HarvardX Data Science Capstone: House Prices Report

sophia rao

2023-12-05

Introduction

House Prices: Advanced Regression Techniques is a competition on Kaggle which has participants use regression techniques to predict sale prices of homes in Ames, Iowa. The given data includes four csv files train.csv, test.csv, sample_submission.csv and data_description.txt. The train dataset includes 1,459 rows (houses) and 81 columns (features). The test dataset includes 1,460 rows (houses) and 81 columns (features). The last variable in the train dataset is the sale price that we are trying to predict for the houses in the test dataset. The samp_submission file format is the Id of the house and predicted SalesPrice separated by a comma. The data_description.txt file includes descriptions of all the variables in the train and test datasets.

Methods

To start, we load the essential packages needed to complete the project.

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

Here, we read in the data from the csv files. For the sake of modeling, any characters will be converted to numeric values. We replace NA values with 0. In the street column, we replace “Pave” with 1. Other values that are not “Pave” will be replaced with 0. The same goes for the lot shape column.”Reg” is replaced with 1, and all other values are replaced with 0.

```
#Read in the training data set
train <- read.csv("train.csv")

#Replace NA with 0- numerics are better for modeling
train <- train %>% mutate_all(~replace(., is.na(.), 0))

#Read in the testing data set
test <- read.csv("test.csv")

#Replace NA with 0- numerics are better for modeling
test <- test %>% mutate_all(~replace(., is.na(.), 0))

#Converting characters to numerics
train$paved[train$Street == "Pave"] <- 1
train$paved[train$Street != "Pave"] <- 0

#Converting characters to numerics
train$regshape[train$LotShape == "Reg"] <- 1
train$regshape[train$LotShape != "Reg"] <- 0
```

Here, we can call the summary function on the train dataframe to get an idea of the data set. This will provide a six number summary for each of the columns in the data set. It will provide the minimum value (Min), the first quartile (1st Qu.), the median (Median), the mean (Mean), the third quartile (3rd Qu.) and the maximum value (Max). With these numbers, we can see how these features vary.

```
summary(train)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   :  0.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 42.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 63.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 57.62
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 79.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##     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   : 103.1
##                                        3rd Qu.: 164.2
##                                        Max.   :1600.0
##   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.   :   0
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1958
##  Median :1.000   Mode  :character   Mode  :character   Median :1977
##  Mean   :0.613                                         Mean   :1869
##  3rd Qu.:1.000                                         3rd Qu.:2001
##  Max.   :3.000                                         Max.   :2010
##  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          paved           regshape
##  Min.   : 34900   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:129975   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :163000   Median :1.0000   Median :1.0000
##  Mean   :180921   Mean   :0.9959   Mean   :0.6336
##  3rd Qu.:214000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :755000   Max.   :1.0000   Max.   :1.0000
```

