home work 2

Blessing Ekereke

1/28/2020

1. Table of Content

2. Objective of Home work

- Data Exploration
- data cleaning
- Missing Values computation
- Data Transformation
- Summary Statistics Calculation
- Modelling

Task 1 Solution

- **Answer in your own words (!) with one sentence:**
- **1. What is the difference between supervised and un-supervised learning?**
- Ans: In Supervised learning, for each observation of the predictor measurement(s) xi, i = 1, . . . , n there is an associated response measurement yi while in unsupervised learning, the observe measurements xi has no associated responses yi.
- **2. What is the difference between prediction and inference?**
- Ans: By Inference, we are interested in understanding the way the dependant variable is affected as the independent variable(s) changes while in prediction, we want to estimate/predict the value of the dependent variable based on the values of the independent variable(s)
- **3. What is the difference between classification and regression?**
- Ans: In classification the outcome are of discreet types could be binary like e.g Yes or No, 1 or 0 0r multi l-level like A,B,C,D while in Regression the outcome variable are of the continous type e.g 1,2000,33,4,56, 567
- **4. Why is it not a good idea to use a linear regression model to predict survival probabilities in the "Titanic" data set?**
- Ans: Linear regression are not robust for probability prediction as they produce negative estimates or estimates gretater than 1 for the outcome probability which are statistically incorrect because probabilities range from 0 - 1 and can never be negative(hence we have the logistic Regression model for this type of prediction).

3. Packages Installation

```
#install.packages('psych')
#install.packages('lemon')
#install.packages('Hmisc')
#install.packages('VIM')
#install.packages('tidyverse')
#install.packages('editrules')
#install.packages('deducorrect')
#install.packages("glmnet")
#install.packages('reshape')
#install.packages('gbm')
#install.packages('corrplot')
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.6.2
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 3.0-2
```

```
library(lemon)
```

```
## Warning: package 'lemon' was built under R version 3.6.2
```

```
knit_print.data.frame <- lemon_print
```

4. data reading

```
rm(list = ls())
data_desc <- read.delim('data_description.txt')
hptrain <- read.csv('hptrain.csv')
hptest <-  read.csv('hptest.csv')
data <- rbind(hptrain[,-81],hptest )
#this was a personal decision inorder to expediously clean both datasets
dim(data)
```

```
## [1] 2919   80
```

5. data Transformation

There is alot of cleaning to be done for the respective attributes of the instances. * First we start by converting the dates into age

```
data$YearBuilt <- 2020 - data$YearBuilt
#This would be the age of the listing as at today
data$YearRemodAdd <- 2020 - data$YearRemodAdd
#This would be the  how long ago from today this house was remoded
data$YrSold <- 2020 - data$YrSold
#How long agao the house was sold
data$GarageYrBlt <- 2020 - data$GarageYrBlt
#How long ago the garage was built
```

