HarvardX Data Science Capstone: House Prices Report

Justin Nielson

May 30, 2019

Table of Contents

1. Introduction

House Prices: Advanced Regression Techniques is an ongoing Kaggle competition using advanced regression techniques to predict sales prices and practice feature engineering, RFs, and gradient boosting. The link to the competition is here.
https://www.kaggle.com/c/house-prices-advanced-regression-techniques

I thought this would be a good choose your own project in that it builds on the linear regression methods that were used in the MovieLens movie recommendation system project. The competition also allows me to get feedback from the Kaggle community on my methods, analysis, and results.

The submissions are evaluated on Root Mean Squared Logarithmic Error (RMSLE) between the logarithm of the predicted value and the logarithm of the observed sales price. The objective is for each Id in the train set, the model must predict the value of the SalePrice variable and minimize the RMSLE. The submission file is them the predicted SalesPrice using your best model on the test set which has unknown values for SalePrice.

The ML model used with the smallest RMSLE was the *Lasso regression model* on the train set with a RMSLE of 0.0978573.

2. Overview

This report contains sections for data exploration, visualization, preprocessing, evaluated machine learning algorithms, and RMSLE analysis sections including methods that were used to transform the data to create the best predicive model.

The results and conclusion sections and the end includes final thoughts on the House Prices project.

2.1. Loading libraries and data

```
# Loading packages for data exploration, visualization, preprocessing,
# machine learning algorithms, and RMSLE analysis

library(tidyverse)
library(caTools)
library(caret)
library(e1071)
library(glmnet)
library(randomForest)
library(xgboost)
library(data.table)
library(lubridate)
library(ggplot2)
library(corrplot)
library(knitr)
library(kableExtra)

# Read House Prices train dataset:
train <- read.csv("train.csv")

# Fill NA with 0 for modeling purposes
Train <- train %>% mutate_all(~replace(., is.na(.), 0))

# Read House Prices test dataset:
test <- read.csv("test.csv")

# Fill NA with 0 for modeling purposes
test <- test %>% mutate_all(~replace(., is.na(.), 0))

#Converting character variables to numeric

train$paved[train$Street == "Pave"] <- 1
train$paved[train$Street != "Pave"] <- 0

train$regshape[train$LotShape == "Reg"] <- 1
train$regshape[train$LotShape != "Reg"] <- 0
```

3. Executive Summary

The House Prices Kaggle data includes four files train.csv, test.csv, sample_submission.csv and data_description.txt. The train dataset includes 1,459 objects or rows of 81 variables were the test dataset includes 1,460 objects or rows of 80 variables. The last variable in the train dataset is the actual SalesPrice that we are trying to predict in the test dataset. The samp_submission file format is the Id of the house and predicted SalesPrice. The data_description.txt file includes descriptions of all the variables in the train and test datasets.

House Prices train dataset

```
summary(train)
```

```
##        Id           MSSubClass       MSZoning     LotFrontage
##  Min.   :   1.0   Min.   : 20.0   C (all):  10   Min.   :  0.00
##  1st Qu.: 365.8   1st Qu.: 20.0   FV     :  65   1st Qu.: 42.00
##  Median : 730.5   Median : 50.0   RH     :  16   Median : 63.00
##  Mean   : 730.5   Mean   : 56.9   RL     :1151   Mean   : 57.62
##  3rd Qu.:1095.2   3rd Qu.: 70.0   RM     : 218   3rd Qu.: 79.00
##  Max.   :1460.0   Max.   :190.0                   Max.   :313.00
##
##     LotArea        Street      Alley      LotShape  LandContour
##  Min.   :  1300   Grvl:   6   0   :1369   IR1:484   Bnk:  63
##  1st Qu.:  7554   Pave:1454   Grvl:  50   IR2: 41   HLS:  50
##  Median :  9478               Pave:  41   IR3: 10   Low:  36
##  Mean   : 10517                           Reg:925   Lvl:1311
##  3rd Qu.: 11602
##  Max.   :215245
##
##   Utilities      LotConfig    LandSlope   Neighborhood   Condition1
##  AllPub:1459   Corner : 263   Gtl:1382   0mes   :225   Norm   :1260
##  NoSeWa:   1   CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81
##                FR2    :  47   Sev:  13   OldTown:113   Artery :  48
##                FR3    :   4              Edwards:100   RRAn   :  26
##                Inside :1052              Somerst: 86   PosN   :  19
##                                          Gilbert: 79   RRAe   :  11
##                                          (Other):707   (Other):  15
##    Condition2     BldgType      HouseStyle   OverallQual
##  Norm   :1445   1Fam  :1220   1Story :726   Min.   : 1.000
##  Feedr  :   6   2fmCon:  31   2Story :445   1st Qu.: 5.000
##  Artery :   2   Duplex:  52   1.5Fin :154   Median : 6.000
##  PosN   :   2   Twnhs :  43   SLvl   : 65   Mean   : 6.099
##  RRNn   :   2   TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.000
##  PosA   :   1                 1.5Unf : 14   Max.   :10.000
##  (Other):   2                 (Other): 19
##   OverallCond      YearBuilt     YearRemodAdd    RoofStyle
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  13
##  1st Qu.:5.000   1st Qu.:1954   1st Qu.:1967   Gable  :1141
##  Median :5.000   Median :1973   Median :1994   Gambrel:  11
```

```
##  Mean   :5.575   Mean   :1971   Mean   :1985   Hip    : 286
##  3rd Qu.:6.000   3rd Qu.:2000   3rd Qu.:2004   Mansard:   7
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :   2
##
##     RoofMatl     Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
##  CompShg:1434   VinylSd:515   VinylSd:504   0      :  8   Min.   :   0.0
##  Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkCmn : 15   1st Qu.:   0.0
##  WdShngl:   6   MetalSd:220   HdBoard:207   BrkFace:445   Median :   0.0
##  WdShake:   5   Wd Sdng:206   Wd Sdng:197   None   :864   Mean   : 103.1
##  ClyTile:   1   Plywood:108   Plywood:142   Stone  :128   3rd Qu.: 164.2
##  Membran:   1   CemntBd: 61   CmentBd: 60                 Max.   :1600.0
##  (Other):   2   (Other):128   (Other):136
##  ExterQual ExterCond  Foundation  BsmtQual BsmtCond  BsmtExposure
##  Ex: 52    Ex:   3   BrkTil:146   0 : 37   0 :  37   0 : 38
##  Fa: 14    Fa:  28   CBlock:634   Ex:121   Fa:  45   Av:221
##  Gd:488    Gd: 146   PConc :647   Fa: 35   Gd:  65   Gd:134
##  TA:906    Po:   1   Slab  : 24   Gd:618   Po:   2   Mn:114
##            TA:1282   Stone :  6   TA:649   TA:1311   No:953
##                      Wood  :  3
##
##  BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
##  0  : 37      Min.   :   0.0   0  :  38     Min.   :   0.00
##  ALQ:220      1st Qu.:   0.0   ALQ:  19     1st Qu.:   0.00
##  BLQ:148      Median : 383.5   BLQ:  33     Median :   0.00
##  GLQ:418      Mean   : 443.6   GLQ:  14     Mean   :  46.55
##  LwQ: 74      3rd Qu.: 712.2   LwQ:  46     3rd Qu.:   0.00
##  Rec:133      Max.   :5644.0   Rec:  54     Max.   :1474.00
##  Unf:430                       Unf:1256
##    BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
##  Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
##  1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
##  Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  0    :   1   Min.   : 334   Min.   :   0   Min.   :  0.000
##  FuseA:  94   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  FuseF:  27   Median :1087   Median :   0   Median :  0.000
##  FuseP:   3   Mean   :1163   Mean   : 347   Mean   :  5.845
##  Mix  :   1   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  SBrkr:1334   Max.   :4692   Max.   :2065   Max.   :572.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
##  Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :1464   Median :0.0000   Median :0.00000   Median :2.000
##  Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
##  3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
```

```
##  Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000
##
##     HalfBath       BedroomAbvGr    Kitche0bvGr    KitchenQual
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000
##
##   TotRmsAbvGrd    Functio0l     Fireplaces    FireplaceQu   GarageType
##  Min.   : 2.000   Maj1:  14   Min.   :0.000   0 :690      0      : 81
##  1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   Ex: 24      2Types :  6
##  Median : 6.000   Min1:  31   Median :1.000   Fa: 33      Attchd :870
##  Mean   : 6.518   Min2:  34   Mean   :0.613   Gd:380      Basment: 19
##  3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   Po: 20      BuiltIn: 88
##  Max.   :14.000   Sev :   1   Max.   :3.000   TA:313      CarPort:  9
##                   Typ :1360                               Detchd :387
##   GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual
##  Min.   :   0   0  : 81      Min.   :0.000   Min.   :   0.0   0 :  81
##  1st Qu.:1958   Fin:352      1st Qu.:1.000   1st Qu.: 334.5   Ex:   3
##  Median :1977   RFn:422      Median :2.000   Median : 480.0   Fa:  48
##  Mean   :1869   Unf:605      Mean   :1.767   Mean   : 473.0   Gd:  14
##  3rd Qu.:2001                3rd Qu.:2.000   3rd Qu.: 576.0   Po:   3
##  Max.   :2010                Max.   :4.000   Max.   :1418.0   TA:1311
##
##  GarageCond PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
##  0 :  81    N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  Ex:   2    P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  Fa:  35    Y:1340     Median :  0.00   Median : 25.00   Median :  0.00
##  Gd:   9               Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
##  Po:   7               3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
##  TA:1326               Max.   :857.00   Max.   :547.00   Max.   :552.00
##
##    X3SsnPorch      ScreenPorch        PoolArea       PoolQC
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   0 :1453
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   Ex:   2
##  Median :  0.00   Median :  0.00   Median :  0.000   Fa:   2
##  Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   Gd:   3
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##    Fence      MiscFeature    MiscVal             MoSold
##  0    :1179   0   :1406   Min.   :    0.00   Min.   : 1.000
##  GdPrv:  59   Gar2:   2   1st Qu.:    0.00   1st Qu.: 5.000
##  GdWo :  54   Othr:   2   Median :    0.00   Median : 6.000
##  MnPrv: 157   Shed:  49   Mean   :   43.49   Mean   : 6.322
##  MnWw :  11   TenC:   1   3rd Qu.:    0.00   3rd Qu.: 8.000
##                           Max.   :15500.00   Max.   :12.000
##
```

```
##      YrSold        SaleType    SaleCondition    SalePrice
##  Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
##  1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
##  Median :2008   COD    :  43   Alloca :  12   Median :163000
##  Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
##  3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
##  Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
##                 (Other):   9
##      paved          regshape
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000
##  Mean   :0.9959   Mean   :0.6336
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

House Prices test dataset

