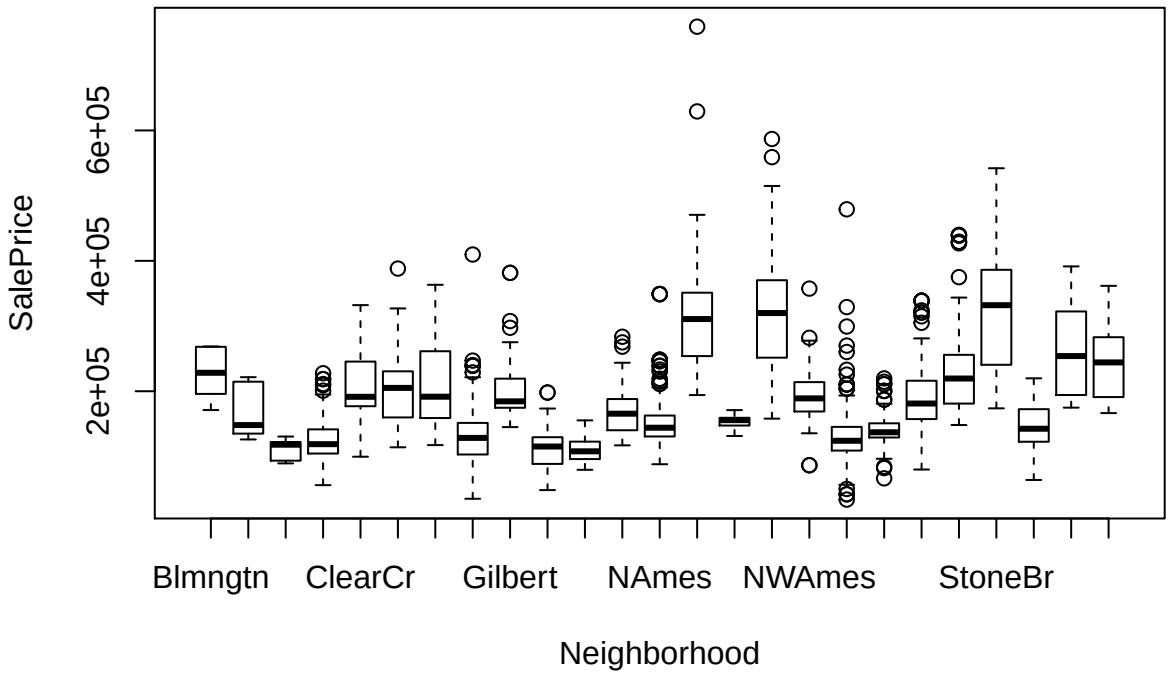

```
boxplot(SalePrice ~ LotConfig)
```



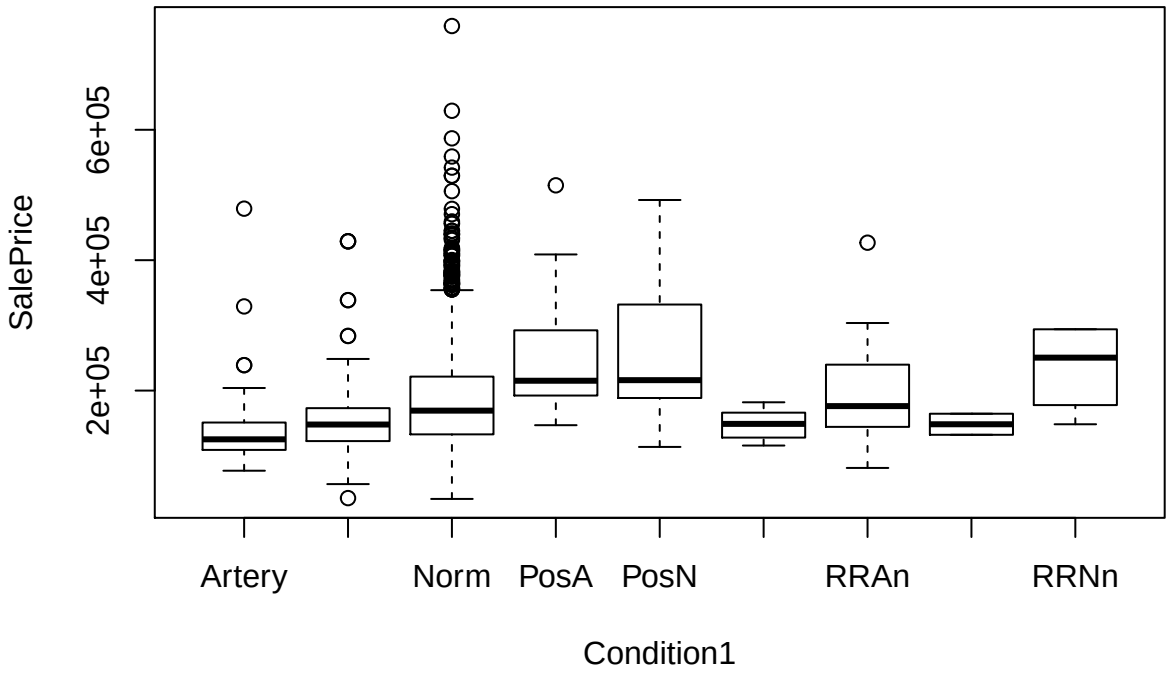

```
boxplot(SalePrice ~ LandSlope)
```



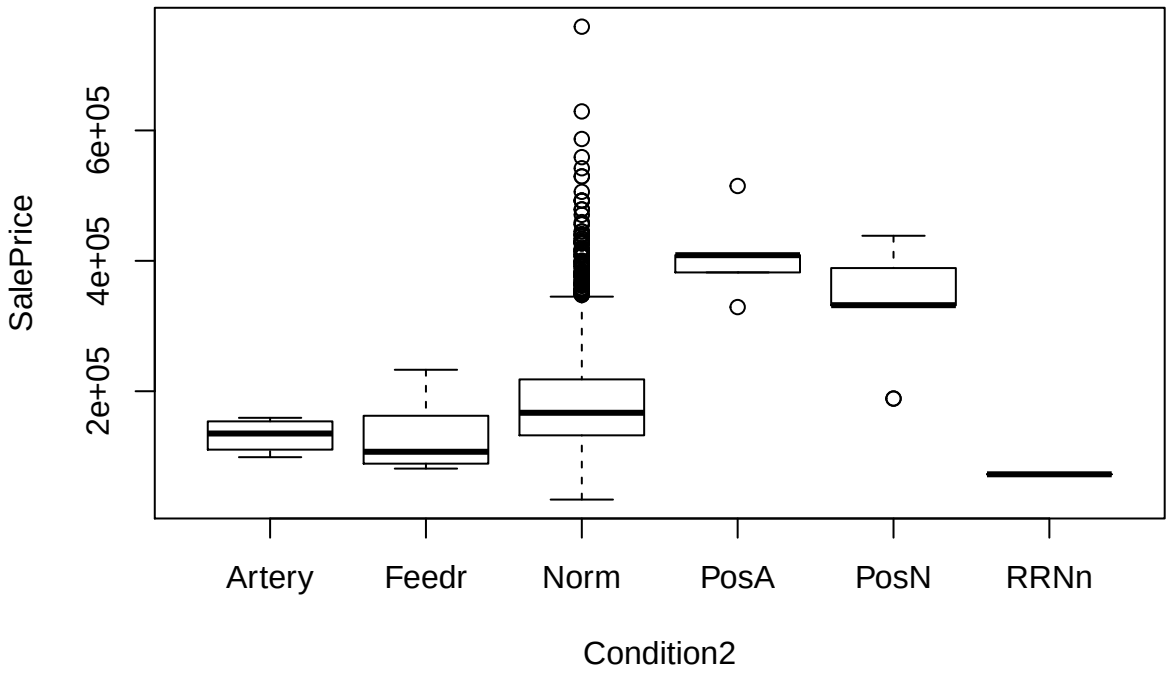

```
boxplot(SalePrice ~ Neighborhood)
```



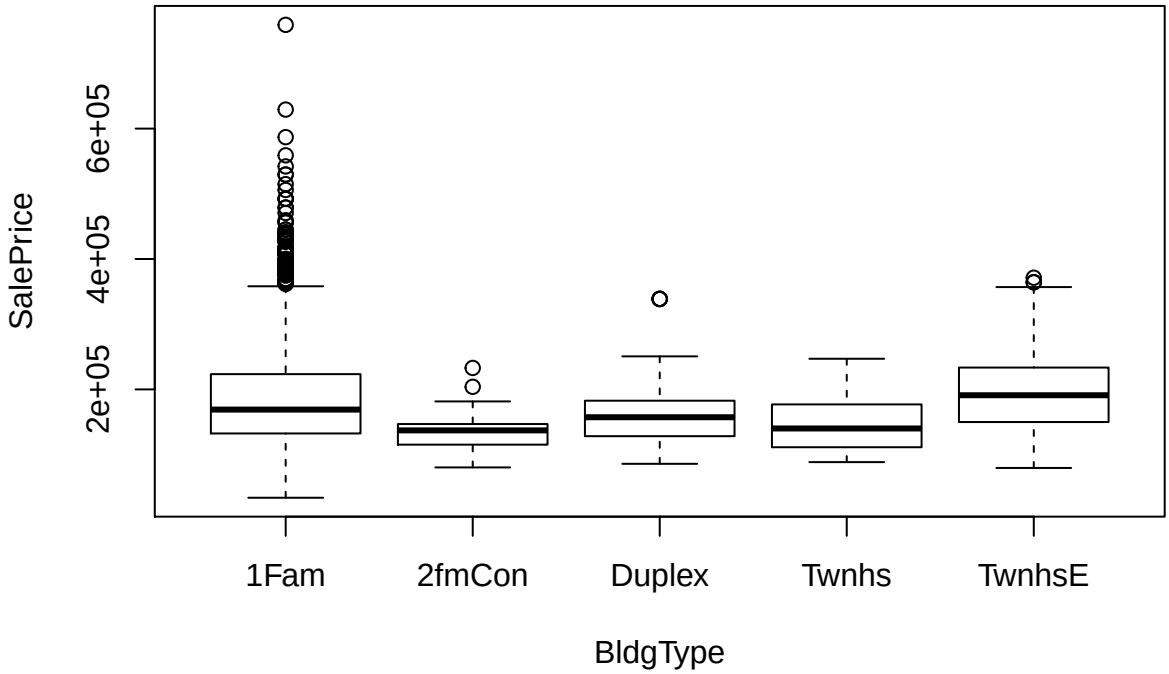

```
boxplot(SalePrice ~ Condition1)
```



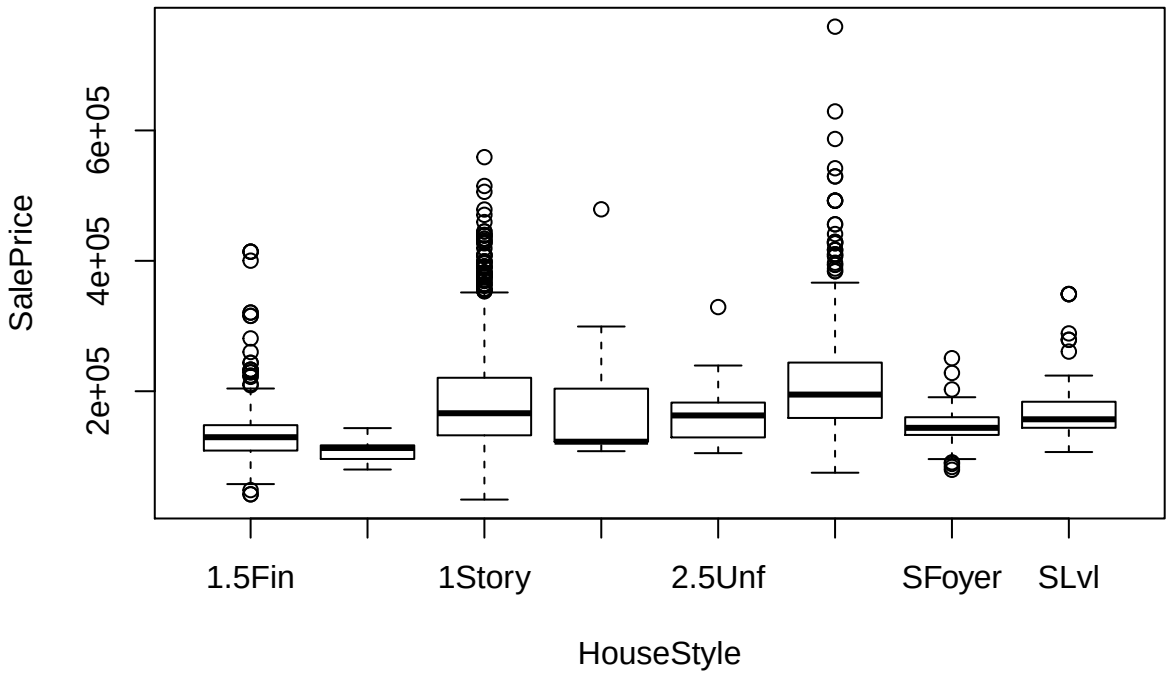

```
boxplot(SalePrice ~ Condition2)
```



