House-Price-Prediction

ShengYa Mei, Binhao Chen

2022-12-15

Problem Statement

The Wisconsin housing market has been unsettling in the year 2022. The median home price sold in Wisconsin had an increase of 9.8% compared to last year and the number of homes sold was down 32.3% year over year (redfin.com). As a result, Zillow's real estate market in the Wisconsin region suffered from the impact and experienced a plunge in houses sold.

Business Application

The machine learning model constructed in this project aims to provide an accurate prediction of housing prices to be used by Zillow Real Estate in optimizing their real-estate marketplace. This model will benefit Zillow directly in their house pricing decisions as well as customers of Zillow in offering them a price that is fair and based. Zillow seeks to improve their housing sales in the upcoming year by setting prices that can accurately reflect the predicted housing market. To do this, Zillow has gathered house sales data in the year 2022 with specific details (features) on the houses sold and the sale price for each of the house sold. This data can be found in Excel file 'train.csv'. Zillow has also collected information from the houses they will be putting on their marketplace in the year 2023 without sale prices set. This data can be found in Excel file 'test.csv'. The goal for Zillow is to build a machine learning model based on the complete data with house sale prices in 'train.csv', then, use this model to predict the price for houses in found in 'test.csv'.

```
rm(list = ls()) # Clear the workspace
```

Import required libraries

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.2
```

```
## corrplot 0.92 loaded
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.2.2
```

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

```
## Warning: package 'lattice' was built under R version 4.2.2
```

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

```
## Warning: package 'survival' was built under R version 4.2.2
```

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

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

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

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

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

```
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

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

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

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

Import test and training data

```
test_dat <- read.csv('test.csv')
train_dat <- read.csv('train.csv')
```

We will extract house id from test data to be used later for submission

```
house_id <- test_dat$Id
```

Exploratory Data Analysis

Explore the imported data sets

```
# Check the dimensions
dim(train_dat)
```

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

```
dim(test_dat)
```

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

```
# Run a summary statistics
summary(train_dat)
```

```
##        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          LandSlope
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1460        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 :1973
##                     Mean   : 6.099   Mean   :5.575   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:1460        Length:1460        Length:1460
##  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:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median :   0.0   Mode  :character
##                                        Mean   : 103.7
##                                        3rd Qu.: 166.0
##                                        Max.   :1600.0
##                                        NA's   :8
##   ExterCond          Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :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          X1stFlrSF
##  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
##
##    X2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
##  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      X3SsnPorch      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           MiscVal
##  Length:1460        Length:1460        Length:1460        Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    0.00
##                                                           Mean   :   43.49
##                                                           3rd Qu.:    0.00
##                                                           Max.   :15500.00
##
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1460        Length:1460
##  1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
```

```
##  Mean   : 6.322   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
##
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##
```

```
# We will not include the ID column since we don't need if for analysis
train_dat <- train_dat[,2:81]
test_dat <- test_dat[,2:80]
```

```
# Check the dimensions again
dim(train_dat)
```

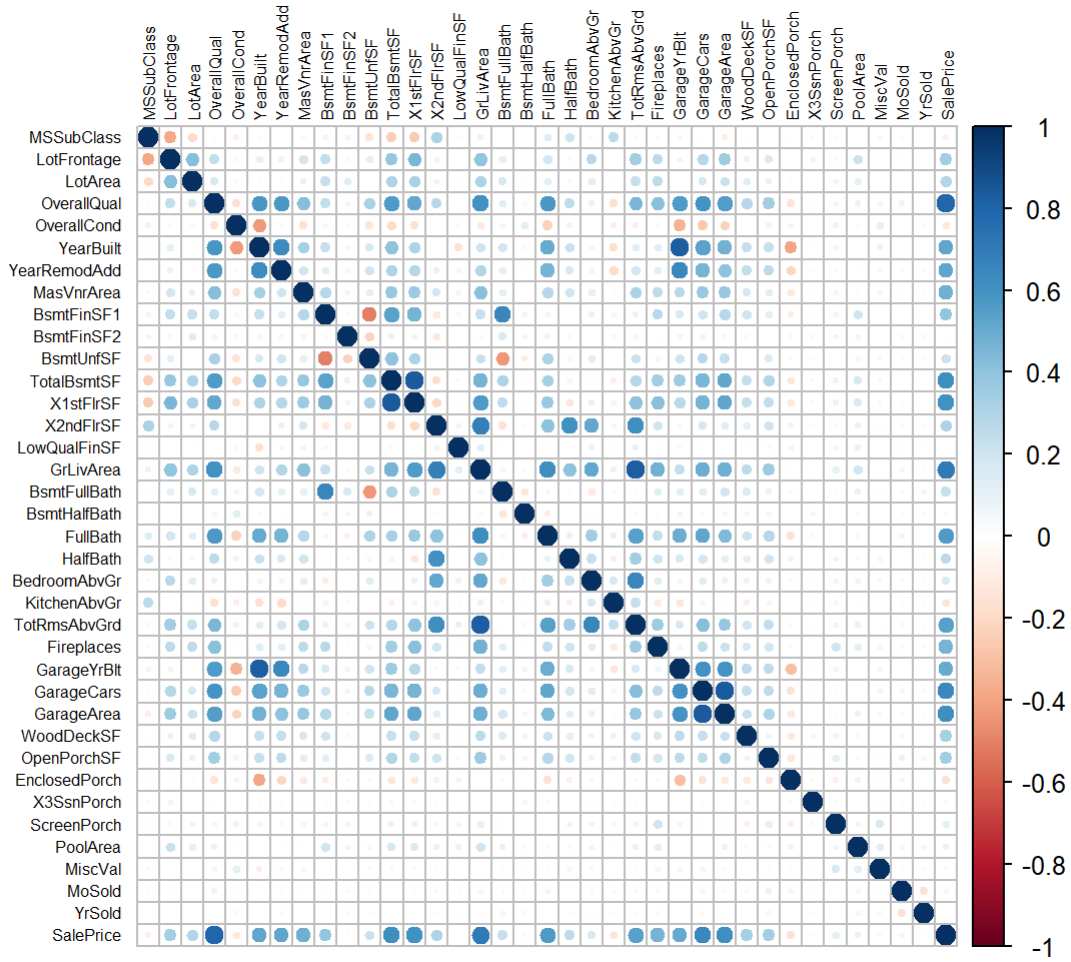
```
## [1] 1460   80
```

```
dim(test_dat)
```

```
## [1] 1459   79
```

