Data 605 - Final project

Jai Jeffryes

5/22/2020

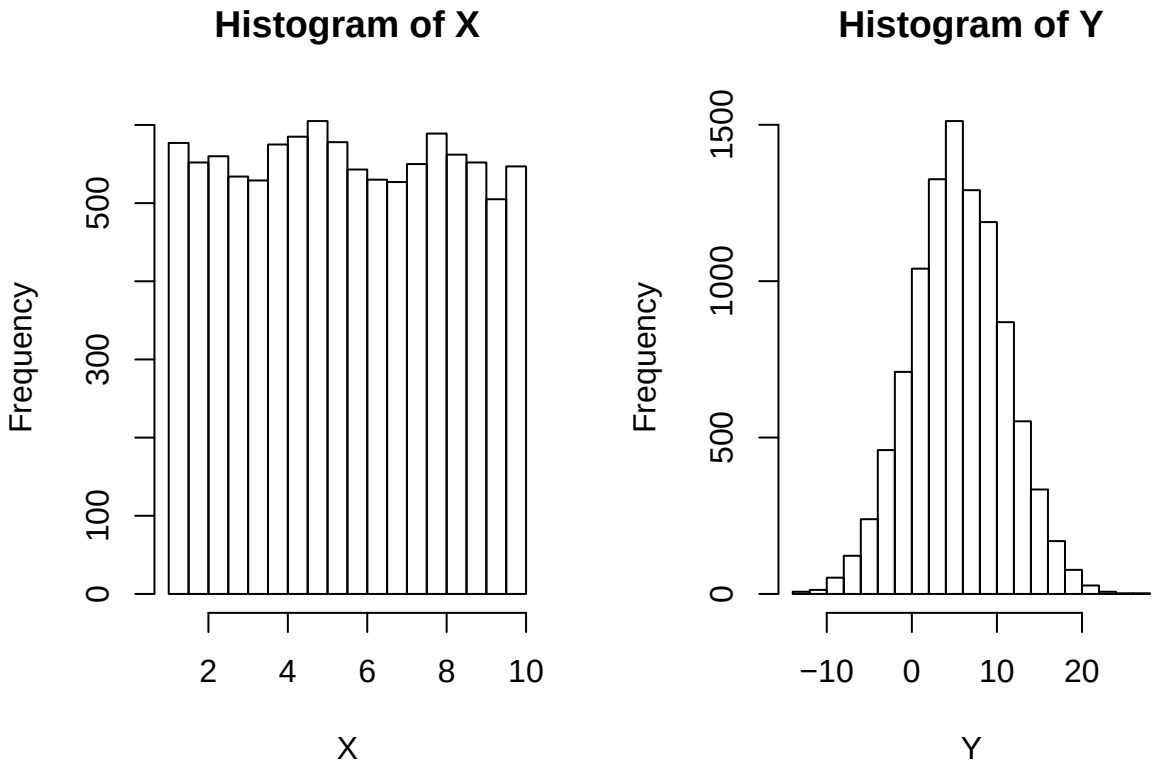
Contents

Problem 1.

Using `R`*, generate a random variable* X *that has 10,000 random uniform numbers from 1 to* N*, where* N *can be any number of your choosing greater than or equal to 6. Then generate a random variable* Y *that has 10,000 random normal numbers with a mean of* $\mu = \sigma = \frac{N+1}{2}$*.*

Probability. Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the median of the X *variable, and the small letter "y" is estimated as the 1st quartile of the* Y *variable. Interpret the meaning of all probabilities.*

```
set.seed(2020)
N <- 10
X <- runif(10000, min = 1, max = N)
Y <- rnorm(10000, mean = (N + 1) / 2, sd = (N + 1) / 2)

par(mfrow = c(1,2))
hist(X)
hist(Y)
```




```
x <- median(X)
y <- quantile(Y, 0.25)
```

1.1: 5 points

a. $P(X > x | X > y)$

$P(X > x|X > y) = \frac{P(X>x \text{ and } X>y)}{P(X>y)}$

```
ans_a <- (sum(X > x & X > y) / length(X)) / (sum(X > y) / length(X))
```

Answer: 0.5468665 is the probability that an observation X is greater than its median given that it is greater than the first quartile of Y.

b. $P(X > x, Y > y)$

```
ans_b <- (sum(X > x) / length(X)) * (sum(Y > y) / length(Y))
```

Answer: 0.375 is the probability that an observation X is greater than its median and an observation Y is greater than its first quartile. That is, $P(X > x) \cap P(Y > y)$.

c. $P(X < x|X > y)$

$P(X < x|X > y) = \frac{P(X<x \text{ and } X>y)}{P(X>y)}$

```
ans_c <- (sum(X < x & X > y) / length(X)) / (sum(X > y) / length(X))
```

Answer: 0.4531335 is the probability that an observation X is less than its median given that it is greater than the first quartile of Y.

1.2: 5 points

Investigate whether $P(X > x \text{ and } Y > y) = P(X > x)P(Y > y)$ by building a table and evaluating the marginal and joint probabilities.

```
# contingency_tbl <- addmargins(prop.table(table(X > x, Y > y)))
contingency_tbl <- table(X > x, Y > y)
contingency_prop <- addmargins(prop.table(contingency_tbl))

dimnames(contingency_prop) <- list(
    c("!(X>x)", "X>x", "Total"),
    c("!(Y>y)", "Y>y", "Total"))
print(contingency_prop)
```

```
##         !(Y>y)    Y>y  Total
## !(X>x) 0.1239 0.3761 0.5000
## X>x    0.1261 0.3739 0.5000
## Total  0.2500 0.7500 1.0000
```

- Marginal:
  - Total of $X > x = 0.5000$, and
  - Total of $Y > y = 0.7500$.
  - $0.5000 \cdot 0.7500 = 0.375$.
- Joint:
  - The joint probability is: 0.3739

The joint probability reduces the marginal probability by -0.29%

1.3: 5 points

Check to see if independence holds by using Fisher's Exact Test and the Chi Square Test. What is the difference between the two? Which is more appropriate?

Notes:

- Fisher's Exact Test, Wikipedia
- Pearson's chi-squared test, Wikipedia
- Chi-squared test, R Bloggers

```
fisher.test(contingency_tbl)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  contingency_tbl
## p-value = 0.6277
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.891325 1.070505
## sample estimates:
## odds ratio
##  0.9768025
```

```
chisq.test(contingency_tbl)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_tbl
## X-squared = 0.2352, df = 1, p-value = 0.6277
```

Conclusion

Both tests led to a p-value of 0.6277, so we reject the null hypothesis.

When sample sizes are small, Fisher's Exact Test is more appropriate than Pearson's Chi-Squared Test. However, as the sample sizes here are 10,000, either test is effective.

Problem 2

You are to register for Kaggle.com (free) and compete in the House Prices: Advanced Regression Techniques competition. https://www.kaggle.com/c/house-prices-advanced-regression-techniques. I want you to do the following.

Reference:

- par(mar, mgp, las). Orientation of barplot labels.

Load data and format it

The project codebook, `data_description.txt`, documents numeric columns containing codes, which are qualitative variables representing categories, rather than quantitative variables. These are converted to factors.

```
library(MASS)
library(Rmisc)
```

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

```
## Loading required package: plyr
```

```
train <- read.csv("train.csv")
test <- read.csv("test.csv")
# Transform codes to factors.
train$MSSubClass <- as.factor(train$MSSubClass)
train$OverallQual <- as.factor(train$OverallQual)
train$OverallCond <- as.factor(train$OverallCond)

test$MSSubClass <- as.factor(test$MSSubClass)
test$OverallQual <- as.factor(test$OverallQual)
test$OverallCond <- as.factor(test$OverallCond)

#train$YearBuilt <- as.factor(train$YearBuilt)
#train$YearRemodAdd <- as.factor(train$YearRemodAdd)
#train$GarageYrBlt <- as.factor(train$GarageYrBlt)
#train$MoSold <- as.factor(train$MoSold)
#train$YrSold <- as.factor(train$YrSold)
```

2.1: 5 points - Descriptive and Inferential Statistics

Provide univariate descriptive statistics and appropriate plots for the training data set

- We summarize counts of categorical variables. Refer to codebook, `data_description.txt`, for decoding.
- We summarize centers, ranges, and IQRs of numerical variables.
- Plots of of the distributions of all variables (except the primary key) follow the summaries. We plot bar charts for categorical variables and box plots for numeric variables.
- Nice to have:
  - Exploratory data analysis would be easier if the summary of each variable appeared next to its plot.
  - It would be helpful to control the pagination, getting the summaries and plots together on a single page and to use up a page fully.

```
dim(train) # Dimensions of the dataset.
```

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

```
sapply(train, class) # List types for each attribute.
```

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

```
str(train) # Structure, including factor levels.
```

```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : Factor w/ 15 levels "20","30","40",..: 6 1 6 7 6 5 1 6 5 15 ...
##  $ 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 7420 ...
##  $ 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 NA 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  : Factor w/ 10 levels "1","2","3","4",..: 7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : Factor w/ 9 levels "1","2","3","4",..: 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 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 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 NA 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    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```
summary(train[-1]) # Counts of factors, 5-number summaries of numeric variables.
```

```
##    MSSubClass     MSZoning     LotFrontage        LotArea
##  20     :536   C (all):  10   Min.   : 21.00   Min.   :  1300
##  60     :299   FV     :  65   1st Qu.: 59.00   1st Qu.:  7554
##  50     :144   RH     :  16   Median : 69.00   Median :  9478
##  120    : 87   RL     :1151   Mean   : 70.05   Mean   : 10517
##  30     : 69   RM     : 218   3rd Qu.: 80.00   3rd Qu.: 11602
##  160    : 63                  Max.   :313.00   Max.   :215245
##  (Other):262                  NA's   :259
##   Street      Alley      LotShape  LandContour  Utilities
##  Grvl:   6   Grvl:  50   IR1:484   Bnk:  63    AllPub:1459
##  Pave:1454   Pave:  41   IR2: 41   HLS:  50    NoSeWa:   1
##              NA's:1369   IR3: 10   Low:  36
##                          Reg:925   Lvl:1311
##
##
##
##    LotConfig    LandSlope   Neighborhood   Condition1     Condition2
##  Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260   Norm   :1445
##  CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81   Feedr  :   6
##  FR2    :  47   Sev:  13   OldTown:113   Artery :  48   Artery :   2
##  FR3    :   4              Edwards:100   RRAn   :  26   PosN   :   2
##  Inside :1052              Somerst: 86   PosN   :  19   RRNn   :   2
##                            Gilbert: 79   RRAe   :  11   PosA   :   1
##                            (Other):707   (Other):  15   (Other):   2
##    BldgType      HouseStyle   OverallQual   OverallCond    YearBuilt
##  1Fam  :1220   1Story :726   5      :397   5      :821   Min.   :1872
##  2fmCon:  31   2Story :445   6      :374   6      :252   1st Qu.:1954
##  Duplex:  52   1.5Fin :154   7      :319   7      :205   Median :1973
##  Twnhs :  43   SLvl   : 65   8      :168   8      : 72   Mean   :1971
##  TwnhsE: 114   SFoyer : 37   4      :116   4      : 57   3rd Qu.:2000
##                1.5Unf : 14   9      : 43   3      : 25   Max.   :2010
##                (Other): 19   (Other): 43   (Other): 28
##   YearRemodAdd    RoofStyle       RoofMatl     Exterior1st   Exterior2nd
##  Min.   :1950   Flat   :  13   CompShg:1434   VinylSd:515   VinylSd:504
##  1st Qu.:1967   Gable  :1141   Tar&Grv:  11   HdBoard:222   MetalSd:214
##  Median :1994   Gambrel:  11   WdShngl:   6   MetalSd:220   HdBoard:207
```