We can do the same thing as above and call the summary function on the test dataframe.

```
summary(test)
```

```
##        Id         MSSubClass       MSZoning          LotFrontage
##  Min.   :1461   Min.   : 20.00   Length:1459        Min.   :  0.00
##  1st Qu.:1826   1st Qu.: 20.00   Class :character   1st Qu.: 44.00
##  Median :2190   Median : 50.00   Mode  :character   Median : 63.00
##  Mean   :2190   Mean   : 57.38                      Mean   : 57.91
##  3rd Qu.:2554   3rd Qu.: 70.00                      3rd Qu.: 78.00
##  Max.   :2919   Max.   :190.00                      Max.   :200.00
##     LotArea         Street             Alley             LotShape
##  Min.   : 1470   Length:1459        Length:1459        Length:1459
##  1st Qu.: 7391   Class :character   Class :character   Class :character
##  Median : 9399   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 9819
##  3rd Qu.:11518
```

```
##  Max.   :56600
##  LandContour         Utilities          LotConfig          LandSlope
##  Length:1459        Length:1459        Length:1459        Length:1459
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1459        Length:1459        Length:1459        Length:1459
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1459        Min.   : 1.000   Min.   :1.000   Min.   :1879
##  Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1953
##  Mode  :character   Median : 6.000   Median :5.000   Median :1973
##                     Mean   : 6.079   Mean   :5.554   Mean   :1971
##                     3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001
##                     Max.   :10.000   Max.   :9.000   Max.   :2010
##   YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
##  Min.   :1950   Length:1459        Length:1459        Length:1459
##  1st Qu.:1963   Class :character   Class :character   Class :character
##  Median :1992   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1984
##  3rd Qu.:2004
##  Max.   :2010
##  Exterior2nd         MasVnrType          MasVnrArea       ExterQual
##  Length:1459        Length:1459        Min.   :   0.00   Length:1459
##  Class :character   Class :character   1st Qu.:   0.00   Class :character
##  Mode  :character   Mode  :character   Median :   0.00   Mode  :character
##                                        Mean   :  99.67
##                                        3rd Qu.: 162.00
##                                        Max.   :1290.00
##   ExterCond          Foundation          BsmtQual           BsmtCond
##  Length:1459        Length:1459        Length:1459        Length:1459
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1459        Length:1459        Min.   :   0.0   Length:1459
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 350.0   Mode  :character
##                                        Mean   : 438.9
##                                        3rd Qu.: 752.0
##                                        Max.   :4010.0
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF     Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0   Length:1459
##  1st Qu.:   0.00   1st Qu.: 219.0   1st Qu.: 784   Class :character
##  Median :   0.00   Median : 460.0   Median : 988   Mode  :character
##  Mean   :  52.58   Mean   : 553.9   Mean   :1045
##  3rd Qu.:   0.00   3rd Qu.: 797.5   3rd Qu.:1304
##  Max.   :1526.00   Max.   :2140.0   Max.   :5095
##   HeatingQC          CentralAir         Electrical          X1stFlrSF
##  Length:1459        Length:1459        Length:1459        Min.   : 407.0
##  Class :character   Class :character   Class :character   1st Qu.: 873.5
##  Mode  :character   Mode  :character   Mode  :character   Median :1079.0
##                                                           Mean   :1156.5
##                                                           3rd Qu.:1382.5
##                                                           Max.   :5095.0
##    X2ndFlrSF     LowQualFinSF        GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :   0.000   Min.   : 407   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:   0.000   1st Qu.:1118   1st Qu.:0.0000
##  Median :   0   Median :   0.000   Median :1432   Median :0.0000
##  Mean   : 326   Mean   :   3.543   Mean   :1486   Mean   :0.4339
##  3rd Qu.: 676   3rd Qu.:   0.000   3rd Qu.:1721   3rd Qu.:1.0000
##  Max.   :1862   Max.   :1064.000   Max.   :5095   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   :0.06511   Mean   :1.571   Mean   :0.3777   Mean   :2.854
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :4.000   Max.   :2.0000   Max.   :6.000
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
```

```
##  Min.   :0.000   Length:1459        Min.   : 3.000   Length:1459
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.042                      Mean   : 6.385
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :2.000                      Max.   :15.000
##    Fireplaces     FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.0000   Length:1459        Length:1459        Min.   :   0
##  1st Qu.:0.0000   Class :character   Class :character   1st Qu.:1956
##  Median :0.0000   Mode  :character   Mode  :character   Median :1977
##  Mean   :0.5812                                        Mean   :1872
##  3rd Qu.:1.0000                                        3rd Qu.:2001
##  Max.   :4.0000                                        Max.   :2207
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1459        Min.   :0.000   Min.   :   0.0   Length:1459
##  Class :character   1st Qu.:1.000   1st Qu.: 317.5   Class :character
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.765   Mean   : 472.4
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :5.000   Max.   :1488.0
##   GarageCond         PavedDrive          WoodDeckSF       OpenPorchSF
##  Length:1459        Length:1459         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 : 28.00
##                                         Mean   :  93.17   Mean   : 48.31
##                                         3rd Qu.: 168.00   3rd Qu.: 72.00
##                                         Max.   :1424.00   Max.   :742.00
##  EnclosedPorch       X3SsnPorch       ScreenPorch        PoolArea
##  Min.   :   0.00   Min.   :  0.000   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :   0.00   Median :  0.000   Median :  0.00   Median :  0.000
##  Mean   :  24.24   Mean   :  1.794   Mean   : 17.06   Mean   :  1.744
##  3rd Qu.:   0.00   3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :1012.00   Max.   :360.000   Max.   :576.00   Max.   :800.000
##     PoolQC             Fence           MiscFeature          MiscVal
##  Length:1459        Length:1459        Length:1459        Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    0.00
##                                                           Mean   :   58.17
##                                                           3rd Qu.:    0.00
##                                                           Max.   :17000.00
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1459        Length:1459
##  1st Qu.: 4.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
##  Mean   : 6.104   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
```