```
summary(test)
```

```
##        Id         MSSubClass        MSZoning     LotFrontage
##  Min.   :1461   Min.   : 20.00   0      :   4   Min.   :  0.00
##  1st Qu.:1826   1st Qu.: 20.00   C (all):  15   1st Qu.: 44.00
##  Median :2190   Median : 50.00   FV     :  74   Median : 63.00
##  Mean   :2190   Mean   : 57.38   RH     :  10   Mean   : 57.91
##  3rd Qu.:2554   3rd Qu.: 70.00   RL     :1114   3rd Qu.: 78.00
##  Max.   :2919   Max.   :190.00   RM     : 242   Max.   :200.00
##
##     LotArea       Street      Alley      LotShape  LandContour
##  Min.   : 1470   Grvl:   6   0   :1352   IR1:484   Bnk:  54
##  1st Qu.: 7391   Pave:1453   Grvl:  70   IR2: 35   HLS:  70
##  Median : 9399               Pave:  37   IR3:  6   Low:  24
##  Mean   : 9819                           Reg:934   Lvl:1311
##  3rd Qu.:11518
##  Max.   :56600
##
##   Utilities      LotConfig    LandSlope   Neighborhood   Condition1
##  0     :   2   Corner : 248   Gtl:1396   0mes   :218   Norm   :1251
##  AllPub:1457   CulDSac:  82   Mod:  60   OldTown:126   Feedr  :  83
##                FR2    :  38   Sev:   3   CollgCr:117   Artery :  44
##                FR3    :  10              Somerst: 96   RRAn   :  24
##                Inside :1081              Edwards: 94   PosN   :  20
##                                          NridgHt: 89   RRAe   :  17
##                                          (Other):719   (Other):  20
##   Condition2     BldgType     HouseStyle   OverallQual      OverallCond
##  Artery:   3   1Fam  :1205   1.5Fin:160   Min.   : 1.000   Min.   :1.000
##  Feedr :   7   2fmCon:  31   1.5Unf:  5   1st Qu.: 5.000   1st Qu.:5.000
##  Norm  :1444   Duplex:  57   1Story:745   Median : 6.000   Median :5.000
##  PosA  :   3   Twnhs :  53   2.5Unf: 13   Mean   : 6.079   Mean   :5.554
```