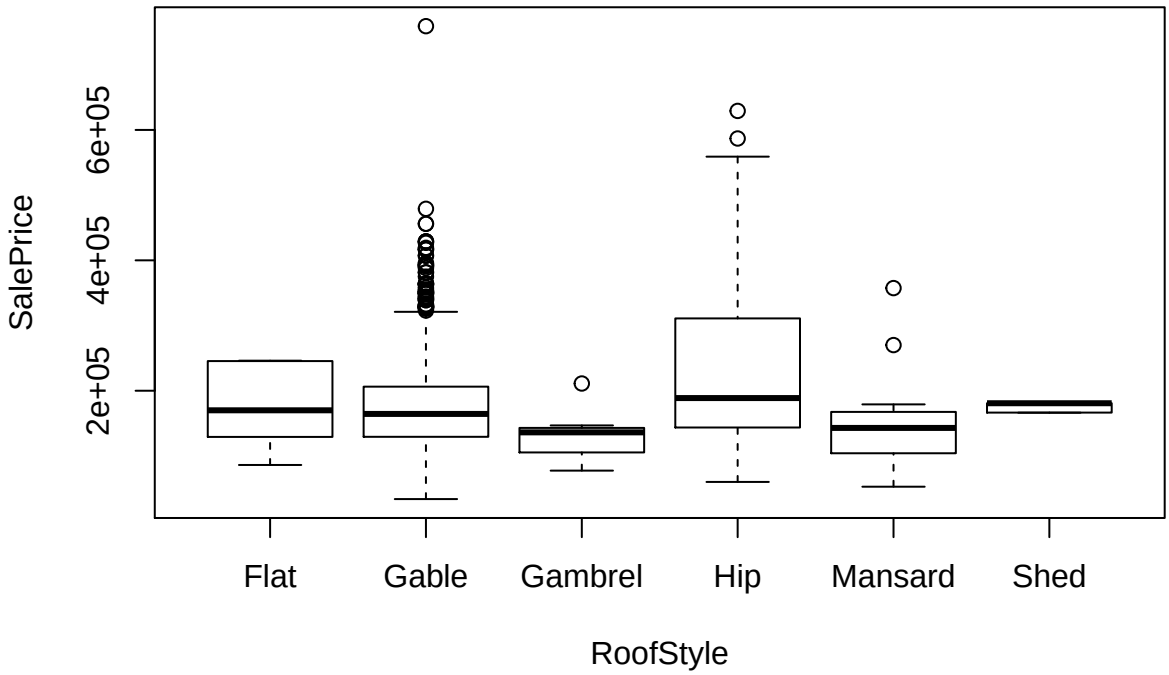

```
boxplot(SalePrice ~ BldgType)
```



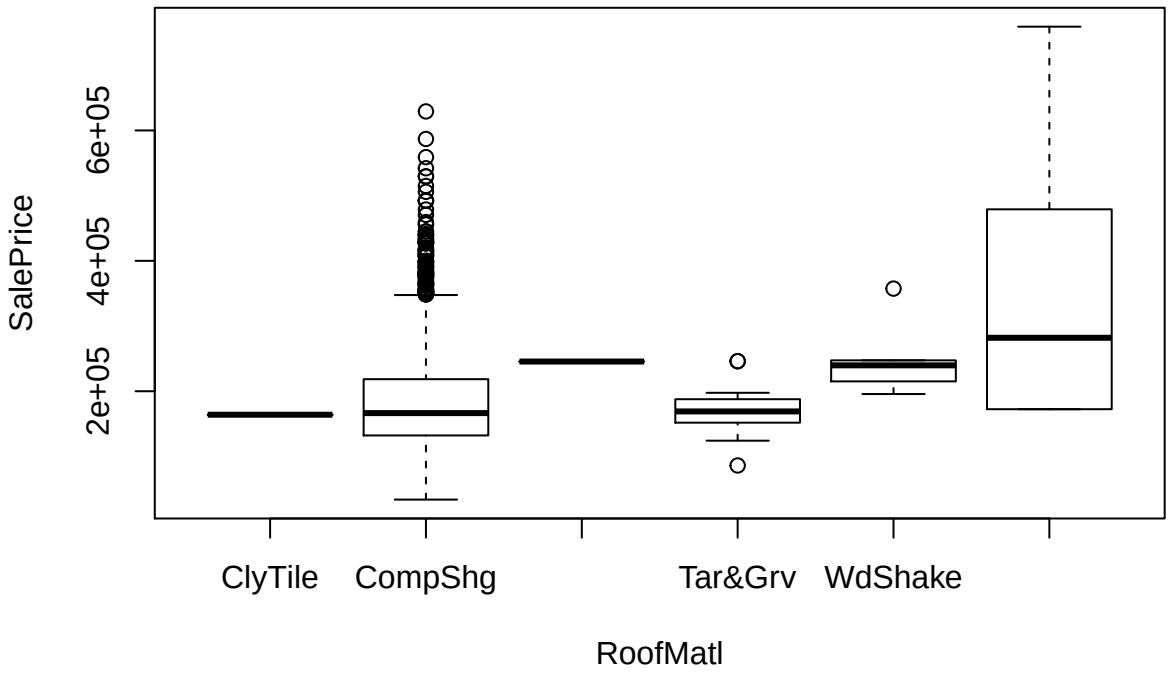

```
boxplot(SalePrice ~ HouseStyle)
```



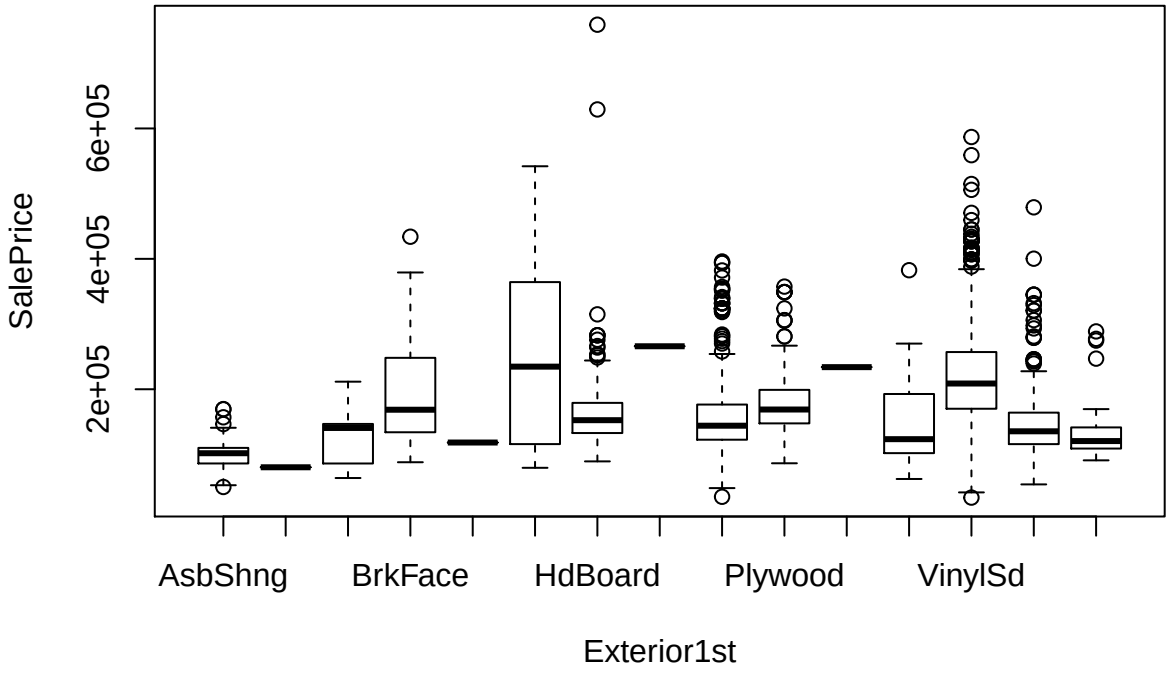

```
boxplot(SalePrice ~ RoofStyle)
```



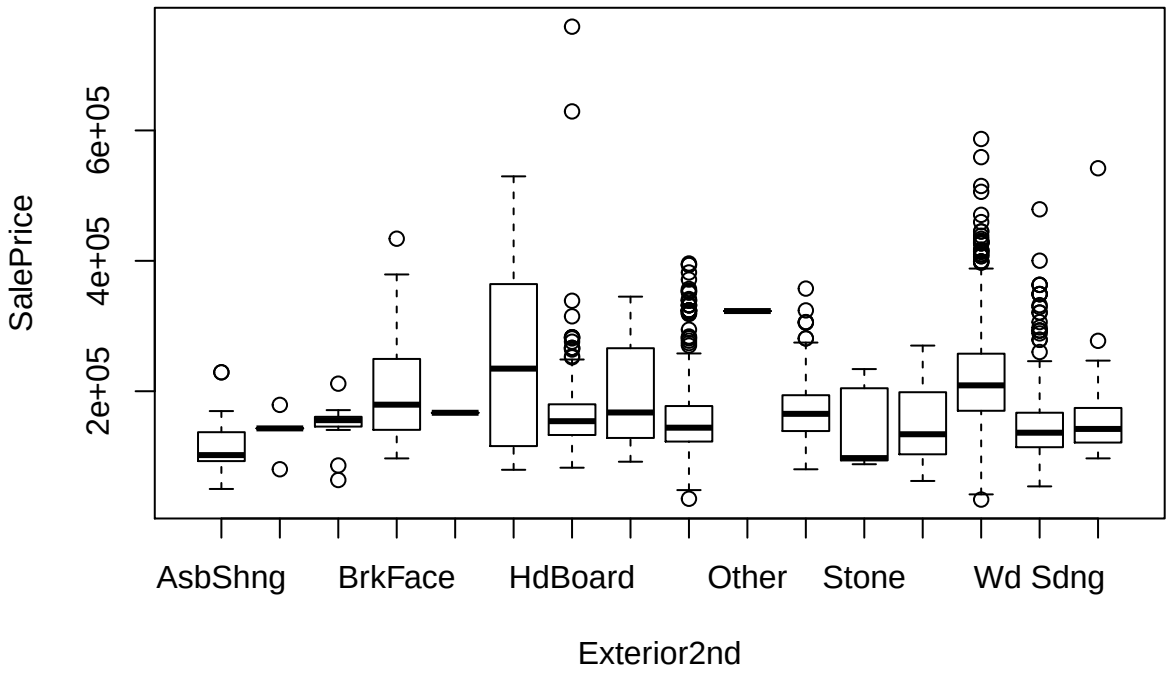

```
boxplot(SalePrice ~ RoofMatl)
```



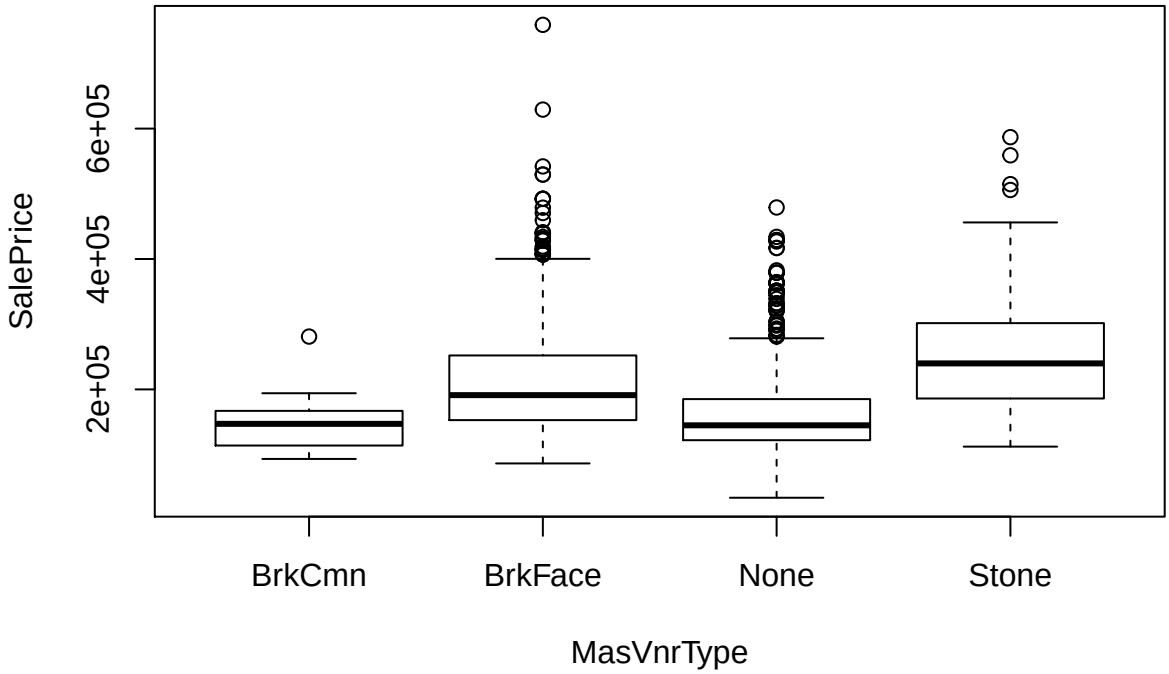

```
boxplot(SalePrice ~ Exterior1st)
```