Below are the contents of the data_description.txt file. It briefly describes what each column represents and what values it can take on. It is important to be familiar with this as we can use this information to decide what factors we want to use to predict the sale price of a house.

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

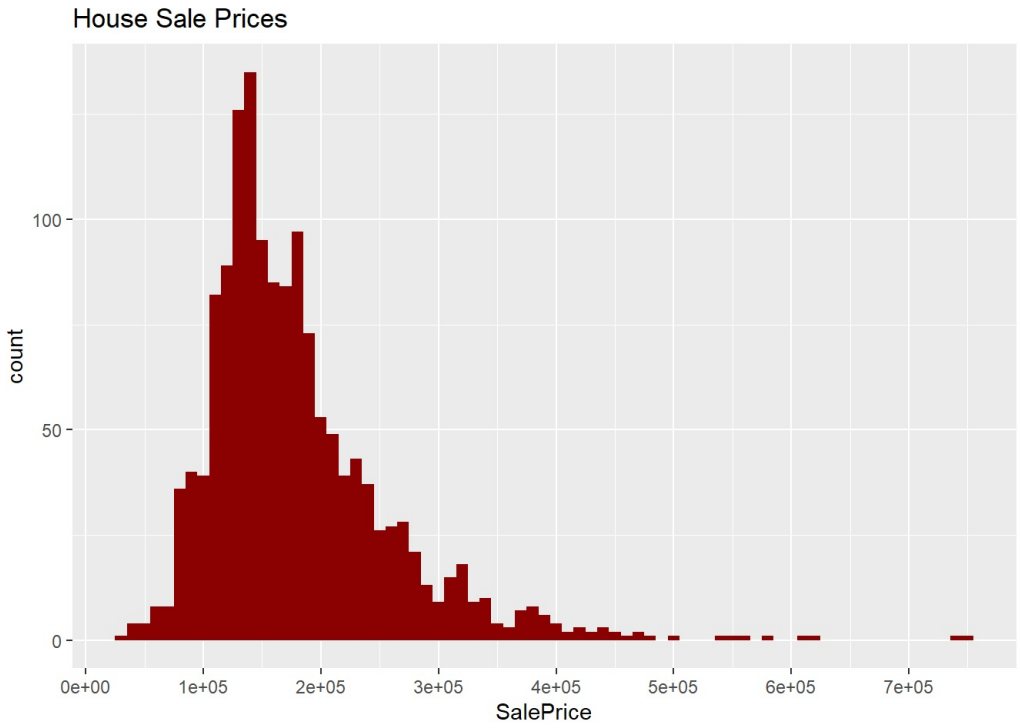
```
WD   Warranty Deed - Conventional
CWD  Warranty Deed - Cash
VWD  Warranty Deed - VA Loan
New  Home just constructed and sold
COD  Court Officer Deed/Estate
Con  Contract 15% Down payment regular terms
ConLw    Contract Low Down payment and low interest
ConLI    Contract Low Interest
ConLD    Contract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

SaleCondition: Condition of sale

```
   Normal   Normal Sales
   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)
```

We can plot a histogram of the frequency of the sale prices of houses. The x axis will represent the sale price, and the y axis will represent the number of times a certain sale price shows up in the data.