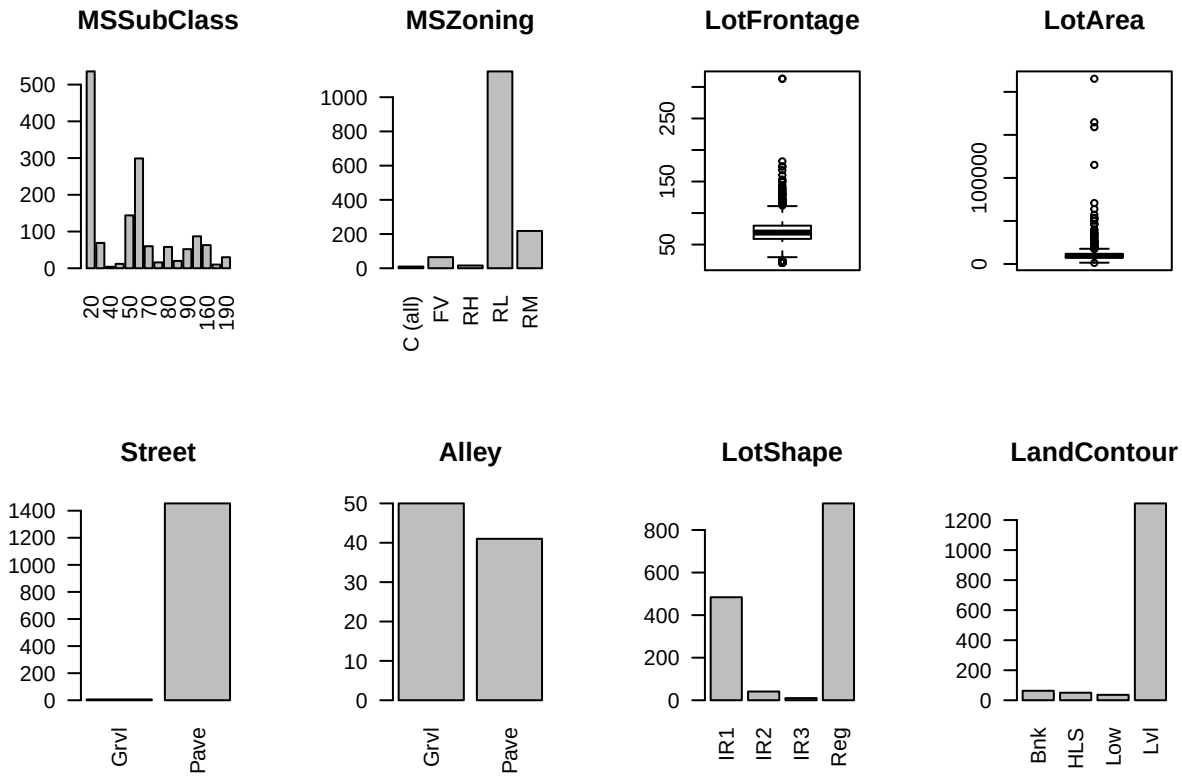
```
##  Mean   :1985   Hip    : 286   WdShake:   5   Wd Sdng:206   Wd Sdng:197
##  3rd Qu.:2004   Mansard:   7   ClyTile:   1   Plywood:108   Plywood:142
##  Max.   :2010   Shed   :   2   Membran:   1   CemntBd: 61   CmentBd: 60
##                                (Other):   2   (Other):128   (Other):136
##    MasVnrType    MasVnrArea     ExterQual ExterCond  Foundation
##  BrkCmn : 15   Min.   :   0.0   Ex: 52    Ex:   3   BrkTil:146
##  BrkFace:445   1st Qu.:   0.0   Fa: 14    Fa:  28   CBlock:634
##  None   :864   Median :   0.0   Gd:488    Gd: 146   PConc :647
##  Stone  :128   Mean   : 103.7   TA:906    Po:   1   Slab  : 24
##  NA's   :  8   3rd Qu.: 166.0             TA:1282   Stone :  6
##                Max.   :1600.0                       Wood  :  3
##                NA's   :8
##  BsmtQual   BsmtCond    BsmtExposure BsmtFinType1   BsmtFinSF1
##  Ex  :121   Fa  :  45   Av  :221     ALQ :220     Min.   :   0.0
##  Fa  : 35   Gd  :  65   Gd  :134     BLQ :148     1st Qu.:   0.0
##  Gd  :618   Po  :   2   Mn  :114     GLQ :418     Median : 383.5
##  TA  :649   TA  :1311   No  :953     LwQ : 74     Mean   : 443.6
##  NA's: 37   NA's:  37   NA's: 38     Rec :133     3rd Qu.: 712.2
##                                      Unf :430     Max.   :5644.0
##                                      NA's: 37
##  BsmtFinType2   BsmtFinSF2        BsmtUnfSF       TotalBsmtSF
##  ALQ :  19    Min.   :   0.00   Min.   :   0.0   Min.   :   0.0
##  BLQ :  33    1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8
##  GLQ :  14    Median :   0.00   Median : 477.5   Median : 991.5
##  LwQ :  46    Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  Rec :  54    3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Unf :1256    Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##  NA's:  38
##   Heating     HeatingQC CentralAir Electrical     X1stFlrSF
##  Floor:   1   Ex:741    N:  95     FuseA:  94   Min.   : 334
##  GasA :1428   Fa: 49    Y:1365     FuseF:  27   1st Qu.: 882
##  GasW :  18   Gd:241               FuseP:   3   Median :1087
##  Grav :   7   Po:  1               Mix  :   1   Mean   :1163
##  OthW :   2   TA:428               SBrkr:1334   3rd Qu.:1391
##  Wall :   4                        NA's :   1   Max.   :4692
##
##    X2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional    Fireplaces
##  Min.   :0.000   Ex:100      Min.   : 2.000   Maj1:  14   Min.   :0.000
```

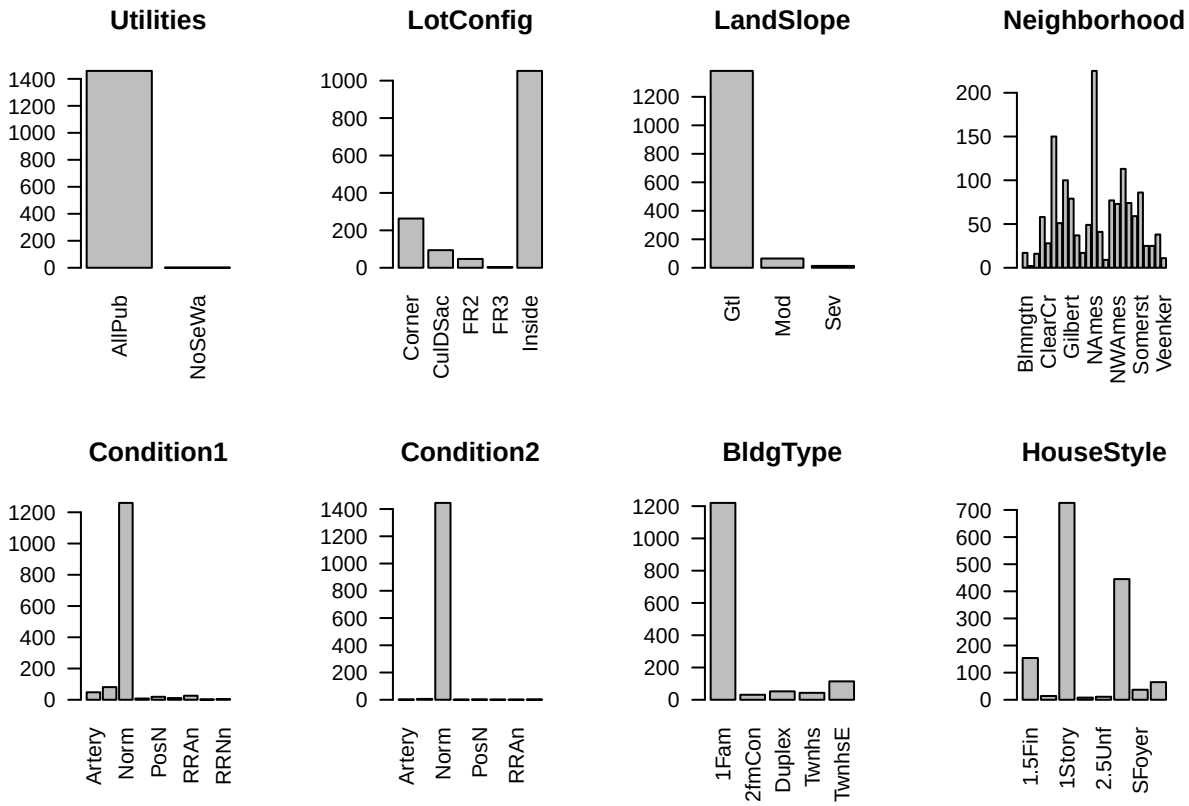
```
##  1st Qu.:1.000   Fa: 39      1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000
##  Median :1.000   Gd:586      Median : 6.000   Min1:  31   Median :1.000
##  Mean   :1.047   TA:735      Mean   : 6.518   Min2:  34   Mean   :0.613
##  3rd Qu.:1.000               3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000
##  Max.   :3.000               Max.   :14.000   Sev :   1   Max.   :3.000
##                                               Typ :1360
##  FireplaceQu   GarageType   GarageYrBlt   GarageFinish   GarageCars
##  Ex  : 24    2Types :  6   Min.   :1900   Fin :352     Min.   :0.000
##  Fa  : 33    Attchd :870   1st Qu.:1961   RFn :422     1st Qu.:1.000
##  Gd  :380    Basment: 19   Median :1980   Unf :605     Median :2.000
##  Po  : 20    BuiltIn: 88   Mean   :1979   NA's: 81     Mean   :1.767
##  TA  :313    CarPort:  9   3rd Qu.:2002                3rd Qu.:2.000
##  NA's:690    Detchd :387   Max.   :2010                Max.   :4.000
##              NA's   : 81   NA's   :81
##    GarageArea     GarageQual  GarageCond  PavedDrive   WoodDeckSF
##  Min.   :   0.0   Ex  :   3   Ex  :   2   N:  90     Min.   :  0.00
##  1st Qu.: 334.5   Fa  :  48   Fa  :  35   P:  30     1st Qu.:  0.00
##  Median : 480.0   Gd  :  14   Gd  :   9   Y:1340     Median :  0.00
##  Mean   : 473.0   Po  :   3   Po  :   7              Mean   : 94.24
##  3rd Qu.: 576.0   TA  :1311   TA  :1326              3rd Qu.:168.00
##  Max.   :1418.0   NA's:  81   NA's:  81              Max.   :857.00
##
##   OpenPorchSF     EnclosedPorch      X3SsnPorch      ScreenPorch
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Median : 25.00   Median :  0.00   Median :  0.00   Median :  0.00
##  Mean   : 46.66   Mean   : 21.95   Mean   :  3.41   Mean   : 15.06
##  3rd Qu.: 68.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00
##  Max.   :547.00   Max.   :552.00   Max.   :508.00   Max.   :480.00
##
##     PoolArea        PoolQC       Fence      MiscFeature    MiscVal
##  Min.   :  0.000   Ex  :   2   GdPrv:  59   Gar2:   2   Min.   :    0.00
##  1st Qu.:  0.000   Fa  :   2   GdWo :  54   Othr:   2   1st Qu.:    0.00
##  Median :  0.000   Gd  :   3   MnPrv: 157   Shed:  49   Median :    0.00
##  Mean   :  2.759   NA's:1453   MnWw :  11   TenC:   1   Mean   :   43.49
##  3rd Qu.:  0.000               NA's :1179   NA's:1406   3rd Qu.:    0.00
##  Max.   :738.000                                       Max.   :15500.00
##
##      MoSold           YrSold        SaleType    SaleCondition
##  Min.   : 1.000   Min.   :2006   WD     :1267   Abnorml: 101
##  1st Qu.: 5.000   1st Qu.:2007   New    : 122   AdjLand:   4
##  Median : 6.000   Median :2008   COD    :  43   Alloca :  12
##  Mean   : 6.322   Mean   :2008   ConLD  :   9   Family :  20
##  3rd Qu.: 8.000   3rd Qu.:2009   ConLI  :   5   Normal :1198
##  Max.   :12.000   Max.   :2010   ConLw  :   5   Partial: 125
##                                  (Other):   9
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##
```

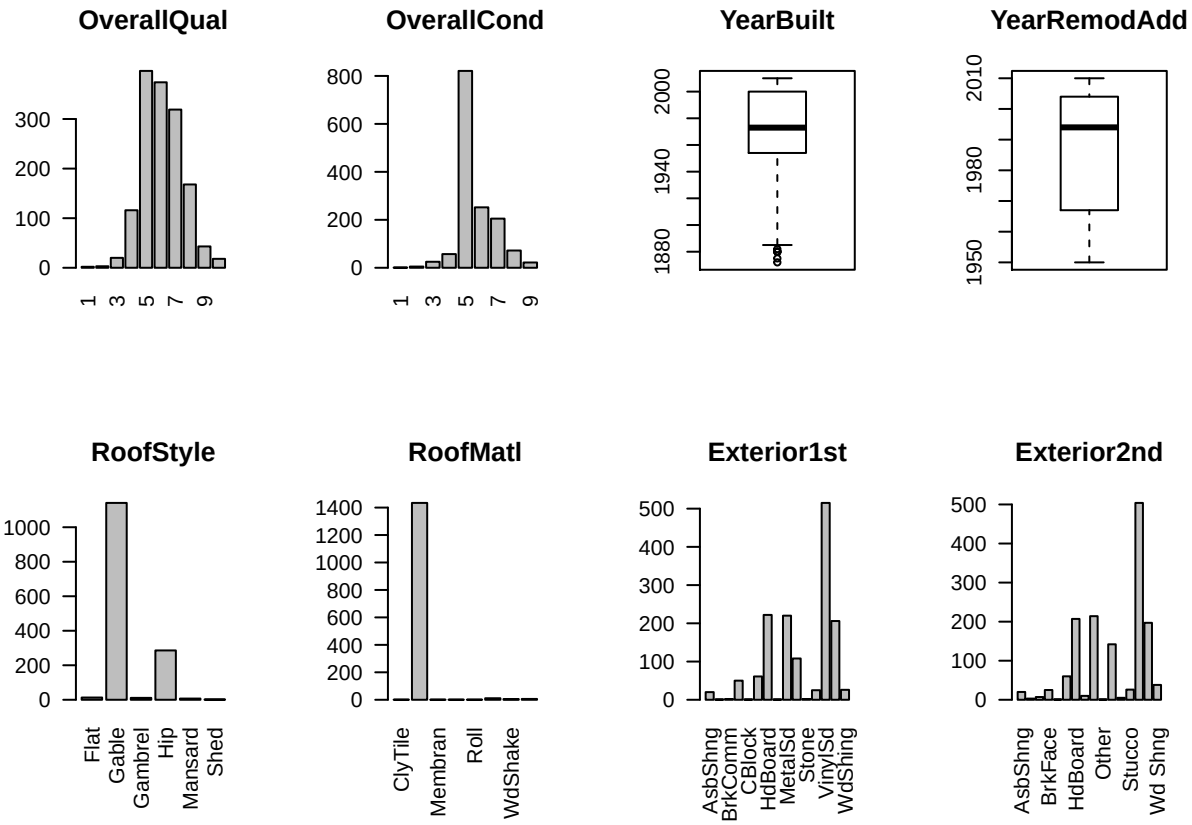
Observations

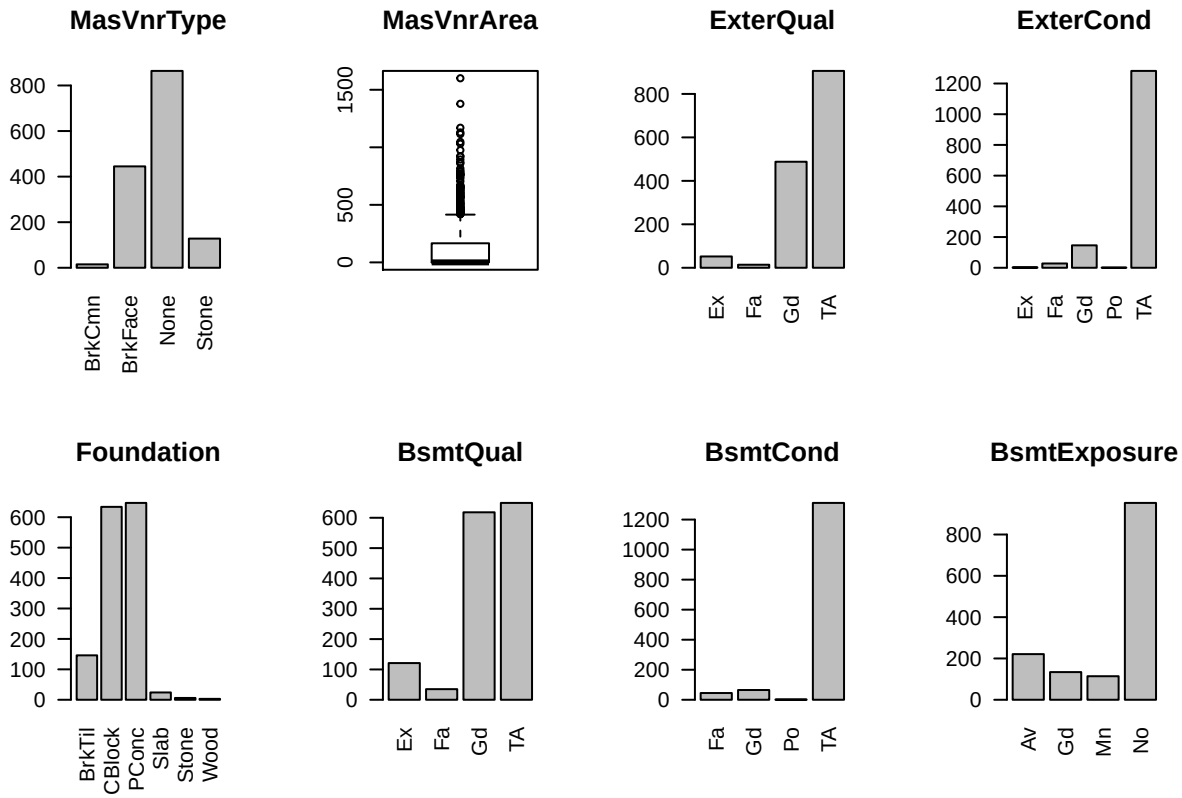
- Data related to garages report many NA values.
- Means of room counts might not be meaningful. For example, number of bathrooms. These might be better understood as ordinal categorical variables.

