House Sale Prediction

Joel

2022-11-19

Summary: This data from Kaggle is referred to as the Ames dataset. Is is composed mostly of single family suburban dwellings that were sold in Ames, Iowa from 2006-2010. It consists of 81 varibables and a total 1456 objected composed of character, integer and numerical data. The goal is to use the train data to build a model to predict the SalePrice (the final variable) in the test set. The purpose of this is to attempt to minimize both the RMSE and run-time in order to have a tool both accurate and efficient at assessing house costs.

The data is already split into two sets, a train set and a test set. The train set consists of 81 variables and a total of 1456 objects composed of either character, numeric or integer type information. However, due to it being taken from a contest dataset, only the train set will be used and split into train and test sets. The test set does not have the SalePrice variable included unfortunately. The data is referred to as the Ames dataset. It is mostly composed of single family suburban dwellings, which were sold in Ames, Iowa in the period 2006-2010. The data was obtained from kaggle.com. The links are below, though you should not need them.

https://www.kaggle.com/datasets/rsizem2/house-prices-ames-cleaned-dataset?resource=download&select=clean_train.csv https://www.kaggle.com/datasets/rsizem2/house-prices-ames-cleaned-dataset?resource=download&select=clean_test.csv

The goal is to use the train data to build a model to predict the SalePrice (the final variable) in the test set, while attempting to minimize both the RMSE and run-time in order to have a tool both accurate and efficient at assessing house costs.

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: caret
##
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following object is masked from 'package:purrr':
##
```

```
##     lift
##
##
## Loading required package: data.table
##
##
## Attaching package: 'data.table'
##
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
##
##
## The following object is masked from 'package:purrr':
##
##     transpose
##
##
## Loading required package: randomForest
##
## randomForest 4.7-1.1
##
## Type rfNews() to see new features/changes/bug fixes.
##
##
## Attaching package: 'randomForest'
##
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
##
## The following object is masked from 'package:ggplot2':
##
##     margin
##
##
## Loading required package: bayesplot
##
## This is bayesplot version 1.9.0
##
## - Online documentation and vignettes at mc-stan.org/bayesplot
##
## - bayesplot theme set to bayesplot::theme_default()
##
##    * Does _not_ affect other ggplot2 plots
##
##    * See ?bayesplot_theme_set for details on theme setting
##
## Loading required package: boot
##
##
```

```
## Attaching package: 'boot'
##
##
## The following object is masked from 'package:lattice':
##
##     melanoma
##
##
## Loading required package: dbarts
##
##
## Attaching package: 'dbarts'
##
##
## The following object is masked from 'package:tidyr':
##
##     extract
##
##
## Loading required package: cowplot

## Warning: package 'cowplot' was built under R version 4.2.2

## Loading required package: rJava
## Loading required package: bartMachine

## Warning: package 'bartMachine' was built under R version 4.2.2

## Loading required package: bartMachineJARs
## Loading required package: missForest

## Warning: package 'missForest' was built under R version 4.2.2

## Welcome to bartMachine v1.3.2! You have 0.54GB memory available.
##
## If you run out of memory, restart R, and use e.g.
## 'options(java.parameters = "-Xmx5g")' for 5GB of RAM before you call
## 'library(bartMachine)'.
```

Exploration and Visualization:

First we ready the data for exploration. The data well be imported in from the github links below.

```
## 'data.frame':    1456 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : chr  "2-STORY 1946+" "1-STORY 1946+" "2-STORY 1946+" "2-STORY 1945-" ...
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ LotFrontage  : num  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  "" "" "" "" ...
##  $ LotShape     : int  0 0 1 1 1 1 0 1 0 0 ...
```

```
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ 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" "WdShing" ...
##  $ MasVnrType   : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : num  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : int  4 3 4 3 4 3 4 3 3 3 ...
##  $ ExterCond    : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Foundation   : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : int  4 4 4 3 4 4 5 4 3 3 ...
##  $ BsmtCond     : int  3 3 3 4 3 3 3 3 3 3 ...
##  $ BsmtExposure : int  0 3 1 0 2 0 2 1 0 0 ...
##  $ BsmtFinType1 : int  6 5 6 5 6 6 6 5 1 6 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : int  1 1 1 1 1 1 1 4 1 1 ...
##  $ 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    : int  5 5 5 4 5 5 5 5 4 5 ...
##  $ CentralAir   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Electrical   : int  5 5 5 5 5 5 5 5 3 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  : int  4 3 4 4 4 3 4 3 3 3 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : int  0 3 3 4 3 0 4 3 3 3 ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : num  2003 1976 2001 1998 2000 ...
##  $ GarageFinish : int  2 2 2 1 2 1 2 2 1 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   : int  3 3 3 3 3 3 3 3 2 4 ...
##  $ GarageCond   : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ PavedDrive   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ 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       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fence        : int  0 0 0 0 0 3 0 0 0 0 ...
##  $ MiscFeature  : chr  "" "" "" "" ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : chr  "Feb" "May" "Sept" "Feb" ...
##  $ 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 ...
```

```
## 'data.frame':    1459 obs. of  80 variables:
##  $ Id           : int  1461 1462 1463 1464 1465 1466 1467 1468 1469 1470 ...
##  $ MSSubClass   : chr  "1-STORY 1946+" "1-STORY 1946+" "2-STORY 1946+" "2-STORY 1946+" ...
##  $ MSZoning     : chr  "RH" "RL" "RL" "RL" ...
##  $ LotFrontage  : num  80 81 74 78 43 75 NA 63 85 70 ...
##  $ LotArea      : int  11622 14267 13830 9978 5005 10000 7980 8402 10176 8400 ...
##  $ Street       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  "" "" "" "" ...
##  $ LotShape     : int  0 1 1 1 1 1 1 1 0 0 ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ LotConfig    : chr  "Inside" "Corner" "Inside" "Inside" ...
##  $ LandSlope    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Neighborhood : chr  "NAmes" "NAmes" "Gilbert" "Gilbert" ...
##  $ Condition1   : chr  "Feedr" "Norm" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "1Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  5 6 5 6 8 6 6 6 7 4 ...
##  $ OverallCond  : int  6 6 5 6 5 5 7 5 5 5 ...
##  $ YearBuilt    : int  1961 1958 1997 1998 1992 1993 1992 1998 1990 1970 ...
##  $ YearRemodAdd : int  1961 1958 1998 1998 1992 1994 2007 1998 1990 1970 ...
##  $ RoofStyle    : chr  "Gable" "Hip" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "Wd Sdng" "VinylSd" "VinylSd" ...
##  $ Exterior2nd  : chr  "VinylSd" "Wd Sdng" "VinylSd" "VinylSd" ...
##  $ MasVnrType   : chr  "None" "BrkFace" "None" "BrkFace" ...
##  $ MasVnrArea   : num  0 108 0 20 0 0 0 0 0 0 ...
##  $ ExterQual    : int  3 3 3 3 4 3 3 3 3 3 ...
##  $ ExterCond    : int  3 3 3 3 3 3 4 3 3 3 ...
##  $ Foundation   : chr  "CBlock" "CBlock" "PConc" "PConc" ...
##  $ BsmtQual     : int  3 3 4 3 4 4 4 4 4 3 ...
##  $ BsmtCond     : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ BsmtExposure : int  0 0 0 0 0 0 0 0 3 0 ...
```

