HarvardX - PH125.9x: Data Science - Housing Price Project

Maria Lucia Cornejo Quenaya

July 27,2021

Contents

1 PROJECT OVERVIEW: 2

1.1 INTRODUCTION . . . 2

1.2 OBJECTIVE . . . 2

1.3 THE DATASET . . . 2

2 THE HOUSING PRICE PROJECT: 3

2.1 DATA LOADING . . . 3

2.2 CREATING TRAIN AND TEST SETS . . . 6

2.3 DATA ANALYSIS AND VISUALIZATIONS . . . 8

2.4 PREDICTIVE MODELING . . . 17

2.4.1 MODEL 1: LINEAR REGRESION MODEL . . . 17

2.4.2 MODEL 2: CART . . . 18

2.4.3 MODEL 3: RANDOM FOREST . . . 19

3 CONCLUSIONS 23

4 BIBLIOGRAPHY 24

5 ENVIRONMENT 25

1 PROJECT OVERVIEW:

1.1 INTRODUCTION

In this project, data from residential housing in Ames, Iowa will be analyzed. The data contains characteristics and prices of these homes, which will be used to determine which variables are most influential in establishing the price of a home, and based on this we will develop and evaluate predictive models that will allow us to predict future prices.

A prediction model such as this one would be very valuable for real estate agents, who could make use of the information provided on a day-to-day basis.

1.2 OBJECTIVE

The objective of this project is to predict the price of residential homes in Ames, Iowa. As well as to define which are the characteristics that influence the price of a house. The accuracy function will be used to determine measures such as: Mean Error, Root Mean Squared Error, Mean Absolute Error, Mean Percentage Error, and Mean Absolute Percentage Error.

1.3 THE DATASET

The dataset has 79 explanatory variables describing most aspects of residential housing in Ames, Iowa. There are a total of 1,460 records. We found 43 factor variables, such as MSZoning, Street, LotShape. Also, 37 integer variables, some of them are: Id, MSSubClass, LotFrontage. This was chosen from the different datasets published in Kaggle.

2 THE HOUSING PRICE PROJECT:

In this section, we will see the development of the project. From data loading, creation of training, testing, and validation data sets, creation and evaluation of prediction models using RMSE, to the final validation of the model with the lowest RMSE.

2.1 DATA LOADING

The first thing to do is to import the data from the csv file. We found that each record in this dataset has a null value, which we must address later, as this could affect our model.

```
#Loading data
getwd()
```

```
## [1] "D:/Biblioteca/Documents"
```

```
train_data <- read.csv("train.csv")

#Checking missing data
missing_rows <- train_data[!complete.cases(train_data),]
head(missing_rows)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4  4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
## 5  5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6  6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
```

```
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
## 4       961       756            0      1717            1            0        1
## 5      1145      1053            0      2198            1            0        2
## 6       796       566            0      1362            1            0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
## 4        0            3            1          Gd            7        Typ
## 5        1            4            1          Gd            9        Typ
## 6        1            1            1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0       <NA>     Attchd        2003          RFn          2
## 2          1         TA     Attchd        1976          RFn          2
## 3          1         TA     Attchd        2001          RFn          2
## 4          1         Gd     Detchd        1998          Unf          3
## 5          1         TA     Attchd        2000          RFn          3
## 6          0       <NA>     Attchd        1993          Unf          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
## 4        642         TA         TA          Y          0          35
## 5        836         TA         TA          Y        192          84
## 6        480         TA         TA          Y         40          30
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
## 4           272          0           0        0   <NA>  <NA>        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0        320           0        0   <NA> MnPrv        Shed
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
## 4       0      2   2006       WD       Abnorml    140000
## 5       0     12   2008       WD        Normal    250000
## 6     700     10   2009       WD        Normal    143000
```

```
nrow(missing_rows)
```

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

2.2 CREATING TRAIN AND TEST SETS

We must now select the variables that would have the greatest impact on housing prices. And thus also, construct a subset of training data for prediction.

```
#############################################################################################
#CREATING TRAIN AND TEST SETS                                                               #
#############################################################################################
#Building subset of train dataset for prediction.
#Showing all variable names
variable_names <- names(train_data)
variable_names
```

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

```
#Selecting important variables by creating a vector that contains variable names
selected_variables <- c('Id','MSZoning','Utilities', 'Neighborhood','BldgType','HouseStyle',
                'OverallQual','OverallCond','YearBuilt', 'ExterQual','ExterCond',
                'BsmtQual','BsmtCond','TotalBsmtSF','Heating','HeatingQC',
                'CentralAir','Electrical','GrLivArea','BedroomAbvGr','KitchenAbvGr',
                'KitchenQual','TotRmsAbvGrd','Functional','Fireplaces','FireplaceQu',
                'GarageArea','GarageQual','GarageCond','OpenPorchSF','PoolArea',
                'Fence','MoSold','YrSold','SaleType','SaleCondition','SalePrice')

#Building subset of train dataset that is used for prediction
selected_train <- train_data[,selected_variables]
head(selected_train)
```

```
##   Id MSZoning Utilities Neighborhood BldgType HouseStyle OverallQual
## 1  1       RL    AllPub      CollgCr     1Fam     2Story           7
## 2  2       RL    AllPub      Veenker     1Fam     1Story           6
## 3  3       RL    AllPub      CollgCr     1Fam     2Story           7
## 4  4       RL    AllPub      Crawfor     1Fam     2Story           7
```

```
## 5  5       RL    AllPub      NoRidge     1Fam     2Story           8
## 6  6       RL    AllPub      Mitchel     1Fam     1.5Fin           5
##   OverallCond YearBuilt ExterQual ExterCond BsmtQual BsmtCond TotalBsmtSF
## 1           5      2003        Gd        TA       Gd       TA         856
## 2           8      1976        TA        TA       Gd       TA        1262
## 3           5      2001        Gd        TA       Gd       TA         920
## 4           5      1915        TA        TA       TA       Gd         756
## 5           5      2000        Gd        TA       Gd       TA        1145
## 6           5      1993        TA        TA       Gd       TA         796
##   Heating HeatingQC CentralAir Electrical GrLivArea BedroomAbvGr KitchenAbvGr
## 1    GasA        Ex          Y      SBrkr      1710            3            1
## 2    GasA        Ex          Y      SBrkr      1262            3            1
## 3    GasA        Ex          Y      SBrkr      1786            3            1
## 4    GasA        Gd          Y      SBrkr      1717            3            1
## 5    GasA        Ex          Y      SBrkr      2198            4            1
## 6    GasA        Ex          Y      SBrkr      1362            1            1
##   KitchenQual TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageArea
## 1          Gd            8        Typ          0        <NA>        548
## 2          TA            6        Typ          1          TA        460
## 3          Gd            6        Typ          1          TA        608
## 4          Gd            7        Typ          1          Gd        642
## 5          Gd            9        Typ          1          TA        836
## 6          TA            5        Typ          0        <NA>        480
##   GarageQual GarageCond OpenPorchSF PoolArea Fence MoSold YrSold SaleType
## 1         TA         TA          61        0  <NA>      2   2008       WD
## 2         TA         TA           0        0  <NA>      5   2007       WD
## 3         TA         TA          42        0  <NA>      9   2008       WD
## 4         TA         TA          35        0  <NA>      2   2006       WD
## 5         TA         TA          84        0  <NA>     12   2008       WD
## 6         TA         TA          30        0 MnPrv     10   2009       WD
##   SaleCondition SalePrice
## 1        Normal    208500
## 2        Normal    181500
## 3        Normal    223500
## 4       Abnorml    140000
## 5        Normal    250000
## 6        Normal    143000
```