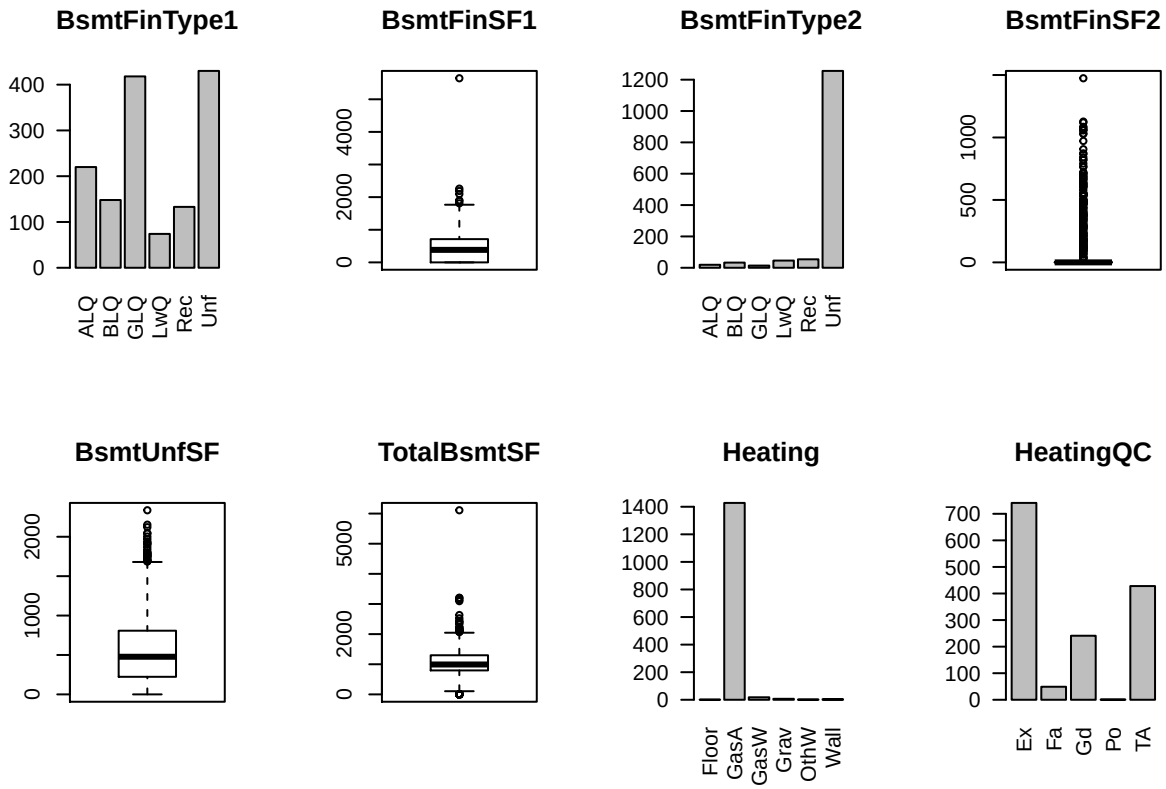
```
par.old <- par(no.readonly = T) #Save base plot parameters
par(mfrow = c(2, 4))

is_col_factor <- sapply(train, class) == "factor"
for (i in 2:length(train)) {
    if (is_col_factor[i]) {
        plot(train[, i], main = names(train[i]), las = 2)
    } else {
        boxplot(train[, i], main = names(train[i]))
    }
}
```

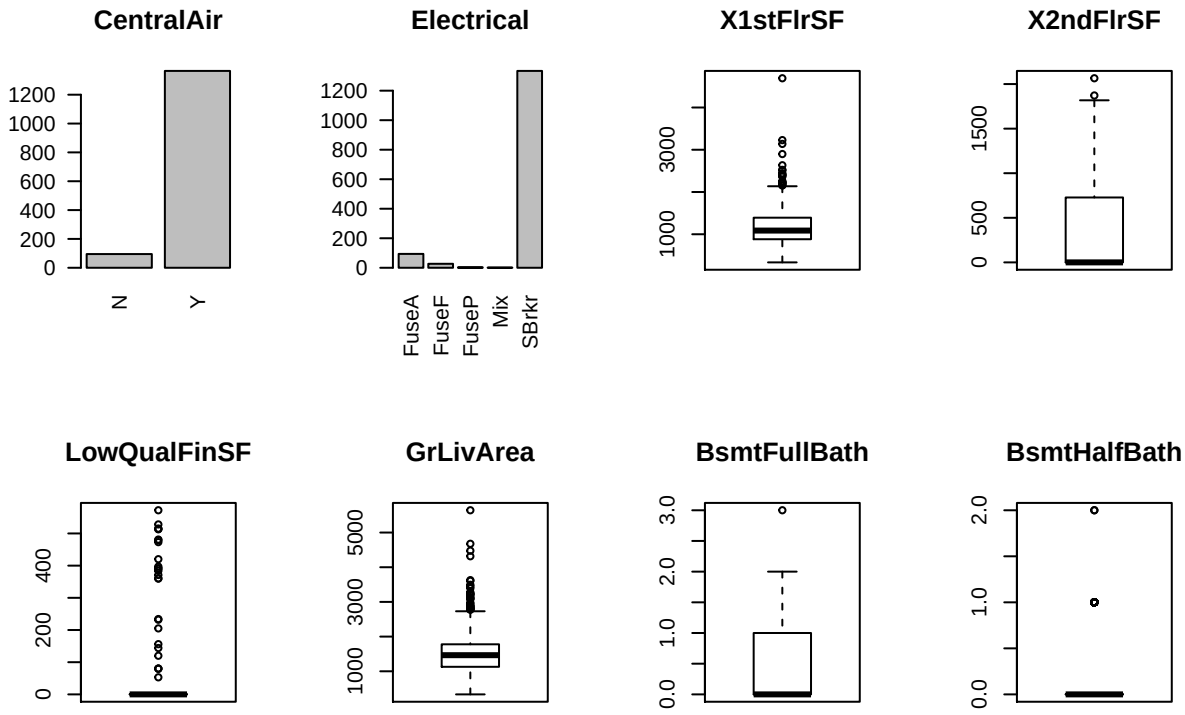


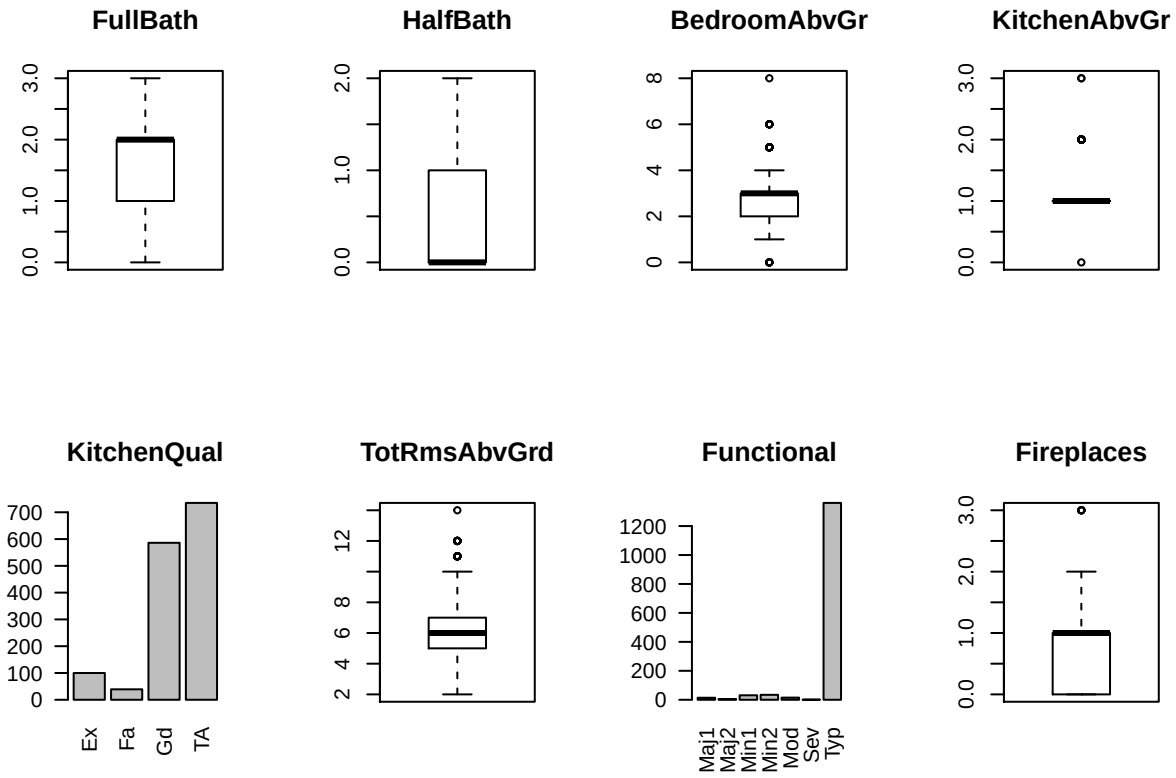

Utilities
1400
1200
1000
800
600
400
200
0
AllPub
NoSeWa
LotConfig
1000
800
600
400
200
0
Corner
CulDSac
FR2
FR3
Inside
LandSlope
1200
1000
800
600
400
200
0
Gtl
Mod
Sev
Neighborhood
200
150
100
50
0
Blmngtn
ClearCr
Gilbert
NAmes
NWAmes
Somerst
Veenker
Condition1
1200
1000
800
600
400
200
0
Artery
Norm
PosN
RRAn
RRNn
Condition2
1400
1200
1000
800
600
400
200
0
Artery
Norm
PosN
RRAn
BldgType
1200
1000
800
600
400
200
0
1Fam
2fmCon
Duplex
Twnhs
TwnhsE
HouseStyle
700
600
500
400
300
200
100
0
1.5Fin
1Story
2.5Unf
SFoyer

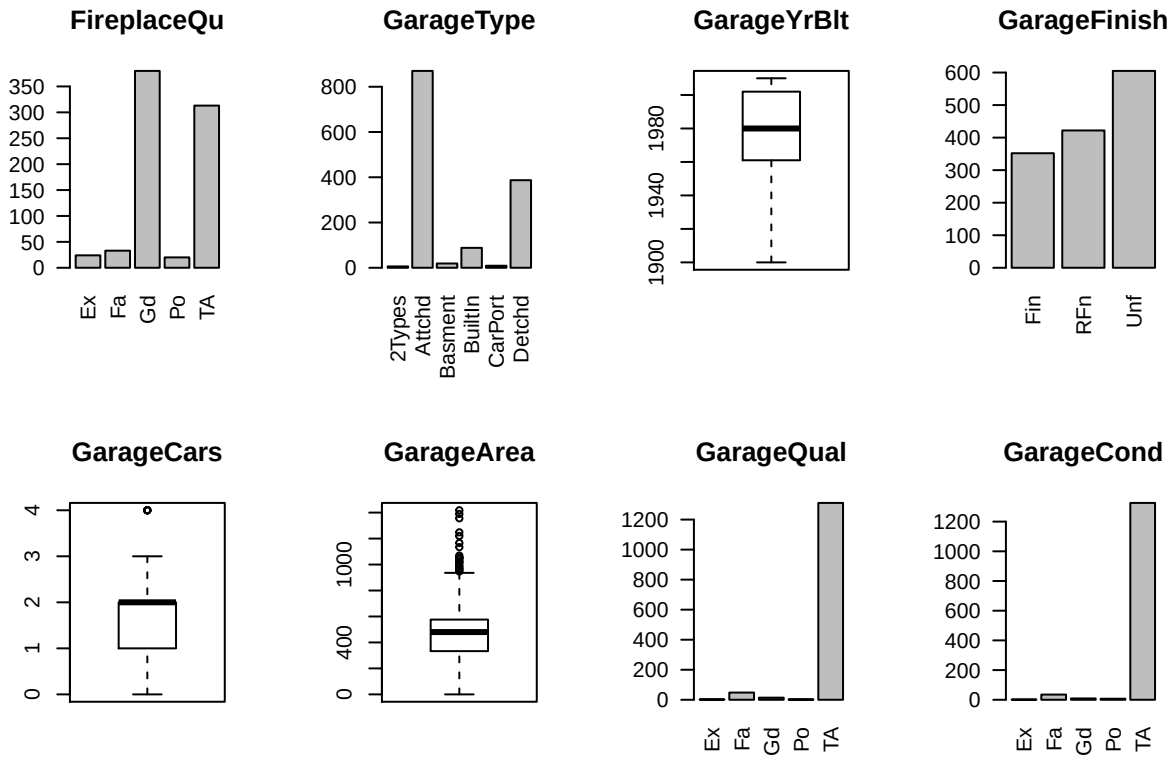

OverallQual
0
100
200
300
1
3
5
7
9
OverallCond
0
200
400
600
800
1
3
5
7
9
YearBuilt
1880
1940
2000
YearRemodAdd
1950
1980
2010
RoofStyle
0
200
400
600
800
1000
Flat
Gable
Gambrel
Hip
Mansard
Shed
RoofMatl
0
200
400
600
800
1000
1200
1400
ClyTile
Membran
Roll
WdShake
Exterior1st
0
100
200
300
400
500
AsbShng
BrkComm
CBlock
HdBoard
MetalSd
Stone
VinylSd
WdShing
Exterior2nd
0
100
200
300
400
500
AsbShng
BrkFace
HdBoard
Other
Stucco
Wd Shng

