House Prices Lasso Regression

Autumn Brinkerhoff

2024-01-05

##Loading libraries

```
library(knitr)
library(ggplot2)
library(plyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

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

```
library(scales)
library(Rmisc)
library(ggrepel)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:randomForest':
##
##     outlier

## The following objects are masked from 'package:scales':
##
##     alpha, rescale

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
##library(xgboost)
```

##Reading the data

```
train <- read.csv("train.csv")
test <- read.csv("test.csv")

dim(train)
```

```
## [1] 1460   81
```

```
dim(test)
```

```
## [1] 1459   80
```

##Data structure

```
test_labels <- test$Id
test$Id <- NULL
train$Id <- NULL

test$SalePrice <- NA
df <- rbind(train, test)
dim(df)
```

```
## [1] 2919   80
```

```
head(df)
```

```
##   MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
## 5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
```

```
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
## 4       961       756            0      1717            1            0        1
## 5      1145      1053            0      2198            1            0        2
## 6       796       566            0      1362            1            0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
## 4        0            3            1          Gd            7        Typ
## 5        1            4            1          Gd            9        Typ
## 6        1            1            1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0       <NA>     Attchd        2003          RFn          2
## 2          1         TA     Attchd        1976          RFn          2
## 3          1         TA     Attchd        2001          RFn          2
## 4          1         Gd     Detchd        1998          Unf          3
## 5          1         TA     Attchd        2000          RFn          3
## 6          0       <NA>     Attchd        1993          Unf          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
## 4        642         TA         TA          Y          0          35
## 5        836         TA         TA          Y        192          84
## 6        480         TA         TA          Y         40          30
##   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>  <NA>        <NA>
## 4           272          0           0        0   <NA>  <NA>        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0        320           0        0   <NA> MnPrv        Shed
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
## 4       0      2   2006       WD       Abnorml    140000
## 5       0     12   2008       WD        Normal    250000
## 6     700     10   2009       WD        Normal    143000
```

```
summary(df)
```

```
##    MSSubClass       MSZoning          LotFrontage        LotArea
```

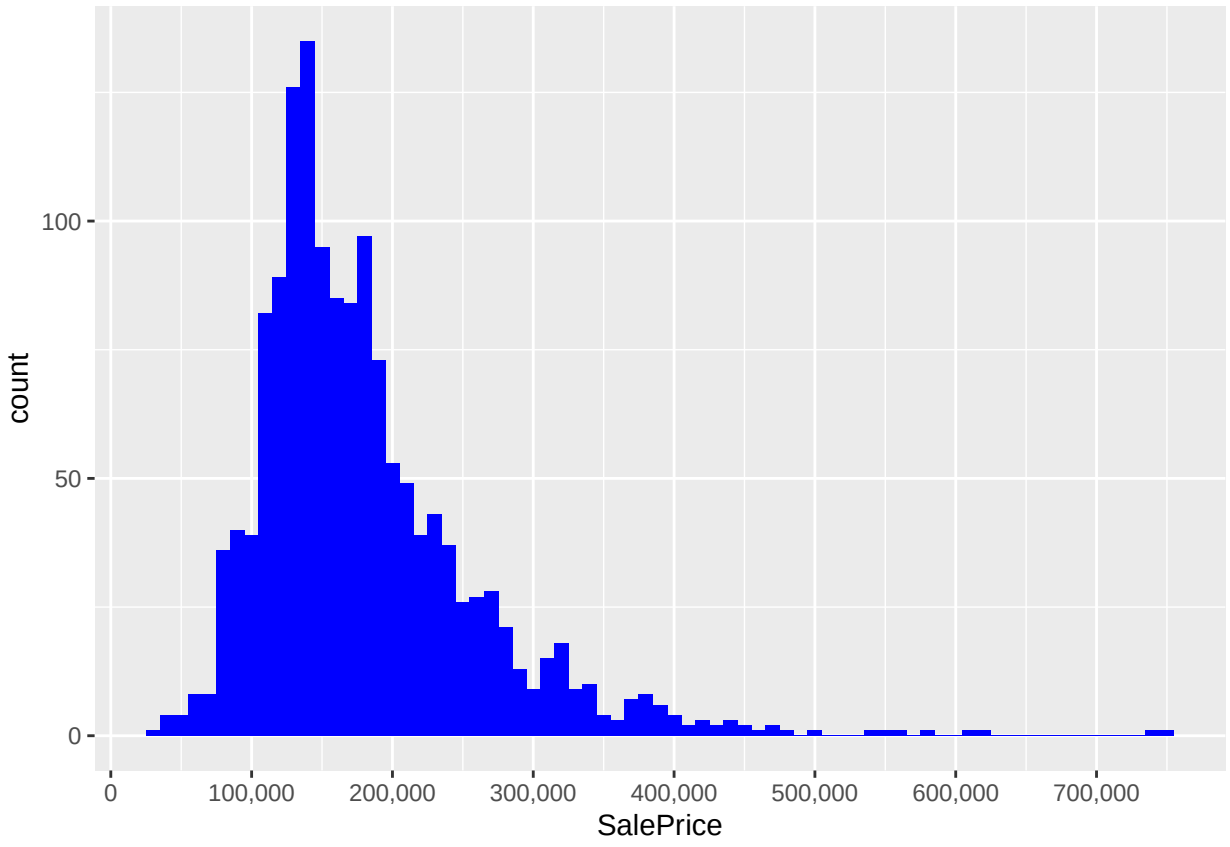
```
##  Min.   : 20.00   Length:2919        Min.   : 21.00   Min.   :  1300
##  1st Qu.: 20.00   Class :character   1st Qu.: 59.00   1st Qu.:  7478
##  Median : 50.00   Mode  :character   Median : 68.00   Median :  9453
##  Mean   : 57.14                      Mean   : 69.31   Mean   : 10168
##  3rd Qu.: 70.00                      3rd Qu.: 80.00   3rd Qu.: 11570
##  Max.   :190.00                      Max.   :313.00   Max.   :215245
##                                      NA's   :486
##     Street              Alley             LotShape         LandContour
##  Length:2919        Length:2919        Length:2919        Length:2919
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Utilities          LotConfig          LandSlope         Neighborhood
##  Length:2919        Length:2919        Length:2919        Length:2919
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Condition1         Condition2          BldgType          HouseStyle
##  Length:2919        Length:2919        Length:2919        Length:2919
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   OverallQual      OverallCond      YearBuilt     YearRemodAdd
##  Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :1950
##  1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954   1st Qu.:1965
##  Median : 6.000   Median :5.000   Median :1973   Median :1993
##  Mean   : 6.089   Mean   :5.565   Mean   :1971   Mean   :1984
##  3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004
##  Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010
##
##   RoofStyle           RoofMatl         Exterior1st        Exterior2nd
##  Length:2919        Length:2919        Length:2919        Length:2919
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   MasVnrType          MasVnrArea      ExterQual          ExterCond
##  Length:2919        Min.   :   0.0   Length:2919        Length:2919
##  Class :character   1st Qu.:   0.0   Class :character   Class :character
##  Mode  :character   Median :   0.0   Mode  :character   Mode  :character
##                     Mean   : 102.2
##                     3rd Qu.: 164.0
##                     Max.   :1600.0
```

```
##                    NA's   :23
##   Foundation          BsmtQual           BsmtCond         BsmtExposure
##  Length:2919        Length:2919        Length:2919        Length:2919
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtFinType1         BsmtFinSF1     BsmtFinType2         BsmtFinSF2
##  Length:2919        Min.   :   0.0   Length:2919        Min.   :   0.00
##  Class :character   1st Qu.:   0.0   Class :character   1st Qu.:   0.00
##  Mode  :character   Median : 368.5   Mode  :character   Median :   0.00
##                     Mean   : 441.4                      Mean   :  49.58
##                     3rd Qu.: 733.0                      3rd Qu.:   0.00
##                     Max.   :5644.0                      Max.   :1526.00
##                     NA's   :1                           NA's   :1
##    BsmtUnfSF       TotalBsmtSF       Heating           HeatingQC
##  Min.   :   0.0   Min.   :   0.0   Length:2919        Length:2919
##  1st Qu.: 220.0   1st Qu.: 793.0   Class :character   Class :character
##  Median : 467.0   Median : 989.5   Mode  :character   Mode  :character
##  Mean   : 560.8   Mean   :1051.8
##  3rd Qu.: 805.5   3rd Qu.:1302.0
##  Max.   :2336.0   Max.   :6110.0
##  NA's   :1        NA's   :1
##   CentralAir         Electrical          X1stFlrSF      X2ndFlrSF
##  Length:2919        Length:2919        Min.   : 334   Min.   :   0.0
##  Class :character   Class :character   1st Qu.: 876   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :1082   Median :   0.0
##                                        Mean   :1160   Mean   : 336.5
##                                        3rd Qu.:1388   3rd Qu.: 704.0
##                                        Max.   :5095   Max.   :2065.0
##
##   LowQualFinSF        GrLivArea     BsmtFullBath     BsmtHalfBath
##  Min.   :   0.000   Min.   : 334   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:   0.000   1st Qu.:1126   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :   0.000   Median :1444   Median :0.0000   Median :0.00000
##  Mean   :   4.694   Mean   :1501   Mean   :0.4299   Mean   :0.06136
##  3rd Qu.:   0.000   3rd Qu.:1744   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1064.000   Max.   :5642   Max.   :3.0000   Max.   :2.00000
##                                    NA's   :2        NA's   :2
##     FullBath        HalfBath       BedroomAbvGr   KitchenAbvGr
##  Min.   :0.000   Min.   :0.0000   Min.   :0.00   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.00   1st Qu.:1.000
##  Median :2.000   Median :0.0000   Median :3.00   Median :1.000
##  Mean   :1.568   Mean   :0.3803   Mean   :2.86   Mean   :1.045
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.00   3rd Qu.:1.000
##  Max.   :4.000   Max.   :2.0000   Max.   :8.00   Max.   :3.000
##
##  KitchenQual         TotRmsAbvGrd     Functional          Fireplaces
##  Length:2919        Min.   : 2.000   Length:2919        Min.   :0.0000
##  Class :character   1st Qu.: 5.000   Class :character   1st Qu.:0.0000
##  Mode  :character   Median : 6.000   Mode  :character   Median :1.0000
##                     Mean   : 6.452                      Mean   :0.5971
```

```
##                    3rd Qu.: 7.000                      3rd Qu.:1.0000
##                    Max.   :15.000                      Max.   :4.0000
##
##  FireplaceQu         GarageType         GarageYrBlt   GarageFinish
##  Length:2919        Length:2919        Min.   :1895   Length:2919
##  Class :character   Class :character   1st Qu.:1960   Class :character
##  Mode  :character   Mode  :character   Median :1979   Mode  :character
##                                        Mean   :1978
##                                        3rd Qu.:2002
##                                        Max.   :2207
##                                        NA's   :159
##    GarageCars      GarageArea      GarageQual         GarageCond
##  Min.   :0.000   Min.   :   0.0   Length:2919        Length:2919
##  1st Qu.:1.000   1st Qu.: 320.0   Class :character   Class :character
##  Median :2.000   Median : 480.0   Mode  :character   Mode  :character
##  Mean   :1.767   Mean   : 472.9
##  3rd Qu.:2.000   3rd Qu.: 576.0
##  Max.   :5.000   Max.   :1488.0
##  NA's   :1       NA's   :1
##   PavedDrive          WoodDeckSF       OpenPorchSF     EnclosedPorch
##  Length:2919        Min.   :   0.00   Min.   :  0.00   Min.   :   0.0
##  Class :character   1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.0
##  Mode  :character   Median :   0.00   Median : 26.00   Median :   0.0
##                     Mean   :  93.71   Mean   : 47.49   Mean   :  23.1
##                     3rd Qu.: 168.00   3rd Qu.: 70.00   3rd Qu.:   0.0
##                     Max.   :1424.00   Max.   :742.00   Max.   :1012.0
##
##    X3SsnPorch       ScreenPorch        PoolArea          PoolQC
##  Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   Length:2919
##  1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000   Class :character
##  Median :  0.000   Median :  0.00   Median :  0.000   Mode  :character
##  Mean   :  2.602   Mean   : 16.06   Mean   :  2.252
##  3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :508.000   Max.   :576.00   Max.   :800.000
##
##     Fence           MiscFeature          MiscVal             MoSold
##  Length:2919        Length:2919        Min.   :    0.00   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:    0.00   1st Qu.: 4.000
##  Mode  :character   Mode  :character   Median :    0.00   Median : 6.000
##                                        Mean   :   50.83   Mean   : 6.213
##                                        3rd Qu.:    0.00   3rd Qu.: 8.000
##                                        Max.   :17000.00   Max.   :12.000
##
##      YrSold       SaleType         SaleCondition        SalePrice
##  Min.   :2006   Length:2919        Length:2919        Min.   : 34900
##  1st Qu.:2007   Class :character   Class :character   1st Qu.:129975
##  Median :2008   Mode  :character   Mode  :character   Median :163000
##  Mean   :2008                                         Mean   :180921
##  3rd Qu.:2009                                         3rd Qu.:214000
##  Max.   :2010                                         Max.   :755000
##                                                       NA's   :1459
```

```
ggplot(data=df[!is.na(df$SalePrice),], aes(x=SalePrice)) +
        geom_histogram(fill="blue", binwidth = 10000) +
```

```
        scale_x_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```



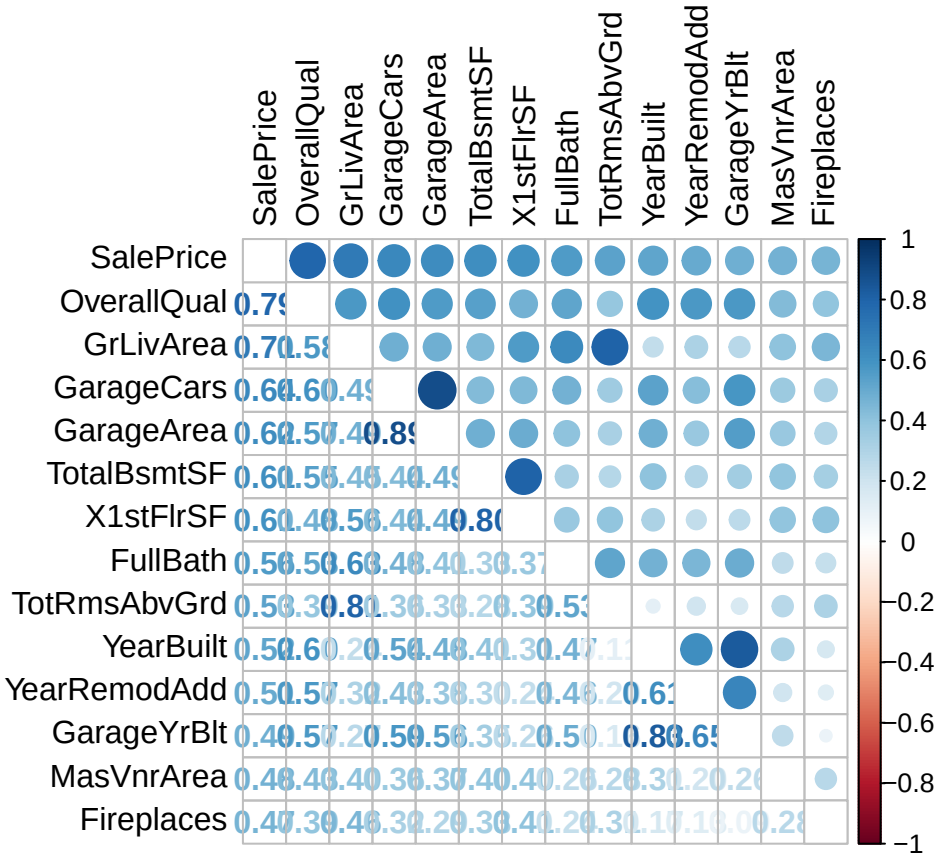

```
summary(df$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   34900  129975  163000  180921  214000  755000    1459
```