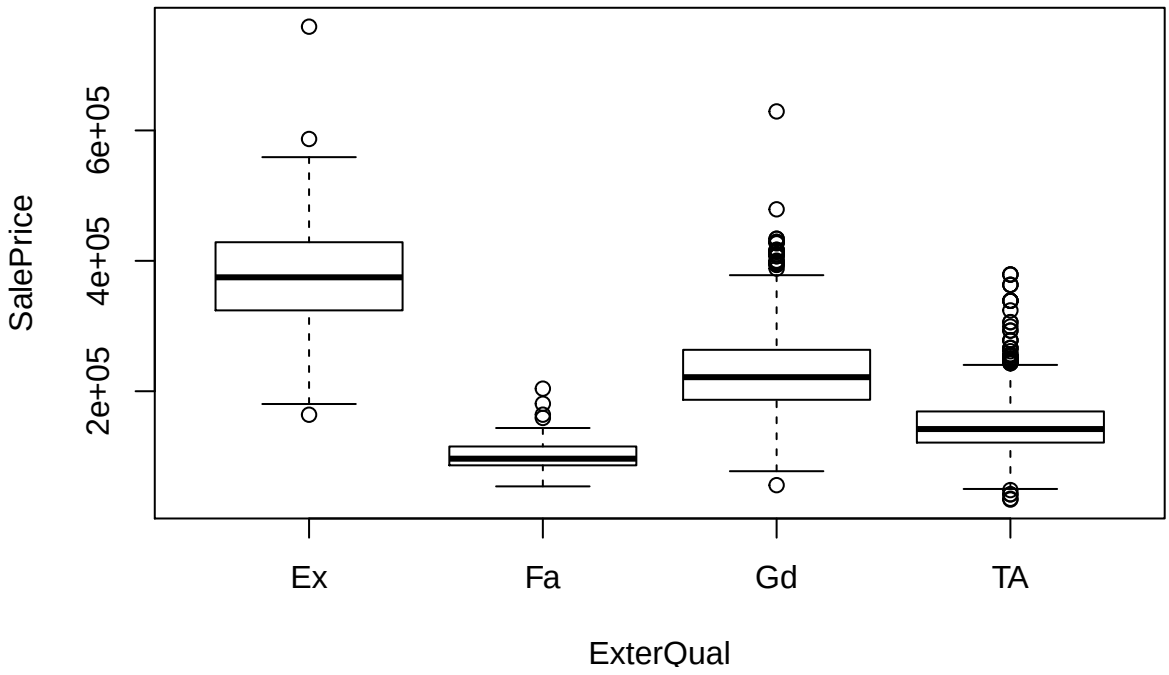

```
boxplot(SalePrice ~ Exterior2nd)
```



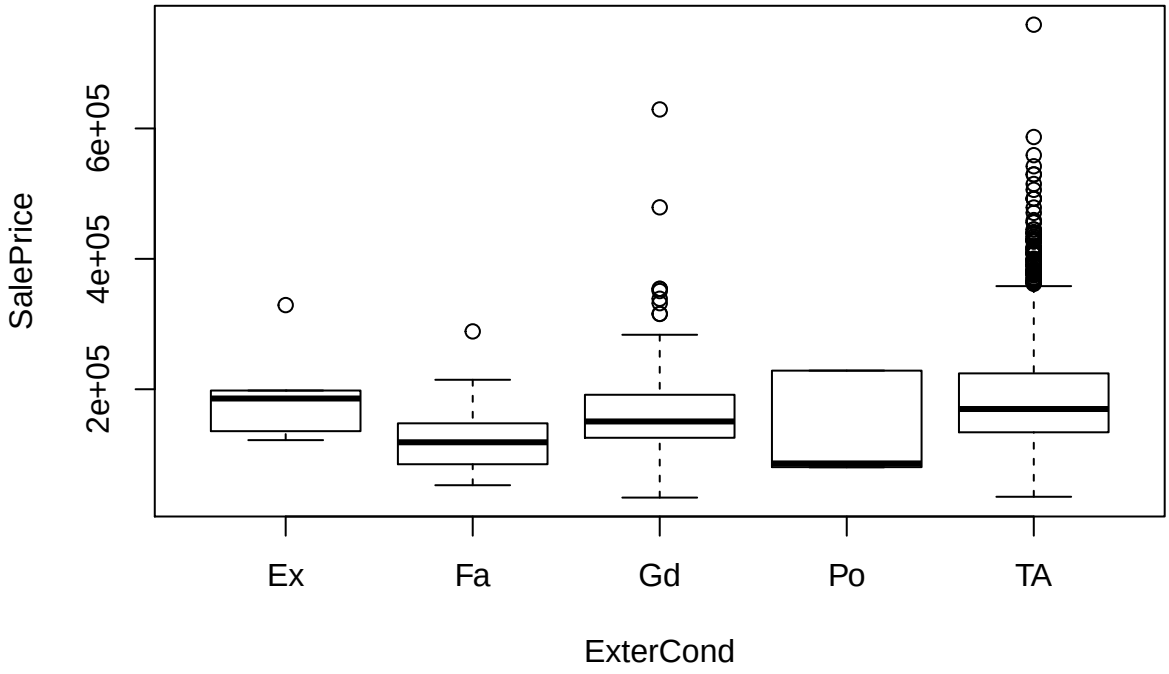

```
boxplot(SalePrice ~ MasVnrType)
```



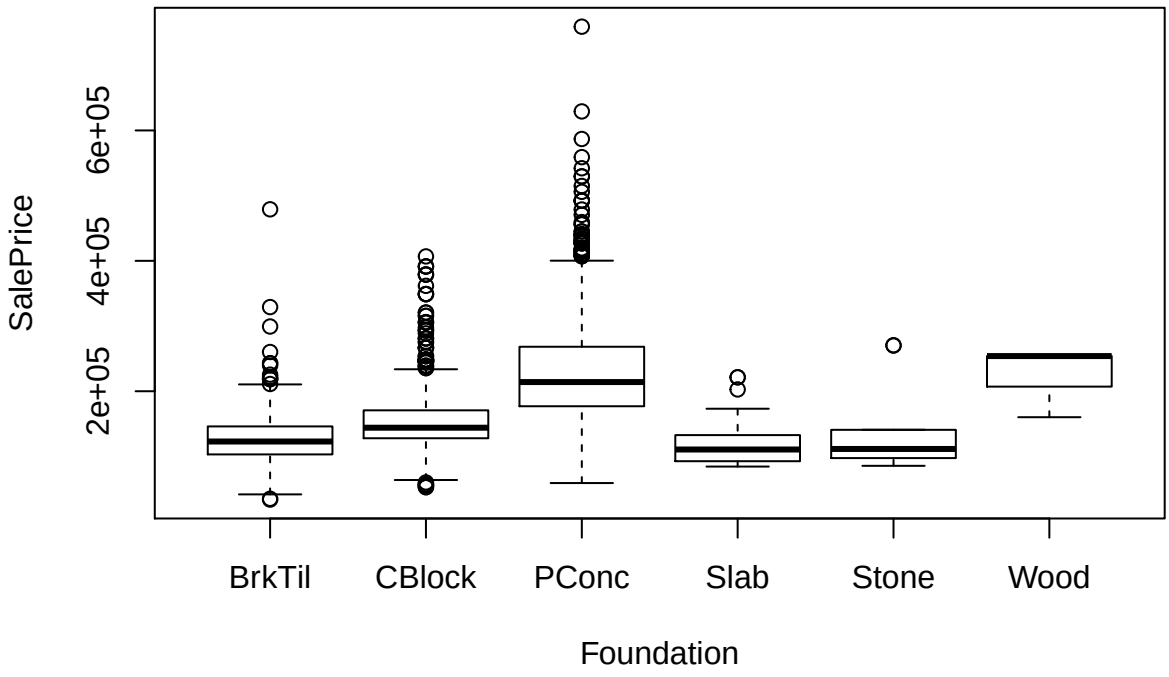

```
boxplot(SalePrice ~ ExterQual)
```



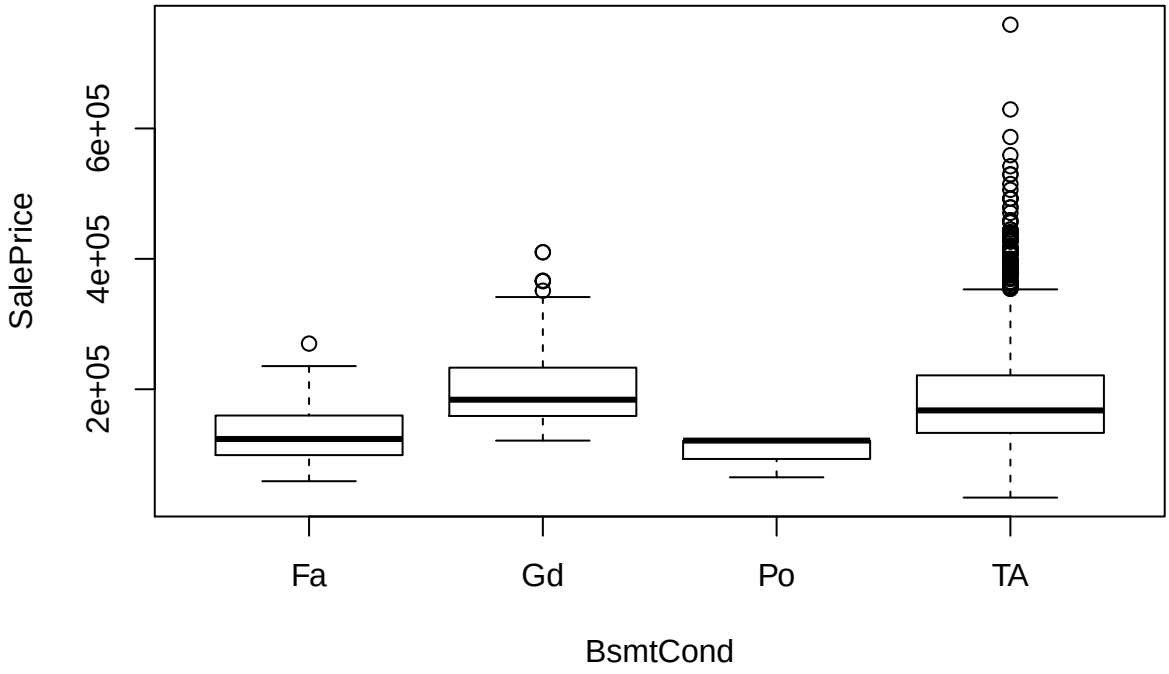

```
boxplot(SalePrice ~ ExterCond)
```



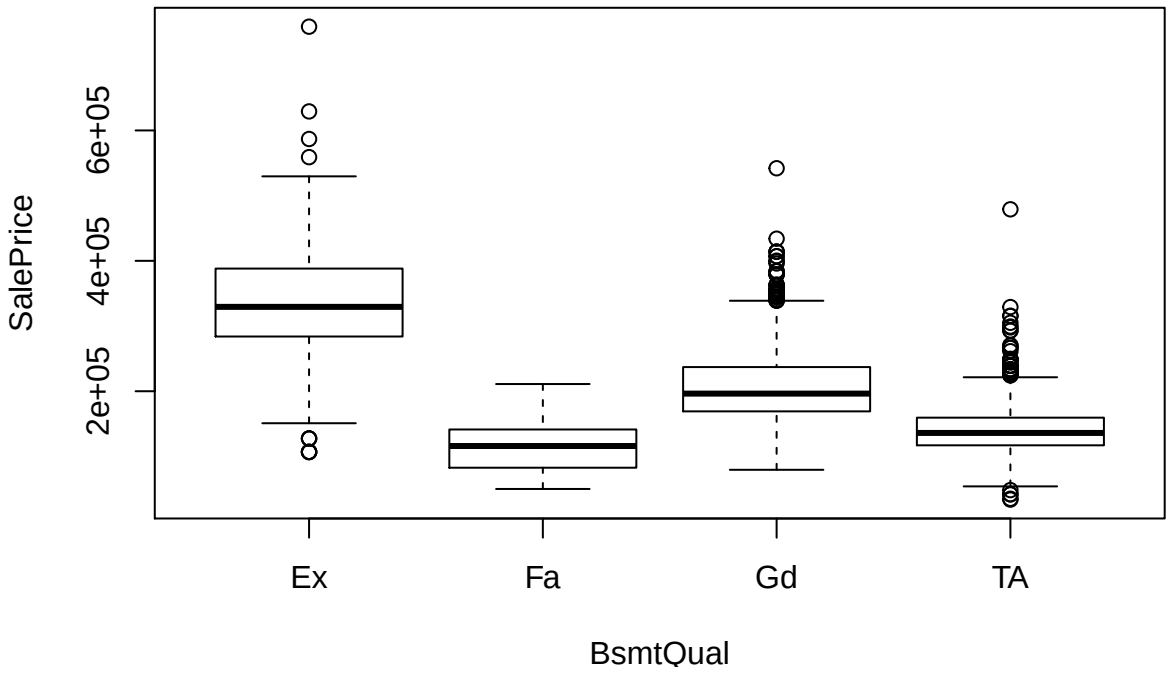

```
boxplot(SalePrice ~ Foundation)
```