- Some of the factor attributes should also be converted from qualitative levels to quantitative levels. We could use the matrix function to convert them to a dummy variables at the onset of building a reg model but that would not be accurate as each levels are qualitative scale. a better option would be converting to a numeric scale

```r
levels(data$ExterQual) <- c(levels(data$ExterQual),1,2,3,4,5)
data$ExterQual[data$ExterQual=='Ex'] <- 5
data$ExterQual[data$ExterQual=='Gd'] <- 4
data$ExterQual[data$ExterQual=='TA'] <- 3
data$ExterQual[data$ExterQual=='Fa'] <- 2
data$ExterQual[data$ExterQual=='Po'] <- 1
data$ExterQual <- droplevels(data$ExterQual)
data$ExterQual <- as.numeric(data$ExterQual)

levels(data$ExterCond) <- c(levels(data$ExterCond),1,2,3,4,5)
data$ExterCond[data$ExterCond=='Ex'] <- 5
data$ExterCond[data$ExterCond=='Gd'] <- 4
data$ExterCond[data$ExterCond=='TA'] <- 3
data$ExterCond[data$ExterCond=='Fa'] <- 2
data$ExterCond[data$ExterCond=='Po'] <- 1
data$ExterCond <- droplevels(data$ExterCond)
data$ExterCond <- as.numeric(data$ExterCond)

levels(data$BsmtQual) <- c(levels(data$BsmtQual),0,1,2,3,4,5)
data$BsmtQual[data$BsmtQual=='Ex'] <- 5
data$BsmtQual[data$BsmtQual=='Gd'] <- 4
data$BsmtQual[data$BsmtQual=='TA'] <- 3
data$BsmtQual[data$BsmtQual=='Fa'] <- 2
data$BsmtQual[data$BsmtQual=='Po'] <- 1
data$BsmtQual[data$BsmtQual=='NA'] <- 0
data$BsmtQual <- droplevels(data$BsmtQual)
data$BsmtQual<- as.numeric(data$BsmtQual)

levels(data$BsmtCond) <- c(levels(data$BsmtCond),0,1,2,3,4,5)
data$BsmtCond[data$BsmtCond=='Ex'] <- 5
data$BsmtCond[data$BsmtCond=='Gd'] <- 4
data$BsmtCond[data$BsmtCond=='TA'] <- 3
data$BsmtCond[data$BsmtCond=='Fa'] <- 2
data$BsmtCond[data$BsmtCond=='Po'] <- 1
data$BsmtCond[data$BsmtCond=='NA'] <- 0
data$BsmtCond <- droplevels(data$BsmtCond)
data$BsmtCond<- as.numeric(data$BsmtCond)

levels(data$BsmtExposure) <- c(levels(data$BsmtExposure),1,2,3,4,5)
data$BsmtExposure[data$BsmtExposure=='Gd'] <- 5
data$BsmtExposure[data$BsmtExposure=='Av'] <- 4
data$BsmtExposure[data$BsmtExposure=='Mn'] <- 3
data$BsmtExposure[data$BsmtExposure=='No'] <- 2
data$BsmtExposure[data$BsmtExposure=='NA'] <- 1
data$BsmtExposure <- droplevels(data$BsmtExposure)
data$BsmtExposure<- as.numeric(data$BsmtExposure)# the numerical scale changed it

levels(data$BsmtFinType1) <- c(levels(data$BsmtFinType1),1,2,3,4,5,6,7)
data$BsmtFinType1[data$BsmtFinType1=='GLQ'] <- 7
data$BsmtFinType1[data$BsmtFinType1=='ALQ'] <- 6
data$BsmtFinType1[data$BsmtFinType1=='BLQ'] <- 5
data$BsmtFinType1[data$BsmtFinType1=='Rec'] <- 4
data$BsmtFinType1[data$BsmtFinType1=='LwQ'] <- 3
```

```r
data$BsmtFinType1[data$BsmtFinType1=='Unf'] <- 2
data$BsmtFinType1[data$BsmtFinType1=='NA'] <- 1
data$BsmtFinType1 <- droplevels(data$BsmtFinType1)
data$BsmtFinType1<- as.numeric(data$BsmtFinType1) # numerical scale changed

levels(data$BsmtFinType2) <- c(levels(data$BsmtFinType2),1,2,3,4,5,6,7)
data$BsmtFinType2[data$BsmtFinType2=='GLQ'] <- 7
data$BsmtFinType2[data$BsmtFinType2=='ALQ'] <- 6
data$BsmtFinType2[data$BsmtFinType2=='BLQ'] <- 5
data$BsmtFinType2[data$BsmtFinType2=='Rec'] <- 4
data$BsmtFinType2[data$BsmtFinType2=='LwQ'] <- 3
data$BsmtFinType2[data$BsmtFinType2=='Unf'] <- 2
data$BsmtFinType2[data$BsmtFinType2=='NA'] <- 1
data$BsmtFinType2 <- droplevels(data$BsmtFinType2)
data$BsmtFinType2<- as.numeric(data$BsmtFinType2)

levels(data$HeatingQC) <- c(levels(data$HeatingQC),1,2,3,4,5)
data$HeatingQC[data$HeatingQC=='Ex'] <- 5
data$HeatingQC[data$HeatingQC=='Gd'] <- 4
data$HeatingQC[data$HeatingQC=='TA'] <- 3
data$HeatingQC[data$HeatingQC=='Fa'] <- 2
data$HeatingQC[data$HeatingQC=='Po'] <- 1
data$HeatingQC <- droplevels(data$HeatingQC)
data$HeatingQC <- as.numeric(data$HeatingQC)

levels(data$CentralAir) <- c(levels(data$CentralAir),-1,0)