2.3 DATA ANALYSIS AND VISUALIZATIONS

We will begin by analyzing the structure of the data set, in order to become familiar with it.

```
##        Id           MSZoning          Utilities         Neighborhood
##  Min.   :   1.0   Length:1460        Length:1460        Length:1460
##  1st Qu.: 365.8   Class :character   Class :character   Class :character
##  Median : 730.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 730.5
##  3rd Qu.:1095.2
##  Max.   :1460.0
##    BldgType          HouseStyle         OverallQual      OverallCond
##  Length:1460        Length:1460        Min.   : 1.000   Min.   :1.000
##  Class :character   Class :character   1st Qu.: 5.000   1st Qu.:5.000
##  Mode  :character   Mode  :character   Median : 6.000   Median :5.000
##                                        Mean   : 6.099   Mean   :5.575
##                                        3rd Qu.: 7.000   3rd Qu.:6.000
##                                        Max.   :10.000   Max.   :9.000
##    YearBuilt     ExterQual          ExterCond          BsmtQual
##  Min.   :1872   Length:1460        Length:1460        Length:1460
##  1st Qu.:1954   Class :character   Class :character   Class :character
##  Median :1973   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1971
##  3rd Qu.:2000
##  Max.   :2010
##    BsmtCond          TotalBsmtSF       Heating           HeatingQC
##  Length:1460        Min.   :   0.0   Length:1460        Length:1460
##  Class :character   1st Qu.: 795.8   Class :character   Class :character
##  Mode  :character   Median : 991.5   Mode  :character   Mode  :character
##                     Mean   :1057.4
##                     3rd Qu.:1298.2
##                     Max.   :6110.0
##   CentralAir         Electrical          GrLivArea     BedroomAbvGr
##  Length:1460        Length:1460        Min.   : 334   Min.   :0.000
##  Class :character   Class :character   1st Qu.:1130   1st Qu.:2.000
##  Mode  :character   Mode  :character   Median :1464   Median :3.000
##                                        Mean   :1515   Mean   :2.866
##                                        3rd Qu.:1777   3rd Qu.:3.000
##                                        Max.   :5642   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          GarageArea      GarageQual
##  Min.   :0.000   Length:1460        Min.   :   0.0   Length:1460
##  1st Qu.:0.000   Class :character   1st Qu.: 334.5   Class :character
##  Median :1.000   Mode  :character   Median : 480.0   Mode  :character
##  Mean   :0.613                      Mean   : 473.0
##  3rd Qu.:1.000                      3rd Qu.: 576.0
##  Max.   :3.000                      Max.   :1418.0
##   GarageCond         OpenPorchSF        PoolArea          Fence
```

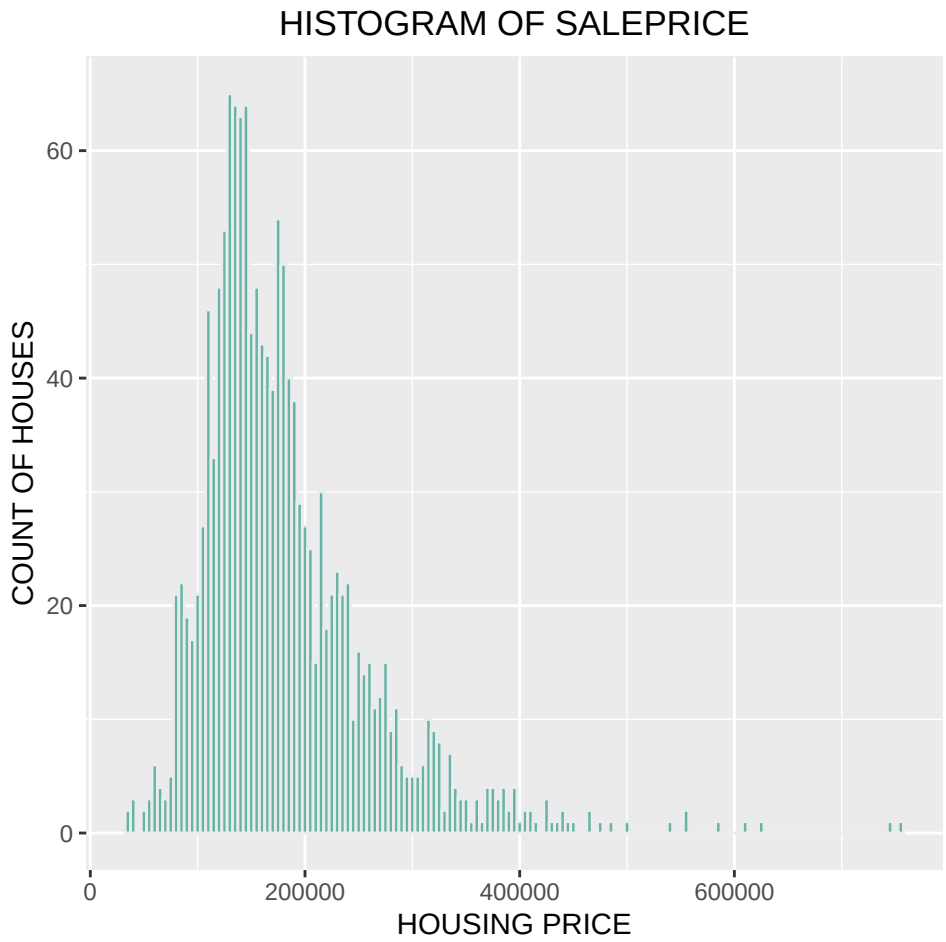
```
##  Length:1460        Min.   :  0.00   Min.   :  0.000   Length:1460
##  Class :character   1st Qu.:  0.00   1st Qu.:  0.000   Class :character
##  Mode  :character   Median : 25.00   Median :  0.000   Mode  :character
##                     Mean   : 46.66   Mean   :  2.759
##                     3rd Qu.: 68.00   3rd Qu.:  0.000
##                     Max.   :547.00   Max.   :738.000
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1460        Length:1460
##  1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
##  Mean   : 6.322   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
```