```
##  $ BsmtFinType1 : int  3 5 6 6 5 1 5 1 6 5 ...
##  $ BsmtFinSF1   : num  468 923 791 602 263 0 935 0 637 804 ...
##  $ BsmtFinType2 : int  3 1 1 1 1 1 1 1 1 3 ...
##  $ BsmtFinSF2   : num  144 0 0 0 0 0 0 0 0 78 ...
##  $ BsmtUnfSF    : num  270 406 137 324 1017 ...
##  $ TotalBsmtSF  : num  882 1329 928 926 1280 ...
##  $ Heating      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : int  3 3 4 5 5 4 5 4 4 3 ...
##  $ CentralAir   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Electrical   : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ X1stFlrSF    : int  896 1329 928 926 1280 763 1187 789 1341 882 ...
##  $ X2ndFlrSF    : int  0 0 701 678 0 892 0 676 0 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  896 1329 1629 1604 1280 1655 1187 1465 1341 882 ...
##  $ BsmtFullBath : num  0 0 0 0 0 0 1 0 1 1 ...
##  $ BsmtHalfBath : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  1 1 2 2 2 2 2 2 1 1 ...
##  $ HalfBath     : int  0 1 1 1 0 1 0 1 1 0 ...
##  $ BedroomAbvGr : int  2 3 3 3 2 3 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ KitchenQual  : int  3 4 3 4 4 3 3 3 4 3 ...
##  $ TotRmsAbvGrd : int  5 6 6 7 5 7 6 7 5 4 ...
##  $ Functional   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fireplaces   : int  0 0 1 1 0 1 0 1 1 0 ...
##  $ FireplaceQu  : int  0 0 3 4 0 3 0 4 1 0 ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Attchd" ...
##  $ GarageYrBlt  : num  1961 1958 1997 1998 1992 ...
##  $ GarageFinish : int  1 1 3 3 2 3 3 3 1 3 ...
##  $ GarageCars   : num  1 1 2 2 2 2 2 2 2 2 ...
##  $ GarageArea   : num  730 312 482 470 506 440 420 393 506 525 ...
##  $ GarageQual   : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ GarageCond   : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ PavedDrive   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ WoodDeckSF   : int  140 393 212 360 0 157 483 0 192 240 ...
##  $ OpenPorchSF  : int  0 36 34 36 82 84 21 75 0 0 ...
##  $ EnclosedPorch: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ScreenPorch  : int  120 0 0 0 144 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fence        : int  3 0 3 0 0 0 4 0 0 3 ...
##  $ MiscFeature  : chr  "" "Gar2" "" "" ...
##  $ MiscVal      : int  0 12500 0 0 0 0 500 0 0 0 ...
##  $ MoSold       : chr  "June" "June" "Mar" "June" ...
##  $ YrSold       : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ SaleType     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Normal" ...
```

```
##    Id      MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
## 1  1   2-STORY 1946+       RL          65    8450   Pave              0
## 2  2   1-STORY 1946+       RL          80    9600   Pave              0
## 3  3   2-STORY 1946+       RL          68   11250   Pave              1
## 4  4   2-STORY 1945-       RL          60    9550   Pave              1
## 5  5   2-STORY 1946+       RL          84   14260   Pave              1
```

```
## 6  6 1-1/2 STORY FIN       RL          85   14115   Pave              1
##   LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2
## 1         Lvl         3    Inside         0      CollgCr       Norm       Norm
## 2         Lvl         3       FR2         0      Veenker      Feedr       Norm
## 3         Lvl         3    Inside         0      CollgCr       Norm       Norm
## 4         Lvl         3    Corner         0      Crawfor       Norm       Norm
## 5         Lvl         3       FR2         0      NoRidge       Norm       Norm
## 6         Lvl         3    Inside         0      Mitchel       Norm       Norm
##   BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle
## 1     1Fam     2Story           7           5      2003         2003     Gable
## 2     1Fam     1Story           6           8      1976         1976     Gable
## 3     1Fam     2Story           7           5      2001         2002     Gable
## 4     1Fam     2Story           7           5      1915         1970     Gable
## 5     1Fam     2Story           8           5      2000         2000     Gable
## 6     1Fam     1.5Fin           5           5      1993         1995     Gable
##   RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond
## 1  CompShg     VinylSd     VinylSd    BrkFace        196         4         3
## 2  CompShg     MetalSd     MetalSd       None          0         3         3
## 3  CompShg     VinylSd     VinylSd    BrkFace        162         4         3
## 4  CompShg     Wd Sdng     WdShing       None          0         3         3
## 5  CompShg     VinylSd     VinylSd    BrkFace        350         4         3
## 6  CompShg     VinylSd     VinylSd       None          0         3         3
##   Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1
## 1      PConc        4        3            0            6        706
## 2     CBlock        4        3            3            5        978
## 3      PConc        4        3            1            6        486
## 4     BrkTil        3        4            0            5        216
## 5      PConc        4        3            2            6        655
## 6       Wood        4        3            0            6        732
##   BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir
## 1            1          0       150         856    GasA         5          1
## 2            1          0       284        1262    GasA         5          1
## 3            1          0       434         920    GasA         5          1
## 4            1          0       540         756    GasA         4          1
## 5            1          0       490        1145    GasA         5          1
## 6            1          0        64         796    GasA         5          1
##   Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath
## 1          5       856       854            0      1710            1
## 2          5      1262         0            0      1262            0
## 3          5       920       866            0      1786            1
## 4          5       961       756            0      1717            1
## 5          5      1145      1053            0      2198            1
## 6          5       796       566            0      1362            1
##   BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual
## 1            0        2        1            3            1           4
## 2            1        2        0            3            1           3
## 3            0        2        1            3            1           4
## 4            0        1        0            3            1           4
## 5            0        2        1            4            1           4
## 6            0        1        1            1            1           3
##   TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt
## 1            8          0          0           0     Attchd        2003
## 2            6          0          1           3     Attchd        1976
## 3            6          0          1           3     Attchd        2001
```

```
## 4            7          0          1           4     Detchd        1998
## 5            9          0          1           3     Attchd        2000
## 6            5          0          0           0     Attchd        1993
##   GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive
## 1            2          2        548          3          3          2
## 2            2          2        460          3          3          2
## 3            2          2        608          3          3          2
## 4            1          3        642          3          3          2
## 5            2          3        836          3          3          2
## 6            1          2        480          3          3          2
##   WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC
## 1          0          61             0          0           0        0      0
## 2        298           0             0          0           0        0      0
## 3          0          42             0          0           0        0      0
## 4          0          35           272          0           0        0      0
## 5        192          84             0          0           0        0      0
## 6         40          30             0        320           0        0      0
##   Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1     0                   0    Feb   2008       WD        Normal    208500
## 2     0                   0    May   2007       WD        Normal    181500
## 3     0                   0   Sept   2008       WD        Normal    223500
## 4     0                   0    Feb   2006       WD       Abnorml    140000
## 5     0                   0    Dec   2008       WD        Normal    250000
## 6     3        Shed     700    Oct   2009       WD        Normal    143000
```

```
##  [1] "Id"            "MSSubClass"    "MSZoning"      "LotFrontage"
##  [5] "LotArea"       "Street"        "Alley"         "LotShape"
##  [9] "LandContour"   "Utilities"     "LotConfig"     "LandSlope"
## [13] "Neighborhood"  "Condition1"    "Condition2"    "BldgType"
## [17] "HouseStyle"    "OverallQual"   "OverallCond"   "YearBuilt"
## [21] "YearRemodAdd"  "RoofStyle"     "RoofMatl"      "Exterior1st"
## [25] "Exterior2nd"   "MasVnrType"    "MasVnrArea"    "ExterQual"
## [29] "ExterCond"     "Foundation"    "BsmtQual"      "BsmtCond"
## [33] "BsmtExposure"  "BsmtFinType1"  "BsmtFinSF1"    "BsmtFinType2"
## [37] "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"   "Heating"
## [41] "HeatingQC"     "CentralAir"    "Electrical"    "X1stFlrSF"
## [45] "X2ndFlrSF"     "LowQualFinSF"  "GrLivArea"     "BsmtFullBath"
## [49] "BsmtHalfBath"  "FullBath"      "HalfBath"      "BedroomAbvGr"
## [53] "KitchenAbvGr"  "KitchenQual"   "TotRmsAbvGrd"  "Functional"
## [57] "Fireplaces"    "FireplaceQu"   "GarageType"    "GarageYrBlt"
## [61] "GarageFinish"  "GarageCars"    "GarageArea"    "GarageQual"
## [65] "GarageCond"    "PavedDrive"    "WoodDeckSF"    "OpenPorchSF"
## [69] "EnclosedPorch" "X3SsnPorch"    "ScreenPorch"   "PoolArea"
## [73] "PoolQC"        "Fence"         "MiscFeature"   "MiscVal"
## [77] "MoSold"        "YrSold"        "SaleType"      "SaleCondition"
## [81] "SalePrice"
```

```
## [1] 1456   81
```

```
##        Id          MSSubClass      MSZoning         LotFrontage
##  Min.   :   1.0   Length:1456        Length:1456        Min.   : 21.00
##  1st Qu.: 364.8   Class :character   Class :character   1st Qu.: 59.00
##  Median : 730.5   Mode  :character   Mode  :character   Median : 69.00
```