```
ggplot(data=train, aes(x=SalePrice)) +
  geom_histogram(fill="red4", binwidth = 10000) +
  scale_x_continuous(breaks= seq(0, 800000, by=100000)) +
  labs(title="House Sale Prices")
```

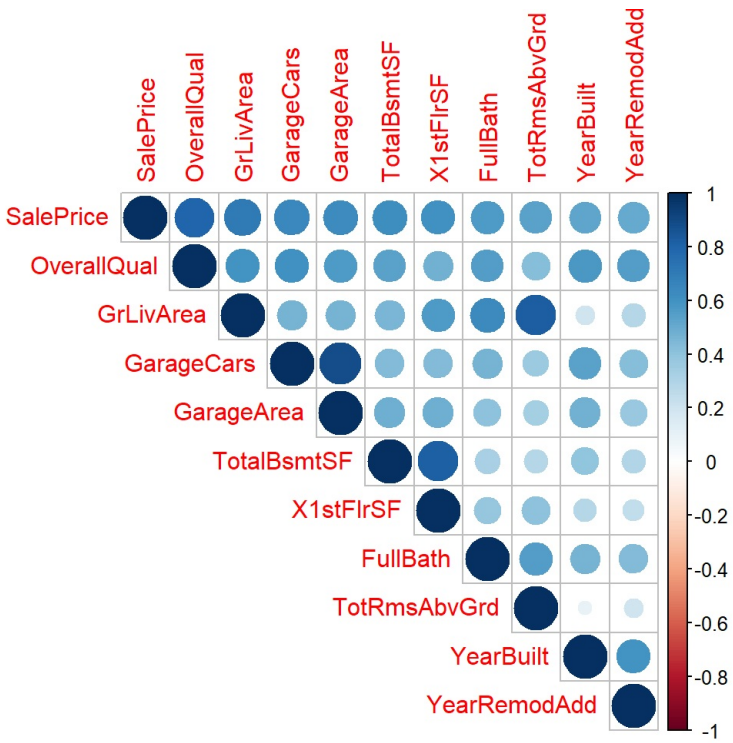


Upon plotting this, we see that the data is right-skewed. From this, we come to the conclusion that less expensive homes are more frequently bought.

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.

Here, we are creating a correlation matrix with SalePrice. The point of this is to evaluate which dependent variables are the most important in predicting the sale price for a house.

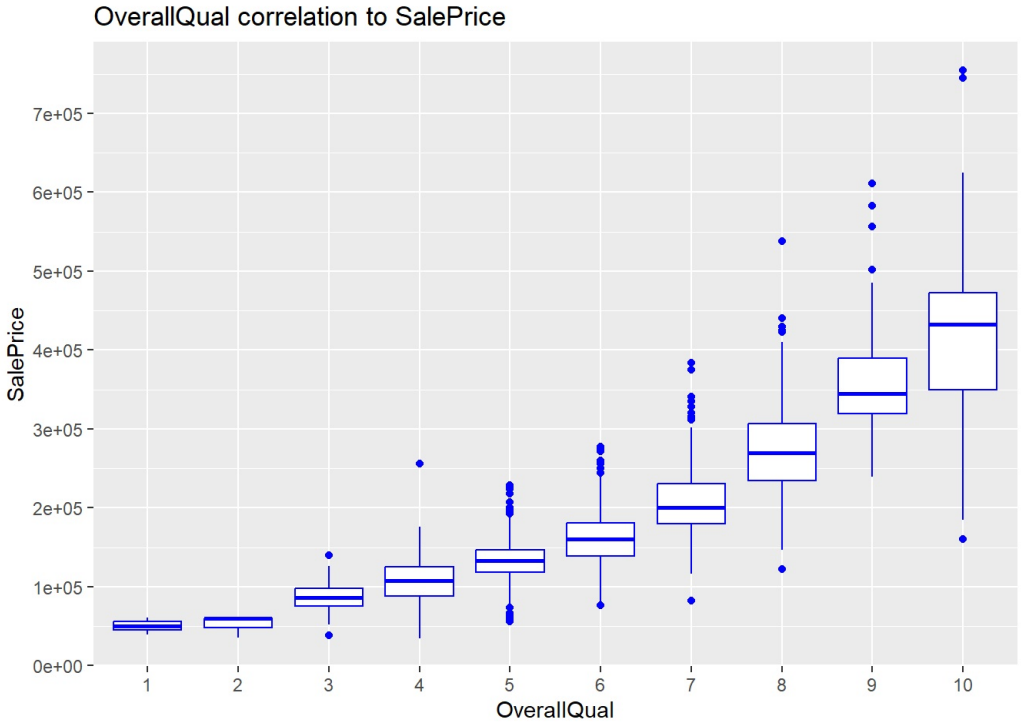
```
num_vars <- which(sapply(train, is.numeric)) #which columnns from train are numeric
num_vars_colnames <- data.table(names(num_vars)) #column names of the numeric variables