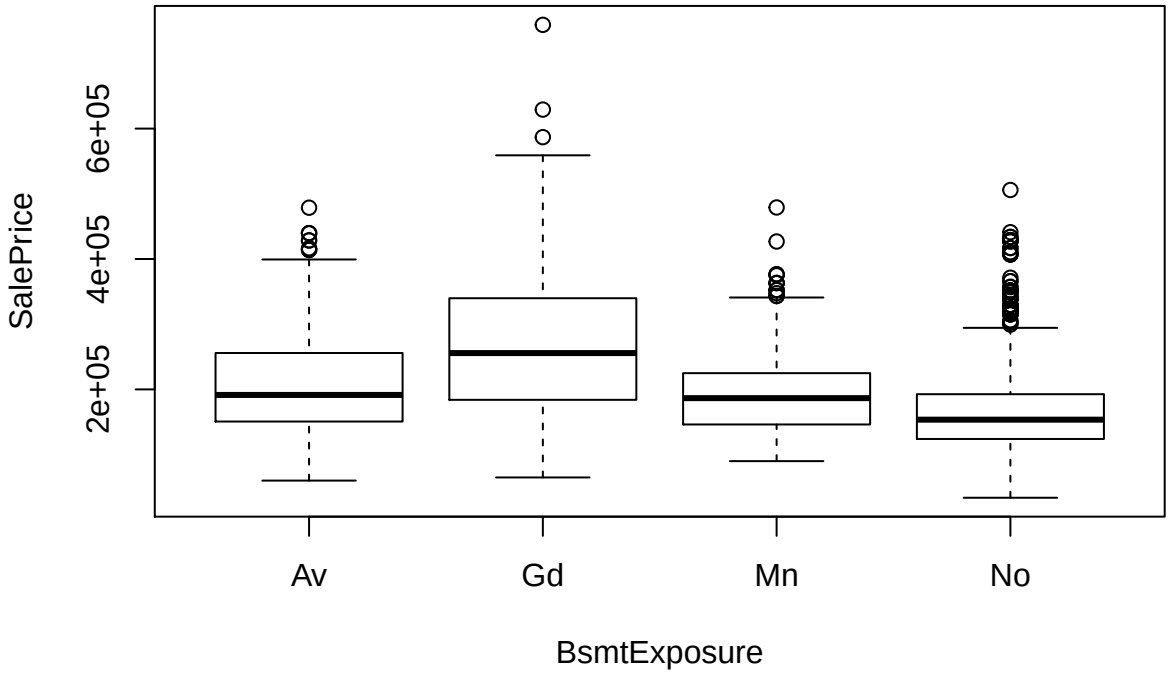

```
boxplot(SalePrice ~ BsmtCond)
```



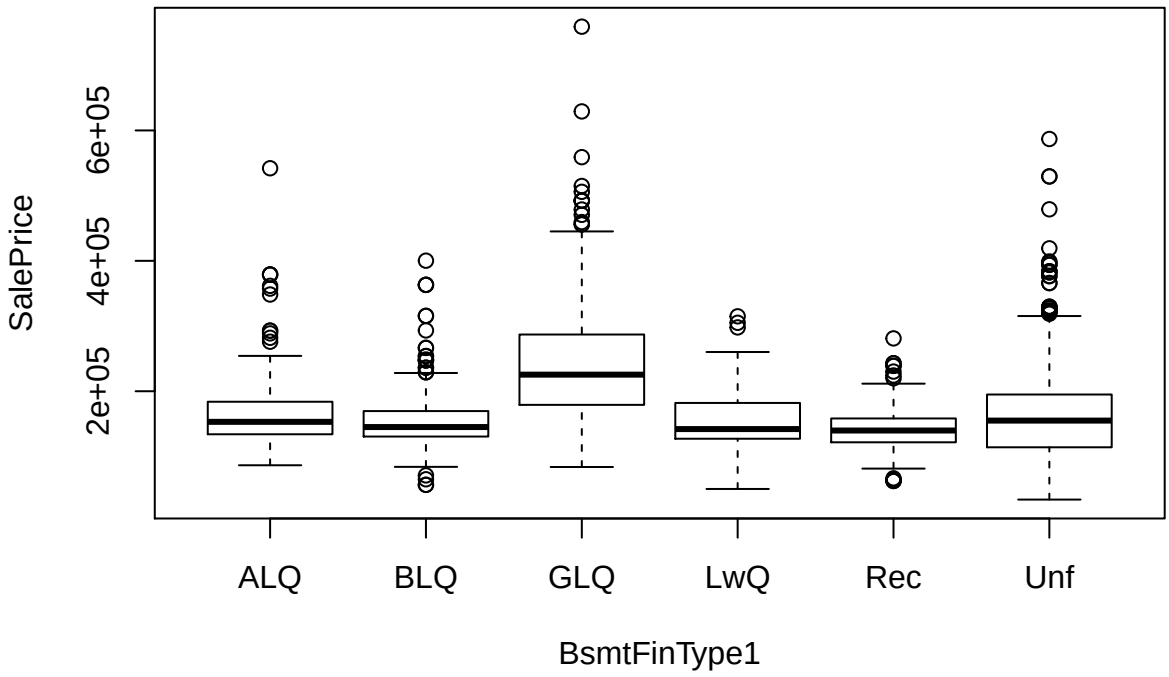

```
boxplot(SalePrice ~ BsmtQual)
```



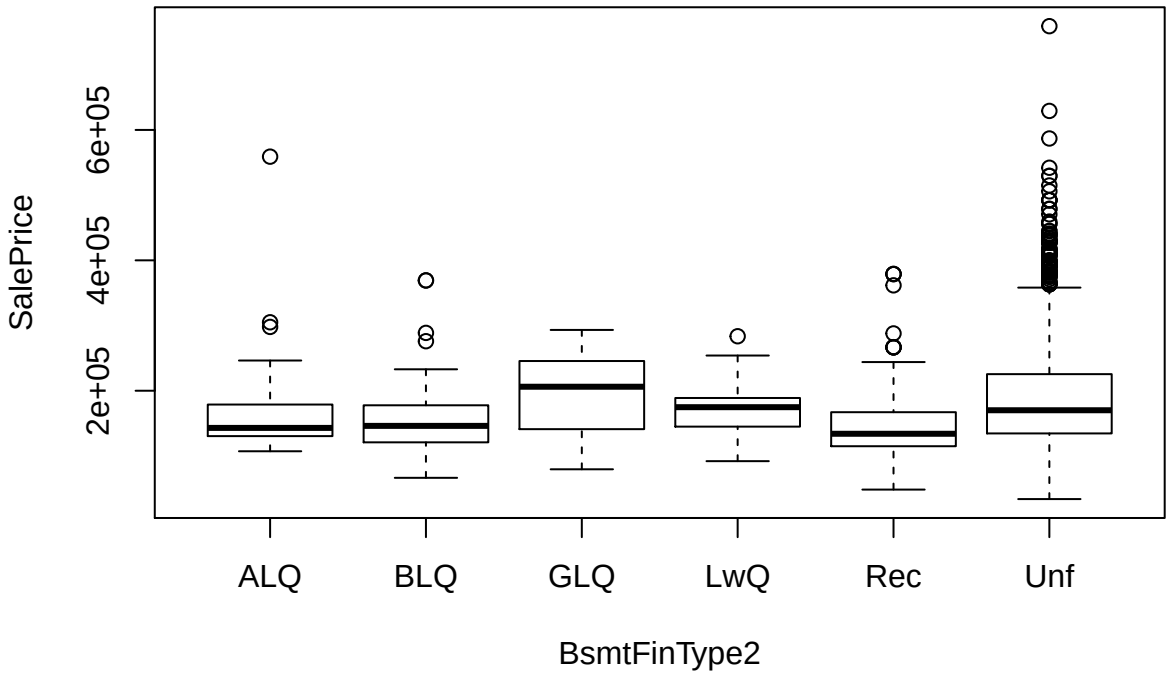

```
boxplot(SalePrice ~ BsmtExposure)
```



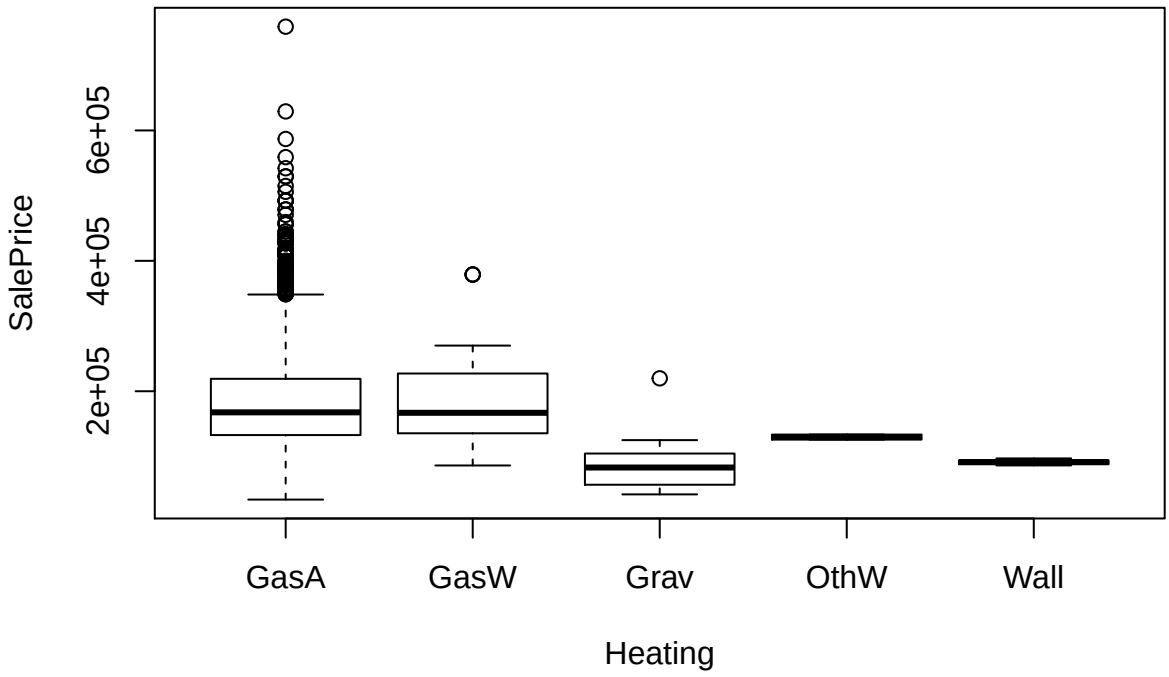

```
boxplot(SalePrice ~ BsmtFinType1)
```