```
##  Mean   : 730.0                                          Mean   : 69.69
##  3rd Qu.:1094.2                                          3rd Qu.: 80.00
##  Max.   :1460.0                                          Max.   :313.00
##                                                          NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1456        Length:1456        Min.   :0.0000
##  1st Qu.:  7539   Class :character   Class :character   1st Qu.:0.0000
##  Median :  9468   Mode  :character   Mode  :character   Median :0.0000
##  Mean   : 10449                                         Mean   :0.4052
##  3rd Qu.: 11588                                         3rd Qu.:1.0000
##  Max.   :215245                                         Max.   :3.0000
##
##  LandContour          Utilities      LotConfig          LandSlope
##  Length:1456        Min.   :1.000   Length:1456        Min.   :0.0000
##  Class :character   1st Qu.:3.000   Class :character   1st Qu.:0.0000
##  Mode  :character   Median :3.000   Mode  :character   Median :0.0000
##                     Mean   :2.999                      Mean   :0.0625
##                     3rd Qu.:3.000                      3rd Qu.:0.0000
##                     Max.   :3.000                      Max.   :2.0000
##
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1456        Length:1456        Length:1456        Length:1456
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1456        Min.   : 1.000   Min.   :1.000   Min.   :1872
##  Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Mode  :character   Median : 6.000   Median :5.000   Median :1972
##                     Mean   : 6.089   Mean   :5.576   Mean   :1971
##                     3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                     Max.   :10.000   Max.   :9.000   Max.   :2010
##
##   YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
##  Min.   :1950   Length:1456        Length:1456        Length:1456
##  1st Qu.:1967   Class :character   Class :character   Class :character
##  Median :1994   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1985
##  3rd Qu.:2004
##  Max.   :2010
##
##  Exterior2nd         MasVnrType          MasVnrArea       ExterQual
##  Length:1456        Length:1456        Min.   :   0.0   Min.   :2.000
##  Class :character   Class :character   1st Qu.:   0.0   1st Qu.:3.000
##  Mode  :character   Mode  :character   Median :   0.0   Median :3.000
##                                        Mean   : 101.5   Mean   :3.392
##                                        3rd Qu.: 163.2   3rd Qu.:4.000
##                                        Max.   :1600.0   Max.   :5.000
##
##    ExterCond      Foundation           BsmtQual        BsmtCond
##  Min.   :1.000   Length:1456        Min.   :0.000   Min.   :0.000
```

```
##  1st Qu.:3.000   Class :character   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Mode  :character   Median :4.000   Median :3.000
##  Mean   :3.084                      Mean   :3.485   Mean   :2.935
##  3rd Qu.:3.000                      3rd Qu.:4.000   3rd Qu.:3.000
##  Max.   :5.000                      Max.   :5.000   Max.   :4.000
##
##   BsmtExposure     BsmtFinType1     BsmtFinSF1      BsmtFinType2
##  Min.   :0.0000   Min.   :0.000   Min.   :   0.0   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:   0.0   1st Qu.:1.000
##  Median :0.0000   Median :4.000   Median : 381.0   Median :1.000
##  Mean   :0.6504   Mean   :3.539   Mean   : 437.0   Mean   :1.283
##  3rd Qu.:1.0000   3rd Qu.:6.000   3rd Qu.: 706.5   3rd Qu.:1.000
##  Max.   :3.0000   Max.   :6.000   Max.   :2188.0   Max.   :6.000
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1456
##  1st Qu.:   0.00   1st Qu.: 222.5   1st Qu.: 795.0   Class :character
##  Median :   0.00   Median : 477.5   Median : 990.5   Mode  :character
##  Mean   :  46.68   Mean   : 567.0   Mean   :1050.7
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1293.8
##  Max.   :1474.00   Max.   :2336.0   Max.   :3206.0
##
##    HeatingQC       CentralAir       Electrical      X1stFlrSF
##  Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   : 334
##  1st Qu.:3.000   1st Qu.:1.0000   1st Qu.:5.000   1st Qu.: 882
##  Median :5.000   Median :1.0000   Median :5.000   Median :1086
##  Mean   :4.143   Mean   :0.9348   Mean   :4.886   Mean   :1157
##  3rd Qu.:5.000   3rd Qu.:1.0000   3rd Qu.:5.000   3rd Qu.:1389
##  Max.   :5.000   Max.   :1.0000   Max.   :5.000   Max.   :3228
##
##    X2ndFlrSF       LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0.0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0.0   1st Qu.:  0.000   1st Qu.:1128   1st Qu.:0.0000
##  Median :   0.0   Median :  0.000   Median :1458   Median :0.0000
##  Mean   : 343.5   Mean   :  5.861   Mean   :1507   Mean   :0.4238
##  3rd Qu.: 728.0   3rd Qu.:  0.000   3rd Qu.:1775   3rd Qu.:1.0000
##  Max.   :1818.0   Max.   :572.000   Max.   :3627   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   :0.05701   Mean   :1.562   Mean   :0.3812   Mean   :2.865
##  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   Min.   :2.000   Min.   : 2.000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 5.000   1st Qu.:0.0000
##  Median :1.000   Median :3.000   Median : 6.000   Median :0.0000
##  Mean   :1.047   Mean   :3.508   Mean   : 6.506   Mean   :0.1648
##  3rd Qu.:1.000   3rd Qu.:4.000   3rd Qu.: 7.000   3rd Qu.:0.0000
##  Max.   :3.000   Max.   :5.000   Max.   :14.000   Max.   :6.0000
##
```

```
##    Fireplaces      FireplaceQu     GarageType         GarageYrBlt
##  Min.   :0.0000   Min.   :0.000   Length:1456        Min.   :1872
##  1st Qu.:0.0000   1st Qu.:0.000   Class :character   1st Qu.:1959
##  Median :1.0000   Median :2.000   Mode  :character   Median :1978
##  Mean   :0.6092   Mean   :1.819                      Mean   :1976
##  3rd Qu.:1.0000   3rd Qu.:4.000                      3rd Qu.:2001
##  Max.   :3.0000   Max.   :5.000                      Max.   :2010
##
##   GarageFinish     GarageCars      GarageArea       GarageQual
##  Min.   :0.000   Min.   :0.000   Min.   :   0.0   Min.   :0.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 329.5   1st Qu.:3.00
##  Median :2.000   Median :2.000   Median : 478.5   Median :3.00
##  Mean   :1.712   Mean   :1.764   Mean   : 471.6   Mean   :2.81
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 576.0   3rd Qu.:3.00
##  Max.   :3.000   Max.   :4.000   Max.   :1390.0   Max.   :5.00
##
##    GarageCond      PavedDrive      WoodDeckSF      OpenPorchSF
##  Min.   :0.000   Min.   :0.000   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:  0.00   1st Qu.:  0.00
##  Median :3.000   Median :2.000   Median :  0.00   Median : 24.00
##  Mean   :2.808   Mean   :1.856   Mean   : 93.83   Mean   : 46.22
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:168.00   3rd Qu.: 68.00
##  Max.   :5.000   Max.   :2.000   Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      X3SsnPorch       ScreenPorch       PoolArea
##  Min.   :  0.00   Min.   :  0.000   Min.   :  0.0   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.0   1st Qu.:  0.000
##  Median :  0.00   Median :  0.000   Median :  0.0   Median :  0.000
##  Mean   : 22.01   Mean   :  3.419   Mean   : 15.1   Mean   :  2.056
##  3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.0   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.000   Max.   :480.0   Max.   :738.000
##
##      PoolQC            Fence        MiscFeature          MiscVal
##  Min.   :0.00000   Min.   :0.0000   Length:1456        Min.   :    0.00
##  1st Qu.:0.00000   1st Qu.:0.0000   Class :character   1st Qu.:    0.00
##  Median :0.00000   Median :0.0000   Mode  :character   Median :    0.00
##  Mean   :0.01168   Mean   :0.5652                      Mean   :   43.95
##  3rd Qu.:0.00000   3rd Qu.:0.0000                      3rd Qu.:    0.00
##  Max.   :5.00000   Max.   :4.0000                      Max.   :15500.00
##
##     MoSold           YrSold       SaleType         SaleCondition
##  Length:1456        Min.   :2006   Length:1456        Length:1456
##  Class :character   1st Qu.:2007   Class :character   Class :character
##  Mode  :character   Median :2008   Mode  :character   Mode  :character
##                     Mean   :2008
##                     3rd Qu.:2009
##                     Max.   :2010
##
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129900
##  Median :163000
##  Mean   :180151
##  3rd Qu.:214000
```

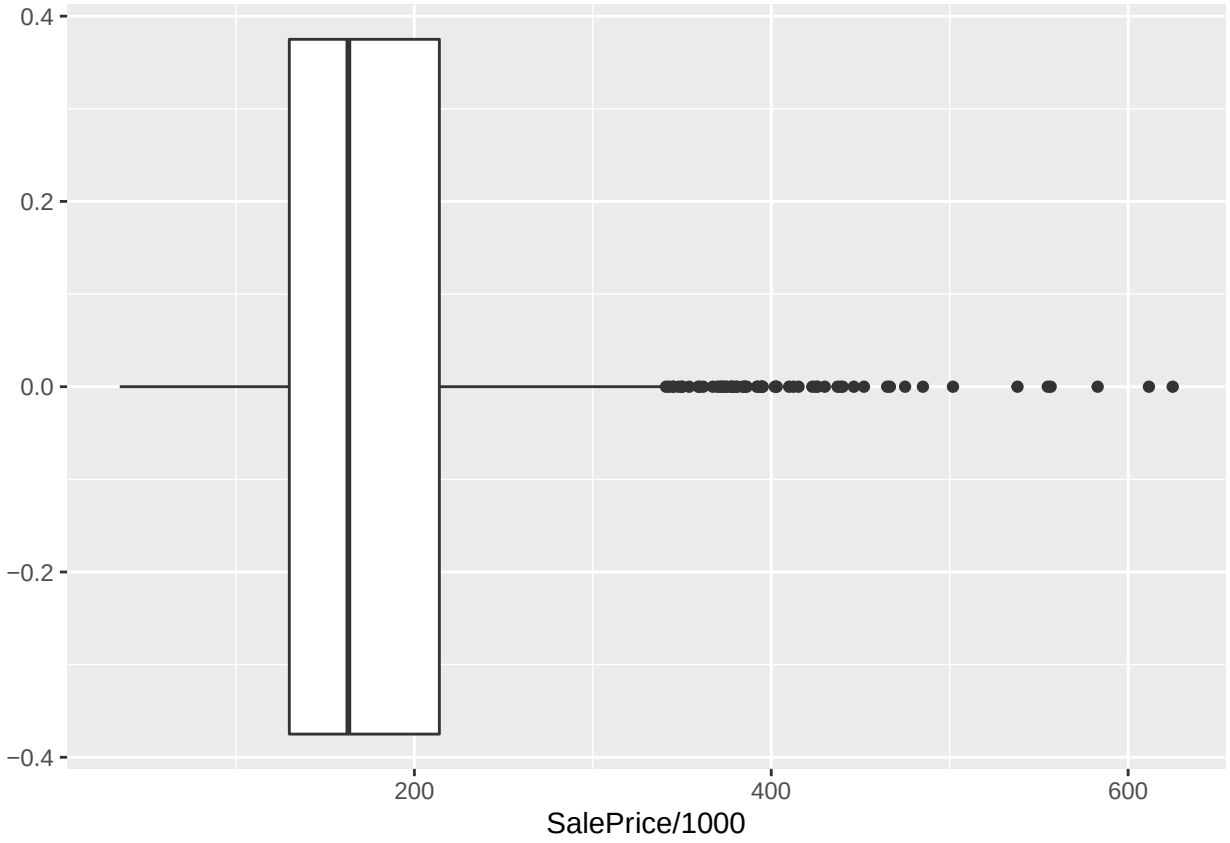
```
##  Max.   :625000
##
```

Many variables have a median of 0, but a non-zero mean and a relatively higher max. This suggests there may few occurances of a particular feature(s) but perhaps they are still relevant.

Here we look at distribution of variable of interest.

