Multiple Regression

Project Team

1. Tejaswini Nutalapati
2. Aditi Bhargava

Loading dataset

```
Train<-read.csv("/Users/tejaswininutalapati/Documents/Multivariate Analysis/Project/DataSet/train.csv")
Test<-read.csv("/Users/tejaswininutalapati/Documents/Multivariate Analysis/Project/DataSet/test.csv")
```

Loading packages

```
library(ggplot2)
library(outliers)

Test$SalePrice <- -1
#combining both the data sets
df <- rbind(Train,Test)

str(df)
```

```
## 'data.frame':    2919 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5
4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7
420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ..
.
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA N
A NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1
4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4
4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1
1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5
1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1
1 ...
```

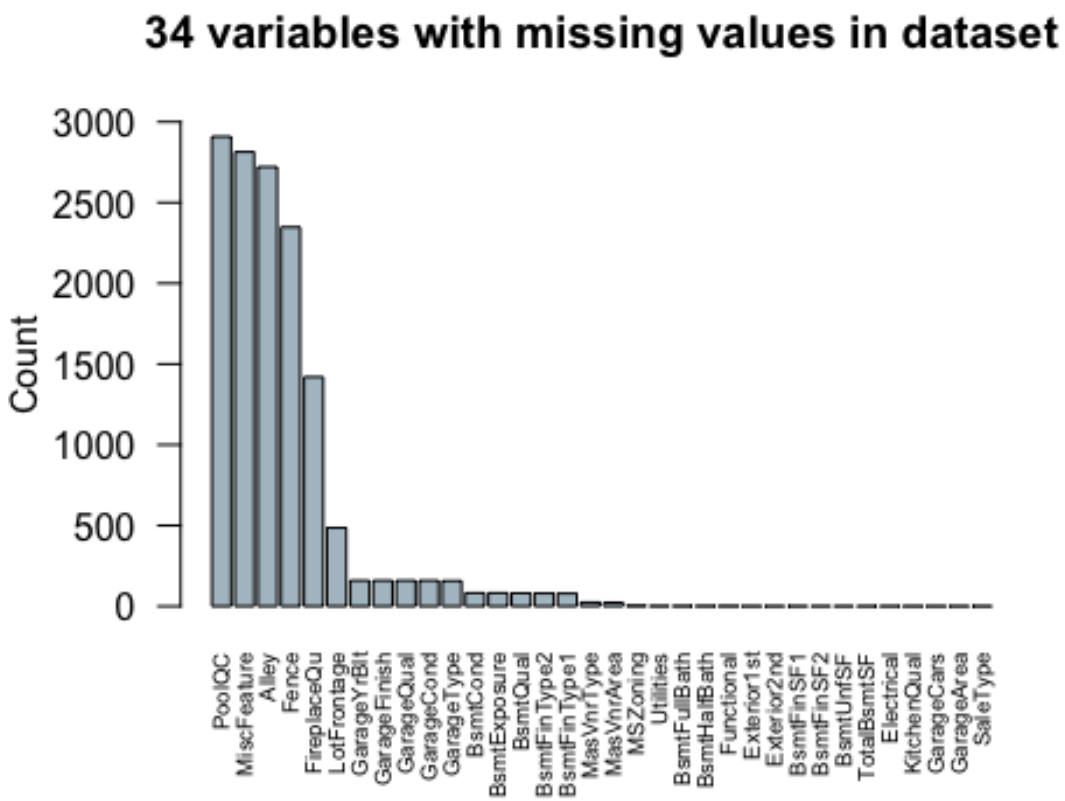
```
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14
12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5
1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3
3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1
2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6
1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 .
..
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 .
..
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2
2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 1
3 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 1
4 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4
4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4
4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5
5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2
1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4
4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4
4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4
4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1
6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2
6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3
```

```
1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2
5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 .
..
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4
4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7
...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5
5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2
6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 .
..
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3
2 ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2
3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5
5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA
NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 N
A NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA
3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
```

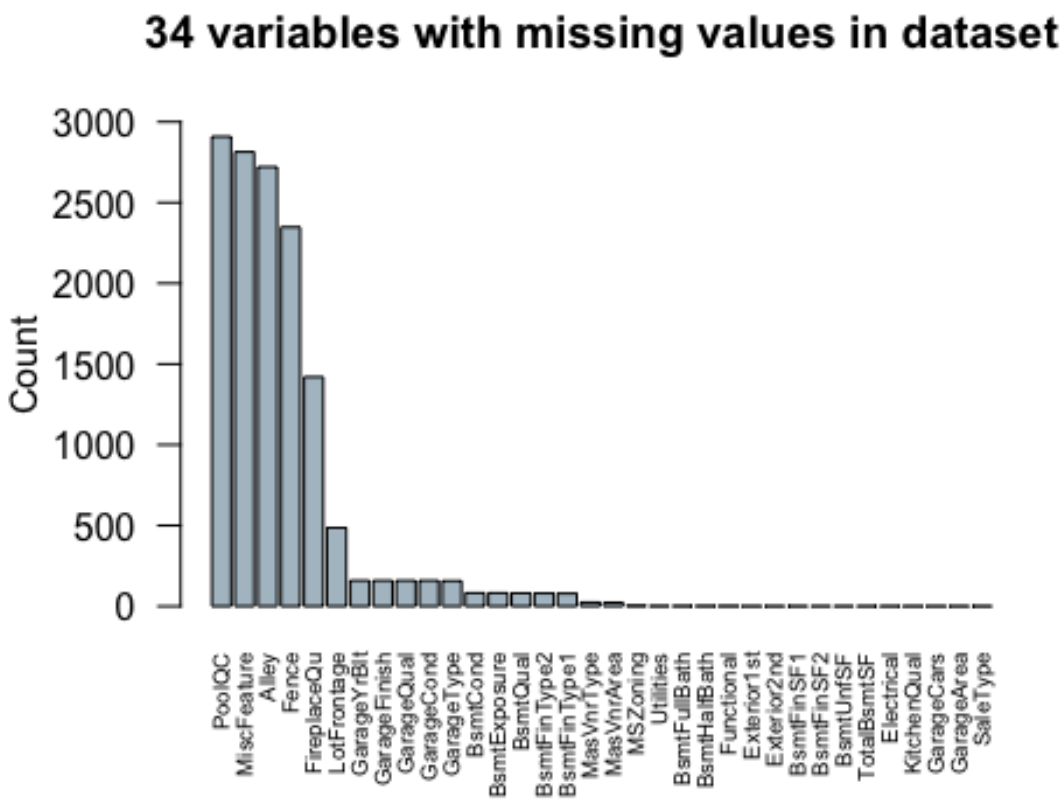
```
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 .
..
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9
9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5
5 1 5 ...
##  $ SalePrice    : num  208500 181500 223500 140000 250000 ...
```

Finding how many variables with missing values are in the dataset

```
options(repr.plot.width=6, repr.plot.height=5)
cMiss = function(x){sum(is.na(x))}
CM <- sort(apply(df,2,cMiss),decreasing=T);
barplot(CM[CM!=0],
        las=2,
        cex.names=0.6,
        ylab="Count",
        ylim=c(0,3000),
        horiz=F,
        col="#AFC0CB",
        main=paste(toString(sum(CM!=0)), "variables with missing values in da
taset"))
```



```r
options(repr.plot.width=6, repr.plot.height=5)
cMiss = function(x){sum(is.na(x))}
CM <- sort(apply(df,2,cMiss),decreasing=T);
barplot(CM[CM!=0],
        las=2,
        cex.names=0.6,
        ylab="Count",
        ylim=c(0,3000),
        horiz=F,
        col="#AFC0CB",
        main=paste(toString(sum(CM!=0)), "variables with missing values in da
taset"))
```

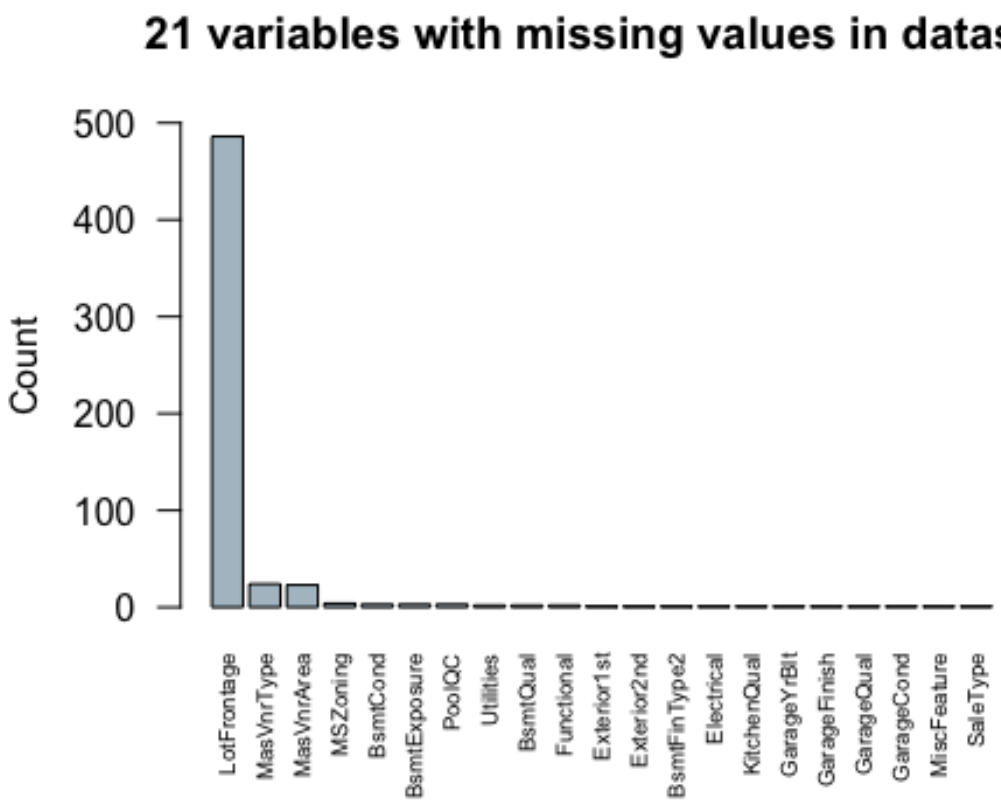



```r
dfClean <-function(df)
{
  # Pool Variable: If PoolQC = NA and PoolArea = 0 , assign factor NoPool
  df$PoolQC <- as.character(df$PoolQC)
  df$PoolQC[df$PoolArea %in% c(0,NA) & is.na(df$PoolQC)] <- "NoPool"
  df$PoolQC <- as.factor(df$PoolQC)
  # MiscFeature Variable: If MiscFeature = NA and MiscVal = 0, assign factor
None
  df$MiscFeature <- as.character(df$MiscFeature)
  df$MiscFeature[df$MiscVal %in% c(0,NA) & is.na(df$MiscFeature)] <- "None"
```

```r
  df$MiscFeature <- as.factor(df$MiscFeature)
  # Alley Variable: If Alley = NA, assign factor NoAccess
  df$Alley <- as.character(df$Alley)
  df$Alley[is.na(df$Alley)] <- "NoAccess"
  df$Alley <- as.factor(df$Alley)
  # Fence Variable: If Fence = NA, assign factor NoFence
  df$Fence <- as.character(df$Fence)
  df$Fence[is.na(df$Fence)] <- "NoFence"
  df$Fence <- as.factor(df$Fence)
  # FireplaceQu Variable: If FireplaceQu = NA and Fireplaces = 0 , assign fac
tor NoFirePlace
  df$FireplaceQu <- as.character(df$FireplaceQu)
  df$FireplaceQu[df$Fireplaces %in% c(0,NA) & is.na(df$FireplaceQu)] <- "NoFi
rePlace"
  df$FireplaceQu <- as.factor(df$FireplaceQu)
  # GarageYrBlt Variable: If GarageYrBlt = NA and GarageArea = 0 assign facto
r NoGarage
  df$GarageYrBlt <- as.character(df$GarageYrBlt)
  df$GarageYrBlt[df$GarageArea %in% c(0,NA) & is.na(df$GarageYrBlt)] <- "NoGa
rage"
  df$GarageYrBlt <- as.factor(df$GarageYrBlt)
  # GarageFinish Variable: If GarageFinish = NA and GarageArea = 0 assign fac
tor NoGarage
  df$GarageFinish <- as.character(df$GarageFinish)
  df$GarageFinish[df$GarageArea %in% c(0,NA) & is.na(df$GarageFinish)] <- "No
Garage"
  df$GarageFinish <- as.factor(df$GarageFinish)
  # GarageQual Variable: If GarageQual = NA and GarageArea = 0 assign factor
NoGarage
  df$GarageQual <- as.character(df$GarageQual)
  df$GarageQual[df$GarageArea %in% c(0,NA) & is.na(df$GarageQual)] <- "NoGara
ge"
  df$GarageQual <- as.factor(df$GarageQual)
  # GarageCond Variable: If GarageCond = NA and GarageArea = 0 assign factor
NoGarage
  df$GarageCond <- as.character(df$GarageCond)
  df$GarageCond[df$GarageArea %in% c(0,NA) & is.na(df$GarageCond)] <- "NoGara
ge"
  df$GarageCond <- as.factor(df$GarageCond)
  # GarageType Variable: If GarageType = NA and GarageArea = 0 assign factor
NoGarage
  df$GarageType <- as.character(df$GarageType)
  df$GarageType[df$GarageArea %in% c(0,NA) & is.na(df$GarageType)] <- "NoGara
ge"
  df$GarageType <- as.factor(df$GarageType)
  df$GarageArea[is.na(df$GarageArea) & df$GarageCars %in% c(0,NA)] <- 0
  df$GarageCars[is.na(df$GarageCars) & df$GarageArea %in% c(0,NA)] <- 0
  # BsmtFullBath Variable: If BsmtFullBath = NA and TotalBsmtSF = 0 assign 0
  df$BsmtFullBath[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtFullBath)] <- 0
  # BsmtHalfBath Variable: If BsmtHalfBath = NA and TotalBsmtSF = 0 assign 0
```

```
  df$BsmtHalfBath[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtHalfBath)] <- 0
  # BsmtFinSF1 Variable: If BsmtFinSF1 = NA and TotalBsmtSF = 0 assign 0
  df$BsmtFinSF1[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtFinSF1)] <- 0
  # BsmtFinSF2 Variable: If BsmtFinSF2 = NA and TotalBsmtSF = 0 assign 0
  df$BsmtFinSF2[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtFinSF2)] <- 0
  # BsmtUnfSF Variable: If BsmtUnfSF = NA and TotalBsmtSF = 0 assign 0
  df$BsmtUnfSF[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtUnfSF)] <- 0
  # TotalBsmtSF Variable: If TotalBsmtSF = NA and TotalBsmtSF = 0 assign 0
  df$TotalBsmtSF[df$TotalBsmtSF %in% c(0,NA) & is.na(df$TotalBsmtSF)] <- 0
  # BsmtQual Variable: If BsmtQual = NA and TotalBsmtSF = 0 assign factor NoB
asement
  df$BsmtQual <- as.character(df$BsmtQual)
  df$BsmtQual[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtQual)] <- "NoBasemen
t"
  df$BsmtQual <- as.factor(df$BsmtQual)
  # BsmtFinType1 Variable: If BsmtFinType1 = NA and TotalBsmtSF = 0 assign fa
ctor NoBasement
  df$BsmtFinType1 <- as.character(df$BsmtFinType1)
  df$BsmtFinType1[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtFinType1)] <- "N
oBasement"
  df$BsmtFinType1 <- as.factor(df$BsmtFinType1)
  # BsmtFinType2 Variable: If BsmtFinType2 = NA and TotalBsmtSF = 0 assign fa
ctor NoBasement
  df$BsmtFinType2 <- as.character(df$BsmtFinType2)
  df$BsmtFinType2[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtFinType2)] <- "N
oBasement"
  df$BsmtFinType2 <- as.factor(df$BsmtFinType2)
  # BsmtExposure Variable: If BsmtExposure = NA and TotalBsmtSF = 0 assign fa
ctor NoBasement
  df$BsmtExposure <- as.character(df$BsmtExposure)
  df$BsmtExposure[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtExposure)] <- "N
oBasement"
  df$BsmtExposure <- as.factor(df$BsmtExposure)
  # BsmtCond Variable: If BsmtCond = NA and TotalBsmtSF = 0 assign factor NoB
asement
  df$BsmtCond <- as.character(df$BsmtCond)
  df$BsmtCond[df$TotalBsmtSF %in% c(0,NA) & is.na(df$BsmtCond)] <- "NoBasemen
t"
  df$BsmtCond <- as.factor(df$BsmtCond)
  return(df)
}

df <- dfClean(df)
PM <- sort(apply(df,2,cMiss),decreasing=T);
barplot(PM[PM!=0],
        las=2,
        cex.names=0.6,
        ylab="Count",
        ylim=c(0,500),
        horiz=F,
```

```
        col="#AFC0CB",
        main=paste(toString(sum(PM!=0)), "variables with missing values in da
taset"))
```




```
#That certainly helped a little bit. Let's see if there's a pattern to the re
maining missing data.
library(VIM);
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

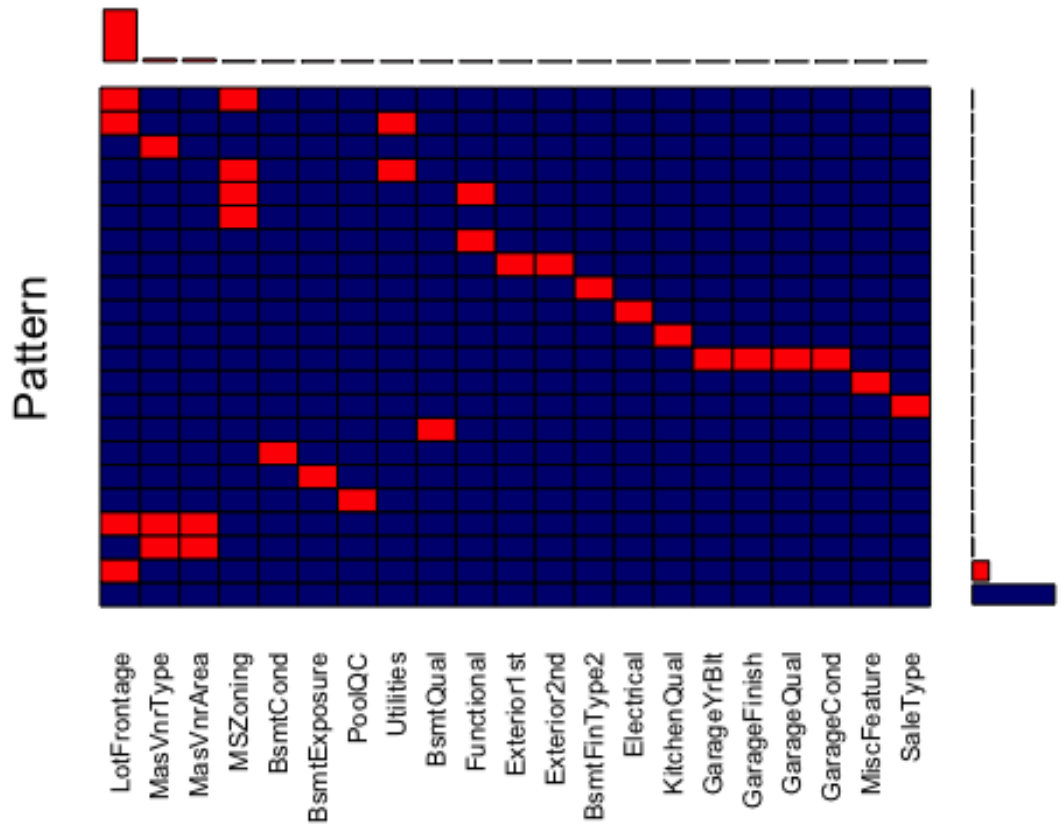
```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statis
tikat/VIM/issues
```

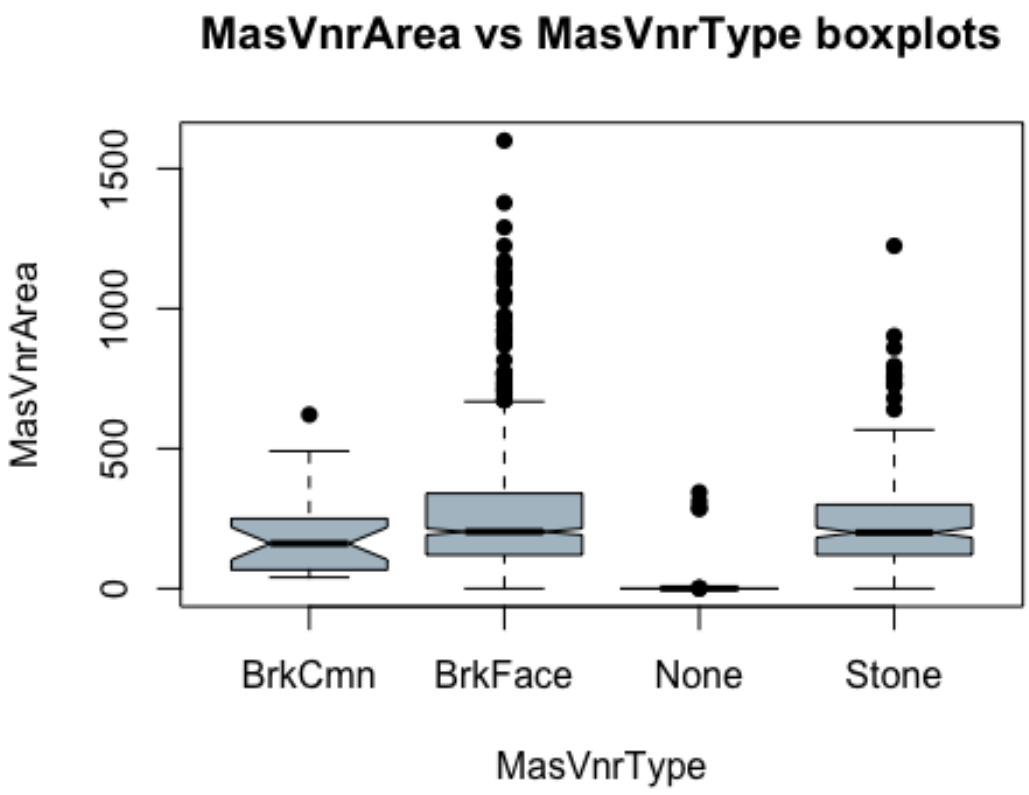
```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
data = df[, names(PM[PM!=0])];
aggr_plot <- aggr(data,
                  col=c('navyblue','red'),
                  bars=T,
                  numbers=T,
                  combined = T,
                  labels=names(data),
                  cex.axis=.7,
                  gap=3,
                  ylab=c("Pattern"),
                  cex.numbers=0.74)

## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
#MasVnrType and MasVnrArea
plot(df[,c("MasVnrType","MasVnrArea")],
     pch=16,
     notch=TRUE,
     main="MasVnrArea vs MasVnrType boxplots",
     col="#AFC0CB")
```




```
df[ (is.na(df$MasVnrType) | is.na(df$MasVnrArea)) ,c("MasVnrType","MasVnrArea
")]
```

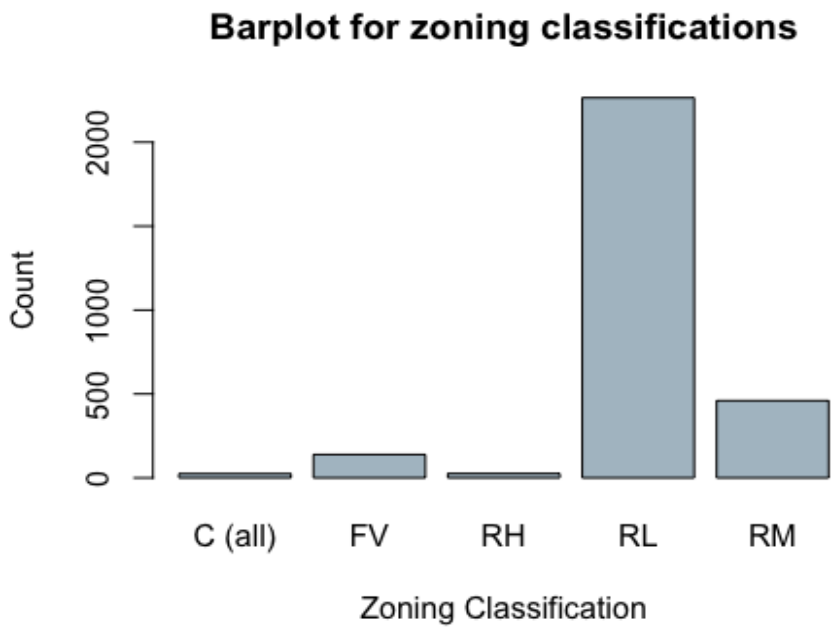
```
##      MasVnrType MasVnrArea
## 235        <NA>         NA
## 530        <NA>         NA
## 651        <NA>         NA
## 937        <NA>         NA
## 974        <NA>         NA
## 978        <NA>         NA
## 1244       <NA>         NA
## 1279       <NA>         NA
## 1692       <NA>         NA
## 1707       <NA>         NA
## 1883       <NA>         NA
## 1993       <NA>         NA
## 2005       <NA>         NA
## 2042       <NA>         NA
## 2312       <NA>         NA
## 2326       <NA>         NA
## 2341       <NA>         NA
## 2350       <NA>         NA
## 2369       <NA>         NA
```

```
## 2593       <NA>          NA
## 2611       <NA>         198
## 2658       <NA>          NA
## 2687       <NA>          NA
## 2863       <NA>          NA
```

```
summary(df[ !(is.na(df$MasVnrType) | is.na(df$MasVnrArea)) ,c("MasVnrType","M
asVnrArea")])
```

```
##    MasVnrType     MasVnrArea
##  BrkCmn :  25   Min.   :   0.0
##  BrkFace: 879   1st Qu.:   0.0
##  None   :1742   Median :   0.0
##  Stone  : 249   Mean   : 102.2
##                 3rd Qu.: 164.0
##                 Max.   :1600.0
```

```
df$MasVnrType <- as.character(df$MasVnrType)
df$MasVnrType[is.na(df$MasVnrType)] <- "None"
df$MasVnrType <- as.factor(df$MasVnrType)
df$MasVnrArea[is.na(df$MasVnrArea)] <- 0

#MSZoning
plot(df$MSZoning,
     col="#AFC0CB",
     xlab="Zoning Classification",
     ylab = "Count",
     main = "Barplot for zoning classifications")
```



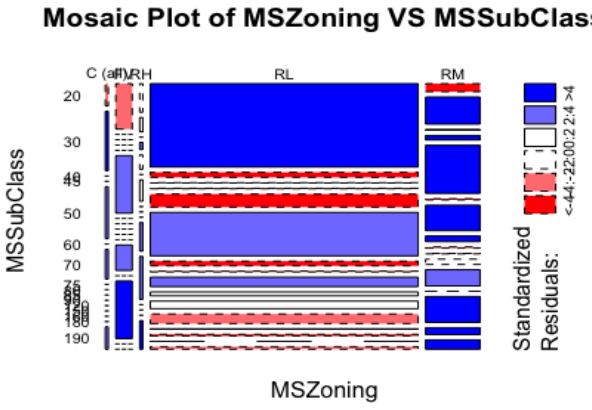
```r
df[ is.na(df$MSZoning) ,c("MSZoning","MSSubClass")]
```

```
##      MSZoning MSSubClass
## 1916     <NA>         30
## 2217     <NA>         20
## 2251     <NA>         70
## 2905     <NA>         20
```

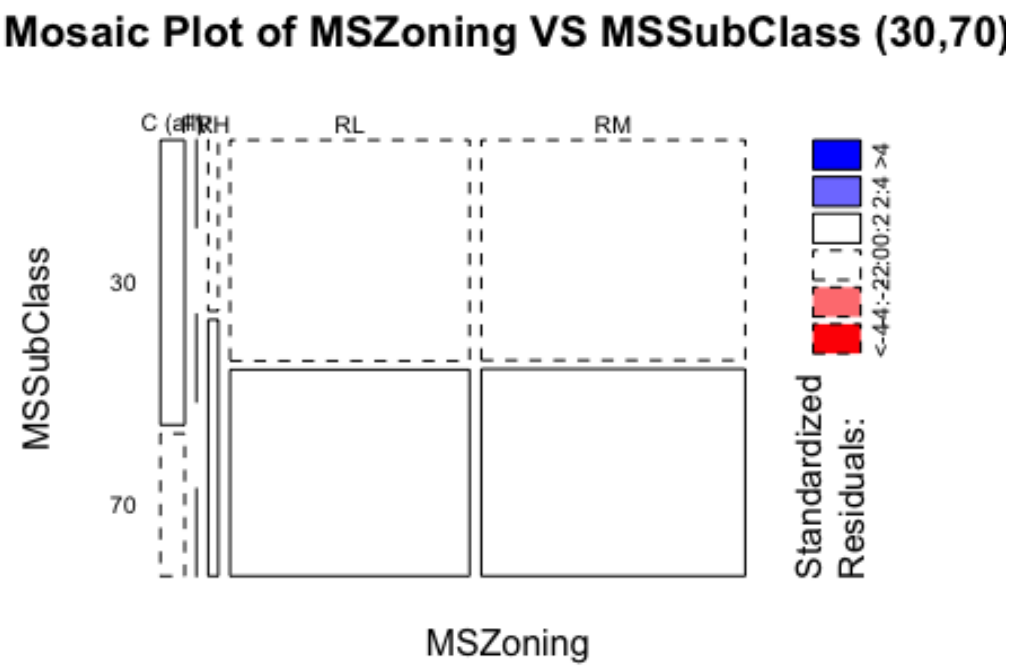
```r
ZoneClassTable <- table(df[ ,c("MSZoning","MSSubClass")])
ZoneClassTable
```

```
##           MSSubClass
## MSZoning    20   30   40   45   50   60   70   75   80   85   90  120  150
160
##   C (all)    3    8    0    0    7    0    4    0    0    0    0    0    0
0
##   FV        34    0    0    0    0   43    0    0    0    0    0   19    0
43
##   RH         4    2    0    1    2    0    3    0    0    0    4    6    0
0
##   RL      1016   61    4    6  159  529   57    9  115   47   92  117    1
21
##   RM        20   67    2   11  119    3   63   14    3    1   13   40    0
64
##           MSSubClass
## MSZoning   180  190
##   C (all)    0    3
##   FV         0    0
##   RH         0    4
##   RL         0   31
##   RM        17   23
```

```r
mosaicplot(ZoneClassTable,
           main="Mosaic Plot of MSZoning VS MSSubClass",
           las=1,
           color=T,
           shade=T)
```



```r
library(DescTools)
GTest(ZoneClassTable)

##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  ZoneClassTable
## G = 1321.9, X-squared df = 60, p-value < 2.2e-16

Table<-table(df[ df$MSSubClass %in% c(30,70) ,c("MSZoning","MSSubClass")])
Table <- Table[ , colSums(Table != 0) > 0 ]
Table

##           MSSubClass
## MSZoning  30 70
##   C (all)  8  4
##   FV       0  0
##   RH       2  3
##   RL      61 57
##   RM      67 63

mosaicplot(Table,
           main="Mosaic Plot of MSZoning VS MSSubClass (30,70)",
           las=1,
           color=T,
           shade=T)
```

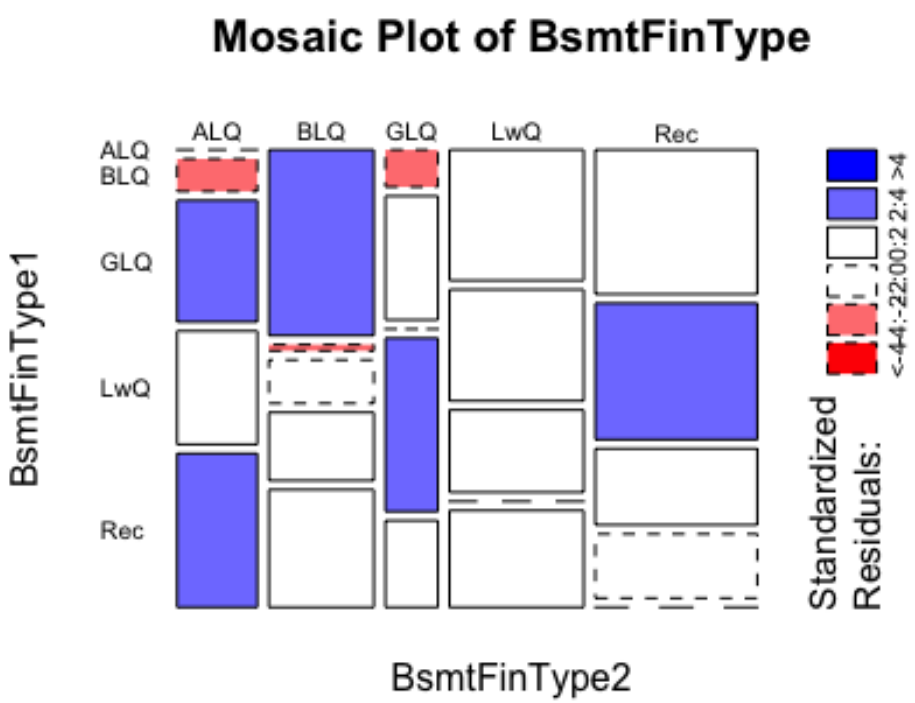


```
Test1<-GTest(Table)
Test1

##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  Table
## G = 1.3625, X-squared df = 4, p-value = 0.8507

paste("At a 95% confidence level, since the p-value =", as.character(round(Te
st1$p.value,2)),
      "> 0.05, we cannot reject the null hypothesis that MSZoning and MSSubCl
ass are independent when MSSubClass = 30 or 70.")