train_num_vars <- train[, num_vars] #new data frame containing only numeric variables from train
cor_num_vars <- cor(train_num_vars, use="pairwise.complete.obs") #correlations of all numeric variables, this is
the correlation matrix

#sorting the correlations in decreasing order (largest to smallest)
cor_sorted <- as.matrix(sort(cor_num_vars[,'SalePrice'], decreasing = TRUE))
#select variables with high correlations. corr>.5
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] #this is a matrix that contains only the correlations greater th
an .5

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



As seen in the correlation plot above, OverallQual has the highest correlation at ~0.8. OverallQual is the rate of the finish and material of the house overall from 1-10, so it makes sense to think that quality would affect price.

To confrim what we concluded in the previous step, we created a boxplot of the correlation.

```
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")
```



With OverallQual on the x axis and Sale Price on the y axis, it is apparent that as the quality rating increases, the sale price increases. The box plots exhibit the variation in price at each level of the rating, but consistently, the mean sale price increases as the rating increases.

Modeling

As a preparation for modeling, we first partition the train dataset. training contains 10% of the data and testing contains the remaining 90%.

```
set.seed(123)

outcome <- train$SalePrice

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

Simple Linear Regression Model

This model simply uses just OverallQual to predict the sale price of a house.

```
# 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)

RMSE_table <- data_frame(Method = "Regression model using OverallQual effect",
                                RMSE = model_rmse)

RMSE_table
```

Method <chr>
Regression model using OverallQual effect

1 row | 1-1 of 2 columns

Multiple Linear Regression Model

Multiple linear regression takes more than one attribute into account. Going back to our correlation calculations, we can determine the highest 10 correlations and use them in our model.

```
set.seed(57)

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)

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

RMSE_table
```

Method <chr>	
Regression model using OverallQual effect	
Regression model using top-10-effect	

2 rows | 1-1 of 2 columns

Backward Elimination Model

In this model we use backwards elimination. It works by iteratively removing features that do not contribute to the prediction or only barely contribute.

```
set.seed(57)

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

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)

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

RMSE_table
```

Method <chr>
Regression model using OverallQual effect
Regression model using top-10-effect
Regression model using backward elimination

3 rows | 1-1 of 2 columns

Random Forest Regression Model

In this model, we will basically use a decision tree framework. What we will do here is create random decision trees using our training data and average their results to gain a new result. This technique often leads to strong predictions.

```
set.seed(57)
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)
RMSE_table <- rbind(RMSE_table,
 data_frame(Method = "Random Forest regression model",
 RMSE = model_rmse))
RMSE_table
```

Method <chr>
Regression model using OverallQual effect
Regression model using top-10-effect
Regression model using backward elimination
Random Forest regression model

4 rows | 1-1 of 2 columns

Conclusion

```
RMSE_table
```

Method <chr>
Regression model using OverallQual effect
Regression model using top-10-effect
Regression model using backward elimination
Random Forest regression model

4 rows | 1-1 of 2 columns

We can see that the more we took into account, the smaller the RMSE became. While OverallQual obviously heavily influences the sale price of a home, it was made clear while working on this project that a more accurate estimate can be found if other factors are considered. The more factors that are considered, the closer the estimate is to the true value.