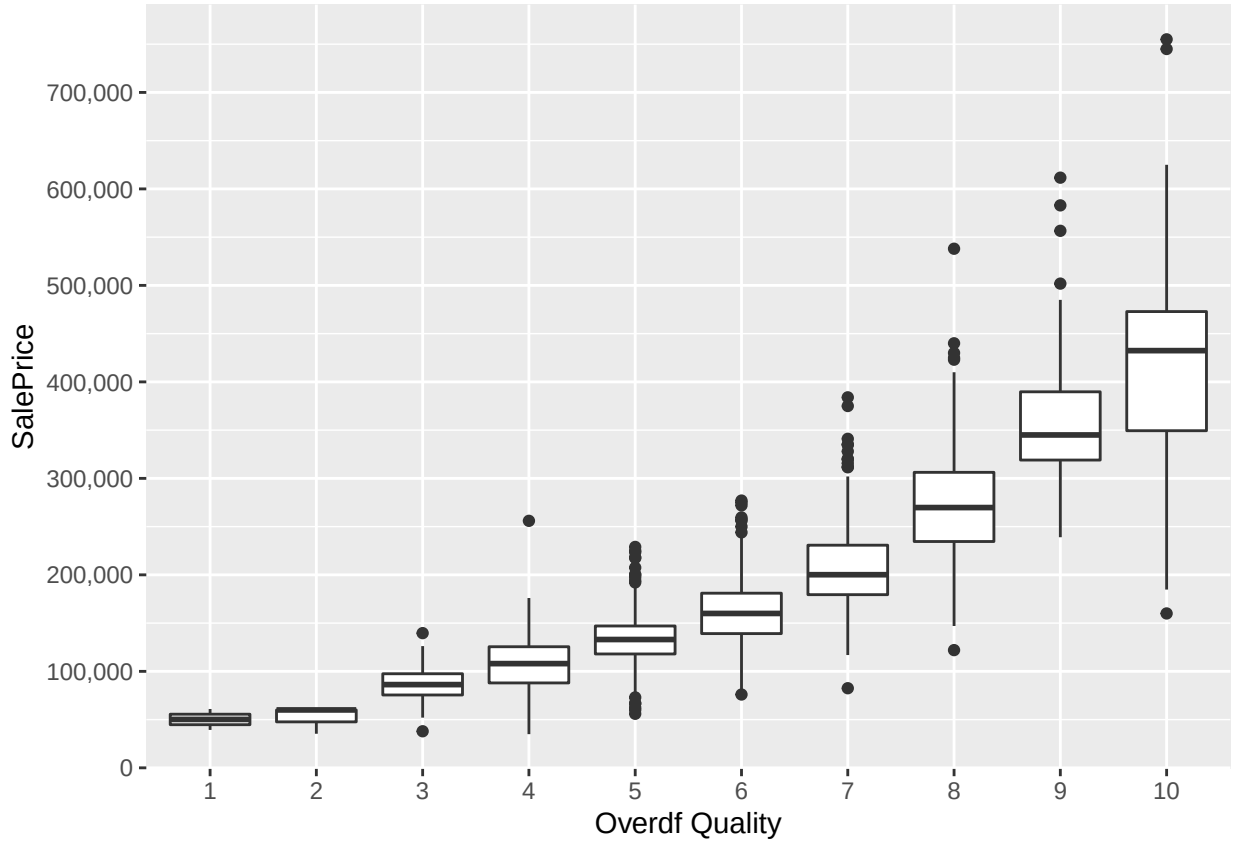
```
numericVars <- which(sapply(df, is.numeric)) #index vector numeric variables
numericVarNames <- names(numericVars) #saving names vector for use later on
cat('There are', length(numericVars), 'numeric variables')
```

```
## There are 37 numeric variables
```

```
df_numVar <- df[, numericVars]
cor_numVar <- cor(df_numVar, use="pairwise.complete.obs") #correlations of df numeric variables with NA


cor_sorted <- as.matrix(sort(cor_numVar[,'SalePrice'], decreasing = TRUE))

CorHigh <- names(which(apply(cor_sorted, 1, function(x) abs(x)>0.4)))
cor_numVar <- cor_numVar[CorHigh, CorHigh]

corrplot.mixed(cor_numVar, tl.col="black", tl.pos = "lt")
```

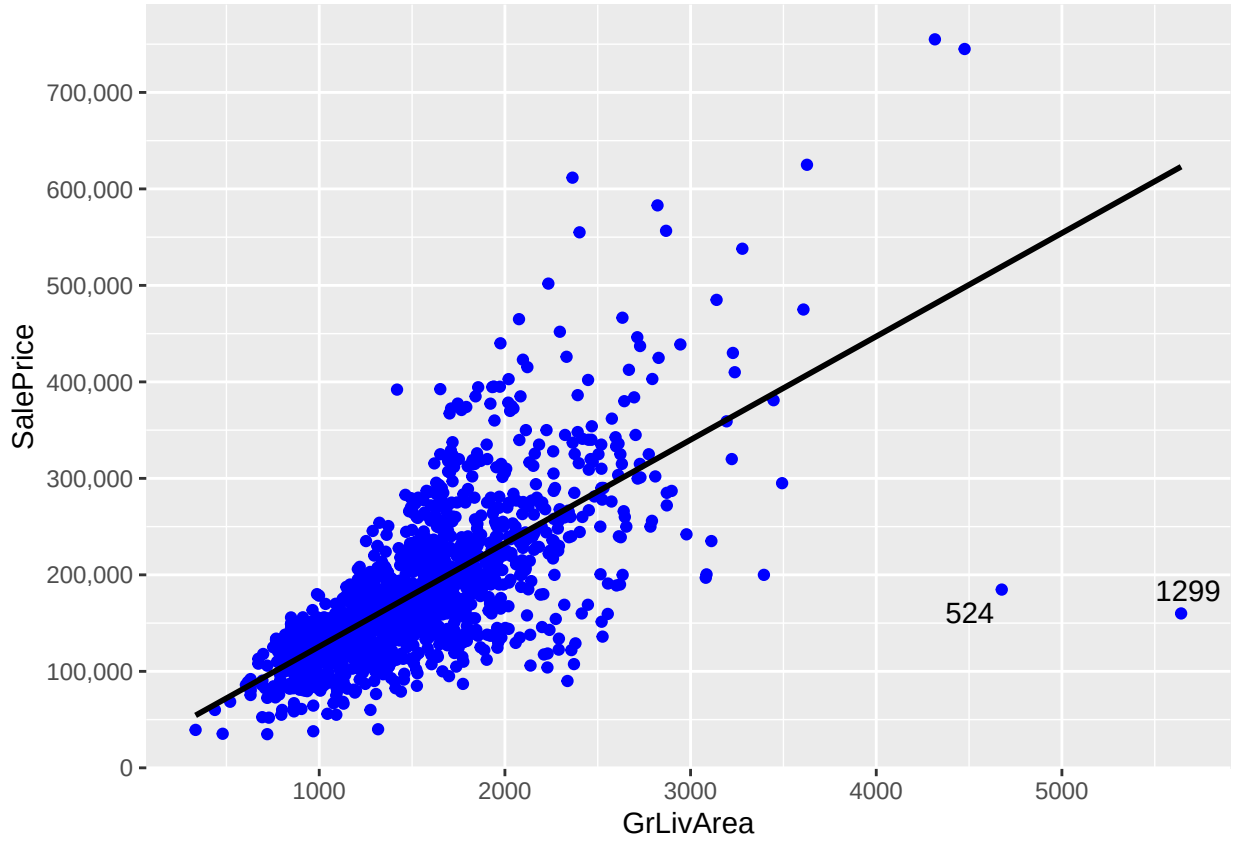



```
ggplot(data=df[!is.na(df$SalePrice),], aes(x=factor(OverallQual), y=SalePrice))+
        geom_boxplot() + labs(x='Overdf Quality') +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```




```
ggplot(data=df[!is.na(df$SalePrice),], aes(x=GrLivArea, y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma) +
        geom_text_repel(aes(label = ifelse(df$GrLivArea[!is.na(df$SalePrice)]>4500, rownames(df), '')))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
df[c(524, 1299), c('SalePrice', 'GrLivArea', 'OverallQual')]
```

```
##      SalePrice GrLivArea OverallQual
## 524     184750      4676          10
## 1299    160000      5642          10
```

```
NAcol <- which(colSums(is.na(df)) > 0)
sort(colSums(sapply(df[NAcol], is.na)), decreasing = TRUE)
```

```
##       PoolQC  MiscFeature        Alley        Fence    SalePrice  FireplaceQu
##         2909         2814         2721         2348         1459         1420
##  LotFrontage  GarageYrBlt GarageFinish   GarageQual   GarageCond   GarageType
##          486          159          159          159          159          157
##     BsmtCond BsmtExposure     BsmtQual BsmtFinType2 BsmtFinType1   MasVnrType
##           82           82           81           80           79           24
##   MasVnrArea     MSZoning    Utilities BsmtFullBath BsmtHalfBath   Functional
##           23            4            2            2            2            2
##  Exterior1st  Exterior2nd   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            1            1            1            1            1            1
##   Electrical  KitchenQual   GarageCars   GarageArea     SaleType
##            1            1            1            1            1
```

##missing data

```
df$PoolQC[is.na(df$PoolQC)] <- 'None'

Qualities <- c('None' = 0, 'Po' = 1, 'Fa' = 2, 'TA' = 3, 'Gd' = 4, 'Ex' = 5)

df$PoolQC<-as.integer(revalue(df$PoolQC, Qualities))
```

```
## The following `from` values were not present in `x`: Po, TA
```

```
table(df$PoolQC)
```

```
##
##    0    2    4    5
## 2909    2    4    4
```

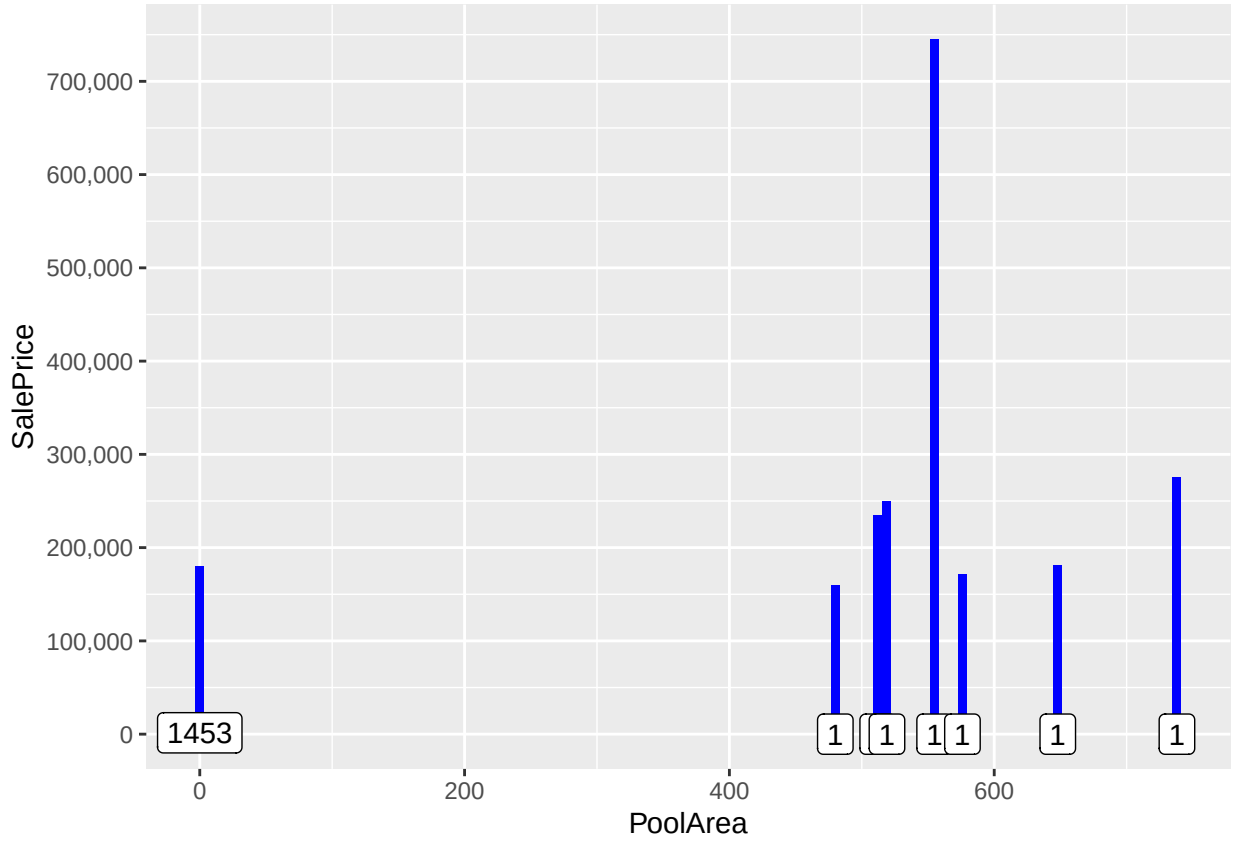
```
df[df$PoolArea>0 & df$PoolQC==0, c('PoolArea', 'PoolQC', 'OverallQual')]
```

```
##      PoolArea PoolQC OverallQual
## 2421      368      0           4
## 2504      444      0           6
## 2600      561      0           3
```

```
df$PoolQC[2421] <- 2
df$PoolQC[2504] <- 3
df$PoolQC[2600] <- 2


ggplot(df[!is.na(df$SalePrice),], aes(x=PoolArea, y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..))
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```

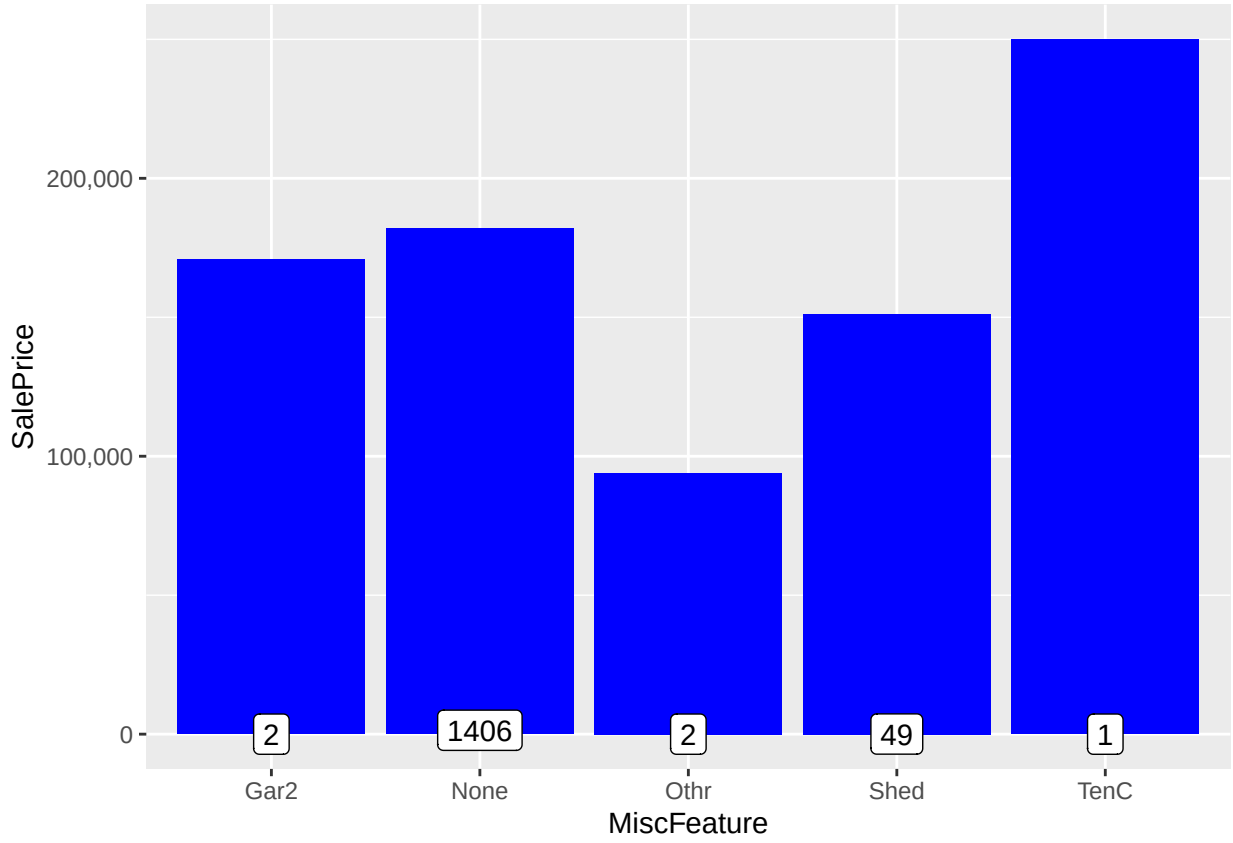


##Miscellaneous Feature

```
df$MiscFeature[is.na(df$MiscFeature)] <- 'None'
df$MiscFeature <- as.factor(df$MiscFeature)

ggplot(df[!is.na(df$SalePrice),], aes(x=MiscFeature, y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..))
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
table(df$MiscFeature)
```