We will generate a correlation plot to see the relationship between each of the features and our outcome of interest, ‘SalePrice’.

```
# Select only numeric columns from our train data.
# We will omit all null values for now
train_dat_numeric <- select_if(train_dat, is.numeric)
corr <- cor(na.omit(train_dat_numeric))
corrplot(corr, tl.cex=0.5, tl.col='black')
```




```
data.frame(cor(na.omit(train_dat_numeric)))
```

```
##                MSSubClass   LotFrontage      LotArea  OverallQual  OverallCond
## MSSubClass     1.000000000 -0.3869395732 -0.198095532  0.029521865 -0.087859316
## LotFrontage   -0.386939573  1.0000000000  0.421184102  0.241322316 -0.046311649
## LotArea       -0.198095532  0.4211841021  1.000000000  0.167524794 -0.034347948
## OverallQual    0.029521865  0.2413223161  0.167524794  1.000000000 -0.163156881
## OverallCond   -0.087859316 -0.0463116489 -0.034347948 -0.163156881  1.000000000
## YearBuilt      0.025799678  0.1097255707  0.029205413  0.589384529 -0.426461858
## YearRemodAdd   0.006645194  0.0864139680  0.026847846  0.570757134  0.039401850
## MasVnrArea     0.040239997  0.1899685917  0.106115431  0.423987651 -0.166762175
## BsmtFinSF1    -0.070388692  0.2413522339  0.230441380  0.249500372 -0.054787769
## BsmtFinSF2    -0.075439002  0.0493053240  0.138233605 -0.068506092  0.042313729
## BsmtUnfSF     -0.145582343  0.1153058755  0.011288124  0.322663328 -0.148629768
## TotalBsmtSF   -0.247781211  0.3876195129  0.302553906  0.563959667 -0.192761549
## X1stFlrSF     -0.252248866  0.4510850287  0.329678689  0.514452946 -0.164250781
## X2ndFlrSF      0.319327639  0.0750038007  0.074611842  0.273196930  0.005984834
## LowQualFinSF   0.024703650  0.0111480855  0.020039426 -0.008118388  0.048719667
## GrLivArea      0.083365382  0.3963060214  0.307163514  0.607466126 -0.112231246
## BsmtFullBath  -0.014681299  0.1180881480  0.179051563  0.126834025 -0.060942551
## BsmtHalfBath   0.012309569  0.0004335725 -0.014281611 -0.053283196  0.122959662
## FullBath       0.131278047  0.1857853092  0.129073367  0.576874691 -0.229848480
## HalfBath       0.203970613  0.0456783485  0.045183485  0.251690396 -0.079023306
## BedroomAbvGr  -0.032971146  0.2704038861  0.137268663  0.094881798  0.004643039
## KitchenAbvGr   0.266012356 -0.0035464716 -0.018941546 -0.178735130 -0.092643738
## TotRmsAbvGrd   0.047209430  0.3484211056  0.237917977  0.451007947 -0.096900905
## Fireplaces    -0.031122227  0.2603208280  0.255754683  0.415293730 -0.022289555
## GarageYrBlt    0.054701367  0.0698781184  0.013730760  0.560425133 -0.343206025
## GarageCars    -0.027410668  0.2865868103  0.172428230  0.593802900 -0.267858830
## GarageArea    -0.092607262  0.3568509373  0.211362399  0.550658903 -0.226346848
## WoodDeckSF    -0.017988416  0.0821656268  0.133576037  0.282512407 -0.010834875
## OpenPorchSF    0.004053970  0.1618151174  0.099170000  0.340679112 -0.076273208
## EnclosedPorch -0.017789898  0.0142610142 -0.023630663 -0.144343973  0.062747824
## X3SsnPorch    -0.039738957  0.0697157667  0.012520261  0.017331014 -0.006860988
## ScreenPorch   -0.021789344  0.0359059844  0.072517046  0.055296043  0.087029841
## PoolArea       0.003166468  0.2117461173  0.109147070  0.080131103 -0.023565580
## MiscVal       -0.040688673  0.0014707377  0.012789992 -0.062063816  0.119772147
## MoSold        -0.027170383  0.0188145348  0.008998481  0.079895095 -0.014236343
## YrSold        -0.012447829  0.0132670710 -0.006903891 -0.008902585  0.041003078
## SalePrice     -0.088031702  0.3442697721  0.299962206  0.797880680 -0.124391232
##                  YearBuilt YearRemodAdd   MasVnrArea   BsmtFinSF1   BsmtFinSF2
```

```
## MSSubClass     0.025799678  0.006645194  0.040239997 -0.070388692 -0.075439002
## LotFrontage    0.109725571  0.086413968  0.189968592  0.241352234  0.049305324
## LotArea        0.029205413  0.026847846  0.106115431  0.230441380  0.138233605
## OverallQual    0.589384529  0.570757134  0.423987651  0.249500372 -0.068506092
## OverallCond   -0.426461858  0.039401850 -0.166762175 -0.054787769  0.042313729
## YearBuilt      1.000000000  0.623171270  0.332189842  0.236940941 -0.054413993
## YearRemodAdd   0.623171270  1.000000000  0.193375602  0.120774417 -0.057024070
## MasVnrArea     0.332189842  0.193375602  1.000000000  0.285331327 -0.075260677
## BsmtFinSF1     0.236940941  0.120774417  0.285331327  1.000000000 -0.035779828
## BsmtFinSF2    -0.054413993 -0.057024070 -0.075260677 -0.035779828  1.000000000
## BsmtUnfSF      0.177545400  0.199892629  0.110067416 -0.502224788 -0.220190489
## TotalBsmtSF    0.409133562  0.308696227  0.384434076  0.530916507  0.094079397
## X1stFlrSF      0.308874836  0.281435959  0.363209260  0.468019759  0.073089633
## X2ndFlrSF     -0.011621305  0.103627388  0.180567317 -0.120822818 -0.111850256
## LowQualFinSF  -0.164358630 -0.053478689 -0.062930449 -0.050823562  0.015458749
## GrLivArea      0.204967302  0.290049515  0.414024201  0.239887620 -0.038541109
## BsmtFullBath   0.182799539  0.111896863  0.110379082  0.651726749  0.168559497
## BsmtHalfBath  -0.049644627 -0.017048664 -0.007035483  0.061963081  0.059147541
## FullBath       0.500494653  0.467562895  0.285560913  0.052313040 -0.082945334
## HalfBath       0.220000423  0.164203213  0.195272679  0.007544717 -0.031684859
## BedroomAbvGr  -0.061580195 -0.075811837  0.114310230 -0.104275130  0.008974681
## KitchenAbvGr  -0.171920229 -0.181802530 -0.023647387 -0.062919672 -0.047693223
## TotRmsAbvGrd   0.121416862  0.181995188  0.315603931  0.080206875 -0.054900179
## Fireplaces     0.133076661  0.125898069  0.252525400  0.270305580  0.022347514
## GarageYrBlt    0.823519546  0.645808468  0.277095408  0.160355947 -0.075477153
## GarageCars     0.532562838  0.462663017  0.375268818  0.196442752 -0.075477080
## GarageArea     0.471285901  0.407470742  0.382162297  0.286656921 -0.047958963
## WoodDeckSF     0.238548109  0.244602168  0.174648597  0.206245716  0.032337560
## OpenPorchSF    0.235432138  0.260521196  0.129531803  0.127900251  0.010517640
## EnclosedPorch -0.392693146 -0.214114825 -0.116832373 -0.105410284  0.047220690
## X3SsnPorch     0.027947576  0.026303516  0.022331253  0.021831074 -0.030848294
## ScreenPorch   -0.063694409 -0.034288042  0.052645658  0.059635214  0.067898778
## PoolArea       0.006716815  0.019307439  0.021647815  0.194349437  0.061211811
## MiscVal       -0.096973392 -0.040419869 -0.054044098  0.003026603  0.014290179
## MoSold         0.013784446  0.026883872  0.015850157 -0.015281479 -0.036101200
## YrSold        -0.004585485  0.041301513 -0.017569233  0.010224175  0.036395269
## SalePrice      0.525393598  0.521253270  0.488658155  0.390300523 -0.028021366
##                   BsmtUnfSF  TotalBsmtSF     X1stFlrSF    X2ndFlrSF
## MSSubClass    -1.455823e-01 -0.247781211 -0.2522488663  0.319327639
```

```
## LotFrontage    1.153059e-01  0.387619513  0.4510850287  0.075003801
## LotArea        1.128812e-02  0.302553906  0.3296786887  0.074611842
## OverallQual    3.226633e-01  0.563959667  0.5144529462  0.273196930
## OverallCond   -1.486298e-01 -0.192761549 -0.1642507808  0.005984834
## YearBuilt      1.775454e-01  0.409133562  0.3088748362 -0.011621305
## YearRemodAdd   1.998926e-01  0.308696227  0.2814359592  0.103627388
## MasVnrArea     1.100674e-01  0.384434076  0.3632092600  0.180567317
## BsmtFinSF1    -5.022248e-01  0.530916507  0.4680197587 -0.120822818
## BsmtFinSF2    -2.201905e-01  0.094079397  0.0730896330 -0.111850256
## BsmtUnfSF      1.000000e+00  0.404510415  0.3149725896 -0.010021847
## TotalBsmtSF    4.045104e-01  1.000000000  0.8359993534 -0.176721795
## X1stFlrSF      3.149726e-01  0.835999353  1.0000000000 -0.208929241
## X2ndFlrSF     -1.002185e-02 -0.176721795 -0.2089292412  1.000000000
## LowQualFinSF   3.899073e-05 -0.047901479 -0.0130255395  0.062411624
## GrLivArea      2.238598e-01  0.464644664  0.5613722585  0.688291551
## BsmtFullBath  -4.312430e-01  0.308962776  0.2571250513 -0.154151901
## BsmtHalfBath  -1.029361e-01 -0.017929115 -0.0109882091 -0.029529053
## FullBath       3.014003e-01  0.330119486  0.3745185176  0.406080546
## HalfBath      -5.711058e-02 -0.060992125 -0.1355978928  0.606336712
## BedroomAbvGr   1.325502e-01  0.027504517  0.1054405332  0.510703044
## KitchenAbvGr  -5.453792e-03 -0.088529206  0.0644553788  0.051635598
## TotRmsAbvGrd   2.165844e-01  0.283676127  0.4053140299  0.617775934
## Fireplaces     5.515445e-02  0.347729684  0.4101442139  0.199343105
## GarageYrBlt    2.089150e-01  0.352876850  0.2790531180  0.049737325
## GarageCars     2.770639e-01  0.459656896  0.4687573955  0.180136044
## GarageArea     2.353287e-01  0.522051222  0.5211829925  0.122756860
## WoodDeckSF     5.391473e-03  0.233663743  0.2376282834  0.114479784
## OpenPorchSF    1.515723e-01  0.291285868  0.2448455634  0.203460285
## EnclosedPorch -3.579056e-02 -0.130223306 -0.1135952632  0.076479404
## X3SsnPorch     2.150183e-02  0.033743492  0.0375045219 -0.027471111
## ScreenPorch   -6.398081e-03  0.080258724  0.0875796921  0.047038918
## PoolArea      -5.389385e-02  0.171488860  0.1517613301  0.094075738
## MiscVal       -3.891536e-02 -0.031075500 -0.0309091885  0.027046536
## MoSold         2.706835e-02 -0.001498092  0.0277310418  0.041485056
## YrSold        -2.673628e-02 -0.003377490  0.0004204947 -0.028009942
## SalePrice      2.131287e-01  0.615612237  0.6079691062  0.306879002
##                LowQualFinSF    GrLivArea BsmtFullBath  BsmtHalfBath
## MSSubClass     2.470365e-02  0.083365382 -0.014681299  0.0123095694
## LotFrontage    1.114809e-02  0.396306021  0.118088148  0.0004335725
```

```