data$CentralAir[data$CentralAir=='Y'] <- 0
data$CentralAir[data$CentralAir=='N'] <- -1
data$CentralAir <- droplevels(data$CentralAir)
data$CentralAir <- as.numeric(data$CentralAir)
data$CentralAir <- data$CentralAir -1 # conversion to 1s and 0s

levels(data$KitchenQual) <- c(levels(data$KitchenQual),1,2,3,4,5)
data$KitchenQual[data$KitchenQual=='Ex'] <- 5
data$KitchenQual[data$KitchenQual=='Gd'] <- 4
data$KitchenQual[data$KitchenQual=='TA'] <- 3
data$KitchenQual[data$KitchenQual=='Fa'] <- 2
data$KitchenQual[data$KitchenQual=='Po'] <- 1
data$KitchenQual<- droplevels(data$KitchenQual)
data$KitchenQual <- as.numeric(data$KitchenQual)# numerical scale changed

levels(data$Functional) <- c(levels(data$Functional),1,2,3,4,5,6)
data$Functional[data$Functional=='Typ'] <- 6
data$Functional[data$Functional=='Min2'] <- 5
data$Functional[data$Functional=='Min1'] <- 5
data$Functional[data$Functional=='Mod'] <- 4
data$Functional[data$Functional=='Maj1'] <- 3
data$Functional[data$Functional=='Maj2'] <- 3
data$Functional[data$Functional=='Sev'] <- 2
data$Functional[data$Functional=='Sal'] <- 1
data$Functional<- droplevels(data$Functional)
data$Functional<- as.numeric(data$Functional)# Numerical scale changed
```

```r
levels(data$FireplaceQu) <- c(levels(data$FireplaceQu),0,1,2,3,4,5)
data$FireplaceQu[data$FireplaceQu=='Ex'] <- 5
data$FireplaceQu[data$FireplaceQu=='Gd'] <- 4
data$FireplaceQu[data$FireplaceQu=='TA'] <- 3
data$FireplaceQu[data$FireplaceQu=='Fa'] <- 2
data$FireplaceQu[data$FireplaceQu=='Po'] <- 1
data$FireplaceQu[data$FireplaceQu=='NA'] <- 0
data$FireplaceQu<- droplevels(data$FireplaceQu)
data$FireplaceQu<- as.numeric(data$FireplaceQu)

levels(data$GarageFinish) <- c(levels(data$GarageFinish),1,2,3,4)
data$GarageFinish[data$GarageFinish=='Fin'] <- 4
data$GarageFinish[data$GarageFinish=='RFn'] <- 3
data$GarageFinish[data$GarageFinish=='Unf'] <- 2
data$GarageFinish[data$GarageFinish=='NA'] <- 1
data$GarageFinish<- droplevels(data$GarageFinish)
data$GarageFinish<- as.numeric(data$GarageFinish)

levels(data$GarageQual) <- c(levels(data$GarageQual),1,2,3,4,5,6)
data$GarageQual[data$GarageQual=='Ex'] <- 6
data$GarageQual[data$GarageQual=='Gd'] <- 5
data$GarageQual[data$GarageQual=='TA'] <- 4
data$GarageQual[data$GarageQual=='Fa'] <- 3
data$GarageQual[data$GarageQual=='Po'] <- 2
data$GarageQual[data$GarageQual=='NA'] <- 1
data$GarageQual<- droplevels(data$GarageQual)
data$GarageQual<- as.numeric(data$GarageQual)

levels(data$GarageCond) <- c(levels(data$GarageCond),1,2,3,4,5,6)
data$GarageCond[data$GarageCond=='Ex'] <- 6
data$GarageCond[data$GarageCond=='Gd'] <- 5
data$GarageCond[data$GarageCond=='TA'] <- 4
data$GarageCond[data$GarageCond=='Fa'] <- 3
data$GarageCond[data$GarageCond=='Po'] <- 2
data$GarageCond[data$GarageCond=='NA'] <- 1
data$GarageCond<- droplevels(data$GarageCond)
data$GarageCond<- as.numeric(data$GarageCond)

levels(data$PavedDrive) <- c(levels(data$PavedDrive),1,2,3)
data$PavedDrive[data$PavedDrive=='Y'] <- 3
data$PavedDrive[data$PavedDrive=='P'] <- 2
data$PavedDrive[data$PavedDrive=='N'] <- 1
data$PavedDrive <- droplevels(data$PavedDrive)
data$PavedDrive <- as.numeric(data$PavedDrive)

levels(data$PoolQC) <- c(levels(data$PoolQC),1,2,3,4,5)
data$PoolQC[data$PoolQC=='Ex'] <- 5
data$PoolQC[data$PoolQC=='Gd'] <- 4
data$PoolQC[data$PoolQC=='TA'] <- 3
data$PoolQC[data$PoolQC=='Fa'] <- 2
data$PoolQC[data$PoolQC=='NA'] <- 1
data$PoolQC<- droplevels(data$PoolQC)
data$PoolQC<- as.numeric(data$PoolQC)
```

```
levels(data$Fence) <- c(levels(data$Fence),1,2,3,4,5)
data$Fence[data$Fence=='GdPrv'] <- 5
data$Fence[data$Fence=='MnPrv'] <- 4
data$Fence[data$Fence=='GdWo'] <- 3
data$Fence[data$Fence=='MnWw'] <- 2
data$Fence[data$Fence=='NA'] <- 1
data$Fence<- droplevels(data$Fence)
data$Fence<- as.numeric(data$Fence)
```