The variable: SalePrice is our target variable and also the dependent variable for the prediction.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

We will now proceed to analyze the data using graphs. Below is a histogram showing the distribution of our target variable - Sales Price, which is skewed to the right. For this reason, a logarithmic term of Sales Price must be generated for the linear regression.

```
library(ggplot2)
#Histogram of SalePrice's distribution
options(scipen=10000)
ggplot(selected_train, aes(x = SalePrice, fill = ..count..)) +
  geom_histogram(binwidth = 5000,fill="#69b3a2", color="#e9ecef") +
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title = "HISTOGRAM OF SALEPRICE", y = "COUNT OF HOUSES", x = "HOUSING PRICE")
```

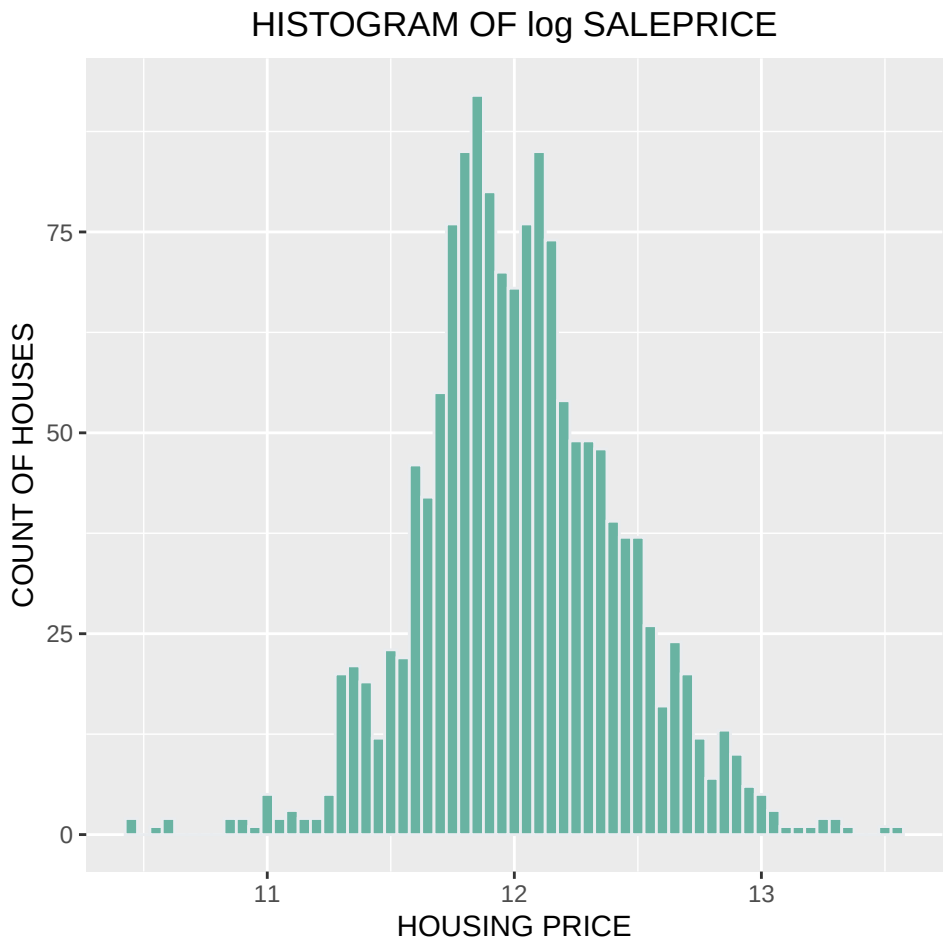



```
#Log term of SalePrice
selected_train$lSalePrice <- log(selected_train$SalePrice)
```

After correction, the new variable lSalePrice presents a normal distribution, which we can observe in the following histogram.

```
library(ggplot2)
#Histogram of log SalePrice distribution
ggplot(selected_train, aes(x = lSalePrice, fill = ..count..)) +
  geom_histogram(binwidth = 0.05,fill="#69b3a2", color="#e9ecef") +
```

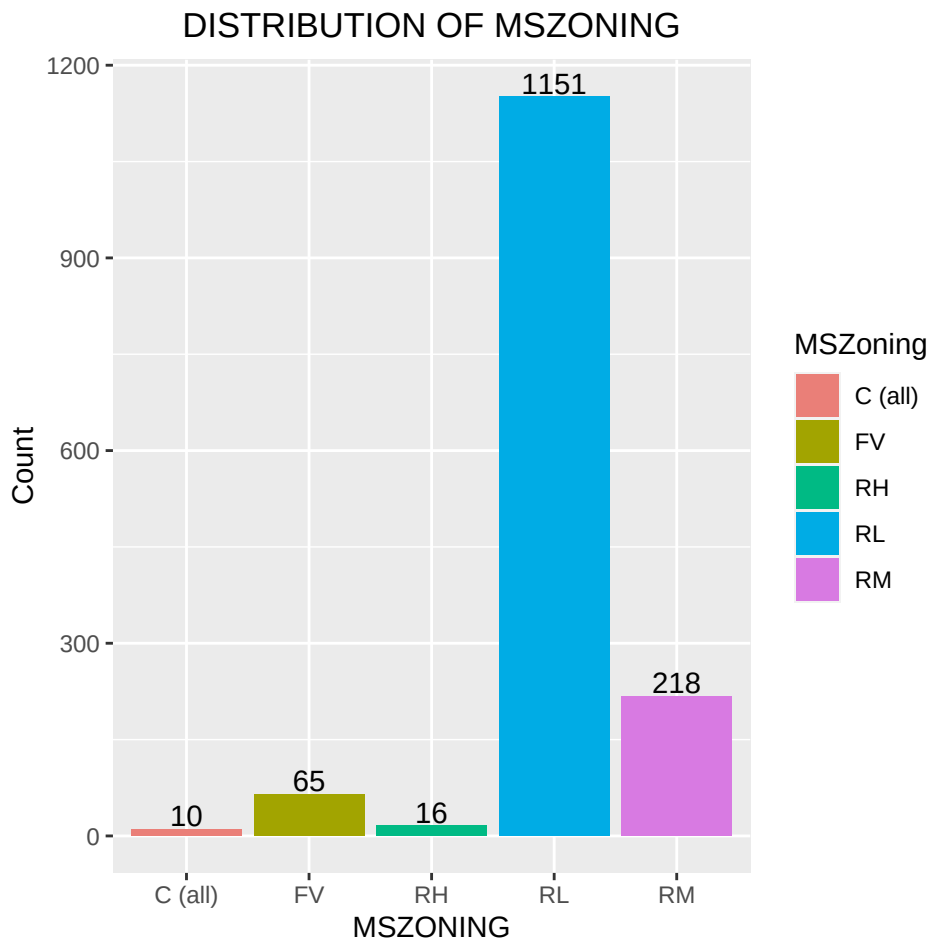
```
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title = "HISTOGRAM OF log SALEPRICE", y = "COUNT OF HOUSES", x = "HOUSING PRICE")
```