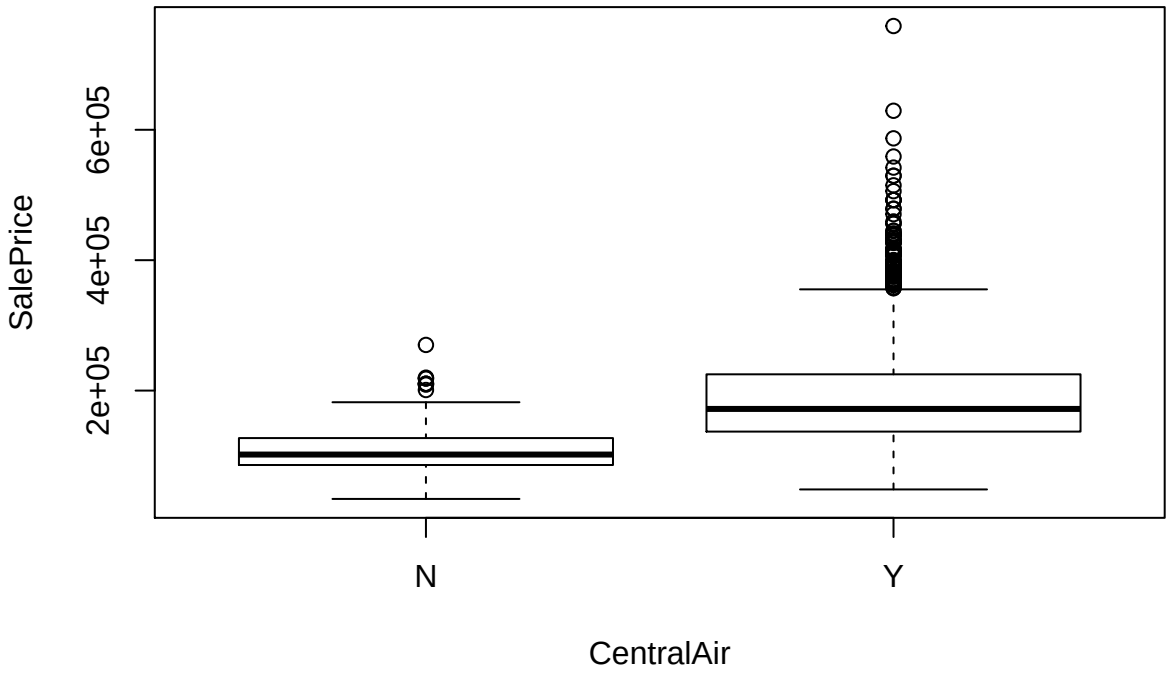

```
boxplot(SalePrice ~ BsmtFinType2)
```



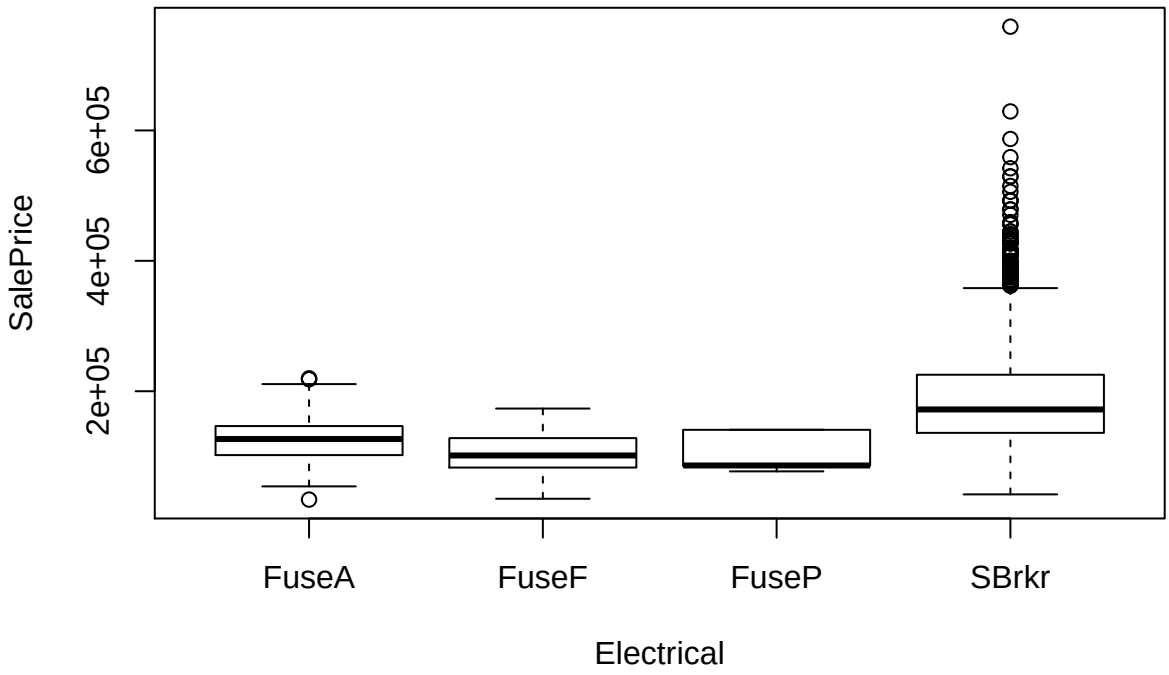

```
boxplot(SalePrice ~ Heating)
```



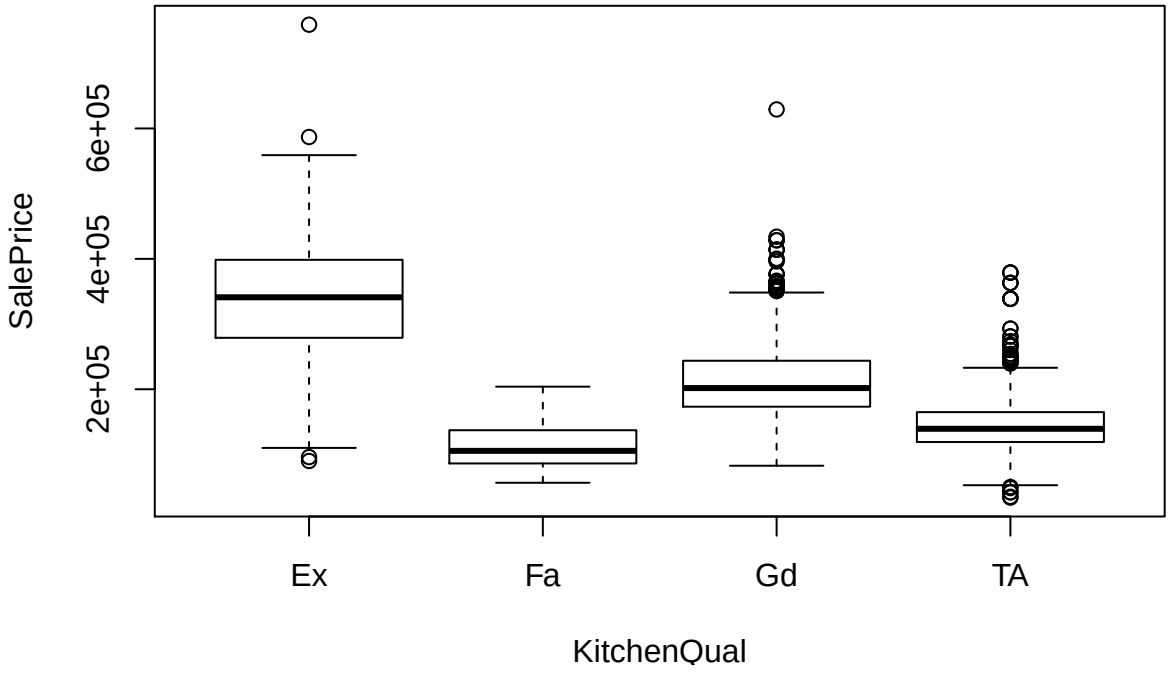

```
boxplot(SalePrice ~ CentralAir)
```



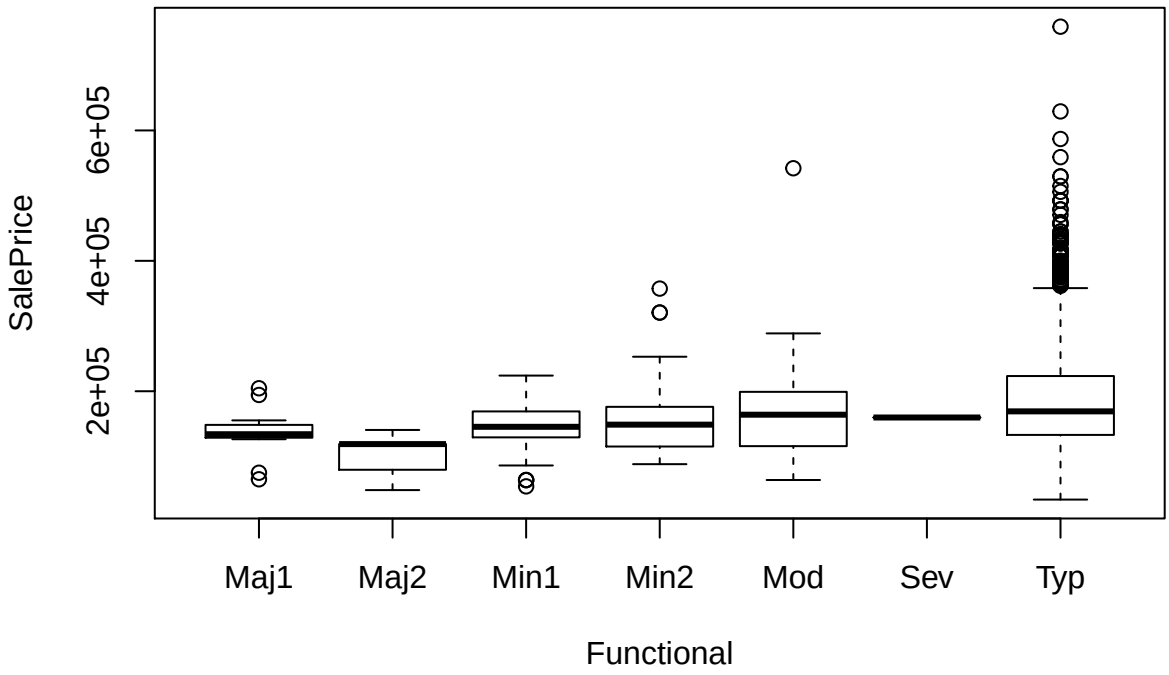

```
boxplot(SalePrice ~ Electrical)
```



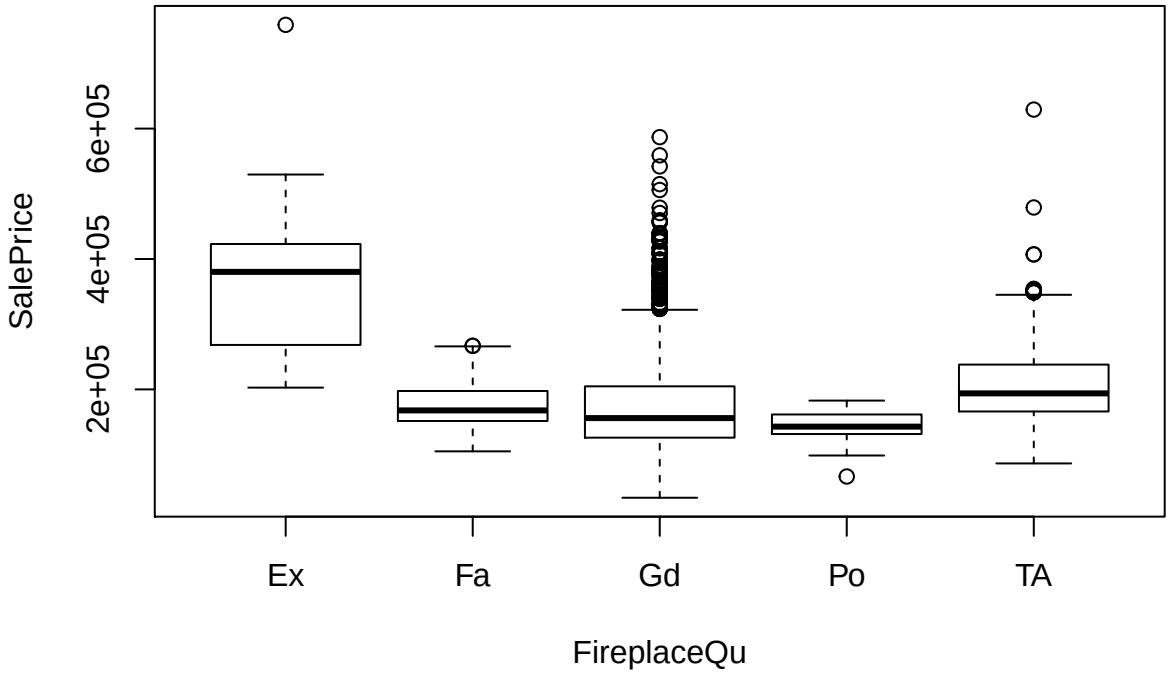

```
boxplot(SalePrice ~ KitchenQual)
```



