Hickman_Homework6

Keith Hickman

November 12, 2017

Intro/Data Discussion:

This dataset comes from home sales in Ames, Iowa from 2006 to 2010, a four year period. There are 1460 observations of 81 variables. The variables are nominal, ordinal, continuous, and discrete.

I found help from several Kaggle submissions. #Load Libraries and Dependencies:

```
##Load Libraries, Data, and perform Initial Analysis
library(data.table)
library(DMwR2)
library(ggplot2)
library(dplyr)
```

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

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## 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 0.84 loaded
```

```
library(forcats)
```

Read in the data

```
train <- fread("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\AppliedDataMining\\HW6\\train.csv")
test <- fread("C:\\Users\\khickman\\Desktop\\Personal\\IUMSDS\\AppliedDataMining\\HW6\\test.csv")
summary(train)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##                                                      NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   LandSlope         Neighborhood        Condition1
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Condition2          BldgType          HouseStyle         OverallQual
##  Length:1460        Length:1460        Length:1460        Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 5.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.000
##                                                           Mean   : 6.099
##                                                           3rd Qu.: 7.000
##                                                           Max.   :10.000
##
##   OverallCond      YearBuilt     YearRemodAdd   RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Length:1460
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Class :character
##  Median :5.000   Median :1973   Median :1994   Mode  :character
##  Mean   :5.575   Mean   :1971   Mean   :1985
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004
##  Max.   :9.000   Max.   :2010   Max.   :2010
##
##    RoofMatl         Exterior1st        Exterior2nd
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   MasVnrType          MasVnrArea      ExterQual          ExterCond
##  Length:1460        Min.   :   0.0   Length:1460        Length:1460
##  Class :character   1st Qu.:   0.0   Class :character   Class :character
##  Mode  :character   Median :   0.0   Mode  :character   Mode  :character
##                     Mean   : 103.7
##                     3rd Qu.: 166.0
##                      Max.   :1600.0
##                      NA's   :8
##   Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##   HeatingQC          CentralAir         Electrical           1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##     2ndFlrSF     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
##  Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.047                      Mean   : 6.518
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :3.000                      Max.   :14.000
##
##    Fireplaces    FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.000   Length:1460        Length:1460        Min.   :1900
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1961
##  Median :1.000   Mode  :character   Mode  :character   Median :1980
##  Mean   :0.613                                         Mean   :1979
##  3rd Qu.:1.000                                         3rd Qu.:2002
##  Max.   :3.000                                         Max.   :2010
##                                                        NA's   :81
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460
##  Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.767   Mean   : 473.0
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :4.000   Max.   :1418.0
##
##   GarageCond         PavedDrive          WoodDeckSF      OpenPorchSF
##  Length:1460        Length:1460        Min.   :  0.00   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median : 25.00
##                                        Mean   : 94.24   Mean   : 46.66
##                                        3rd Qu.:168.00   3rd Qu.: 68.00
##                                        Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      3SsnPorch       ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     PoolQC             Fence           MiscFeature
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     MiscVal             MoSold          YrSold       SaleType
##  Min.   :    0.00   Min.   : 1.000   Min.   :2006   Length:1460
##  1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   Class :character
##  Median :    0.00   Median : 6.000   Median :2008   Mode  :character
##  Mean   :   43.49   Mean   : 6.322   Mean   :2008
##  3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :15500.00   Max.   :12.000   Max.   :2010
##
##  SaleCondition        SalePrice
##  Length:1460        Min.   : 34900
##  Class :character   1st Qu.:129975
##  Mode  :character   Median :163000
##                     Mean   :180921
##                     3rd Qu.:214000
##                     Max.   :755000
##
```

```r
str(train)
```

```
## Classes 'data.table' and 'data.frame':   1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ 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       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  NA NA NA NA ...
##  $ LotShape     : chr  "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr  "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : chr  "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : chr  "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : chr  "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr  "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr  "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ 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      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 1stFlrSF     : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ 2ndFlrSF     : 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  : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ 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   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ 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 ...
##  $ 3SsnPorch    : 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       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA 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     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

The train dataset has 1460 observations of 81 variables, with the 81st variable being the target (not included in test dataset); Sale Price. There are several apparently discrete variables that are categorized as continuous, as noted below. Roughly half of the 81 variables are continuous, so we have a dataset that will lend itself well to regression analysis. With the relatively high number of variables, we may want to reduce dimensionality with PCA as well. First, I need to understand and correctly name the variables.

Transform Data Types

```
#OverallQual
#GarageCars
#MoSold
#YrSold
#GarageYrBlt
#Fireplaces
```

```r
#BsmtFullBath
#BsmtHalfBath
#FullBath
#HalfBath
#BedroomAbvGr
#KitchenAbvGr

train$OverallQual <- as.factor(train$OverallQual)
train$GarageCars <- as.factor(train$GarageCars)
train$MoSold <- as.factor(train$MoSold)
train$YrSold <- as.factor(train$YrSold)
train$GarageYrBlt <- as.factor(train$GarageYrBlt)
train$Fireplaces <- as.factor(train$Fireplaces)
train$BsmtFullBath <- as.factor(train$BsmtFullBath)
train$BsmtHalfBath <- as.factor(train$BsmtHalfBath)
train$FullBath <- as.factor(train$FullBath)
train$HalfBath <- as.factor(train$HalfBath)
train$BedroomAbvGr <- as.factor(train$BedroomAbvGr)
train$KitchenAbvGr <- as.factor(train$KitchenAbvGr)
summary(train)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##                                                      NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   LandSlope         Neighborhood        Condition1
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Condition2          BldgType          HouseStyle         OverallQual
##  Length:1460        Length:1460        Length:1460        5      :397
##  Class :character   Class :character   Class :character   6      :374
##  Mode  :character   Mode  :character   Mode  :character   7      :319
##                                                           8      :168
##                                                           4      :116
##                                                           9      : 43
##                                                           (Other): 43
##   OverallCond      YearBuilt     YearRemodAdd   RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Length:1460
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Class :character
##  Median :5.000   Median :1973   Median :1994   Mode  :character
##  Mean   :5.575   Mean   :1971   Mean   :1985
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004
##  Max.   :9.000   Max.   :2010   Max.   :2010
##
##    RoofMatl         Exterior1st        Exterior2nd
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   MasVnrType          MasVnrArea      ExterQual          ExterCond
##  Length:1460        Min.   :   0.0   Length:1460        Length:1460
##  Class :character   1st Qu.:   0.0   Class :character   Class :character
##  Mode  :character   Median :   0.0   Mode  :character   Mode  :character
##                     Mean   : 103.7
##                     3rd Qu.: 166.0
##                     Max.   :1600.0
##                     NA's   :8
##   Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
```

```
##
##   HeatingQC          CentralAir         Electrical           1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##    2ndFlrSF     LowQualFinSF       GrLivArea    BsmtFullBath BsmtHalfBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   0:856        0:1378
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1:588        1:  80
##  Median :   0   Median :  0.000   Median :1464   2: 15        2:   2
##  Mean   : 347   Mean   :  5.845   Mean   :1515   3:  1
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777
##  Max.   :2065   Max.   :572.000   Max.   :5642
##
##  FullBath HalfBath  BedroomAbvGr KitchenAbvGr KitchenQual
##  0:  9    0:913    3      :804   0:   1       Length:1460
##  1:650    1:535    2      :358   1:1392       Class :character
##  2:768    2: 12    4      :213   2:  65       Mode  :character
##  3: 33             1      : 50   3:   2
##                    5      : 21
##                    6      :  7
##                    (Other):  7
##   TotRmsAbvGrd     Functional        Fireplaces FireplaceQu
##  Min.   : 2.000   Length:1460        0:690      Length:1460
##  1st Qu.: 5.000   Class :character   1:650      Class :character
##  Median : 6.000   Mode  :character   2:115      Mode  :character
##  Mean   : 6.518                      3:  5
##  3rd Qu.: 7.000
##  Max.   :14.000
##
##   GarageType         GarageYrBlt   GarageFinish       GarageCars
##  Length:1460        2005   :  65   Length:1460        0: 81
##  Class :character   2006   :  59   Class :character   1:369
##  Mode  :character   2004   :  53   Mode  :character   2:824
##                     2003   :  50                      3:181
##                     2007   :  49                      4:  5
##                     (Other):1103
##                     NA's   :  81
##    GarageArea      GarageQual         GarageCond         PavedDrive
##  Min.   :   0.0   Length:1460        Length:1460        Length:1460
##  1st Qu.: 334.5   Class :character   Class :character   Class :character
##  Median : 480.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 473.0
##  3rd Qu.: 576.0
##  Max.   :1418.0
##
##    WoodDeckSF      OpenPorchSF     EnclosedPorch      3SsnPorch
##  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 :  0.00   Median : 25.00   Median :  0.00   Median :  0.00
##  Mean   : 94.24   Mean   : 46.66   Mean   : 21.95   Mean   :  3.41
##  3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00   3rd Qu.:  0.00  
##  Max.   :857.00   Max.   :547.00   Max.   :552.00   Max.   :508.00  
##                                                                     
##   ScreenPorch        PoolArea          PoolQC             Fence          
##  Min.   :  0.00   Min.   :  0.000   Length:1460        Length:1460       
##  1st Qu.:  0.00   1st Qu.:  0.000   Class :character   Class :character  
##  Median :  0.00   Median :  0.000   Mode  :character   Mode  :character  
##  Mean   : 15.06   Mean   :  2.759                                        
##  3rd Qu.:  0.00   3rd Qu.:  0.000                                        
##  Max.   :480.00   Max.   :738.000                                        
##                                                                          
##  MiscFeature           MiscVal             MoSold     YrSold    
##  Length:1460        Min.   :    0.00   6      :253   2006:314  
##  Class :character   1st Qu.:    0.00   7      :234   2007:329  
##  Mode  :character   Median :    0.00   5      :204   2008:304  
##                     Mean   :   43.49   4      :141   2009:338  
##                     3rd Qu.:    0.00   8      :122   2010:175  
##                     Max.   :15500.00   3      :106             
##                                        (Other):400             
##    SaleType         SaleCondition        SalePrice     
##  Length:1460        Length:1460        Min.   : 34900  
##  Class :character   Class :character   1st Qu.:129975  
##  Mode  :character   Mode  :character   Median :163000  
##                                        Mean   :180921  
##                                        3rd Qu.:214000  
##                                        Max.   :755000  
## 
```