If we talk about the price of the house, the value of the house is generally related to two types of elements: internal and external. The internal elements refer to key characteristics of the house itself, for example the total surface area or the number of rooms. And the external elements, the environment is one of the key factors.

The variable indicating the housing environment in the data set would be MSZoning. We will now analyze the values of this variable:

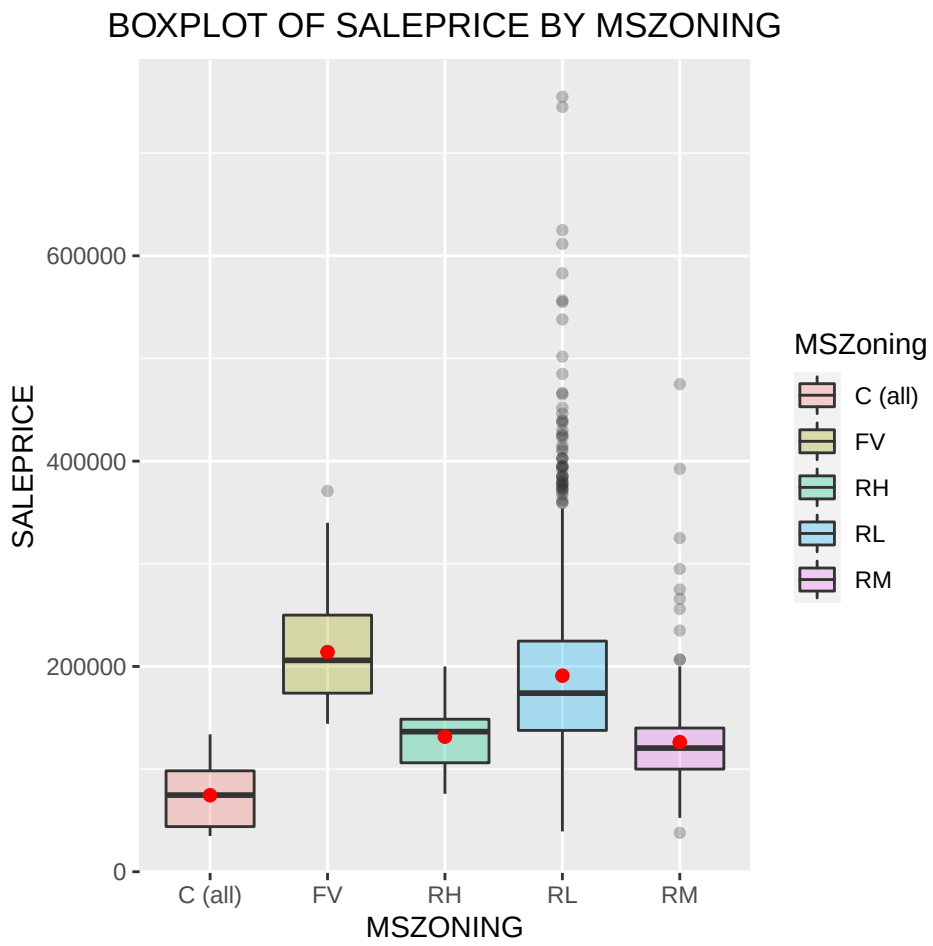
```
library(ggplot2)
#Counting house by MSZoning
options(repr.plot.width=5, repr.plot.height=5)
ggplot(selected_train, aes(x = MSZoning, fill = MSZoning )) +
  geom_bar()+scale_fill_hue(c = 80)+
  theme(plot.title = element_text(hjust = 0.5),legend.position="right", legend.background = element_rec
  (size=0.5))+geom_text(stat='count',aes(label=..count..),vjust=-0.15)+
  labs(title = "DISTRIBUTION OF MSZONING",y="Count",x="MSZONING")
```




```
##
## C (all)      FV      RH      RL      RM
##      10      65      16    1151     218
```

We will now analyze the relationship between MSZoning and our target variable SalePrice. Then, we will add our target variable to the analysis. How is the housing price in each category? We will use a boxplot to show the distribution of prices in each MSZoning.

```
#Boxplot of SalePrice by MSZoning, adding average value of SalePrice as red point
library(ggplot2)
ggplot(selected_train, aes(x=MSZoning, y=SalePrice, fill=MSZoning)) +
  geom_boxplot(alpha=0.3) +
  stat_summary(fun=mean, geom="point", shape=20, size=3, color="red", fill="red")+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title = "BOXPLOT OF SALEPRICE BY MSZONING",y="SALEPRICE",x="MSZONING")
```



The boxplot shows the distribution of SalePrice by MSZoning. The houses located in "Floating Village Residential" have the highest average sale price, and then followed by "Residential Low Density". While "Commercial" sales have the houses with the lowest average sale price.

Somewhat oddly, the commercial or urban area has the lowest average sales price, while the rural areas have the highest price. It could be that the selling price is also related to the size of the houses. The variable indicating the size of houses in the dataset is called GrLivArea. We will then proceed to analyze it.