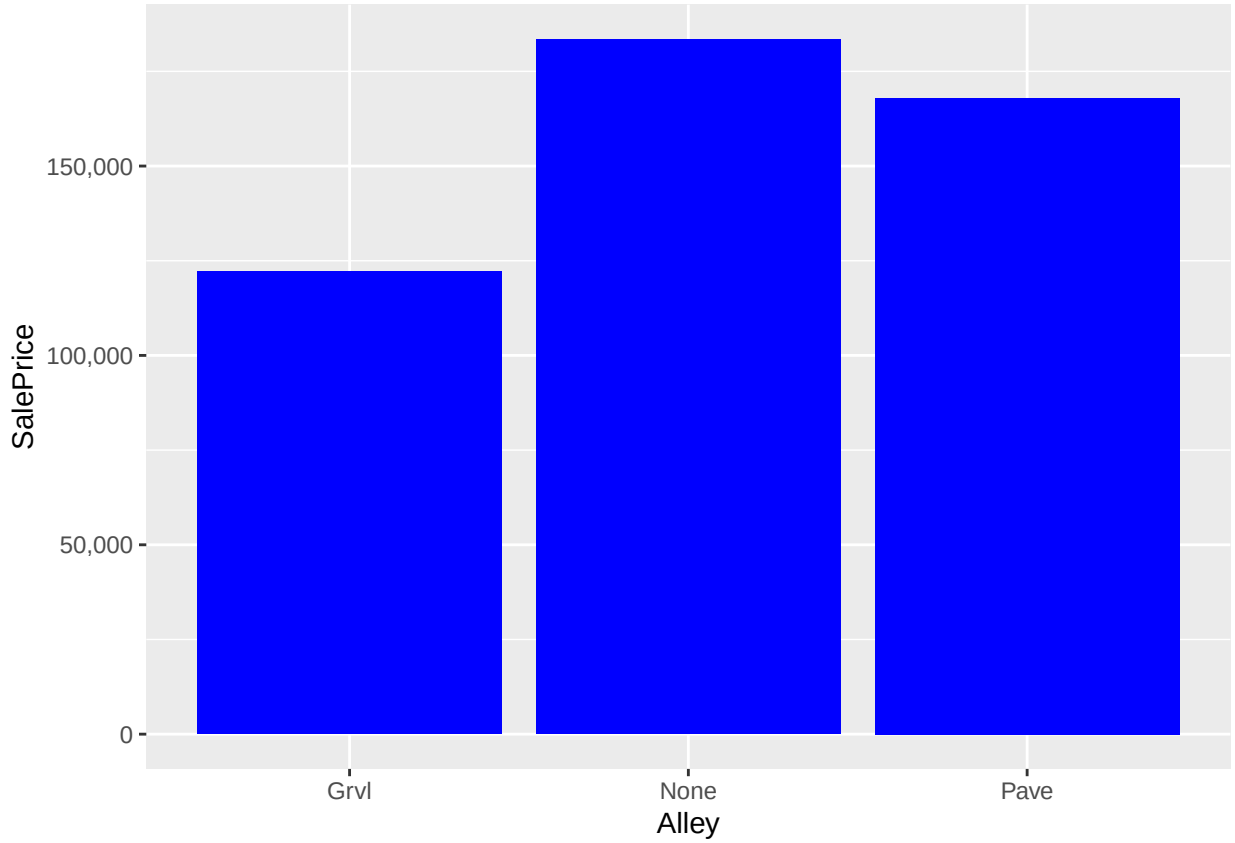
```
##
## Gar2 None Othr Shed TenC
##    5 2814    4   95    1
```

##Alley

```
df$Alley[is.na(df$Alley)] <- 'None'
df$Alley <- as.factor(df$Alley)

ggplot(df[!is.na(df$SalePrice),], aes(x=Alley, y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue')+
        scale_y_continuous(breaks= seq(0, 200000, by=50000), labels = comma)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
table(df$Alley)
```

```
##
## Grvl None Pave
##  120 2721   78
```

##FireplaceQu

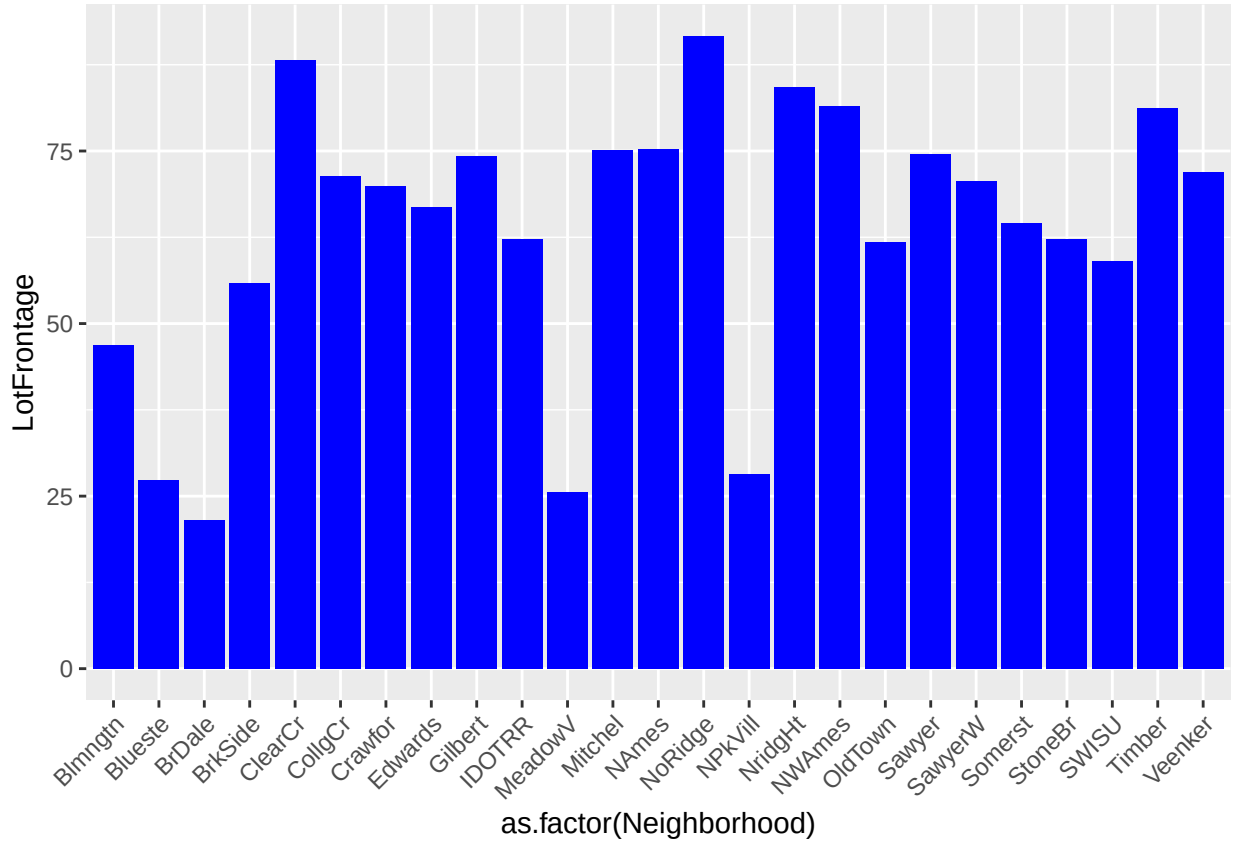
```
df$FireplaceQu[is.na(df$FireplaceQu)] <- 'None'
df$FireplaceQu<-as.integer(revalue(df$FireplaceQu, Qualities))
table(df$FireplaceQu)
```

```
##
##    0    1    2    3    4    5
## 1420   46   74  592  744   43
```

```
ggplot(df[!is.na(df$LotFrontage),], aes(x=as.factor(Neighborhood), y=LotFrontage)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
for (i in 1:nrow(df)){
        if(is.na(df$LotFrontage[i])){
                df$LotFrontage[i] <- as.integer(median(df$LotFrontage[df$Neighborhood==df$Neighborhood[i]
        }
}


df$LotShape<-as.integer(revalue(df$LotShape, c('IR3'=0, 'IR2'=1, 'IR1'=2, 'Reg'=3)))
table(df$LotShape)
```

```
##
##    0    1    2    3
##   16   76  968 1859
```

```
length(which(is.na(df$GarageType) & is.na(df$GarageFinish) & is.na(df$GarageCond) & is.na(df$GarageQual
```

```
## [1] 157
```

```
kable(df[!is.na(df$GarageType) & is.na(df$GarageFinish), c('GarageCars', 'GarageArea', 'GarageType', 'Ga
```

	GarageCars	GarageArea	GarageType	GarageCond	GarageQual	GarageFinish
2127	1	360	Detchd	NA	NA	NA

	GarageCars	GarageArea	GarageType	GarageCond	GarageQual	GarageFinish
2577	NA	NA	Detchd	NA	NA	NA

```
df$GarageCond[2127] <- names(sort(-table(df$GarageCond)))[1]
df$GarageQual[2127] <- names(sort(-table(df$GarageQual)))[1]
df$GarageFinish[2127] <- names(sort(-table(df$GarageFinish)))[1]

#display "fixed" house
kable(df[2127, c('GarageYrBlt', 'GarageCars', 'GarageArea', 'GarageType', 'GarageCond', 'GarageQual', '
```

	GarageYrBlt	GarageCars	GarageArea	GarageType	GarageCond	GarageQual	GarageFinish
2127	NA	1	360	Detchd	TA	TA	Unf

```
#fixing 3 values for house 2577
df$GarageCars[2577] <- 0
df$GarageArea[2577] <- 0
df$GarageType[2577] <- NA

#check if NAs of the character variables are now df 158
length(which(is.na(df$GarageType) & is.na(df$GarageFinish) & is.na(df$GarageCond) & is.na(df$GarageQual)
```

```
## [1] 158
```

```
df$GarageType[is.na(df$GarageType)] <- 'No Garage'
df$GarageType <- as.factor(df$GarageType)
table(df$GarageType)
```

```
##
##    2Types    Attchd   Basment   BuiltIn   CarPort    Detchd No Garage
##        23      1723        36       186        15       778       158
```

```
df$GarageFinish[is.na(df$GarageFinish)] <- 'None'
Finish <- c('None'=0, 'Unf'=1, 'RFn'=2, 'Fin'=3)

df$GarageFinish<-as.integer(revalue(df$GarageFinish, Finish))
table(df$GarageFinish)
```

```
##
##    0    1    2    3
##  158 1231  811  719
```

```
df$GarageQual[is.na(df$GarageQual)] <- 'None'
df$GarageQual<-as.integer(revalue(df$GarageQual, Qualities))
table(df$GarageQual)
```

```
##
##    0    1    2    3    4    5
##  158    5  124 2605   24    3
```

```
df$GarageCond[is.na(df$GarageCond)] <- 'None'
df$GarageCond<-as.integer(revalue(df$GarageCond, Qualities))
table(df$GarageCond)
```

```
##
##    0    1    2    3    4    5
##  158   14   74 2655   15    3
```

##Basement

```
length(which(is.na(df$BsmtQual) & is.na(df$BsmtCond) & is.na(df$BsmtExposure) & is.na(df$BsmtFinType1) &
```

```
## [1] 79
```

```
#Find the additional NAs; BsmtFinType1 is the one with 79 NAs
df[!is.na(df$BsmtFinType1) & (is.na(df$BsmtCond)|is.na(df$BsmtQual)|is.na(df$BsmtExposure)|is.na(df$Bsmt
```

```
##      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2
## 333        Gd       TA           No          GLQ         <NA>
## 949        Gd       TA         <NA>          Unf          Unf
## 1488       Gd       TA         <NA>          Unf          Unf
## 2041       Gd     <NA>           Mn          GLQ          Rec
## 2186       TA     <NA>           No          BLQ          Unf
## 2218     <NA>       Fa           No          Unf          Unf
## 2219     <NA>       TA           No          Unf          Unf
## 2349       Gd       TA         <NA>          Unf          Unf
## 2525       TA     <NA>           Av          ALQ          Unf
```

```
#Imputing modes.
df$BsmtFinType2[333] <- names(sort(-table(df$BsmtFinType2)))[1]
df$BsmtExposure[c(949, 1488, 2349)] <- names(sort(-table(df$BsmtExposure)))[1]
df$BsmtCond[c(2041, 2186, 2525)] <- names(sort(-table(df$BsmtCond)))[1]
df$BsmtQual[c(2218, 2219)] <- names(sort(-table(df$BsmtQual)))[1]
```

```
df$BsmtQual[is.na(df$BsmtQual)] <- 'None'
df$BsmtQual<-as.integer(revalue(df$BsmtQual, Qualities))
```

```
## The following `from` values were not present in `x`: Po
```

```
table(df$BsmtQual)
```

```
##
##    0    2    3    4    5
##   79   88 1285 1209  258
```

```
df$BsmtCond[is.na(df$BsmtCond)] <- 'None'
df$BsmtCond<-as.integer(revalue(df$BsmtCond, Qualities))
```

```
## The following `from` values were not present in `x`: Ex
```

```
table(df$BsmtCond)
```

```
##
##    0    1    2    3    4
##   79    5  104 2609  122
```

```
df$BsmtExposure[is.na(df$BsmtExposure)] <- 'None'
Exposure <- c('None'=0, 'No'=1, 'Mn'=2, 'Av'=3, 'Gd'=4)

df$BsmtExposure<-as.integer(revalue(df$BsmtExposure, Exposure))
table(df$BsmtExposure)
```

```
##
##    0    1    2    3    4
##   79 1907  239  418  276
```

```
df$BsmtFinType1[is.na(df$BsmtFinType1)] <- 'None'
FinType <- c('None'=0, 'Unf'=1, 'LwQ'=2, 'Rec'=3, 'BLQ'=4, 'ALQ'=5, 'GLQ'=6)

df$BsmtFinType1<-as.integer(revalue(df$BsmtFinType1, FinType))
table(df$BsmtFinType1)
```

```
##
##   0   1   2   3   4   5   6
##  79 851 154 288 269 429 849
```

```
df$BsmtFinType2[is.na(df$BsmtFinType2)] <- 'None'
FinType <- c('None'=0, 'Unf'=1, 'LwQ'=2, 'Rec'=3, 'BLQ'=4, 'ALQ'=5, 'GLQ'=6)

df$BsmtFinType2<-as.integer(revalue(df$BsmtFinType2, FinType))
table(df$BsmtFinType2)
```

```
##
##    0    1    2    3    4    5    6
##   79 2494   87  105   68   52   34
```

```
#display remaining NAs. Using BsmtQual as a reference for the 79 houses without basement agreed upon ea
df[(is.na(df$BsmtFullBath)|is.na(df$BsmtHalfBath)|is.na(df$BsmtFinSF1)|is.na(df$BsmtFinSF2)|is.na(df$Bs
```

```
##      BsmtQual BsmtFullBath BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF
## 2121        0           NA           NA         NA         NA        NA
## 2189        0           NA           NA          0          0         0
##      TotalBsmtSF
## 2121          NA
## 2189           0
```

```
df$BsmtFullBath[is.na(df$BsmtFullBath)] <-0
table(df$BsmtFullBath)
```

```
##
##    0    1    2    3
## 1707 1172   38    2
```

```
df$BsmtHalfBath[is.na(df$BsmtHalfBath)] <-0
table(df$BsmtHalfBath)
```

```
##
##    0    1    2
## 2744  171    4
```

```
df$BsmtFinSF1[is.na(df$BsmtFinSF1)] <-0

df$BsmtFinSF2[is.na(df$BsmtFinSF2)] <-0

df$BsmtUnfSF[is.na(df$BsmtUnfSF)] <-0

df$TotalBsmtSF[is.na(df$TotalBsmtSF)] <-0
```

##Masonry

```
#check if the 23 houses with veneer area NA are also NA in the veneer type
length(which(is.na(df$MasVnrType) & is.na(df$MasVnrArea)))
```

```
## [1] 23
```

```
df[is.na(df$MasVnrType) & !is.na(df$MasVnrArea), c('MasVnrType', 'MasVnrArea')]
```

```
##      MasVnrType MasVnrArea
## 2611       <NA>        198
```

```
#fix this veneer type by imputing the mode
df$MasVnrType[2611] <- names(sort(-table(df$MasVnrType)))[2] #taking the 2nd value as the 1st is 'none'
df[2611, c('MasVnrType', 'MasVnrArea')]
```

```
##      MasVnrType MasVnrArea
## 2611    BrkFace        198
```

```
df$MasVnrType[is.na(df$MasVnrType)] <- 'None'

df[!is.na(df$SalePrice),] %>% group_by(MasVnrType) %>% summarise(median = median(SalePrice), counts=n())
```

```
## # A tibble: 4 x 3
##   MasVnrType median counts
##   <chr>       <dbl>  <int>
## 1 BrkCmn     139000     15
## 2 None       143125    872
## 3 BrkFace    181000    445
## 4 Stone      246839    128
```

```
Masonry <- c('None'=0, 'BrkCmn'=0, 'BrkFace'=1, 'Stone'=2)
df$MasVnrType<-as.integer(revalue(df$MasVnrType, Masonry))
table(df$MasVnrType)
```

```
##
##    0    1    2
## 1790  880  249
```

```
df$MasVnrArea[is.na(df$MasVnrArea)] <-0
```