```
## [1] 163000
```

```
## [1] 56338.8
```



Next we check for columns with NAs

```
##           Id    MSSubClass      MSZoning   LotFrontage       LotArea
##            0             0             0           259             0
##       Street         Alley      LotShape   LandContour     Utilities
##            0             0             0             0             0
##    LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##            0             0             0             0             0
##     BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##            0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##            0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##            0             0             0             0             0
```

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

In the readme this value is described as "Linear feet of street connected to property". Let's find out more about LotFrontage before we decide on what to do with the NAs

```
## [1] "numeric"
```

```
##  [1] Id            MSSubClass    MSZoning      LotFrontage   LotArea
##  [6] Street        Alley         LotShape      LandContour   Utilities
## [11] LotConfig     LandSlope     Neighborhood  Condition1    Condition2
## [16] BldgType      HouseStyle    OverallQual   OverallCond   YearBuilt
## [21] YearRemodAdd  RoofStyle     RoofMatl      Exterior1st   Exterior2nd
## [26] MasVnrType    MasVnrArea    ExterQual     ExterCond     Foundation
## [31] BsmtQual      BsmtCond      BsmtExposure  BsmtFinType1  BsmtFinSF1
## [36] BsmtFinType2  BsmtFinSF2    BsmtUnfSF     TotalBsmtSF   Heating
## [41] HeatingQC     CentralAir    Electrical    X1stFlrSF     X2ndFlrSF
## [46] LowQualFinSF  GrLivArea     BsmtFullBath  BsmtHalfBath  FullBath
## [51] HalfBath      BedroomAbvGr  KitchenAbvGr  KitchenQual   TotRmsAbvGrd
## [56] Functional    Fireplaces    FireplaceQu   GarageType    GarageYrBlt
## [61] GarageFinish  GarageCars    GarageArea    GarageQual    GarageCond
## [66] PavedDrive    WoodDeckSF    OpenPorchSF   EnclosedPorch X3SsnPorch
## [71] ScreenPorch   PoolArea      PoolQC        Fence         MiscFeature
## [76] MiscVal       MoSold        YrSold        SaleType      SaleCondition
## [81] SalePrice
## <0 rows> (or 0-length row.names)
```

```
##
##  21  24  30  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
##  23  19   6   5   1  10   9   6   5   1   1  12   6   4  12   9   3   1   5   6
##  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##   4  57  15  14  10   6  17   5  12   7  13 143   8   9  17  19  44  15  12  19
##  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88
```

```
##  11  70  12  17  18  15  53  11   9  25  17  69   6  12   5   9  40  10   5  10
##  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
##   6  23   6  10   8   6   7   8   2   8   3  16   2   4   3   2   6   1   7   3
## 109 110 111 112 114 115 116 118 120 121 122 124 128 129 130 134 137 138 140 141
##   2   6   1   1   2   2   2   2   7   2   2   2   1   2   2   2   1   1   1   1
## 144 149 150 152 153 168 174 182 313
##   1   1   1   1   1   1   2   1   1
```

It looks like there are no instances where LotFrontage = 0, so the NAs are likely instances where the property does not directly abut a street. We will change those instances to a quantity, 0, rather than leave it as NA.