I'm 100% sure there is an easier way to do this - possibly with an sapply or c() function.

Now that the datatypes are set, we can move on with our analysis. I found an interesting method to separate continuous and discrete variables on Kaggle:

Discretize Variables

```
cat_var <- names(train)[which(sapply(train, is.character))]
cat_var
```

```
##  [1] "MSZoning"      "Street"        "Alley"         "LotShape"     
##  [5] "LandContour"   "Utilities"     "LotConfig"     "LandSlope"    
##  [9] "Neighborhood"  "Condition1"    "Condition2"    "BldgType"     
## [13] "HouseStyle"    "RoofStyle"     "RoofMatl"      "Exterior1st"  
## [17] "Exterior2nd"   "MasVnrType"    "ExterQual"     "ExterCond"    
## [21] "Foundation"    "BsmtQual"      "BsmtCond"      "BsmtExposure" 
## [25] "BsmtFinType1"  "BsmtFinType2"  "Heating"       "HeatingQC"    
## [29] "CentralAir"    "Electrical"    "KitchenQual"   "Functional"   
## [33] "FireplaceQu"   "GarageType"    "GarageFinish"  "GarageQual"   
## [37] "GarageCond"    "PavedDrive"    "PoolQC"        "Fence"        
## [41] "MiscFeature"   "SaleType"      "SaleCondition"
```

```
numeric_var <- names(train)[which(sapply(train, is.numeric))]
numeric_var
```

```
##  [1] "Id"            "MSSubClass"    "LotFrontage"   "LotArea"      
##  [5] "OverallCond"   "YearBuilt"     "YearRemodAdd"  "MasVnrArea"
##  [9] "BsmtFinSF1"    "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"
## [13] "1stFlrSF"      "2ndFlrSF"      "LowQualFinSF"  "GrLivArea"
## [17] "TotRmsAbvGrd"  "GarageArea"    "WoodDeckSF"    "OpenPorchSF"
## [21] "EnclosedPorch" "3SsnPorch"     "ScreenPorch"   "PoolArea"
## [25] "MiscVal"       "SalePrice"
```

```
str(numeric_var)
```

```
##  chr [1:26] "Id" "MSSubClass" "LotFrontage" "LotArea" ...
```

Missing features in the Data

At first impression, there appear to be several variables with many missing values. However, several of them are not necessarily missing, only listed as "NA" when the true value should be "none". For instance, "Alley"; "NA"" might mean that we don't have an alley, not that the values are missing. I'll refactor this and other variables so that "NA" becomes None.

Found a good Analysis/transformation of variables from a Kaggle user here: https://www.kaggle.com/sidraina89/regularized-regression-housing-pricing

```
Missing_indices <- sapply(train,function(x)sum(is.na(x)))
Missing_Summary <- data.frame(index = names(train),Missing_Values=Missing_indices)
Missing_Summary[Missing_Summary$Missing_Values > 0,]
```

```
##                     index Missing_Values
## LotFrontage   LotFrontage            259
## Alley               Alley           1369
## MasVnrType     MasVnrType              8
## MasVnrArea     MasVnrArea              8
## BsmtQual         BsmtQual             37
## BsmtCond         BsmtCond             37
## BsmtExposure BsmtExposure             38
## BsmtFinType1 BsmtFinType1             37
## BsmtFinType2 BsmtFinType2             38
## Electrical     Electrical              1
## FireplaceQu   FireplaceQu            690
## GarageType     GarageType             81
## GarageYrBlt   GarageYrBlt             81
## GarageFinish GarageFinish             81
## GarageQual     GarageQual             81
## GarageCond     GarageCond             81
## PoolQC             PoolQC           1453
## Fence               Fence           1179
## MiscFeature   MiscFeature           1406
```

Immediately, several variables stand out that strongly suggest "NA" does not always mean "missing". Additionally, the dataset description points to this conclusion. Alley, PoolQC, Fence, and MiscFeature all have a high number of NAs, but it's also very probable that a high number of homes in our dataset don't have those features at all. Let's refactor those variables from "NA" to none:

```
train$Alley[which(is.na(train$Alley))] <- "None"
table(train$Alley)
```

```
##
## Grvl None Pave
##   50 1369   41
```

It worked! I understand I'll have to do this same method on the test dataset as well. There were a few Kagglers who combined the two datasets to do the transform, but I don't want to do that for quality reasons - I'll do it the longer way.

```
train$Alley[which(is.na(train$Alley))] <- "None"
train$MoSold[which(is.na(train$MoSold))] <- "None"
```

```
## Warning in `[<-.factor`(`*tmp*`, which(is.na(train$MoSold)), value =
## structure(c(2L, : invalid factor level, NA generated
```

```
train$Fireplaces[which(is.na(train$Fireplaces))] <- "None"
```

```
## Warning in `[<-.factor`(`*tmp*`, which(is.na(train$Fireplaces)), value =
## structure(c(1L, : invalid factor level, NA generated
```

```
#Transform Garage Characteristics on homes that have no garages:
train$GarageCond[which(is.na(train$GarageCond))] <- "None"
train$GarageYrBlt[which(is.na(train$GarageYrBuilt))] <- "None"
```

```
## Warning in `[<-.factor`(`*tmp*`, which(is.na(train$GarageYrBuilt)), value =
## structure(c(90L, : invalid factor level, NA generated
```

```
## Warning in is.na(train$GarageYrBuilt): is.na() applied to non-(list or
## vector) of type 'NULL'
```

```
train$GarageType[which(is.na(train$GarageType))] <- "None"
train$GarageCars[which(is.na(train$GarageCars))] <- "None"
```

```
## Warning in `[<-.factor`(`*tmp*`, which(is.na(train$GarageCars)), value =
## structure(c(3L, : invalid factor level, NA generated
```

```
train$GarageFinish[which(is.na(train$GarageFinish))] <- "None"
train$GarageQual[which(is.na(train$GarageQual))] <- "None"


## Check to make sure it's still working as intended:
table(train$GarageQual)
```

```
##
##   Ex   Fa   Gd None   Po   TA
##    3   48   14   81    3 1311
```

Surprisingly, there are 9 homes listed with 0 full baths. Based on the other characteristics of the rows, these look like they might be missing rather than 0. With a large number of variables, let's do a PCA to reduce the dimensionality. Using prcomp() and princomp().