Unbinding the data to wit's state before binding took plcae

```
train <- data[1:1460,]
train$SalePrice <- hptrain$SalePrice
test <- data[1461 :2919,]
dim(train)
```

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

6. data Cleaning

```
# Cleaning the Training Dataset
colSums(is.na(train)) # check for sum of NAs in each column
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0           259             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0          1369             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##             8             8             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            37            37            38            37             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            38             0             0             0             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             1             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0           690            81            81
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##            81             0             0            81            81
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          1453          1179          1406
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
```

```
##             0             0             0             0             0
##     SalePrice
##             0
```

```r
train<-train[colSums(is.na(train))< 690]
colSums(is.na(train))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0           259             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             0             0             8
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##             8             0             0             0            37
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##            37            38            37             0            38
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             1             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             0             0             0
##    Fireplaces    GarageType   GarageYrBlt  GarageFinish    GarageCars
##             0            81            81            81             0
##    GarageArea    GarageQual    GarageCond    PavedDrive    WoodDeckSF
##             0            81            81             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

```r
M_train <- na.omit(train)
dim(M_train)# in the end we have 1094 instances with complete atributes
```

```
## [1] 1094   76
```

```r
#Cleaning the Test Dataset
colSums(is.na(test)) # check for sum of NAs in each column
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             4           227             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0          1352             0             0             2
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
```

```
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             1             1
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##            16            15             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            44            45            44            42             1
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            42             1             1             1             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             2             2             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             1             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             2             0           730            76            78
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##            78             1             1            78            78
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          1456          1169          1408
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             1             0
```

```
test<-test[colSums(is.na(test))< 730]
colSums(is.na(test))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             4           227             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             2             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             1             1            16
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##            15             0             0             0            44
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##            45            44            42             1            42
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             1             1             1             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             2             2             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             1             0             2
##    Fireplaces    GarageType   GarageYrBlt  GarageFinish    GarageCars
##             0            76            78            78             1
##    GarageArea    GarageQual    GarageCond    PavedDrive    WoodDeckSF
##             1            78            78             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
```

```
##            0            0            0            0            0
##      MiscVal       MoSold       YrSold     SaleType SaleCondition
##            0            0            0            1            0
```