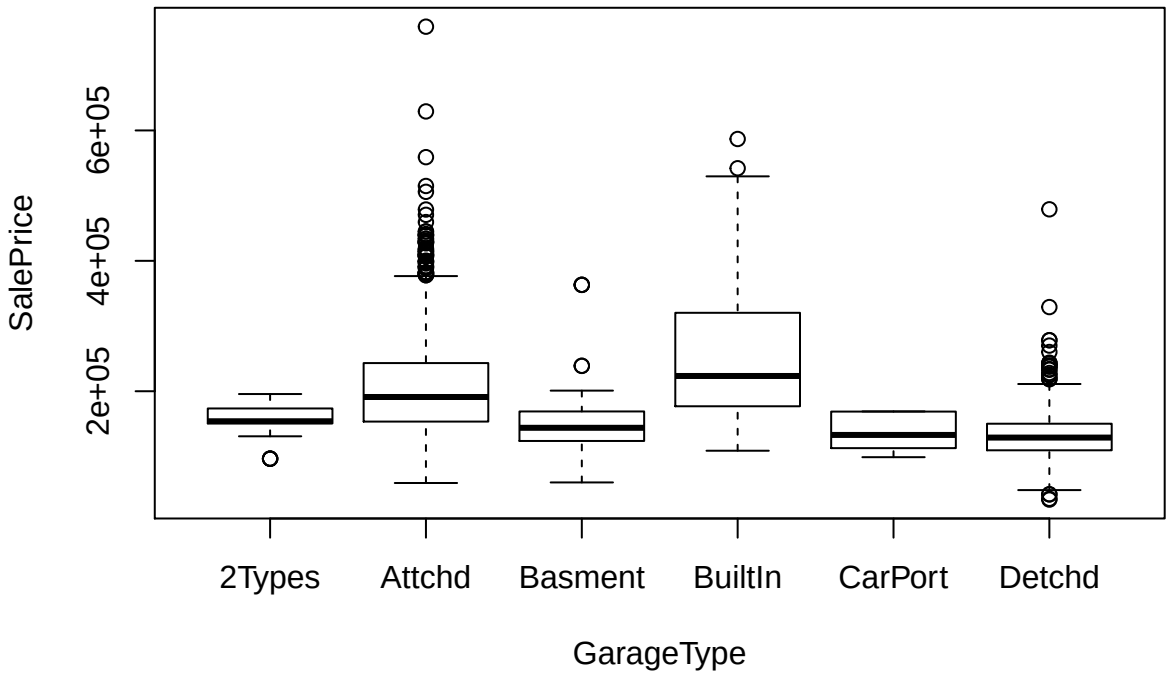

```
boxplot(SalePrice ~ Functional)
```



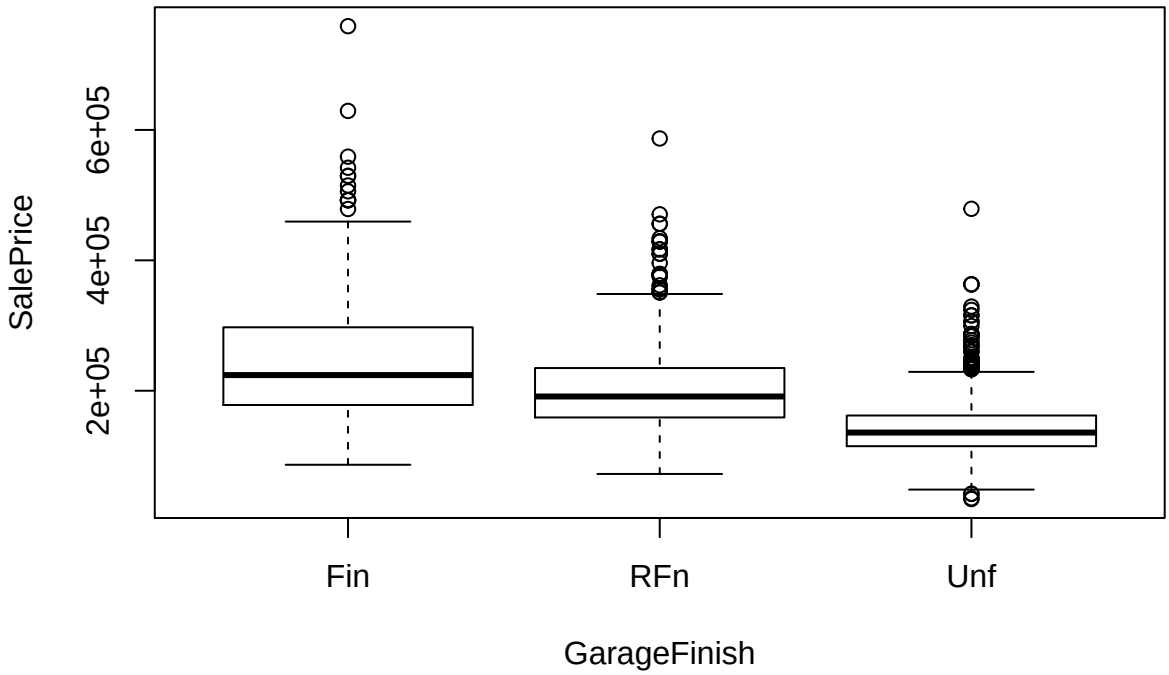

```
boxplot(SalePrice ~ FireplaceQu)
```



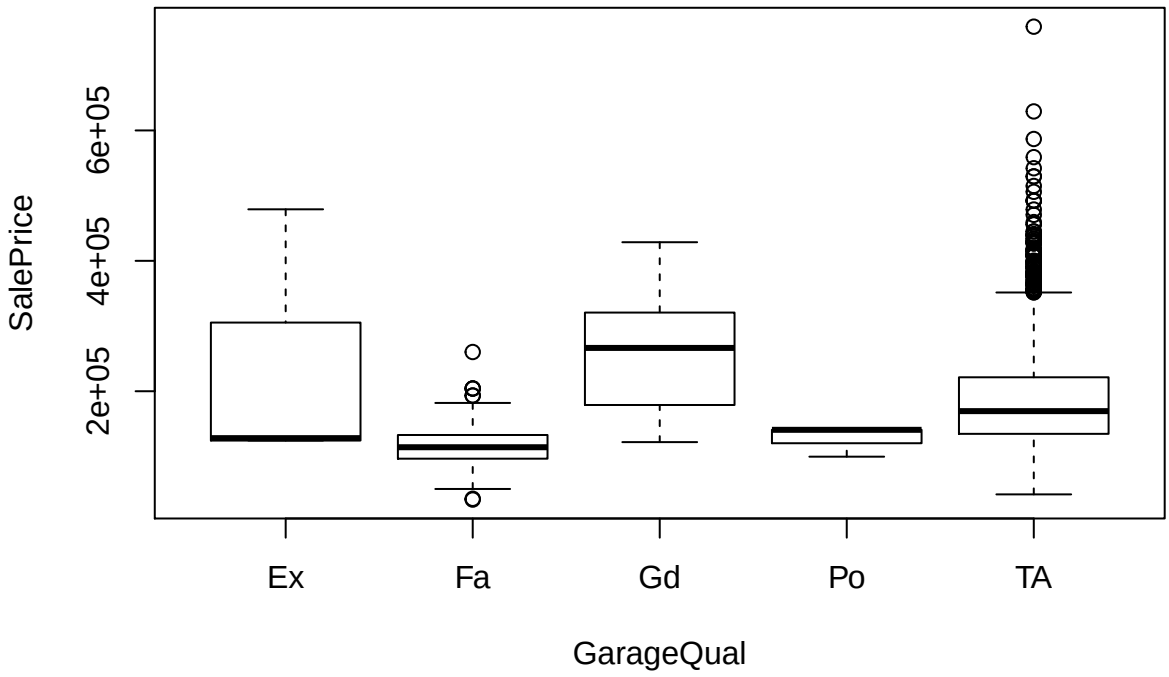

```
boxplot(SalePrice ~ GarageType)
```



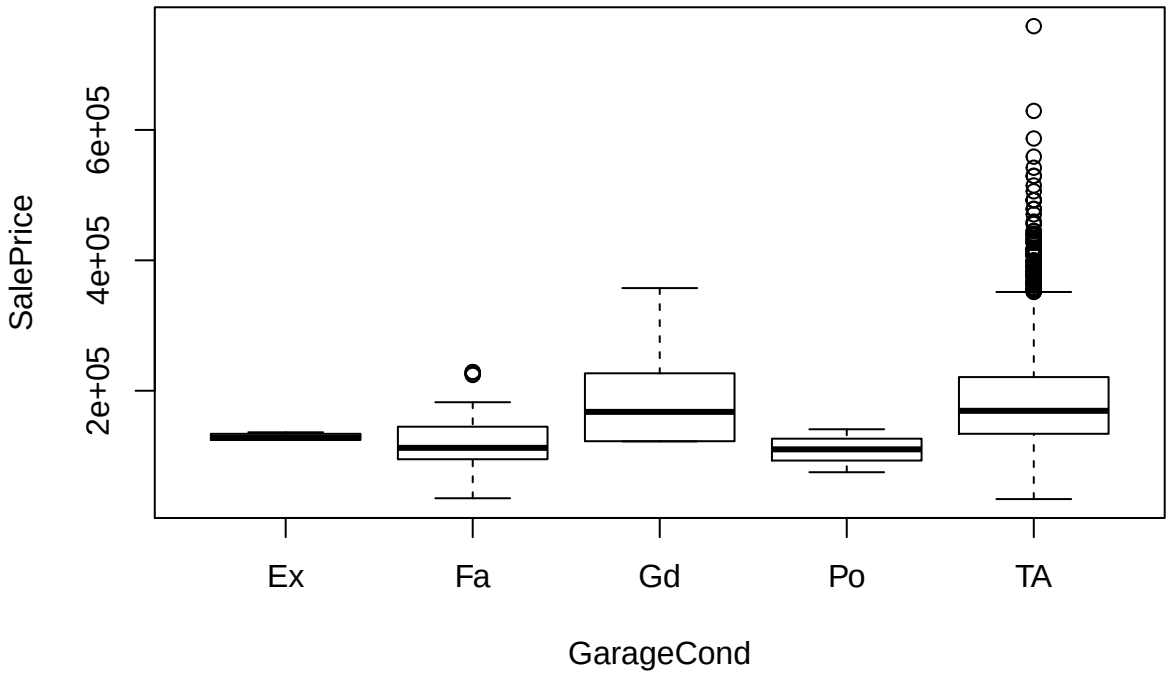

```
boxplot(SalePrice ~ GarageFinish)
```



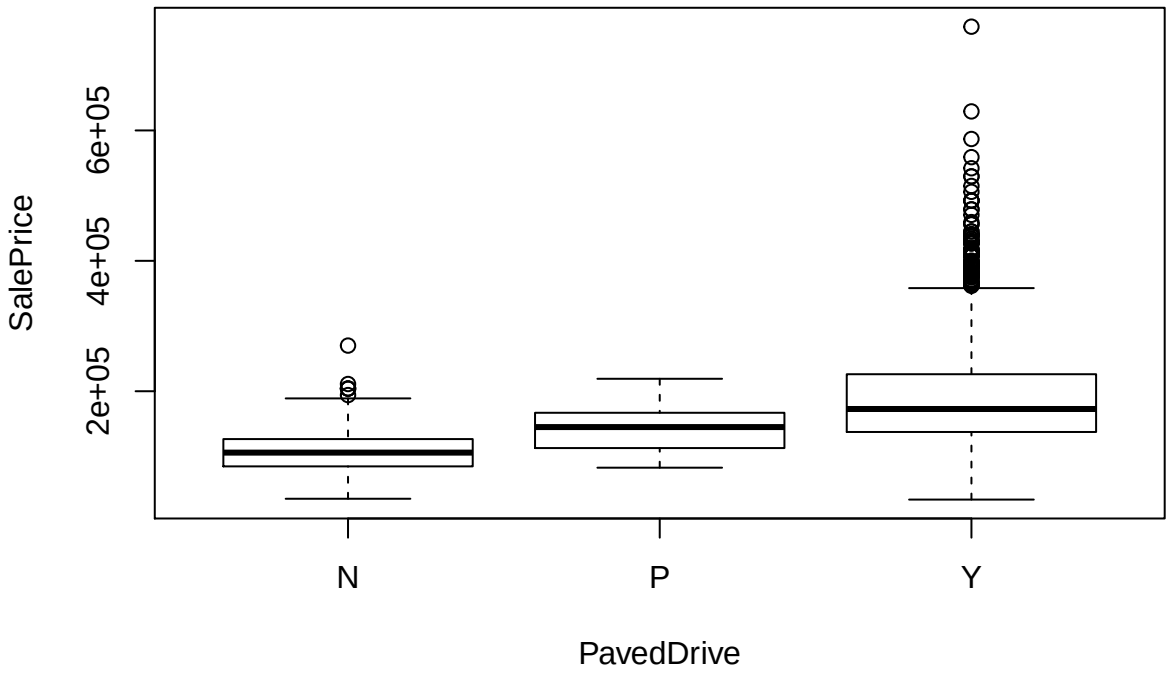

```
boxplot(SalePrice ~ GarageQual)
```