##MSZoning

```
#imputing the mode
df$MSZoning[is.na(df$MSZoning)] <- names(sort(-table(df$MSZoning)))[1]
df$MSZoning <- as.factor(df$MSZoning)
table(df$MSZoning)
```

```
##
## C (all)      FV      RH      RL      RM
##      25     139      26    2269     460
```

##Kitchen

```
df$KitchenQual[is.na(df$KitchenQual)] <- 'TA' #replace with most common value
df$KitchenQual<-as.integer(revalue(df$KitchenQual, Qualities))
```

```
## The following `from` values were not present in `x`: None, Po
```

```
table(df$KitchenQual)
```

```
##
##    2    3    4    5
##   70 1493 1151  205
```

```
table(df$KitchenAbvGr)
```

```
##
##    0    1    2    3
##    3 2785  129    2
```

##Utilities

```
table(df$Utilities)
```

```
##
## AllPub NoSeWa
##   2916      1
```

```
kable(df[is.na(df$Utilities) | df$Utilities=='NoSeWa', 1:9])
```

	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
945	20	RL	82	14375	Pave	None	2	Lvl	NoSeWa
1916	30	RL	109	21780	Grvl	None	3	Lvl	NA
1946	20	RL	64	31220	Pave	None	2	Bnk	NA

```
df$Utilities <- NULL
```

##Home functionality

```
#impute mode for the 1 NA
df$Functional[is.na(df$Functional)] <- names(sort(-table(df$Functional)))[1]

df$Functional <- as.integer(revalue(df$Functional, c('Sal'=0, 'Sev'=1, 'Maj2'=2, 'Maj1'=3, 'Mod'=4, 'Mi
```

```
## The following `from` values were not present in `x`: Sal
```

```
table(df$Functional)
```

```
##
##    1    2    3    4    5    6    7
##    2    9   19   35   70   65 2719
```

##exterior variables

```
#imputing mode
df$Exterior1st[is.na(df$Exterior1st)] <- names(sort(-table(df$Exterior1st)))[1]

df$Exterior1st <- as.factor(df$Exterior1st)
table(df$Exterior1st)
```

```
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd Plywood
##      44       2       6      87       2     126     442       1     450     221
##   Stone  Stucco VinylSd Wd Sdng WdShing
##       2      43    1026     411      56
```

```
#imputing mode
df$Exterior2nd[is.na(df$Exterior2nd)] <- names(sort(-table(df$Exterior2nd)))[1]

df$Exterior2nd <- as.factor(df$Exterior2nd)
table(df$Exterior2nd)
```

```
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd   Other
##      38       4      22      47       3     126     406      15     447       1
## Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng
##     270       6      47    1015     391      81
```

```
df$ExterCond<-as.integer(revalue(df$ExterCond, Qualities))
```

```
## The following `from` values were not present in `x`: None
```

```
sum(table(df$ExterCond))
```

```
## [1] 2919
```

##Electrical system

```
#imputing mode
df$Electrical[is.na(df$Electrical)] <- names(sort(-table(df$Electrical)))[1]

df$Electrical <- as.factor(df$Electrical)
table(df$Electrical)
```

```
##
## FuseA FuseF FuseP   Mix SBrkr
##   188    50     8     1  2672
```

```
sum(table(df$Electrical))
```

```
## [1] 2919
```

##Fence

```
df$Fence[is.na(df$Fence)] <- 'None'
table(df$Fence)
```

```
##
## GdPrv  GdWo MnPrv  MnWw  None
##   118   112   329    12  2348
```

```
df[!is.na(df$SalePrice),] %>% group_by(Fence) %>% summarise(median = median(SalePrice), counts=n())
```

```
## # A tibble: 5 x 3
##   Fence median counts
##   <chr>  <dbl>  <int>
## 1 GdPrv 167500     59
## 2 GdWo  138750     54
## 3 MnPrv 137450    157
## 4 MnWw  130000     11
## 5 None  173000   1179
```

```
df$Fence <- as.factor(df$Fence)
```

##SaleType

```
#imputing mode
df$SaleType[is.na(df$SaleType)] <- names(sort(-table(df$SaleType)))[1]

df$SaleType <- as.factor(df$SaleType)
table(df$SaleType)
```

```
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD
##    87     5    26     9     8    12   239     7  2526
```

```
df$SaleCondition <- as.factor(df$SaleCondition)
table(df$SaleCondition)
```

```
##
## Abnorml AdjLand  Alloca  Family  Normal Partial
##     190      12      24      46    2402     245
```

```
sum(table(df$SaleCondition))
```

```
## [1] 2919
```

```
Charcol <- names(df[,sapply(df, is.character)])
Charcol
```

```
##  [1] "Street"       "LandContour"  "LotConfig"    "LandSlope"    "Neighborhood"
##  [6] "Condition1"   "Condition2"   "BldgType"     "HouseStyle"   "RoofStyle"
## [11] "RoofMatl"     "ExterQual"    "Foundation"   "Heating"      "HeatingQC"
## [16] "CentralAir"   "PavedDrive"
```

```
cat('There are', length(Charcol), 'remaining columns with character values')
```

```
## There are 17 remaining columns with character values
```

##Foundation

```
#No ordinality, so converting into factors
df$Foundation <- as.factor(df$Foundation)
table(df$Foundation)
```

```
##
## BrkTil CBlock  PConc   Slab  Stone   Wood
##    311   1235   1308     49     11      5
```

```
sum(table(df$Foundation))
```

```
## [1] 2919
```

##Heating

```
#No ordinality, so converting into factors
df$Heating <- as.factor(df$Heating)
table(df$Heating)
```

```
##
## Floor  GasA  GasW  Grav  OthW  Wall
##     1  2874    27     9     2     6
```

```
sum(table(df$Heating))
```

```
## [1] 2919
```

##RoofStyle

```
#No ordinality, so converting into factors
df$RoofStyle <- as.factor(df$RoofStyle)
table(df$RoofStyle)
```

```
##
##    Flat   Gable Gambrel     Hip Mansard    Shed
##      20    2310      22     551      11       5
```

```
sum(table(df$RoofStyle))
```

```
## [1] 2919
```

##LandContour

```
#No ordinality, so converting into factors
df$LandContour <- as.factor(df$LandContour)
table(df$LandContour)
```

```
##
##  Bnk  HLS  Low  Lvl
##  117  120   60 2622
```

```
sum(table(df$LandContour))
```

```
## [1] 2919
```

##BldgType

```
#No ordinality, so converting into factors
df$BldgType <- as.factor(df$BldgType)
table(df$BldgType)
```

```
##
##   1Fam 2fmCon Duplex  Twnhs TwnhsE
##   2425     62    109     96    227
```

```
sum(table(df$BldgType))
```

```
## [1] 2919
```

##Neighborhood

```
#No ordinality, so converting into factors
df$Neighborhood <- as.factor(df$Neighborhood)
table(df$Neighborhood)
```

```
##
## Blmngtn Blueste  BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert  IDOTRR
##      28      10      30     108      44     267     103     194     165      93
## MeadowV Mitchel   NAmes NoRidge NPkVill NridgHt  NWAmes OldTown  Sawyer SawyerW
##      37     114     443      71      23     166     131     239     151     125
## Somerst StoneBr   SWISU  Timber Veenker
##     182      51      48      72      24
```

```
sum(table(df$Neighborhood))
```

```
## [1] 2919
```

##Street

```
#Ordinal, so label encoding
df$Street<-as.integer(revalue(df$Street, c('Grvl'=0, 'Pave'=1)))
table(df$Street)
```

```
##
##    0    1
##   12 2907
```

```
sum(table(df$Street))
```

```
## [1] 2919
```

```
#Ordinal, so label encoding
df$PavedDrive<-as.integer(revalue(df$PavedDrive, c('N'=0, 'P'=1, 'Y'=2)))
table(df$PavedDrive)
```

```
##
##    0    1    2
##  216   62 2641
```

```
sum(table(df$PavedDrive))
```

```
## [1] 2919
```

```
df$MoSold <- as.factor(df$MoSold)
```

```
ys <- ggplot(df[!is.na(df$SalePrice),], aes(x=as.factor(YrSold), y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue')+
        scale_y_continuous(breaks= seq(0, 800000, by=25000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..)) +
        coord_cartesian(ylim = c(0, 200000)) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePric
```

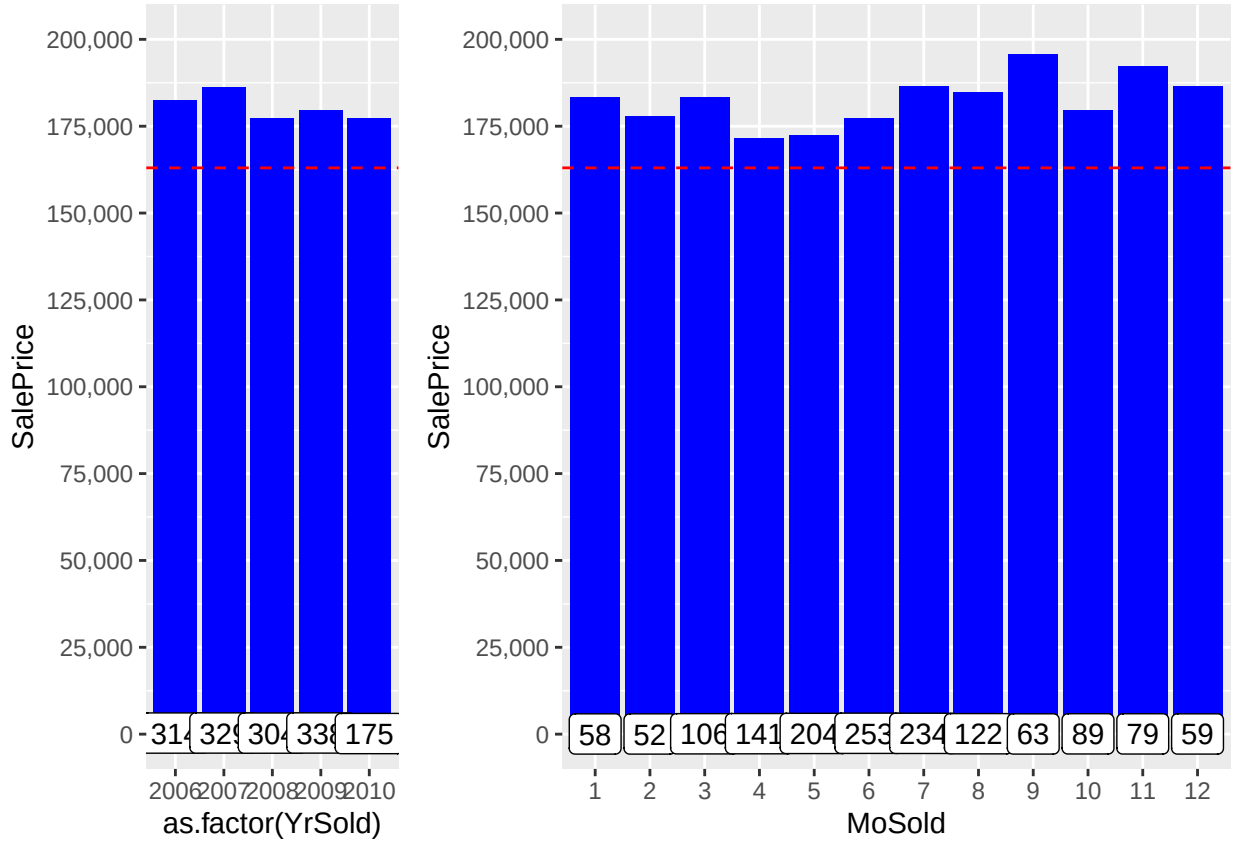
```
## Warning: Ignoring unknown parameters: fun.y
```

```
ms <- ggplot(df[!is.na(df$SalePrice),], aes(x=MoSold, y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue')+
        scale_y_continuous(breaks= seq(0, 800000, by=25000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..)) +
        coord_cartesian(ylim = c(0, 200000)) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePric
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
grid.arrange(ys, ms, widths=c(1,2))
```

```
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
```



```
df$MSSubClass <- as.factor(df$MSSubClass)

#revalue for better readability
df$MSSubClass<-revalue(df$MSSubClass, c('20'='1 story 1946+', '30'='1 story 1945-', '40'='1 story unf a

str(df$MSSubClass)
```

```
##  Factor w/ 16 levels "1 story 1946+",..: 6 1 6 7 6 5 1 6 5 16 ...
```

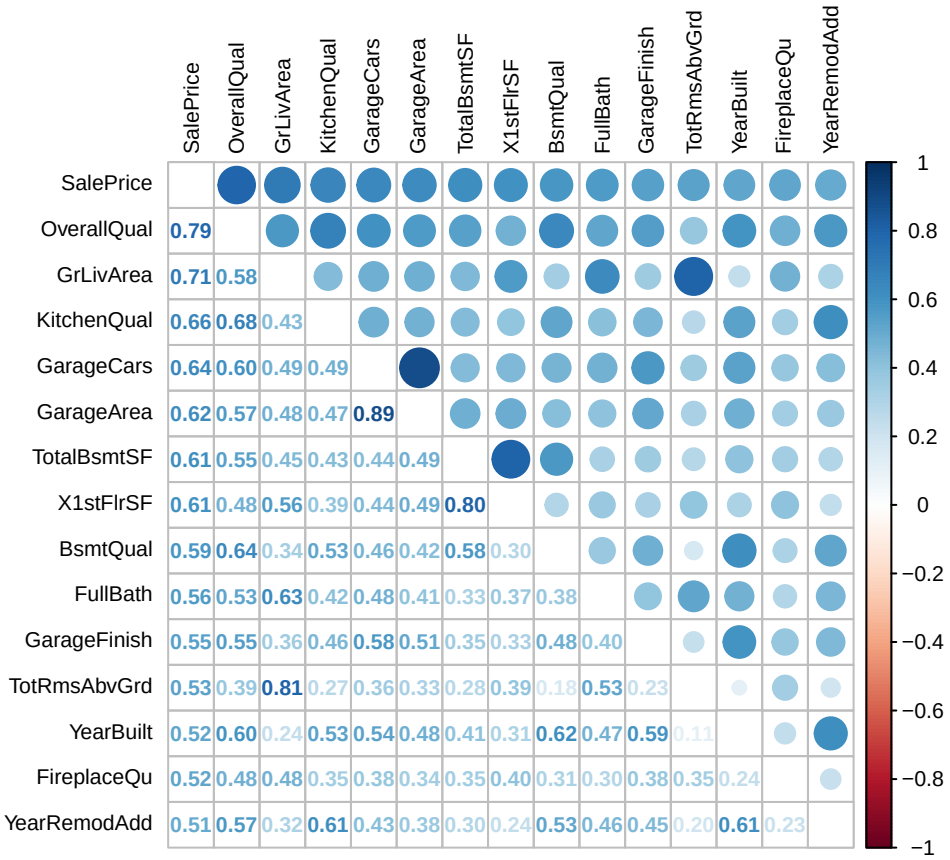
#Visualization

```
numericVars <- which(sapply(df, is.numeric)) #index vector numeric variables
factorVars <- which(sapply(df, is.factor)) #index vector factor variables
cat('There are', length(numericVars), 'numeric variables, and', length(factorVars), 'categoric variables
```

```
## There are 52 numeric variables, and 18 categoric variables
```

#Correlations

```
df_numVar <- df[, numericVars]
cor_numVar <- cor(df_numVar, use="pairwise.complete.obs") #correlations of df numeric variables

#sort on decreasing correlations with SalePrice
cor_sorted <- as.matrix(sort(cor_numVar[,'SalePrice'], decreasing = TRUE))
 #select only high corelations
CorHigh <- names(which(apply(cor_sorted, 1, function(x) abs(x)>0.5)))
cor_numVar <- cor_numVar[CorHigh, CorHigh]

corrplot.mixed(cor_numVar, tl.col="black", tl.pos = "lt", tl.cex = 0.7,cl.cex = .7, number.cex=.7)
```