```
M_test <- na.omit(test)
dim(M_test)
```

```
## [1] 1108   75
```

- The train dataset contains 1460 observations and 81 variables, after cleaning the dataset sshrinked to 1094 instances and 76 variables and the test 1108

Summary Statistics

- Since the R visual aid is not competent enough top display summary stats on 76 dimension, i will use my sentiments to discern 10 dimensions which are well deserving of exploration

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.6.1
```

```
describe(M_train)
```

```
##                vars    n     mean      sd  median  trimmed     mad
## Id                1 1094   727.38  420.96   723.5   726.72  541.89
## MSSubClass        2 1094    56.13   41.98    50.0    48.33   44.48
## MSZoning*         3 1094     4.03    0.66     4.0     4.07    0.00
## LotFrontage       4 1094    70.76   24.51    70.0    69.65   14.83
## LotArea           5 1094 10132.35 8212.25  9444.5  9497.02 2793.96
## Street*           6 1094     2.00    0.06     2.0     2.00    0.00
## LotShape*         7 1094     3.12    1.34     4.0     3.28    0.00
## LandContour*      8 1094     3.78    0.71     4.0     4.00    0.00
## Utilities*        9 1094     1.00    0.00     1.0     1.00    0.00
## LotConfig*       10 1094     4.14    1.57     5.0     4.42    0.00
## LandSlope*       11 1094     1.05    0.24     1.0     1.00    0.00
## Neighborhood*    12 1094    13.29    5.93    13.0    13.29    7.41
## Condition1*      13 1094     3.03    0.90     3.0     3.00    0.00
## Condition2*      14 1094     3.01    0.26     3.0     3.00    0.00
## BldgType*        15 1094     1.49    1.21     1.0     1.13    0.00
## HouseStyle*      16 1094     4.03    1.89     3.0     4.04    1.48
## OverallQual      17 1094     6.25    1.37     6.0     6.20    1.48
## OverallCond      18 1094     5.58    1.07     5.0     5.46    0.00
## YearBuilt        19 1094    47.59   31.19    45.0    44.44   38.55
## YearRemodAdd     20 1094    34.08   20.93    25.0    32.34   17.79
## RoofStyle*       21 1094     2.44    0.84     2.0     2.30    0.00
## RoofMatl*        22 1094     2.06    0.57     2.0     2.00    0.00
## Exterior1st*     23 1094    10.76    3.15    13.0    11.08    1.48
## Exterior2nd*     24 1094    11.46    3.54    14.0    11.80    1.48
## MasVnrType*      25 1094     2.79    0.63     3.0     2.75    0.00
## MasVnrArea       26 1094   109.86  190.67     0.0    67.49    0.00
## ExterQual        27 1094     2.44    0.59     2.0     2.38    0.00
## ExterCond        28 1094     3.09    0.33     3.0     3.00    0.00
## Foundation*      29 1094     2.39    0.72     2.0     2.46    1.48
## BsmtQual         30 1094     2.60    0.71     3.0     2.54    1.48
## BsmtCond         31 1094     3.01    0.29     3.0     3.00    0.00
## BsmtExposure     32 1094     1.67    1.04     1.0     1.47    0.00
## BsmtFinType1     33 1094     3.60    2.07     4.0     3.62    2.97
```

```
## BsmtFinSF1       34 1094    448.19   468.73    384.5    386.96   570.06
## BsmtFinType2     35 1094      1.27     0.87      1.0      1.01     0.00
## BsmtFinSF2       36 1094     45.25   159.08      0.0      1.22     0.00
## BsmtUnfSF        37 1094    606.12   445.83    525.0    559.72   416.61
## TotalBsmtSF      38 1094   1099.56   415.85   1023.0   1063.85   356.57
## Heating*         39 1094      2.02     0.17      2.0      2.00     0.00
## HeatingQC        40 1094      4.22     0.94      5.0      4.31     0.00
## CentralAir       41 1094      0.95     0.22      1.0      1.00     0.00
## Electrical*      42 1094      4.71     1.01      5.0      5.00     0.00
## X1stFlrSF        43 1094   1173.81   387.68   1097.0   1143.02   353.60
## X2ndFlrSF        44 1094    356.54   439.26      0.0    296.17     0.00
## LowQualFinSF     45 1094      4.68    42.10      0.0      0.00     0.00
## GrLivArea        46 1094   1535.03   526.12   1480.0   1484.76   459.61
## BsmtFullBath     47 1094      0.42     0.51      0.0      0.39     0.00
## BsmtHalfBath     48 1094      0.06     0.24      0.0      0.00     0.00
## FullBath         49 1094      1.58     0.55      2.0      1.57     0.00
## HalfBath         50 1094      0.39     0.50      0.0      0.35     0.00
## BedroomAbvGr     51 1094      2.86     0.76      3.0      2.85     