## [1] "At a 95% confidence level, since the p-value = 0.85 > 0.05, we cannot
reject the null hypothesis that MSZoning and MSSubClass are independent when
MSSubClass = 30 or 70."

df$MSZoning <- as.character(df$MSZoning)
df$MSZoning[is.na(df$MSZoning)] <- "RL"
df$MSZoning <- as.factor(df$MSZoning)

#Basement
MissBsmt = c('BsmtCond','BsmtExposure','BsmtQual','BsmtFinType2')
df[!complete.cases(df[,names(df) %in% MissBsmt]),names(df) %in% names(df)[whi
ch(grepl("Bsmt",names(df)))]]

##      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 333        Gd       TA           No          GLQ       1124         <NA>
## 949        Gd       TA         <NA>          Unf          0          Unf
## 1488       Gd       TA         <NA>          Unf          0          Unf
## 2041       Gd     <NA>           Mn          GLQ       1044          Rec
## 2186       TA     <NA>           No          BLQ       1033          Unf
## 2218     <NA>       Fa           No          Unf          0          Unf
## 2219     <NA>       TA           No          Unf          0          Unf
## 2349       Gd       TA         <NA>          Unf          0          Unf
## 2525       TA     <NA>           Av          ALQ        755          Unf
##      BsmtFinSF2 BsmtUnfSF TotalBsmtSF BsmtFullBath BsmtHalfBath
## 333         479      1603        3206            1            0
## 949           0       936         936            0            0
## 1488          0      1595        1595            0            0
## 2041        382         0        1426            1            0
## 2186          0        94        1127            0            1
## 2218          0       173         173            0            0
## 2219          0       356         356            0            0
## 2349          0       725         725            0            0
## 2525          0       240         995            0            0

#BsmtExposure
df$BsmtExposure <- as.character(df$BsmtExposure)
df$BsmtExposure[is.na(df$BsmtExposure)]<-"No"
```

```r
df$BsmtExposure <- as.factor(df$BsmtExposure)

#BsmtFinType2
BsmtFinQuality<-table(df[ !(df$BsmtFinType2 %in% c("NoBasement","Unf") | df$B
smtFinType1 %in% c("NoBasement","Unf")) ,c("BsmtFinType2","BsmtFinType1")])
BsmtFinQuality<-BsmtFinQuality[rowSums(BsmtFinQuality != 0) > 0 , colSums(Bsm
tFinQuality != 0) > 0]
BsmtFinQuality
```

```
##              BsmtFinType1
## BsmtFinType2 ALQ BLQ GLQ LwQ Rec
##          ALQ   0   4  15  14  19
##          BLQ  30   1   7  11  19
##          GLQ   3  10   0  14   7
##          LwQ  27  23  17   0  20
##          Rec  36  34  19  16   0
```

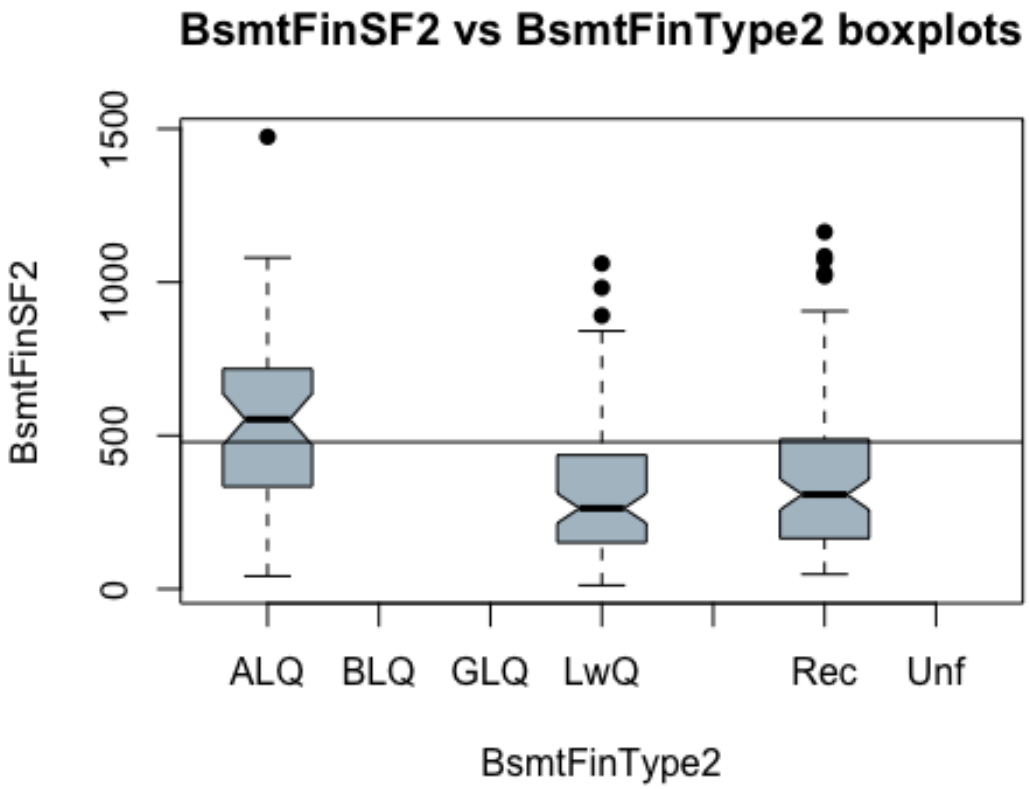
```r
mosaicplot(BsmtFinQuality,
           main="Mosaic Plot of BsmtFinType",
           las=1,
           color=T,
           shade=T)
```




```r
TestQ<-GTest(BsmtFinQuality)
TestQ
```

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  BsmtFinQuality
## G = 184.71, X-squared df = 16, p-value < 2.2e-16
```

```
plot(df[df$BsmtFinType2 %in% c("ALQ","LwQ", "Rec"),c("BsmtFinType2","BsmtFinS
F2")],
     pch=16,
     notch=TRUE,
     main="BsmtFinSF2 vs BsmtFinType2 boxplots",
     col="#AFC0CB")
abline(h=df[is.na(df$BsmtFinType2) ,c("BsmtFinSF2")])
```

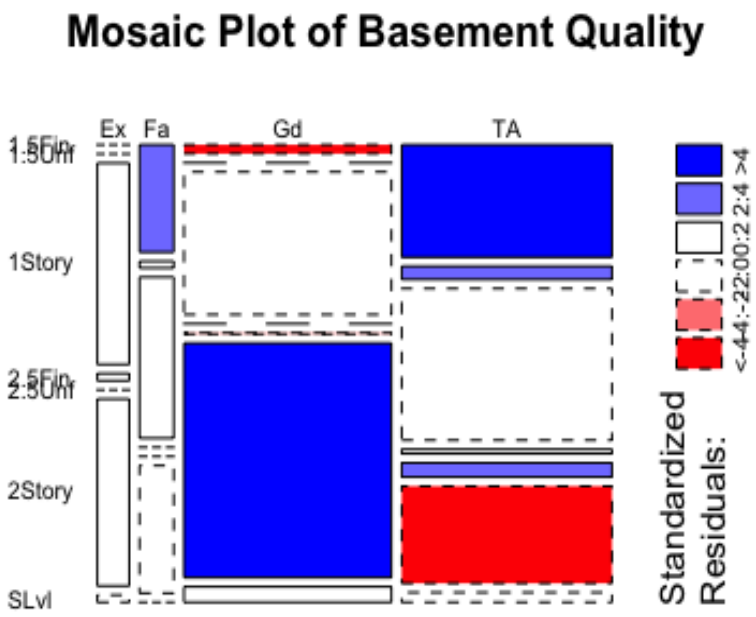



```
df$BsmtFinType2 <- as.character(df$BsmtFinType2)
df$BsmtFinType2[is.na(df$BsmtFinType2)]<-"ALQ"
df$BsmtFinType2 <- as.factor(df$BsmtFinType2)

#BsmtQual
BsmtQualUnf<-table(df$BsmtQual[df$BsmtUnfSF==df$TotalBsmtSF & df$TotalBsmtSF>
0],df$HouseStyle[df$BsmtUnfSF==df$TotalBsmtSF & df$TotalBsmtSF>0])
BsmtQualUnf<-BsmtQualUnf[rowSums(BsmtQualUnf != 0) > 0 , colSums(BsmtQualUnf
```

```
!= 0) > 0]
BsmtQualUnf
```

```
##
##      1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SLvl
##   Ex      0      0     28      1      0     26    1
##   Fa     16      1     24      0      0     19    0
##   Gd      8      0    129      0      1    212   14
##   TA    103     12    139      4     13     89    9
```

```
mosaicplot(BsmtQualUnf,
           main="Mosaic Plot of Basement Quality",
           las=1,
           color=T,
           shade=T)
```




```
TestQ2<-GTest(BsmtQualUnf)
TestQ2
```

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  BsmtQualUnf
## G = 220.7, X-squared df = 18, p-value < 2.2e-16
```

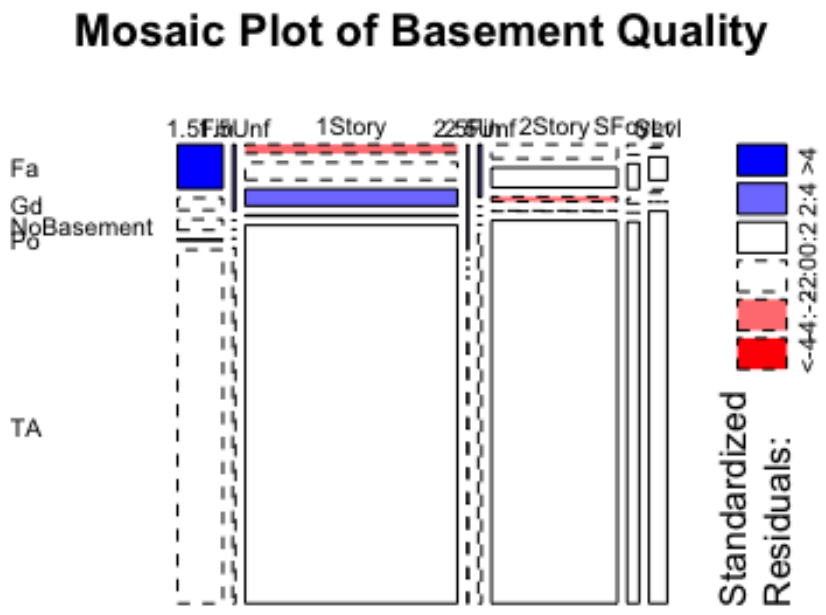
```
df$HouseStyle[is.na(df$BsmtQual)]
```

```
## [1] 2Story 1.5Fin
## Levels: 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer SLvl
```

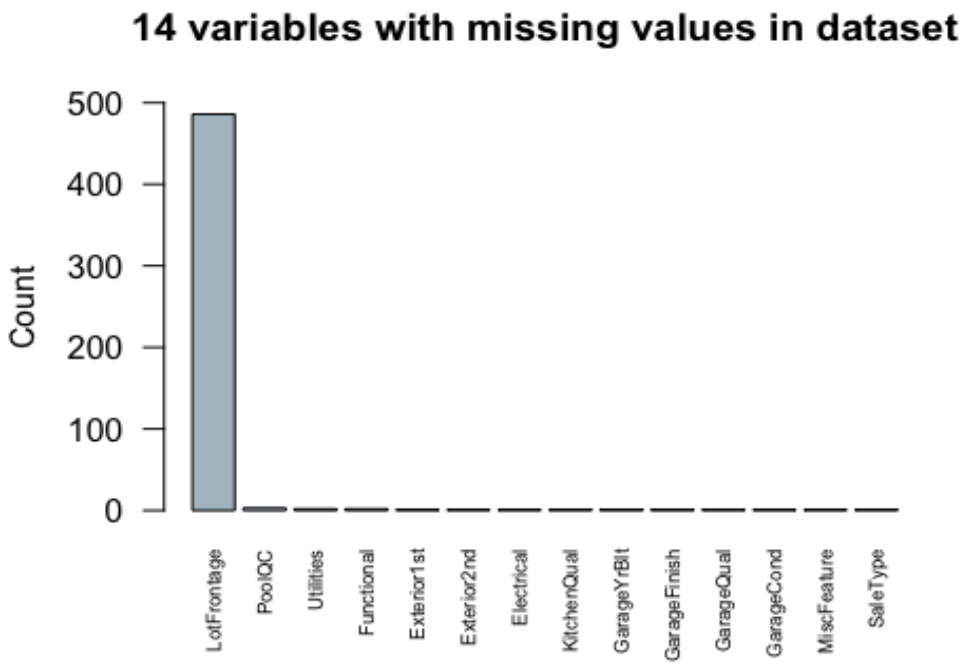
```r
df$BsmtQual <- as.character(df$BsmtQual)
df$BsmtQual[is.na(df$BsmtQual) & df$HouseStyle == "2Story"]<-"Gd"
df$BsmtQual[is.na(df$BsmtQual) & df$HouseStyle == "1.5Fin"]<-"TA"
df$BsmtQual <- as.factor(df$BsmtQual)

#BsmtCond
TableBsmtCond<-table(df$HouseStyle,df$BsmtCond)
TableBsmtCond<-TableBsmtCond[rowSums(TableBsmtCond != 0) > 0 , colSums(TableB
smtCond != 0) > 0]
TableBsmtCond
```