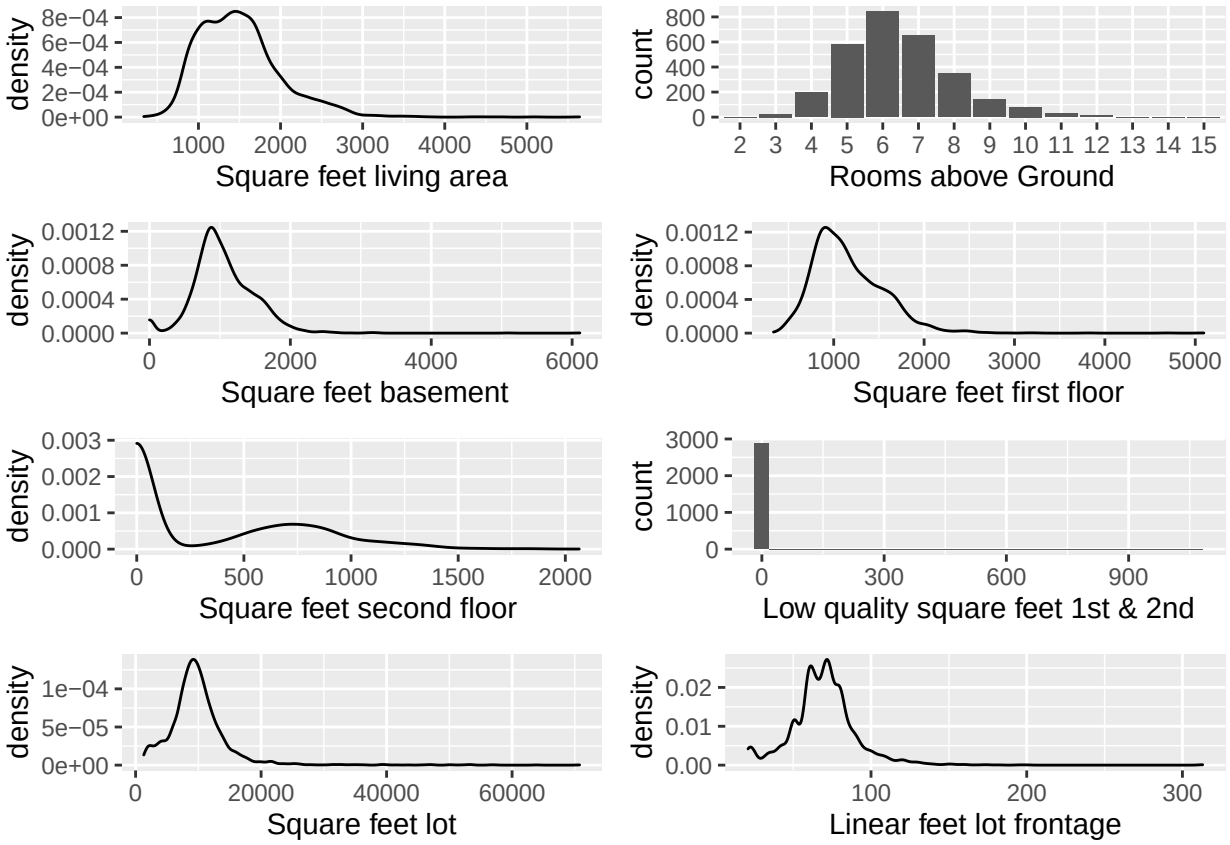

```
#is.na(df$SalePrice)
```

```
s1 <- ggplot(data= df, aes(x=GrLivArea)) +
        geom_density() + labs(x='Square feet living area')
s2 <- ggplot(data=df, aes(x=as.factor(TotRmsAbvGrd))) +
        geom_histogram(stat='count') + labs(x='Rooms above Ground')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
s3 <- ggplot(data= df, aes(x=X1stFlrSF)) +
        geom_density() + labs(x='Square feet first floor')
s4 <- ggplot(data= df, aes(x=X2ndFlrSF)) +
        geom_density() + labs(x='Square feet second floor')
s5 <- ggplot(data= df, aes(x=TotalBsmtSF)) +
        geom_density() + labs(x='Square feet basement')
s6 <- ggplot(data= df[df$LotArea<100000,], aes(x=LotArea)) +
        geom_density() + labs(x='Square feet lot')
s7 <- ggplot(data= df, aes(x=LotFrontage)) +
        geom_density() + labs(x='Linear feet lot frontage')
s8 <- ggplot(data= df, aes(x=LowQualFinSF)) +
        geom_histogram() + labs(x='Low quality square feet 1st & 2nd')

layout <- matrix(c(1,2,5,3,4,8,6,7),4,2,byrow=TRUE)
multiplot(s1, s2, s3, s4, s5, s6, s7, s8, layout=layout)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
cor(df$GrLivArea, (df$X1stFlrSF + df$X2ndFlrSF + df$LowQualFinSF))
```

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

```
head(df[df$LowQualFinSF>0, c('GrLivArea', 'X1stFlrSF', 'X2ndFlrSF', 'LowQualFinSF')])
```

```
##     GrLivArea X1stFlrSF X2ndFlrSF LowQualFinSF
## 52       1176       816         0          360
## 89       1526      1013         0          513
## 126       754       520         0          234
## 171      1382       854         0          528
## 186      3608      1518      1518          572
## 188      1656       808       704          144
```

```
n1 <- ggplot(df[!is.na(df$SalePrice),], aes(x=Neighborhood, y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePric
```

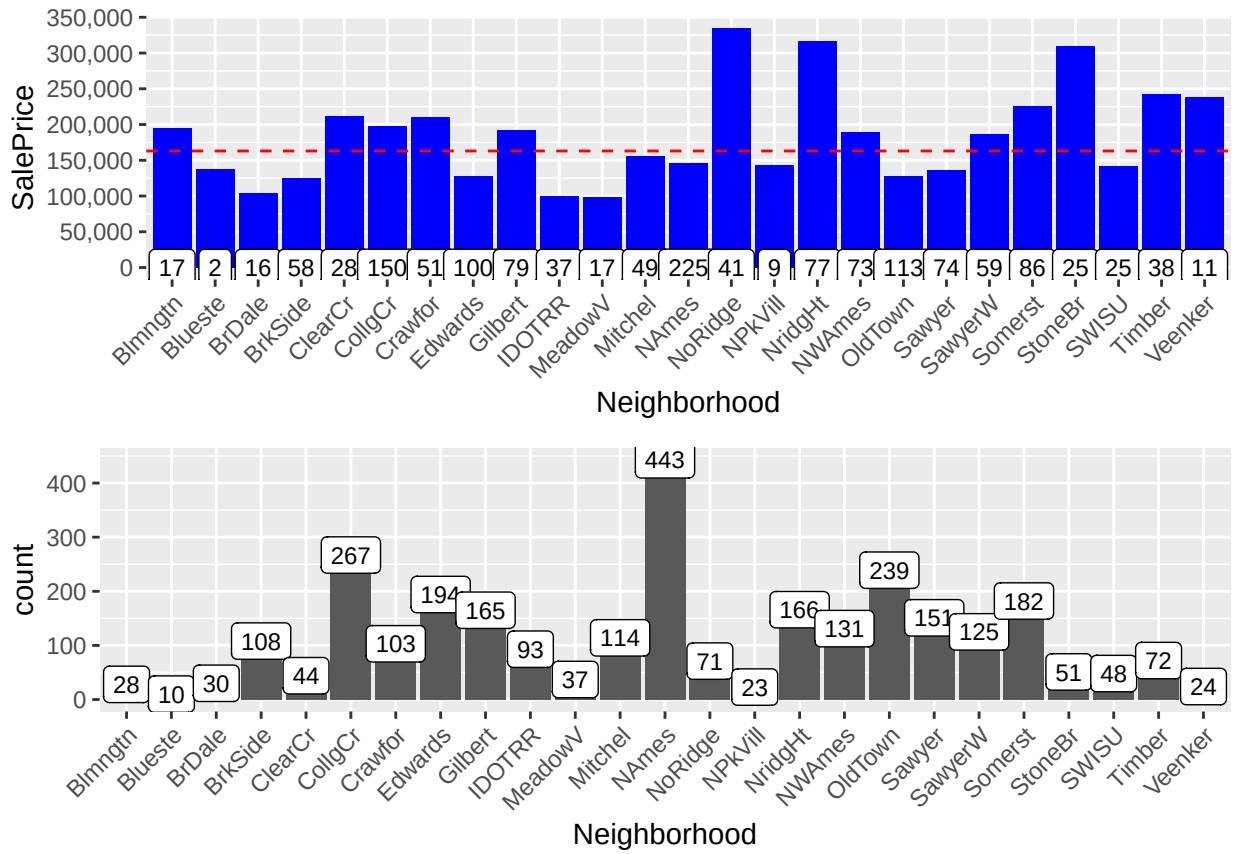
```
## Warning: Ignoring unknown parameters: fun.y
```

```
n2 <- ggplot(data=df, aes(x=Neighborhood)) +
        geom_histogram(stat='count')+
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3)+
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
grid.arrange(n1, n2)
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
q1 <- ggplot(data=df, aes(x=as.factor(OverallQual))) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
q2 <- ggplot(data=df, aes(x=as.factor(ExterQual))) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
q3 <- ggplot(data=df, aes(x=as.factor(BsmtQual))) +
       geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
q4 <- ggplot(data=df, aes(x=as.factor(KitchenQual))) +
       geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
q5 <- ggplot(data=df, aes(x=as.factor(GarageQual))) +
       geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

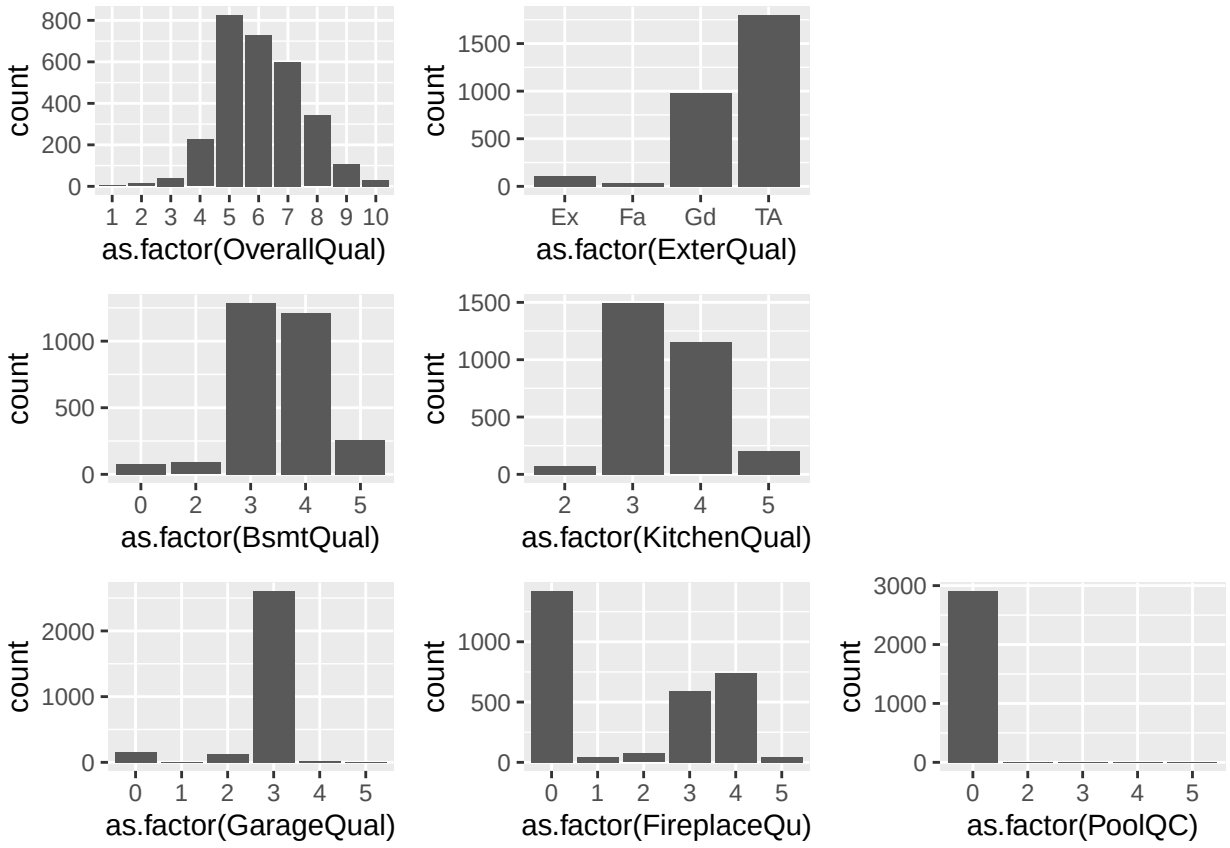
```
q6 <- ggplot(data=df, aes(x=as.factor(FireplaceQu))) +
       geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
q7 <- ggplot(data=df, aes(x=as.factor(PoolQC))) +
       geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
layout <- matrix(c(1,2,8,3,4,8,5,6,7),3,3,byrow=TRUE)
multiplot(q1, q2, q3, q4, q5, q6, q7, layout=layout)
```




```
ms1 <- ggplot(df[!is.na(df$SalePrice),], aes(x=MSSubClass, y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePric
```

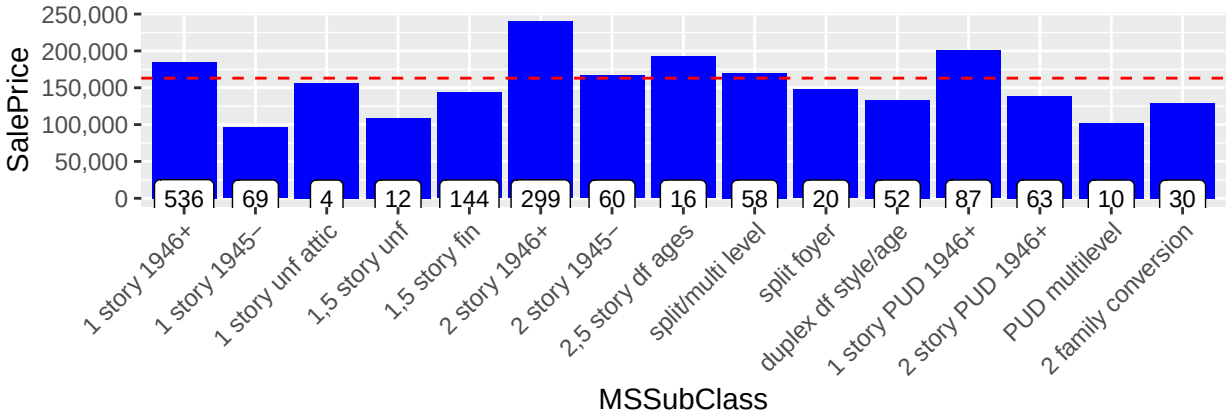
```
## Warning: Ignoring unknown parameters: fun.y
```

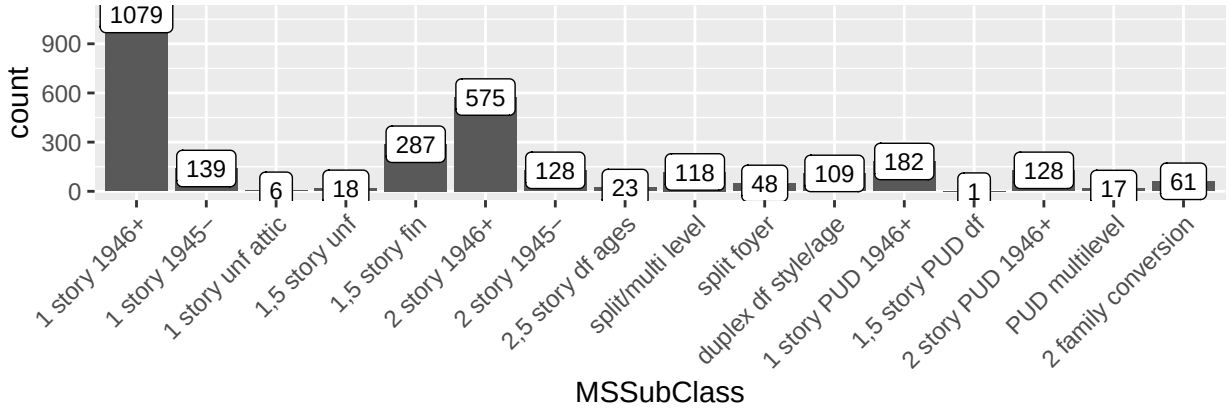
```
ms2 <- ggplot(data=df, aes(x=MSSubClass)) +
        geom_histogram(stat='count')+
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3) +
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
grid.arrange(ms1, ms2)
```