Further transformations and PCA

```
train.numeric <- train[,.SD, .SDcols =numeric_var]
train.numeric <- train.numeric

summary(train.numeric)
```

```
##        Id             MSSubClass     LotFrontage        LotArea
##  Min.   :    1.0   Min.   : 20.0   Min.   : 21.00   Min.   :  1300
##  1st Qu.: 365.8   1st Qu.: 20.0   1st Qu.: 59.00   1st Qu.:  7554
##  Median : 730.5   Median : 50.0   Median : 69.00   Median :  9478
##  Mean   : 730.5   Mean   : 56.9   Mean   : 70.05   Mean   : 10517
##  3rd Qu.:1095.2   3rd Qu.: 70.0   3rd Qu.: 80.00   3rd Qu.: 11602
##  Max.   :1460.0   Max.   :190.0   Max.   :313.00   Max.   :215245
##                                   NA's   :259
##   OverallCond      YearBuilt     YearRemodAdd    MasVnrArea
##  Min.   :1.000   Min.   :1872   Min.   :1950   Min.   :   0.0
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   1st Qu.:   0.0
##  Median :5.000   Median :1973   Median :1994   Median :   0.0
##  Mean   :5.575   Mean   :1971   Mean   :1985   Mean   : 103.7
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   3rd Qu.: 166.0
##  Max.   :9.000   Max.   :2010   Max.   :2010   Max.   :1600.0
##                                                NA's   :8
##    BsmtFinSF1       BsmtFinSF2        BsmtUnfSF       TotalBsmtSF
##  Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8
##  Median : 383.5   Median :   0.00   Median : 477.5   Median : 991.5
##  Mean   : 443.6   Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.: 712.2   3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :5644.0   Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##     1stFlrSF       2ndFlrSF     LowQualFinSF       GrLivArea
##  Min.   : 334   Min.   :   0   Min.   :  0.000   Min.   : 334
##  1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130
##  Median :1087   Median :   0   Median :  0.000   Median :1464
##  Mean   :1163   Mean   : 347   Mean   :  5.845   Mean   :1515
##  3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777
##  Max.   :4692   Max.   :2065   Max.   :572.000   Max.   :5642
##
##   TotRmsAbvGrd      GarageArea       WoodDeckSF      OpenPorchSF
##  Min.   : 2.000   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 5.000   1st Qu.: 334.5   1st Qu.:  0.00   1st Qu.:  0.00
##  Median : 6.000   Median : 480.0   Median :  0.00   Median : 25.00
##  Mean   : 6.518   Mean   : 473.0   Mean   : 94.24   Mean   : 46.66
##  3rd Qu.: 7.000   3rd Qu.: 576.0   3rd Qu.:168.00   3rd Qu.: 68.00
##  Max.   :14.000   Max.   :1418.0   Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      3SsnPorch       ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     MiscVal           SalePrice
##  Min.   :    0.00   Min.   : 34900
##  1st Qu.:    0.00   1st Qu.:129975
##  Median :    0.00   Median :163000
##  Mean   :   43.49   Mean   :180921
##  3rd Qu.:    0.00   3rd Qu.:214000
##  Max.   :15500.00   Max.   :755000
##
```

There aren't any variables with 20% or more NAs, so we can't use manyNAs. Let's use most frequent values for the missing numbers.

The variables `LotFrontage`, `MasVnrArea`, `BsmtFinSF1`, `BsmtFinSF2`, `BsmtUnfSF`, `TotalBsmtSF`, and `GarageArea` all exhibit NAs.

Before we do this, we need to know whether to use the mean or median for LotFrontage. Here, we appear to have a normally distributed variable, so it probably won't matter much, and we'll use mean.

```
train.numeric[is.na(train.numeric$LotFrontage), "LotFrontage"] <- mean(train.numeric$LotFrontage, na.rm
```

```
## Warning in `[<-.data.table`(`*tmp*`, is.na(train.numeric$LotFrontage),
## "LotFrontage", : Coerced 'double' RHS to 'integer' to match the column's
## type; may have truncated precision. Either change the target column to
## 'double' first (by creating a new 'double' vector length 1460 (nrows of
## entire table) and assign that; i.e. 'replace' column), or coerce RHS to
## 'integer' (e.g. 1L, NA_[real|integer]_, as.*, etc) to make your intent
## clear and for speed. Or, set the column type correctly up front when you
## create the table and stick to it, please.
```

```
summary(train.numeric)
```

```
##        Id           MSSubClass     LotFrontage        LotArea
##  Min.   :   1.0   Min.   : 20.0   Min.   : 21.00   Min.   :  1300
##  1st Qu.: 365.8   1st Qu.: 20.0   1st Qu.: 60.00   1st Qu.:  7554
##  Median : 730.5   Median : 50.0   Median : 70.00   Median :  9478
##  Mean   : 730.5   Mean   : 56.9   Mean   : 70.04   Mean   : 10517
##  3rd Qu.:1095.2   3rd Qu.: 70.0   3rd Qu.: 79.00   3rd Qu.: 11602
##  Max.   :1460.0   Max.   :190.0   Max.   :313.00   Max.   :215245
##
##   OverallCond      YearBuilt     YearRemodAdd    MasVnrArea
##  Min.   :1.000   Min.   :1872   Min.   :1950   Min.   :   0.0
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   1st Qu.:   0.0
##  Median :5.000   Median :1973   Median :1994   Median :   0.0
##  Mean   :5.575   Mean   :1971   Mean   :1985   Mean   : 103.7
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   3rd Qu.: 166.0
##  Max.   :9.000   Max.   :2010   Max.   :2010   Max.   :1600.0
##                                                NA's   :8
##    BsmtFinSF1       BsmtFinSF2        BsmtUnfSF       TotalBsmtSF
##  Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8
##  Median : 383.5   Median :   0.00   Median : 477.5   Median : 991.5
##  Mean   : 443.6   Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.: 712.2   3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :5644.0   Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##     1stFlrSF       2ndFlrSF     LowQualFinSF       GrLivArea
##  Min.   : 334   Min.   :   0   Min.   :  0.000   Min.   : 334
##  1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130
##  Median :1087   Median :   0   Median :  0.000   Median :1464
##  Mean   :1163   Mean   : 347   Mean   :  5.845   Mean   :1515
##  3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777
##  Max.   :4692   Max.   :2065   Max.   :572.000   Max.   :5642
##
##   TotRmsAbvGrd      GarageArea       WoodDeckSF      OpenPorchSF
##  Min.   : 2.000   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 5.000   1st Qu.: 334.5   1st Qu.:  0.00   1st Qu.:  0.00
##  Median : 6.000   Median : 480.0   Median :  0.00   Median : 25.00
##  Mean   : 6.518   Mean   : 473.0   Mean   : 94.24   Mean   : 46.66
##  3rd Qu.: 7.000   3rd Qu.: 576.0   3rd Qu.:168.00   3rd Qu.: 68.00
##  Max.   :14.000   Max.   :1418.0   Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      3SsnPorch       ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     MiscVal            SalePrice
##  Min.   :    0.00   Min.   : 34900
##  1st Qu.:    0.00   1st Qu.:129975
##  Median :    0.00   Median :163000
##  Mean   :   43.49   Mean   :180921
##  3rd Qu.:    0.00   3rd Qu.:214000
##  Max.   :15500.00   Max.   :755000
##
```

```
## That worked, so we'll continue with the remaining variables.

train.numeric[is.na(train.numeric$MasVnrArea), "MasVnrArea"] <- mean(train.numeric$MasVnrArea, na.rm = T
```

```
## Warning in `[<-.data.table`(`*tmp*`, is.na(train.numeric$MasVnrArea),
## "MasVnrArea", : Coerced 'double' RHS to 'integer' to match the column's
## type; may have truncated precision. Either change the target column to
## 'double' first (by creating a new 'double' vector length 1460 (nrows of
## entire table) and assign that; i.e. 'replace' column), or coerce RHS to
## 'integer' (e.g. 1L, NA_[real|integer]_, as.*, etc) to make your intent
## clear and for speed. Or, set the column type correctly up front when you
## create the table and stick to it, please.
```

```
train.numeric[is.na(train.numeric$BsmtFinSF2), "BsmtFinSF2"] <- mean(train.numeric$BsmtFinSF2, na.rm = T

train.numeric[is.na(train.numeric$BsmtUnfSF), "BsmtUnfSF"] <- mean(train.numeric$BsmtUnfSF, na.rm = TRU

train.numeric[is.na(train.numeric$TotalBsmtSF), "TotalBsmtSF"] <- mean(train.numeric$TotalBsmtSF, na.rm

train.numeric[is.na(train.numeric$BsmtUnfSF), "BsmtUnfSF"] <- mean(train.numeric$BsmtUnfSF, na.rm = TRU

train.numeric[is.na(train.numeric$GarageArea), "GarageArea"] <- mean(train.numeric$GarageArea, na.rm = T
```