## LotArea         2.003943e-02  0.307163514  0.179051563 -0.0142816112
## OverallQual    -8.118388e-03  0.607466126  0.126834025 -0.0532831964
## OverallCond     4.871967e-02 -0.112231246 -0.060942551  0.1229596616
## YearBuilt      -1.643586e-01  0.204967302  0.182799539 -0.0496446267
## YearRemodAdd   -5.347869e-02  0.290049515  0.111896863 -0.0170486639
## MasVnrArea     -6.293045e-02  0.414024201  0.110379082 -0.0070354830
## BsmtFinSF1     -5.082356e-02  0.239887620  0.651726749  0.0619630806
## BsmtFinSF2      1.545875e-02 -0.038541109  0.168559497  0.0591475414
## BsmtUnfSF       3.899073e-05  0.223859800 -0.431242987 -0.1029361323
## TotalBsmtSF    -4.790148e-02  0.464644664  0.308962776 -0.0179291150
## X1stFlrSF      -1.302554e-02  0.561372258  0.257125051 -0.0109882091
## X2ndFlrSF       6.241162e-02  0.688291551 -0.154151901 -0.0295290526
## LowQualFinSF    1.000000e+00  0.122080919 -0.019259921  0.0100396331
## GrLivArea       1.220809e-01  1.000000000  0.058979198 -0.0320322871
## BsmtFullBath   -1.925992e-02  0.058979198  1.000000000 -0.1398663536
## BsmtHalfBath    1.003963e-02 -0.032032287 -0.139866354  1.0000000000
## FullBath       -1.658038e-02  0.614887255 -0.058482039 -0.0907666267
## HalfBath       -7.842062e-03  0.407133422 -0.025007125  0.0025271343
## BedroomAbvGr    8.226367e-02  0.511939699 -0.128883935  0.0268916096
## KitchenAbvGr   -2.274915e-02  0.088959367 -0.032582312 -0.0121503307
## TotRmsAbvGrd    1.023476e-01  0.824312123 -0.019047216 -0.0403646055
## Fireplaces      2.149048e-02  0.471059867  0.134397492  0.0276569967
## GarageYrBlt    -4.632193e-02  0.243733841  0.125844218 -0.0734153115
## GarageCars     -2.338131e-02  0.494631363  0.133960898 -0.0497478397
## GarageArea      5.708720e-03  0.487549600  0.189142978 -0.0510976682
## WoodDeckSF     -1.737401e-02  0.269702612  0.157509947  0.0540663661
## OpenPorchSF     3.296761e-02  0.353534109  0.081622526 -0.0603466204
## EnclosedPorch   6.098785e-02 -0.014873875 -0.042630596  0.0008535239
## X3SsnPorch      2.170723e-03  0.004822758 -0.007893439  0.0564022782
## ScreenPorch     5.647180e-02  0.108453071  0.023857189 -0.0073231014
## PoolArea        9.908857e-02  0.198551141  0.104349485  0.0315034089
## MiscVal         1.523129e-02  0.001066968  0.027641102 -0.0109465220
## MoSold         -2.664535e-02  0.053070805 -0.030282362  0.0270365971
## YrSold         -1.625314e-02 -0.024436087  0.058467480 -0.0498512058
## SalePrice      -1.481983e-03  0.705153567  0.236737407 -0.0365126645
##                     FullBath     HalfBath BedroomAbvGr  KitchenAbvGr
## MSSubClass      1.312780e-01  0.203970613 -0.032971146  0.2660123561
## LotFrontage     1.857853e-01  0.045678349  0.270403886 -0.0035464716
## LotArea         1.290734e-01  0.045183485  0.137268663 -0.0189415460
```

```
## OverallQual    5.768747e-01  0.251690396  0.094881798 -0.1787351305
## OverallCond   -2.298485e-01 -0.079023306  0.004643039 -0.0926437383
## YearBuilt      5.004947e-01  0.220000423 -0.061580195 -0.1719202285
## YearRemodAdd   4.675629e-01  0.164203213 -0.075811837 -0.1818025296
## MasVnrArea     2.855609e-01  0.195272679  0.114310230 -0.0236473865
## BsmtFinSF1     5.231304e-02  0.007544717 -0.104275130 -0.0629196720
## BsmtFinSF2    -8.294533e-02 -0.031684859  0.008974681 -0.0476932234
## BsmtUnfSF      3.014003e-01 -0.057110576  0.132550217 -0.0054537917
## TotalBsmtSF    3.301195e-01 -0.060992125  0.027504517 -0.0885292057
## X1stFlrSF      3.745185e-01 -0.135597893  0.105440533  0.0644553788
## X2ndFlrSF      4.060805e-01  0.606336712  0.510703044  0.0516355977
## LowQualFinSF  -1.658038e-02 -0.007842062  0.082263670 -0.0227491458
## GrLivArea      6.148873e-01  0.407133422  0.511939699  0.0889593667
## BsmtFullBath  -5.848204e-02 -0.025007125 -0.128883935 -0.0325823122
## BsmtHalfBath  -9.076663e-02  0.002527134  0.026891610 -0.0121503307
## FullBath       1.000000e+00  0.105591235  0.343285855  0.1029583343
## HalfBath       1.055912e-01  1.000000000  0.229674240 -0.0980039193
## BedroomAbvGr   3.432859e-01  0.229674240  1.000000000  0.1705534204
## KitchenAbvGr   1.029583e-01 -0.098003919  0.170553420  1.0000000000
## TotRmsAbvGrd   5.404489e-01  0.343551792  0.650284589  0.2368675212
## Fireplaces     2.433556e-01  0.205242951  0.131282237 -0.1090753580
## GarageYrBlt    4.997305e-01  0.175464529 -0.052059314 -0.1355612621
## GarageCars     5.208570e-01  0.188582285  0.133568409  0.0581928750
## GarageArea     4.452408e-01  0.122798810  0.095113006  0.0277129837
## WoodDeckSF     2.150276e-01  0.114152565  0.077918216 -0.0998316572
## OpenPorchSF    2.862477e-01  0.194015990  0.079124036 -0.0601334776
## EnclosedPorch -1.645476e-01 -0.080585912  0.040681496  0.0134112511
## X3SsnPorch     3.205087e-02 -0.002422323 -0.029136458 -0.0232988631
## ScreenPorch    1.414558e-03  0.073305999  0.063660090 -0.0503077605
## PoolArea       4.960819e-02  0.042099482  0.073360852 -0.0151142862
## MiscVal       -2.939724e-02 -0.055379373  0.046522897 -0.0006403304
## MoSold         7.230451e-02 -0.007637235  0.031267773  0.0300006891
## YrSold        -1.518289e-05 -0.020875110 -0.026035312  0.0284630300
## SalePrice      5.666274e-01  0.268560303  0.166813894 -0.1404974454
##               TotRmsAbvGrd   Fireplaces  GarageYrBlt  GarageCars  GarageArea
## MSSubClass      0.04720943 -0.031122227  0.054701367 -0.02741067 -0.09260726
## LotFrontage     0.34842111  0.260320828  0.069878118  0.28658681  0.35685094
## LotArea         0.23791798  0.255754683  0.013730760  0.17242823  0.21136240
## OverallQual     0.45100795  0.415293730  0.560425133  0.59380290  0.55065890
```

```
## OverallCond    -0.09690091 -0.022289555 -0.343206025 -0.26785883 -0.22634685
## YearBuilt       0.12141686  0.133076661  0.823519546  0.53256284  0.47128590
## YearRemodAdd    0.18199519  0.125898069  0.645808468  0.46266302  0.40747074
## MasVnrArea      0.31560393  0.252525400  0.277095408  0.37526882  0.38216230
## BsmtFinSF1      0.08020688  0.270305580  0.160355947  0.19644275  0.28665692
## BsmtFinSF2     -0.05490018  0.022347514 -0.075477153 -0.07547708 -0.04795896
## BsmtUnfSF       0.21658444  0.055154452  0.208915005  0.27706388  0.23532868
## TotalBsmtSF     0.28367613  0.347729684  0.352876850  0.45965690  0.52205122
## X1stFlrSF       0.40531403  0.410144214  0.279053118  0.46875740  0.52118299
## X2ndFlrSF       0.61777593  0.199343105  0.049737325  0.18013604  0.12275686
## LowQualFinSF    0.10234764  0.021490479 -0.046321925 -0.02338131  0.00570872
## GrLivArea       0.82431212  0.471059867  0.243733841  0.49463136  0.48754960
## BsmtFullBath   -0.01904722  0.134397492  0.125844218  0.13396090  0.18914298
## BsmtHalfBath   -0.04036461  0.027656997 -0.073415312 -0.04974784 -0.05109767
## FullBath        0.54044893  0.243355640  0.499730486  0.52085700  0.44524076
## HalfBath        0.34355179  0.205242951  0.175464529  0.18858228  0.12279881
## BedroomAbvGr    0.65028459  0.131282237 -0.052059314  0.13356841  0.09511301
## KitchenAbvGr    0.23686752 -0.109075358 -0.135561262  0.05819288  0.02771298
## TotRmsAbvGrd    1.00000000  0.352047792  0.167206751  0.42396283  0.38192956
## Fireplaces      0.35204779  1.000000000  0.064578978  0.25268507  0.21655099
## GarageYrBlt     0.16720675  0.064578978  1.000000000  0.60090342  0.59263525
## GarageCars      0.42396283  0.252685069  0.600903418  1.00000000  0.83941492
## GarageArea      0.38192956  0.216550993  0.592635246  0.83941492  1.00000000
## WoodDeckSF      0.19052652  0.177762561  0.255915956  0.23427620  0.22395499
## OpenPorchSF     0.24671363  0.185274131  0.257141006  0.25813718  0.30255823
## EnclosedPorch  -0.03165122 -0.034478393 -0.308277701 -0.15188609 -0.11574897
## X3SsnPorch     -0.02390409 -0.001001989  0.019842482  0.02014079  0.01530622
## ScreenPorch     0.07089430  0.192128653 -0.067595752  0.02513541  0.02644616
## PoolArea        0.09338651  0.117107761 -0.009295071  0.01282888  0.08087138
## MiscVal         0.02649503  0.054826316 -0.053295449 -0.06959231 -0.03699294
## MoSold          0.04309712  0.048788120  0.009232878  0.05748115  0.03759656
## YrSold         -0.02481218 -0.031402273  0.009596052 -0.03350744 -0.01620605
## SalePrice       0.54706736  0.461872689  0.504753018  0.64703361  0.61932962
##                 WoodDeckSF OpenPorchSF EnclosedPorch   X3SsnPorch  ScreenPorch
## MSSubClass    -0.017988416  0.00405397 -0.0177898980 -0.039738957 -0.021789344
## LotFrontage    0.082165627  0.16181512  0.0142610142  0.069715767  0.035905984
## LotArea        0.133576037  0.09917000 -0.0236306632  0.012520261  0.072517046
## OverallQual    0.282512407  0.34067911 -0.1443439735  0.017331014  0.055296043
## OverallCond   -0.010834875 -0.07627321  0.0627478237 -0.006860988  0.087029841
```

```
## YearBuilt      0.238548109  0.23543214 -0.3926931459  0.027947576 -0.063694409
## YearRemodAdd   0.244602168  0.26052120 -0.2141148247  0.026303516 -0.034288042
## MasVnrArea     0.174648597  0.12953180 -0.1168323725  0.022331253  0.052645658
## BsmtFinSF1     0.206245716  0.12790025 -0.1054102836  0.021831074  0.059635214
## BsmtFinSF2     0.032337560  0.01051764  0.0472206899 -0.030848294  0.067898778
## BsmtUnfSF      0.005391473  0.15157230 -0.0357905581  0.021501827 -0.006398081
## TotalBsmtSF    0.233663743  0.29128587 -0.1302233060  0.033743492  0.080258724
## X1stFlrSF      0.237628283  0.24484556 -0.1135952632  0.037504522  0.087579692
## X2ndFlrSF      0.114479784  0.20346028  0.0764794041 -0.027471111  0.047038918
## LowQualFinSF  -0.017374011  0.03296761  0.0609878453  0.002170723  0.056471796
## GrLivArea      0.269702612  0.35353411 -0.0148738747  0.004822758  0.108453071
## BsmtFullBath   0.157509947  0.08162253 -0.0426305960 -0.007893439  0.023857189
## BsmtHalfBath   0.054066366 -0.06034662  0.0008535239  0.056402278 -0.007323101
## FullBath       0.215027625  0.28624773 -0.1645476442  0.032050870  0.001414558
## HalfBath       0.114152565  0.19401599 -0.0805859117 -0.002422323  0.073305999
## BedroomAbvGr   0.077918216  0.07912404  0.0406814964 -0.029136458  0.063660090
## KitchenAbvGr  -0.099831657 -0.06013348  0.0134112511 -0.023298863 -0.050307761
## TotRmsAbvGrd   0.190526524  0.24671363 -0.0316512181 -0.023904094  0.070894298
## Fireplaces     0.177762561  0.18527413 -0.0344783935 -0.001001989  0.192128653
## GarageYrBlt    0.255915956  0.25714101 -0.3082777012  0.019842482 -0.067595752
## GarageCars     0.234276205  0.25813718 -0.1518860875  0.020140787  0.025135406
## GarageArea     0.223954993  0.30255823 -0.1157489718  0.015306215  0.026446162
## WoodDeckSF     1.000000000  0.07552504 -0.1210606440 -0.053825448 -0.087574843
## OpenPorchSF    0.075525042  1.00000000 -0.1305655132 -0.010350664  0.112442842
## EnclosedPorch -0.121060644 -0.13056551  1.0000000000 -0.034375570 -0.081550145
## X3SsnPorch    -0.053825448 -0.01035066 -0.0343755695  1.000000000 -0.031359297
## ScreenPorch   -0.087574843  0.11244284 -0.0815501446 -0.031359297  1.000000000
## PoolArea       0.033075524  0.03378559  0.0763415650 -0.008214593  0.067356042
## MiscVal       -0.007100697  0.02884250  0.0287954966  0.024613679  0.169856874
## MoSold         0.041547155  0.08976692 -0.0610833426  0.022260093  0.012859261
## YrSold         0.014809970 -0.05303516 -0.0011845771  0.020730677 -0.004118063
## SalePrice      0.336855121  0.34335381 -0.1548432035  0.030776594  0.110426815
##                   PoolArea       MiscVal       MoSold        YrSold
## MSSubClass     0.003166468 -0.0406886729 -0.027170383 -1.244783e-02
## LotFrontage    0.211746117  0.0014707377  0.018814535  1.326707e-02
## LotArea        0.109147070  0.0127899922  0.008998481 -6.903891e-03
## OverallQual    0.080131103 -0.0620638160  0.079895095 -8.902585e-03
## OverallCond   -0.023565580  0.1197721471 -0.014236343  4.100308e-02
## YearBuilt      0.006716815 -0.0969733924  0.013784446 -4.585485e-03
```