```
##  PosN  :   2   TwnhsE: 113   2Story:427   3rd Qu.: 7.000   3rd Qu.:6.000
##                              SFoyer: 46   Max.   :10.000   Max.   :9.000
##                              SLvl  : 63
##    YearBuilt     YearRemodAdd    RoofStyle       RoofMatl     Exterior1st
##  Min.   :1879   Min.   :1950   Flat   :   7   CompShg:1442   VinylSd:510
##  1st Qu.:1953   1st Qu.:1963   Gable  :1169   Tar&Grv:  12   MetalSd:230
##  Median :1973   Median :1992   Gambrel:  11   WdShake:   4   HdBoard:220
##  Mean   :1971   Mean   :1984   Hip    : 265   WdShngl:   1   Wd Sdng:205
##  3rd Qu.:2001   3rd Qu.:2004   Mansard:   4                  Plywood:113
##  Max.   :2010   Max.   :2010   Shed   :   3                  CemntBd: 65
##                                                             (Other):116
##   Exterior2nd    MasVnrType    MasVnrArea      ExterQual ExterCond
##  VinylSd:510   0      : 16   Min.   :   0.00   Ex: 55    Ex:   9
##  MetalSd:233   BrkCmn : 10   1st Qu.:   0.00   Fa: 21    Fa:  39
##  HdBoard:199   BrkFace:434   Median :   0.00   Gd:491    Gd: 153
##  Wd Sdng:194   None   :878   Mean   :  99.67   TA:892    Po:   2
##  Plywood:128   Stone  :121   3rd Qu.: 162.00             TA:1256
##  CmentBd: 66                 Max.   :1290.00
##  (Other):129
##   Foundation  BsmtQual BsmtCond  BsmtExposure BsmtFinType1
##  BrkTil:165   0 : 44   0 :  45   0 : 44       0  : 42
##  CBlock:601   Ex:137   Fa:  59   Av:197       ALQ:209
##  PConc :661   Fa: 53   Gd:  57   Gd:142       BLQ:121
##  Slab  : 25   Gd:591   Po:   3   Mn:125       GLQ:431
##  Stone :  5   TA:634   TA:1295   No:951       LwQ: 80
##  Wood  :  2                                  Rec:155
##                                              Unf:421
##    BsmtFinSF1     BsmtFinType2   BsmtFinSF2        BsmtUnfSF
##  Min.   :   0.0   0  :  42     Min.   :   0.00   Min.   :   0.0
##  1st Qu.:   0.0   ALQ:  33     1st Qu.:   0.00   1st Qu.: 219.0
##  Median : 350.0   BLQ:  35     Median :   0.00   Median : 460.0
##  Mean   : 438.9   GLQ:  20     Mean   :  52.58   Mean   : 553.9
##  3rd Qu.: 752.0   LwQ:  41     3rd Qu.:   0.00   3rd Qu.: 797.5
##  Max.   :4010.0   Rec:  51     Max.   :1526.00   Max.   :2140.0
##                   Unf:1237
##   TotalBsmtSF   Heating     HeatingQC CentralAir Electrical
##  Min.   :   0   GasA:1446   Ex:752    N: 101     FuseA:  94
##  1st Qu.: 784   GasW:   9   Fa: 43    Y:1358     FuseF:  23
##  Median : 988   Grav:   2   Gd:233               FuseP:   5
##  Mean   :1045   Wall:   2   Po:  2               SBrkr:1337
##  3rd Qu.:1304               TA:429
##  Max.   :5095
##
##    X1stFlrSF        X2ndFlrSF     LowQualFinSF        GrLivArea
##  Min.   : 407.0   Min.   :   0   Min.   :   0.000   Min.   : 407
##  1st Qu.: 873.5   1st Qu.:   0   1st Qu.:   0.000   1st Qu.:1118
##  Median :1079.0   Median :   0   Median :   0.000   Median :1432
##  Mean   :1156.5   Mean   : 326   Mean   :   3.543   Mean   :1486
##  3rd Qu.:1382.5   3rd Qu.: 676   3rd Qu.:   0.000   3rd Qu.:1721
##  Max.   :5095.0   Max.   :1862   Max.   :1064.000   Max.   :5095
```

```
##
##   BsmtFullBath     BsmtHalfBath        FullBath        HalfBath
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :2.000   Median :0.0000
##  Mean   :0.4339   Mean   :0.06511   Mean   :1.571   Mean   :0.3777
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :3.0000   Max.   :2.00000   Max.   :4.000   Max.   :2.0000
##
##   BedroomAbvGr    Kitche0bvGr    KitchenQual  TotRmsAbvGrd
##  Min.   :0.000   Min.   :0.000   0 :  1      Min.   : 3.000
##  1st Qu.:2.000   1st Qu.:1.000   Ex:105      1st Qu.: 5.000
##  Median :3.000   Median :1.000   Fa: 31      Median : 6.000
##  Mean   :2.854   Mean   :1.042   Gd:565      Mean   : 6.385
##  3rd Qu.:3.000   3rd Qu.:1.000   TA:757      3rd Qu.: 7.000
##  Max.   :6.000   Max.   :2.000               Max.   :15.000
##
##    Functio0l      Fireplaces     FireplaceQu   GarageType   GarageYrBlt
##  Typ    :1357   Min.   :0.0000   0 :730      0      : 76   Min.   :   0
##  Min2   :  36   1st Qu.:0.0000   Ex: 19      2Types : 17   1st Qu.:1956
##  Min1   :  34   Median :0.0000   Fa: 41      Attchd :853   Median :1977
##  Mod    :  20   Mean   :0.5812   Gd:364      Basment: 17   Mean   :1872
##  Maj1   :   5   3rd Qu.:1.0000   Po: 26      BuiltIn: 98   3rd Qu.:2001
##  Maj2   :   4   Max.   :4.0000   TA:279      CarPort:  6   Max.   :2207
##  (Other):   3                               Detchd :392
##  GarageFinish   GarageCars      GarageArea     GarageQual GarageCond
##  0  : 78      Min.   :0.000   Min.   :   0.0   0 :  78    0 :  78
##  Fin:367      1st Qu.:1.000   1st Qu.: 317.5   Fa:  76    Ex:   1
##  RFn:389      Median :2.000   Median : 480.0   Gd:  10    Fa:  39
##  Unf:625      Mean   :1.765   Mean   : 472.4   Po:   2    Gd:   6
##               3rd Qu.:2.000   3rd Qu.: 576.0   TA:1293    Po:   7
##               Max.   :5.000   Max.   :1488.0              TA:1328
##
##  PavedDrive   WoodDeckSF       OpenPorchSF     EnclosedPorch
##  N: 126     Min.   :   0.00   Min.   :  0.00   Min.   :   0.00
##  P:  32     1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00
##  Y:1301     Median :   0.00   Median : 28.00   Median :   0.00
##             Mean   :  93.17   Mean   : 48.31   Mean   :  24.24
##             3rd Qu.: 168.00   3rd Qu.: 72.00   3rd Qu.:   0.00
##             Max.   :1424.00   Max.   :742.00   Max.   :1012.00
##
##    X3SsnPorch       ScreenPorch        PoolArea       PoolQC
##  Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   0 :1456
##  1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000   Ex:   2
##  Median :  0.000   Median :  0.00   Median :  0.000   Gd:   1
##  Mean   :  1.794   Mean   : 17.06   Mean   :  1.744
##  3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :360.000   Max.   :576.00   Max.   :800.000
##
##    Fence      MiscFeature    MiscVal             MoSold
```