```
## No summary function supplied, defaulting to `mean_se()`
```






```
#correct error
df$GarageYrBlt[2593] <- 2007 #this must have been a typo. GarageYrBlt=2207, YearBuilt=2006, YearRemodAd

g1 <- ggplot(data=df[df$GarageCars !=0,], aes(x=GarageYrBlt)) +
        geom_histogram()
g2 <- ggplot(data=df, aes(x=as.factor(GarageCars))) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
g3 <- ggplot(data= df, aes(x=GarageArea)) +
        geom_density()
g4 <- ggplot(data=df, aes(x=as.factor(GarageCond))) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
g5 <- ggplot(data=df, aes(x=GarageType)) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
g6 <- ggplot(data=df, aes(x=as.factor(GarageQual))) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

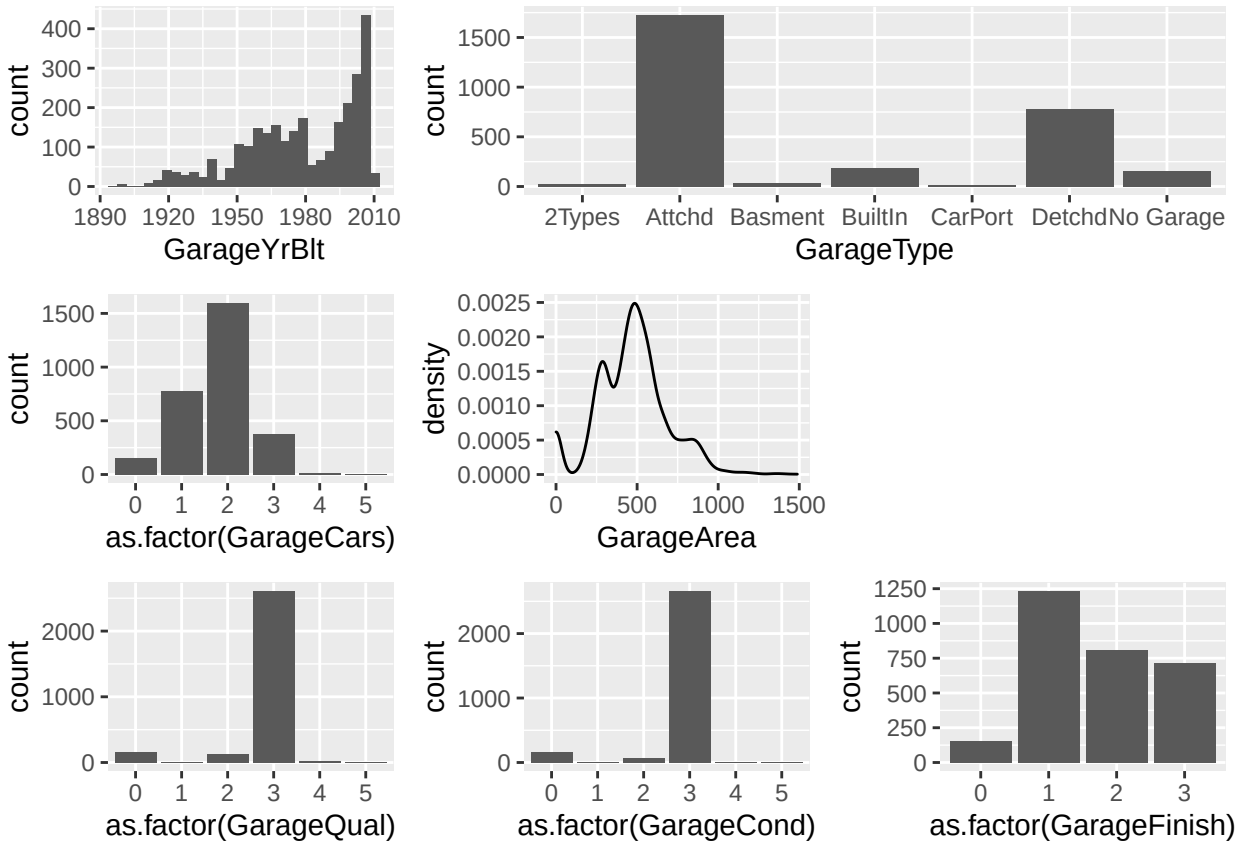
```
g7 <- ggplot(data=df, aes(x=as.factor(GarageFinish))) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
layout <- matrix(c(1,5,5,2,3,8,6,4,7),3,3,byrow=TRUE)
multiplot(g1, g2, g3, g4, g5, g6, g7, layout=layout)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```




```
b1 <- ggplot(data=df, aes(x=BsmtFinSF1)) +
        geom_histogram() + labs(x='Type 1 finished square feet')
b2 <- ggplot(data=df, aes(x=BsmtFinSF2)) +
        geom_histogram()+ labs(x='Type 2 finished square feet')
b3 <- ggplot(data=df, aes(x=BsmtUnfSF)) +
```

```
        geom_histogram()+ labs(x='Unfinished square feet')
b4 <- ggplot(data=df, aes(x=as.factor(BsmtFinType1))) +
        geom_histogram(stat='count')+ labs(x='Rating of Type 1 finished area')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
b5 <- ggplot(data=df, aes(x=as.factor(BsmtFinType2))) +
        geom_histogram(stat='count')+ labs(x='Rating of Type 2 finished area')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
b6 <- ggplot(data=df, aes(x=as.factor(BsmtQual))) +
        geom_histogram(stat='count')+ labs(x='Height of the basement')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
b7 <- ggplot(data=df, aes(x=as.factor(BsmtCond))) +
        geom_histogram(stat='count')+ labs(x='Rating of general condition')
```

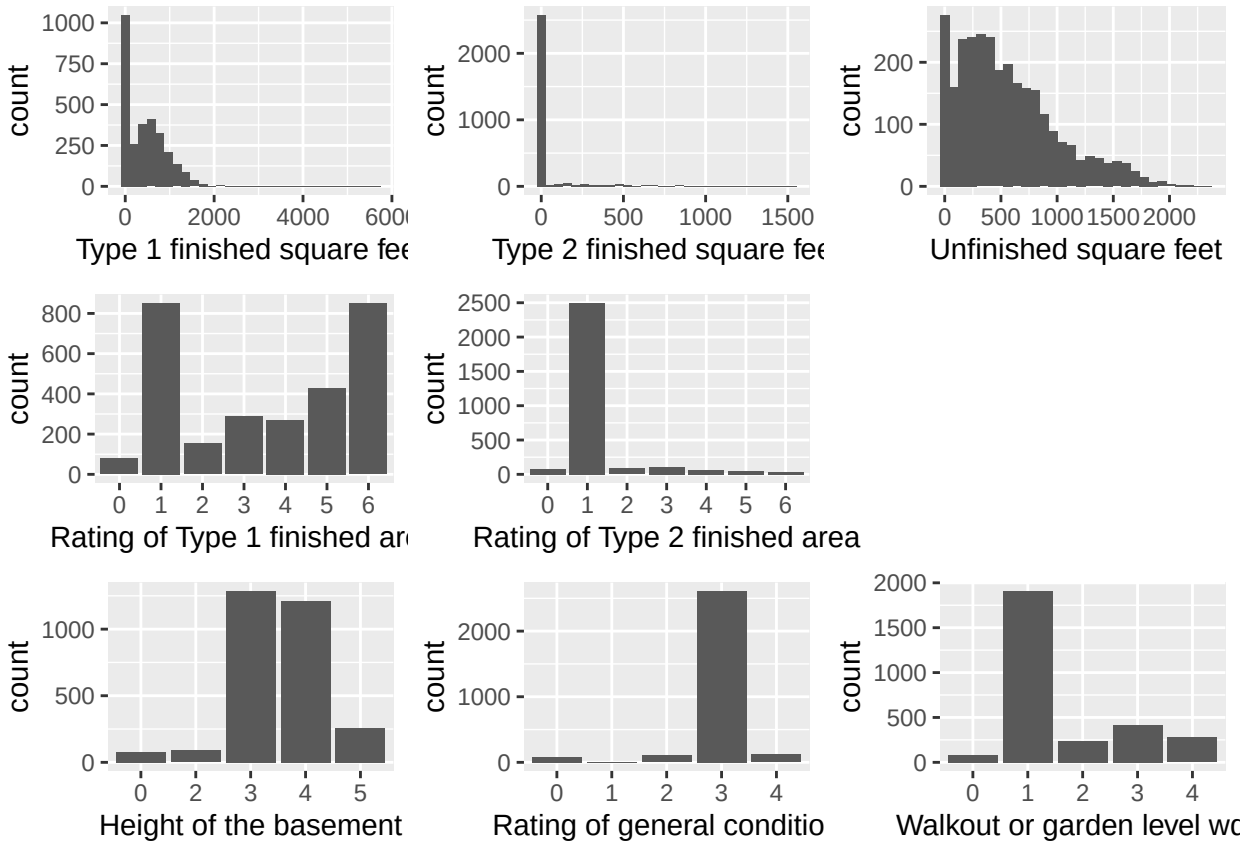
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
b8 <- ggplot(data=df, aes(x=as.factor(BsmtExposure))) +
        geom_histogram(stat='count')+ labs(x='Walkout or garden level wdfs')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
layout <- matrix(c(1,2,3,4,5,9,6,7,8),3,3,byrow=TRUE)
multiplot(b1, b2, b3, b4, b5, b6, b7, b8, layout=layout)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

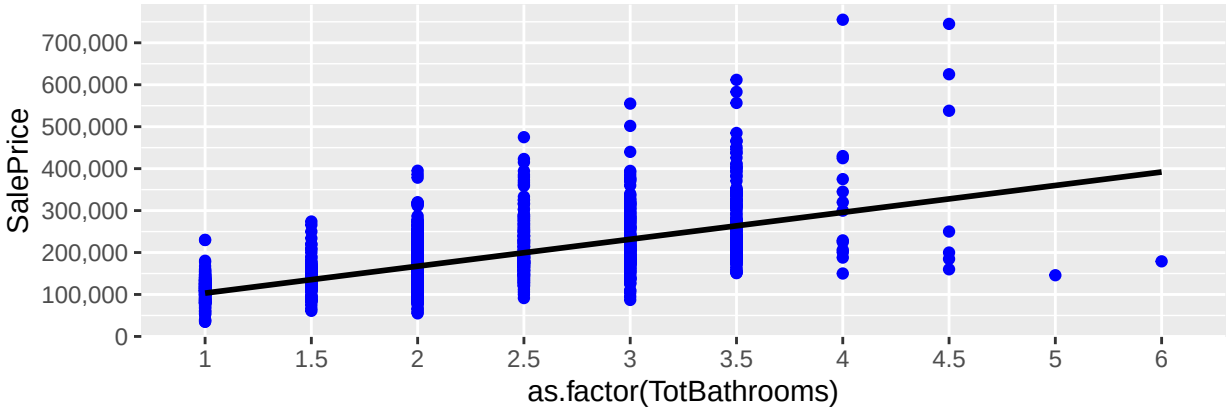


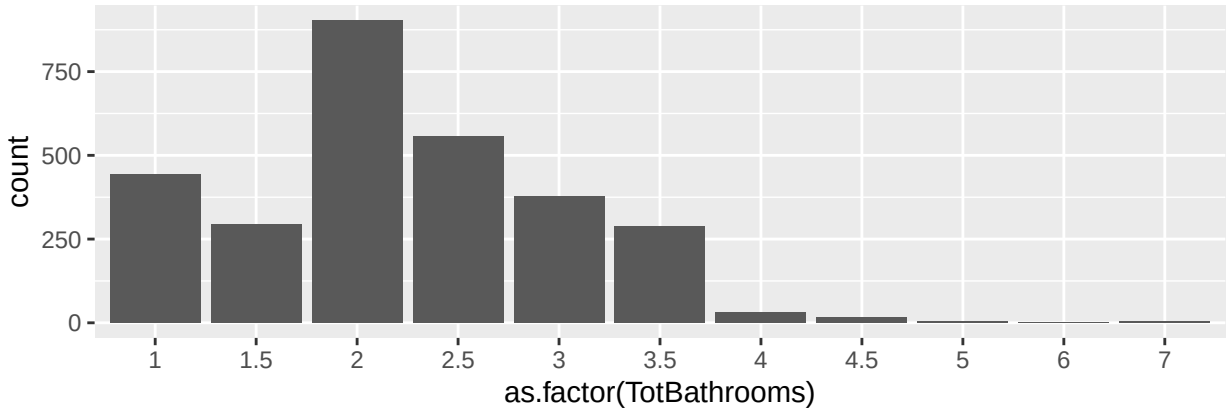

```
df$TotBathrooms <- df$FullBath + (df$HalfBath*0.5) + df$BsmtFullBath + (df$BsmtHalfBath*0.5)

tb1 <- ggplot(data=df[!is.na(df$SalePrice),], aes(x=as.factor(TotBathrooms), y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
tb2 <- ggplot(data=df, aes(x=as.factor(TotBathrooms))) +
        geom_histogram(stat='count')
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
grid.arrange(tb1, tb2)
```

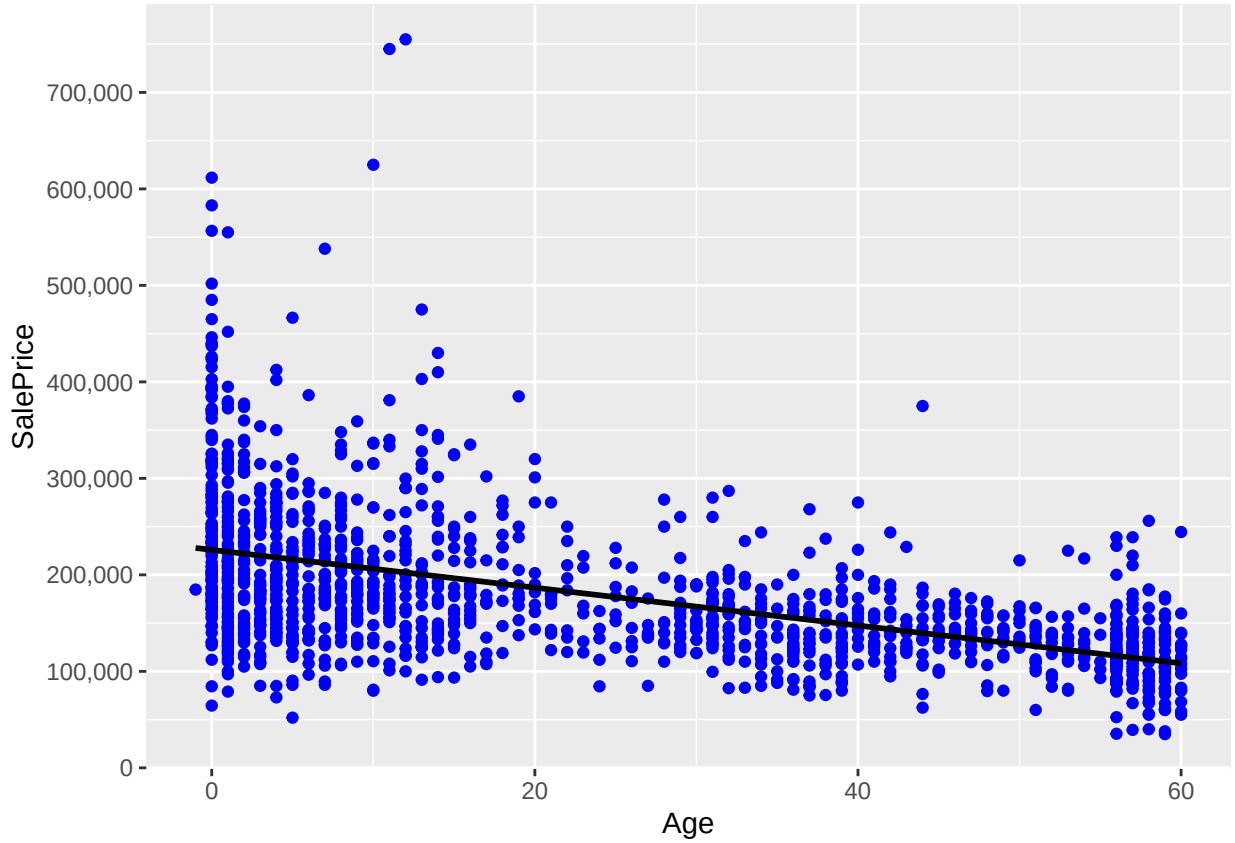
```
## `geom_smooth()` using formula 'y ~ x'
```






```
df$Remod <- ifelse(df$YearBuilt==df$YearRemodAdd, 0, 1) #0=No Remodeling, 1=Remodeling
df$Age <- as.numeric(df$YrSold)-df$YearRemodAdd

ggplot(data=df[!is.na(df$SalePrice),], aes(x=Age, y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