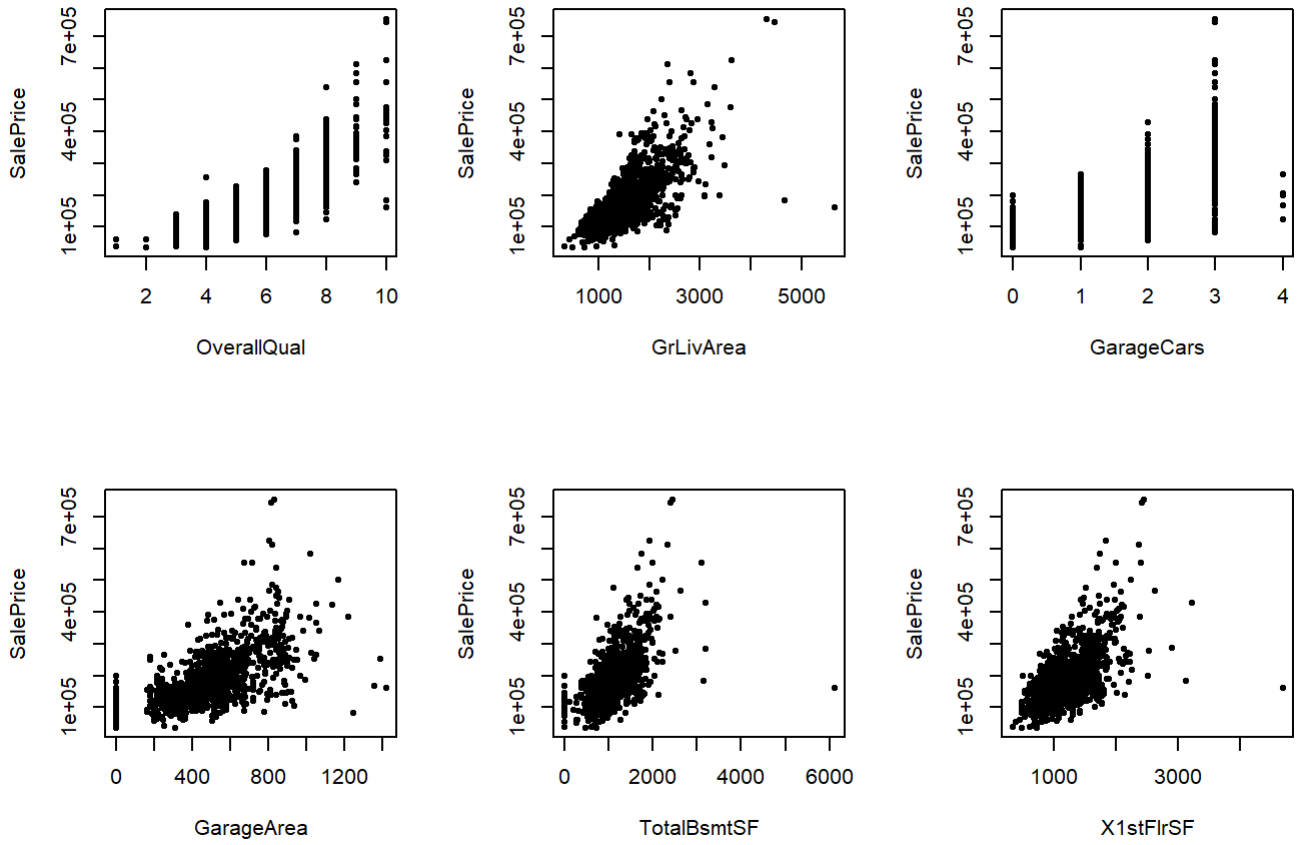
```
## YearRemodAdd   0.019307439 -0.0404198692  0.026883872  4.130151e-02
## MasVnrArea     0.021647815 -0.0540440976  0.015850157 -1.756923e-02
## BsmtFinSF1     0.194349437  0.0030266034 -0.015281479  1.022418e-02
## BsmtFinSF2     0.061211811  0.0142901786 -0.036101200  3.639527e-02
## BsmtUnfSF     -0.053893848 -0.0389153564  0.027068355 -2.673628e-02
## TotalBsmtSF    0.171488860 -0.0310755000 -0.001498092 -3.377490e-03
## X1stFlrSF      0.151761330 -0.0309091885  0.027731042  4.204947e-04
## X2ndFlrSF      0.094075738  0.0270465356  0.041485056 -2.800994e-02
## LowQualFinSF   0.099088571  0.0152312894 -0.026645350 -1.625314e-02
## GrLivArea      0.198551141  0.0010669684  0.053070805 -2.443609e-02
## BsmtFullBath   0.104349485  0.0276411019 -0.030282362  5.846748e-02
## BsmtHalfBath   0.031503409 -0.0109465220  0.027036597 -4.985121e-02
## FullBath       0.049608194 -0.0293972414  0.072304513 -1.518289e-05
## HalfBath       0.042099482 -0.0553793731 -0.007637235 -2.087511e-02
## BedroomAbvGr   0.073360852  0.0465228967  0.031267773 -2.603531e-02
## KitchenAbvGr  -0.015114286 -0.0006403304  0.030000689  2.846303e-02
## TotRmsAbvGrd   0.093386510  0.0264950274  0.043097116 -2.481218e-02
## Fireplaces     0.117107761  0.0548263163  0.048788120 -3.140227e-02
## GarageYrBlt   -0.009295071 -0.0532954488  0.009232878  9.596052e-03
## GarageCars     0.012828877 -0.0695923062  0.057481155 -3.350744e-02
## GarageArea     0.080871376 -0.0369929442  0.037596558 -1.620605e-02
## WoodDeckSF     0.033075524 -0.0071006971  0.041547155  1.480997e-02
## OpenPorchSF    0.033785590  0.0288425004  0.089766922 -5.303516e-02
## EnclosedPorch  0.076341565  0.0287954966 -0.061083343 -1.184577e-03
## X3SsnPorch    -0.008214593  0.0246136793  0.022260093  2.073068e-02
## ScreenPorch    0.067356042  0.1698568742  0.012859261 -4.118063e-03
## PoolArea       1.000000000  0.1286840123 -0.054872361 -5.388769e-02
## MiscVal        0.128684012  1.0000000000  0.020067062  3.410608e-02
## MoSold        -0.054872361  0.0200670615  1.000000000 -1.505766e-01
## YrSold        -0.053887689  0.0341060793 -0.150576612  1.000000e+00
## SalePrice      0.092488120 -0.0360412372  0.051568064 -1.186882e-02
##                  SalePrice
## MSSubClass    -0.088031702
## LotFrontage    0.344269772
## LotArea        0.299962206
## OverallQual    0.797880680
## OverallCond   -0.124391232
## YearBuilt      0.525393598
## YearRemodAdd   0.521253270
```

```
## MasVnrArea     0.488658155
## BsmtFinSF1     0.390300523
## BsmtFinSF2    -0.028021366
## BsmtUnfSF      0.213128680
## TotalBsmtSF    0.615612237
## X1stFlrSF      0.607969106
## X2ndFlrSF      0.306879002
## LowQualFinSF  -0.001481983
## GrLivArea      0.705153567
## BsmtFullBath   0.236737407
## BsmtHalfBath  -0.036512665
## FullBath       0.566627442
## HalfBath       0.268560303
## BedroomAbvGr   0.166813894
## KitchenAbvGr  -0.140497445
## TotRmsAbvGrd   0.547067360
## Fireplaces     0.461872689
## GarageYrBlt    0.504753018
## GarageCars     0.647033611
## GarageArea     0.619329622
## WoodDeckSF     0.336855121
## OpenPorchSF    0.343353812
## EnclosedPorch -0.154843204
## X3SsnPorch     0.030776594
## ScreenPorch    0.110426815
## PoolArea       0.092488120
## MiscVal       -0.036041237
## MoSold         0.051568064
## YrSold        -0.011868823
## SalePrice      1.000000000
```