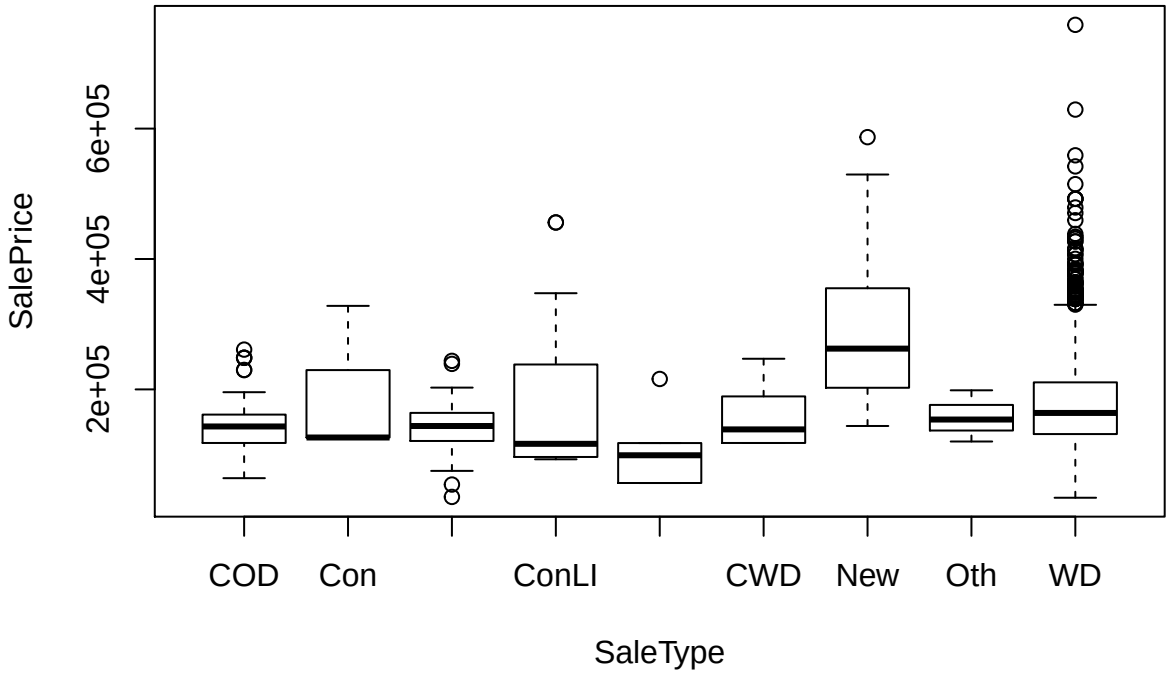

```
boxplot(SalePrice ~ GarageCond)
```




```
boxplot(SalePrice ~ PavedDrive)
```




```
boxplot(SalePrice ~ SaleType)
```



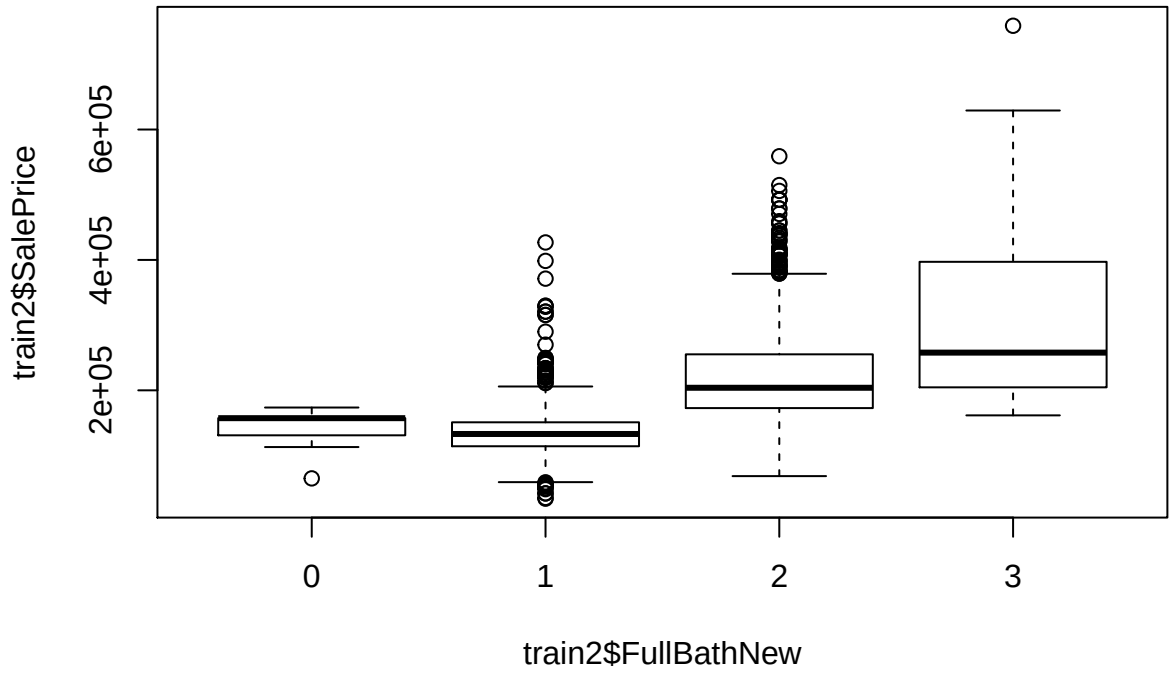

```
#model with numerical and dummy
model2 <- lm(SalePrice ~ LotArea + OverallQual + Age + GrLivArea + FullBath + TotRmsAbvGrd + Fireplaces

#transformation
tSalePrice <- log(SalePrice)
model3 <- lm(tSalePrice ~ LotArea + OverallQual + Age + GrLivArea + FullBath + TotRmsAbvGrd + Fireplace
summary(model3)

##
## Call:
```

```
## lm(formula = tSalePrice ~ LotArea + OverallQual + Age + GrLivArea +
##     FullBath + TotRmsAbvGrd + Fireplaces + GarageArea, data = train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91819 -0.07990 -0.00005  0.09415  0.66238
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.051e+01  1.855e-02 566.495  < 2e-16 ***
## LotArea       1.185e-05  1.069e-06  11.082  < 2e-16 ***
## OverallQual   1.426e-01  3.057e-03  46.652  < 2e-16 ***
## Age          -1.448e-03  1.415e-04 -10.232  < 2e-16 ***
## GrLivArea     1.556e-04  1.238e-05  12.574  < 2e-16 ***
## FullBath      4.883e-02  7.742e-03   6.307 3.36e-10 ***
## TotRmsAbvGrd  1.224e-02  3.241e-03   3.776 0.000163 ***
## Fireplaces    5.900e-02  5.610e-03  10.516  < 2e-16 ***
## GarageArea    3.443e-04  2.007e-05  17.157  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1536 on 2491 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8463
## F-statistic:  1721 on 8 and 2491 DF,  p-value: < 2.2e-16
```

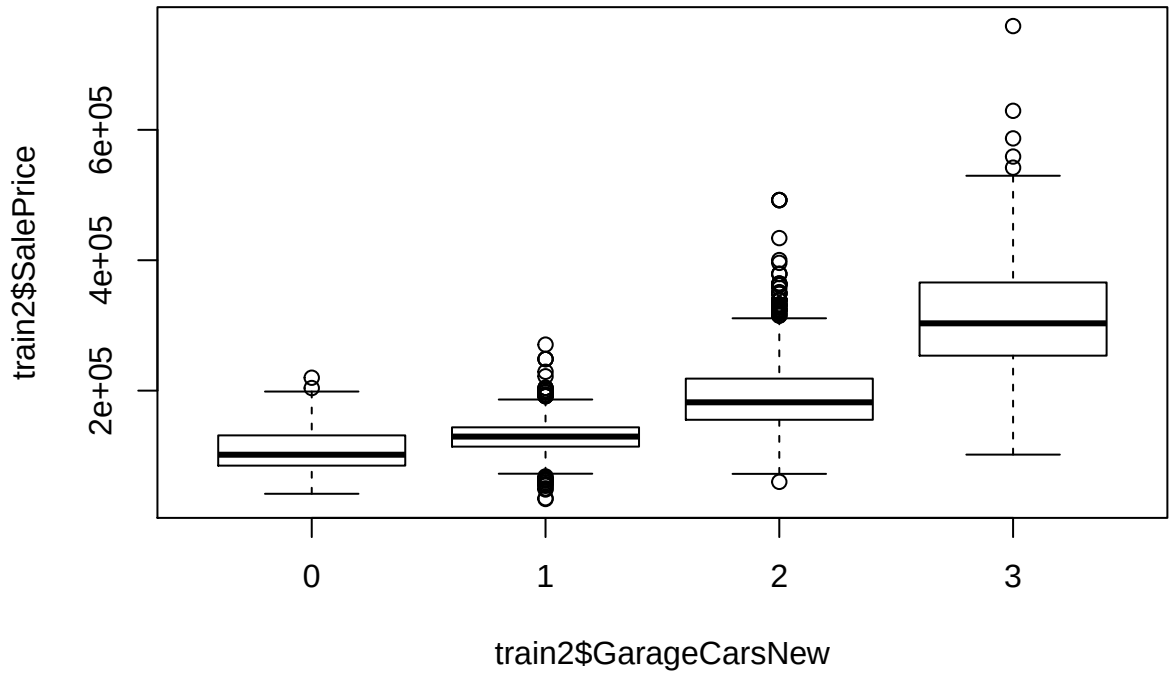
```
#I convert FullBath and GarageCars into dummy
dim(train2)
```

```
## [1] 2500   78
```

```
train2[79] <- data.frame("FullBathNew" = factor(train2$FullBath))
boxplot(train2$SalePrice ~ train2$FullBathNew)
```



```
train2[80] <- data.frame("GarageCarsNew" = factor(train2$GarageCars))
boxplot(train2$SalePrice ~ train2$GarageCarsNew)
```




```
dim(train2)
```

```
## [1] 2500   80
```

```
dim(test2)
```

```
## [1] 1500   76
```

```
test2[77] <- data.frame("FullBathNew" = factor(test2$FullBath))
test2[78] <- data.frame("GarageCarsNew" = factor(test2$GarageCars))
dim(test2)
```