```
##  0    :1169   0   :1408   Min.   :    0.00   Min.   : 1.000
##  GdPrv:  59   Gar2:   3   1st Qu.:    0.00   1st Qu.: 4.000
##  GdWo :  58   Othr:   2   Median :    0.00   Median : 6.000
##  MnPrv: 172   Shed:  46   Mean   :   58.17   Mean   : 6.104
##  MnWw :   1               3rd Qu.:    0.00   3rd Qu.: 8.000
##                           Max.   :17000.00   Max.   :12.000
##
##      YrSold        SaleType    SaleCondition
##  Min.   :2006   WD     :1258   Abnorml:  89
##  1st Qu.:2007   New    : 117   AdjLand:   8
##  Median :2008   COD    :  44   Alloca :  12
##  Mean   :2008   ConLD  :  17   Family :  26
##  3rd Qu.:2009   CWD    :   8   Normal :1204
##  Max.   :2010   ConLI  :   4   Partial: 120
##                 (Other):  11
```

House Prices Data Description

MSSubClass: Identifies the type of dwelling involved in the sale.

```
    20  1-STORY 1946 & NEWER ALL STYLES
    30  1-STORY 1945 & OLDER
    40  1-STORY W/FINISHED ATTIC ALL AGES
    45  1-1/2 STORY - UNFINISHED ALL AGES
    50  1-1/2 STORY FINISHED ALL AGES
    60  2-STORY 1946 & NEWER
    70  2-STORY 1945 & OLDER
    75  2-1/2 STORY ALL AGES
    80  SPLIT OR MULTI-LEVEL
    85  SPLIT FOYER
    90  DUPLEX - ALL STYLES AND AGES
   120  1-STORY PUD (Planned Unit Development) - 1946 & NEWER
   150  1-1/2 STORY PUD - ALL AGES
   160  2-STORY PUD - 1946 & NEWER
   180  PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
   190  2 FAMILY CONVERSION - ALL STYLES AND AGES
```

MSZoning: Identifies the general zoning classification of the sale.

```
   A    Agriculture
   C    Commercial
   FV   Floating Village Residential
   I    Industrial
   RH   Residential High Density
   RL   Residential Low Density
   RP   Residential Low Density Park
   RM   Residential Medium Density
```

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

```
   Grvl Gravel
   Pave Paved
```

Alley: Type of alley access to property

```
   Grvl Gravel
   Pave Paved
   NA   No alley access
```

LotShape: General shape of property

```
   Reg  Regular
   IR1  Slightly irregular
   IR2  Moderately Irregular
   IR3  Irregular
```

LandContour: Flatness of the property

```
   Lvl  Near Flat/Level
   Bnk  Banked - Quick and significant rise from street grade to building
   HLS  Hillside - Significant slope from side to side
   Low  Depression
```

Utilities: Type of utilities available

```
   AllPub   All public Utilities (E,G,W,& S)
   NoSewr   Electricity, Gas, and Water (Septic Tank)
   NoSeWa   Electricity and Gas Only
   ELO  Electricity only
```

LotConfig: Lot configuration

```
   Inside   Inside lot
   Corner   Corner lot
   CulDSac  Cul-de-sac
   FR2  Frontage on 2 sides of property
   FR3  Frontage on 3 sides of property
```

LandSlope: Slope of property

```
   Gtl  Gentle slope
   Mod  Moderate Slope
   Sev  Severe Slope
```

Neighborhood: Physical locations within Ames city limits

```
   Blmngtn  Bloomington Heights
   Blueste  Bluestem
   BrDale   Briardale
   BrkSide  Brookside
   ClearCr  Clear Creek
   CollgCr  College Creek
   Crawfor  Crawford
   Edwards  Edwards
   Gilbert  Gilbert
   IDOTRR   Iowa DOT and Rail Road
   MeadowV  Meadow Village
   Mitchel  Mitchell
   Names    North Ames
   NoRidge  Northridge
   NPkVill  Northpark Villa
   NridgHt  Northridge Heights
   NWAmes   Northwest Ames
   OldTown  Old Town
   SWISU    South & West of Iowa State University
   Sawyer   Sawyer
   SawyerW  Sawyer West
   Somerst  Somerset
   StoneBr  Stone Brook
   Timber   Timberland
   Veenker  Veenker
```

Condition1: Proximity to various conditions

```
   Artery   Adjacent to arterial street
   Feedr    Adjacent to feeder street
   Norm Normal
   RRNn Within 200' of North-South Railroad
   RRAn Adjacent to North-South Railroad
   PosN Near positive off-site feature--park, greenbelt, etc.
   PosA Adjacent to postive off-site feature
   RRNe Within 200' of East-West Railroad
   RRAe Adjacent to East-West Railroad
```

Condition2: Proximity to various conditions (if more than one is present)

```
   Artery   Adjacent to arterial street
   Feedr    Adjacent to feeder street
   Norm Normal
   RRNn Within 200' of North-South Railroad
   RRAn Adjacent to North-South Railroad
   PosN Near positive off-site feature--park, greenbelt, etc.
   PosA Adjacent to postive off-site feature
   RRNe Within 200' of East-West Railroad
   RRAe Adjacent to East-West Railroad
```

BldgType: Type of dwelling

```
   1Fam Single-family Detached
   2FmCon   Two-family Conversion; originally built as one-family dwelling
   Duplx    Duplex
   TwnhsE   Townhouse End Unit
   TwnhsI   Townhouse Inside Unit
```

HouseStyle: Style of dwelling

```
   1Story   One story
   1.5Fin   One and one-half story: 2nd level finished
   1.5Unf   One and one-half story: 2nd level unfinished
   2Story   Two story
   2.5Fin   Two and one-half story: 2nd level finished
   2.5Unf   Two and one-half story: 2nd level unfinished
   SFoyer   Split Foyer
   SLvl Split Level
```

OverallQual: Rates the overall material and finish of the house

```
   10   Very Excellent
   9    Excellent
   8    Very Good
   7    Good
   6    Above Average
   5    Average
   4    Below Average
   3    Fair
   2    Poor
   1    Very Poor
```

OverallCond: Rates the overall condition of the house

```
   10   Very Excellent
   9    Excellent
   8    Very Good
   7    Good
   6    Above Average
   5    Average
   4    Below Average
   3    Fair
   2    Poor
   1    Very Poor
```

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

```
   Flat Flat
   Gable    Gable
   Gambrel  Gabrel (Barn)
   Hip  Hip
   Mansard  Mansard
   Shed Shed
```

RoofMatl: Roof material