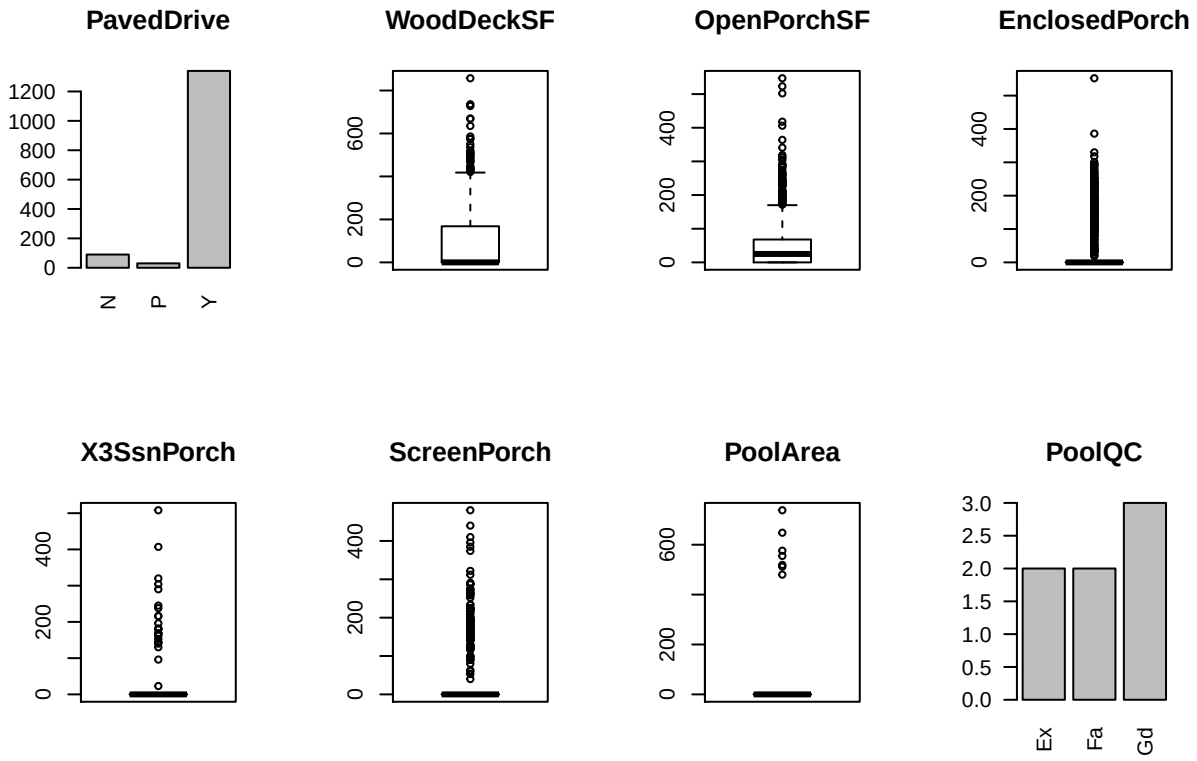

MasVnrType
800
600
400
200
0
BrkCmn
BrkFace
None
Stone
MasVnrArea
1500
500
0
ExterQual
800
600
400
200
0
Ex
Fa
Gd
TA
ExterCond
1200
1000
800
600
400
200
0
Ex
Fa
Gd
Po
TA
Foundation
600
500
400
300
200
100
0
BrkTil
CBlock
PConc
Slab
Stone
Wood
BsmtQual
600
500
400
300
200
100
0
Ex
Fa
Gd
TA
BsmtCond
1200
1000
800
600
400
200
0
Fa
Gd
Po
TA
BsmtExposure
800
600
400
200
0
Av
Gd
Mn
No

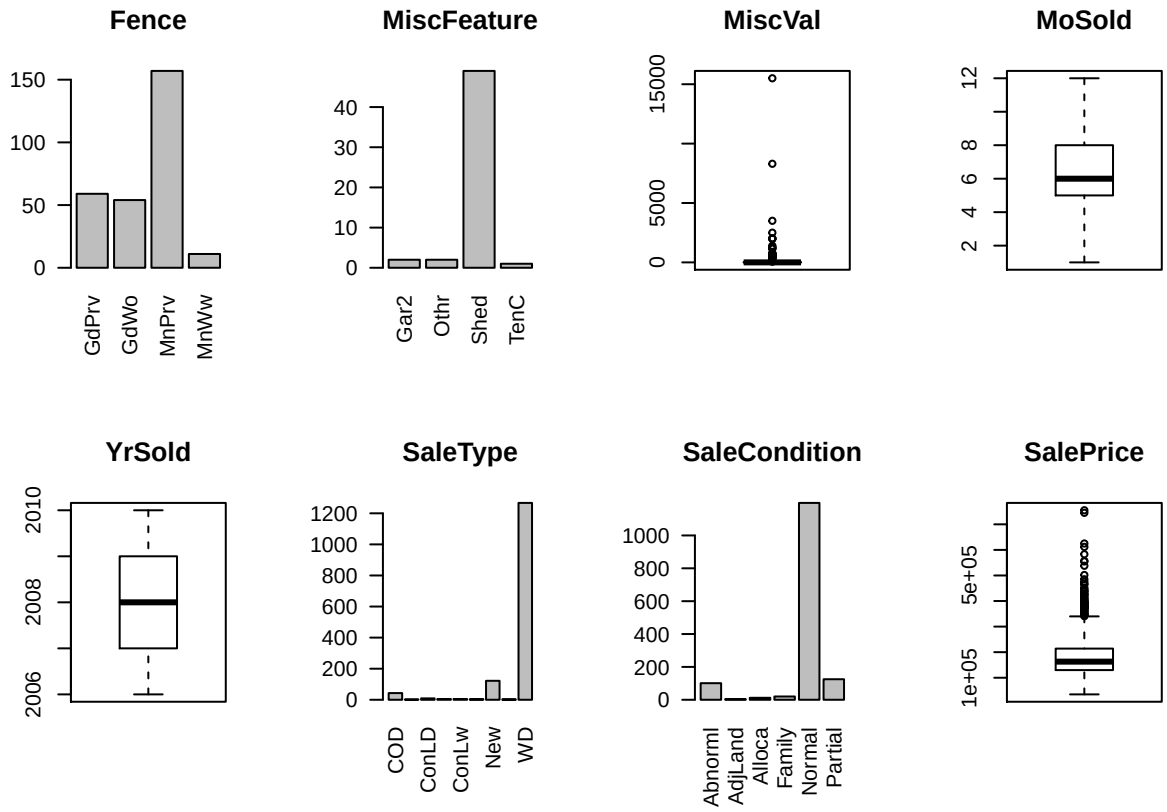

BsmtFinType1
400
300
200
100
0
ALQ
BLQ
GLQ
LwQ
Rec
Unf
BsmtFinSF1
4000
2000
0
BsmtFinType2
1200
1000
800
600
400
200
0
ALQ
BLQ
GLQ
LwQ
Rec
Unf
BsmtFinSF2
1000
500
0
BsmtUnfSF
2000
1000
0
TotalBsmtSF
5000
2000
0
Heating
1400
1200
1000
800
600
400
200
0
Floor
GasA
GasW
Grav
OthW
Wall
HeatingQC
700
600
500
400
300
200
100
0
Ex
Fa
Gd
Po
TA


CentralAir
1200
1000
800
600
400
200
0
N
Y
Electrical
1200
1000
800
600
400
200
0
FuseA
FuseF
FuseP
Mix
SBrkr
X1stFlrSF
3000
1000
X2ndFlrSF
1500
500
0
LowQualFinSF
400
200
0
GrLivArea
5000
3000
1000
BsmtFullBath
3.0
2.0
1.0
0.0
BsmtHalfBath
2.0
1.0
0.0


FullBath
0.0
1.0
2.0
3.0
HalfBath
0.0
1.0
2.0
BedroomAbvGr
0
2
4
6
8
KitchenAbvGr
0.0
1.0
2.0
3.0
KitchenQual
0
100
200
300
400
500
600
700
Ex
Fa
Gd
TA
TotRmsAbvGrd
2
4
6
8
12
Functional
0
200
400
600
800
1000
1200
Maj1
Maj2
Min1
Min2
Mod
Sev
Typ
Fireplaces
0.0
1.0
2.0
3.0


FireplaceQu
350
300
250
200
150
100
50
0
Ex
Fa
Gd
Po
TA
GarageType
800
600
400
200
0
2Types
Attchd
Basment
BuiltIn
CarPort
Detchd
GarageYrBlt
1900
1940
1980
GarageFinish
600
500
400
300
200
100
0
Fin
RFn
Unf
GarageCars
0
1
2
3
4
GarageArea
0
400
1000
GarageQual
1200
1000
800
600
400
200
0
Ex
Fa
Gd
Po
TA
GarageCond
1200
1000
800
600
400
200
0
Ex
Fa
Gd
Po
TA


PavedDrive
1200
1000
800
600
400
200
0
N
P
Y
WoodDeckSF
0
200
600
OpenPorchSF
0
200
400
EnclosedPorch
0
200
400
X3SsnPorch
0
200
400
ScreenPorch
0
200
400
PoolArea
0
200
600
PoolQC
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Ex
Fa
Gd



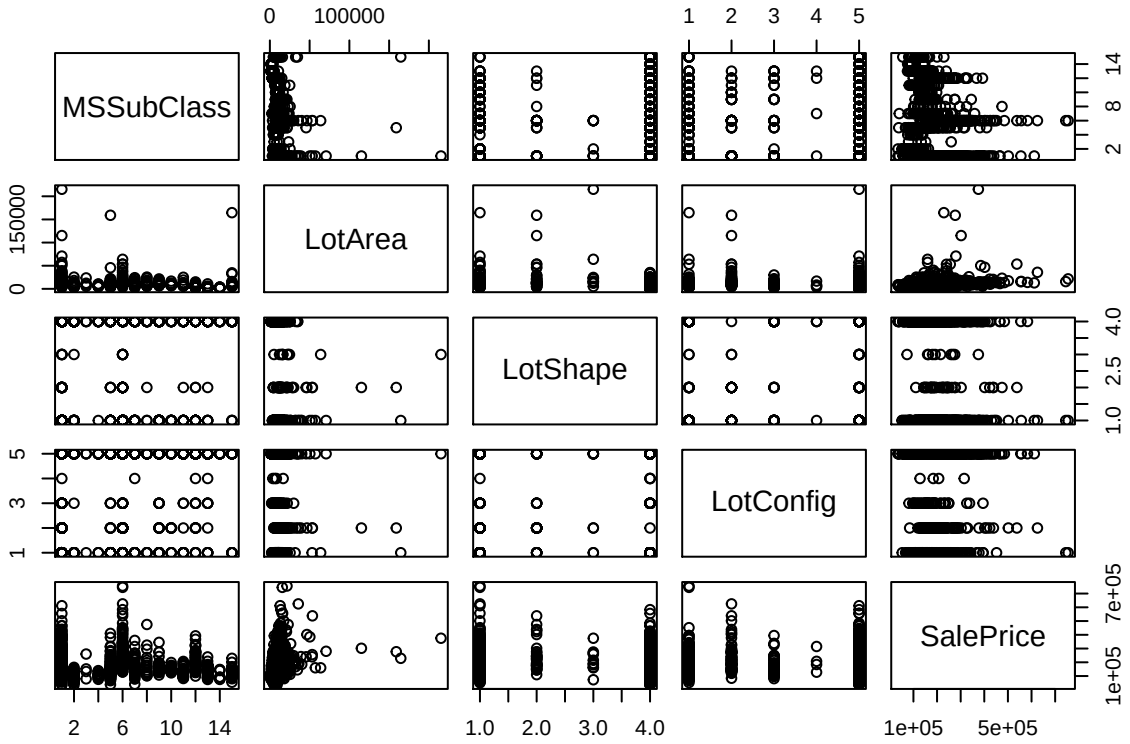

```
par(par.old) #Restore base plot parameters
```

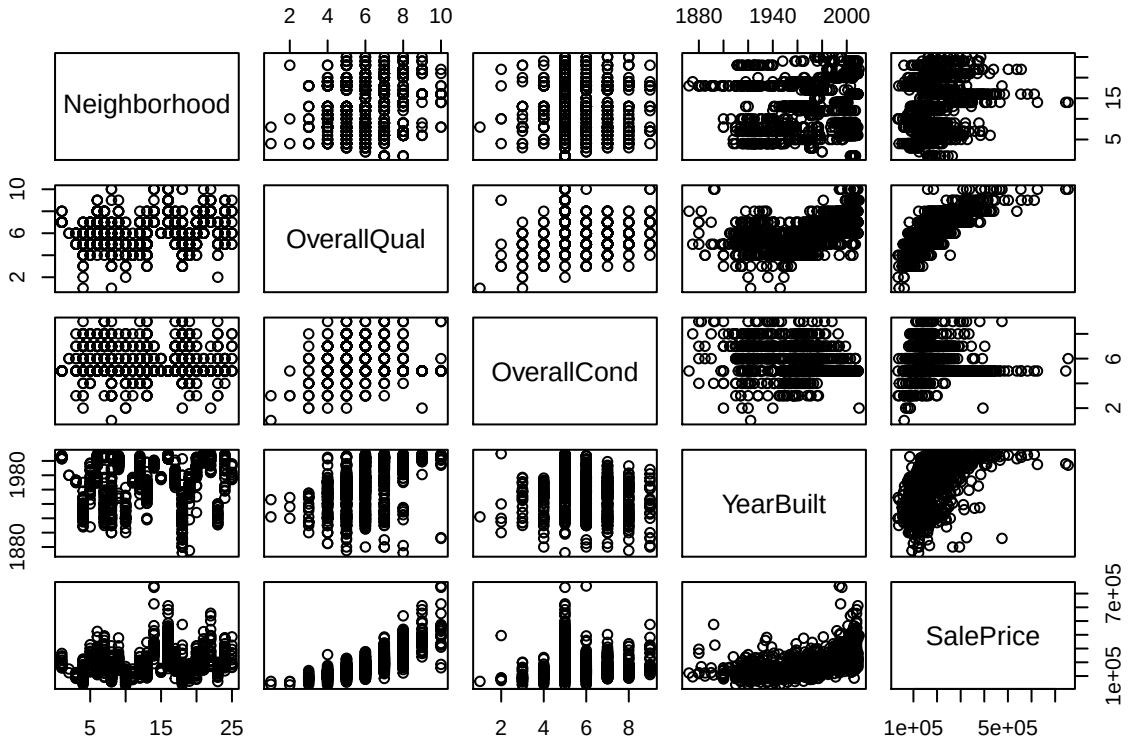
Provide a scatterplot matrix for at least two of the independent variables and the dependent variable