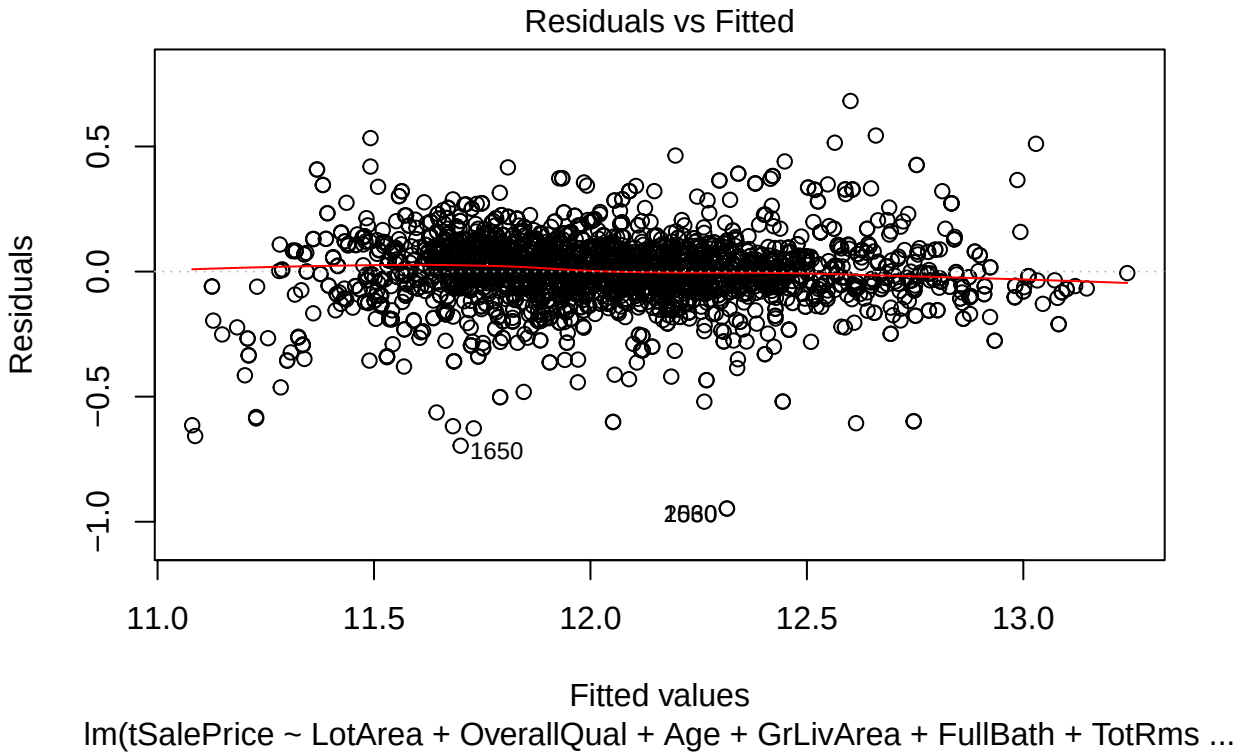
```
## [1] 1500   78
```

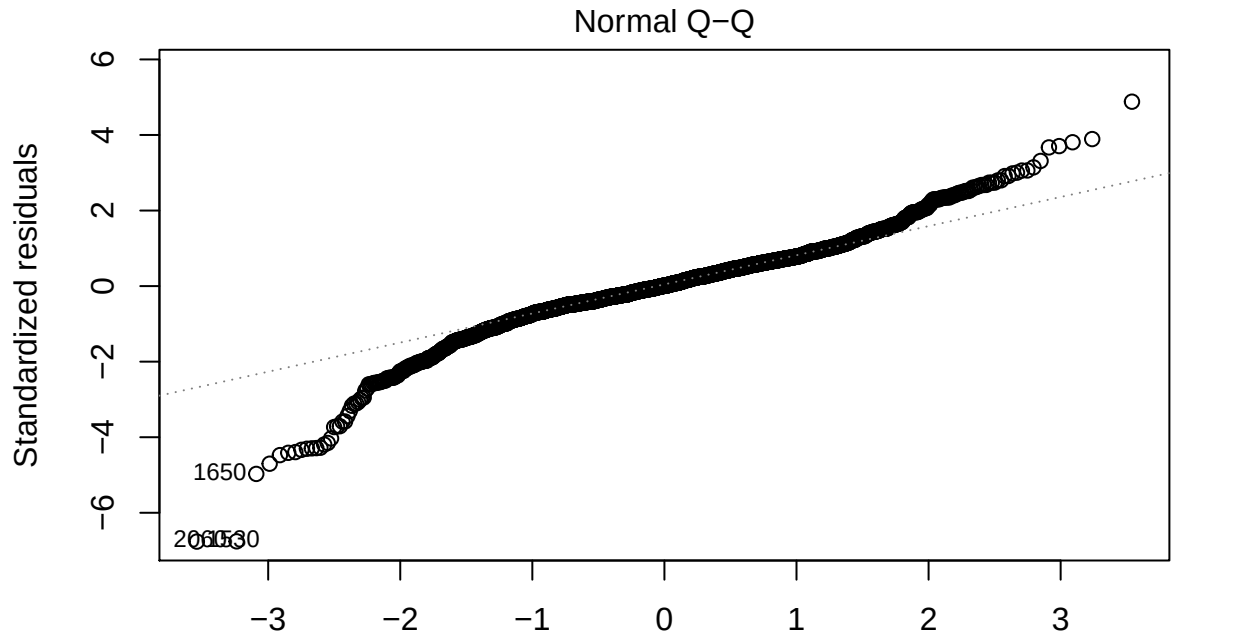
```
#Final model
model4 <- lm(tSalePrice ~ LotArea + OverallQual + Age + GrLivArea + FullBath + TotRmsAbvGrd + Fireplace
summary(model4)
```

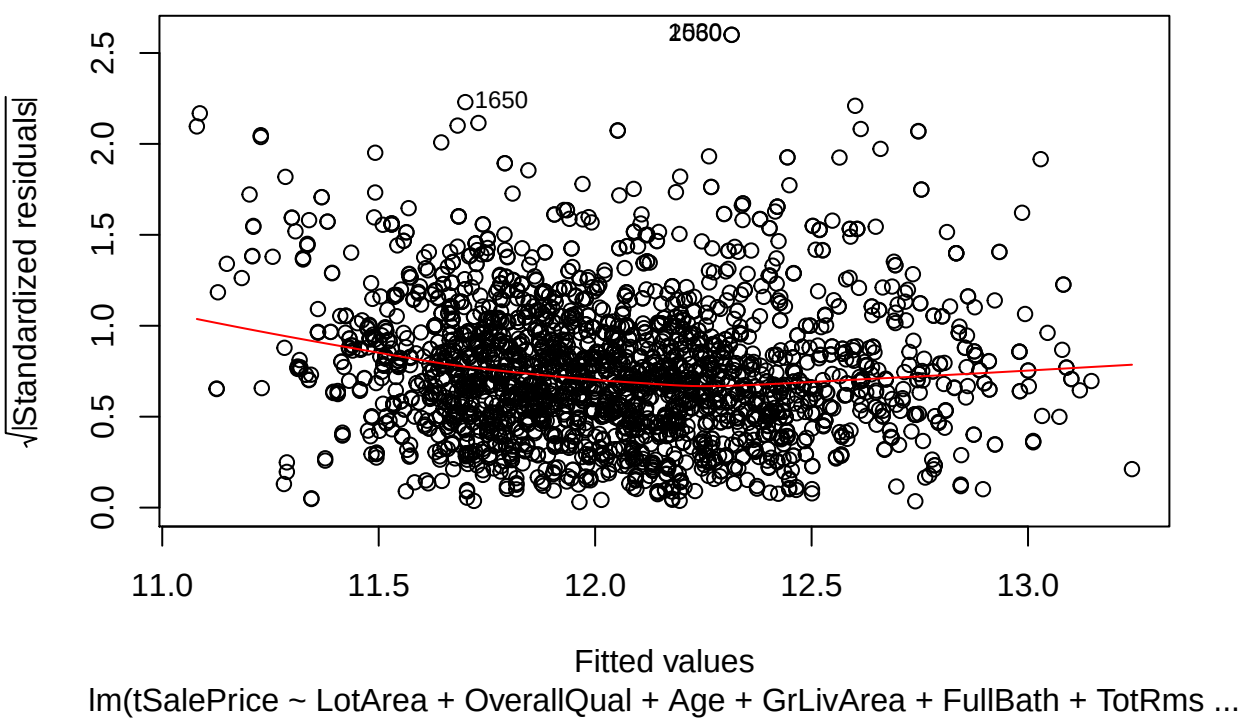
```
##
## Call:
## lm(formula = tSalePrice ~ LotArea + OverallQual + Age + GrLivArea +
##     FullBath + TotRmsAbvGrd + Fireplaces + GarageCars + BsmtFinSF1,
##     data = train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94774 -0.06624  0.00142  0.07934  0.68181
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.050e+01  1.700e-02 617.698  < 2e-16 ***
## LotArea       1.016e-05  9.737e-07  10.435  < 2e-16 ***
```

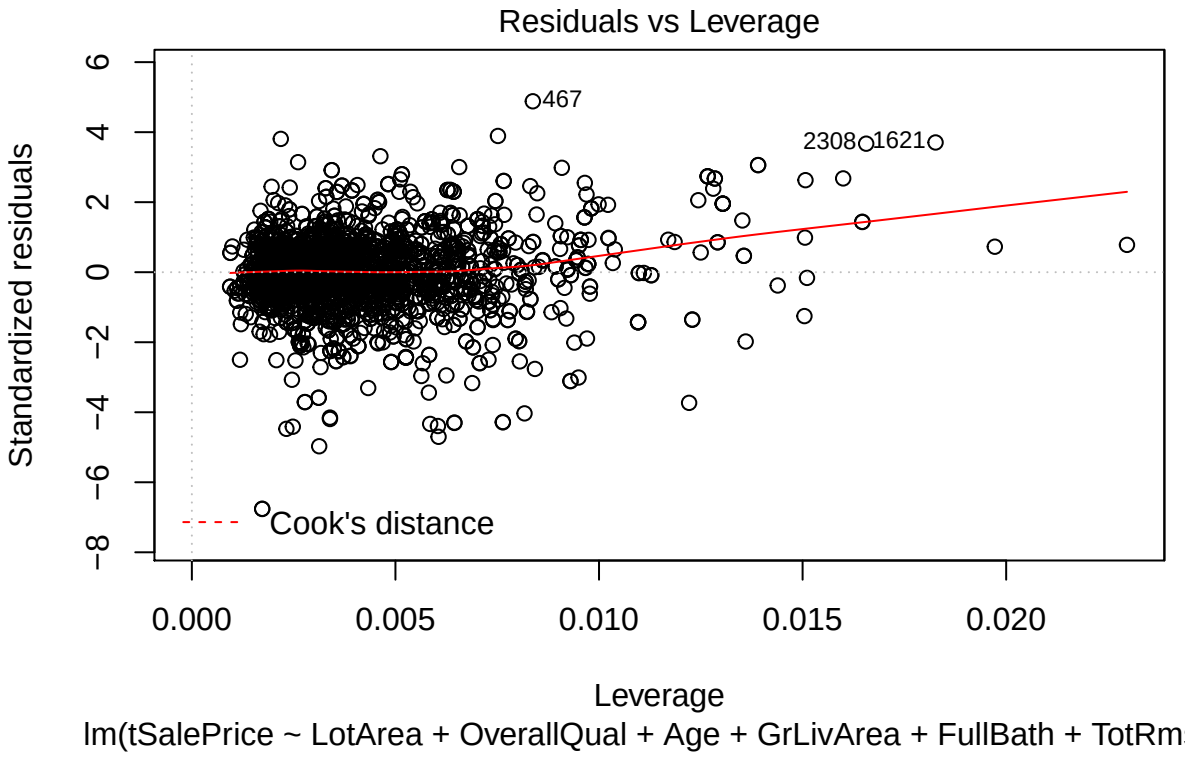
```
## OverallQual   1.341e-01  2.856e-03  46.955  < 2e-16 ***
## Age          -9.926e-04  1.312e-04  -7.564 5.44e-14 ***
## GrLivArea     1.542e-04  1.127e-05  13.681  < 2e-16 ***
## FullBath      5.566e-02  7.147e-03   7.787 9.97e-15 ***
## TotRmsAbvGrd  1.618e-02  2.966e-03   5.454 5.41e-08 ***
## Fireplaces    3.154e-02  5.214e-03   6.050 1.66e-09 ***
## GarageCars    7.841e-02  5.248e-03  14.941  < 2e-16 ***
## BsmtFinSF1    1.685e-04  7.134e-06  23.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1403 on 2490 degrees of freedom
## Multiple R-squared:  0.8723, Adjusted R-squared:  0.8718
## F-statistic:  1890 on 9 and 2490 DF,  p-value: < 2.2e-16
```

```
plot(model4)
```




Normal Q−Q
Standardized residuals
6
4
2
0
−2
−4
−6
1650
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(tSalePrice ~ LotArea + OverallQual + Age + GrLivArea + FullBath + TotRms ...


Scale−Location
√|Standardized residuals|
2.5
2.0
1.5
1.0
0.5
0.0
1650
11.0
11.5
12.0
12.5
13.0
Fitted values
lm(tSalePrice ~ LotArea + OverallQual + Age + GrLivArea + FullBath + TotRms ...




```
prediction = exp(predict(model4, newdata = test2))
prediction[1:10]
```

```
##        1        2        3        4        5        6        7        8
## 171266.1 175534.7 144937.6 142543.8 120112.1 207577.5 105137.8 169698.7
##        9       10
## 101879.8 110559.0
```

```
saleprice_prediction = data.frame(Ob = 1:1500, SalePrice = prediction)
write.csv(saleprice_prediction, file = 'Yuqing_Yang_101A-saleprice-predictions.csv', row.names = FALSE)
```