Let's check whether our transform worked:

```
summary(train.numeric)
```

```
##        Id            MSSubClass     LotFrontage        LotArea
##  Min.   :   1.0   Min.   : 20.0   Min.   : 21.00   Min.   :  1300
##  1st Qu.: 365.8   1st Qu.: 20.0   1st Qu.: 60.00   1st Qu.:  7554
##  Median : 730.5   Median : 50.0   Median : 70.00   Median :  9478
##  Mean   : 730.5   Mean   : 56.9   Mean   : 70.04   Mean   : 10517
##  3rd Qu.:1095.2   3rd Qu.: 70.0   3rd Qu.: 79.00   3rd Qu.: 11602
##  Max.   :1460.0   Max.   :190.0   Max.   :313.00   Max.   :215245  
##   OverallCond      YearBuilt     YearRemodAdd    MasVnrArea    
##  Min.   :1.000   Min.   :1872   Min.   :1950   Min.   :   0.0  
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   1st Qu.:   0.0  
##  Median :5.000   Median :1973   Median :1994   Median :   0.0  
##  Mean   :5.575   Mean   :1971   Mean   :1985   Mean   : 103.7  
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   3rd Qu.: 164.2  
##  Max.   :9.000   Max.   :2010   Max.   :2010   Max.   :1600.0  
##    BsmtFinSF1       BsmtFinSF2        BsmtUnfSF       TotalBsmtSF    
##  Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Min.   :   0.0  
##  1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8  
##  Median : 383.5   Median :   0.00   Median : 477.5   Median : 991.5  
##  Mean   : 443.6   Mean   :  46.55   Mean   : 567.2   Mean   :1057.4  
##  3rd Qu.: 712.2   3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2  
##  Max.   :5644.0   Max.   :1474.00   Max.   :2336.0   Max.   :6110.0  
##     1stFlrSF       2ndFlrSF     LowQualFinSF       GrLivArea   
##  Min.   : 334   Min.   :   0   Min.   :  0.000   Min.   : 334  
##  1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130  
##  Median :1087   Median :   0   Median :  0.000   Median :1464  
##  Mean   :1163   Mean   : 347   Mean   :  5.845   Mean   :1515  
##  3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777  
##  Max.   :4692   Max.   :2065   Max.   :572.000   Max.   :5642  
##   TotRmsAbvGrd      GarageArea       WoodDeckSF      OpenPorchSF    
##  Min.   : 2.000   Min.   :   0.0   Min.   :  0.00   Min.   :  0.00  
##  1st Qu.: 5.000   1st Qu.: 334.5   1st Qu.:  0.00   1st Qu.:  0.00  
##  Median : 6.000   Median : 480.0   Median :  0.00   Median : 25.00  
##  Mean   : 6.518   Mean   : 473.0   Mean   : 94.24   Mean   : 46.66  
##  3rd Qu.: 7.000   3rd Qu.: 576.0   3rd Qu.:168.00   3rd Qu.: 68.00  
##  Max.   :14.000   Max.   :1418.0   Max.   :857.00   Max.   :547.00  
##  EnclosedPorch      3SsnPorch       ScreenPorch        PoolArea     
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000  
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000  
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000  
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759  
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000  
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000  
##     MiscVal           SalePrice     
##  Min.   :    0.00   Min.   : 34900  
##  1st Qu.:    0.00   1st Qu.:129975  
##  Median :    0.00   Median :163000  
##  Mean   :   43.49   Mean   :180921  
##  3rd Qu.:    0.00   3rd Qu.:214000  
##  Max.   :15500.00   Max.   :755000
```