```
##
##             Fa   Gd NoBasement   Po   TA
##   1.5Fin    33    9          8    1  263
##   1.5Unf     3    0          0    0   16
##   1Story    31   60         59    3 1316
##   2.5Fin     2    0          0    0    6
##   2.5Unf     3    0          0    0   21
##   2Story    29   41         10    1  791
##   SFoyer     2    5          1    0   75
##   SLvl       1    7          1    0  118
```

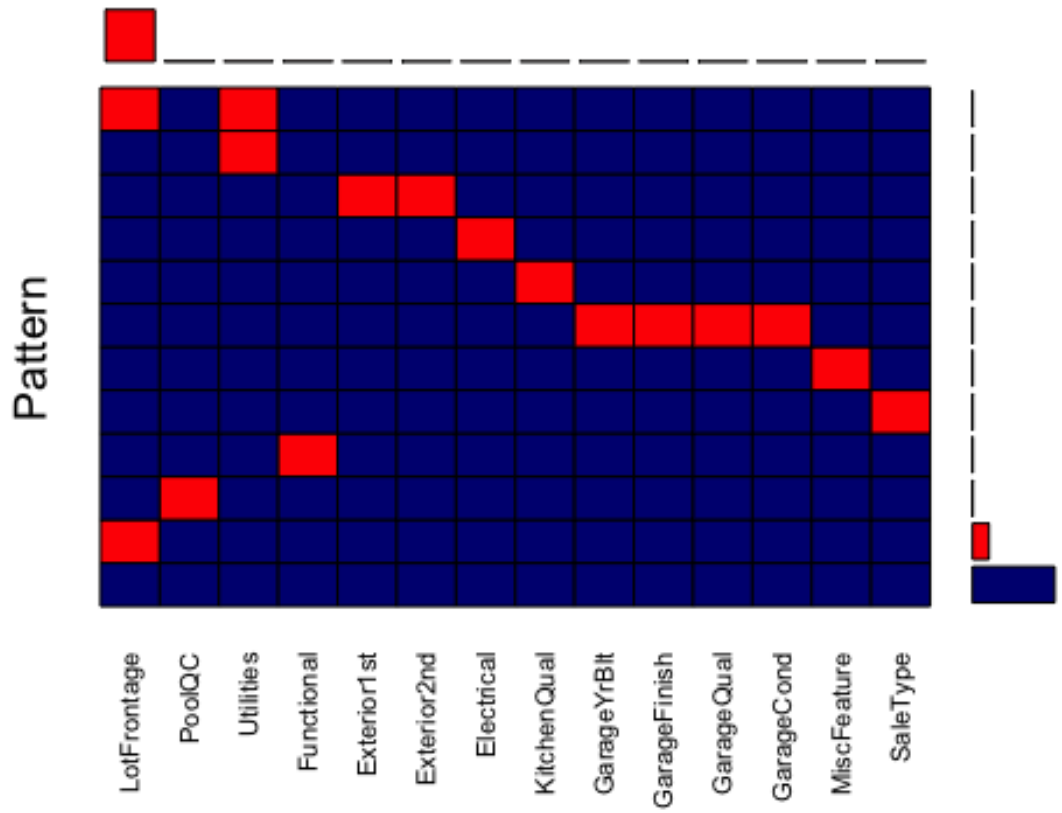
```r
mosaicplot(TableBsmtCond,
           main="Mosaic Plot of Basement Quality",
           las=1,
           color=T,
           shade=T)
```



```
TestQ2<-GTest(TableBsmtCond)
TestQ2

##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  TableBsmtCond
## G = 89.202, X-squared df = 28, p-value = 2.64e-08

df$HouseStyle[is.na(df$BsmtCond)]

## [1] 1Story 1Story SLvl
## Levels: 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer SLvl

df$BsmtCond <- as.character(df$BsmtCond)
df$BsmtCond[is.na(df$BsmtCond)]<-"TA"
df$BsmtCond <- as.factor(df$BsmtCond)
PM <- sort(apply(df,2,cMiss),decreasing=T);
barplot(PM[PM!=0],
        las=2,
        cex.names=0.6,
        ylab="Count",
        ylim=c(0,500),
        horiz=F,
        col="#AFC0CB",
        main=paste(toString(sum(PM!=0)), "variables with missing values in da
taset"))
```

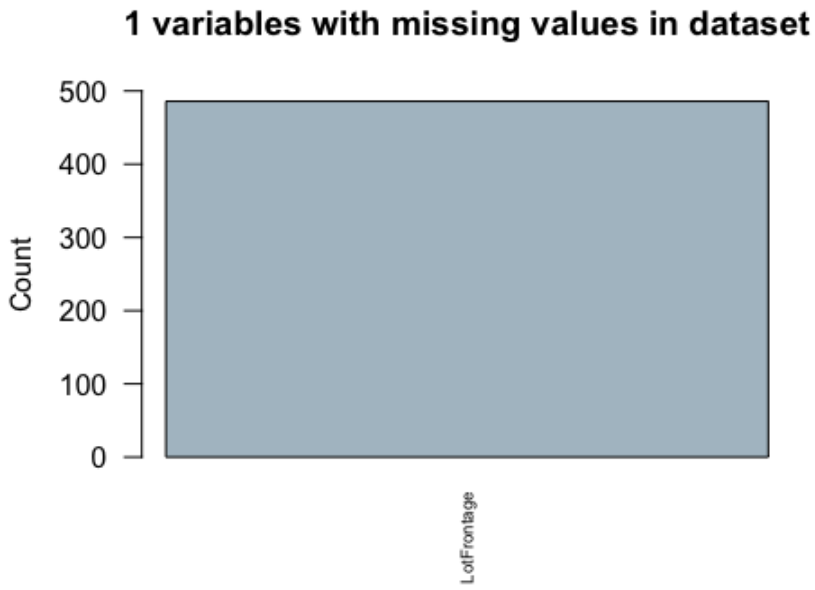


```
data = df[, names(PM[PM!=0])];
aggr_plot <- aggr(data,
                  col=c('navyblue','red'),
                  bars=T,
                  numbers=T,
                  combined = T,
                  labels=names(data),
                  cex.axis=.7,
                  gap=3,
                  ylab=c("Pattern"),
                  cex.numbers=0.74)

## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```

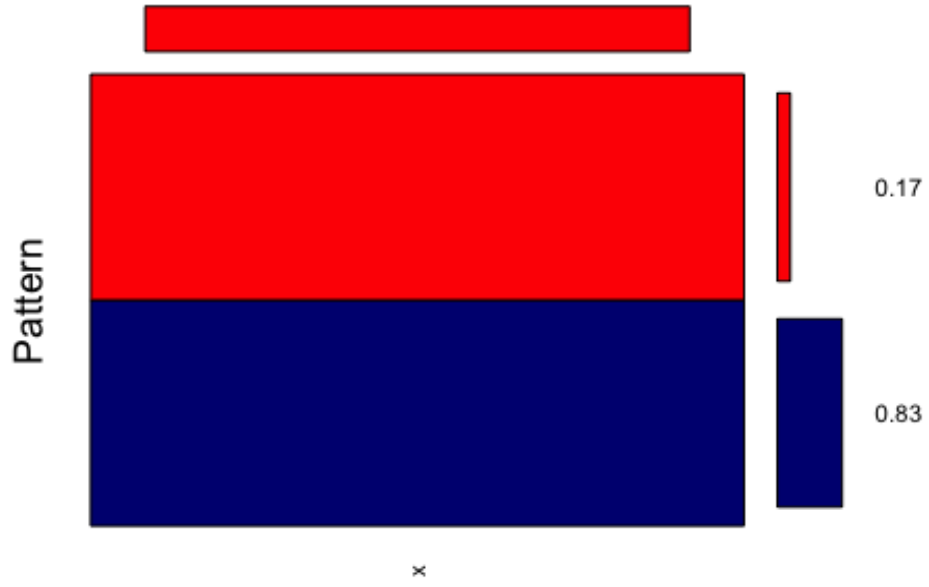



```
#The rest
fillMiss<- function(x)
{
  ux <- unique(x[!is.na(x)])
  x <- as.character(x)
  mode <- ux[which.max(tabulate(match(x[!is.na(x)], ux)))]
  x[is.na(x)] <- as.character(mode)
  x <- as.factor(x)
```

```
  return(x)
}
df[,sapply(df,function(x){!(is.numeric(x))}) ]<-as.data.frame(apply(df[,sappl
y(df,function(x){!(is.numeric(x))}) ],2,fillMiss))
PM <- sort(apply(df,2,cMiss),decreasing=T);
barplot(PM[PM!=0],
        las=2,
        cex.names=0.6,
        ylab="Count",
        ylim=c(0,500),
        horiz=F,
        col="#AFC0CB",
        main=paste(toString(sum(PM!=0)), "variables with missing values in da
taset"))
```



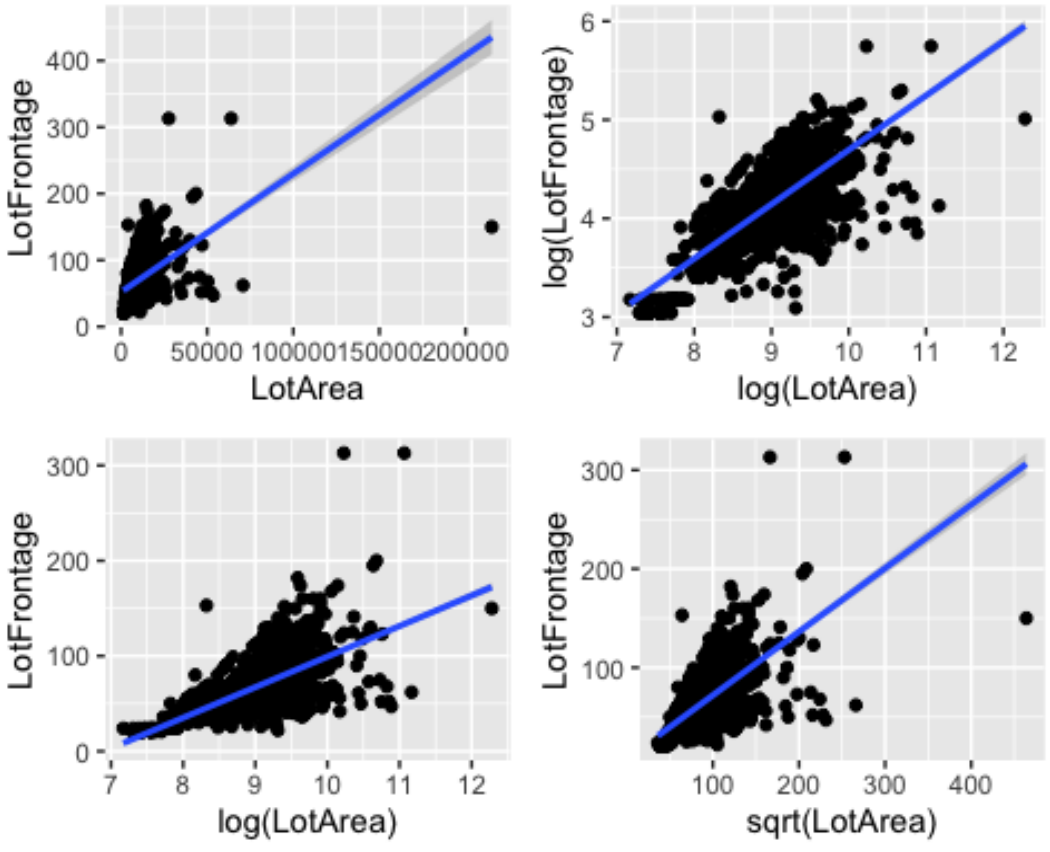

```
data = df[, names(PM[PM!=0])];
aggr_plot <- aggr(data,
                  col=c('navyblue','red'),
                  bars=T,
                  numbers=T,
                  combined = T,
                  labels=names(data),
                  cex.axis=.7,
                  gap=3,
                  ylab=c("Pattern"),
                  cex.numbers=0.74)
```




```
#LotFrontage Imputation
#Let's investigate this variable further. Maybe we could use a regression imp
utation technique to impute the missing LotFrontage variables.
#Since LotFrontage is defined as "Linear feet of street connected to property
", we would suspect that this variable would be related to quantities like "L
otArea", "Street", "LotShape", "LandContour", "LotConfig", "LandSlope", "Neig
hborhood", "BldgType".
```

```
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL)
{
  library(grid)
  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)
  numPlots = length(plots)
  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout))
  {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }
  if (numPlots==1)
  {
    print(plots[[1]])
  }
  else
  {
    # Set up the page
```

```
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))
    # Make each plot, in the correct location
    for (i in 1:numPlots)
    {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))
      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}

p1<-ggplot(df, aes(LotArea, LotFrontage)) + geom_point() + geom_smooth(method
= "lm", se = T)
p2<-ggplot(df, aes(log(LotArea), LotFrontage)) + geom_point() + geom_smooth(m
ethod = "lm", se = T)
p3<-ggplot(df, aes(log(LotArea), log(LotFrontage))) + geom_point() + geom_smo
oth(method = "lm", se = T)
p4<-ggplot(df, aes(sqrt(LotArea), LotFrontage)) + geom_point() + geom_smooth(
method = "lm", se = T)
multiplot(p1, p2, p3, p4, cols=2)
```




```
chisq.out.test(df$LotArea,opposite=F)
```

```
##
##  chi-squared test for outlier
##
## data:  df$LotArea
## X-squared = 676.1, p-value < 2.2e-16
## alternative hypothesis: highest value 215245 is an outlier
```

```
chisq.out.test(df$LotFrontage,opposite=F)
```

```
##
##  chi-squared test for outlier
##
## data:  df$LotFrontage
## X-squared = 108.97, p-value < 2.2e-16
## alternative hypothesis: highest value 313 is an outlier
```

```
chisq.out.test(df$LotArea,opposite=T)
```

```
##
##  chi-squared test for outlier
##
## data:  df$LotArea
## X-squared = 1.2643, p-value = 0.2608
## alternative hypothesis: lowest value 1300 is an outlier
```

```
chisq.out.test(df$LotFrontage,opposite=T)
```

```
##
##  chi-squared test for outlier
##
## data:  df$LotFrontage
## X-squared = 4.2817, p-value = 0.03853
## alternative hypothesis: lowest value 21 is an outlier
```