```
   ClyTile  Clay or Tile
   CompShg  Standard (Composite) Shingle
   Membran  Membrane
   Metal    Metal
   Roll Roll
   Tar&Grv  Gravel & Tar
   WdShake  Wood Shakes
   WdShngl  Wood Shingles
```

Exterior1st: Exterior covering on house

```
   AsbShng  Asbestos Shingles
   AsphShn  Asphalt Shingles
   BrkComm  Brick Common
   BrkFace  Brick Face
   CBlock   Cinder Block
   CemntBd  Cement Board
   HdBoard  Hard Board
   ImStucc  Imitation Stucco
   MetalSd  Metal Siding
   Other    Other
   Plywood  Plywood
   PreCast  PreCast
   Stone    Stone
   Stucco   Stucco
   VinylSd  Vinyl Siding
   Wd Sdng  Wood Siding
   WdShing  Wood Shingles
```

Exterior2nd: Exterior covering on house (if more than one material)

```
   AsbShng  Asbestos Shingles
   AsphShn  Asphalt Shingles
   BrkComm  Brick Common
   BrkFace  Brick Face
   CBlock   Cinder Block
   CemntBd  Cement Board
   HdBoard  Hard Board
   ImStucc  Imitation Stucco
   MetalSd  Metal Siding
```

```
   Other    Other
   Plywood  Plywood
   PreCast  PreCast
   Stone    Stone
   Stucco   Stucco
   VinylSd  Vinyl Siding
   Wd Sdng  Wood Siding
   WdShing  Wood Shingles
```

MasVnrType: Masonry veneer type

```
   BrkCmn   Brick Common
   BrkFace  Brick Face
   CBlock   Cinder Block
   None None
   Stone    Stone
```

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

```
   Ex   Excellent
   Gd   Good
   TA   Average/Typical
   Fa   Fair
   Po   Poor
```

ExterCond: Evaluates the present condition of the material on the exterior

```
   Ex   Excellent
   Gd   Good
   TA   Average/Typical
   Fa   Fair
   Po   Poor
```

Foundation: Type of foundation

```
   BrkTil   Brick & Tile
   CBlock   Cinder Block
   PConc    Poured Contrete
   Slab Slab
   Stone    Stone
   Wood Wood
```

BsmtQual: Evaluates the height of the basement

```
   Ex   Excellent (100+ inches)
   Gd   Good (90-99 inches)
   TA   Typical (80-89 inches)
```

```
   Fa   Fair (70-79 inches)
   Po   Poor (<70 inches
   NA   No Basement
```

BsmtCond: Evaluates the general condition of the basement

```
   Ex   Excellent
   Gd   Good
   TA   Typical - slight dampness allowed
   Fa   Fair - dampness or some cracking or settling
   Po   Poor - Severe cracking, settling, or wetness
   NA   No Basement
```

BsmtExposure: Refers to walkout or garden level walls

```
   Gd   Good Exposure
   Av   Average Exposure (split levels or foyers typically score average or
above)
   Mn   Mimimum Exposure
   No   No Exposure
   NA   No Basement
```

BsmtFinType1: Rating of basement finished area

```
   GLQ  Good Living Quarters
   ALQ  Average Living Quarters
   BLQ  Below Average Living Quarters
   Rec  Average Rec Room
   LwQ  Low Quality
   Unf  Unfinshed
   NA   No Basement
```

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

```
   GLQ  Good Living Quarters
   ALQ  Average Living Quarters
   BLQ  Below Average Living Quarters
   Rec  Average Rec Room
   LwQ  Low Quality
   Unf  Unfinshed
   NA   No Basement
```

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

```
   Floor    Floor Furnace
   GasA Gas forced warm air furnace
   GasW Gas hot water or steam heat
   Grav Gravity furnace
   OthW Hot water or steam heat other than gas
   Wall Wall furnace
```

HeatingQC: Heating quality and condition

```
   Ex   Excellent
   Gd   Good
   TA   Average/Typical
   Fa   Fair
   Po   Poor
```

CentralAir: Central air conditioning

```
   N    No
   Y    Yes
```

Electrical: Electrical system

```
   SBrkr    Standard Circuit Breakers & Romex
   FuseA    Fuse Box over 60 AMP and all Romex wiring (Average)
   FuseF    60 AMP Fuse Box and mostly Romex wiring (Fair)
   FuseP    60 AMP Fuse Box and mostly knob & tube wiring (poor)
   Mix  Mixed
```

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

```
   Ex   Excellent
   Gd   Good
   TA   Typical/Average
   Fa   Fair
   Po   Poor
```

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

```
   Typ  Typical Functionality
   Min1 Minor Deductions 1
   Min2 Minor Deductions 2
   Mod  Moderate Deductions
   Maj1 Major Deductions 1
   Maj2 Major Deductions 2
   Sev  Severely Damaged
   Sal  Salvage only
```

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

```
   Ex   Excellent - Exceptional Masonry Fireplace
   Gd   Good - Masonry Fireplace in main level
   TA   Average - Prefabricated Fireplace in main living area or Masonry
Fireplace in basement
   Fa   Fair - Prefabricated Fireplace in basement
   Po   Poor - Ben Franklin Stove
   NA   No Fireplace
```

GarageType: Garage location

```
   2Types   More than one type of garage
   Attchd   Attached to home
   Basment  Basement Garage
   BuiltIn  Built-In (Garage part of house - typically has room above garage)
   CarPort  Car Port
   Detchd   Detached from home
   NA   No Garage
```

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

```
   Fin  Finished
   RFn  Rough Finished
   Unf  Unfinished
   NA   No Garage
```

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

```
   Ex   Excellent
   Gd   Good
   TA   Typical/Average
   Fa   Fair
   Po   Poor
   NA   No Garage
```

GarageCond: Garage condition

```
   Ex   Excellent
   Gd   Good
   TA   Typical/Average
   Fa   Fair
   Po   Poor
   NA   No Garage
```

PavedDrive: Paved driveway

```
   Y    Paved
   P    Partial Pavement
   N    Dirt/Gravel
```

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

```
   Ex   Excellent
   Gd   Good
```

```
   TA   Average/Typical
   Fa   Fair
   NA   No Pool
```

Fence: Fence quality

```
   GdPrv    Good Privacy
   MnPrv    Minimum Privacy
   GdWo Good Wood
   MnWw Minimum Wood/Wire
   NA   No Fence
```

MiscFeature: Miscellaneous feature not covered in other categories

```
   Elev Elevator
   Gar2 2nd Garage (if not described in garage section)
   Othr Other
   Shed Shed (over 100 SF)
   TenC Tennis Court
   NA   None
```

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

```
   WD   Warranty Deed - Conventional
   CWD  Warranty Deed - Cash
   VWD  Warranty Deed - VA Loan
   New  Home just constructed and sold
   COD  Court Officer Deed/Estate
   Con  Contract 15% Down payment regular terms
   ConLw    Contract Low Down payment and low interest
   ConLI    Contract Low Interest
   ConLD    Contract Low Down
   Oth  Other
```

SaleCondition: Condition of sale