```
##    [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [217] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [229] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [253] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [277] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [349] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [361] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [373] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [385] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [397] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [409] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [421] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [433] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [445] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [457] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [469] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [481] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [493] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [505] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##  [517] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [529] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [541] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [553] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [565] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [577] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [589] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [601] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [613] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [625] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [637] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [649] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [661] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [673] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [685] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [697] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [709] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [721] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [733] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [745] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [757] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [769] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [781] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [793] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [805] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [817] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [829] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [841] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [853] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [865] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [877] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [889] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [901] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [913] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [925] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [937] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [949] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [961] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [973] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [985] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [997] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1009] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1021] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1033] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1045] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1057] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1069] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1081] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1093] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1105] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1117] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1129] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1141] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1153] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
## [1165] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1189] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1201] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1213] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1225] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1237] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1249] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1261] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1273] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1285] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1297] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1309] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1321] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1333] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1345] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1357] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1369] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1381] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1393] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1405] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1417] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1429] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1441] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [1453] FALSE FALSE FALSE FALSE
```

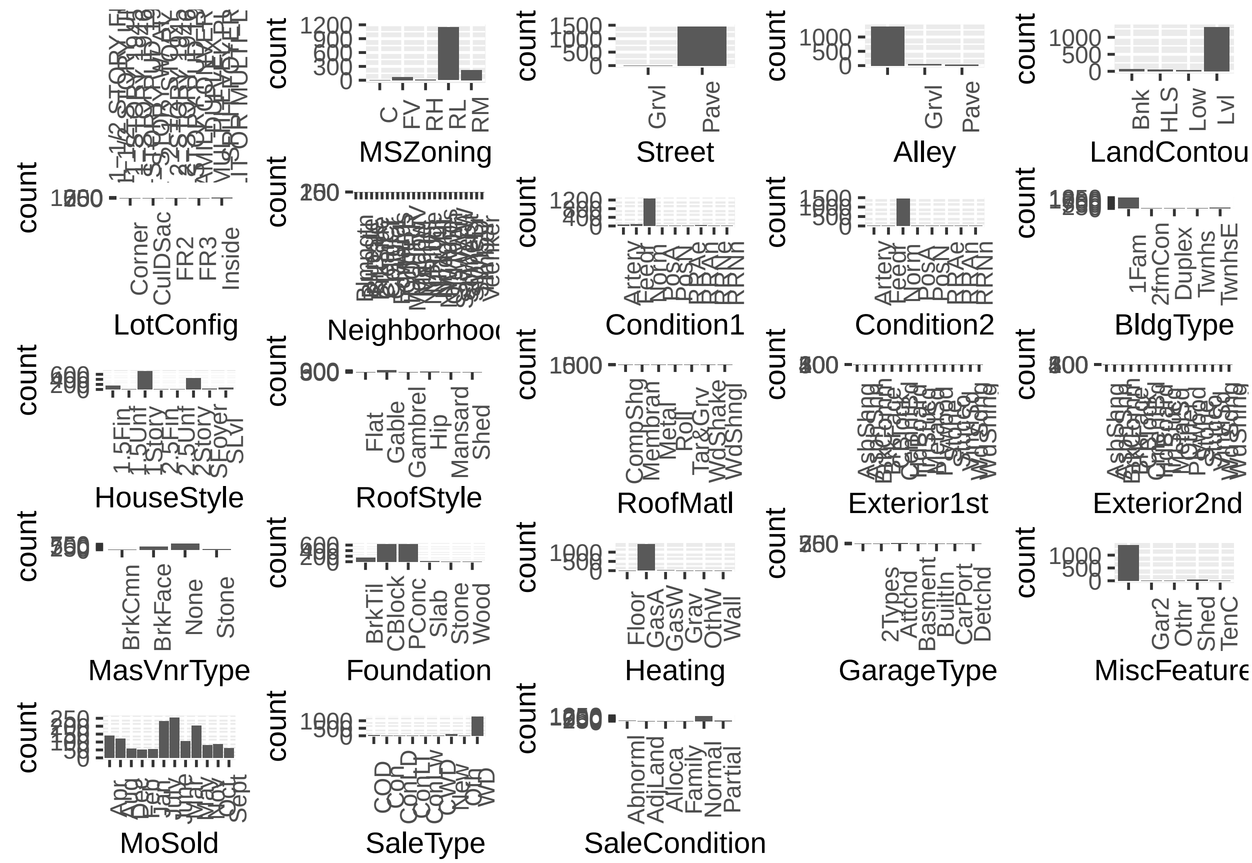
Success, the NAs have been replaced by 0's.

Next we will plot by category to get any idea of what the rest of the many variable distributions look like. We plot the character content variables first.

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

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

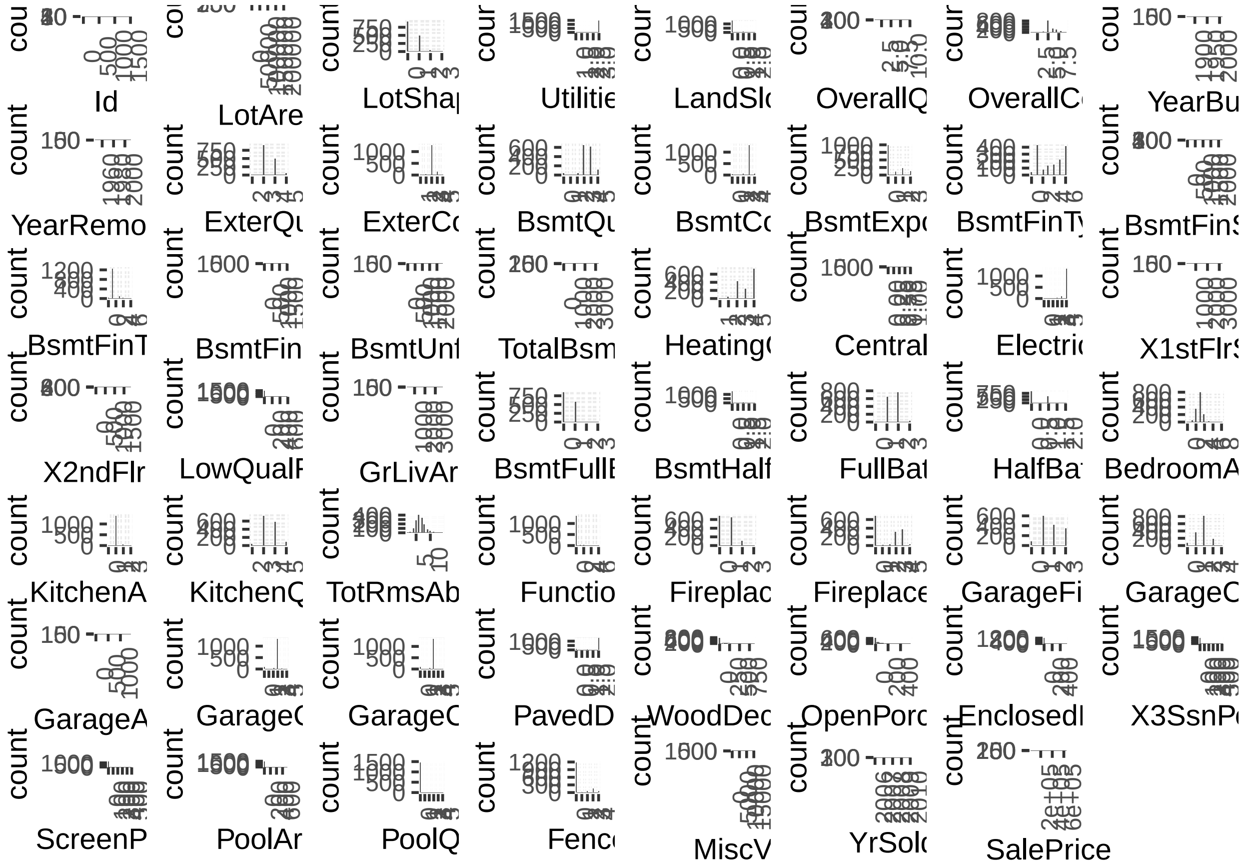
count
Id
MSSubC
MSZon
LotFront
LotAre
Stree
Alle
LotSha
LandCo
Utiliti
LotCon
LandSl
Neighbor
Conditi
Conditi
BldgTy
HouseS
OverallC
OverallC
YearBu
YearRemo
RoofSt
RoofM
Exterio
Exterior
MasVnr
MasVnrA
ExterQ
ExterC
Founda
BsmtQ
BsmtC
BsmtExp
BsmtFinT
BsmtFin
BsmtFin
BsmtFir
BsmtUn
TotalBsm
Heati
Heating
Centra
Electri
X1stFlr
X2ndFl
LowQual
GrLivA
BsmtFull
BsmtHal
FullBa
HalfBa
Bedroom
KitchenA
Kitchen
TotRmsAb
Functio
Firepla
Fireplac
Garage
GarageY
GarageF
Garage
GarageA
Garage
Garage
PavedD
WoodDe
OpenPor
Enclosed
X3SsnP
ScreenP
PoolA
PoolQ
Fenc
MiscFea
MiscV
MoSo
YrSol
SaleTy
SaleCon
SalePr



Here we plot integer variables.

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

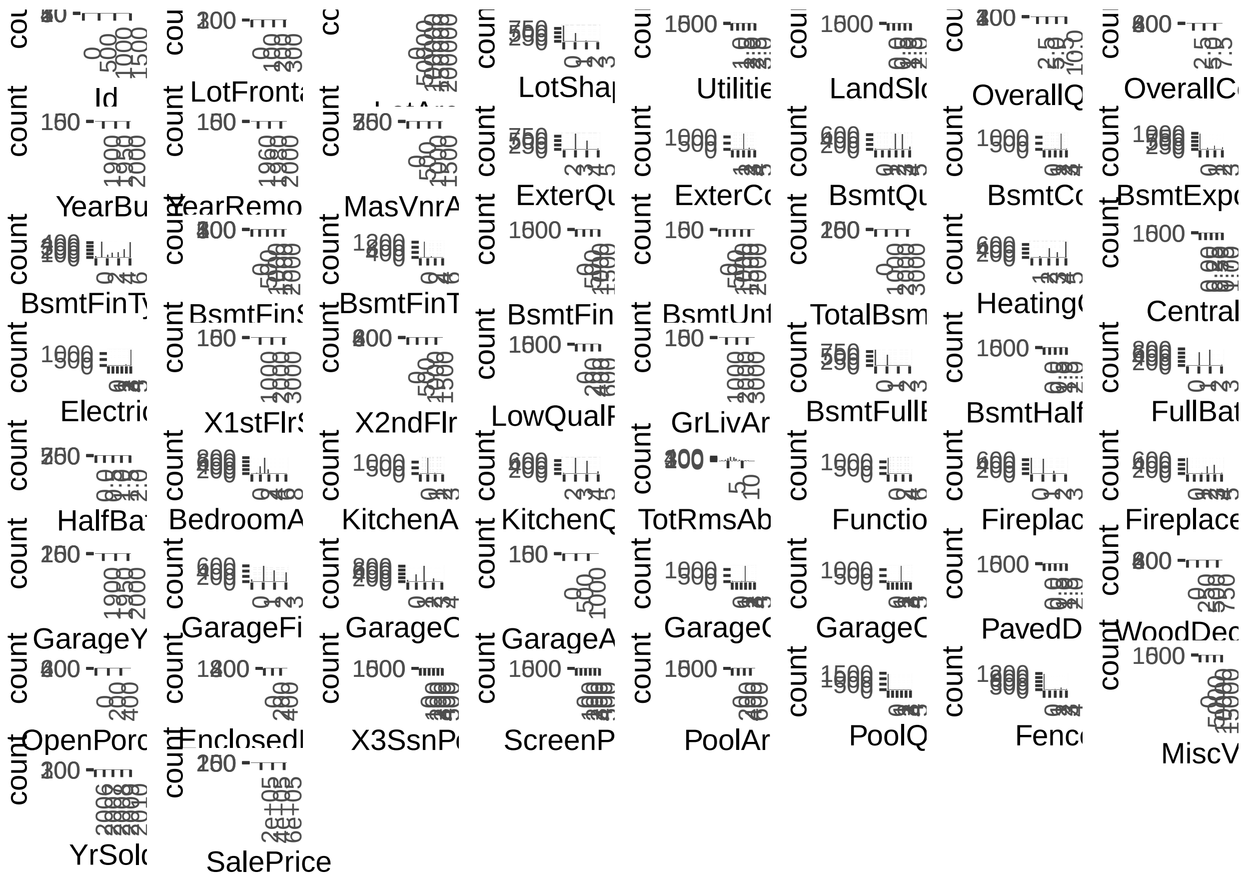
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using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



Finally, here area plots for numeric variables. There is some overlap with the integers.

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

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



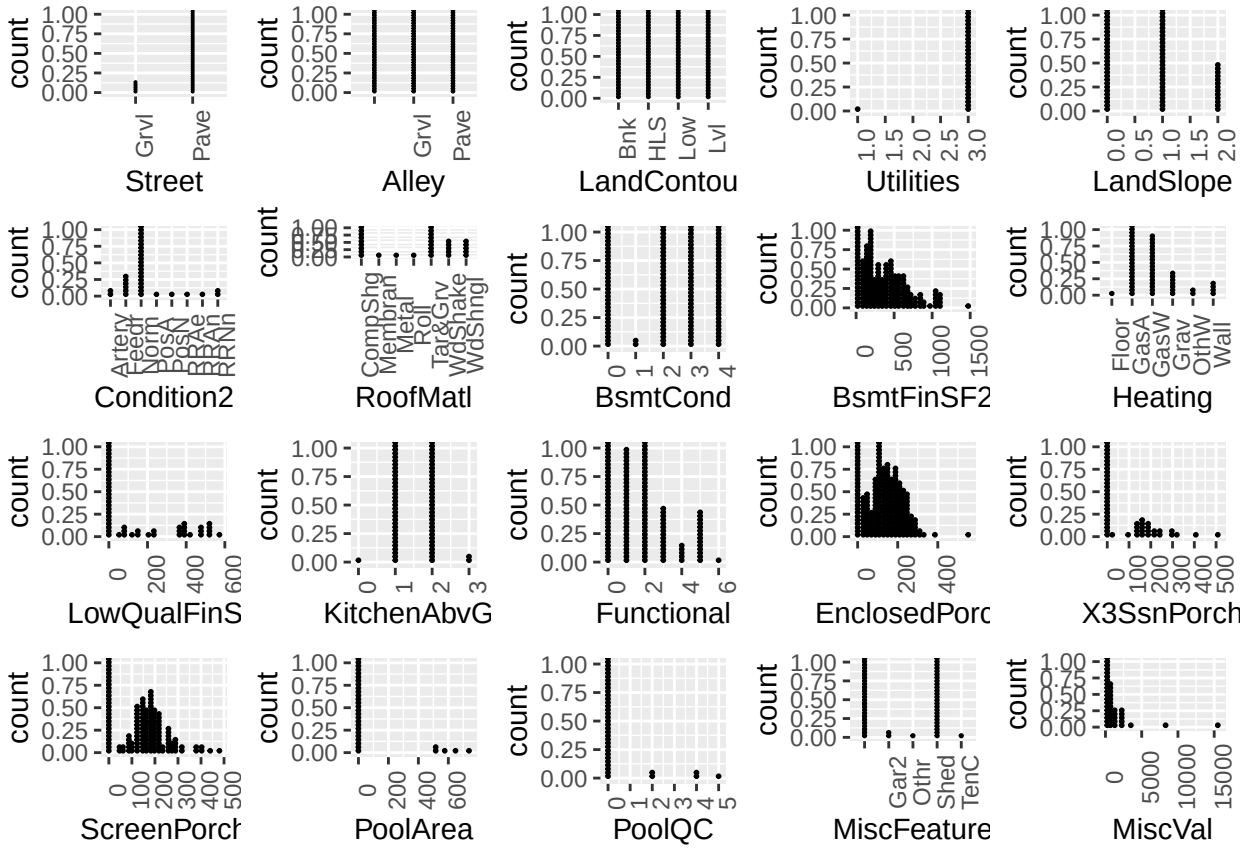
There are many variables that appear to have very low variance.