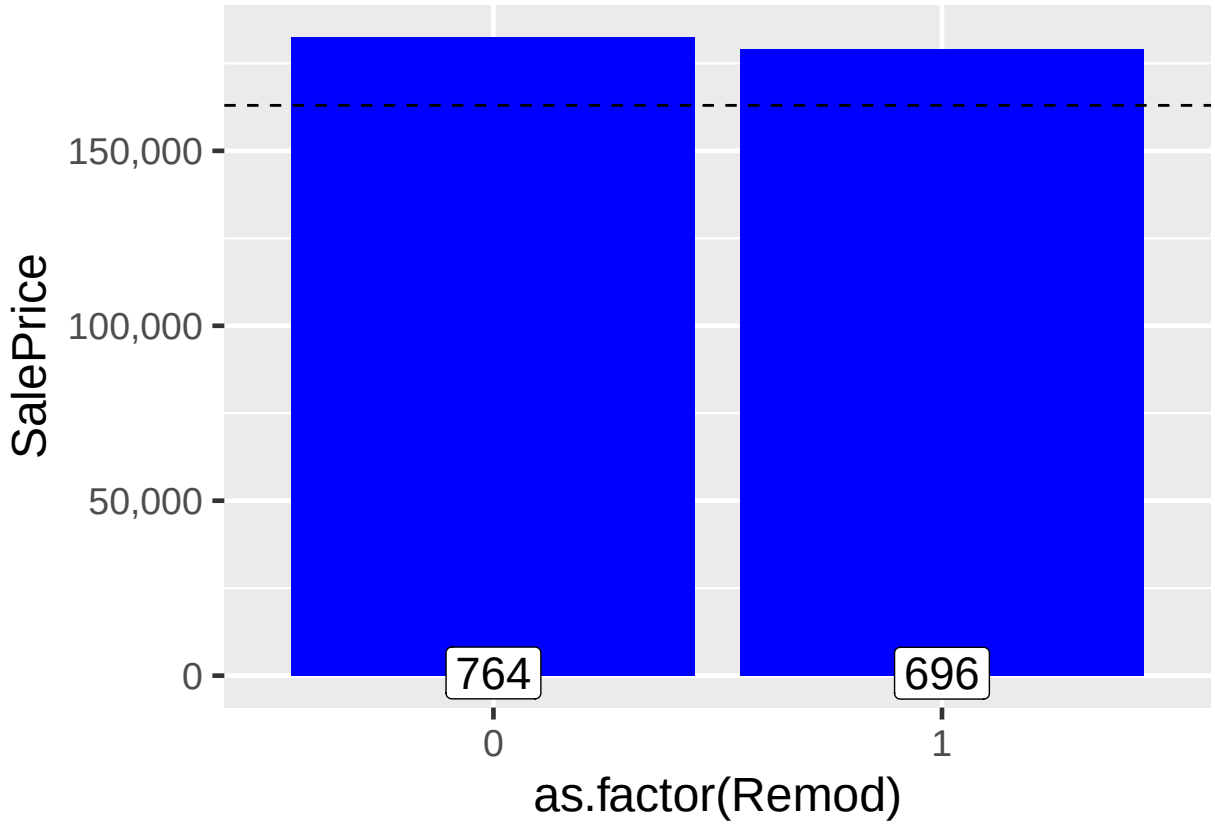

```
cor(df$SalePrice[!is.na(df$SalePrice)], df$Age[!is.na(df$SalePrice)])
```

```
## [1] -0.5090787
```

```
ggplot(df[!is.na(df$SalePrice),], aes(x=as.factor(Remod), y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=6) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
        theme_grey(base_size = 18) +
        geom_hline(yintercept=163000, linetype="dashed") #dashed line is median SalePrice
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```



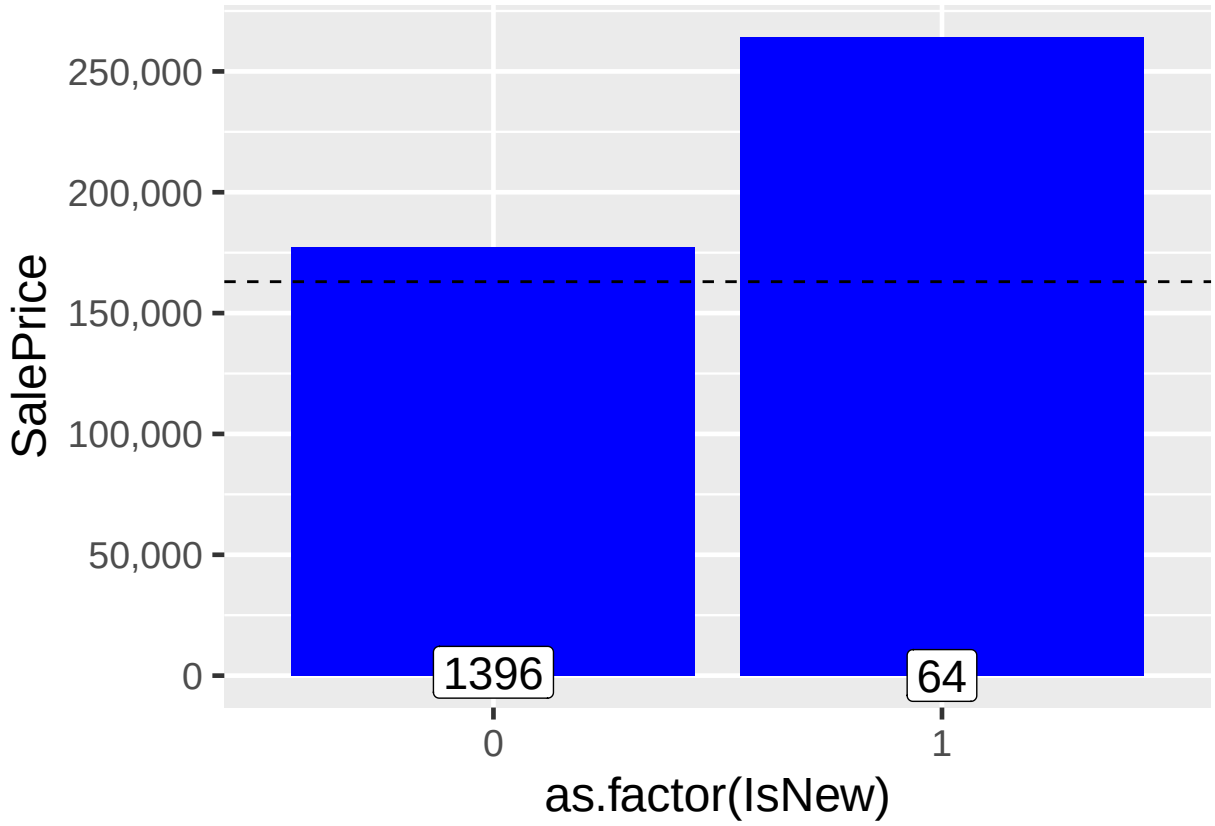

```
df$IsNew <- ifelse(df$YrSold==df$YearBuilt, 1, 0)
table(df$IsNew)
```

```
##
##    0    1
## 2803  116
```

```
ggplot(df[!is.na(df$SalePrice),], aes(x=as.factor(IsNew), y=SalePrice)) +
        geom_bar(stat='summary', fun.y = "median", fill='blue') +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=6) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
        theme_grey(base_size = 18) +
        geom_hline(yintercept=163000, linetype="dashed") #dashed line is median SalePrice
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
## No summary function supplied, defaulting to `mean_se()`
```




```
df$YrSold <- as.factor(df$YrSold) #the numeric version is now not needed anymore
```

```
nb1 <- ggplot(df[!is.na(df$SalePrice),], aes(x=reorder(Neighborhood, SalePrice, FUN=median), y=SalePric
        geom_bar(stat='summary', fun.y = "median", fill='blue') + labs(x='Neighborhood', y='Median Sale
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePric
```

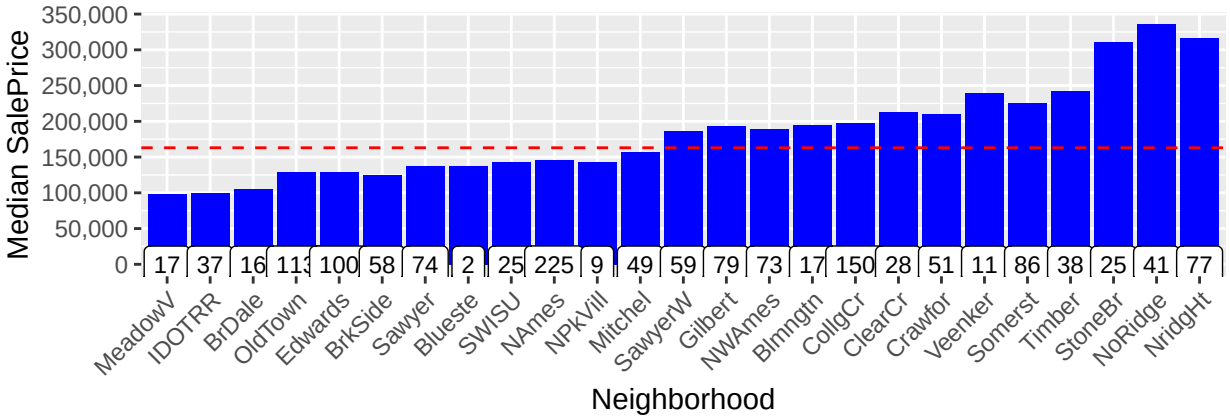
```
## Warning: Ignoring unknown parameters: fun.y
```

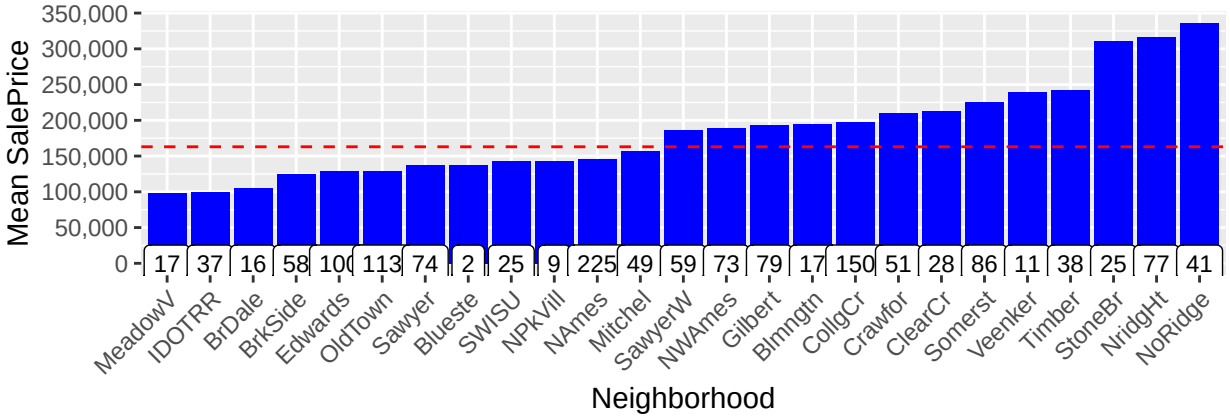
```
nb2 <- ggplot(df[!is.na(df$SalePrice),], aes(x=reorder(Neighborhood, SalePrice, FUN=mean), y=SalePrice)
        geom_bar(stat='summary', fun.y = "mean", fill='blue') + labs(x='Neighborhood', y="Mean SalePric
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=50000), labels = comma) +
        geom_label(stat = "count", aes(label = ..count.., y = ..count..), size=3) +
        geom_hline(yintercept=163000, linetype="dashed", color = "red") #dashed line is median SalePric
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
grid.arrange(nb1, nb2)
```

```
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
```



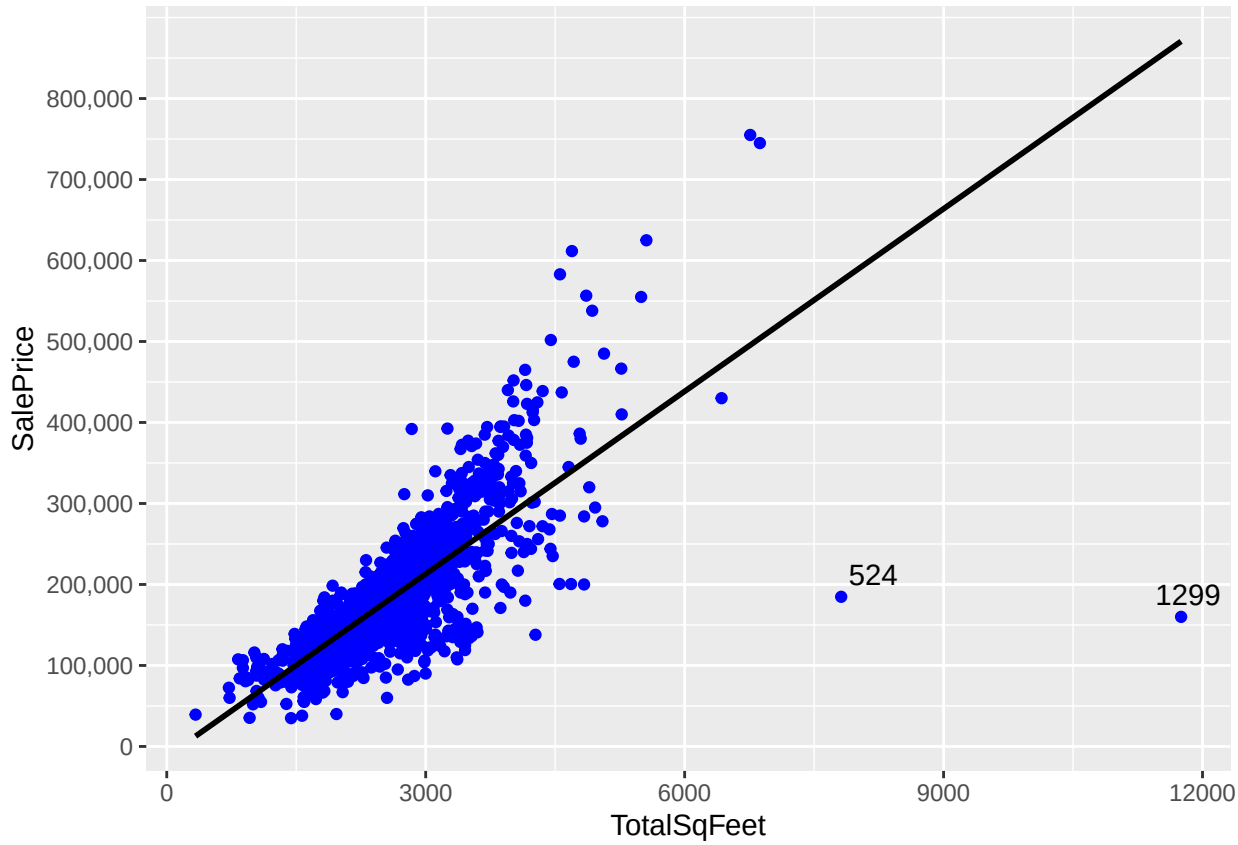



```
df$NeighRich[df$Neighborhood %in% c('StoneBr', 'NridgHt', 'NoRidge')] <- 2
df$NeighRich[!df$Neighborhood %in% c('MeadowV', 'IDOTRR', 'BrDale', 'StoneBr', 'NridgHt', 'NoRidge')] <-
df$NeighRich[df$Neighborhood %in% c('MeadowV', 'IDOTRR', 'BrDale')] <- 0

table(df$NeighRich)
```

```
##
##    0    1    2
##  160 2471  288
```

```
df$TotalSqFeet <- df$GrLivArea + df$TotalBsmtSF

ggplot(data=df[!is.na(df$SalePrice),], aes(x=TotalSqFeet, y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma) +
        geom_text_repel(aes(label = ifelse(df$GrLivArea[!is.na(df$SalePrice)]>4500, rownames(df), '')))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(df$SalePrice, df$TotalSqFeet, use= "pairwise.complete.obs")
```