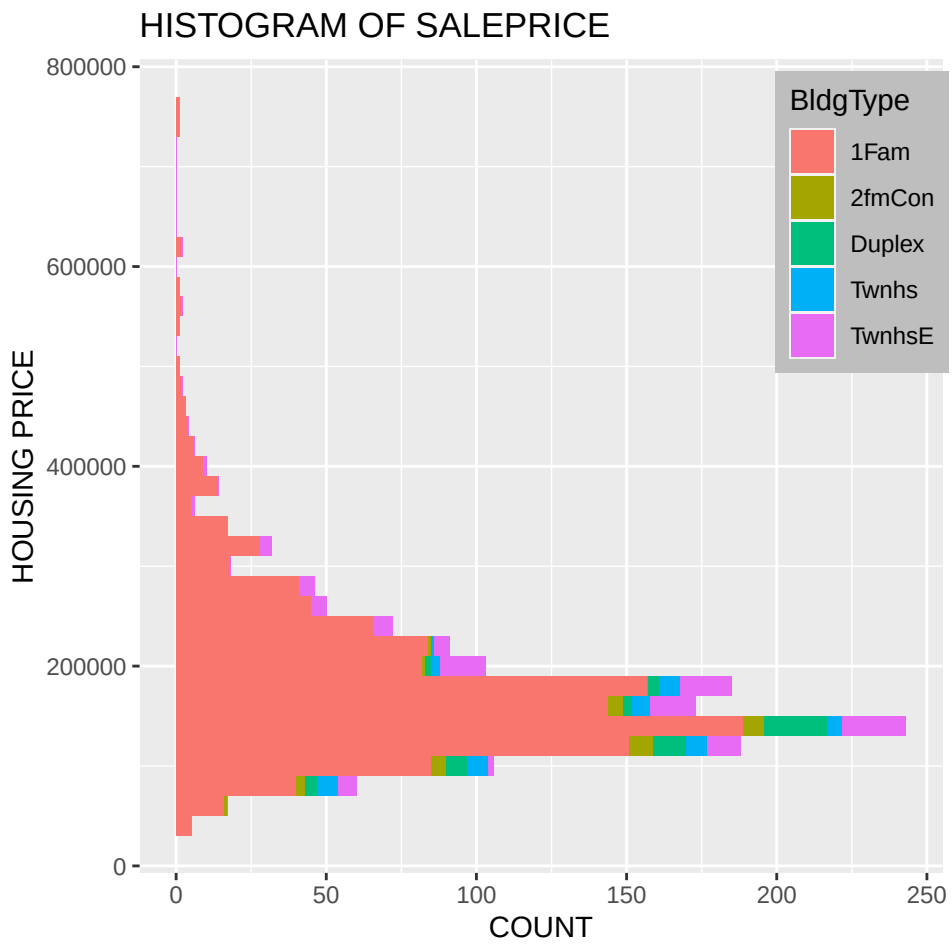
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

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

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

```
##   MSZoning     size
## 1  C (all) 1191.400
## 2       FV 1574.538
## 3       RH 1510.125
## 4       RL 1551.646
## 5       RM 1322.073
```

We can confirm the statement we made earlier. Now we will relate our variable SalePrice, with respect to the type of housing, that is to say with the variable: BldfType.

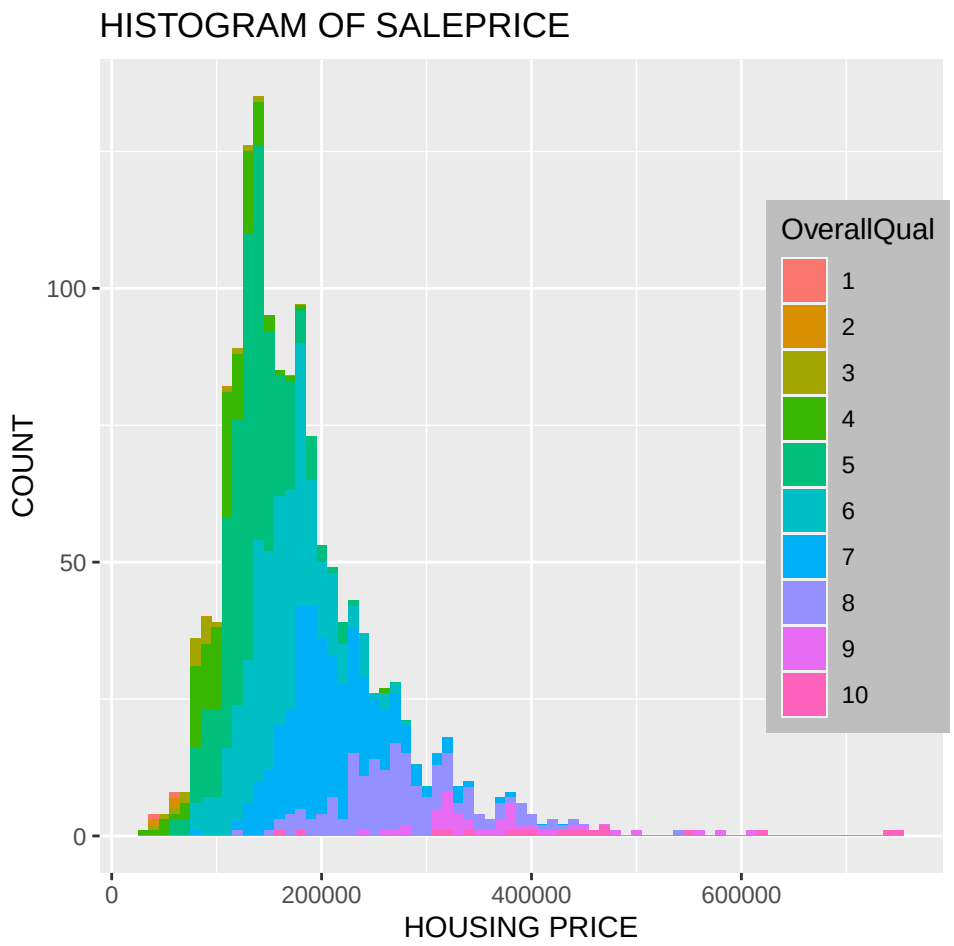
```
##   BldgType Total Max_price Min_price
## 1     1Fam  1220    755000     34900
## 2   2fmCon    31    228950     55000
## 3   Duplex    52    206300     82000
## 4    Twnhs    43    230000     75000
## 5   TwnhsE   114    392500     75500
```

```
library(ggplot2)
#Distribution of SalePrice by BldfType
ggplot(selected_train, aes(SalePrice)) +
  geom_histogram(aes(fill = BldgType), position = position_stack(reverse = TRUE), binwidth = 20000) +
  coord_flip() +
  theme(legend.position=c(0.9,0.8),
  legend.background = element_rect(fill="grey", size=0.5))+
  labs(title = "HISTOGRAM OF SALEPRICE",y="COUNT",x="HOUSING PRICE")
```



Thanks to the graph, we see that most of the prices of single-family houses, range between 50000 and 300000. While two-family houses, duplexes, semi-detached houses and interior semi-detached houses, generally their prices are between 75,000 and 21,000 euros. The highest and the lowest price correspond to the type of detached single-family house We will now look at the relationship between our target variable and the variable that qualifies the material and overall finish of the house: OverallQual.

```
library(ggplot2)
#Distribution of SalePrice by OverallQual
ggplot(selected_train, aes(x = SalePrice,fill = as.factor(OverallQual))) +
  geom_histogram(position = "stack", binwidth = 10000) +
  scale_fill_discrete(name="OverallQual")+
  theme(legend.position=c(0.9,0.5),
  legend.background = element_rect(fill="grey",size=0.5))+
  labs(title = "HISTOGRAM OF SALEPRICE",y="COUNT",x="HOUSING PRICE")
```



The graph shows that the higher rate of overall quality, the higher house sale price.