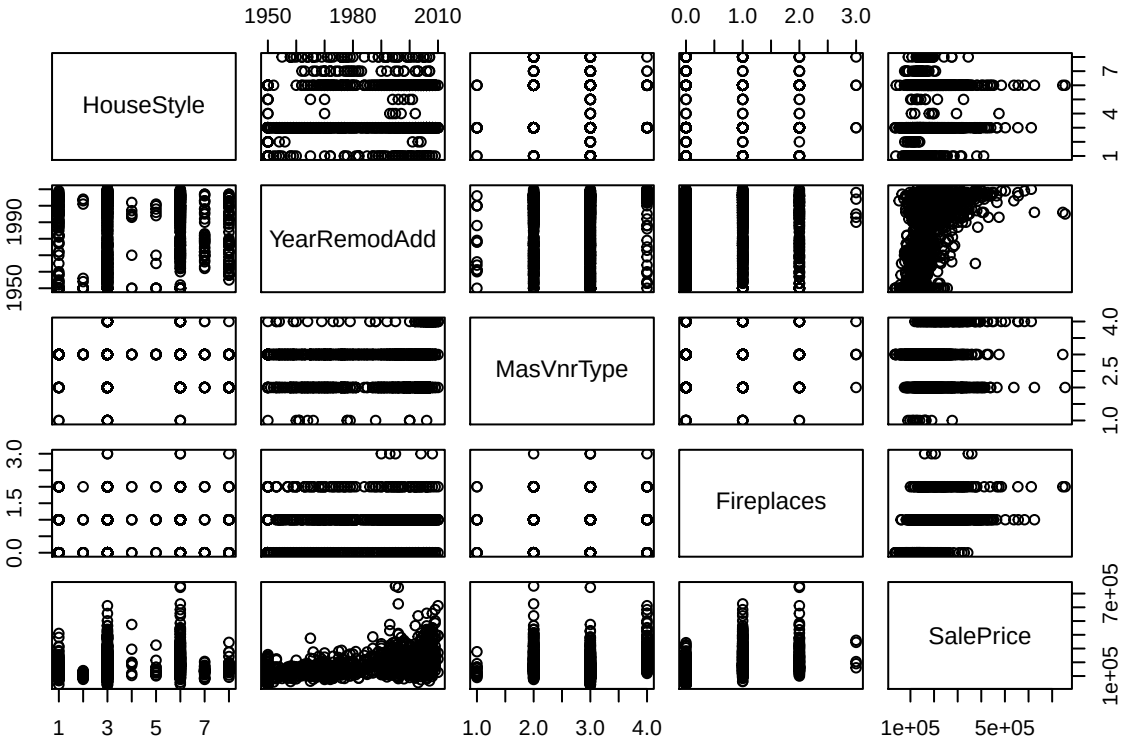
We're interested in potential predictors of a home's sale price. We examine some of the more obvious variables. In addition, upon review of the distribution plots, we examine some categorical variables with many levels and observations within them. The reasoning is that categories with few levels or a preponderance of observations on one level have less predictive influence.

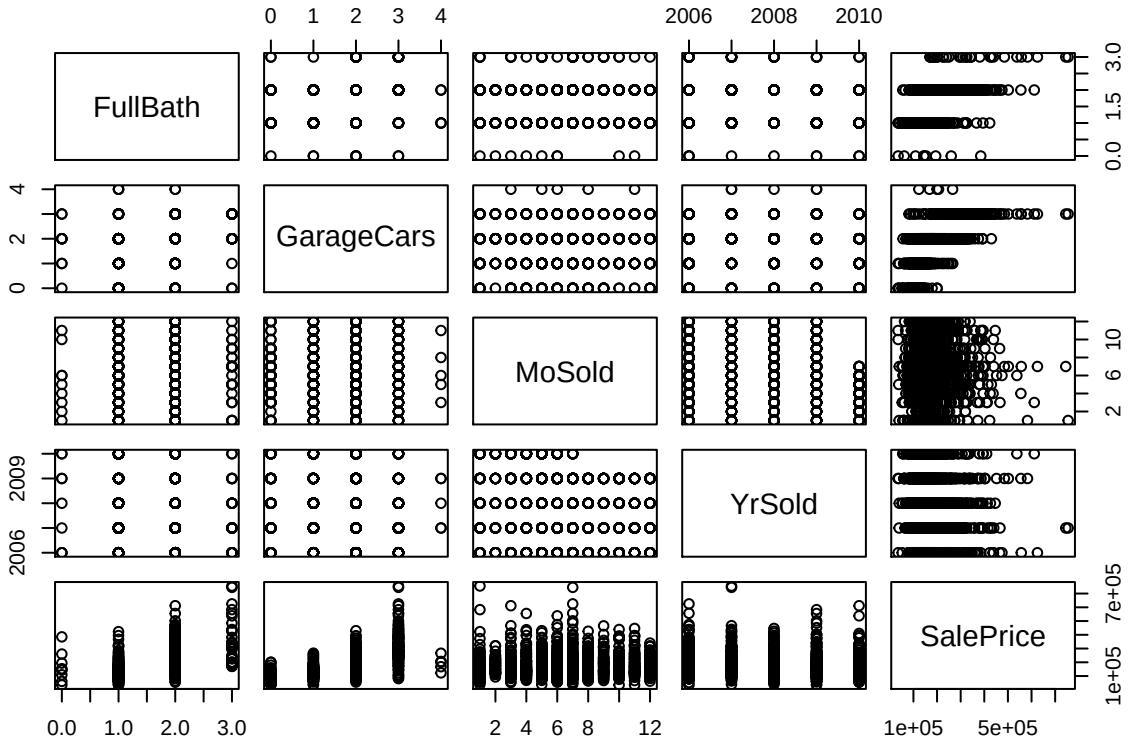
We made a list of columns of interest and looped through them, making pairs plots.

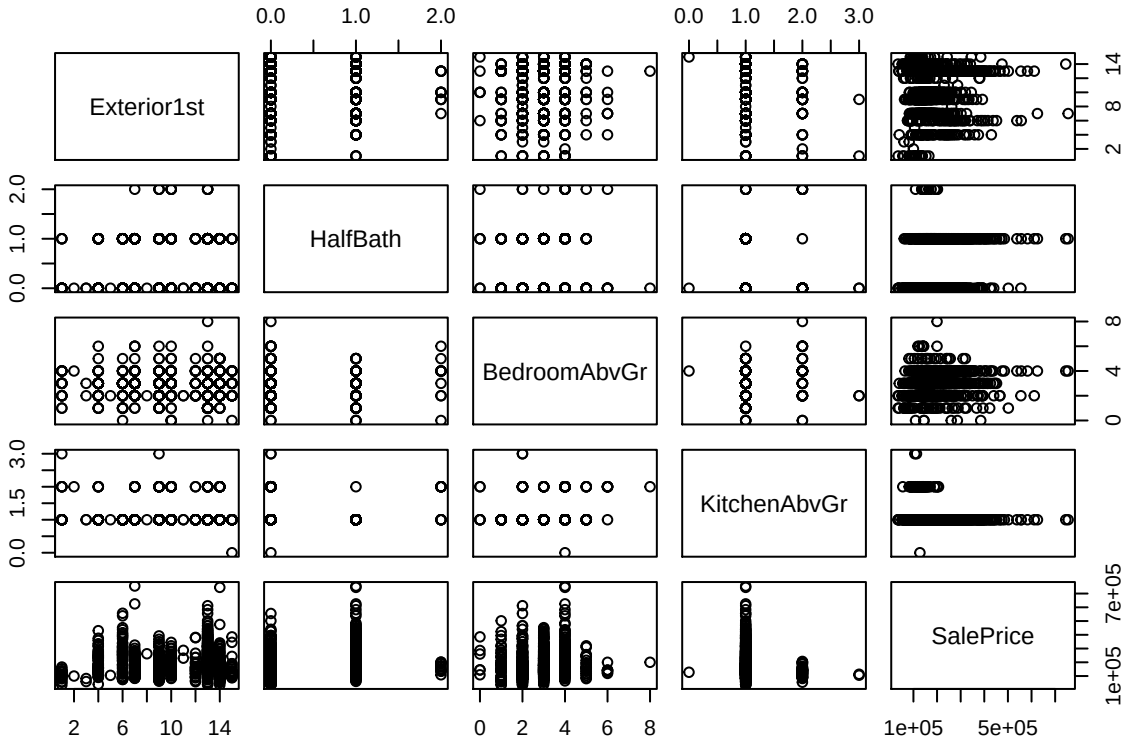
```
# Columns to pair against SalePrice.
pair_cols <- c("MSSubClass", "LotArea", "LotShape", "LotConfig",
               "Neighborhood", "OverallQual", "OverallCond", "YearBuilt",
               "YearRemodAdd", "MasVnrType", "Fireplaces", "HouseStyle",
               "YrSold", "MoSold", "GarageCars", "FullBath",
               "Exterior1st", "HalfBath", "BedroomAbvGr", "KitchenAbvGr",
               "TotRmsAbvGrd", "BsmtFinType1", "BsmtFinSF1", "TotalBsmtSF",
               "X1stFlrSF", "GrLivArea", "Exterior2nd", "Functional")

# Pair 4 columns and SalePrice at a time.
for(i in seq(from = 1, to = length(pair_cols), by = 4)) {
  pairs(train[which(colnames(train) %in% c(pair_cols[i:(i + 3)], "SalePrice"))])
}
```

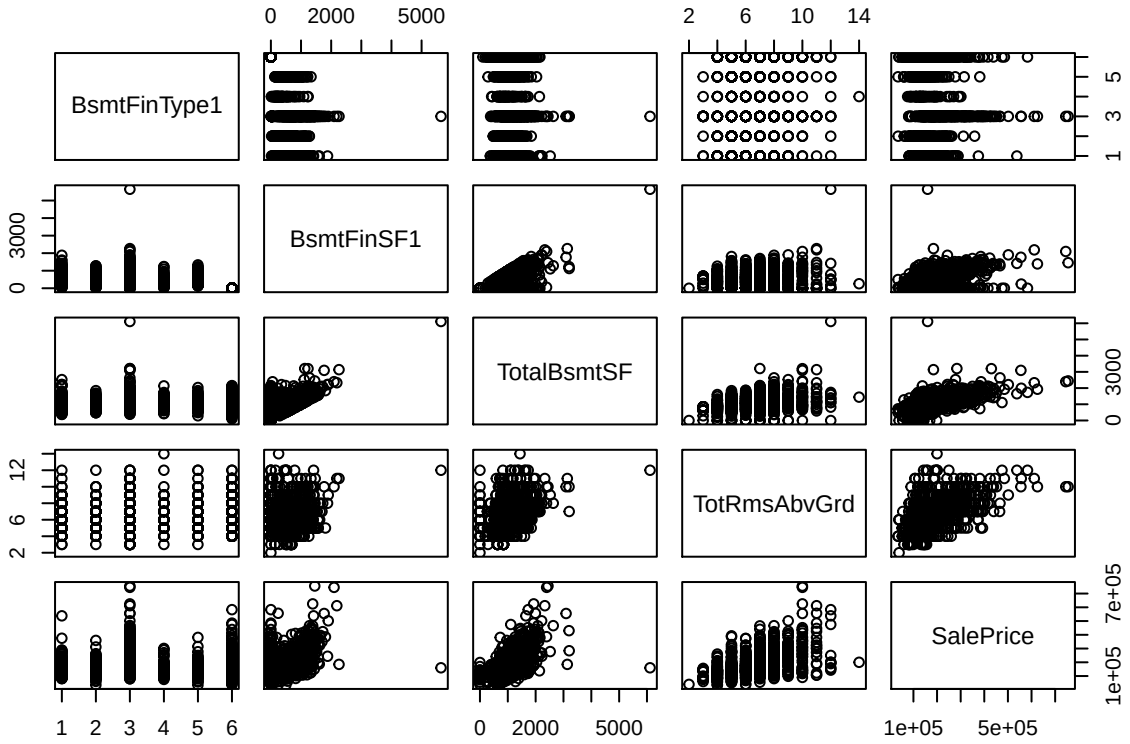

MSSubClass
LotArea
LotShape
LotConfig
SalePrice
0
100000
1
2
3
4
5
2
8
14
0
150000
1.0
2.5
4.0
1
3
5
1e+05
7e+05
2
6
10
14
1.0
2.0
3.0
4.0
1e+05
5e+05