We see that features ‘OverallQual’, ‘GrLivArea’, ‘GarageCars’, ‘GarageArea’, ‘TotalBsmtSF’ and ‘X1stFlrSF’ have a positive correlation of over 0.6 with ‘SalePrice’. We will generate scatterplot to visual their relationships.

```
par(mfrow=c(2,3))
attach(train_dat)

plot(OverallQual, SalePrice, pch=20)
plot(GrLivArea, SalePrice, pch=20)
plot(GarageCars, SalePrice, pch=20)
plot(GarageArea, SalePrice, pch=20)
plot(TotalBsmtSF, SalePrice, pch=20)
plot(X1stFlrSF, SalePrice, pch=20)
```




```
detach(train_dat)
```

We see some prominent outliers in 'GrLivArea', 'TotalBsmtSF' and 'X1stFlrSF'. We will go ahead and remove them from our train data to avoid skewed results and under-performing models.

```
# We find that the one outlier point in both 'TotalBsmtSF' and 'X1stFlrSF' plot are
# from the same record and this record is the same as one of the two outliers in 'GrLivArea'
# plot. We will go ahead and remove them
train_dat[train_dat$TotalBsmtSF > 5000 & train_dat$X1stFlrSF > 4000, ]
```

```
##      MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1299         60       RL         313   63887   Pave  <NA>      IR3         Bnk
##      Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1299    AllPub    Corner       Gtl      Edwards      Feedr       Norm     1Fam
##      HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle
## 1299     2Story          10           5      2008         2008       Hip
##      RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond
## 1299  ClyTile      Stucco      Stucco      Stone        796        Ex        TA
##      Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1
## 1299      PConc       Ex       TA           Gd          GLQ       5644
##      BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir
## 1299          Unf          0       466        6110    GasA        Ex          Y
##      Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath
## 1299      SBrkr      4692       950            0      5642            2
##      BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual
## 1299            0        2        1            3           1          Ex
##      TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt
## 1299           12        Typ          3          Gd     Attchd        2008
##      GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive
## 1299          Fin          2       1418         TA         TA          Y
##      WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea
## 1299        214         292             0          0           0      480
##      PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition
## 1299     Gd  <NA>        <NA>       0      1   2008      New       Partial
##      SalePrice
## 1299    160000
```

```
train_dat[train_dat$GrLivArea > 4000 & train_dat$SalePrice < 300000, ]
```

```
##      MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 524          60       RL         130   40094   Pave  <NA>      IR1         Bnk
## 1299         60       RL         313   63887   Pave  <NA>      IR3         Bnk
##      Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 524     AllPub    Inside       Gtl      Edwards       PosN       PosN     1Fam
## 1299    AllPub    Corner       Gtl      Edwards      Feedr       Norm     1Fam
##      HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle
## 524      2Story          10           5      2007         2008       Hip
## 1299     2Story          10           5      2008         2008       Hip
##      RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond
## 524   CompShg     CemntBd     CmentBd      Stone        762        Ex        TA
## 1299  ClyTile      Stucco      Stucco      Stone        796        Ex        TA
##      Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1
## 524       PConc       Ex       TA           Gd          GLQ       2260
## 1299      PConc       Ex       TA           Gd          GLQ       5644
##      BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir
## 524           Unf          0       878        3138    GasA        Ex          Y
## 1299          Unf          0       466        6110    GasA        Ex          Y
##      Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath
## 524       SBrkr      3138      1538            0      4676            1
## 1299      SBrkr      4692       950            0      5642            2
##      BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual
## 524             0        3        1            3            1          Ex
## 1299            0        2        1            3            1          Ex
##      TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt
## 524            11        Typ          1          Gd    BuiltIn        2007
## 1299           12        Typ          3          Gd     Attchd        2008
##      GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive
## 524           Fin          3        884         TA         TA          Y
## 1299          Fin          2       1418         TA         TA          Y
##      WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea
## 524         208         406             0          0           0        0
## 1299        214         292             0          0           0      480
##      PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition
## 524    <NA>  <NA>        <NA>       0     10   2007      New       Partial
## 1299     Gd  <NA>        <NA>       0      1   2008      New       Partial
##      SalePrice
```

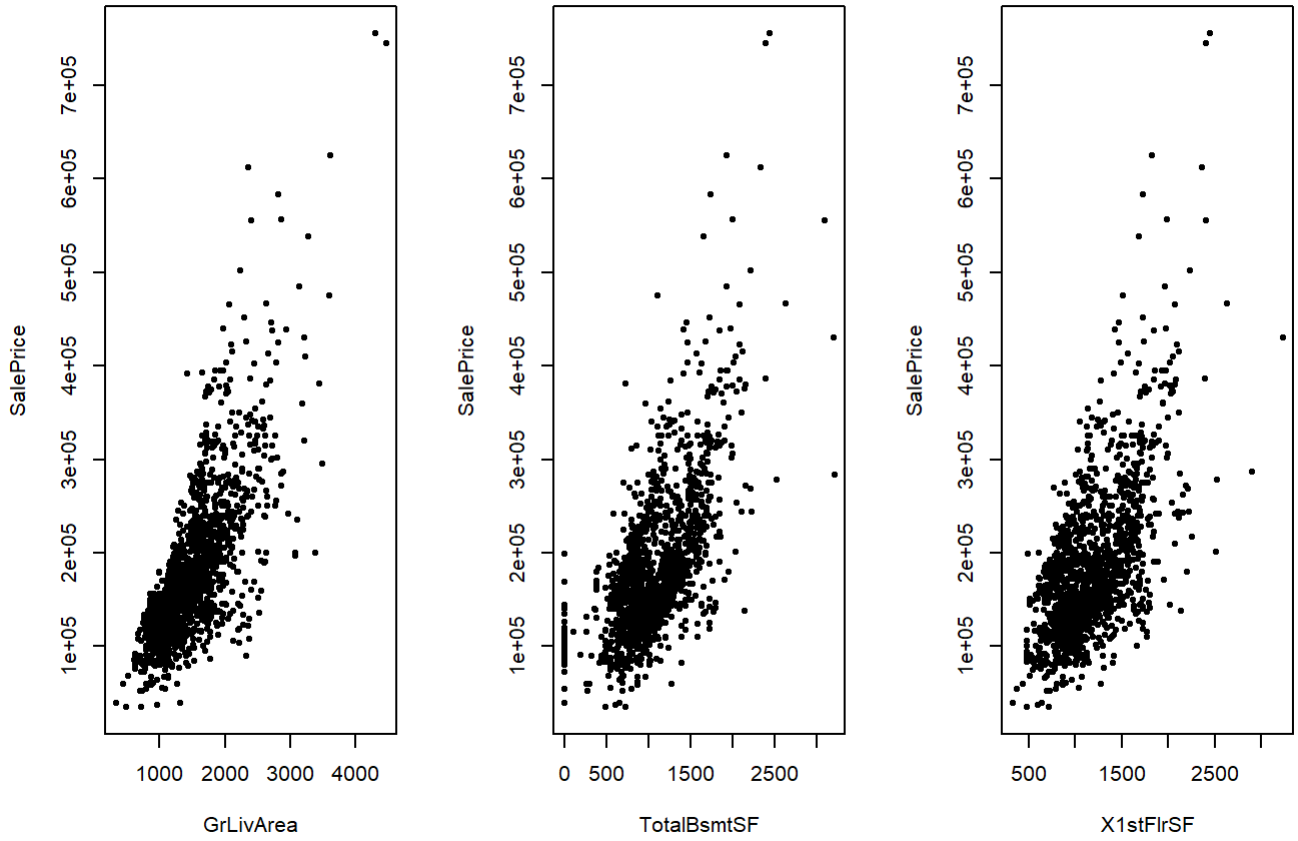
```
## 524     184750
## 1299    160000
```

```
# Remove outliers
train_dat <- train_dat[-c(524,1299),]
```