2.4 PREDICTIVE MODELING

Predictive modeling is a mathematical approach to create a statistical model to forecast future behavior based on input test data. Steps involved in predictive modeling: - Algorithm Selection: When we have the structured dataset, and we want to estimate the continuous or categorical outcome then we use supervised machine learning methodologies like regression and classification techniques. When we have unstructured data and want to predict the clusters of items to which a particular input test sample belongs, we use unsupervised algorithms. An actual data scientist applies multiple algorithms to get a more accurate model.

-Train Model: After assigning the algorithm and getting the data handy, we train our model using the input data applying the preferred algorithm. It is an action to determine the correspondence between independent variables, and the prediction targets.

-Model Prediction: We make predictions by giving the input test data to the trained model. We measure the accuracy by using a cross-validation strategy or ROC curve which performs well to derive model output for test data.

Next, let's look at the different models that were developed and the accuracy evaluation of each one of them.

2.4.1 MODEL 1: LINEAR REGRESION MODEL

In the linear regression model, the relationships between the dependent and independent variables are expressed by an equation with coefficients. The objective of this model is to minimize the sum of the squared residuals. Sixteen variables were selected for this model.

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

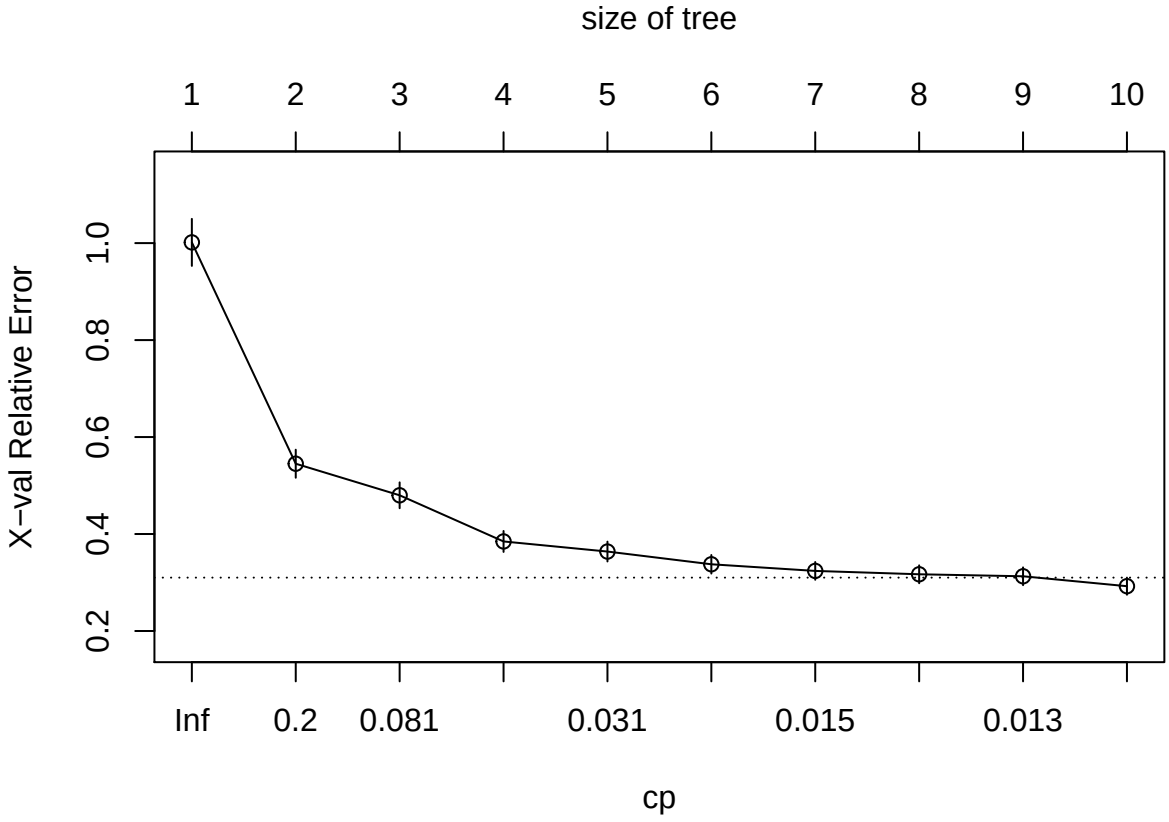
```
##
## Call:
## lm(formula = lSalePrice ~ . - SalePrice, data = model_linear_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98613 -0.07164  0.00209  0.08015  0.55020
##
## Coefficients:
##                 Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)  17.50305706  7.11375055   2.460              0.01402 *
## OverallQual   0.08057086  0.00575656  13.996 < 0.0000000000000002 ***
## OverallCond   0.05664296  0.00489313  11.576 < 0.0000000000000002 ***
## YearBuilt     0.00317718  0.00024216  13.120 < 0.0000000000000002 ***
## ExterCond2    0.02627048  0.01170750   2.244              0.02503 *
## TotalBsmtSF   0.00011153  0.00001344   8.301 0.000000000000000286 ***
## HeatingQC2   -0.01828236  0.00407563  -4.486 0.000007992540141142 ***
## CentralAir2   0.06343196  0.02300441   2.757              0.00592 **
## GrLivArea     0.00020261  0.00001946  10.414 < 0.0000000000000002 ***
## BedroomAbvGr -0.00455622  0.00848589  -0.537              0.59143
## KitchenAbvGr -0.06642318  0.02533811  -2.621              0.00887 **
## TotRmsAbvGrd  0.01726182  0.00623236   2.770              0.00570 **
## Fireplaces    0.06899560  0.00854558   8.074 0.000000000000001704 ***
## GarageArea    0.00023841  0.00002956   8.064 0.000000000000001832 ***
## OpenPorchSF   0.00001953  0.00007922   0.247              0.80529
```

```
## PoolArea     -0.00078143  0.00014053  -5.561 0.00000033400863871 ***
## YrSold       -0.00664527  0.00353924  -1.878              0.06069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1585 on 1151 degrees of freedom
## Multiple R-squared:  0.8467, Adjusted R-squared:  0.8446
## F-statistic: 397.3 on 16 and 1151 DF,  p-value: < 0.00000000000000022
```

```
##                    ME      RMSE       MAE        MPE      MAPE
## Test set 0.007261273 0.1538444 0.1075528 0.04271564 0.9029266
```