0.00
## KitchenAbvGr     52 1094      1.03     0.19      1.0      1.00     0.00
## KitchenQual      53 1094      2.56     0.67      2.0      2.50     1.48
## TotRmsAbvGrd     54 1094      6.57     1.58      6.0      6.44     1.48
## Functional       55 1094      4.90     0.43      5.0      5.00     0.00
## Fireplaces       56 1094      0.61     0.63      1.0      0.54     1.48
## GarageType*      57 1094      3.33     1.81      2.0      3.17     0.00
## GarageYrBlt      58 1094     41.43    25.93     38.0     38.62    31.88
## GarageFinish     59 1094      1.81     0.81      2.0      1.76     1.48
## GarageCars       60 1094      1.88     0.66      2.0      1.84     0.00
## GarageArea       61 1094    503.76   192.26    484.0    489.63   182.36
## GarageQual       62 1094      2.97     0.27      3.0      3.00     0.00
## GarageCond       63 1094      2.97     0.25      3.0      3.00     0.00
## PavedDrive       64 1094      2.89     0.43      3.0      3.00     0.00
## WoodDeckSF       65 1094     94.34   122.62      0.0     73.38     0.00
## OpenPorchSF      66 1094     46.95    64.82     28.0     34.19    41.51
## EnclosedPorch    67 1094     22.05    61.57      0.0      3.96     0.00
## X3SsnPorch       68 1094      3.27    29.66      0.0      0.00     0.00
## ScreenPorch      69 1094     16.50    58.46      0.0      0.00     0.00
## PoolArea         70 1094      3.01    40.71      0.0      0.00     0.00
## MiscVal          71 1094     23.55   167.14      0.0      0.00     0.00
## MoSold           72 1094      6.34     2.69      6.0      6.28     2.97
## YrSold           73 1094     12.21     1.33     12.0     12.27     1.48
## SaleType*        74 1094      8.48     1.54      9.0      8.87     0.00
## SaleCondition*   75 1094      4.82     1.07      5.0      5.01     0.00
## SalePrice        76 1094 187033.26 83165.33 165750.0 175479.82 57450.75
##                  min    max  range   skew kurtosis     se
## Id                 1   1460   1459   0.02    -1.19  12.73
## MSSubClass        20    190    170   1.42     1.58   1.27
## MSZoning*          1      5      4  -1.71     5.70   0.02
## LotFrontage       21    313    292   2.22    17.96   0.74
## LotArea         1300 215245 213945  15.47   359.81 248.29
## Street*            1      2      1 -16.42   268.01   0.00
## LotShape*          1      4      3  -0.90    -1.16   0.04
## LandContour*       1      4      3  -3.17     8.65   0.02
## Utilities*         1      1      0    NaN      NaN   0.00
## LotConfig*         1      5      4  -1.35    -0.06   0.05
```

```
## LandSlope*        1     3     2  5.14   28.45   0.01
## Neighborhood*     1    25    24 -0.04   -1.06   0.18
## Condition1*       1     9     8  2.95   15.41   0.03
## Condition2*       1     8     7 13.24  274.18   0.01
## BldgType*         1     5     4  2.27    3.43   0.04
## HouseStyle*       1     8     7  0.27   -0.99   0.06
## OverallQual       2    10     8  0.30   -0.14   0.04
## OverallCond       2     9     7  0.86    1.04   0.03
## YearBuilt        10   140   130  0.63   -0.55   0.94
## YearRemodAdd     10    70    60  0.58   -1.23   0.63
## RoofStyle*        1     5     4  1.35    0.03   0.03
## RoofMatl*         1     8     7  9.11   84.11   0.02
## Exterior1st*      1    15    14 -0.80   -0.26   0.10
## Exterior2nd*      1    16    15 -0.77   -0.43   0.11
## MasVnrType*       1     4     3  0.00   -0.26   0.02
## MasVnrArea        0  1600  1600  2.69   10.03   5.76
## ExterQual         1     4     3  0.77   -0.16   0.02
## ExterCond         2     5     3  1.88    6.29   0.01
## Foundation*       1     6     5 -0.07    0.98   0.02
## BsmtQual          1     4     3  0.26   -0.43   0.02
## BsmtCond          1     4     3  0.14   10.59   0.01
## BsmtExposure      1     4     3  1.18   -0.16   0.03
## BsmtFinType1      1     6     5 -0.14   -1.64   0.06
## BsmtFinSF1        0  5644  5644  1.93   13.29  14.17
## BsmtFinType2      1     6     5  3.61   13.02   0.03
## BsmtFinSF2        0  1474  1474  4.36   21.45   4.81
## BsmtUnfSF         0  2336  2336  0.88    0.35  13.48
## TotalBsmtSF     105  6110  6005  2.31   19.48  12.57
## Heating*          2     5     3 10.14  125.20   0.01
## HeatingQC         1     5     4 -0.65   -1.04   0.03
## CentralAir        