```
   Normal   Normal Sale
   Abnorml  Abnormal Sale -  trade, foreclosure, short sale
   AdjLand  Adjoining Land Purchase
   Alloca   Allocation - two linked properties with separate deeds, typically
condo with a garage unit
   Family   Sale between family members
```

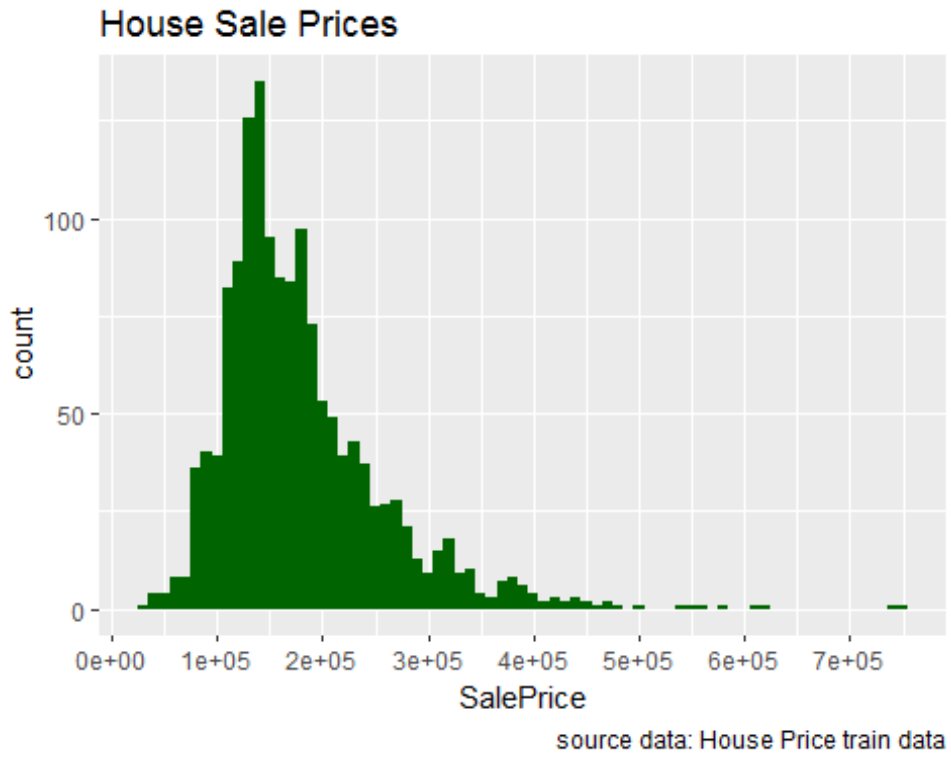
```
   Partial  Home was not completed when last assessed (associated with New
Homes)
```

SalePrice: the property's sale price in dollars. This is the target variable that you're trying to predict.

4. Methods and Analysis: Data exploration and visualization

As you can see from the chart below, the SalePrice in the train dataset is right skewed. This is not unusual since the most expensive homes greater than 400K have lower volume of sales as compared with the bulk of sales in the 100K to 300K range.

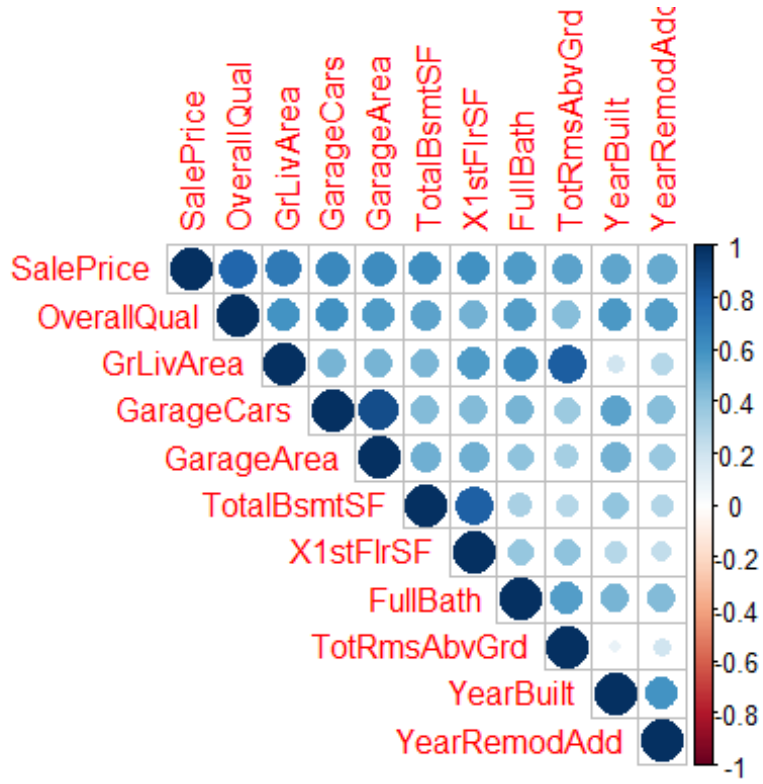
```
ggplot(data=train, aes(x=SalePrice)) +
  geom_histogram(fill="dark green", binwidth = 10000) +
  scale_x_continuous(breaks= seq(0, 800000, by=100000)) +
  labs(title="House Sale Prices",
       caption = "source data: House Price train data")
```



In evaluating the dependent variables that are most important in predicting SalePrice I created a correlation matrix with SalePrice. The correlation matrix table below shows that there are 10 variables out of 37 numeric variables in the train dataset with a correlation of at least 0.5 and are greater than 0.

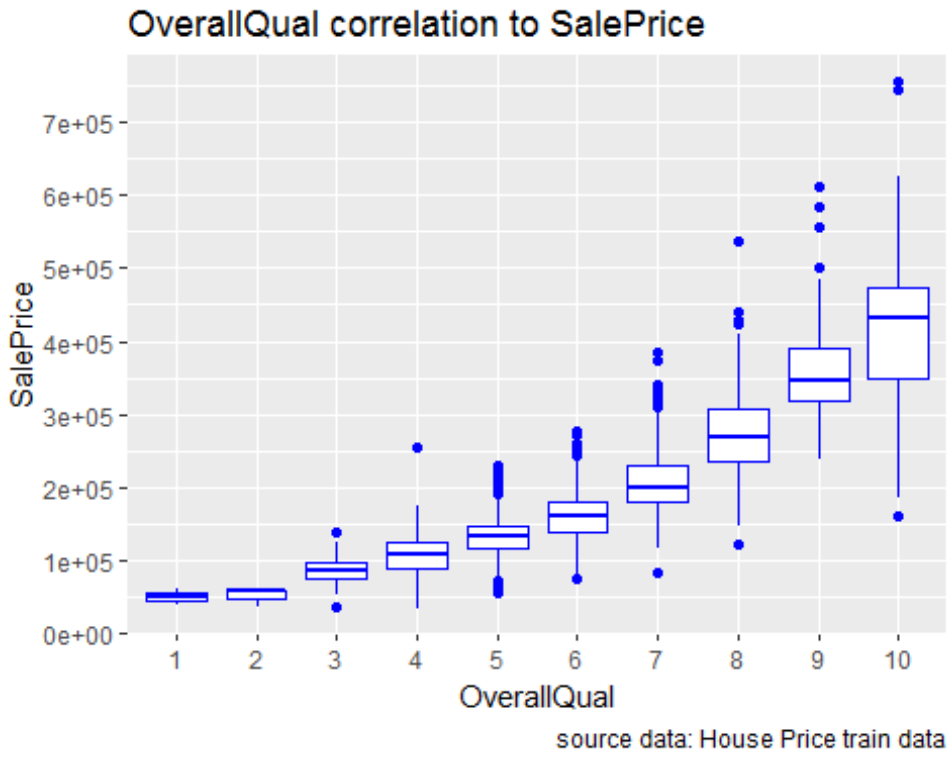
```
num_vars <- which(sapply(train, is.numeric)) #index vector numeric variables
num_vars_colnames <- data.table(names(num_vars)) #column names of the numeric
variables
```