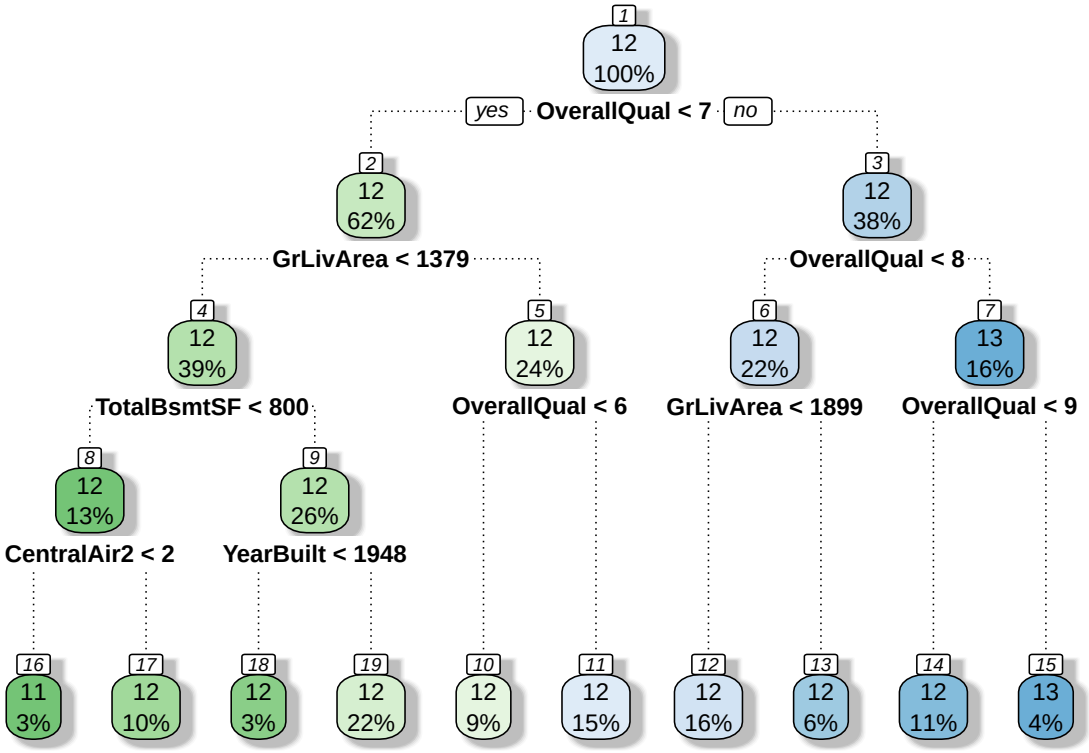
As we can see, the RMSE obtained in this first model is acceptable: 0.225640. Let's see if we can improve it.

2.4.2 MODEL 2: CART




```
##
## Regression tree:
## rpart(formula = lSalePrice ~ . - SalePrice, data = model_linear_train,
##     control = rpart.control(cp = 0.01))
##
## Variables actually used in tree construction:
## [1] CentralAir2 GrLivArea   OverallQual TotalBsmtSF YearBuilt
```

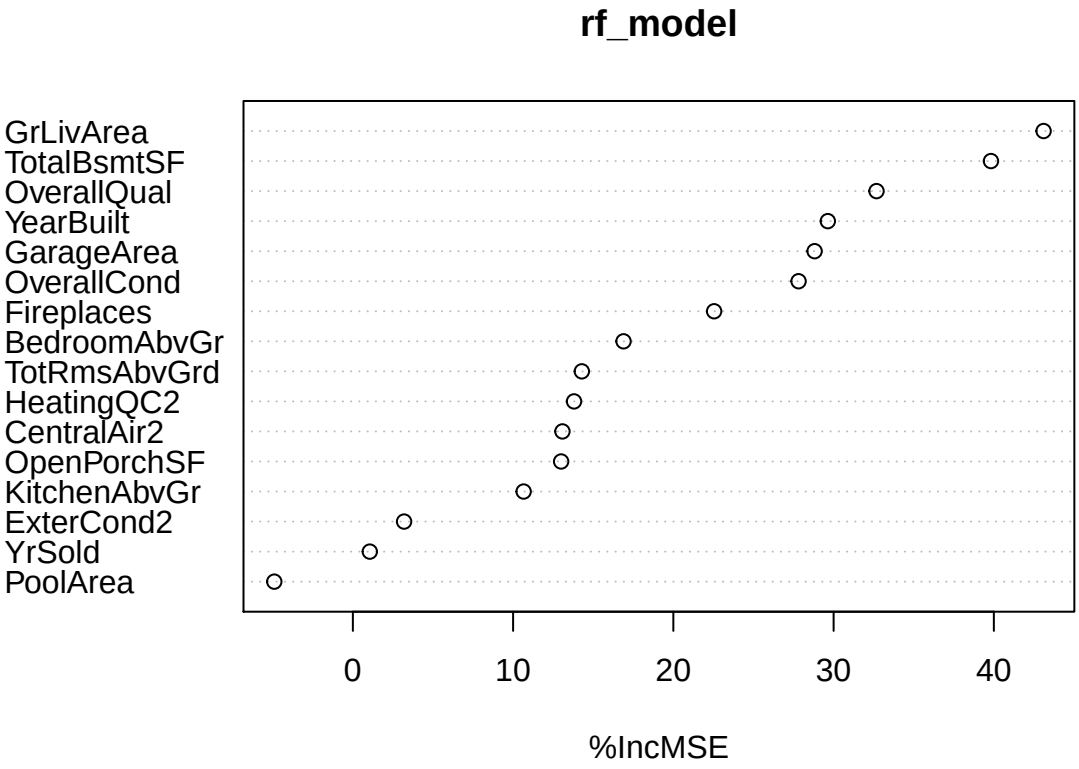
```
##
## Root node error: 188.62/1168 = 0.16149
##
## n= 1168
##
##          CP nsplit rel error  xerror     xstd
## 1  0.456916      0   1.00000 1.00151 0.048514
## 2  0.084271      1   0.54308 0.54500 0.028941
## 3  0.078437      2   0.45881 0.47987 0.026685
## 4  0.042483      3   0.38038 0.38467 0.021622
## 5  0.023327      4   0.33789 0.36382 0.020452
## 6  0.015481      5   0.31457 0.33757 0.019182
## 7  0.014402      6   0.29908 0.32393 0.018320
## 8  0.013422      7   0.28468 0.31695 0.018477
## 9  0.013194      8   0.27126 0.31284 0.018345
## 10 0.010000      9   0.25807 0.29241 0.017702
```