Now that we don't have any NAs and all numeric variables, let's conduct the PCA.

```r
pca.train <- prcomp(train.numeric)
pca.train2 <- princomp(train.numeric)
summary(pca.train2)
```

```
## Importance of components:
##                              Comp.1       Comp.2       Comp.3       Comp.4
## Standard deviation     7.946179e+04 9.619525e+03 5.856634e+02 5.386349e+02
## Proportion of Variance 9.853398e-01 1.444031e-02 5.352609e-05 4.527498e-05
## Cumulative Proportion  9.853398e-01 9.997801e-01 9.998336e-01 9.998789e-01
##                               Comp.5       Comp.6       Comp.7       Comp.8
## Standard deviation     4.943939e+02 4.396728e+02 4.178152e+02 2.238917e+02
## Proportion of Variance 3.814306e-05 3.016675e-05 2.724192e-05 7.822497e-06
## Cumulative Proportion  9.999170e-01 9.999472e-01 9.999744e-01 9.999823e-01
##                               Comp.9      Comp.10      Comp.11      Comp.12
## Standard deviation     1.764332e+02 1.661207e+02 1.509872e+02 1.175744e+02
## Proportion of Variance 4.857693e-06 4.306423e-06 3.557538e-06 2.157222e-06
## Cumulative Proportion  9.999871e-01 9.999914e-01 9.999950e-01 9.999971e-01
##                              Comp.13      Comp.14      Comp.15      Comp.16
## Standard deviation     6.390520e+01 5.827439e+01 5.533403e+01 5.315321e+01
## Proportion of Variance 6.372968e-07 5.299377e-07 4.778085e-07 4.408880e-07
## Cumulative Proportion  9.999978e-01 9.999983e-01 9.999988e-01 9.999992e-01
##                              Comp.17      Comp.18      Comp.19      Comp.20
## Standard deviation     3.940492e+01 3.884311e+01 2.905478e+01 2.229807e+01
## Proportion of Variance 2.423094e-07 2.354493e-07 1.317360e-07 7.758965e-08
## Cumulative Proportion  9.999995e-01 9.999997e-01 9.999998e-01 9.999999e-01
##                              Comp.21      Comp.22      Comp.23      Comp.24
## Standard deviation     1.741653e+01 1.422433e+01 9.071284e-01 8.677139e-01
## Proportion of Variance 4.733610e-08 3.157427e-08 1.284124e-10 1.174958e-10
## Cumulative Proportion  1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
##                              Comp.25      Comp.26
## Standard deviation     3.217073e-05 2.990967e-05
## Proportion of Variance 1.615069e-19 1.396023e-19
## Cumulative Proportion  1.000000e+00 1.000000e+00
```

```
loadings(pca.train2)
```

```
## 
## Loadings:
##               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Id                                               -0.403  0.914       
## MSSubClass                                                           
## LotFrontage                                                          
## LotArea              -0.999                                          
## OverallCond                                                          
## YearBuilt                                                            
## YearRemodAdd                                                         
## MasVnrArea                                                      0.107
## BsmtFinSF1                   0.523  0.393        -0.344 -0.156  0.383
## BsmtFinSF2                                                     -0.368
## BsmtUnfSF                   -0.193 -0.751                       0.342
## TotalBsmtSF                  0.357 -0.345        -0.322 -0.153  0.358
## 1stFlrSF                     0.277 -0.277        -0.336 -0.126 -0.545
## 2ndFlrSF                    -0.605  0.252        -0.293 -0.141  0.311
## LowQualFinSF                                                         
## GrLivArea                   -0.338               -0.639 -0.277 -0.254
## TotRmsAbvGrd                                                         
## GarageArea                                                           
## WoodDeckSF                                                           
## OpenPorchSF                                                          
## EnclosedPorch                                                        
## 3SsnPorch                                                            
## ScreenPorch                                                          
## PoolArea                                                             
## MiscVal                             0.138 -0.989                     
## SalePrice      0.999
##               Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
## Id
## MSSubClass
## LotFrontage
## LotArea
## OverallCond
## YearBuilt
## YearRemodAdd
## MasVnrArea     0.264  0.401   0.866
## BsmtFinSF1     0.119 -0.117
## BsmtFinSF2    -0.701  0.327   0.108
## BsmtUnfSF      0.107
## TotalBsmtSF   -0.474  0.123
## 1stFlrSF       0.245 -0.107                                  -0.263
## 2ndFlrSF      -0.156                                          -0.246
## LowQualFinSF                                          0.107   0.727
## GrLivArea                                                     0.218
## TotRmsAbvGrd
## GarageArea     0.305  0.817  -0.478
## WoodDeckSF                           -0.991
## OpenPorchSF                                  -0.724   0.683
## EnclosedPorch                                  0.639   0.660
## 3SsnPorch
## ScreenPorch                                  -0.227  -0.254   0.515
## PoolArea
## MiscVal
## SalePrice
##               Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21 Comp.22
## Id
## MSSubClass            -0.963                  -0.112   0.215
## LotFrontage             0.207                   0.102   0.963  -0.115
## LotArea
## OverallCond
## YearBuilt                                      0.840           0.496
## YearRemodAdd                                   0.484  -0.140  -0.858
## MasVnrArea
## BsmtFinSF1
## BsmtFinSF2
## BsmtUnfSF
## TotalBsmtSF
## 1stFlrSF      -0.152
## 2ndFlrSF      -0.149
## LowQualFinSF   0.449
## GrLivArea      0.149
## TotRmsAbvGrd
## GarageArea
## WoodDeckSF
## OpenPorchSF
## EnclosedPorch -0.327                           0.154
## 3SsnPorch                             -0.997
## ScreenPorch   -0.774
## PoolArea                      -0.991
## MiscVal
## SalePrice
##               Comp.23 Comp.24 Comp.25 Comp.26
## Id
## MSSubClass
## LotFrontage
## LotArea
## OverallCond   -0.956  -0.290
## YearBuilt
## YearRemodAdd
## MasVnrArea
## BsmtFinSF1                     0.214   0.452
## BsmtFinSF2                     0.214   0.452
## BsmtUnfSF                      0.214   0.452
## TotalBsmtSF                   -0.214  -0.452
## 1stFlrSF                       0.452  -0.214
## 2ndFlrSF                       0.452  -0.214
## LowQualFinSF                   0.452  -0.214
## GrLivArea                     -0.452   0.214
## TotRmsAbvGrd   0.291  -0.957
## GarageArea
## WoodDeckSF
## OpenPorchSF
## EnclosedPorch
## 3SsnPorch
## ScreenPorch
## PoolArea
## MiscVal
## SalePrice
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.038  0.038  0.038  0.038  0.038  0.038  0.038  0.038
## Cumulative Var  0.038  0.077  0.115  0.154  0.192  0.231  0.269  0.308
##                Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
## SS loadings     1.000   1.000   1.000   1.000   1.000   1.000   1.000
## Proportion Var  0.038   0.038   0.038   0.038   0.038   0.038   0.038
## Cumulative Var  0.346   0.385   0.423   0.462   0.500   0.538   0.577
##                Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21 Comp.22
## SS loadings      1.000   1.000   1.000   1.000   1.000   1.000   1.000
## Proportion Var   0.038   0.038   0.038   0.038   0.038   0.038   0.038
## Cumulative Var   0.615   0.654   0.692   0.731   0.769   0.808   0.846
##                Comp.23 Comp.24 Comp.25 Comp.26
## SS loadings      1.000   1.000   1.000   1.000
## Proportion Var   0.038   0.038   0.038   0.038
## Cumulative Var   0.885   0.923   0.962   1.000
```

There's not much to be gained from using PCA apparently. Each variable explains an equal 4% of the variance.

We'll likely have to log transform the sale price variable to fit it to a linear regression model. The variable is right-skewed and non-normal.

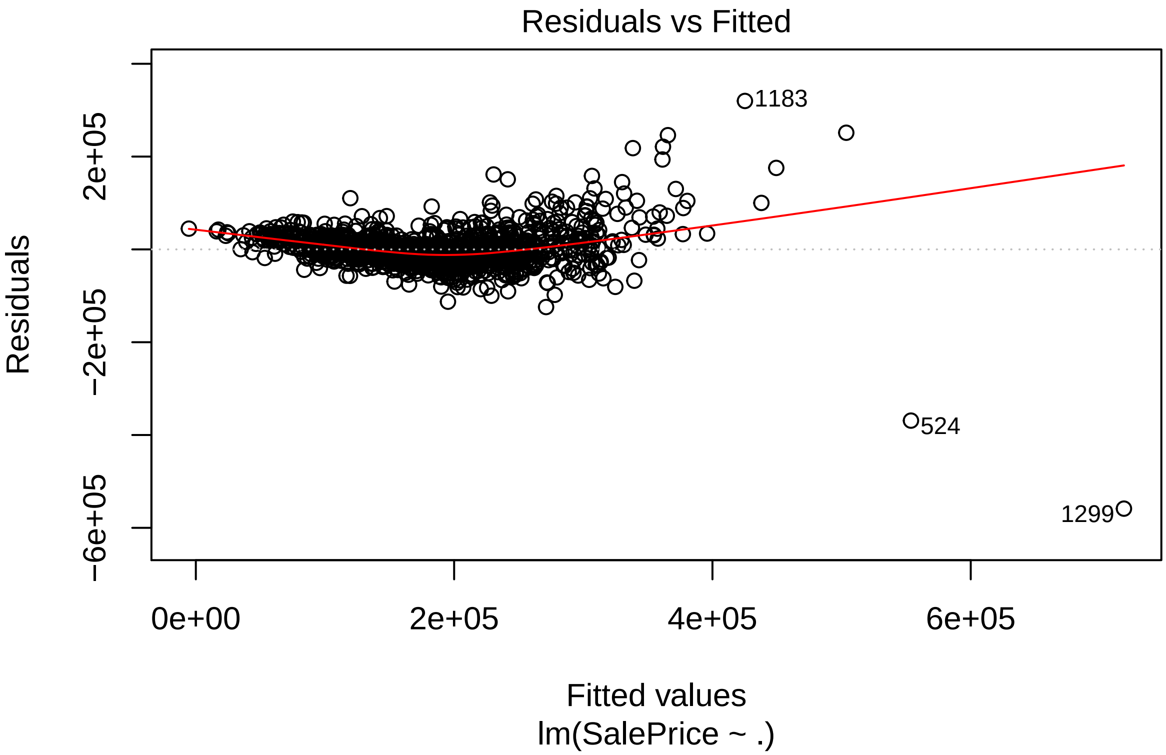
Initial linear model on continuous variables:

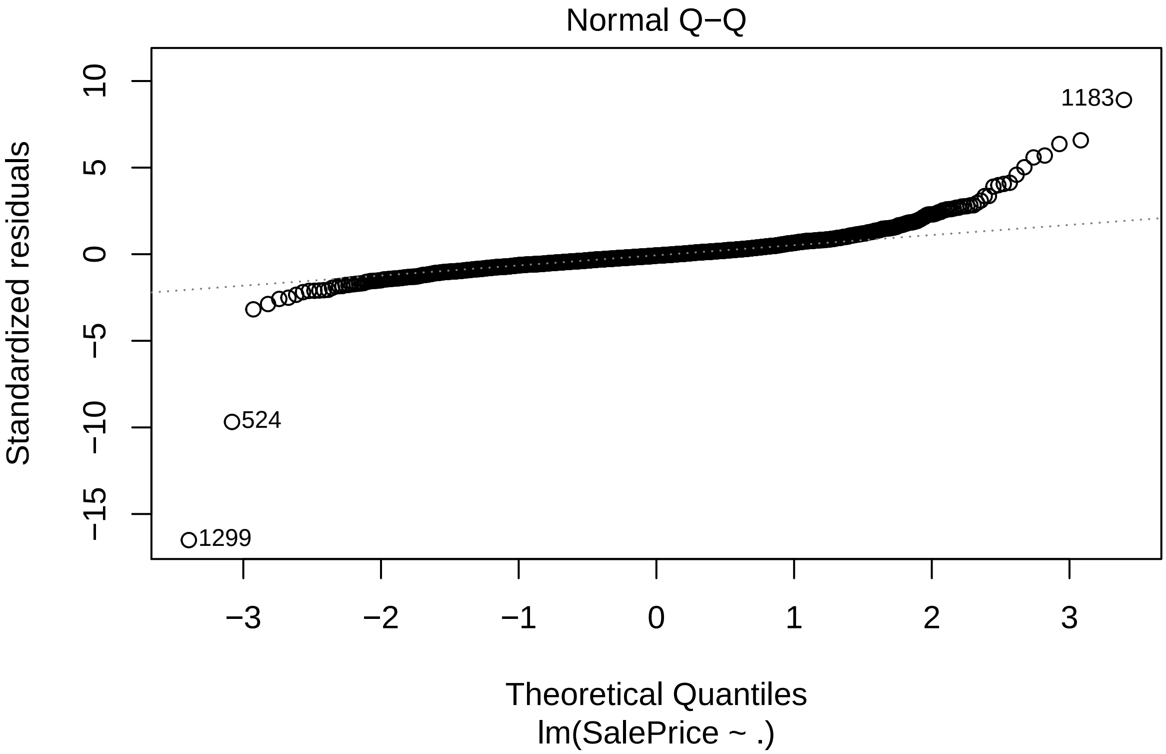
```
lm.sales <- lm(SalePrice ~ ., data = train.numeric)
summary(lm.sales)
```

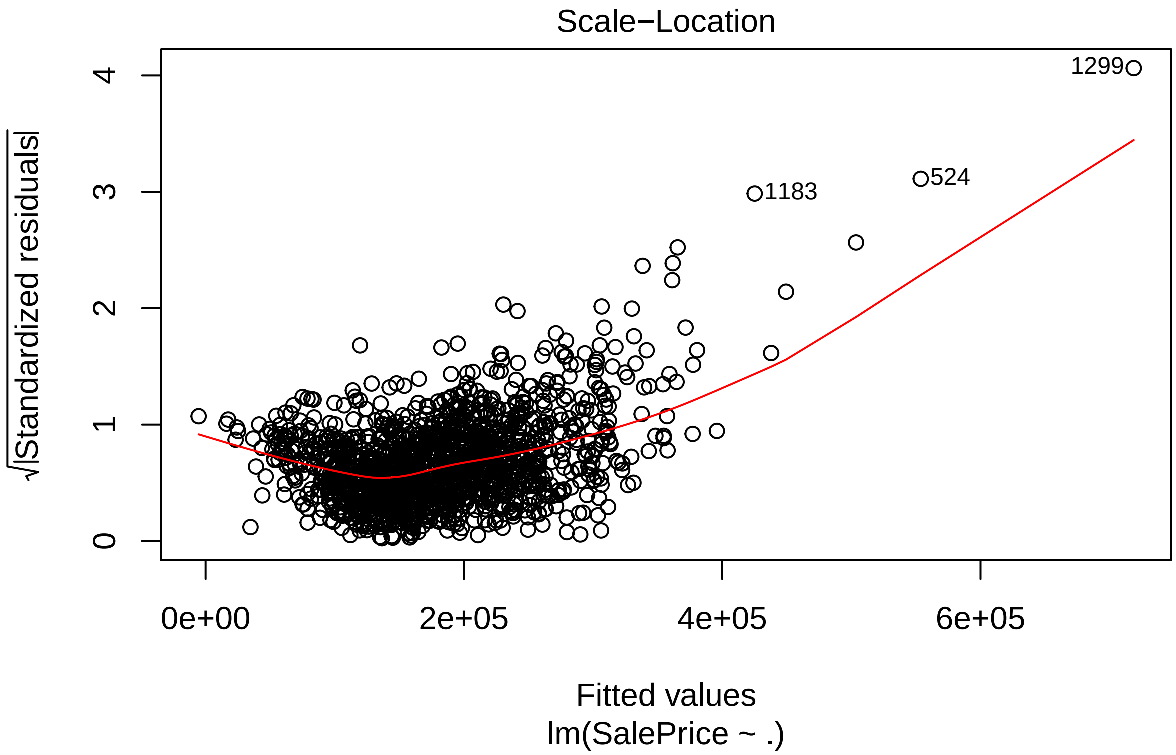
```
##
## Call:
## lm(formula = SalePrice ~ ., data = train.numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -558620  -17856   -3067   12973  319848
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.166e+06  1.199e+05 -18.070  < 2e-16 ***
## Id            -2.263e+00  2.462e+00  -0.919 0.358207
## MSSubClass    -1.499e+02  2.860e+01  -5.243 1.82e-07 ***
## LotFrontage   -1.199e+02  5.797e+01  -2.069 0.038699 *
## LotArea        3.927e-01  1.136e-01   3.455 0.000565 ***
## OverallCond    6.388e+03  1.124e+03   5.684 1.59e-08 ***
## YearBuilt      6.566e+02  5.918e+01  11.096  < 2e-16 ***
## YearRemodAdd   4.350e+02  7.115e+01   6.113 1.25e-09 ***
## MasVnrArea     3.974e+01  6.667e+00   5.960 3.17e-09 ***
## BsmtFinSF1     3.608e+01  4.684e+00   7.703 2.47e-14 ***
## BsmtFinSF2     1.864e+01  7.686e+00   2.424 0.015457 *
## BsmtUnfSF      2.286e+01  4.573e+00   5.000 6.45e-07 ***
## TotalBsmtSF           NA         NA      NA       NA
## `1stFlrSF`     6.665e+01  6.057e+00  11.003  < 2e-16 ***
## `2ndFlrSF`     6.658e+01  4.427e+00  15.039  < 2e-16 ***
## LowQualFinSF   5.132e+01  2.208e+01   2.325 0.020225 *
## GrLivArea             NA         NA      NA       NA
## TotRmsAbvGrd   7.396e+02  1.180e+03   0.627 0.530825
## GarageArea     4.755e+01  6.436e+00   7.389 2.51e-13 ***
## WoodDeckSF     3.133e+01  8.950e+00   3.501 0.000478 ***
## OpenPorchSF    1.593e+01  1.690e+01   0.943 0.345852
## EnclosedPorch  5.235e+01  1.886e+01   2.776 0.005572 **
## `3SsnPorch`    2.415e+01  3.542e+01   0.682 0.495440
## ScreenPorch    8.041e+01  1.911e+01   4.208 2.73e-05 ***
## PoolArea      -5.070e+01  2.676e+01  -1.895 0.058311 .
## MiscVal       -1.958e+00  2.085e+00  -0.939 0.347920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39260 on 1436 degrees of freedom
## Multiple R-squared:  0.7596, Adjusted R-squared:  0.7557
## F-statistic: 197.3 on 23 and 1436 DF,  p-value: < 2.2e-16
```