```
train_num_vars <- train[, num_vars]
cor_num_vars <- cor(train_num_vars, use="pairwise.complete.obs")
#correlations of all numeric variables

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

cor_num_vars <- cor_num_vars[high_cor, high_cor]

corrplot(cor_num_vars, type = "upper")
```



As shown from the above dependent variable correlation matrix to SalePrice, OverallQual has the highest correlation at of the numeric variables at 0.8. We will plot this correlation in a box plot below to visualize the accuracy if OverallQual as the primary predictor of SalePrice.

```
ggplot(data=train[!is.na(train$SalePrice),], aes(x=factor(OverallQual),
y=SalePrice))+
        geom_boxplot(col='blue') + labs(x='OverallQual') +
        scale_y_continuous(breaks= seq(0, 800000, by=100000))+
        labs(title="OverallQual correlation to SalePrice", caption = "source
data: House Price train data")
```



5. Evaluated Machine Learning Algorithms

5.1 Model Prepping

Before we can start modeling, we need to first partition the train dataset and here I will use the caret package.

```
set.seed(123)

outcome <- train$SalePrice

partition <- createDataPartition(y=outcome,
                                 p=.5,
                                 list=F)
training <- train[partition,]
testing <- train[-partition,]
```

5.2 Simple Linear Regression Model

The simple linear regression model that I will generate first uses the OverallQual bias or the *OverallQual effect* b_o dependent variable on the training_set to predict the `SalePrice` (Y).

The *OverallQual effect model* is calculated as follows:

$$Y = \mu + b_o + \varepsilon$$

where:

* (μ) is the mean SalePrice for all Houses.

* (b_o) effects or bias, OverallQual effect.

* (ε) are independent errors sampled from the same distribution centered at 0.

The resulting RMSLE from this *OverallQual effect model* was 0.2690968. We can likely do better with adding more dependent variables to the model.

```
# Fitting Simple Regression Model to the train set.
set.seed(123)

OQ_effect_model <- lm(SalePrice ~ OverallQual, data = training)

summary(OQ_effect_model)

##
## Call:
## lm(formula = SalePrice ~ OverallQual, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -196886  -27298    -550   21036  299939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -94796       7622  -12.44   <2e-16 ***
## OverallQual    45168       1220   37.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46130 on 729 degrees of freedom
## Multiple R-squared:  0.6529, Adjusted R-squared:  0.6524
## F-statistic:  1371 on 1 and 729 DF,  p-value: < 2.2e-16

prediction <- predict(OQ_effect_model, testing, type="response")

prediction_log <- log(prediction)

testing_log <- log(testing$SalePrice)

model_rmse <- RMSE(testing_log, prediction_log)

RMSLE_table <- data_frame(Method = "Regression model using OverallQual
effect",
                                RMSLE = model_rmse)

RMSLE_table %>% knitr::kable(caption = "RMSLEs")
```

RMSLEs

Method	RMSLE
Regression model using OverallQual effect	0.2680968

5.3 Multiple Linear Regression Model

The next regression model that I will generate uses the the top 10 most correlated numeric variables to SalePrice on the train set to predict the `rating` Y for the test set.

The *top-10-effect model* is calculated as follows:

$$Y_{u,i} = \mu + b_1 + b_2 + b_3 + b_4 + b_5 + b_6 + b_7 + b_8 + b_9 + b_{10} + \varepsilon$$

where:

* (μ) **is the mean SalePrice for all Houses.**

* (b_1) **effects or bias, OverallQual effect.**

* (b_2) **effects or bias, GrLivArea effect.**

* (b_3) **effects or bias, GarageCars effect.**

* (b_4) **effects or bias, GarageArea effect.**

* (b_5) **effects or bias, TotalBsmtSF effect.**

* (b_6) **effects or bias, X1stFlrSF effect.**

* (b_7) **effects or bias, FullBath effect.**

* (b_8) **effects or bias, TotRmsAbvGrd effect.**

* (b_9) **effects or bias, YearBuilt effect.**

* (b_{10}) **effects or bias, YearRemodAdd effect.**

* (ε) **are independent errors sampled from the same distribution centered at 0.**

The resulting RMSLE from this *top-10-effect model* on the test_set was an improvement on the *OverallQual effect model* RMSLE at 0.1912581. Let's see if we can do better with backward elimination.

```
set.seed(123)

top10_effect_model <- lm(SalePrice ~ OverallQual + GrLivArea + GarageCars +
                           GarageArea + TotalBsmtSF + X1stFlrSF + FullBath +
                           TotRmsAbvGrd + YearBuilt + YearRemodAdd, data =
training)
```

```
summary(top10_effect_model)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + GarageCars +
##     GarageArea + TotalBsmtSF + X1stFlrSF + FullBath + TotRmsAbvGrd +
##     YearBuilt + YearRemodAdd, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397852  -18994   -2084   16901  258017
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.086e+06  1.865e+05  -5.819 8.89e-09 ***
## OverallQual   2.253e+04  1.683e+03  13.383  < 2e-16 ***
## GrLivArea     3.319e+01  6.013e+00   5.519 4.76e-08 ***
## GarageCars    1.341e+04  4.208e+03   3.188 0.001494 **
## GarageArea    9.853e+00  1.406e+01   0.701 0.483805
## TotalBsmtSF   1.231e+01  6.385e+00   1.928 0.054292 .
## X1stFlrSF     1.355e+01  7.257e+00   1.868 0.062183 .
## FullBath     -3.829e+03  3.773e+03  -1.015 0.310457
## TotRmsAbvGrd  3.143e+03  1.635e+03   1.922 0.055013 .
## YearBuilt     2.071e+02  7.242e+01   2.859 0.004369 **
## YearRemodAdd  3.015e+02  8.722e+01   3.457 0.000579 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38120 on 720 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7626
## F-statistic: 235.4 on 10 and 720 DF,  p-value: < 2.2e-16
```