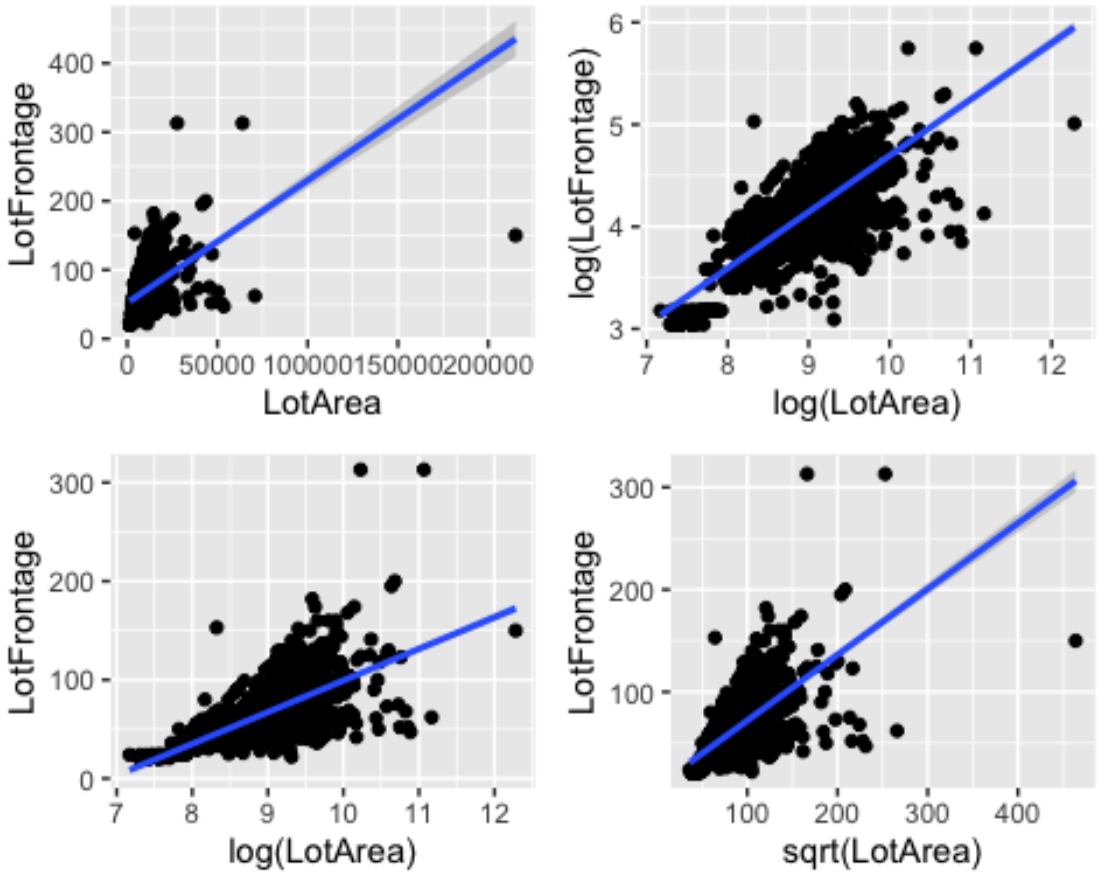
```
grubbs.test(df$LotArea,type=11)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  df$LotArea
## G = 27.12630, U = 0.76779, p-value < 2.2e-16
## alternative hypothesis: 1300 and 215245 are outliers
```

```
grubbs.test(df$LotFrontage,type=11)
```

```
##
##  Grubbs test for two opposite outliers
##
## data:  df$LotFrontage
## G = 12.50808, U = 0.95342, p-value < 2.2e-16
## alternative hypothesis: 21 and 313 are outliers
```

```
p1<-ggplot(df  , aes(LotArea, LotFrontage)) + geom_point() + geom_smooth(meth
od = "lm", se = T)
p2<-ggplot(df, aes(log(LotArea), LotFrontage)) + geom_point() + geom_smooth(m
ethod = "lm", se = T)
p3<-ggplot(df, aes(log(LotArea), log(LotFrontage))) + geom_point() + geom_smo
oth(method = "lm", se = T)
p4<-ggplot(df, aes(sqrt(LotArea), LotFrontage)) + geom_point() + geom_smooth(
method = "lm", se = T)
multiplot(p1, p2, p3, p4, cols=2)
```




```
cor(as.numeric(df$LotArea),as.numeric(df$LotFrontage),use="complete.obs")

## [1] 0.4898956

cor(log(as.numeric(df$LotArea)),log(as.numeric(df$LotFrontage)),use="complete
.obs")

## [1] 0.7662858

cor(log(as.numeric(df$LotArea)),as.numeric(df$LotFrontage),use="complete.obs"
)

## [1] 0.6835123
```

```r
cor(sqrt(as.numeric(df$LotArea)),as.numeric(df$LotFrontage),use="complete.obs
")
```

```
## [1] 0.647658
```

```r
#We have cleaned the dataset, imputed the missing values
str(df)
```

```
## 'data.frame':    2919 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5
4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7
420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ..
.
##  $ Alley        : Factor w/ 3 levels "Grvl","NoAccess",..: 2 2 2 2 2 2 2 2
2 2 ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1
4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4
4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1
1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5
1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1
1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14
12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5
1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3
3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1
2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6
1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 .
..
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 .
..
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2
```

```
2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 1
3 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 1
4 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4
4 3 3 ...
##  $ MasVnrArea   : num  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4
4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5
5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2
1 1 ...
##  $ BsmtQual     : Factor w/ 5 levels "Ex","Fa","Gd",..: 3 3 3 5 3 3 1 3 5
5 ...
##  $ BsmtCond     : Factor w/ 5 levels "Fa","Gd","NoBasement",..: 5 5 5 2 5
5 5 5 5 5 ...
##  $ BsmtExposure : Factor w/ 5 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4
4 ...
##  $ BsmtFinType1 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1
7 3 ...
##  $ BsmtFinSF1   : num  706 978 486 216 655 ...
##  $ BsmtFinType2 : Factor w/ 7 levels "ALQ","BLQ","GLQ",..: 7 7 7 7 7 7 7 2
7 7 ...
##  $ BsmtFinSF2   : num  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : num  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : num  856 1262 920 756 1145 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3
1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2
5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 .
..
##  $ BsmtFullBath : num  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4
4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7
```

```
...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 6 levels "Ex","Fa","Gd",..: 4 6 6 3 6 4 3 6 6
6 ...
##  $ GarageType   : Factor w/ 7 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2
6 2 ...
##  $ GarageYrBlt  : Factor w/ 104 levels "1895","1896",..: 95 68 93 90 92 85
96 65 24 32 ...
##  $ GarageFinish : Factor w/ 4 levels "Fin","NoGarage",..: 3 3 3 4 3 4 3 3
4 3 ...
##  $ GarageCars   : num  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : num  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 6 levels "Ex","Fa","Gd",..: 6 6 6 6 6 6 6 6 2
3 ...
##  $ GarageCond   : Factor w/ 6 levels "Ex","Fa","Gd",..: 6 6 6 6 6 6 6 6 6
6 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 4 levels "Ex","Fa","Gd",..: 4 4 4 4 4 4 4 4 4
4 ...
##  $ Fence        : Factor w/ 5 levels "GdPrv","GdWo",..: 5 5 5 5 5 3 5 5 5
5 ...
##  $ MiscFeature  : Factor w/ 5 levels "Gar2","None",..: 2 2 2 2 2 4 2 4 2 2
...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 .
..
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9
9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5
5 1 5 ...
##  $ SalePrice    : num  208500 181500 223500 140000 250000 ...
```

Loading packages for regression

```
library(car)

## Loading required package: carData

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

```
## The following object is masked from 'package:DescTools':
##
##     Recode

library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

library(ggplot2)
library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library(qqplotr)

##
## Attaching package: 'qqplotr'

## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine

library(MASS)
library(gvlma)
library(leaps)

fit <- lm (SalePrice~MSZoning+I(LotArea^2)+Street+
              LotConfig+LandSlope+Condition1+OverallQual+OverallCond+
              RoofStyle+RoofMatl+MasVnrArea+ExterQual+BsmtQual+BsmtExposure+B
smtFinSF2 +
              BsmtUnfSF+X1stFlrSF+X2ndFlrSF+BedroomAbvGr+KitchenQual+GarageQu
al+
              ScreenPorch+PoolArea+PoolQC,data=df)

summary(fit)

##
## Call:
## lm(formula = SalePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig +
##     LandSlope + Condition1 + OverallQual + OverallCond + RoofStyle +
##     RoofMatl + MasVnrArea + ExterQual + BsmtQual + BsmtExposure +
##     BsmtFinSF2 + BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr +
##     KitchenQual + GarageQual + ScreenPorch + PoolArea + PoolQC,
##     data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -341292  -83668   -5511   77989  418061
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.781e+04  1.563e+05   0.178  0.85883
## MSZoningFV              1.623e+04  2.361e+04   0.688  0.49176
## MSZoningRH              2.964e+04  2.890e+04   1.026  0.30520
## MSZoningRL              2.779e+04  2.178e+04   1.276  0.20216
## MSZoningRM              1.181e+04  2.197e+04   0.538  0.59081
.

.
## ScreenPorch            -2.410e+01  3.383e+01  -0.712  0.47625
## PoolArea                4.158e+01  1.120e+02   0.371  0.71053
## PoolQCFa                5.949e+04  9.107e+04   0.653  0.51362
## PoolQCGd               -3.973e+04  8.622e+04  -0.461  0.64498
## PoolQCNoPool           -2.747e+04  6.455e+04  -0.426  0.67044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 99260 on 2855 degrees of freedom
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.1313
## F-statistic:     8 on 63 and 2855 DF,  p-value: < 2.2e-16
```

```
coefficients(fit)
```

```
##            (Intercept)              MSZoningFV              MSZoningRH
##           2.780848e+04           1.623213e+04           2.964091e+04
##              MSZoningRL              MSZoningRM           I(LotArea^2)
##           2.778636e+04           1.181323e+04           6.154794e-06
##              StreetPave       LotConfigCulDSac           LotConfigFR2
.

.
##          -2.409991e+01           4.157891e+01           5.949329e+04
##                PoolQCGd           PoolQCNoPool
##          -3.973143e+04          -2.747183e+04
```

```
#str(df)
confint(fit,level=0.95)
```

```
##                                2.5 %        97.5 %
## (Intercept)            -2.787315e+05  3.343485e+05
## MSZoningFV             -3.005618e+04  6.252044e+04
## MSZoningRH             -2.703222e+04  8.631404e+04
.
```

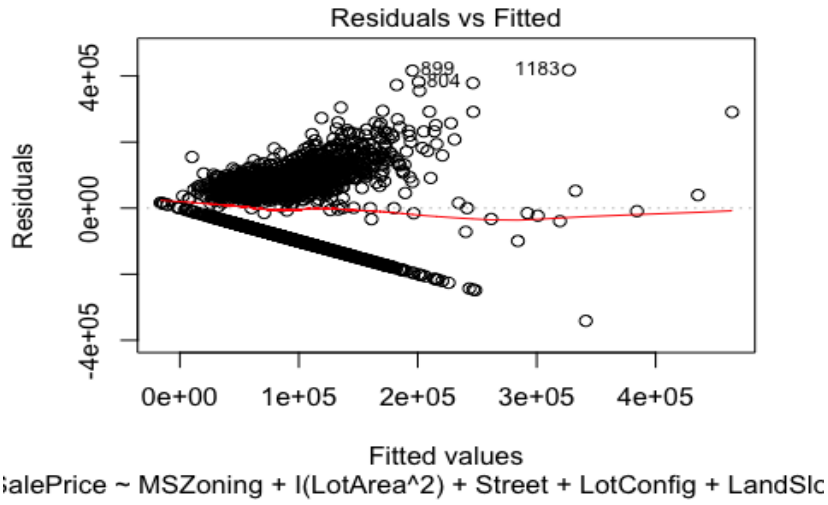
```
.
## PoolQCGd                -2.087973e+05  1.293345e+05
## PoolQCNoPool            -1.540409e+05  9.909728e+04
```

Predicted Values

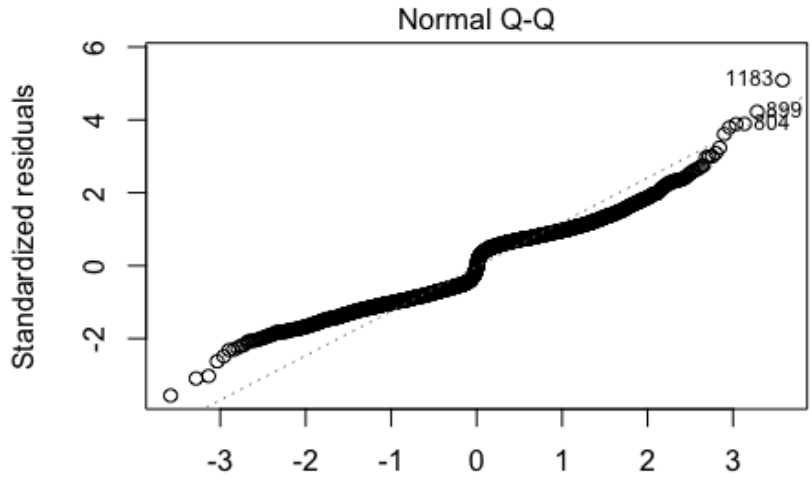
```
fitted(fit)

##           1           2           3           4           5           6
## 108276.1023  84220.0187 110927.0940 106836.6611 155744.2913  86758.0079
##           7           8           9          10          11          12
## 126057.5886  95918.1020  67538.7647  69773.1369  60537.3915 176720.1078
##          13          14          15          16          17          18
.

.

##          85          86          87          88          89          90
##  86923.1747 156345.9965 109555.0150  93426.1067  16977.3672  46461.8187
##          91          92          93          94          95          96
##  41605.2663  73602.5363  67965.3749  91379.0660 105157.1843 115273.1327

.

.