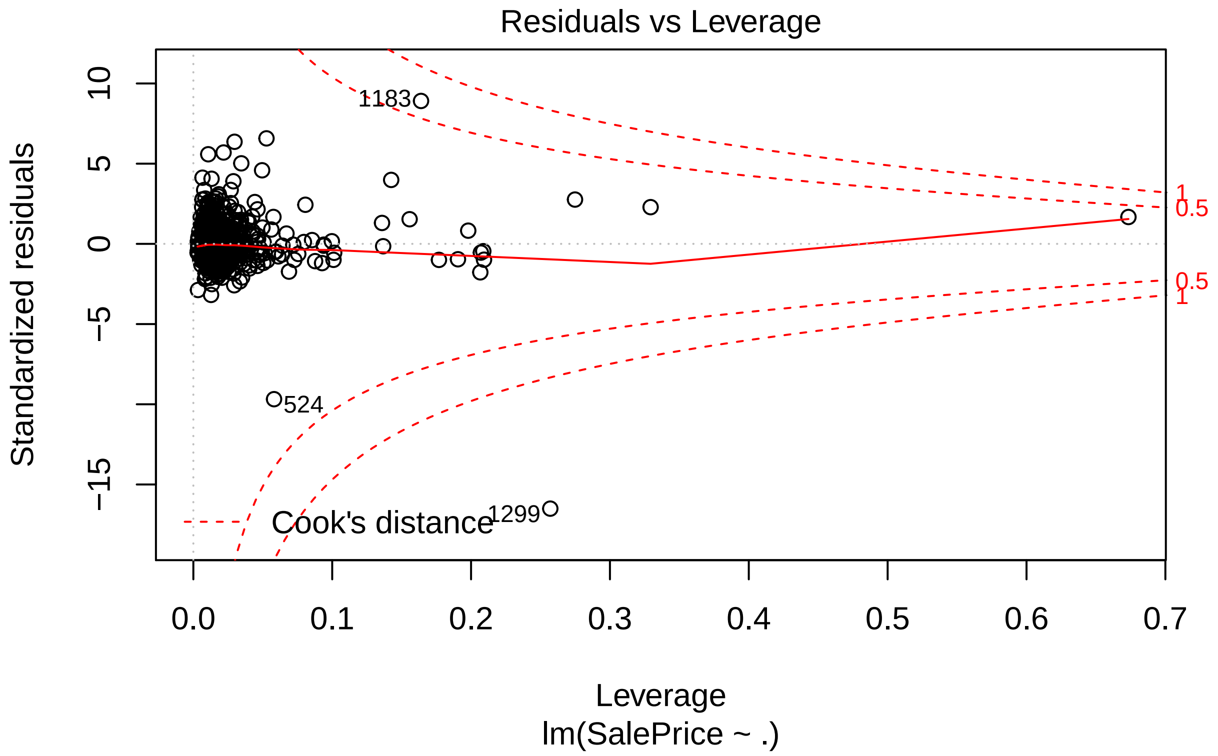
Analysis of the Model:

```
plot(lm.sales)
```


Residuals vs Fitted
Residuals
2e+05
−2e+05
−6e+05
0e+00
2e+05
4e+05
6e+05
Fitted values
lm(SalePrice ~ .)
1183
524
1299


Normal Q−Q
Standardized residuals
10
5
0
−5
−10
−15
1183
524
1299
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(SalePrice ~ .)


Scale−Location
√|Standardized residuals|
0
1
2
3
4
0e+00
2e+05
4e+05
6e+05
Fitted values
lm(SalePrice ~ .)
1299
1183
524



It appears that the QQnorm follows a t-distribution. The curve appears normal, except at the tails. It looks like we have some outliers at both ends. Thus a linear model may not be the best predictor. Additionally, the variable LotFrontage had the highest significance code. This is likely due to frontage space correlating with the size of the homes, which in turn, correlate with sales price.