We will plot scatterplot again for these 3 features to check if outliers are removed

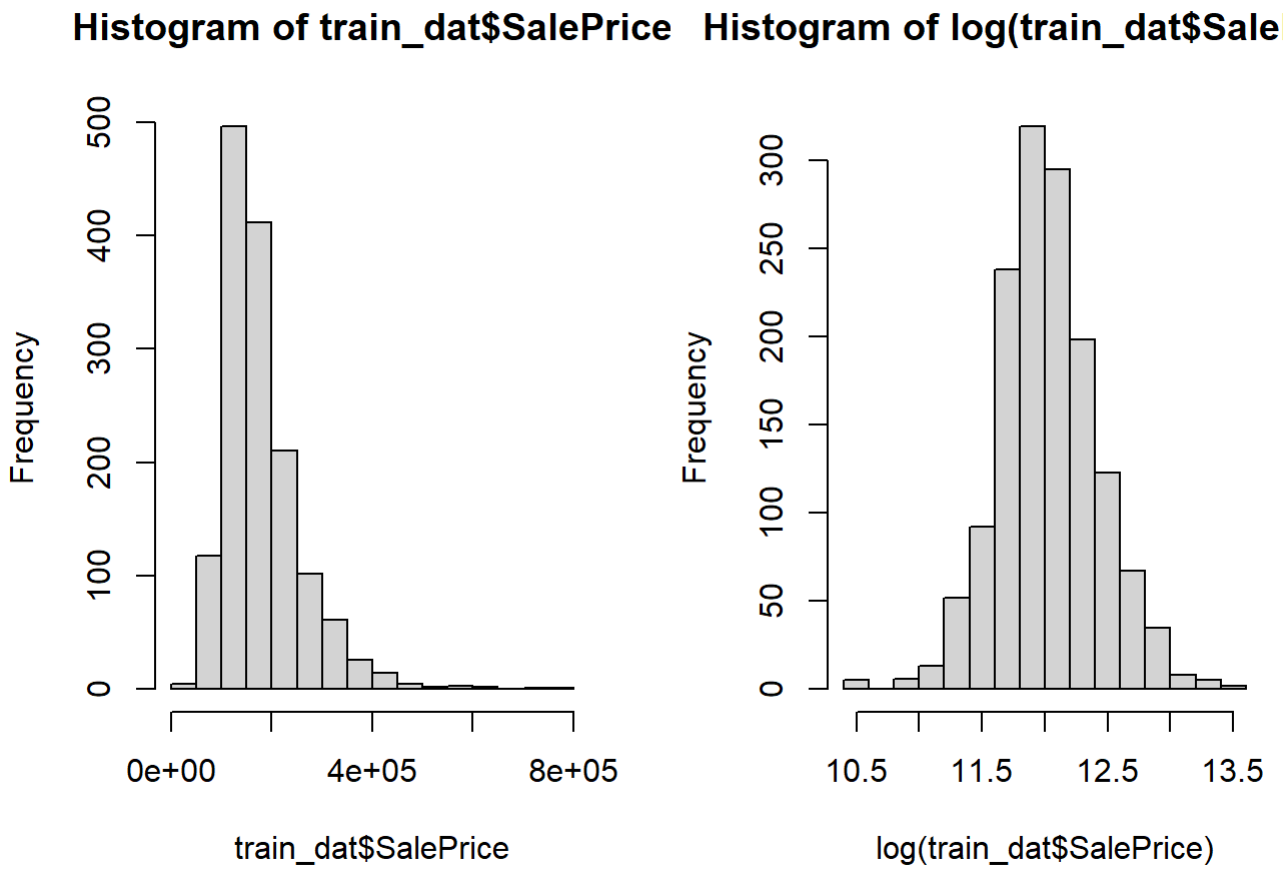
```
par(mfrow=c(1,3))
attach(train_dat)

# Outliers are removed
plot(GrLivArea, SalePrice, pch=20)
plot(TotalBsmtSF, SalePrice, pch=20)
plot(X1stFlrSF, SalePrice, pch=20)
```




```
detach(train_dat)
```

```
par(mfrow=c(1,2))
hist(train_dat$SalePrice)
hist(log(train_dat$SalePrice))
```



Data Cleaning

Check how many missing values we have in total and in each parameter in train data

```
paste('Total Missing Value:', sum(is.na(train_dat)))
```

```
## [1] "Total Missing Value: 6958"
```

```
var_na <- colnames(train_dat)
df_na <- data.frame(var_na, sapply(train_dat, function(x) sum(is.na(x))))
colnames(df_na) <- c('parameter', 'na_count')
df_na <- filter(df_na, df_na[,2]>0)
df_na <- df_na[order(-df_na$na_count),]
df_na
```

```
##                parameter na_count
## PoolQC             PoolQC     1452
## MiscFeature   MiscFeature     1404
## Alley              Alley     1367
## Fence              Fence     1177
## FireplaceQu   FireplaceQu      690
## LotFrontage   LotFrontage      259
## GarageType     GarageType       81
## GarageYrBlt   GarageYrBlt       81
## GarageFinish GarageFinish       81
## GarageQual     GarageQual       81
## GarageCond     GarageCond       81
## BsmtExposure BsmtExposure       38
## BsmtFinType2 BsmtFinType2       38
## BsmtQual         BsmtQual       37
## BsmtCond         BsmtCond       37
## BsmtFinType1 BsmtFinType1       37
## MasVnrType     MasVnrType        8
## MasVnrArea     MasVnrArea        8
## Electrical     Electrical        1
```

Most of the NA values are explained in data set description where NA means there are none of the feature present at that household. For these NA values, we will replace them with ‘none’. For missing categorical features, we will replace the NA value with the mode.

We will create a mode function to deal with some of our missing categorical values.

```
# create a mode function
mode <- function(x) {
   uniqv <- unique(x)
   uniqv[which.max(tabulate(match(x, uniqv)))]
}
```

```
train_dat <- train_dat %>%
  mutate(PoolQC = ifelse(is.na(PoolQC),'None',PoolQC),
         MiscFeature = ifelse(is.na(MiscFeature),'None',MiscFeature),
         Alley = ifelse(is.na(Alley),'None',Alley),
         Fence = ifelse(is.na(Fence),'None',Fence),
         FireplaceQu = ifelse(is.na(FireplaceQu),'None',FireplaceQu),
         GarageType = ifelse(is.na(GarageType),'None',GarageType),
         GarageYrBlt = ifelse(is.na(GarageYrBlt),0,GarageYrBlt),
         GarageFinish = ifelse(is.na(GarageFinish),'None',GarageFinish),
         GarageQual = ifelse(is.na(GarageQual),'None',GarageQual),
         GarageCond = ifelse(is.na(GarageCond),'None',GarageCond),
         BsmtExposure = ifelse(is.na(BsmtExposure),'None',BsmtExposure),
         BsmtFinType2 = ifelse(is.na(BsmtFinType2),'None',BsmtFinType2),
         BsmtQual = ifelse(is.na(BsmtQual),'None',BsmtQual),
         BsmtCond = ifelse(is.na(BsmtCond),'None',BsmtCond),
         BsmtFinType1 = ifelse(is.na(BsmtFinType1),'None',BsmtFinType1),
         MasVnrType = ifelse(is.na(MasVnrType),'None',MasVnrType),
         MasVnrArea = ifelse(is.na(MasVnrArea),0,MasVnrArea),
         Electrical = ifelse(is.na(Electrical),mode(train_dat$Electrical),Electrical)
         )
```