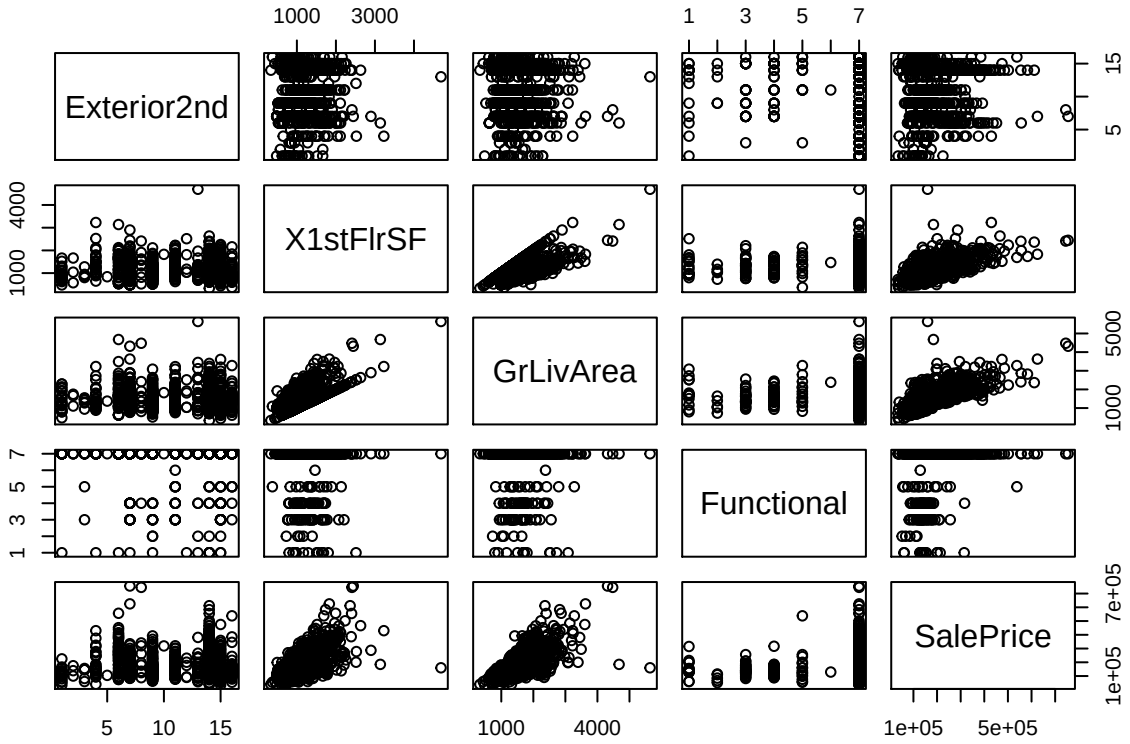

Neighborhood
OverallQual
OverallCond
YearBuilt
SalePrice
2 4 6 8 10
1880 1940 2000
5 15
2 6 10
2 6
1880 1980
5 15 25
2 4 6 8
1e+05 5e+05
1e+05 7e+05


HouseStyle
YearRemodAdd
MasVnrType
Fireplaces
SalePrice
1950
1980
2010
0.0
1.0
2.0
3.0
1
4
7
1990
1950
4.0
2.5
1.0
3.0
1.5
0.0
1
3
5
7
1.0
2.0
3.0
4.0
1e+05
7e+05
5e+05


FullBath
GarageCars
MoSold
YrSold
SalePrice
0 1 2 3 4
2006 2008 2010
0.0 1.5 3.0
0 2 4
2 6 10
2006 2009
1e+05 7e+05
0.0 1.0 2.0 3.0
2 4 6 8 12
1e+05 5e+05


Exterior1st
HalfBath
BedroomAbvGr
KitchenAbvGr
SalePrice
0.0
1.0
2.0
3.0
0
2
4
6
8
10
12
14
1.5
1e+05
5e+05
7e+05


BsmtFinType1
BsmtFinSF1
TotalBsmtSF
TotRmsAbvGrd
SalePrice
0 2000 5000
2 6 10 14
1 3 5
0 3000
0 3000
2 6 12
1e+05 7e+05
1 2 3 4 5 6
0 2000 5000
1e+05 5e+05



Observations

- The quantitative variables appear to correlate to SalePrice.
- Of the categorical variables, we are unsure how to interpret the relation of nominal variables to SalePrice. Nevertheless, the nominal categories appear to differentiate on SalePrice. Perhaps that indicates they are predictive.
- The quantitative variable LotArea appears to correlate to SalePrice. However, there are outliers that might merit elimination to reduce their influence on a regression line.

Derive a correlation matrix for any three quantitative variables in the dataset

```
cor_cols <- c("LotArea", "GrLivArea", "SalePrice")
train_cor <- cor(train[which(colnames(train) %in% cor_cols)])
train_cor
```

```
##             LotArea GrLivArea SalePrice
## LotArea   1.0000000 0.2631162 0.2638434
## GrLivArea 0.2631162 1.0000000 0.7086245
## SalePrice 0.2638434 0.7086245 1.0000000
```

Test the hypotheses that the correlations between each pairwise set of variables is 0 and provide an 80% confidence interval

For our three quantitative variables, LotArea, GrLivArea, and SalePrice there are $\binom{3}{2} = 3$ pairwise sets.

- LotArea ~ GrLivArea
- LotArea ~ SalePrice
- GrLivArea ~ SalePrice

Reference

- Hypothesis Testing: Correlations.

```
alpha <- 0.80
cor_LotArea_GrLivArea <- cor.test(train$LotArea, train$GrLivArea, conf.level = alpha)
cor_LotArea_SalePrice <- cor.test(train$LotArea, train$SalePrice, conf.level = alpha)
cor_GrLivArea_SalePrice <- cor.test(train$GrLivArea, train$SalePrice, conf.level = alpha)
cor_LotArea_GrLivArea
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$LotArea and train$GrLivArea
## t = 10.414, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.2315997 0.2940809
## sample estimates:
##       cor
## 0.2631162
```

```
cor_LotArea_SalePrice
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$LotArea and train$SalePrice
## t = 10.445, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.2323391 0.2947946
## sample estimates:
##       cor
## 0.2638434
```

```
cor_GrLivArea_SalePrice
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$GrLivArea and train$SalePrice
## t = 38.348, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.6915087 0.7249450
## sample estimates:
##       cor
## 0.7086245
```

Discuss the meaning of your analysis

Would you be worried about familywise error? Why or why not?

The issue of familywise error pertains to multiple hypothesis tests. When testing multiple hypotheses, the chance of observing rare events increases, which increases the likelihood of Type I errors.

If this were a study under our review, we would be concerned with the weak choice of confidence interval. However, these correlations lie far from an r value of 0, so the likelihood of familywise error is low.

It is interesting to observe the results using a more narrow confidence interval and application of the Bonferroni Correction.

```
alpha <- 0.95
m <- 3 # Hypothesis count
cor_LotArea_GrLivArea <- cor.test(train$LotArea, train$GrLivArea, conf.level = alpha / m)
cor_LotArea_SalePrice <- cor.test(train$LotArea, train$SalePrice, conf.level = alpha / m)
cor_GrLivArea_SalePrice <- cor.test(train$GrLivArea, train$SalePrice, conf.level = alpha / m)
cor_LotArea_GrLivArea
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$LotArea and train$GrLivArea
## t = 10.414, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 31.66667 percent confidence interval:
##  0.2531416 0.2730348
## sample estimates:
##       cor
## 0.2631162
```

```
cor_LotArea_SalePrice
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$LotArea and train$SalePrice
## t = 10.445, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 31.66667 percent confidence interval:
##  0.2538729 0.2737578
## sample estimates:
##       cor
## 0.2638434
```

```
cor_GrLivArea_SalePrice
```

```
##
##  Pearson's product-moment correlation
##
## data:  train$GrLivArea and train$SalePrice
## t = 38.348, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```

```
## 31.66667 percent confidence interval:
##  0.7032637 0.7139047
## sample estimates:
##       cor
## 0.7086245
```

The confidence intervals still lie far from 0 with p-values approaching 0. We conclude that there are significant correlations between these quantitative variables.

Reference

- Family-wise error rate on Wikipedia.
- Bonferroni correction

2.2: 5 points - Linear Algebra and Correlation

- *Invert your correlation matrix from above. (This is known as the precision matrix and contains variance inflation factors on the diagonal.)*

```
train_cor
```

```
##             LotArea GrLivArea SalePrice
## LotArea   1.0000000 0.2631162 0.2638434
## GrLivArea 0.2631162 1.0000000 0.7086245
## SalePrice 0.2638434 0.7086245 1.0000000
```

```
train_cor_inv <- solve(train_cor)
train_cor_inv
```

```
##              LotArea  GrLivArea  SalePrice
## LotArea    1.0884485 -0.1664868 -0.1692033
## GrLivArea -0.1664868  2.0340972 -1.3974846
## SalePrice -0.1692033 -1.3974846  2.0349350
```

- *Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix.*

```
train_cor %*% train_cor_inv
```

```
##                 LotArea     GrLivArea    SalePrice
## LotArea    1.000000e+00  0.000000e+00 0.000000e+00
## GrLivArea  1.387779e-17  1.000000e+00 2.220446e-16
## SalePrice -2.775558e-17 -2.220446e-16 1.000000e+00
```

```
train_cor_inv %*% train_cor
```

```
##                 LotArea   GrLivArea    SalePrice
## LotArea    1.000000e+00 1.94289e-16 1.387779e-16
## GrLivArea -1.110223e-16 1.00000e+00 0.000000e+00
## SalePrice -1.110223e-16 0.00000e+00 1.000000e+00
```

- *Conduct LU decomposition on the matrix.*

```
factor_lu <- function(A) {
    # Accept a matrix, A, and factor it into lower and upper
    # triangular matrices.
    # Assumptions:
    #   A is square.
    #   A row reduces to U in row echelon form without row permutation.

    L <- diag(nrow(A))
    for (j in 1:(ncol(A) - 1)) {
        # print(paste("j:", j))
        i <- j
        # print(paste("i:", i))
        for (k in i:(nrow(A) - 1)) {
            # print(paste("k:", k))
            multiplier <- -(A[k + 1, j] / A[i, j])
            A[k + 1, ] <- multiplier * A[i, ] + A[k + 1, ]
            L[k + 1, j] <- -multiplier
        }
    }

    return(list("L" = L, "U" = A))
}

factor_lu(train_cor)
```

```
## $L
##           [,1]      [,2] [,3]
## [1,] 1.0000000 0.0000000    0
## [2,] 0.2631162 1.0000000    0
## [3,] 0.2638434 0.6867466    1
##
## $U
##           LotArea GrLivArea SalePrice
## LotArea         1 0.2631162 0.2638434
## GrLivArea       0 0.9307699 0.6392030
## SalePrice       0 0.0000000 0.4914162
```

2.3: 5 points - Calculus-Based Probability & Statistics

Many times, it makes sense to fit a closed form distribution to data.

- *Select a variable in the Kaggle.com training dataset that is skewed to the right, shift it so that the minimum value is absolutely above zero if necessary.*

```
par(mfrow = c(1, 2))
boxplot(train$GrLivArea)
hist(train$GrLivArea)
```


```
min(train$GrLivArea)
```

```
## [1] 334
```

GrLivArea is skewed to the right and its minimum is 334.

- *Then load the MASS package and run* ***fitdistr*** *to fit an exponential probability density function. (See https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html).*

```
train_GrLivArea_fit <- fitdistr(x = train$GrLivArea, densfun = "exponential")
```

- *Find the optimal value of* λ *for this distribution, and then take 1000 samples from this exponential distribution using this value (e.g.,* ***rexp(1000, λ)****).*

```
lambda <- train_GrLivArea_fit$estimate
lambda
```

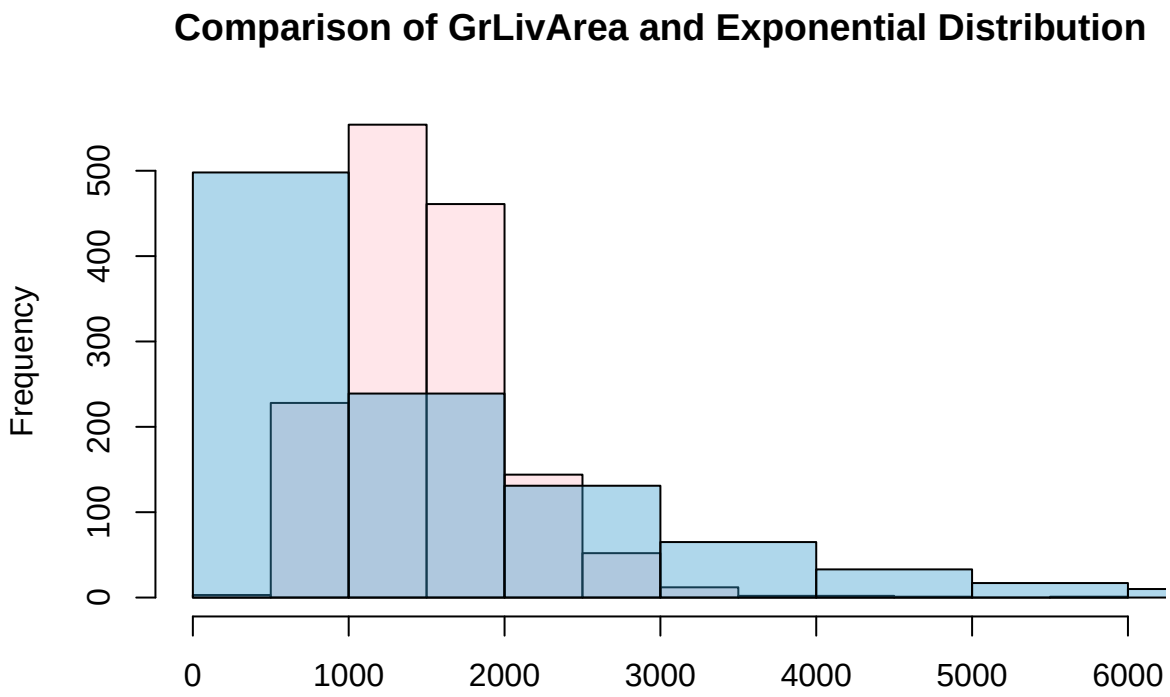
```
##        rate
## 0.000659864
```