```
##  [1] "Street"        "Alley"          "LandContour"   "Utilities"
##  [5] "LandSlope"     "Condition2"     "RoofMatl"      "BsmtCond"
##  [9] "BsmtFinSF2"    "Heating"        "LowQualFinSF"  "KitchenAbvGr"
## [13] "Functional"    "EnclosedPorch" "X3SsnPorch"    "ScreenPorch"
## [17] "PoolArea"      "PoolQC"         "MiscFeature"   "MiscVal"
```

```
##  [1]  6  7  9 10 12 15 23 32 37 40 46 53 56 69 70 71 72 73 75 76
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
```



There are a number of "quality" assessment variables. It appears to be a subjective assessment made of particular features rather than just identifying the presence or size/type of a particular feature.

```
## BsmtQual ExterQual FireplaceQu GarageQual HeatingQC KitchenQual LowQualFinSF OverallQual PoolQC
```

Condition 1 and 2 and Sale condition are not quality assessments. We'll remove those from the list.

```
## BsmtCond Condition1 Condition2 ExterCond GarageCond OverallCond SaleCondition
```

```
##  chr [1:7] "BsmtCond" "Condition1" "Condition2" "ExterCond" "GarageCond" ...
```

```
## BsmtQual ExterQual FireplaceQu GarageQual HeatingQC KitchenQual LowQualFinSF OverallQual PoolQC BsmtC
```

Now we'll split the dataset into train and test sets. Sometimes character data can be problematic, we'll change it to factors first.

The 81st column is the SaleSprice