For feature ‘LotFrontage’, we see that it takes numerical values and that there is no specified values for NA. In this case, we will find the median of ‘LotFrontage’ after grouping by feature ‘Neighborhood’ then we will use this median to fill out missing ‘LotFrontage’ based on the ‘Neighborhood’ they are in.

```
# First we will create a temporary dataframe that removes all null vales
# in the 'LotFrontage' column in train set
LotFrontage_subset <- train_dat[,c('LotFrontage')]
temp <- train_dat[complete.cases(LotFrontage_subset),]
```

```
# We then group 'LotFrontage' by 'Neighborhood' to find the median in each 'Neighborhood'
LotFrontage_median <- temp %>%
   group_by(Neighborhood)%>%
     summarise_each(funs(median), LotFrontage)
```

```
## Warning: `summarise_each_()` was deprecated in dplyr 0.7.0.
## ℹ Please use `across()` instead.
## ℹ The deprecated feature was likely used in the dplyr package.
##   Please report the issue at <]8;;https://github.com/tidyverse/dplyr/issueshttps://github.com/tidyverse/dplyr/issues]
8;;>.
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## ℹ Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
LotFrontage_median
```

```
## # A tibble: 25 × 2
##    Neighborhood LotFrontage
##    <chr>              <dbl>
##  1 Blmngtn             43
##  2 Blueste             24
##  3 BrDale              21
##  4 BrkSide             52
##  5 ClearCr             80
##  6 CollgCr             70
##  7 Crawfor             74
##  8 Edwards             64.5
##  9 Gilbert             65
## 10 IDOTRR              60
## # … with 15 more rows
```

```
# Lastly, we will perform a left join on missing 'LotFrontage' values in the cleaned train data with the median we found
train_dat <- left_join(train_dat, LotFrontage_median, by = 'Neighborhood') %>%
  mutate(LotFrontage = ifelse(is.na(LotFrontage.x), LotFrontage.y, LotFrontage.x)) %>%
  select(-LotFrontage.y, -LotFrontage.x) # remove duplicate columns
```

We will perform the same cleaning method for our test data set.

```
paste('Total Missing Value:', sum(is.na(test_dat)))
```

```
## [1] "Total Missing Value: 7000"
```

```
var_na <- colnames(test_dat)
df_na <- data.frame(var_na, sapply(test_dat, function(x) sum(is.na(x))))
colnames(df_na) <- c('parameter', 'na_count')
df_na <- filter(df_na, df_na[,2]>0)
df_na <- df_na[order(-df_na$na_count),]
df_na
```

```
##                parameter na_count
## PoolQC            PoolQC     1456
## MiscFeature   MiscFeature     1408
## Alley              Alley     1352
## Fence              Fence     1169
## FireplaceQu   FireplaceQu      730
## LotFrontage   LotFrontage      227
## GarageYrBlt   GarageYrBlt       78
## GarageFinish GarageFinish       78
## GarageQual     GarageQual       78
## GarageCond     GarageCond       78
## GarageType     GarageType       76
## BsmtCond         BsmtCond       45
## BsmtQual         BsmtQual       44
## BsmtExposure BsmtExposure       44
## BsmtFinType1 BsmtFinType1       42
## BsmtFinType2 BsmtFinType2       42
## MasVnrType     MasVnrType       16
## MasVnrArea     MasVnrArea       15
## MSZoning         MSZoning        4
## Utilities       Utilities        2
## BsmtFullBath BsmtFullBath        2
## BsmtHalfBath BsmtHalfBath        2
## Functional     Functional        2
## Exterior1st   Exterior1st        1
## Exterior2nd   Exterior2nd        1
## BsmtFinSF1     BsmtFinSF1        1
## BsmtFinSF2     BsmtFinSF2        1
## BsmtUnfSF       BsmtUnfSF        1
## TotalBsmtSF   TotalBsmtSF        1
## KitchenQual   KitchenQual        1
## GarageCars     GarageCars        1
## GarageArea     GarageArea        1
## SaleType         SaleType        1
```

```
test_dat <- test_dat %>%
  mutate(PoolQC = ifelse(is.na(PoolQC),'None',PoolQC),
         MiscFeature = ifelse(is.na(MiscFeature),'None',MiscFeature),
         Alley = ifelse(is.na(Alley),'None',Alley),
         Fence = ifelse(is.na(Fence),'None',Fence),
         FireplaceQu = ifelse(is.na(FireplaceQu),'None',FireplaceQu),
         GarageYrBlt = ifelse(is.na(GarageYrBlt),0,GarageYrBlt),
         GarageFinish = ifelse(is.na(GarageFinish),'None',GarageFinish),
         GarageQual = ifelse(is.na(GarageQual),'None',GarageQual),
         GarageCond = ifelse(is.na(GarageCond),'None',GarageCond),
         GarageType = ifelse(is.na(GarageType),'None',GarageType),
         BsmtExposure = ifelse(is.na(BsmtExposure),'None',BsmtExposure),
         BsmtFinType2 = ifelse(is.na(BsmtFinType2),'None',BsmtFinType2),
         BsmtQual = ifelse(is.na(BsmtQual),'None',BsmtQual),
         BsmtCond = ifelse(is.na(BsmtCond),'None',BsmtCond),
         BsmtFinType1 = ifelse(is.na(BsmtFinType1),'None',BsmtFinType1),
         MasVnrType = ifelse(is.na(MasVnrType),'None',MasVnrType),
         MasVnrArea = ifelse(is.na(MasVnrArea),0,MasVnrArea),
         MSZoning = ifelse(is.na(MSZoning),mode(test_dat$MSZoning),MSZoning),
         Utilities = ifelse(is.na(Utilities),'AllPub',Utilities), # We will assume the two missing values for utilities
are 'AllPub' since there are no other variations
         BsmtFullBath = ifelse(is.na(BsmtFullBath),0,BsmtFullBath),
         BsmtHalfBath = ifelse(is.na(BsmtHalfBath),0,BsmtHalfBath),
         Functional = ifelse(is.na(Functional),mode(test_dat$Functional),Functional),
         Exterior1st = ifelse(is.na(Exterior1st),mode(test_dat$Exterior1st),Exterior1st),
         Exterior2nd = ifelse(is.na(Exterior2nd),mode(test_dat$Exterior2nd),Exterior2nd),
         BsmtFinSF1 = ifelse(is.na(BsmtFinSF1),0,BsmtFinSF1),
         BsmtFinSF2 = ifelse(is.na(BsmtFinSF2),0,BsmtFinSF2),
         BsmtUnfSF = ifelse(is.na(BsmtUnfSF),0,BsmtUnfSF),
         TotalBsmtSF = ifelse(is.na(TotalBsmtSF),0,TotalBsmtSF),
         KitchenQual = ifelse(is.na(KitchenQual),mode(test_dat$KitchenQual),KitchenQual),
         GarageCars = ifelse(is.na(GarageCars),0,GarageCars),
         GarageArea = ifelse(is.na(GarageArea),0,GarageArea),
         SaleType = ifelse(is.na(SaleType),mode(test_dat$SaleType),SaleType)
         )
```

```
# First we will create a temporary dataframe that removes all null vales in the 'LotFrontage' column in test set
LotFrontage_subset <- test_dat[,c('LotFrontage')]
temp <- test_dat[complete.cases(LotFrontage_subset),]
```

```
# We then group 'LotFrontage' by 'Neighborhood' to find the median in each 'Neighborhood'
LotFrontage_median <- temp %>%
   group_by(Neighborhood)%>%
     summarise_each(funs(median), LotFrontage)
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## ℹ Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
LotFrontage_median
```

```
## # A tibble: 25 × 2
##    Neighborhood LotFrontage
##    <chr>              <dbl>
##  1 Blmngtn             43
##  2 Blueste             24
##  3 BrDale              21
##  4 BrkSide             51
##  5 ClearCr             87
##  6 CollgCr             70
##  7 Crawfor             66
##  8 Edwards             64.5
##  9 Gilbert             63
## 10 IDOTRR              60
## # … with 15 more rows
```

```
# Lastly, we will perform a left join on missing 'LotFrontage' values in the cleaned test data with the median we found
test_dat <- left_join(test_dat, LotFrontage_median, by = 'Neighborhood') %>%
  mutate(LotFrontage = ifelse(is.na(LotFrontage.x), LotFrontage.y, LotFrontage.x)) %>%
  select(-LotFrontage.y, -LotFrontage.x) # remove duplicate columns
```

We will make sure the total number of missing values in both train and test data is now zero

```
paste('Total Missing Value in train data:', sum(is.na(train_dat)))
```

```
## [1] "Total Missing Value in train data: 0"
```

```
paste('Total Missing Value in test data:', sum(is.na(test_dat)))
```

```
## [1] "Total Missing Value in test data: 0"
```

Some of the numerial variables needs to be in categorical type. We will transform those

```
str(train_dat)
```

```
## 'data.frame':    1458 obs. of  80 variables:
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : chr  "RL" "RL" "RL" "RL" ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  "None" "None" "None" "None" ...
##  $ 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   : num  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" ...
##  $ 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  : 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  "None" "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : num  2003 1976 2001 1998 2000 ...
##  $ 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 ...
##  $ 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       : chr  "None" "None" "None" "None" ...
##  $ Fence        : chr  "None" "None" "None" "None" ...
##  $ MiscFeature  : chr  "None" "None" "None" "None" ...
##  $ 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 ...
##  $ LotFrontage  : num  65 80 68 60 84 85 75 80 51 50 ...
```

We will store our train outcome variable 'SalePrice' separately

```
Y.trn <- train_dat[, 79]
```

We will stack train and test together and transform them together

```
full_dat <- rbind(train_dat[, c(1:78, 80)], test_dat)
```

Lable Encoding

Before we feed in our train data into ML models, we first need to transform our categorical variables into numerical attributes which can be processed by the models. We perform lable encoding using as.factor

```
# Transform to categorical features
full_dat$MSSubClass <- as.factor(full_dat$MSSubClass)
full_dat$OverallQual<- as.factor(full_dat$OverallQual)
full_dat$OverallCond<- as.factor(full_dat$OverallCond)
```