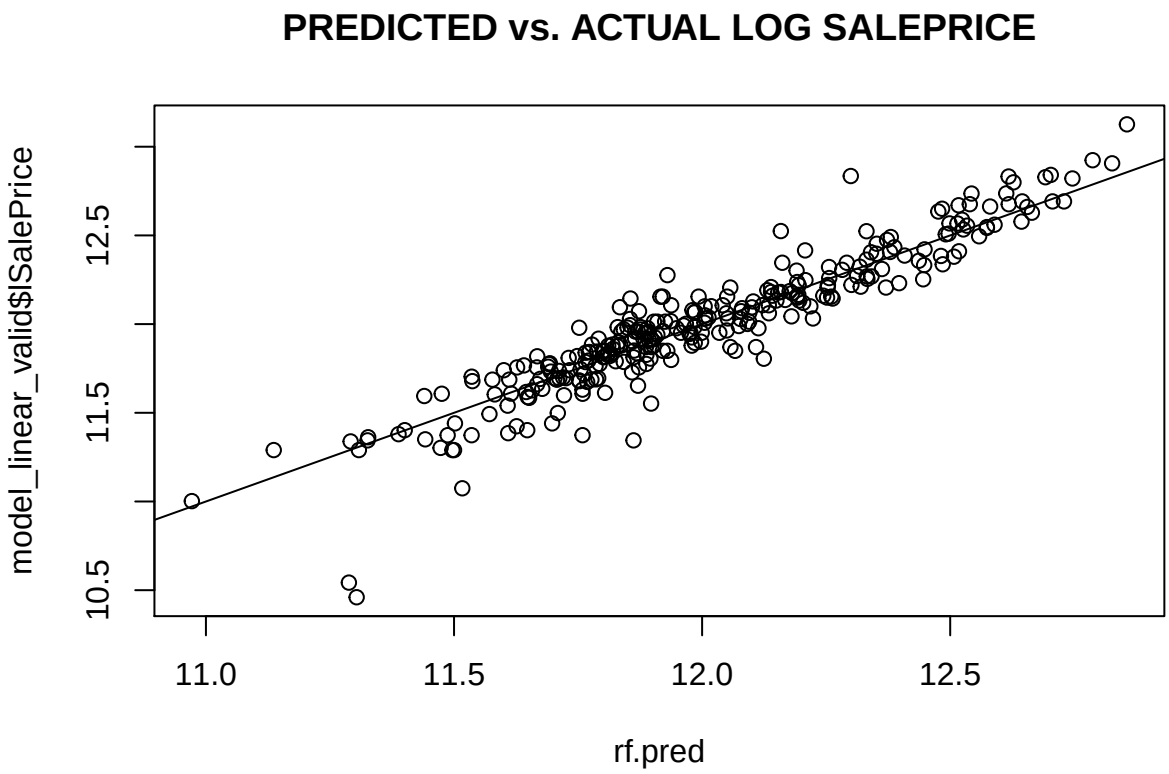
2.4.3 MODEL 3: RANDOM FOREST

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'
```

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

## The following object is masked from 'package:ggplot2':
##
##     margin
```




```
##                       ME      RMSE        MAE        MPE      MAPE
## Test set -0.0005328495 0.1383988 0.09484976 -0.0239557 0.7978943
```



The RMSE value obtained by using this model is 0.1641079.
RESULTS AND DISCUSSION Taking into account the RMSE value, we see that the third model using

Random Forest had the lowest RMSE, 0.1641079. This is the model that offers the highest accuracy.

3 CONCLUSIONS

1. Just as important as understanding the problem is understanding the data available to us. That is why we conducted an exploratory analysis, which through graphs, correlations and descriptive statistics allowed us to better understand what story the data are telling us. It also helps to estimate whether the data we have are sufficient, and relevant, to build a model.
2. We had to experiment with the training data in order to find the most effective and efficient method. This is very important, as it sometimes happens that apparently more modest models turn out to be better than extremely complex and versatile models. In general, the latter are very time-consuming and their results will not always be better than simple or modest models.
3. The last model based on Random Forest was the most effective, achieving the lowest RMSE.
4. The final RMSE value was 0.1641079.
5. As future work, the use of other popular methods could be evaluated. But, we must take into account the dynamics of the environment and the availability of information, in order to choose the most appropriate for our case.
6. Also, once we have the predicted valuations, we must make decisions based on this information and then use causal inference techniques to establish cause-effect relationships between our decisions and the events that occur after implementing them. Thus, we will understand whether the decision taken based on the predicted information has worked or not, and whether we should attribute success or failure to it.

4 BIBLIOGRAPHY

Wikipedia contributors. (2021, July 22). Linear regression. In Wikipedia, The Free Encyclopedia. Retrieved 10:25, July 27, 2021, from https://en.wikipedia.org/w/index.php?title=Linear_regression&oldid=1034938319

Wikipedia contributors. (2021, July 12). Random forest. In Wikipedia, The Free Encyclopedia. Retrieved 11:33, July 28, 2021, from https://en.wikipedia.org/w/index.php?title=Random_forest&oldid=1033187858

Wikipedia contributors. (2021, February 5). Classification Tree Method. In Wikipedia, The Free Encyclopedia. Retrieved 12:14, July 28, 2021, from https://en.wikipedia.org/w/index.php?title=Classification_Tree_Method&oldid=1004975321

5 ENVIRONMENT

```
print("Operating System:")
```

```
## [1] "Operating System:"
```

```
version
```

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          4
## minor          0.5
## year           2021
## month          03
## day            31
## svn rev        80133
## language       R
## version.string R version 4.0.5 (2021-03-31)
## nickname       Shake and Throw
```