```
##   SalePrice
## 1    208500
```

```
## 2     181500
## 3     223500
## 4     140000
## 5     250000
## 6     143000
```

We'll establish the definition of RMSE for the validation step later on.

Modeling:

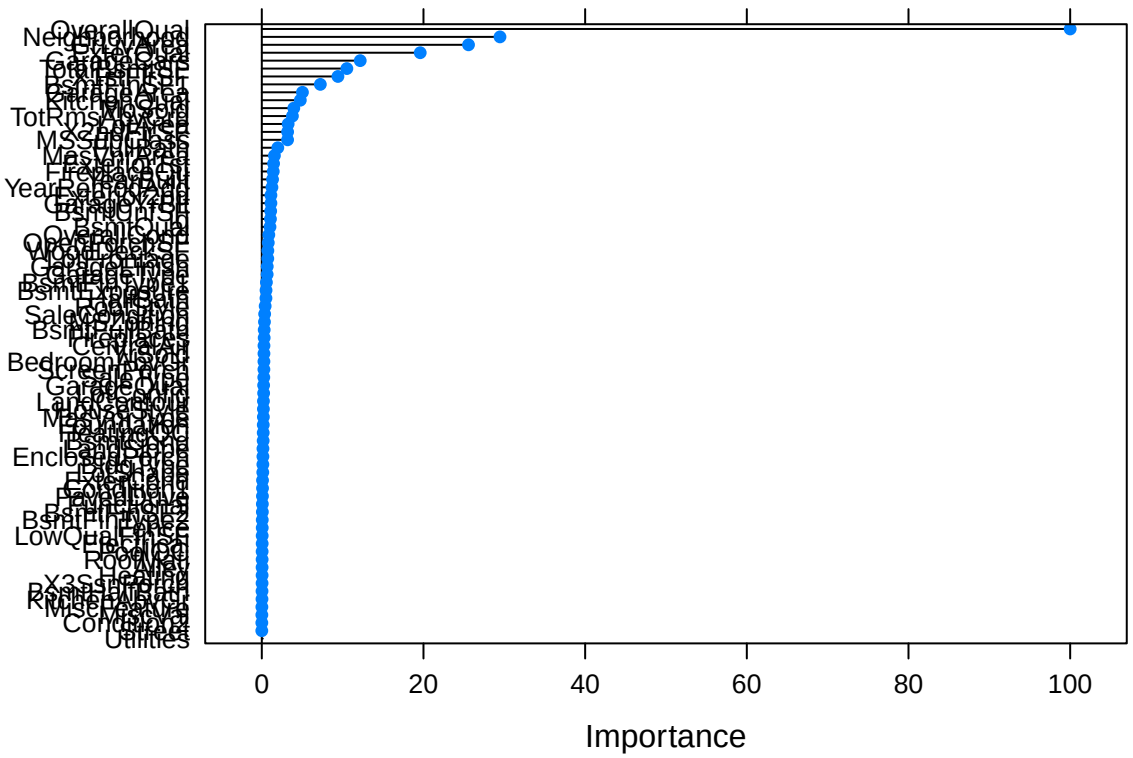
A random forest and a BART model will be used. A Bayesian Additive Regression Tree model is another ensemble of trees model that is similar to gradient boosting models. However, it weakens the effect of any given tree by its priors, attempting to decrease the risk of over-fitting that sometimes plague gradient boosting or forest models. Whereas random forest uses subsets of the data to build trees, which are combined to form predictions, BART uses a set number of small trees that weakly influence the result.

First Model: Develop a random forest model. Parameters to help control for low variance variables and highly correlated variables have been set in order to reduce computational resources for parameters with little value. For now the ntree value is kept low to reduce computational expense. Cross-validation measures are also included within the model. We then check the results and identify which variables are considered most important by the model. In the R-script more steps are described, but here the first and final RF models are summarized.

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): invalid mtry:
## reset to within valid range

## [1] 26261.72

##
## Call:
##  randomForest(x = x, y = y, ntree = 50, mtry = param$mtry)
##                Type of random forest: regression
##                      Number of trees: 50
## No. of variables tried at each split: 41
##
##           Mean of squared residuals: 780584207
##                     % Var explained: 87.04
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 78)
##
##               Overall
## OverallQual  100.000
## Neighborhood  29.465
## GrLivArea     25.573
## ExterQual     19.597
## GarageCars    12.176
## TotalBsmtSF   10.525
## X1stFlrSF      9.418
## BsmtFinSF1     7.241
## GarageArea     5.029
## KitchenQual    4.765
## MoSold         3.960
## TotRmsAbvGrd   3.787
## LotArea        3.260
## X2ndFlrSF      3.192
## MSSubClass     3.191
## FullBath       1.978
## MasVnrArea     1.561
## Exterior1st    1.460
## FireplaceQu    1.420
## YearBuilt      1.331
```

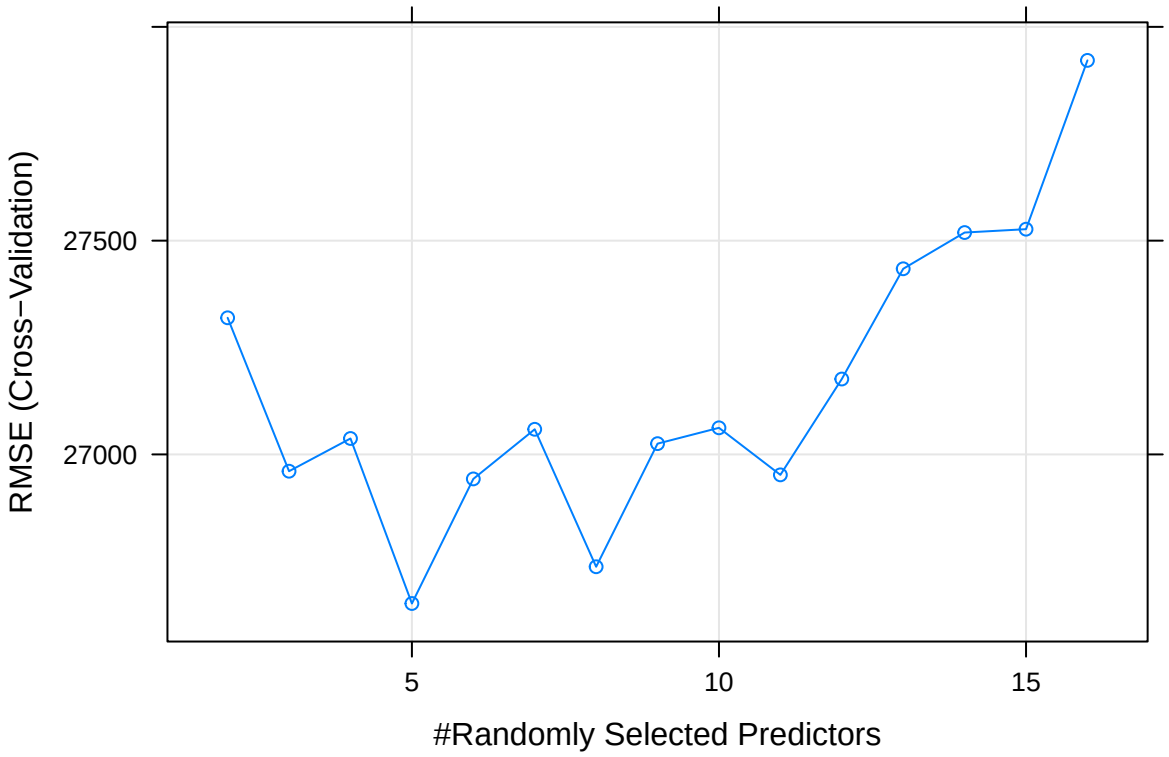


This took about 7 min to run. Let's see if we can reduce the number of variables to improve run-time without losing much predictive value. Again, in the R-script this is demonstrated more step-wise, but we will skip ahead to the final model settled on.

```
## note: only 15 unique complexity parameters in default grid. Truncating the grid to 15 .
```

```
## [1] 26651.32
```

```
##                 Overall
## OverallQual  100.0000000
## GrLivArea     54.3495626
## ExterQual     40.3230552
## GarageCars    36.9076067
## X1stFlrSF     14.8955092
## GarageArea    15.0542118
## TotalBsmtSF   19.9499457
## BsmtQual       8.8511767
## KitchenQual    8.3705674
## BsmtFinSF1    12.5633267
## LotArea        7.4548071
## YearBuilt     16.1940656
## X2ndFlrSF      3.1660596
## FireplaceQu    0.5999287
## FullBath       0.0000000
## YearRemodAdd   1.8257353
```



```
##
## Call:
##  randomForest(x = x, y = y, ntree = 150, mtry = param$mtry)
##                Type of random forest: regression
##                      Number of trees: 150
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 716559292
##                     % Var explained: 88.1
```