```
## [1] 0.7789588
```

```
cor(df$SalePrice[-c(524, 1299)], df$TotalSqFeet[-c(524, 1299)], use= "pairwise.complete.obs")
```

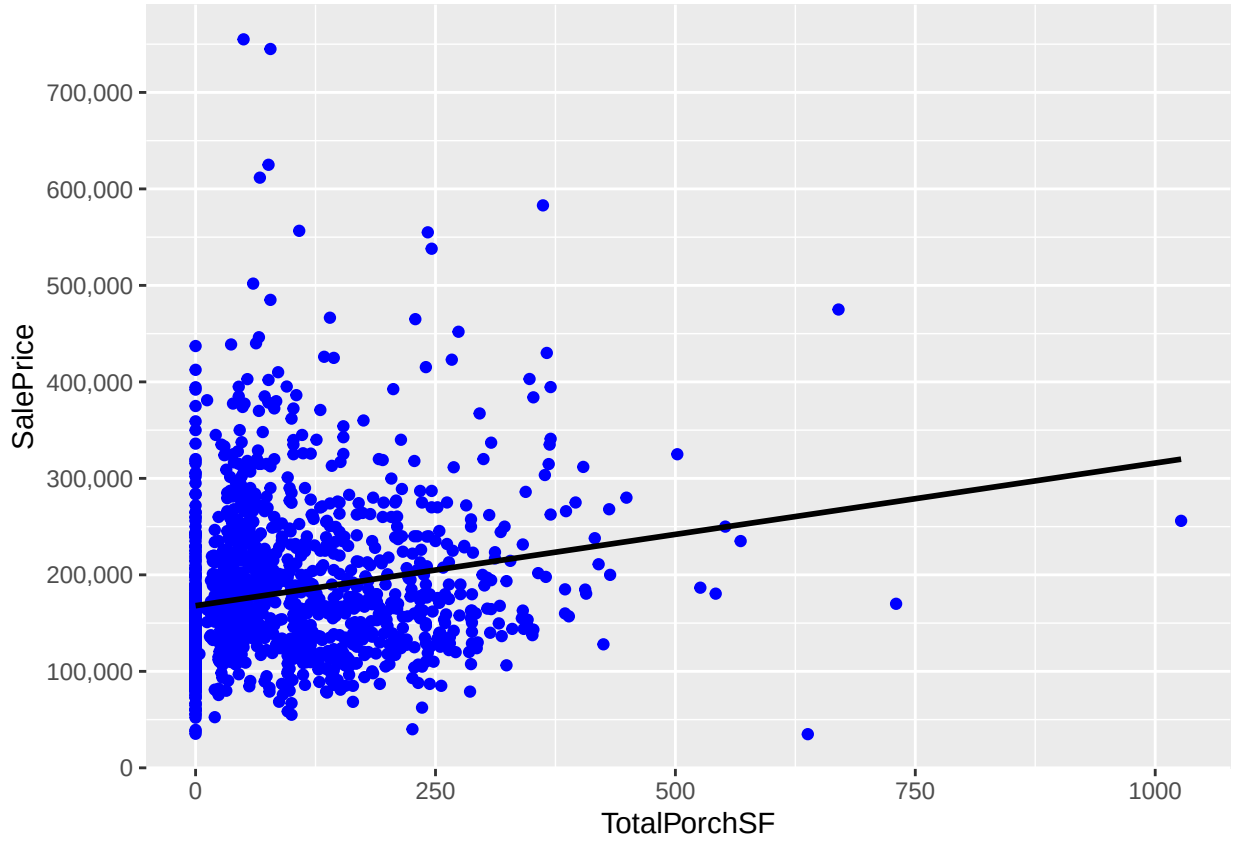
```
## [1] 0.829042
```

```
df$TotalPorchSF <- df$OpenPorchSF + df$EnclosedPorch + df$X3SsnPorch + df$ScreenPorch

cor(df$SalePrice, df$TotalPorchSF, use= "pairwise.complete.obs")
```

```
## [1] 0.1957389
```

```
ggplot(data=df[!is.na(df$SalePrice),], aes(x=TotalPorchSF, y=SalePrice))+
        geom_point(col='blue') + geom_smooth(method = "lm", se=FALSE, color="black", aes(group=1)) +
        scale_y_continuous(breaks= seq(0, 800000, by=100000), labels = comma)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



#Preparing data for modeling

```
dropVars <- c('YearRemodAdd', 'GarageYrBlt', 'GarageArea', 'GarageCond', 'TotalBsmtSF', 'TotalRmsAbvGrd

df <- df[,!(names(df) %in% dropVars)]


df <- df[-c(524, 1299),]
```

```
numericVarNames <- numericVarNames[!(numericVarNames %in% c('MSSubClass', 'MoSold', 'YrSold', 'SalePric
numericVarNames <- append(numericVarNames, c('Age', 'TotalPorchSF', 'TotBathrooms', 'TotalSqFeet'))

DFnumeric <- df[, names(df) %in% numericVarNames]

DFfactors <- df[, !(names(df) %in% numericVarNames)]
DFfactors <- DFfactors[, names(DFfactors) != 'SalePrice']

cat('There are', length(DFnumeric), 'numeric variables, and', length(DFfactors), 'factor variables')
```

```
## There are 32 numeric variables, and 47 factor variables
```

```
for(i in 1:ncol(DFnumeric)){
        if (abs(skew(DFnumeric[,i]))>0.8){
                DFnumeric[,i] <- log(DFnumeric[,i] +1)
        }
```

```
}


PreNum <- preProcess(DFnumeric, method=c("center", "scale"))
print(PreNum)
```

```
## Created from 2917 samples and 32 variables
##
## Pre-processing:
##   - centered (32)
##   - ignored (0)
##   - scaled (32)
```

```
DFnorm <- predict(PreNum, DFnumeric)
dim(DFnorm)
```

```
## [1] 2917   32
```

```
DFdummies <- as.data.frame(model.matrix(~.-1, DFfactors))
dim(DFdummies)
```

```
## [1] 2917  204
```

```
#check if some values are absent in the test set
ZerocolTest <- which(colSums(DFdummies[(nrow(df[!is.na(df$SalePrice),])+1):nrow(df),])==0)
colnames(DFdummies[ZerocolTest])
```

```
##  [1] "Condition2RRAe"     "Condition2RRAn"     "Condition2RRNn"
##  [4] "HouseStyle2.5Fin"   "RoofMatlMembran"    "RoofMatlMetal"
##  [7] "RoofMatlRoll"       "Exterior1stImStucc" "Exterior1stStone"
## [10] "Exterior2ndOther"   "HeatingOthW"        "ElectricalMix"
## [13] "MiscFeatureTenC"
```

```
DFdummies <- DFdummies[,-ZerocolTest] #removing predictors
```

```
#check if some values are absent in the train set
ZerocolTrain <- which(colSums(DFdummies[1:nrow(df[!is.na(df$SalePrice),]),])==0)
colnames(DFdummies[ZerocolTrain])
```

```
## [1] "MSSubClass1,5 story PUD df"
```

```
DFdummies <- DFdummies[,-ZerocolTrain] #removing predictor

fewOnes <- which(colSums(DFdummies[1:nrow(df[!is.na(df$SalePrice),]),])<10)
colnames(DFdummies[fewOnes])
```

```
##  [1] "MSSubClass1 story unf attic" "LotConfigFR3"
##  [3] "NeighborhoodBlueste"         "NeighborhoodNPkVill"
##  [5] "Condition1PosA"              "Condition1RRNe"
```

```
##  [7] "Condition1RRNn"              "Condition2Feedr"
##  [9] "Condition2PosA"              "Condition2PosN"
## [11] "RoofStyleMansard"            "RoofStyleShed"
## [13] "RoofMatlWdShake"             "RoofMatlWdShngl"
## [15] "Exterior1stAsphShn"          "Exterior1stBrkComm"
## [17] "Exterior1stCBlock"           "Exterior2ndAsphShn"
## [19] "Exterior2ndBrk Cmn"          "Exterior2ndCBlock"
## [21] "Exterior2ndStone"            "FoundationStone"
## [23] "FoundationWood"              "HeatingGrav"
## [25] "HeatingWall"                 "HeatingQCPo"
## [27] "ElectricalFuseP"             "GarageTypeCarPort"
## [29] "MiscFeatureOthr"             "SaleTypeCon"
## [31] "SaleTypeConLD"               "SaleTypeConLI"
## [33] "SaleTypeConLw"               "SaleTypeCWD"
## [35] "SaleTypeOth"                 "SaleConditionAdjLand"
```

```
DFdummies <- DFdummies[,-fewOnes] #removing predictors
dim(DFdummies)
```

```
## [1] 2917  154
```

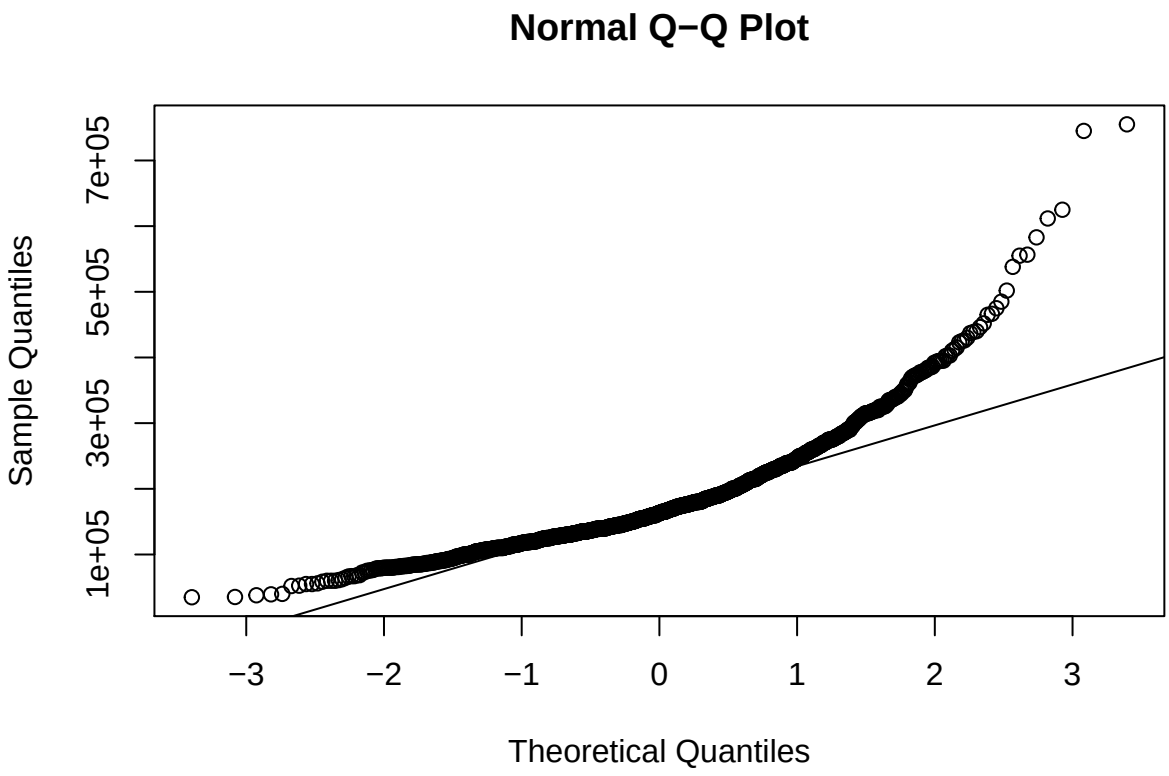
```
combined <- cbind(DFnorm, DFdummies) #combining df (now numeric) predictors into one dataframe
```

##Dealing with skewness of response variable

```
skew(df$SalePrice)
```

```
## [1] 1.877427
```

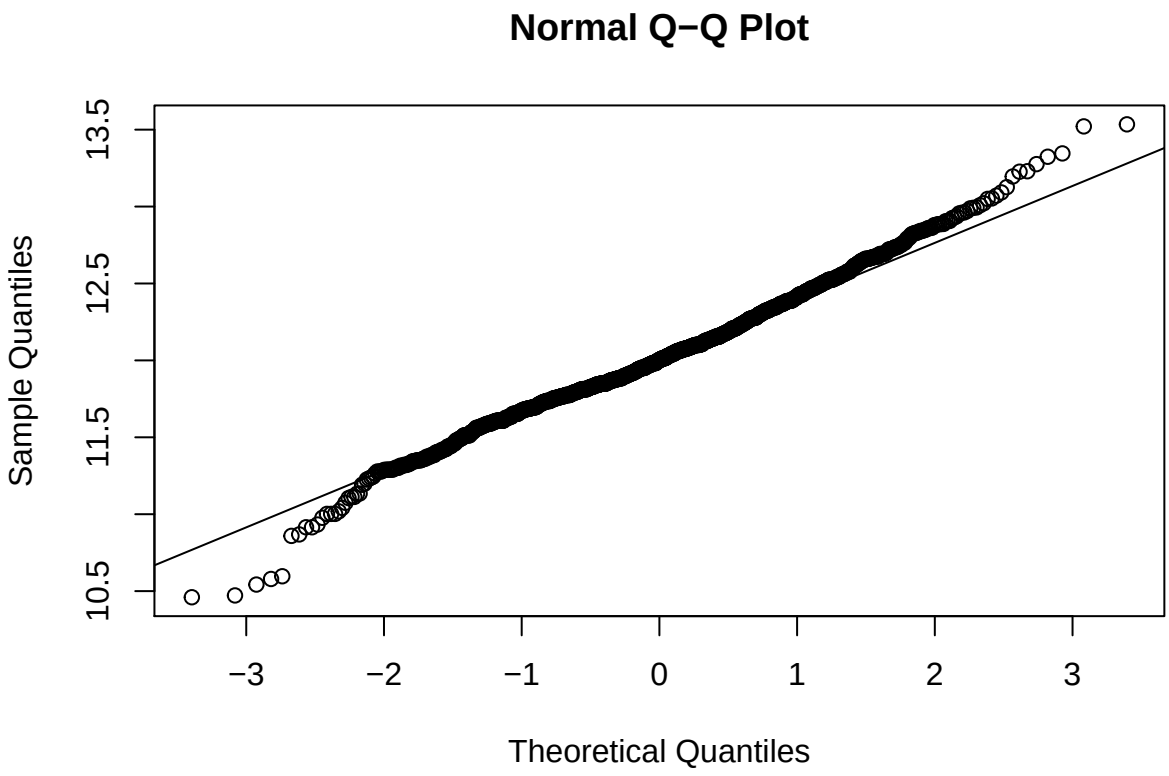
```
qqnorm(df$SalePrice)
qqline(df$SalePrice)
```




```
df$SalePrice <- log(df$SalePrice) #default is the natural logarithm, "+1" is not necessary as there are
skew(df$SalePrice)
```

```
## [1] 0.1213182
```

```
qqnorm(df$SalePrice)
qqline(df$SalePrice)
```




```
train1 <- combined[!is.na(df$SalePrice),]
test1 <- combined[is.na(df$SalePrice),]
```

#Lasso regression model

```
set.seed(27042018)

my_control <-trainControl(method="cv", number=5)
lassoGrid <- expand.grid(alpha = 1, lambda = seq(0.001,0.1,by = 0.0005))

lasso_mod <- train(x=train1, y= df$SalePrice[!is.na(df$SalePrice)],
                   method='glmnet',
                   trControl= my_control,
                   tuneGrid=lassoGrid)
lasso_mod$bestTune
```

```
##   alpha lambda
## 1     1  0.001
```

```
min(lasso_mod$results$RMSE)
```

```
## [1] 0.1126066
```

```
#print(lasso_mod$results)
summary(lasso_mod$results)
```

```
##      alpha       lambda             RMSE           Rsquared
##  Min.   :1   Min.   :0.00100   Min.   :0.1126   Min.   :0.8335
##  1st Qu.:1   1st Qu.:0.02575   1st Qu.:0.1382   1st Qu.:0.8452
##  Median :1   Median :0.05050   Median :0.1601   Median :0.8651
##  Mean   :1   Mean   :0.05050   Mean   :0.1588   Mean   :0.8689
##  3rd Qu.:1   3rd Qu.:0.07525   3rd Qu.:0.1810   3rd Qu.:0.8887
##  Max.   :1   Max.   :0.10000   Max.   :0.2001   Max.   :0.9207
##       MAE              RMSESD           RsquaredSD           MAESD
##  Min.   :0.07883   Min.   :0.006354   Min.   :0.008675   Min.   :0.002049
##  1st Qu.:0.09629   1st Qu.:0.010558   1st Qu.:0.017703   1st Qu.:0.004762
##  Median :0.11199   Median :0.010635   Median :0.020017   Median :0.005069
##  Mean   :0.11215   Mean   :0.010328   Mean   :0.018967   Mean   :0.004958
##  3rd Qu.:0.12900   3rd Qu.:0.010762   3rd Qu.:0.021636   3rd Qu.:0.005413
##  Max.   :0.14414   Max.   :0.011166   Max.   :0.022866   Max.   :0.006713
```

```
lassoVarImp <- varImp(lasso_mod,scale=F)
lassoImportance <- lassoVarImp$importance

varsSelected <- length(which(lassoImportance$Overdf!=0))
varsNotSelected <- length(which(lassoImportance$Overdf==0))

cat('Lasso uses', varsSelected, 'variables in its model, and did not select', varsNotSelected, 'variable
```

```
## Lasso uses 0 variables in its model, and did not select 0 variables.
```

```
LassoPred <- predict(lasso_mod, test1)
predictions_lasso <- exp(LassoPred) #need to reverse the log to the real values
head(predictions_lasso)
```

```
##     1461     1462     1463     1464     1465     1466
## 115022.3 162468.2 179228.5 199205.0 204180.9 168945.3
```