.
```

#diagnostic plots

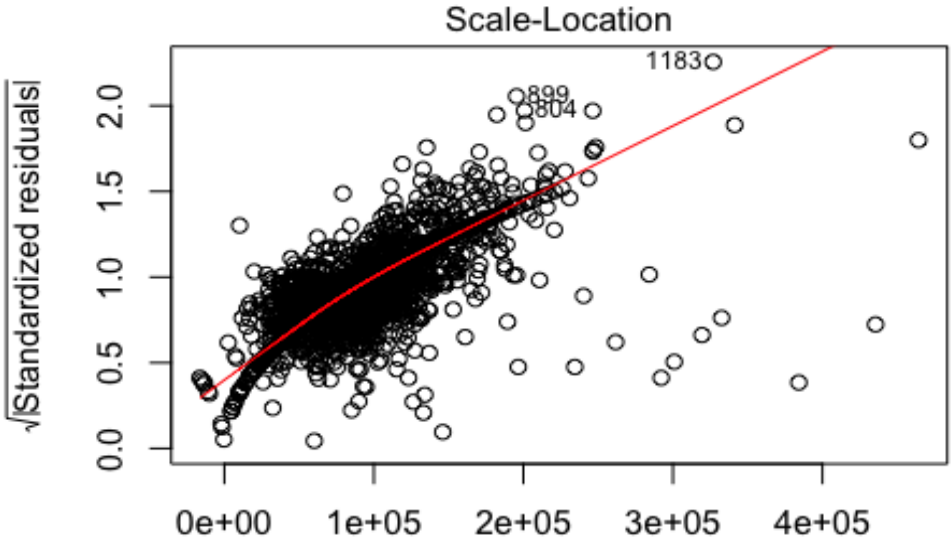
```
plot(fit)
##   121, 272, 1276, 1299
```



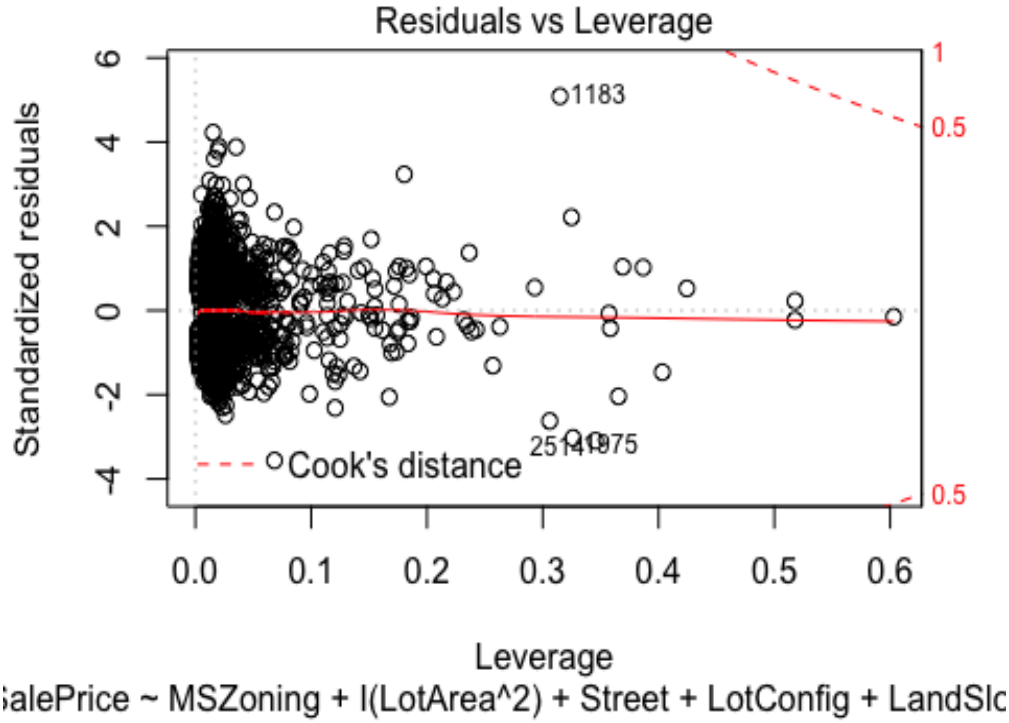
alePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig + LandSlc



alePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig + LandSlc
121, 272, 1276, 1299

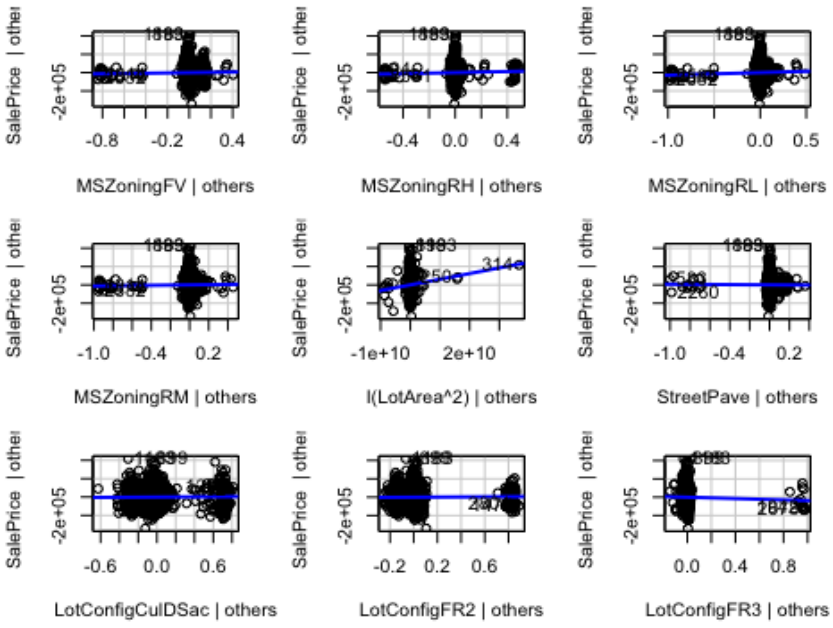


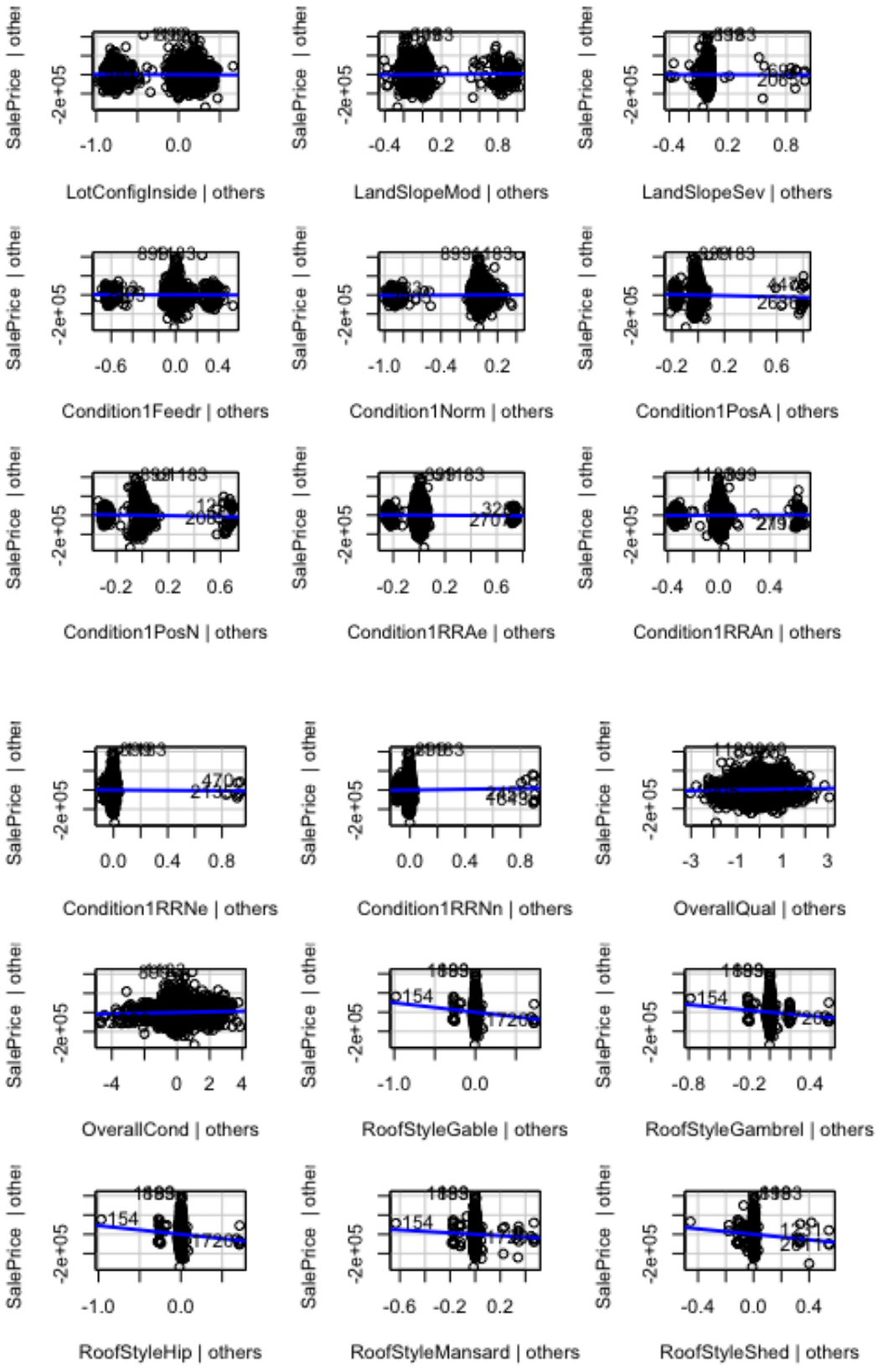
alePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig + LandSlc

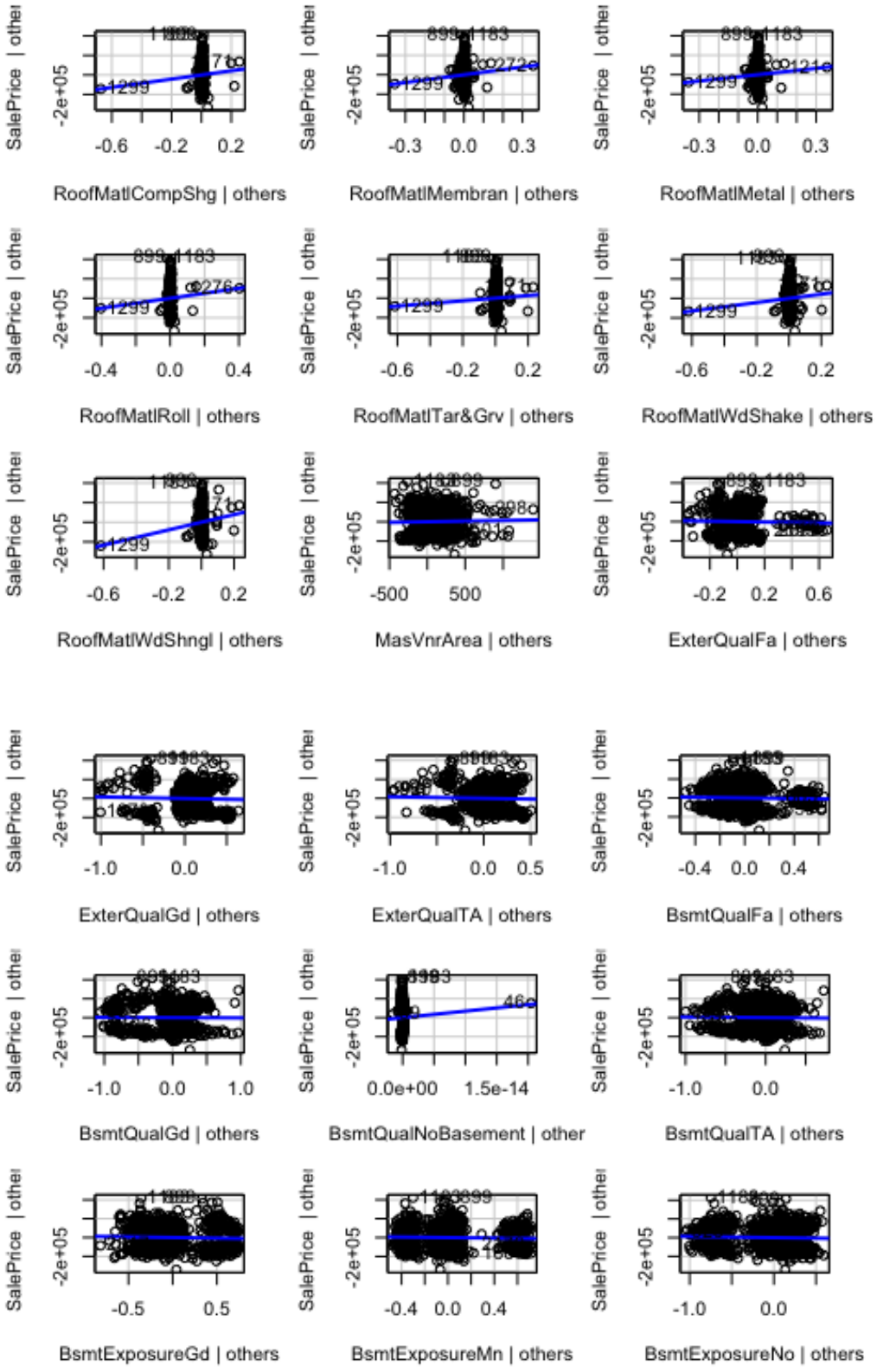


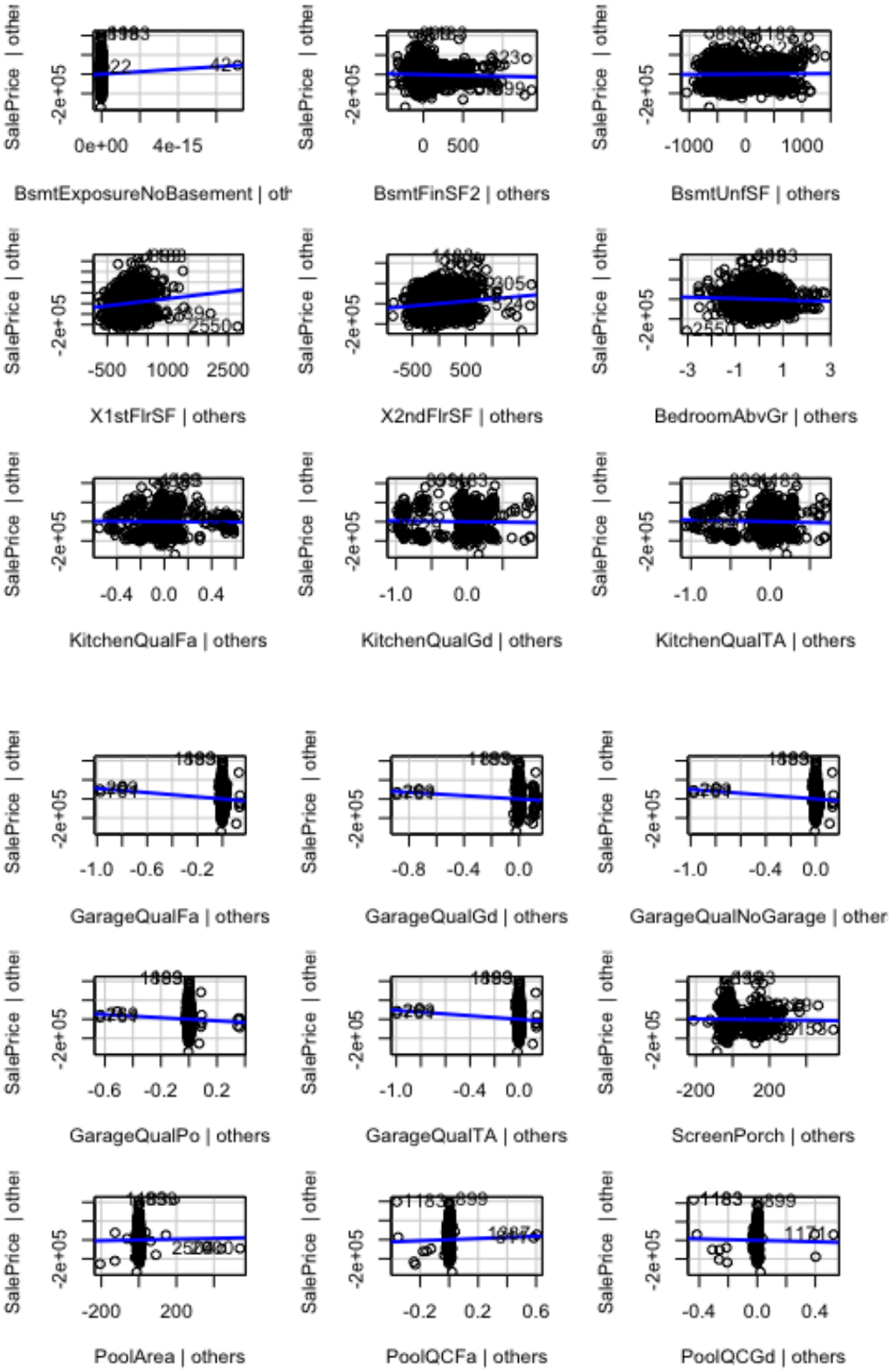
Influential Observations

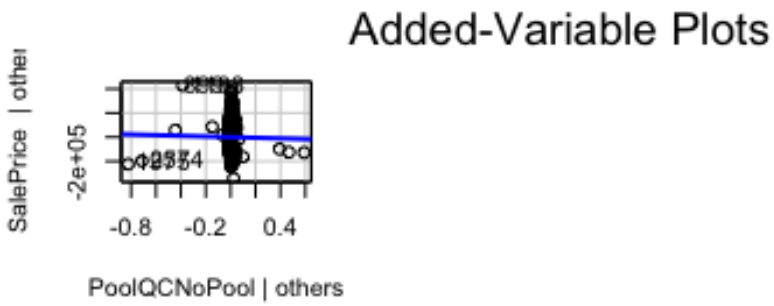
```
# added variable plots
avPlots(fit)
```




SalePrice | others
-2e+05
LotConfigInside | others
LandSlopeMod | others
LandSlopeSev | others
Condition1Feedr | others
Condition1Norm | others
Condition1PosA | others
Condition1PosN | others
Condition1RRAe | others
Condition1RRAn | others
Condition1RRNe | others
Condition1RRNn | others
OverallQual | others
OverallCond | others
RoofStyleGable | others
RoofStyleGambrel | others
RoofStyleHip | others
RoofStyleMansard | others
RoofStyleShed | others

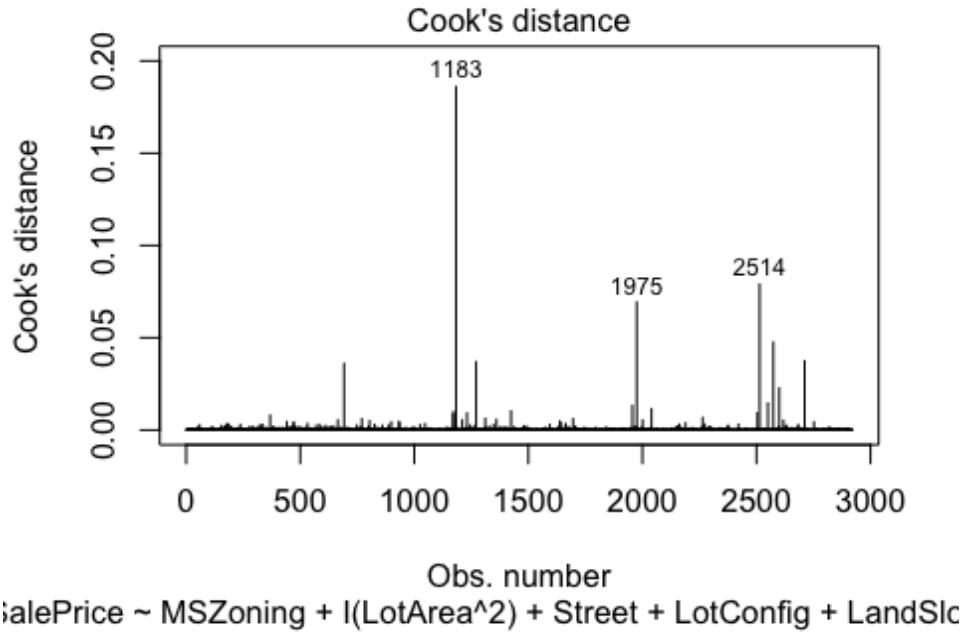

SalePrice | others
-2e+05
RoofMatlCompShg | others
RoofMatlMembran | others
RoofMatlMetal | others
RoofMatlRoll | others
RoofMatlTar&Grv | others
RoofMatlWdShake | others
RoofMatlWdShngl | others
MasVnrArea | others
ExterQualFa | others
ExterQualGd | others
ExterQualTA | others
BsmtQualFa | others
BsmtQualGd | others
BsmtQualNoBasement | other
BsmtQualTA | others
BsmtExposureGd | others
BsmtExposureMn | others
BsmtExposureNo | others


SalePrice | other
-2e+05
0e+00 4e-15
BsmtExposureNoBasement | oth
0 500
BsmtFinSF2 | others
-1000 0 1000
BsmtUnfSF | others
-500 1000 2500
X1stFlrSF | others
-500 500
X2ndFlrSF | others
-3 -1 1 3
BedroomAbvGr | others
-0.4 0.0 0.4
KitchenQualFa | others
-1.0 0.0
KitchenQualGd | others
-1.0 0.0
KitchenQualTA | others
-1.0 -0.6 -0.2
GarageQualFa | others
-0.8 -0.4 0.0
GarageQualGd | others
-1.0 -0.4 0.0
GarageQualNoGarage | other
-0.6 -0.2 0.2
GarageQualPo | others
-1.0 -0.4 0.0
GarageQualTA | others
-200 200
ScreenPorch | others
-200 200
PoolArea | others
-0.2 0.2 0.6
PoolQCFa | others
-0.4 0.0 0.4
PoolQCGd | others



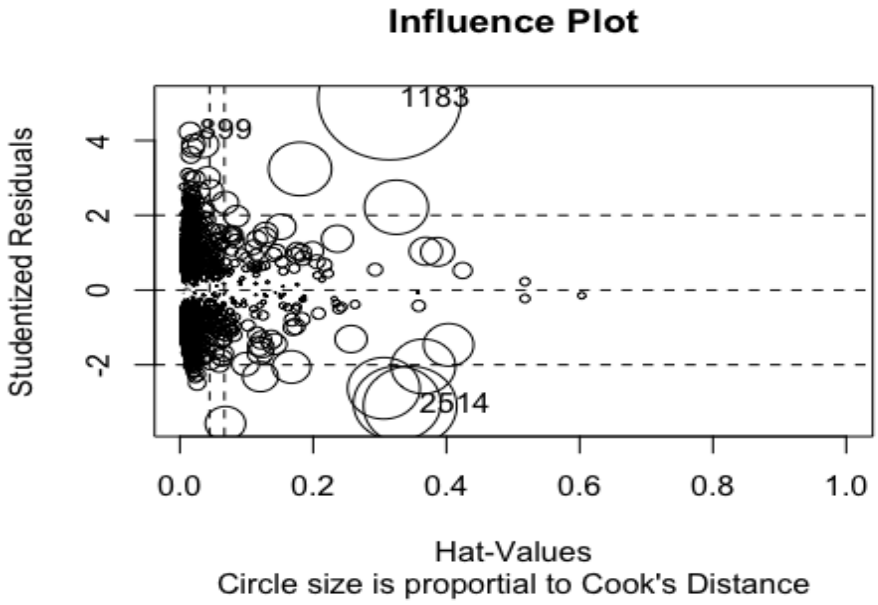
Cook's D plot