Given that we have so many variables that are skewed, I'd like to try a decision tree against the performance of our linear model. Additionally, given the presence of so many categorical variables, we might see better performance.

```
library(rpart)
rt.sales <- rpart(SalePrice ~ ., data=train)
summary(rt.sales)
```

```
## Call:
## rpart(formula = SalePrice ~ ., data = train)
##   n= 1460
##
##            CP nsplit rel error    xerror       xstd
## 1  0.45437625      0 1.0000000 1.0029646 0.07641545
## 2  0.12050238      1 0.5456237 0.5475873 0.04133926
## 3  0.06288535      2 0.4251214 0.4314978 0.04020626
## 4  0.03404531      3 0.3622360 0.3812884 0.02967265
## 5  0.03312902      4 0.3281907 0.3735602 0.02997738
## 6  0.02066528      5 0.2950617 0.3452027 0.02896458
## 7  0.01856908      6 0.2743964 0.3285062 0.02860839
## 8  0.01392462      7 0.2558273 0.3185870 0.03008518
## 9  0.01094828      8 0.2419027 0.3041927 0.03031834
## 10 0.01000000      9 0.2309544 0.2981945 0.02955018
##
## Variable importance
##  OverallQual Neighborhood   GarageCars  KitchenQual   GarageArea
##           28           14            8            7            7
##  TotalBsmtSF    ExterQual    GrLivArea     BsmtQual    YearBuilt
##            7            4            4            4            4
##   Foundation  GarageYrBlt     2ndFlrSF TotRmsAbvGrd     1stFlrSF
##            3            3            2            1            1
## BedroomAbvGr YearRemodAdd   HouseStyle
##            1            1            1
##
## Node number 1: 1460 observations,    complexity param=0.4543763
##   mean=180921.2, MSE=6.306789e+09
##   left son=2 (1231 obs) right son=3 (229 obs)
##   Primary splits:
##       OverallQual  splits as  LLLLLLLRRR, improve=0.4543763, (0 missing)
##       ExterQual    splits as  RLRL, improve=0.3786159, (0 missing)
##       GarageCars   splits as  LLLRL, improve=0.3717557, (0 missing)
##       Neighborhood splits as  LLLLLLLLLLLLLRLRLLLLRRLRR, improve=0.3460826, (0 missing)
##       GrLivArea    < 1488   to the left,  improve=0.3280228, (0 missing)
##   Surrogate splits:
##       GarageCars   splits as  LLLRL, agree=0.899, adj=0.354, (0 split)
##       Neighborhood splits as  LLLLLLLLLLLLLRLRLLLLLRLLL, agree=0.896, adj=0.336, (0 split)
##       GarageArea   < 690.5  to the left,  agree=0.888, adj=0.288, (0 split)
##       KitchenQual  splits as  RLLL, agree=0.884, adj=0.262, (0 split)
##       TotalBsmtSF  < 1560.5 to the left,  agree=0.880, adj=0.236, (0 split)
##
## Node number 2: 1231 observations,    complexity param=0.1205024
##   mean=157832.4, MSE=2.426929e+09
##   left son=4 (713 obs) right son=5 (518 obs)
##   Primary splits:
##       Neighborhood splits as  RLLLRRRLRLLLLRLRRLLRRRLRR, improve=0.3713998, (0 missing)
##       OverallQual  splits as  LLLLLLR---, improve=0.3586420, (0 missing)
##       GrLivArea    < 1413   to the left,  improve=0.3344517, (0 missing)
##       FullBath     splits as  LLRR, improve=0.3160671, (0 missing)
##       KitchenQual  splits as  RLRL, improve=0.2793628, (0 missing)
##   Surrogate splits:
##       YearBuilt   < 1971.5 to the left,  agree=0.833, adj=0.602, (0 split)
##       BsmtQual    splits as  RLRL,       agree=0.794, adj=0.510, (0 split)
##       ExterQual   splits as  RLRL,       agree=0.773, adj=0.459, (0 split)
##       Foundation  splits as  LLRLLL,     agree=0.767, adj=0.446, (0 split)
##       OverallQual splits as  LLLLLLR---, agree=0.760, adj=0.431, (0 split)
##
## Node number 3: 229 observations,    complexity param=0.06288535
##   mean=305035.9, MSE=8.893039e+09
##   left son=6 (168 obs) right son=7 (61 obs)
##   Primary splits:
##       OverallQual  splits as  -------LRR, improve=0.2843315, (0 missing)
##       TotalBsmtSF  < 1846   to the left,  improve=0.2508392, (0 missing)
##       1stFlrSF     < 1829.5 to the left,  improve=0.2362698, (0 missing)
##       TotRmsAbvGrd < 9.5    to the left,  improve=0.2297492, (0 missing)
##       GarageCars   splits as  LLLR-,      improve=0.2272069, (0 missing)
##   Surrogate splits:
```

```
##       ExterQual   splits as  R-LL,       agree=0.865, adj=0.492, (0 split)
##       KitchenQual splits as  R-LL,       agree=0.821, adj=0.328, (0 split)
##       BsmtQual    splits as  R-LL,       agree=0.795, adj=0.230, (0 split)
##       TotalBsmtSF < 1818   to the left,  agree=0.786, adj=0.197, (0 split)
##       MasVnrArea  < 662    to the left,  agree=0.773, adj=0.148, (0 split)
##
## Node number 4: 713 observations,    complexity param=0.02066528
##   mean=132242.5, MSE=1.226583e+09
##   left son=8 (410 obs) right son=9 (303 obs)
##   Primary splits:
##       1stFlrSF    < 1050.5 to the left,  improve=0.2175785, (0 missing)
##       GrLivArea   < 1377   to the left,  improve=0.2131900, (0 missing)
##       OverallQual splits as  LLLLRRR---, improve=0.1819676, (0 missing)
##       Fireplaces  splits as  LRRR,       improve=0.1722454, (0 missing)
##       TotalBsmtSF < 1050.5 to the left,  improve=0.1707600, (0 missing)
##   Surrogate splits:
##       TotalBsmtSF < 1050.5 to the left,  agree=0.850, adj=0.647, (0 split)
##       GrLivArea   < 1051   to the left,  agree=0.735, adj=0.376, (0 split)
##       GarageType  splits as  RRRLRLL,    agree=0.697, adj=0.287, (0 split)
##       GarageArea  < 440.5  to the left,  agree=0.669, adj=0.221, (0 split)
##       LotArea     < 9100.5 to the left,  agree=0.663, adj=0.208, (0 split)
##
## Node number 5: 518 observations,    complexity param=0.03312902
##   mean=193055.7, MSE=1.937105e+09
##   left son=10 (350 obs) right son=11 (168 obs)
##   Primary splits:
##       GrLivArea    < 1719   to the left,  improve=0.3040093, (0 missing)
##       OverallQual  splits as  ---LLLR---, improve=0.2202643, (0 missing)
##       2ndFlrSF     < 881.5  to the left,  improve=0.1953041, (0 missing)
##       FullBath     splits as  RLRR,       improve=0.1826389, (0 missing)
##       TotRmsAbvGrd < 6.5    to the left,  improve=0.1569699, (0 missing)
##   Surrogate splits:
##       2ndFlrSF     < 855.5  to the left,  agree=0.865, adj=0.583, (0 split)
##       TotRmsAbvGrd < 7.5    to the left,  agree=0.805, adj=0.399, (0 split)
##       BedroomAbvGr splits as  LLLLRRR-, agree=0.766, adj=0.280, (0 split)
##       HouseStyle   splits as  RLLRLRLL, agree=0.743, adj=0.208, (0 split)
##       GarageYrBlt  splits as  --R-R--RRLLL-LL--R-LR---R-LLRLLRR--RLL-RLLRRLRL--LLRLRRLLLLLLLLLLLLL-LL
##
## Node number 6: 168 observations,    complexity param=0.01856908
##   mean=274735.5, MSE=4.058766e+09
##   left son=12 (103 obs) right son=13 (65 obs)
##   Primary splits:
##       GrLivArea  < 1971.5 to the left,  improve=0.2507543, (0 missing)
##       BsmtFinSF1 < 1225.5 to the left,  improve=0.1998414, (0 missing)
##       GarageArea < 662.5  to the left,  improve=0.1815255, (0 missing)
##       1stFlrSF   < 1888   to the left,  improve=0.1776766, (0 missing)
##       WoodDeckSF < 238.5  to the left,  improve=0.1750817, (0 missing)
##   Surrogate splits:
##       2ndFlrSF     < 874.5  to the left,  agree=0.845, adj=0.600, (0 split)
##       BedroomAbvGr splits as  LLLLRR--, agree=0.815, adj=0.523, (0 split)
##       TotRmsAbvGrd < 7.5    to the left,  agree=0.815, adj=0.523, (0 split)
##       HouseStyle   splits as  R-L--R-L, agree=0.762, adj=0.385, (0 split)
##       Neighborhood splits as  L----LL-L--LRR-LRL-LLL-LL, agree=0.756, adj=0.369, (0 split)
##
```

```
## Node number 7: 61 observations,    complexity param=0.03404531
##   mean=388486.1, MSE=1.27146e+10
##   left son=14 (49 obs) right son=15 (12 obs)
##   Primary splits:
##       GarageYrBlt  splits as  -------------------------------------------------------------------L--
##       GrLivArea    < 2229   to the left,  improve=0.2362506, (0 missing)
##       Neighborhood splits as  -----L-LL----R-L-L--LR-LL, improve=0.2217872, (0 missing)
##       BedroomAbvGr splits as  LRLLR---, improve=0.2157154, (0 missing)
##       Fireplaces   splits as  LLRL, improve=0.1978834, (0 missing)
##   Surrogate splits:
##       YearRemodAdd < 2008.5 to the left,  agree=0.885, adj=0.417, (0 split)
##       Neighborhood splits as  -----L-LL----R-L-L--LL-LL, agree=0.852, adj=0.250, (0 split)
##       Exterior1st  splits as  -----LR-L--LLRL, agree=0.852, adj=0.250, (0 split)
##       2ndFlrSF     < 1667   to the left,  agree=0.852, adj=0.250, (0 split)
##       GrLivArea    < 3042.5 to the left,  agree=0.852, adj=0.250, (0 split)
##
## Node number 8: 410 observations
##   mean=118198.6, MSE=7.937454e+08
##
## Node number 9: 303 observations
##   mean=151245.7, MSE=1.18427e+09
##
## Node number 10: 350 observations,    complexity param=0.01392462
##   mean=176242.8, MSE=1.077342e+09
##   left son=20 (63 obs) right son=21 (287 obs)
##   Primary splits:
##       GrLivArea   < 1120   to the left,  improve=0.3400342, (0 missing)
##       TotalBsmtSF < 1272.5 to the left,  improve=0.2960171, (0 missing)
##       1stFlrSF    < 1199.5 to the left,  improve=0.2510921, (0 missing)
##       FullBath    splits as  RLR-,       improve=0.2452918, (0 missing)
##       OverallQual splits as  ---LLRR---, improve=0.2261882, (0 missing)
##   Surrogate splits:
##       FullBath     splits as  RLR-, agree=0.860, adj=0.222, (0 split)
##       OverallQual  splits as  ---LLRR---, agree=0.857, adj=0.206, (0 split)
##       MSZoning     splits as  -RLRL, agree=0.854, adj=0.190, (0 split)
##       TotRmsAbvGrd < 3.5    to the left,  agree=0.840, adj=0.111, (0 split)
##       GarageYrBlt  splits as  ---------RRR-RL----R------LRRRR-----LL-RRL--RRR--RR-RRRRRRRRLRRRRRR-L
##
## Node number 11: 168 observations,    complexity param=0.01094828
##   mean=228082.4, MSE=1.912509e+09
##   left son=22 (92 obs) right son=23 (76 obs)
##   Primary splits:
##       GarageYrBlt splits as  --R-L--RL-----R--L--R---R--RRR-RL--L---L-LLLRR-----RLLLL-RLLRLLLLL---RL
##       BsmtFinSF1  < 860.5  to the left,  improve=0.2913054, (0 missing)
##       TotalBsmtSF < 1107.5 to the left,  improve=0.2841683, (0 missing)
##       1stFlrSF    < 1177.5 to the left,  improve=0.1852893, (0 missing)
##       OverallQual splits as  ---LLLR---, improve=0.1838032, (0 missing)
##   Surrogate splits:
##       GarageArea   < 566    to the left,  agree=0.689, adj=0.316, (1 split)
##       Neighborhood splits as  ----LLL-L----R-RL--RRL-RL, agree=0.671, adj=0.276, (0 split)
##       GarageCars   splits as  -RLR-, agree=0.665, adj=0.263, (0 split)
##       BsmtFinSF1   < 817.5  to the left,  agree=0.647, adj=0.224, (0 split)
##       YearBuilt    < 2004.5 to the left,  agree=0.641, adj=0.211, (0 split)
##
```

```
## Node number 12: 103 observations
##   mean=249392.5, MSE=2.332109e+09
##
## Node number 13: 65 observations
##   mean=314894.6, MSE=4.164354e+09
##
## Node number 14: 49 observations
##   mean=353009.9, MSE=5.562902e+09
##
## Node number 15: 12 observations
##   mean=533347.2, MSE=1.579351e+10
##
## Node number 20: 63 observations
##   mean=135391.3, MSE=3.698856e+08
##
## Node number 21: 287 observations
##   mean=185210.3, MSE=7.858899e+08
##
## Node number 22: 92 observations
##   mean=205817.9, MSE=8.918343e+08
##
## Node number 23: 76 observations
##   mean=255034.1, MSE=1.821605e+09
```