0     1     1 -3.98   13.89   0.01
## Electrical*       1     5     4 -3.25    8.69   0.03
## X1stFlrSF       438  4692  4254  1.37    6.33  11.72
## X2ndFlrSF         0  2065  2065  0.79   -0.54  13.28
## LowQualFinSF      0   572   572  9.87  101.05   1.27
## GrLivArea       438  5642  5204  1.55    6.11  15.91
## BsmtFullBath      0     2     2  0.53   -1.20   0.02
## BsmtHalfBath      0     2     2  4.04   15.44   0.01
## FullBath          0     3     3  0.02   -0.85   0.02
## HalfBath          0     2     2  0.61   -1.26   0.02
## BedroomAbvGr      0     6     6  0.02    1.29   0.02
## KitchenAbvGr      1     3     2  5.57   31.83   0.01
## KitchenQual       1     4     3  0.41   -0.38   0.02
## TotRmsAbvGrd      3    12     9  0.72    0.70   0.05
## Functional        2     5     3 -5.13   28.06   0.01
## Fireplaces        0     3     3  0.63   -0.16   0.02
## GarageType*       1     6     5  0.70   -1.40   0.05
## GarageYrBlt      10   120   110  0.66   -0.55   0.78
## GarageFinish      1     3     2  0.36   -1.40   0.02
## GarageCars        1     4     3  0.21   -0.43   0.02
## GarageArea      160  1418  1258  0.72    0.79   5.81
## GarageQual        1     5     4 -1.26   22.50   0.01
## GarageCond        1     5     4 -3.12   35.30   0.01
## PavedDrive        1     3     2 -3.91   13.82   0.01
```

```
## WoodDeckSF         0    857    857   1.52      3.25    3.71
## OpenPorchSF        0    547    547   2.38      8.84    1.96
## EnclosedPorch      0    552    552   3.16     11.27    1.86
## X3SsnPorch         0    508    508  11.04    140.41    0.90
## ScreenPorch        0    480    480   3.95     16.90    1.77
## PoolArea           0    648    648  13.58    184.46    1.23
## MiscVal            0   2500   2500   9.65    108.26    5.05
## MoSold             1     12     11   0.17     -0.43    0.08
## YrSold            10     14      4  -0.12     -1.20    0.04
## SaleType*          1      9      8  -3.74     14.30    0.05
## SaleCondition*     1      6      5  -2.81      7.61    0.03
## SalePrice      35311 755000 719689   1.93      6.37 2514.40
```

```r
describe(M_test)
```

```
##                  vars    n    mean      sd median trimmed     mad  min   max
## Id                  1 1108 2185.00  424.75 2195.5 2184.69  542.63 1461  2919
## MSSubClass          2 1108   56.89   42.83   50.0   49.25   44.48   20   190
## MSZoning*           3 1108    4.05    0.65    4.0    4.09    0.00    1     5
## LotFrontage         4 1108   68.63   22.04   68.0   68.12   17.79   21   195
## LotArea             5 1108 9459.20 4211.98 9350.0 9260.06 3008.94 1484 51974
## Street*             6 1108    2.00    0.05    2.0    2.00    0.00    1     2
## LotShape*           7 1108    3.10    1.36    4.0    3.25    0.00    1     4
## LandContour*        8 1108    3.77    0.70    4.0    4.00    0.00    1     4
## Utilities*          9 1108    1.00    0.00    1.0    1.00    0.00    1     1
## LotConfig*         10 1108    4.17    1.55    5.0    4.46    0.00    1     5
## LandSlope*         11 1108    1.05    0.22    1.0    1.00    0.00    1     3
## Neighborhood*      12 1108   13.53    5.73   13.0   13.56    7.41    1    25
## Condition1*        13 1108    3.04    0.82    3.0    3.00    0.00    1     9
## Condition2*        14 1108    3.00    0.14    3.0    3.00    0.00    1     5
## BldgType*          15 1108    1.52    1.25    1.0    1.17    0.00    1     5
## HouseStyle*        16 1108    3.95    1.91    3.0    3.93    0.00    1     8
## OverallQual        17 1108    6.19    1.42    6.0    6.14    1.48    2    10
## OverallCond        18 1108    5.60    1.05    5.0    5.47    0.00    1     9
## YearBuilt          19 1108   47.66   30.60   46.0   44.72   38.55   10   141
## YearRemodAdd       20 1108   35.42   21.17   26.5   33.99   20.02   10    70
## RoofStyle*         21 1108    2.39    0.80    2.0    2.24    0.00    1     6
## RoofMatl*          22 1108    2.03    0.40    2.0    2.00    0.00    2     8
## Exterior1st*       23 1108   10.61    3.19   13.0   10.91    1.48    1    15
## Exterior2nd*       24 1108   11.28    3.59   14.0   11.61    1.48    1    16
## MasVnrType*        25 1108    2.79    0.61    3.0    2.75    0.00    1     4
## MasVnrArea         26 1108  105.15  180.82    0.0   