```
train_GrLivArea_exp <- rexp(1000, lambda)
```

- *Plot a histogram and compare it with a histogram of your original variable.*

We overlaid the histogram of GrLivArea in pink and the Exponential Distribution in blue. They are both right-skewed, but we don't see what is meaningful about this simulation. The simulation is displaced, but even if we shift it to the right, the two curves don't match well.

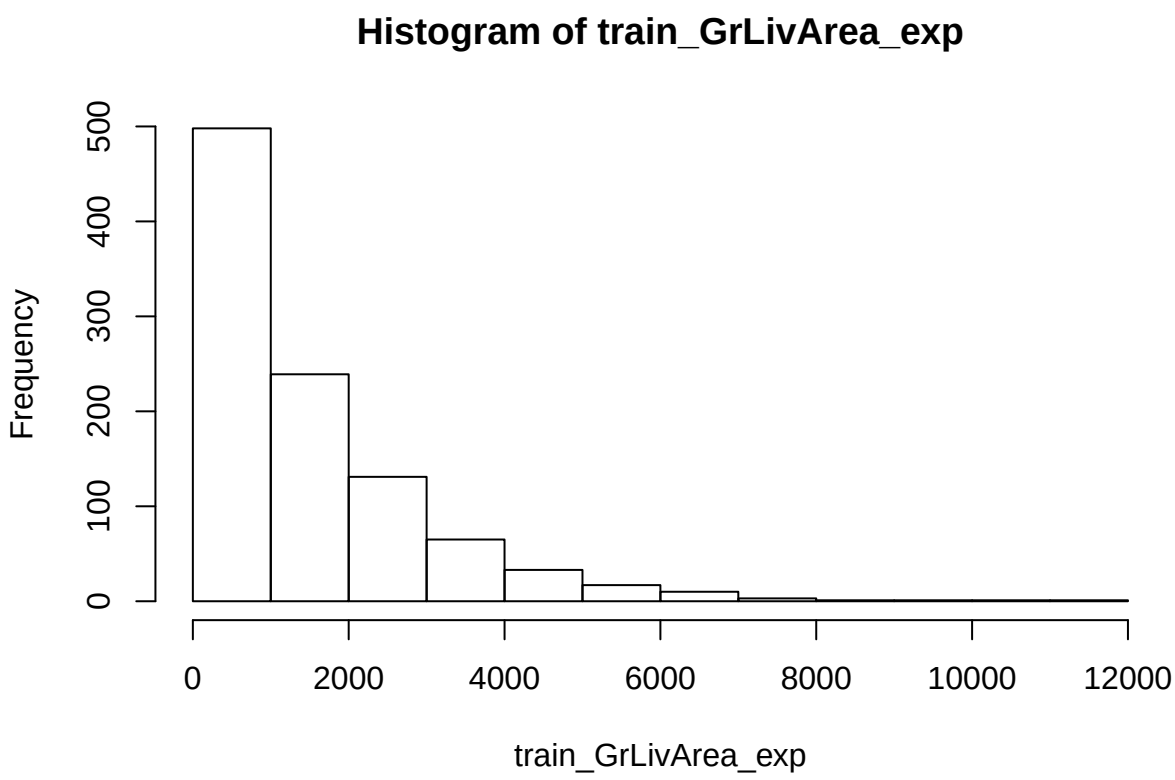
```
hist(train$GrLivArea, col = rgb(255, 192, 203, maxColorValue = 255, alpha = 100),
     main = "Comparison of GrLivArea and Exponential Distribution",
     xlab = "")
hist(train_GrLivArea_exp, add = T, col = rgb(50, 153, 204, maxColorValue = 255, alpha = 100))
```



- *Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF).*

The question is ambiguous, indicating to use both the exponential PDF and the CDF. We obtain the quantiles using `qexp()` and supply the value for the exponential distribution's rate λ which we computed when simulating GRLivArea.

```
hist(train_GrLivArea_exp)
```




```
qexp(c(0.05, 0.95), lambda)
```

```
## [1]   77.73313 4539.92351
```

- *Also generate a 95% confidence interval from the empirical data, assuming normality.*

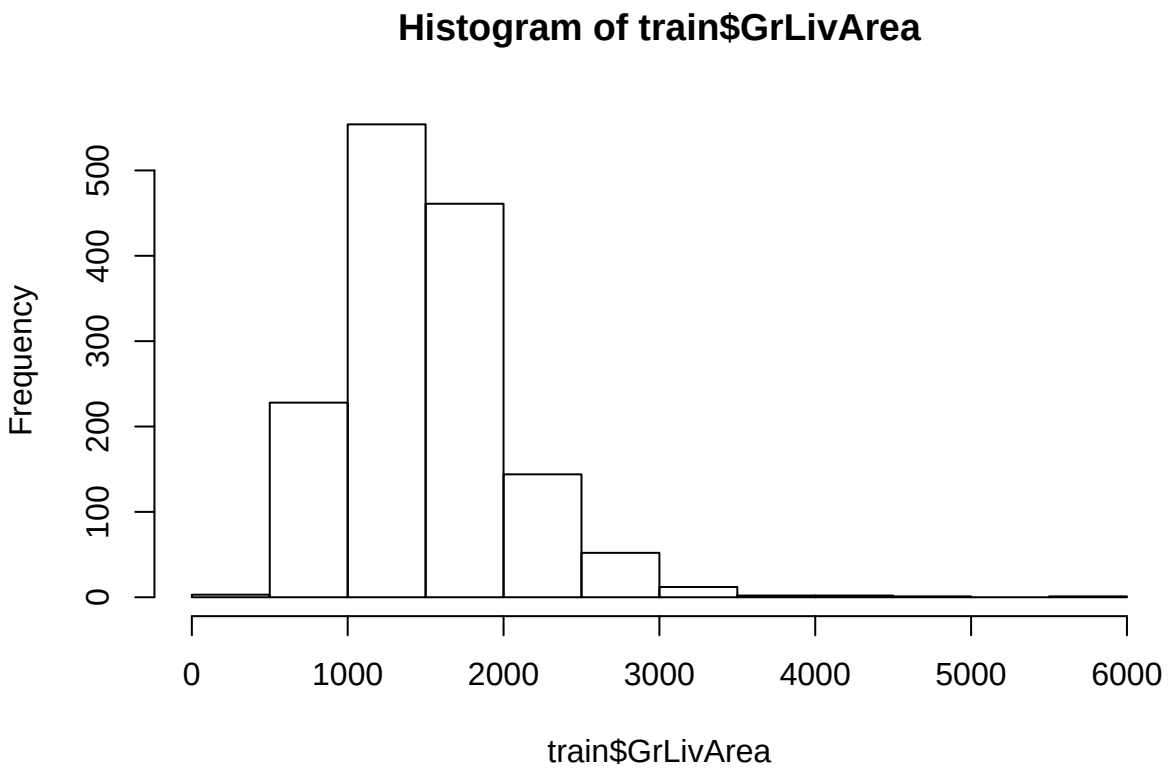
This question, too, is ambiguous, omitting specification of the point estimate. We compute the confidence interval for the mean.

```
CI(train$GrLivArea, ci = 0.95)
```

```
##    upper     mean    lower
## 1542.440 1515.464 1488.487
```

- *Finally, provide the empirical 5th percentile and 95th percentile of the data. Discuss.*

```
hist(train$GrLivArea)
```




```
quantile(train$GrLivArea, c(0.05, 0.95))
```

```
##     5%    95%
##  848.0 2466.1
```

The simulation omits the left tail of the empirical data and the 0.95 percentiles don't match. It appears such simulation requires a form of additional normalization.

2.4: 10 points - Modeling

Build some type of multiple regression model and submit your model to the competition board. Provide your complete model summary and results with analysis. Report your Kaggle.com user name and score.

We used forward selection to evaluate the performance of predictors and arrived at this model.

Note: The model is sensitive to new factor levels. After performing prediction, we removed some categorical variables.

```
model <- SalePrice ~ LotArea + LotShape + Neighborhood + OverallQual + YearBuilt + YearRemodAdd + MasVn
data <- train

summary(lm(model, data))
```

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

```
## lm(formula = model, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401716  -11686    -329   11331  244409
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -8.356e+05  1.681e+05  -4.972 7.45e-07 ***
## LotArea              5.536e-01  1.015e-01   5.454 5.82e-08 ***
## LotShapeIR2          9.984e+03  5.224e+03   1.911 0.056195 .
## LotShapeIR3         -4.433e+04  1.048e+04  -4.228 2.51e-05 ***
## LotShapeReg         -2.037e+03  1.936e+03  -1.052 0.293026
## NeighborhoodBlueste -5.218e+03  2.350e+04  -0.222 0.824303
## NeighborhoodBrDale  -1.318e+04  1.175e+04  -1.122 0.261991
## NeighborhoodBrkSide  1.257e+04  9.995e+03   1.257 0.208913
## NeighborhoodClearCr  1.947e+04  1.053e+04   1.849 0.064711 .
## NeighborhoodCollgCr  1.172e+04  8.302e+03   1.412 0.158121
## NeighborhoodCrawfor  3.754e+04  9.869e+03   3.804 0.000149 ***
## NeighborhoodEdwards -5.151e+03  9.215e+03  -0.559 0.576277
## NeighborhoodGilbert  3.365e+03  8.905e+03   0.378 0.705567
## NeighborhoodIDOTRR  -4.684e+03  1.055e+04  -0.444 0.657186
## NeighborhoodMeadowV -1.763e+04  1.280e+04  -1.378 0.168433
## NeighborhoodMitchel  3.041e+03  9.420e+03   0.323 0.746896
## NeighborhoodNAmes    8.378e+03  8.813e+03   0.951 0.341942
## NeighborhoodNoRidge  5.544e+04  9.753e+03   5.684 1.60e-08 ***
## NeighborhoodNPkVill  2.141e+03  1.598e+04   0.134 0.893438
## NeighborhoodNridgHt  3.798e+04  8.896e+03   4.270 2.09e-05 ***
## NeighborhoodNWAmes   4.414e+03  9.040e+03   0.488 0.625458
## NeighborhoodOldTown -4.968e+02  9.710e+03  -0.051 0.959200
## NeighborhoodSawyer   4.700e+03  9.356e+03   0.502 0.615529
## NeighborhoodSawyerW  1.020e+04  9.105e+03   1.120 0.262783
## NeighborhoodSomerst  1.878e+04  8.753e+03   2.145 0.032090 *
## NeighborhoodStoneBr  5.133e+04  1.049e+04   4.893 1.11e-06 ***
## NeighborhoodSWISU    1.095e+03  1.132e+04   0.097 0.922954
## NeighborhoodTimber   1.214e+04  9.580e+03   1.267 0.205333
## NeighborhoodVeenker  3.234e+04  1.270e+04   2.546 0.010989 *
## OverallQual2        -2.672e+03  2.874e+04  -0.093 0.925933
## OverallQual3         4.544e+03  2.308e+04   0.197 0.843959
## OverallQual4         1.494e+04  2.224e+04   0.672 0.501894
## OverallQual5         1.713e+04  2.219e+04   0.772 0.440301
## OverallQual6         2.063e+04  2.226e+04   0.927 0.354279
## OverallQual7         3.286e+04  2.243e+04   1.465 0.143201
## OverallQual8         5.916e+04  2.269e+04   2.607 0.009226 **
## OverallQual9         1.212e+05  2.332e+04   5.197 2.33e-07 ***
## OverallQual10        1.375e+05  2.434e+04   5.647 1.98e-08 ***
## YearBuilt            1.427e+02  7.091e+01   2.013 0.044320 *
## YearRemodAdd         2.902e+02  5.627e+01   5.158 2.86e-07 ***
## MasVnrTypeBrkFace    1.893e+04  8.337e+03   2.270 0.023348 *
## MasVnrTypeNone       1.889e+04  8.234e+03   2.294 0.021948 *
## MasVnrTypeStone      1.933e+04  8.841e+03   2.187 0.028934 *
## Fireplaces           6.769e+03  1.636e+03   4.138 3.72e-05 ***
## GarageCars           1.040e+04  1.556e+03   6.685 3.34e-11 ***
## FullBath             6.500e+03  2.621e+03   2.480 0.013238 *
```

```
## HalfBath             2.925e+03  2.452e+03   1.193 0.233199
## BedroomAbvGr        -2.181e+03  1.619e+03  -1.347 0.178074
## KitchenAbvGr        -2.740e+04  4.392e+03  -6.238 5.86e-10 ***
## TotRmsAbvGrd         3.845e+03  1.142e+03   3.367 0.000780 ***
## BsmtFinSF1           1.524e+01  2.314e+00   6.585 6.44e-11 ***
## TotalBsmtSF          7.810e+00  3.790e+00   2.061 0.039521 *
## X1stFlrSF            3.358e+00  4.967e+00   0.676 0.499113
## GrLivArea            3.244e+01  4.479e+00   7.243 7.25e-13 ***
## Exterior2ndAsphShn   1.054e+04  1.945e+04   0.542 0.587849
## Exterior2ndBrk Cmn  -1.261e+04  1.693e+04  -0.745 0.456364
## Exterior2ndBrkFace   1.448e+04  9.823e+03   1.474 0.140673
## Exterior2ndCBlock   -5.766e+02  3.168e+04  -0.018 0.985481
## Exterior2ndCmentBd   8.270e+03  9.211e+03   0.898 0.369441
## Exterior2ndHdBoard  -1.241e+02  7.834e+03  -0.016 0.987360
## Exterior2ndImStucc   2.158e+04  1.252e+04   1.723 0.085065 .
## Exterior2ndMetalSd  -8.999e+02  7.622e+03  -0.118 0.906033
## Exterior2ndOther     5.443e+03  3.243e+04   0.168 0.866733
## Exterior2ndPlywood   5.933e+01  7.968e+03   0.007 0.994061
## Exterior2ndStone     4.106e+03  1.745e+04   0.235 0.814030
## Exterior2ndStucco   -1.648e+04  9.666e+03  -1.705 0.088482 .
## Exterior2ndVinylSd   5.378e+03  7.783e+03   0.691 0.489691
## Exterior2ndWd Sdng   2.483e+03  7.649e+03   0.325 0.745483
## Exterior2ndWd Shng  -5.037e+03  8.924e+03  -0.564 0.572538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30650 on 1383 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.8576, Adjusted R-squared:  0.8506
## F-statistic: 122.5 on 68 and 1383 DF,  p-value: < 2.2e-16
```

```
require(car)
```

```
## Loading required package: car
```

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

```
require(lmtest)
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
require(Hmisc)
```

```
## Loading required package: Hmisc
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:plyr':
##
##     is.discrete, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
require(ResourceSelection)
```

```
## Loading required package: ResourceSelection
```

```
## ResourceSelection 0.3-5   2019-07-22
```

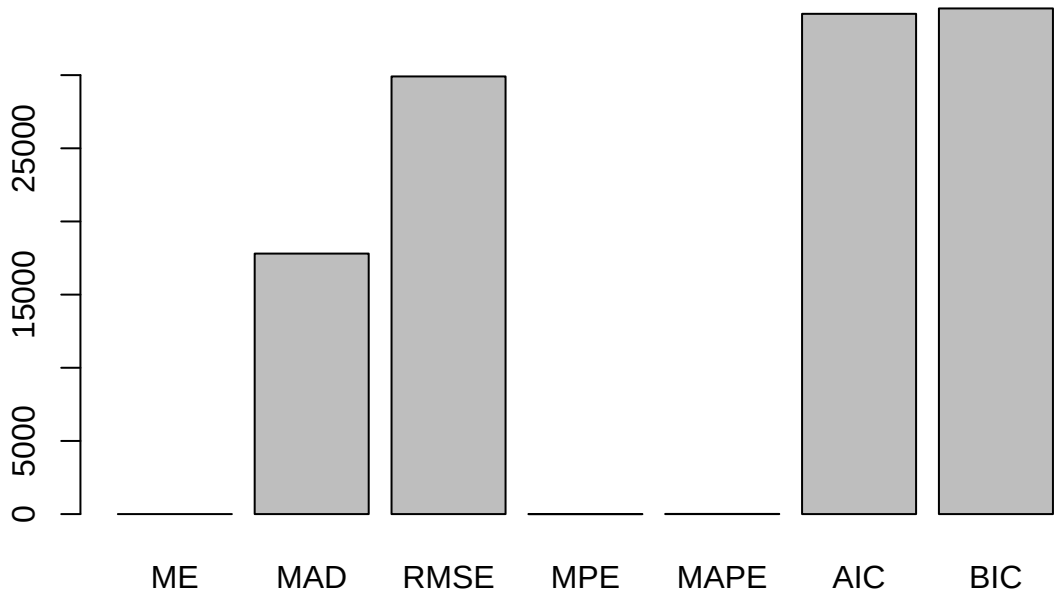
```
#specify model, data, and level
#model <- SalePrice ~ GrLivArea + TotRmsAbvGrd
#data <- train
level<-.95