```
prediction <- predict(top10_effect_model, testing, type="response")

prediction_log <- log(prediction)

testing_log <- log(testing$SalePrice)

model_rmse <- RMSE(testing_log, prediction_log)

RMSLE_table <- rbind(RMSLE_table,
                     data_frame(Method = "Regression model using top-10-
effect",
                                RMSLE = model_rmse))

RMSLE_table %>% knitr::kable(caption = "RMSLEs")
```

RMSLEs

Method	RMSLE
Regression model using OverallQual effect	0.2680968
Regression model using top-10-effect	0.1914210

5.4 Backward Elimination Linear Regression Model

The next regression model that I will generate uses backwards elimination to pick the most signification bias variables to SalePrice on the train set to predict the `rating` Y for the test set.

The *backward-elimination model* is calculated by taking the 5 most significant variables from the top-10 model:

$$Y_{u,i} = \mu + b_1 + b_2 + b_3 + b_4 + b_5 + \varepsilon$$

where:

* (μ) **is the mean SalePrice for all Houses.**

* (b_1) **effects or bias, OverallQual effect.**

* (b_2) **effects or bias, GrLivArea effect.**

* (b_3) **effects or bias, GarageCars effect. .**

* (b_4) **effects or bias, YearBuilt effect.**

* (b_5) **effects or bias, YearRemodAdd effect.**

* (ε) **are independent errors sampled from the same distribution centered at 0.**

The resulting RMSLE from this *backward elimination model* on the test_set was not an improvement and in a sense we did go backward at 0.2039798. Let's see if we can do better with a random forest model.

```
set.seed(123)

top10_effect_model <- lm(SalePrice ~ OverallQual + GrLivArea + GarageCars +
YearBuilt + YearRemodAdd, data = training)

summary(top10_effect_model)

##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + GarageCars +
##     YearBuilt + YearRemodAdd, data = training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -323701  -22245   -2232   18080  285361
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.021e+06  1.767e+05  -5.778 1.12e-08 ***
## OverallQual   2.391e+04  1.670e+03  14.322  < 2e-16 ***
## GrLivArea     4.640e+01  3.729e+00  12.443  < 2e-16 ***
## GarageCars    1.787e+04  2.671e+03   6.690 4.46e-11 ***
## YearBuilt     2.231e+02  6.825e+01   3.268  0.00113 **
## YearRemodAdd  2.591e+02  8.843e+01   2.930  0.00350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39040 on 725 degrees of freedom
## Multiple R-squared:  0.7527, Adjusted R-squared:  0.751
## F-statistic: 441.4 on 5 and 725 DF,  p-value: < 2.2e-16
```

```
prediction <- predict(top10_effect_model, testing, type="response")

prediction_log <- log(prediction)

testing_log <- log(testing$SalePrice)

model_rmse <- RMSE(testing_log, prediction_log)

RMSLE_table <- rbind(RMSLE_table,
                     data_frame(Method = "Regression model using backward
elimination",
                                RMSLE = model_rmse))

RMSLE_table %>% knitr::kable(caption = "RMSLEs")
```

RMSLEs

Method	RMSLE
Regression model using OverallQual effect	0.2680968
Regression model using top-10-effect	0.1914210
Regression model using backward elimination	0.2039798

5.5 Random Forest Regression Model

The next regression model that I will generate uses all variables to SalePrice on the training set using the randomForest package.

The resulting RMSLE from this *Random Forest model* to predict SalePrice on the testing set resulted in a RMSLE of 0.1384136. Now let us try a regularization method.

```
set.seed(123)
rf_model <- randomForest(SalePrice ~ ., data = training)

prediction <- predict(rf_model, testing)

prediction_log <- log(prediction)

testing_log <- log(testing$SalePrice)

model_rmse <- RMSE(testing_log, prediction_log)

RMSLE_table <- rbind(RMSLE_table,
                     data_frame(Method = "Random Forest regression model",
                                RMSLE = model_rmse))

RMSLE_table %>% knitr::kable(caption = "RMSLEs")
```

RMSLEs

Method	RMSLE
Regression model using OverallQual effect	0.2680968
Regression model using top-10-effect	0.1914210
Regression model using backward elimination	0.2039798
Random Forest regression model	0.1384136

5.6 Lasso Regression Model

The Lasso Regression model on all variables in the training set to seeks to minimize the RMSLE using cross validation to pick the optimal λ.

According to 'Statistics How To', in the Lasso Regression model, or Least Absolute Shrinkage and Selection Operator, it performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. A tuning parameter, λ controls the strength of the L1 penalty. λ is basically the amount of shrinkage.

```
# Lasso regularization method on the training set.

set.seed(123)

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

lasso_model <- train(SalePrice ~ ., data = training, method='glmnet',
trControl= my_control, tuneGrid=lassoGrid)
```

```
lambda_opt <- lasso_model$bestTune

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

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

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

## Lasso uses 235 variables in its model, and did not select 29 variables.

prediction <- predict(lasso_model, training)

prediction_log <- log(prediction)

training_log <- log(training$SalePrice)

model_rmse <- RMSE(training_log, prediction_log)

RMSLE_table <- rbind(RMSLE_table,
                     data_frame(Method = "Lasso regression model",
                                RMSLE = model_rmse))

RMSLE_table %>% knitr::kable(caption = "RMSLEs")
```

RMSLEs

Method	RMSLE
Regression model using OverallQual effect	0.2680968
Regression model using top-10-effect	0.1914210
Regression model using backward elimination	0.2039798
Random Forest regression model	0.1384136
Lasso regression model	0.0978573

6. Results:

The resulting RMSLE from this *Lasso regression model* on the training set brought the RMSLE down to 0.0978573. That is good enough for top 10 on the current Kaggle leaderboard!

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/leaderboard

7. Conclusion:

The main learning objective of the HarvardX: Introduction to Data Science program was to give aspiring data scientists like myself the tools in R to run analytic models using machine learning to make predictions and solve real world problems. This was a fascinating journey over the past six months and I look forward to improving my data science skills in R and Python through my work and personally through Kaggle competitions that I have an interest in.

My "Harvard_Data_Science_House Prices Github repository" is **in this link**

References

*** Irizzary,R., 2019. Introduction to Data Science,**

github page,https://rafalab.github.io/dsbook/

*** Statistics How To, https://www.statisticshowto.datasciencecentral.com/lasso-regression/**