We will split them back to train and test data

```
train_dat <- full_dat[1:1458,]
test_dat <- full_dat[1459:2917,]
```

Add the outcome variable 'SalePrice' back to train data

```
train_dat$SalePrice <- Y.trn
```

Model training

```
# We will create data matrix for our train and test data to be used later
X.tst <- data.matrix(test_dat[, 1:79])
X.trn <- data.matrix(train_dat[, 1:79])
```

Install and run required libraries

```
# install.packages("caret", dependencies = TRUE)
# install.packages("randomForest")
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.2
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.2.2
```

```
## 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
```

Random Forest Algorithm

```
# Set a random seed
set.seed(42)
# Training using 'random forest' algorithm
rf_model <- train(SalePrice ~., data = train_dat, method = 'rf', trControl = trainControl(method = 'cv', number = 5))
# Use 5 folds for cross-validation
rf_model
```

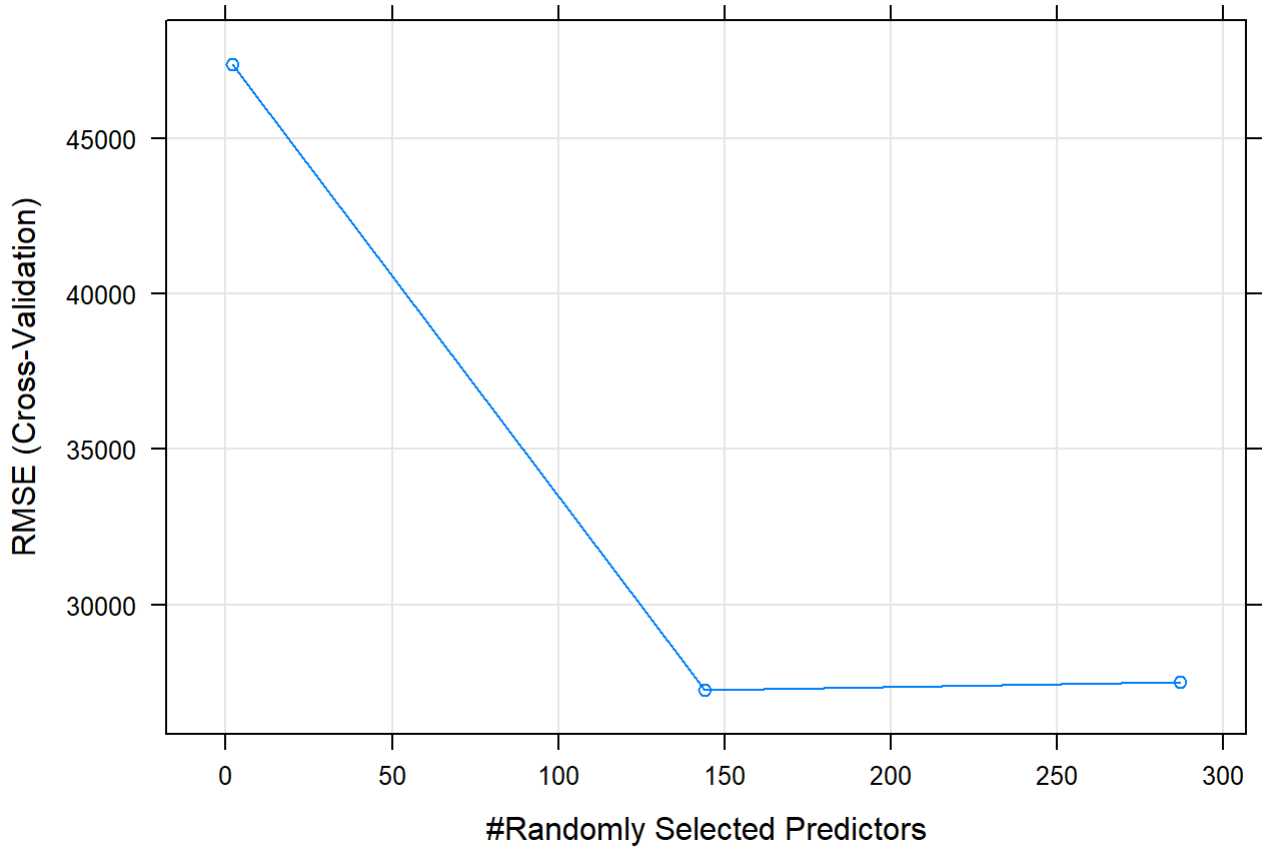
```
## Random Forest
##
## 1458 samples
##   79 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 1167, 1167, 1166, 1166, 1166
## Resampling results across tuning parameters:
##
##   mtry  RMSE      Rsquared   MAE
##     2   47403.93  0.7921125  30176.90
##   144   27235.38  0.8915446  17109.41
##   287   27497.41  0.8872381  17391.55
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 144.
```

The best random forest model generated has a r^2 of 89.15% and RMSE of 27235.38 and MAE of 17109.41. We will now predict the sale price for our test data using this rf model

```
test_dat$SalePrice_rf <- predict(rf_model, newdata = test_dat)
# predict test data using rf_model
```

If we plot our rf_model, we can see the point where the machine chose to be the best number of predictors with the least RMSE. This shows our bias and variance trade-off. We want a model that is as simple as possible and as complex as necessary.

```
plot(rf_model)
```



Now, we will see if we can improve our model further with lower RMSE by running a Gradient Boost

```
# We will now fit a boosted tree learner to the data
#GRADIENT BOOSTING
library(xgboost)
```

```
## Warning: package 'xgboost' was built under R version 4.2.2
```

```
##
## Attaching package: 'xgboost'
```

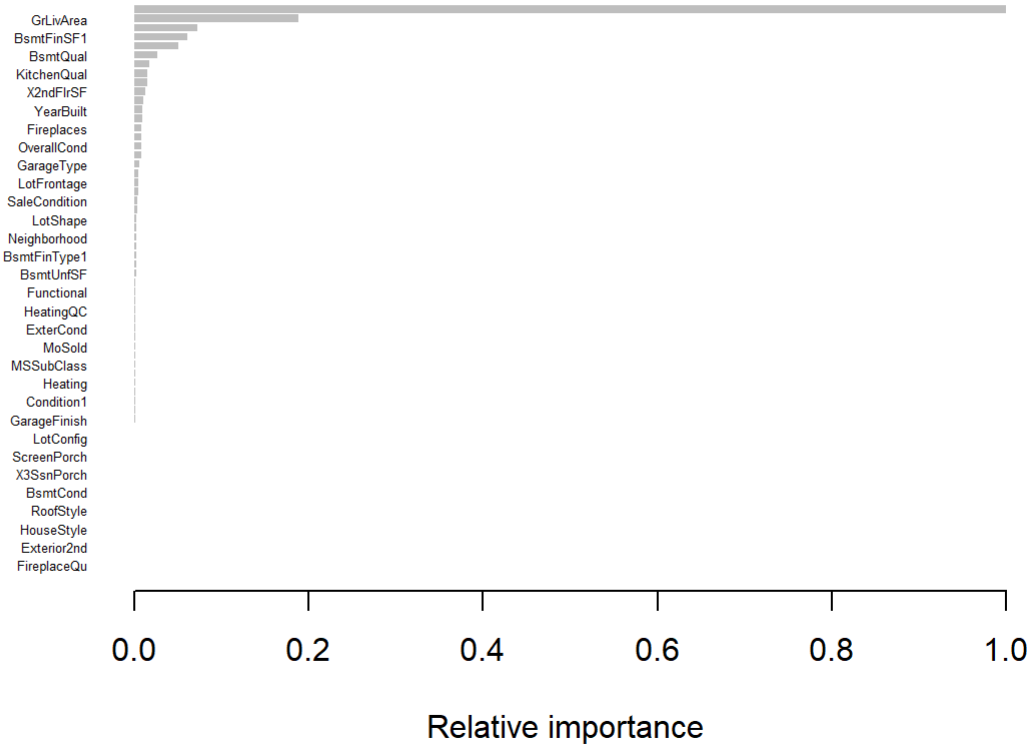
```
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
parm <- list(nthread=2, max_depth=2, eta=0.10)
# xgboost takes in data matrix and not dataframe so we will use the data matrices we created earlier
bt_model <- xgboost(parm, data=X.trn, label=Y.trn, verbose=2, nrounds=10)
```

```
## [1]  train-rmse:141325.400711
## [2]  train-rmse:101755.866546
## [3]  train-rmse:73971.145460
## [4]  train-rmse:54384.955619
## [5]  train-rmse:40560.586077
## [6]  train-rmse:31056.475242
## [7]  train-rmse:24280.090232
## [8]  train-rmse:19738.704612
## [9]  train-rmse:16599.367110
## [10] train-rmse:14384.717832
```

```
# we can evaluate the outcomes and particularly the variable importance: We can then plot the importance.
imp <- xgb.importance(feature_names=colnames(X.trn), model=bt_model)

xgb.plot.importance(imp, rel_to_first = TRUE, xlab = "Relative importance")
```



We did get a significantly lower RMSE of 14384.717832 after running through 10 iterations. We will now use this boosted tree model to predict our test data

```
test_dat$SalePrice_bt <- predict(bt_model, newdata = X.tst)
```

We will now create our submission CSV file

```
submission <- data.frame(cbind(house_id, test_dat$SalePrice_bt))
colnames(submission) <- c('Id', 'SalePrice')
write.csv(submission, "E:\\MSBA\\Machine Learning Course\\Final Project House Prices\\house_price_submission.csv", row.names
=FALSE)
```