#function
regressit <- function(model,data,level) #model is the functional relationship among variables
{
#   kdepairs(data) #scatterplot
    lm.fit <- lm(model, data, y = TRUE) #runregression
    confint(lm.fit, level = level ) #confidenceintervals
    r1 <- summary(lm.fit) #regression summary
    r2 <- anova(lm.fit) #ANOVA tableforregression
    r3 <- coefficients(lm.fit) #coefficients
    r4 <- lm.fit$residuals #residuals
    r5 <- shapiro.test(r4) #test for normality of residuals
    r6 <- bptest(lm.fit)  #Breusch-Pagan-Godfrey-Test for homoscedasticity, lmtest
    r7a <- ncvTest(lm.fit) #non-constant variance test for homoscedasticity
    r7 <- dwtest(lm.fit) #independence of residuals #test for independence of errors
    #r8 <- runs.test(lm.fit$residuals) #test for randomness of errors, lawstat
    #r9 <- vif(lm.fit) #look at collinearity, car package
    me <- mean(lm.fit$residuals)
    mad <- mean(abs(lm.fit$residuals))
```

```
    mse <- mean(lm.fit$residuals^2)
    rmse <- sqrt(mse)
    mpe <- 100*mean(lm.fit$residuals / lm.fit$y)
    mape <- 100*mean(abs(lm.fit$residuals) / lm.fit$y)
    aic <- AIC(lm.fit)
    bic <- BIC(lm.fit)
    all <- c(me,mad,rmse,mpe,mape,aic,bic)
    names(all) <- c("ME","MAD","RMSE","MPE","MAPE", "AIC","BIC")
    #names(r9) <- c("VIF")
    barplot_all <- barplot(all)
    par(mfrow = c(2,2))
    diagnostics <- plot(lm.fit)
    #rlist <- list(r1,r2,r3,r4,r5,r6,r7,r7a,r9,all)
    rlist <- list(lm.fit, r1,r2,r3,r5,r6,r7,r7a,all) #Omit residuals
    return(rlist)
}
```

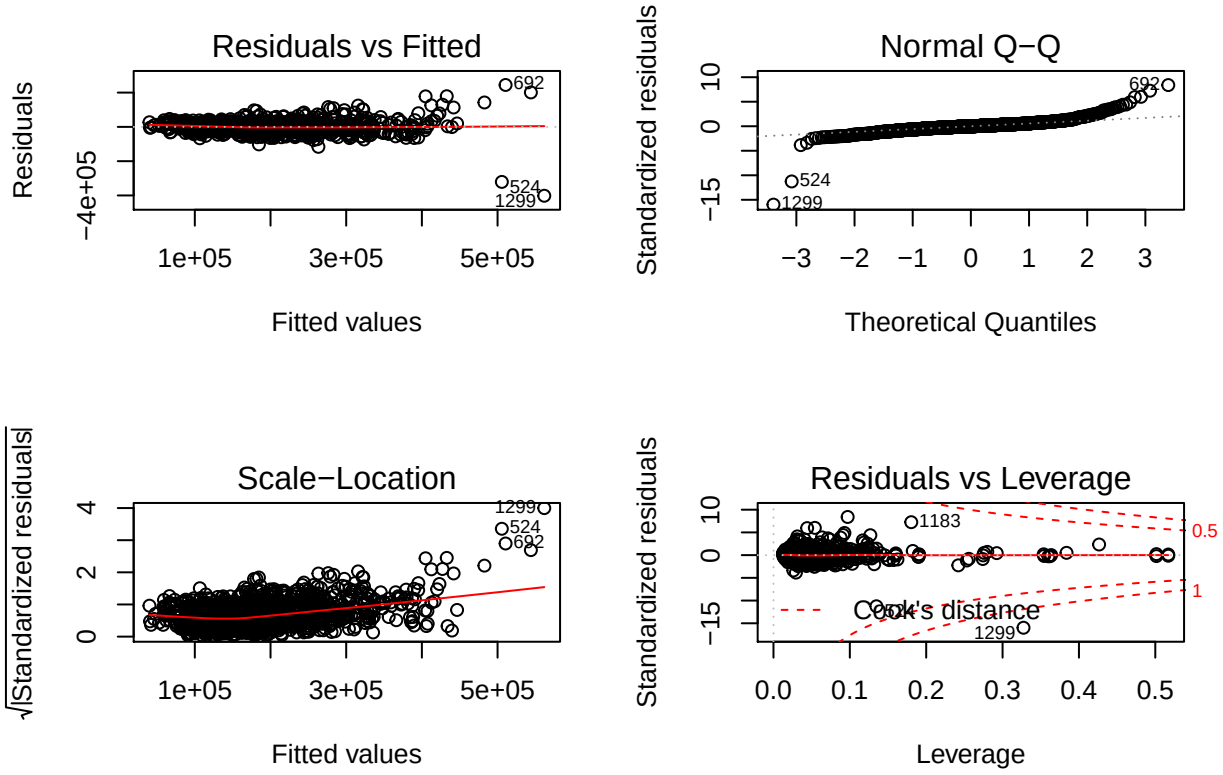
Here is a more extensive report, though understanding of the interpretation of some of these metrics will wait for now.

```
set.seed(2020)
lm_results <- regressit(model,data,level)
```

```
## Warning: not plotting observations with leverage one:
##   594, 1363
```



```
## Warning: not plotting observations with leverage one:
##   594, 1363
```



Summary

```
print(lm_results[[2]])
```

```
##
## Call:
## lm(formula = model, data = data, y = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401716  -11686    -329   11331  244409
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -8.356e+05  1.681e+05  -4.972 7.45e-07 ***
## LotArea                5.536e-01  1.015e-01   5.454 5.82e-08 ***
## LotShapeIR2            9.984e+03  5.224e+03   1.911 0.056195 .
## LotShapeIR3           -4.433e+04  1.048e+04  -4.228 2.51e-05 ***
## LotShapeReg           -2.037e+03  1.936e+03  -1.052 0.293026
## NeighborhoodBlueste -5.218e+03  2.350e+04  -0.222 0.824303
## NeighborhoodBrDale  -1.318e+04  1.175e+04  -1.122 0.261991
```

```
## NeighborhoodBrkSide  1.257e+04  9.995e+03   1.257 0.208913
## NeighborhoodClearCr  1.947e+04  1.053e+04   1.849 0.064711 .
## NeighborhoodCollgCr  1.172e+04  8.302e+03   1.412 0.158121
## NeighborhoodCrawfor  3.754e+04  9.869e+03   3.804 0.000149 ***
## NeighborhoodEdwards -5.151e+03  9.215e+03  -0.559 0.576277
## NeighborhoodGilbert  3.365e+03  8.905e+03   0.378 0.705567
## NeighborhoodIDOTRR  -4.684e+03  1.055e+04  -0.444 0.657186
## NeighborhoodMeadowV -1.763e+04  1.280e+04  -1.378 0.168433
## NeighborhoodMitchel  3.041e+03  9.420e+03   0.323 0.746896
## NeighborhoodNAmes    8.378e+03  8.813e+03   0.951 0.341942
## NeighborhoodNoRidge  5.544e+04  9.753e+03   5.684 1.60e-08 ***
## NeighborhoodNPkVill  2.141e+03  1.598e+04   0.134 0.893438
## NeighborhoodNridgHt  3.798e+04  8.896e+03   4.270 2.09e-05 ***
## NeighborhoodNWAmes   4.414e+03  9.040e+03   0.488 0.625458
## NeighborhoodOldTown -4.968e+02  9.710e+03  -0.051 0.959200
## NeighborhoodSawyer   4.700e+03  9.356e+03   0.502 0.615529
## NeighborhoodSawyerW  1.020e+04  9.105e+03   1.120 0.262783
## NeighborhoodSomerst  1.878e+04  8.753e+03   2.145 0.032090 *
## NeighborhoodStoneBr  5.133e+04  1.049e+04   4.893 1.11e-06 ***
## NeighborhoodSWISU    1.095e+03  1.132e+04   0.097 0.922954
## NeighborhoodTimber   1.214e+04  9.580e+03   1.267 0.205333
## NeighborhoodVeenker  3.234e+04  1.270e+04   2.546 0.010989 *
## OverallQual2        -2.672e+03  2.874e+04  -0.093 0.925933
## OverallQual3         4.544e+03  2.308e+04   0.197 0.843959
## OverallQual4         1.494e+04  2.224e+04   0.672 0.501894
## OverallQual5         1.713e+04  2.219e+04   0.772 0.440301
## OverallQual6         2.063e+04  2.226e+04   0.927 0.354279
## OverallQual7         3.286e+04  2.243e+04   1.465 0.143201
## OverallQual8         5.916e+04  2.269e+04   2.607 0.009226 **
## OverallQual9         1.212e+05  2.332e+04   5.197 2.33e-07 ***
## OverallQual10        1.375e+05  2.434e+04   5.647 1.98e-08 ***
## YearBuilt            1.427e+02  7.091e+01   2.013 0.044320 *
## YearRemodAdd         2.902e+02  5.627e+01   5.158 2.86e-07 ***
## MasVnrTypeBrkFace    1.893e+04  8.337e+03   2.270 0.023348 *
## MasVnrTypeNone       1.889e+04  8.234e+03   2.294 0.021948 *
## MasVnrTypeStone      1.933e+04  8.841e+03   2.187 0.028934 *
## Fireplaces           6.769e+03  1.636e+03   4.138 3.72e-05 ***
## GarageCars           1.040e+04  1.556e+03   6.685 3.34e-11 ***
## FullBath             6.500e+03  2.621e+03   2.480 0.013238 *
## HalfBath             2.925e+03  2.452e+03   1.193 0.233199
## BedroomAbvGr        -2.181e+03  1.619e+03  -1.347 0.178074
## KitchenAbvGr        -2.740e+04  4.392e+03  -6.238 5.86e-10 ***
## TotRmsAbvGrd         3.845e+03  1.142e+03   3.367 0.000780 ***
## BsmtFinSF1           1.524e+01  2.314e+00   6.585 6.44e-11 ***
## TotalBsmtSF          7.810e+00  3.790e+00   2.061 0.039521 *
## X1stFlrSF            3.358e+00  4.967e+00   0.676 0.499113
## GrLivArea            3.244e+01  4.479e+00   7.243 7.25e-13 ***
## Exterior2ndAsphShn   1.054e+04  1.945e+04   0.542 0.587849
## Exterior2ndBrk Cmn  -1.261e+04  1.693e+04  -0.745 0.456364
## Exterior2ndBrkFace   1.448e+04  9.823e+03   1.474 0.140673
## Exterior2ndCBlock   -5.766e+02  3.168e+04  -0.018 0.985481
## Exterior2ndCmentBd   8.270e+03  9.211e+03   0.898 0.369441
## Exterior2ndHdBoard  -1.241e+02  7.834e+03  -0.016 0.987360
## Exterior2ndImStucc   2.158e+04  1.252e+04   1.723 0.085065 .
```

```
## Exterior2ndMetalSd  -8.999e+02  7.622e+03  -0.118 0.906033
## Exterior2ndOther     5.443e+03  3.243e+04   0.168 0.866733
## Exterior2ndPlywood   5.933e+01  7.968e+03   0.007 0.994061
## Exterior2ndStone     4.106e+03  1.745e+04   0.235 0.814030
## Exterior2ndStucco   -1.648e+04  9.666e+03  -1.705 0.088482 .
## Exterior2ndVinylSd   5.378e+03  7.783e+03   0.691 0.489691
## Exterior2ndWd Sdng   2.483e+03  7.649e+03   0.325 0.745483
## Exterior2ndWd Shng  -5.037e+03  8.924e+03  -0.564 0.572538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30650 on 1383 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.8576, Adjusted R-squared:  0.8506
## F-statistic: 122.5 on 68 and 1383 DF,  p-value: < 2.2e-16
```

Analysis

- There is some bias in the residuals, with a median value of 189.
- F statistic. The value (34.57) is distant from 0 and its p-value is near zero. We reject the null hypothesis, which states the variability in the data is due to chance.
- R^2. The value of 0.878 means that the model explains 28% of the variability in the data.
- Residual standard error. The model is off, on average by 30440 dollars.
- The ratio for the estimate of the intercept to its standard error is -0.9471269, below the desired range of 5-10 which indicates instability.
- The diagnostic plot Residuals vs. Fitted shows a good result for the model. The line is flat with no discernable pattern until some outlier values at the high end.
- The Normal Q-Q plot reveals some non-linearity above 2 standard deviations.
- The Scale-Location plot shows mostly consistent variance. Again, there is a departure at the high end.
- The Residuals vs Leverage plate reveals a few outlier values that exert influence on the regression. They might be data quality issues. If not, they might merit scrubbing to improve the predictability of the model.

Prediction

It was interesting to discover that our predictions had missing values. For now, we imputed those with the mean prediction.

```
set.seed(2020)
lm.fit <- lm_results[[1]]
predictions <- predict(lm.fit, test)

pred_mean <- mean(predictions, na.rm = T)
is_predictions_na <- is.na(predictions)
predictions[is_predictions_na] <- pred_mean

submission <- as.data.frame(cbind(Id = test[1], SalePrice = predictions))
write.csv(x = submission, file = "submission.csv", row.names = F, quote = F)
```

Kaggle

- User name: pnojai

- Score: 0.16412
- Position on leaderboard: 3604