```
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(df)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
```

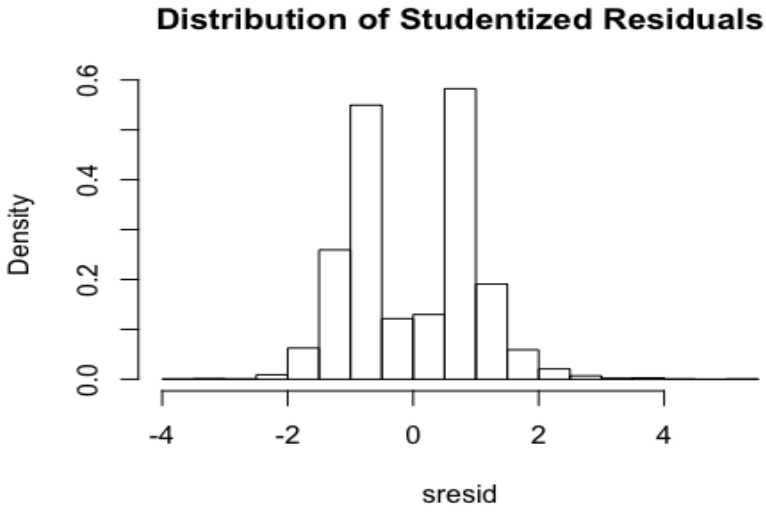


Influence Plot

```
influencePlot(fit, id.method="identify", main="Influence Plot", sub="Circle s
ize is proportial to Cook's Distance" )
```

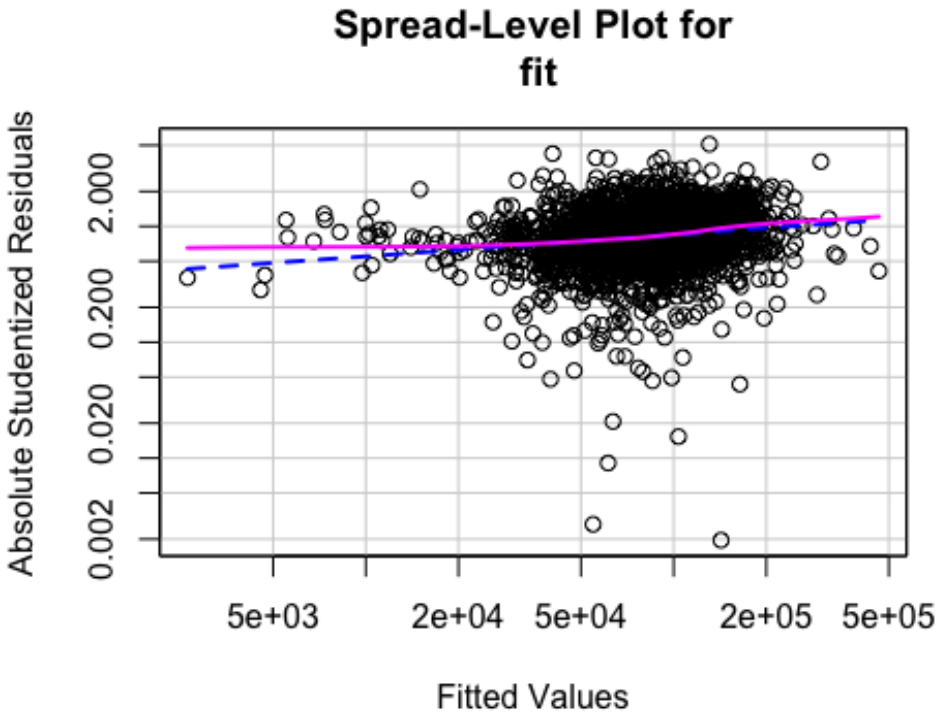



```
##         StudRes        Hat       CookD
## 121         NaN 1.00000000         NaN
## 272         NaN 1.00000000         NaN
## 899    4.237423 0.01522266 0.004311263
## 1183   5.110882 0.31491507 0.185975667
## 2514  -3.098459 0.34554705 0.078964872

# distribution of studentized residuals
sresid <- studres(fit)

## Warning in sqrt((n - p - sr^2)/(n - p - 1)): NaNs produced

hist(sresid, freq=FALSE, main="Distribution of Studentized Residuals")
```



Evaluate homoscedasticity

```
# non-constant error variance test
ncvTest(fit)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 932.7669, Df = 1, p = < 2.22e-16

# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)

## Warning in spreadLevelPlot.lm(fit):
## 9 negative fitted values removed
```




```
##
## Suggested power transformation:  0.81225
```

#Non-independence of Errors

```
# Test for Autocorrelated Errors
durbinWatsonTest(fit)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.8373218     0.3243811       0
##  Alternative hypothesis: rho != 0
```

Global test of model assumptions

```
fit

##
## Call:
## lm(formula = SalePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig +
##     LandSlope + Condition1 + OverallQual + OverallCond + RoofStyle +
##     RoofMatl + MasVnrArea + ExterQual + BsmtQual + BsmtExposure +
##     BsmtFinSF2 + BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr +
##     KitchenQual + GarageQual + ScreenPorch + PoolArea + PoolQC,
##     data = df)
##
## Coefficients:
##            (Intercept)              MSZoningFV              MSZoningRH
##              2.781e+04               1.623e+04               2.964e+04
##             MSZoningRL              MSZoningRM            I(LotArea^2)
##              2.779e+04               1.181e+04               6.155e-06
##             StreetPave       LotConfigCulDSac             LotConfigFR2
##             -7.166e+03               6.363e+03               7.290e+03
##           LotConfigFR3        LotConfigInside             LandSlopeMod
##             -3.727e+04              -4.120e+03               7.602e+03
##           LandSlopeSev        Condition1Feedr         Condition1Norm
##             -1.347e+03              -2.845e+03               4.841e+03
##         Condition1PosA         Condition1PosN         Condition1RRAe
##             -3.175e+04              -2.236e+04              -1.004e+04
##         Condition1RRAn         Condition1RRNe         Condition1RRNn
##              5.386e+03              -1.070e+04               2.042e+04
##            OverallQual             OverallCond         RoofStyleGable
##              4.105e+03               3.456e+03              -9.797e+04
##       RoofStyleGambrel           RoofStyleHip        RoofStyleMansard
##             -9.676e+04              -8.999e+04              -7.571e+04
##          RoofStyleShed        RoofMatlCompShg         RoofMatlMembran
##             -1.424e+05               2.125e+05               2.661e+05
##          RoofMatlMetal           RoofMatlRoll        RoofMatlTar&Grv
##              2.170e+05               2.563e+05               1.352e+05
##        RoofMatlWdShake        RoofMatlWdShngl              MasVnrArea
##              2.158e+05               3.897e+05               1.337e+01
##            ExterQualFa             ExterQualGd             ExterQualTA
##             -2.333e+04              -1.718e+04              -1.725e+04
##             BsmtQualFa              BsmtQualGd     BsmtQualNoBasement
##             -1.796e+04              -2.749e+03              -3.259e+04
##             BsmtQualTA          BsmtExposureGd          BsmtExposureMn
##             -7.289e+03              -1.387e+04              -1.413e+04
##         BsmtExposureNo  BsmtExposureNoBasement              BsmtFinSF2
##             -1.242e+04                      NA              -1.894e+01
##              BsmtUnfSF             X1stFlrSF              X2ndFlrSF
##              3.572e+00              4.527e+01              4.693e+01
##           BedroomAbvGr          KitchenQualFa          KitchenQualGd
##             -6.974e+03              -9.360e+03              -9.284e+03
##          KitchenQualTA           GarageQualFa           GarageQualGd
```

```
##             -1.810e+04              -1.100e+05              -8.755e+04
##     GarageQualNoGarage            GarageQualPo            GarageQualTA
##             -9.865e+04              -8.261e+04              -9.380e+04
##            ScreenPorch                PoolArea                  PoolQCFa
##             -2.410e+01              4.158e+01               5.949e+04
##               PoolQCGd            PoolQCNoPool
##             -3.973e+04              -2.747e+04
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = SalePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig +
##     LandSlope + Condition1 + OverallQual + OverallCond + RoofStyle +
##     RoofMatl + MasVnrArea + ExterQual + BsmtQual + BsmtExposure +
##     BsmtFinSF2 + BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr +
##     KitchenQual + GarageQual + ScreenPorch + PoolArea + PoolQC,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -341292  -83668   -5511   77989  418061
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.781e+04  1.563e+05   0.178  0.85883
## MSZoningFV              1.623e+04  2.361e+04   0.688  0.49176
## MSZoningRH              2.964e+04  2.890e+04   1.026  0.30520
## MSZoningRL              2.779e+04  2.178e+04   1.276  0.20216
## MSZoningRM              1.181e+04  2.197e+04   0.538  0.59081
## I(LotArea^2)            6.155e-06  1.934e-06   3.183  0.00147 **
## StreetPave             -7.166e+03  3.150e+04  -0.228  0. 105e+03

.

.
## OverallCond             3.456e+03  1.848e+03   1.871  0.06149 .
## RoofStyleGable         -9.797e+04  4.316e+04  -2.270  0.02328 *
## RoofStyleGambrel       -9.676e+04  4.824e+04  -2.006  0.04499 *
.

.
## PoolArea                4.158e+01  1.120e+02   0.371  0.71053
## PoolQCFa                5.949e+04  9.107e+04   0.653  0.51362
## PoolQCGd               -3.973e+04  8.622e+04  -0.461  0.64498
## PoolQCNoPool           -2.747e+04  6.455e+04  -0.426  0.67044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 99260 on 2855 degrees of freedom
```

```
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.1313
## F-statistic:     8 on 63 and 2855 DF,  p-value: < 2.2e-16

fit1 <- fit
fit2 <- lm (SalePrice~MSZoning+I(LotArea^2)+Street+
              RoofStyle+RoofMatl+MasVnrArea+ExterQual+BsmtQual+BsmtExposure+Bs
mtFinSF2 +
              BsmtUnfSF+X1stFlrSF+X2ndFlrSF+BedroomAbvGr+KitchenQual+GarageQua
l+
              ScreenPorch+PoolArea+PoolQC,data=df)
# compare models
anova(fit1, fit2)

## Analysis of Variance Table
##
## Model 1: SalePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig + LandSl
ope +
##     Condition1 + OverallQual + OverallCond + RoofStyle + RoofMatl +
##     MasVnrArea + ExterQual + BsmtQual + BsmtExposure + BsmtFinSF2 +
##     BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr + KitchenQual +
##     GarageQual + ScreenPorch + PoolArea + PoolQC
## Model 2: SalePrice ~ MSZoning + I(LotArea^2) + Street + RoofStyle + RoofMa
tl +
##     MasVnrArea + ExterQual + BsmtQual + BsmtExposure + BsmtFinSF2 +
##     BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr + KitchenQual +
##     GarageQual + ScreenPorch + PoolArea + PoolQC
##   Res.Df        RSS  Df   Sum of Sq      F Pr(>F)
## 1   2855 2.8129e+13
## 2   2871 2.8314e+13 -16 -1.8511e+11 1.1742 0.2807

step <- stepAIC(fit, direction="both")

## Start:  AIC=67232.39
## SalePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig + LandSlope +
##     Condition1 + OverallQual + OverallCond + RoofStyle + RoofMatl +
##     MasVnrArea + ExterQual + BsmtQual + BsmtExposure + BsmtFinSF2 +
##     BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr + KitchenQual +
##     GarageQual + ScreenPorch + PoolArea + PoolQC
##
##                Df  Sum of Sq        RSS   AIC
## - Condition1    8 6.8394e+10 2.8197e+13 67223
## - PoolQC        3 1.4338e+10 2.8143e+13 67228
.

.
## - RoofMatl      7 2.8778e+11 2.8417e+13 67248
## - X1stFlrSF     1 3.2759e+11 2.8457e+13 67264
## - X2ndFlrSF     1 4.6803e+11 2.8597e+13 67279
##
## Step:  AIC=67223.48
## SalePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig + LandSlope +
```

```
##     OverallQual + OverallCond + RoofStyle + RoofMatl + MasVnrArea +
##     ExterQual + BsmtQual + BsmtExposure + BsmtFinSF2 + BsmtUnfSF +
##     X1stFlrSF + X2ndFlrSF + BedroomAbvGr + KitchenQual + GarageQual +
##     ScreenPorch + PoolArea + PoolQC
##
##                 Df  Sum of Sq        RSS   AIC
## - ExterQual      3 1.4416e+10 2.8212e+13 67219
## - PoolQC         3 1.5331e+10 2.8213e+13 67219
## - GarageQual     5 5.5792e+10 2.8253e+13 67219
## - BsmtQual       3 2.0073e+10 2.8218e+13 67220
## - LotConfig      4 4.3075e+10 2.8241e+13 67220
## - LandSlope      2 6.5522e+09 2.8204e+13 67220
## - KitchenQual    3 3.6932e+10 2.8234e+13 67221
## - Street         1 7.1861e+08 2.8198e+13 67222
## - PoolArea       1 1.0784e+09 2.8199e+13 67222
## - BsmtUnfSF      1 5.5263e+09 2.8203e+13 67222
## - ScreenPorch    1 5.7465e+09 2.8203e+13 67222
## - RoofStyle      5 8.5628e+10 2.8283e+13 67222
## - MasVnrArea     1 1.1600e+10 2.8209e+13 67223
## - BsmtExposure   3 5.7124e+10 2.8255e+13 67223
## <none>                        2.8197e+13 67223
## - BsmtFinSF2     1 2.4691e+10 2.8222e+13 67224
## - OverallQual    1 2.5470e+10 2.8223e+13 67224
## - OverallCond    1 3.3582e+10 2.8231e+13 67225
## - MSZoning       4 9.3724e+10 2.8291e+13 67225
## - BedroomAbvGr   1 5.0718e+10 2.8248e+13 67227
## - I(LotArea^2)   1 9.9063e+10 2.8297e+13 67232
## + Condition1     8 6.8394e+10 2.8129e+13 67232
## - RoofMatl       7 2.7935e+11 2.8477e+13 67238
## - X1stFlrSF      1 3.1013e+11 2.8508e+13 67253
## - X2ndFlrSF      1 4.6652e+11 2.8664e+13 67269
##

## Step:  AIC=67200.11
## Step:  AIC=67193.99
## SalePrice ~ MSZoning + I(LotArea^2) + OverallQual + OverallCond +
##     RoofMatl + MasVnrArea + BsmtExposure + BsmtFinSF2 + X1stFlrSF +
##     X2ndFlrSF + BedroomAbvGr + KitchenQual + PoolArea
##
##                 Df  Sum of Sq        RSS   AIC
## - BsmtExposure   4 7.6969e+10 2.8532e+13 67194
## <none>                        2.8455e+13 67194
## - PoolArea       1 2.1262e+10 2.8476e+13 67194
## - MasVnrArea     1 2.3513e+10 2.8478e+13 67194
## + RoofStyle      5 9.0880e+10 2.8364e+13 67195
## - BsmtFinSF2     1 2.8728e+10 2.8483e+13 67195
## - OverallCond    1 2.9519e+10 2.8484e+13 67195
## - KitchenQual    3 7.5286e+10 2.8530e+13 67196
## + ScreenPorch    1 1.4467e+09 2.8453e+13 67196
## + BsmtUnfSF      1 1.4304e+09 2.8453e+13 67196
```

```
## + Street        1 3.6872e+08 2.8454e+13 67196
## + BsmtQual      3 2.8895e+10 2.8426e+13 67197
## + LotConfig     4 4.7192e+10 2.8408e+13 67197
## + LandSlope     2 5.6954e+09 2.8449e+13 67197
## + ExterQual     3 2.2159e+10 2.8433e+13 67198
## - BedroomAbvGr  1 5.7835e+10 2.8513e+13 67198
## + GarageQual    5 5.4369e+10 2.8400e+13 67198
## + PoolQC        3 1.1687e+10 2.8443e+13 67199
## - MSZoning      4 1.3359e+11 2.8588e+13 67200
## - OverallQual   1 9.4432e+10 2.8549e+13 67202
## + Condition1    8 6.5324e+10 2.8389e+13 67203
## - I(LotArea^2)  1 1.4550e+11 2.8600e+13 67207
## - RoofMatl      7 2.9990e+11 2.8755e+13 67211
## - X1stFlrSF     1 3.9528e+11 2.8850e+13 67232
## - X2ndFlrSF     1 4.8649e+11 2.8941e+13 67241
##

.