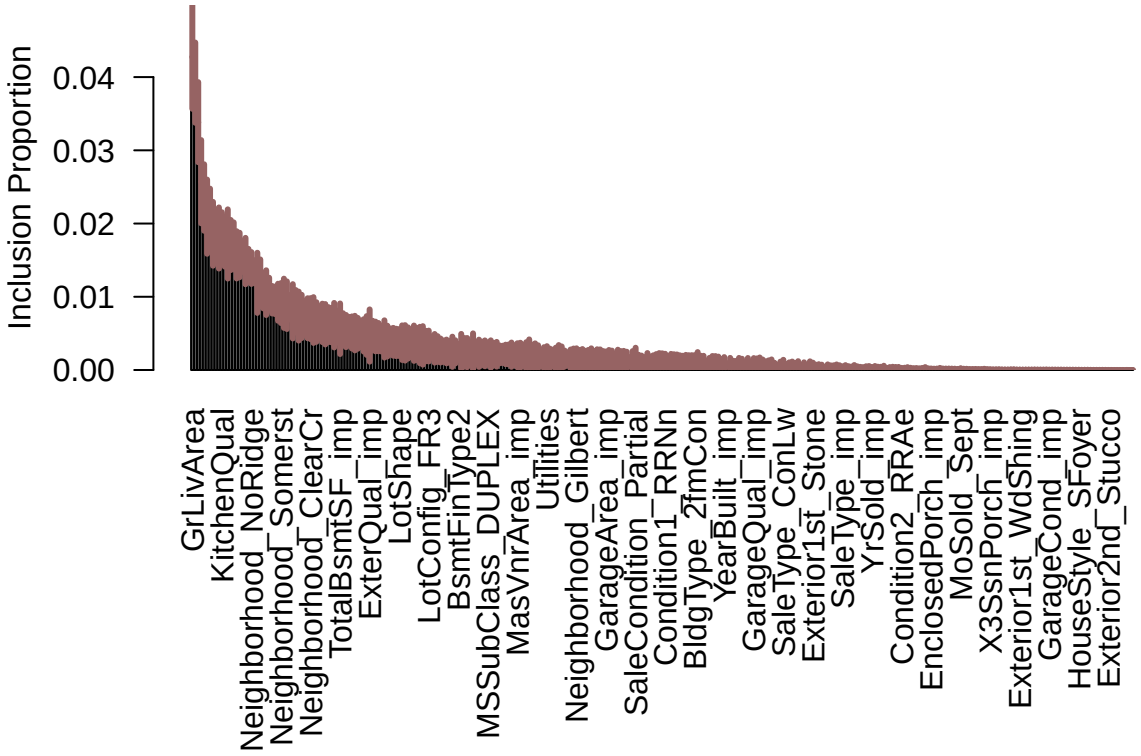
This model ran in about 3 minutes had similar predictive value to our initial model with an RMSE of around 20000. This is maybe not so useful to individuals on a budget looking for a bargain, but perhaps could serve those looking at higher cost purchases.

Bart Model- While BartMachine is in fact available in the caret package currently, there seems to be better control of it using it on its own. Controls were set in place to reduce memory usage, though it slowed down the processing. On a machine with more RAM, this would be less of an issue. We can also look at what variables/features seemed to have greater importance to the model, like with RF. Many of those important variables are similar in both models.

```
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via
## evaluating in sample data...done

## bartMachine v1.3.2 for regression
##
## training data size: n = 1164 and p = 319
## built in 38 secs on 1 core, 100 trees, 250 burn-in and 1000 post. samples
##
## sigsq est for y beforehand: 424589603.827
## avg sigsq estimate after burn-in: 76678032.9235
##
## in-sample statistics:
##  L1 = 10537309.83
##  L2 = 160626192916.25
##  rmse = 11747.13
##  Pseudo-Rsq = 0.9771
## p-val for shapiro-wilk test of normality of residuals: 0
## p-val for zero-mean noise: 0.99722

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.
```

```
## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.
```

```
## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .

## Warning in build_bart_machine(X, y, num_trees = num_trees, num_burn_in =
## bart_machine$num_burn_in, : No missing entries in the training data to impute.

## .
```




```
##    GrLivArea  OverallQual  TotalBsmtSF   BsmtFinSF1    X1stFlrSF BsmtExposure
##   0.04274154   0.03923838   0.03384204   0.02570924   0.02354956   0.02093776
##      LotArea     FullBath  OverallCond    X2ndFlrSF  KitchenQual     BsmtQual
##   0.02056962   0.01859962   0.01848274   0.01802562   0.01796546   0.01753683
## TotRmsAbvGrd   GarageCars YearRemodAdd
##   0.01721090   0.01706767   0.01695206
```

The initial performance appears better than the RF model. More trials and strategies were employed in the R script.

cross-validate, ~2.5 min

```
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...
```

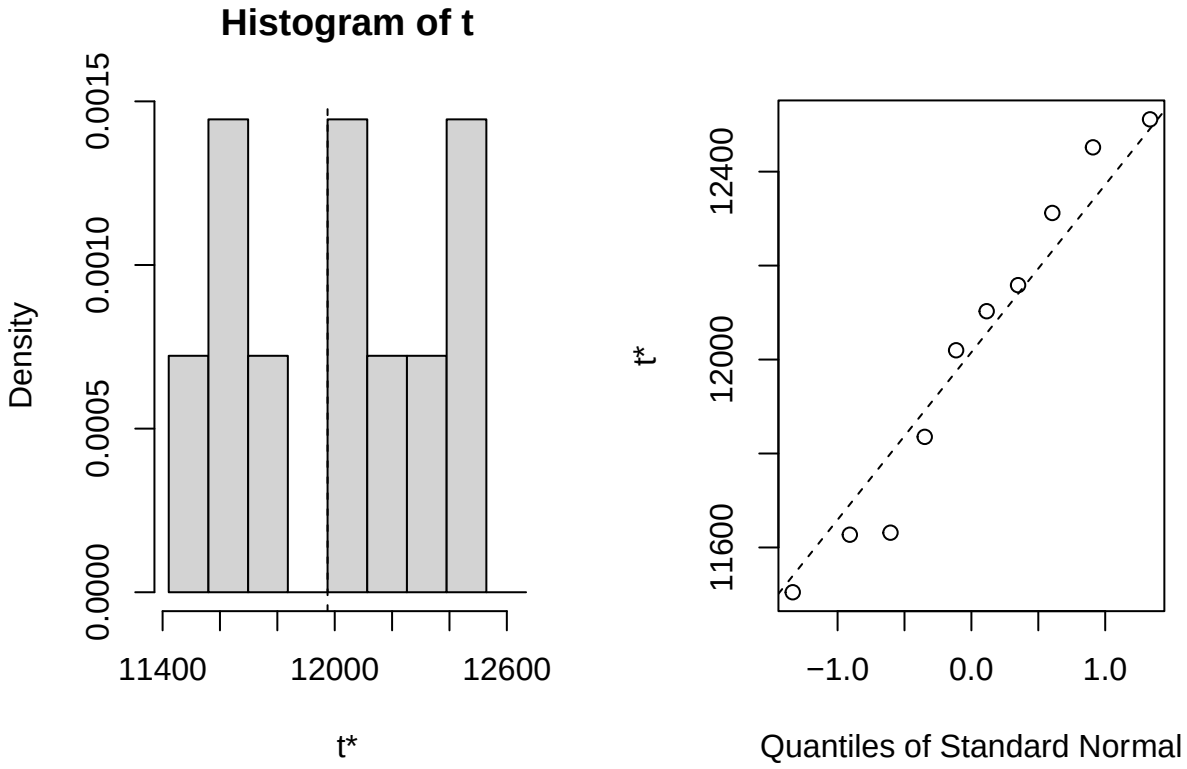
```
## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.
```

```
## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
```

```
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via r
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via r
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via r
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via r
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...
```

```
## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via r
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via r
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via r
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
```

```
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via :
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via :
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via :
## evaluating in sample data...done
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...

## Warning in (function (X = NULL, y = NULL, Xy = NULL, num_trees = 50, num_burn_in
## = 250, : No missing entries in the training data to impute.

## bartMachine after rf imputations...
## bartMachine before preprocess...
## bartMachine after preprocess... 319 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON. Missing values imputed via :
## evaluating in sample data...done
```



The results appear similar to the original model.

impute_missingness_with_rf_impute =TRUE was attempted to be used to assist with the slight differences in low frequency occurrences between the test and train sets, but predict complained of row differences, so it had to be removed.

```
## bartMachine initializing with 100 trees...
## bartMachine vars checked...
## bartMachine java init...
## bartMachine factors created...
## bartMachine before preprocess...
## bartMachine after preprocess... 239 total features...
## bartMachine sigsq estimated...
## bartMachine training data finalized...
## Now building bartMachine for regression...Covariate importance prior ON.
## evaluating in sample data...done
```

Results:

Test the models against the holdout data

RF

```
## [1] 20202.35
```

BART

```
## Warning in pre_process_new_data(new_data, bart_machine): The following features were found in records
##     Condition2_PosA, Condition2_RRNn, RoofMatl_Metal, Exterior1st_CBlock, Exterior2nd_CBlock, Heating
##   These features will be ignored during prediction.
```

```
## [1] 18912.7
```

Conclusion

The results ended up being about the same with both models, though BART seemed to perform a little better and faster than random forest, though the smaller RMSE may be due to random variablilty as well. Perhaps more of difference would be determined if more data was available. It could be because I have had more practice with random forest and caret, but it does seem more new-user friendly than BartMachine or dbarts (this package was abandoned because of errors with predict I could not solve.)

The models, as noted above, may be useful in aiding an estimation for a price with data contextualized for a particular region, but are likely not very helpful in lower cost houses.

I do plan on doing more research in the use of, and which situations they are ideal for, BART style models and interpreting the other aspects of its output. For example, L2 loss function minimizes the squared differences between the estimated and existing target values, helping you determine how much outliers are effecting your model.