.
## Step:  AIC=67193.87
## SalePrice ~ MSZoning + I(LotArea^2) + OverallQual + OverallCond +
##     RoofMatl + MasVnrArea + BsmtFinSF2 + X1stFlrSF + X2ndFlrSF +
##     BedroomAbvGr + KitchenQual + PoolArea
##
##                Df  Sum of Sq        RSS   AIC
## <none>                       2.8532e+13 67194
## + BsmtExposure  4 7.6969e+10 2.8455e+13 67194
## - PoolArea      1 2.3425e+10 2.8555e+13 67194
## - BsmtFinSF2    1 2.5108e+10 2.8557e+13 67194
## - MasVnrArea    1 2.7022e+10 2.8559e+13 67195
## - OverallCond   1 2.7746e+10 2.8559e+13 67195
## + RoofStyle     5 8.5758e+10 2.8446e+13 67195
## + BsmtUnfSF     1 2.5226e+09 2.8529e+13 67196
## + ScreenPorch   1 1.7580e+09 2.8530e+13 67196
## + Street        1 6.0743e+08 2.8531e+13 67196
## + BsmtQual      4 5.6119e+10 2.8476e+13 67196
## - KitchenQual   3 8.1273e+10 2.8613e+13 67196
## + LotConfig     4 4.6629e+10 2.8485e+13 67197
## + LandSlope     2 6.6429e+09 2.8525e+13 67197
## - BedroomAbvGr  1 5.7815e+10 2.8590e+13 67198
## + ExterQual     3 1.9750e+10 2.8512e+13 67198
## + GarageQual    5 5.5633e+10 2.8476e+13 67198
## + PoolQC        3 1.0790e+10 2.8521e+13 67199
## - MSZoning      4 1.4172e+11 2.8673e+13 67200
## + Condition1    8 7.1810e+10 2.8460e+13 67203
## - OverallQual   1 1.3667e+11 2.8668e+13 67206
## - I(LotArea^2)  1 1.4784e+11 2.8680e+13 67207
## - RoofMatl      7 2.9439e+11 2.8826e+13 67210
```

```
## - X1stFlrSF     1 3.7272e+11 2.8904e+13 67230
## - X2ndFlrSF     1 4.4630e+11 2.8978e+13 67237
```

```
step$anova # display results
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## SalePrice ~ MSZoning + I(LotArea^2) + Street + LotConfig + LandSlope +
##     Condition1 + OverallQual + OverallCond + RoofStyle + RoofMatl +
##     MasVnrArea + ExterQual + BsmtQual + BsmtExposure + BsmtFinSF2 +
##     BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr + KitchenQual +
##     GarageQual + ScreenPorch + PoolArea + PoolQC
##
## Final Model:
## SalePrice ~ MSZoning + I(LotArea^2) + OverallQual + OverallCond +
##     RoofMatl + MasVnrArea + BsmtFinSF2 + X1stFlrSF + X2ndFlrSF +
##     BedroomAbvGr + KitchenQual + PoolArea
##
##
##              Step Df    Deviance Resid. Df   Resid. Dev      AIC
## 1                                     2855 2.812904e+13 67232.39
## 2    - Condition1  8 68394129476      2863 2.819744e+13 67223.48
## 3     - ExterQual  3 14415767073      2866 2.821185e+13 67218.97
## 4        - PoolQC  3 15254644176      2869 2.822711e+13 67214.55
## 5    - GarageQual  5 54502308767      2874 2.828161e+13 67210.18
## 6     - LotConfig  4 41913350536      2878 2.832352e+13 67206.50
## 7     - LandSlope  2  5776560667      2880 2.832930e+13 67203.09
## 8      - BsmtQual  3 29314002129      2883 2.835861e+13 67200.11
## 9        - Street  1   349187666      2884 2.835896e+13 67198.15
## 10    - BsmtUnfSF  1  2023853143      2885 2.836099e+13 67196.36
## 11  - ScreenPorch  1  2852212945      2886 2.836384e+13 67194.65
## 12    - RoofStyle  5 90879867922      2891 2.845472e+13 67193.99
## 13 - BsmtExposure  4 76969434725      2895 2.853169e+13 67193.87
```

```
leaps<-regsubsets(SalePrice~MSZoning+I(LotArea^2)+Street+
                    RoofStyle+RoofMatl+MasVnrArea+ExterQual+BsmtQual+BsmtExpo
sure+BsmtFinSF2 +
                    BsmtUnfSF+X1stFlrSF+X2ndFlrSF+BedroomAbvGr+KitchenQual+Ga
rageQual+
                    ScreenPorch+PoolArea+PoolQC,data=df,nbest=10)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.
in =
## force.in, : 1 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
summary(leaps)
```

```
## Subset selection object
## Call: regsubsets.formula(SalePrice ~ MSZoning + I(LotArea^2) + Street +
##     RoofStyle + RoofMatl + MasVnrArea + ExterQual + BsmtQual +
##     BsmtExposure + BsmtFinSF2 + BsmtUnfSF + X1stFlrSF + X2ndFlrSF +
##     BedroomAbvGr + KitchenQual + GarageQual + ScreenPorch + PoolArea +
##     PoolQC, data = df, nbest = 10)
## 48 Variables  (and intercept)
##                       Forced in Forced out
## MSZoningFV                FALSE      FALSE
## MSZoningRH                FALSE      FALSE
## MSZoningRL                FALSE      FALSE
## MSZoningRM                FALSE      FALSE
## I(LotArea^2)              FALSE      FALSE
## StreetPave                FALSE      FALSE
## RoofStyleGable            FALSE      FALSE
## RoofStyleGambrel          FALSE      FALSE
## RoofStyleHip              FALSE      FALSE
## RoofStyleMansard          FALSE      FALSE
## RoofStyleShed             FALSE      FALSE
## RoofMatlCompShg           FALSE      FALSE
## RoofMatlMembran           FALSE      FALSE
## RoofMatlMetal             FALSE      FALSE
## RoofMatlRoll              FALSE      FALSE
## RoofMatlTar&Grv           FALSE      FALSE
## RoofMatlWdShake           FALSE      FALSE
## RoofMatlWdShngl           FALSE      FALSE
## MasVnrArea                FALSE      FALSE
## ExterQualFa               FALSE      FALSE
## ExterQualGd               FALSE      FALSE
## ExterQualTA               FALSE      FALSE
## BsmtQualFa                FALSE      FALSE
## BsmtQualGd                FALSE      FALSE
## BsmtQualNoBasement        FALSE      FALSE
## BsmtQualTA                FALSE      FALSE
## BsmtExposureGd            FALSE      FALSE
## BsmtExposureMn            FALSE      FALSE
## BsmtExposureNo            FALSE      FALSE
## BsmtFinSF2                FALSE      FALSE
## BsmtUnfSF                 FALSE      FALSE
## X1stFlrSF                 FALSE      FALSE
## X2ndFlrSF                 FALSE      FALSE
## BedroomAbvGr              FALSE      FALSE
## KitchenQualFa             FALSE      FALSE
## KitchenQualGd             FALSE      FALSE
## KitchenQualTA             FALSE      FALSE
## GarageQualFa              FALSE      FALSE
## GarageQualGd              FALSE      FALSE
## GarageQualNoGarage        FALSE      FALSE
## GarageQualPo              FALSE      FALSE
## GarageQualTA              FALSE      FALSE
```

```
## ScreenPorch                FALSE      FALSE
## PoolArea                   FALSE      FALSE
## PoolQCFa                   FALSE      FALSE
## PoolQCGd                   FALSE      FALSE
## PoolQCNoPool               FALSE      FALSE
## BsmtExposureNoBasement     FALSE      FALSE
## 10 subsets of each size up to 9
## Selection Algorithm: exhaustive
##           MSZoningFV MSZoningRH MSZoningRL MSZoningRM I(LotArea^2) StreetP
ave
## 1  ( 1 )  " "        " "        " "        " "        " "           " "
## 1  ( 2 )  " "        " "        " "        " "        " "           " "
## 1  ( 3 )  " "        " "        " "        " "        " "           " "
.

.
## 9  ( 8 )  " "        " "        "*"        " "        "*"           " "
## 9  ( 9 )  " "        " "        "*"        " "        "*"           " "
## 9  ( 10 ) " "        " "        "*"        " "        "*"           " "
##           RoofStyleGable RoofStyleGambrel RoofStyleHip RoofStyleMansard
## 1  ( 1 )  " "             " "               " "           " "
## 1  ( 2 )  " "             " "               " "           " "
## 1  ( 3 )  " "             " "               " "           " "
## 1  ( 4 )  " "             " "               " "           " "
.

.
## 9  ( 7 )  "*"             " "               " "           " "
## 9  ( 8 )  "*"             " "               " "           " "
## 9  ( 9 )  " "             " "               "*"           " "
## 9  ( 10 ) " "             " "               "*"           " "
##           RoofStyleShed RoofMatlCompShg RoofMatlMembran RoofMatlMetal
## 1  ( 1 )  " "           " "             " "             " "
## 1  ( 2 )  " "           " "             " "             " "
## 1  ( 3 )  " "           " "             " "             " "
.

.
## 9  ( 7 )  " "           " "             " "             " "
## 9  ( 8 )  " "           " "             " "             " "
## 9  ( 9 )  " "           " "             " "             " "
## 9  ( 10 ) " "           " "             " "             " "
##           RoofMatlRoll RoofMatlTar&Grv RoofMatlWdShake RoofMatlWdShngl
## 1  ( 1 )  " "          " "             " "             " "
## 1  ( 2 )  " "          " "             " "             " "
## 1  ( 3 )  " "          " "             " "             " "
## 1  ( 4 )  " "          " "             " "             " "
.

.
## 9  ( 7 )  " "          " "             " "             "*"
```

```
## 9  ( 8 )  " "          " "             " "             "*"
## 9  ( 9 )  " "          " "             " "             "*"
## 9  ( 10 ) " "          " "             " "             "*"
##           MasVnrArea ExterQualFa ExterQualGd ExterQualTA BsmtQualFa BsmtQu
alGd
## 1  ( 1 )  " "        " "         " "         " "         " "        " "
## 1  ( 2 )  " "        " "         " "         "*"         " "        " "
## 1  ( 3 )  " "        " "         " "         " "         " "        " "
## 1  ( 4 )  "*"        " "         " "         " "         " "        " "
## 1  ( 5 )  " "        " "         "*"         " "         " "        " "
## 1  ( 6 )  " "        " "         " "         " "         " "        " "
.

.
## 9  ( 3 )  "*"        " "         " "         " "         " "        " "
## 9  ( 4 )  " "        "*"         "*"         "*"         " "        " "
## 9  ( 5 )  "*"        " "         " "         " "         " "        " "
## 9  ( 6 )  " "        " "         " "         " "         " "        " "
## 9  ( 7 )  " "        " "         " "         " "         " "        " "
## 9  ( 8 )  " "        " "         " "         " "         " "        " "
## 9  ( 9 )  " "        " "         " "         " "         " "        " "
## 9  ( 10 ) " "        " "         " "         " "         " "        " "
##           BsmtQualNoBasement BsmtQualTA BsmtExposureGd BsmtExposureMn
## 1  ( 1 )  " "                " "        " "            " "
## 1  ( 2 )  " "                " "        " "            " "
## 1  ( 3 )  " "                " "        " "            " "
## 1  ( 4 )  " "                " "        " "            " "
## 1  ( 5 )  " "                " "        " "            " "
.

.
## 9  ( 9 )  " "                " "        " "            " "
## 9  ( 10 ) "*"                " "        " "            " "
##           BsmtExposureNo BsmtExposureNoBasement BsmtFinSF2 BsmtUnfSF X1stF
lrSF
## 1  ( 1 )  " "            " "                  " "        " "       "*"
## 1  ( 2 )  " "            " "                  " "        " "       " "
## 1  ( 3 )  " "            " "                  " "        " "       " "
## 1  ( 4 )  " "            " "                  " "        " "       " "
## 1  ( 5 )  " "            " "                  " "        " "       " "
.
## 9  ( 6 )  " "            " "                  " "        " "       "*"
## 9  ( 7 )  " "            "*"                  " "        " "       "*"
## 9  ( 8 )  " "            " "                  " "        " "       "*"
## 9  ( 9 )  " "            "*"                  " "        " "       "*"
## 9  ( 10 ) " "            " "                  " "        " "       "*"
##           X2ndFlrSF BedroomAbvGr KitchenQualFa KitchenQualGd KitchenQualTA
## 1  ( 1 )  " "       " "          " "           " "           " "
## 1  ( 2 )  " "       " "          " "           " "           " "
```

```
## 1  ( 3 )  " "       " "          " "           " "           "*"
## 1  ( 4 )  " "       " "          " "           " "           "
.
.
## 9  ( 5 )  "*"       "*"          " "           " "           "*"
## 9  ( 6 )  "*"       "*"          " "           " "           "*"
## 9  ( 7 )  "*"       "*"          " "           " "           "*"
## 9  ( 8 )  "*"       "*"          " "           " "           "*"
## 9  ( 9 )  "*"       "*"          " "           " "           "*"
## 9  ( 10 ) "*"       "*"          " "           " "           "*"
##           GarageQualFa GarageQualGd GarageQualNoGarage GarageQualPo
## 1  ( 1 )  " "          " "          " "                " "
## 1  ( 2 )  " "          " "          " "                " "
## 1  ( 3 )  " "          " "          " "                " "
## 1  ( 4 )  " "          " "          " "                " "
.
.
## 9  ( 6 )  "*"          " "          " "                " "
## 9  ( 7 )  " "          " "          " "                " "
## 9  ( 8 )  " "          " "          " "                " "
## 9  ( 9 )  " "          " "          " "                " "
## 9  ( 10 ) " "          " "          " "                " "
##           GarageQualTA ScreenPorch PoolArea PoolQCFa PoolQCGd PoolQCNoPool
## 1  ( 1 )  " "          " "         " "      " "      " "      " "
## 1  ( 2 )  " "          " "         " "      " "      " "      " "
## 1  ( 3 )  " "          " "         " "      " "      " "      " "
## 1  ( 4 )  " "          " "         " "      " "      " "      " "
.
.
## 9  ( 5 )  " "          " "         " "      " "      " "      " "
## 9  ( 6 )  " "          " "         " "      " "      " "      " "
## 9  ( 7 )  " "          " "         " "      " "      " "      " "
## 9  ( 8 )  " "          " "         " "      " "      " "      " "
## 9  ( 9 )  " "          " "         " "      " "      " "      " "
## 9  ( 10 ) " "          " "         " "      " "      " "      " "
```