63.85    0.00    0  1290
## ExterQual          27 1108    2.44    0.60    2.0    2.38    0.00    1     4
## ExterCond          28 1108    3.10    0.38    3.0    3.02    0.00    1     5
## Foundation*        29 1108    2.36    0.70    2.0    2.44    1.48    1     6
## BsmtQual           30 1108    2.57    0.73    3.0    2.51    1.48    1     4
## BsmtCond           31 1108    3.00    0.29    3.0    3.00    0.00    1     4
## BsmtExposure       32 1108    1.67    1.06    1.0    1.47    0.00    1     4
## BsmtFinType1       33 1108    3.63    2.05    4.0    3.67    2.97    1     6
## BsmtFinSF1         34 1108  450.79  464.80  364.5  387.53  540.41    0  4010
## BsmtFinType2       35 1108    1.35    1.01    1.0    1.04    0.00    1     6
## BsmtFinSF2         36 1108   54.14  174.72    0.0    3.63    0.00    0  1393
## BsmtUnfSF          37 1108  577.40  434.73  480.0  529.21  397.34    0  2140
## TotalBsmtSF        38 1108 1082.33  424.67  993.5 1050.74  359.53  160  5095
## Heating*           39 1108    2.01    0.08    2.0    2.00    0.00    2     3
```

```
## HeatingQC        40 1108    4.21    0.93    5.0    4.30    0.00   1     5
## CentralAir       41 1108    0.96    0.20    1.0    1.00    0.00   0     1
## Electrical*      42 1108    4.70    1.03    5.0    5.00    0.00   1     5
## X1stFlrSF        43 1108 1156.38  406.46 1072.0 1121.26  355.82 407  5095
## X2ndFlrSF        44 1108  323.16  410.39    0.0  263.87    0.00   0  1862
## LowQualFinSF     45 1108    3.29   45.42    0.0    0.00    0.00   0  1064
## GrLivArea        46 1108 1482.82  480.51 1429.0 1434.00  426.25 407  5095
## BsmtFullBath     47 1108    0.45    0.52    0.0    0.42    0.00   0     2
## BsmtHalfBath     48 1108    0.06    0.25    0.0    0.00    0.00   0     2
## FullBath         49 1108    1.56    0.54    2.0    1.56    0.00   0     4
## HalfBath         50 1108    0.37    0.49    0.0    0.33    0.00   0     2
## BedroomAbvGr     51 1108    2.82    0.78    3.0    2.79    0.00   0     6
## KitchenAbvGr     52 1108    1.02    0.15    1.0    1.00    0.00   1     2
## KitchenQual      53 1108    2.54    0.67    2.0    2.47    0.00   1     4
## TotRmsAbvGrd     54 1108    6.36    1.49    6.0    6.25    1.48   3    15
## Functional       55 1108    4.93    0.33    5.0    5.00    0.00   2     5
## Fireplaces       56 1108    0.59    0.64    1.0    0.51    1.48   0     4
## GarageType*      57 1108    3.38    1.83    2.0    3.24    0.00   1     6
## GarageYrBlt      58 1108   42.37   26.36   41.0   39.59   34.10  10   124
## GarageFinish     59 1108    1.81    0.83    2.0    1.76    1.48   1     3
## GarageCars       60 1108    1.85    0.68    2.0    1.80    0.00   1     5
## GarageArea       61 1108  496.51  196.93  480.0  478.62  189.77 100  1488
## GarageQual       62 1108    2.95    0.26    3.0    3.00    0.00   1     4
## GarageCond       63 1108    2.97    0.23    3.0    3.00    0.00   1     5
## PavedDrive       64 1108    2.85    0.50    3.0    3.00    0.00   1     3
## WoodDeckSF       65 1108   93.95  123.63    0.0   72.40    0.00   0   870
## OpenPorchSF      66 1108   48.60   67.62   28.0   34.78   41.51   0   570
## EnclosedPorch    67 1108   23.44   66.78    0.0    5.42    0.00   0  1012
## X3SsnPorch       68 1108    1.77   20.16    0.0    0.00    0.00   0   360
## ScreenPorch      69 1108   17.76   57.19    0.0    0.10    0.00   0   576
## PoolArea         70 1108    1.79   30.68    0.0    0.00    0.00   0   800
## MiscVal          71 1108   63.86  711.77    0.0    0.00    0.00   0 17000
## MoSold           72 1108    6.11    2.75    6.0    6.05    2.97   1    12
## YrSold           73 1108   12.23    1.32   12.0   12.28    1.48  10    14
## SaleType*        74 1108    8.45    1.64    9.0    8.88    0.00   1     9
## SaleCondition*   75 1108    4.83    1.02    5.0    5.00    0.00   1     6
##                range   skew kurtosis     se
## Id              1458   0.01    -1.21  12.76
## MSSubClass       170   1.35     1.24   1.29
## MSZoning*          4  -1.55     4.93   0.02
## LotFrontage      174   0.54     2.01   0.66
## LotArea        50490   2.22    16.00 126.54
## Street*            1 -19.11   363.67   0.00
## LotShape*          3  -0.87    -1.22   0.04
## LandContour*       3  -3.01     7.71   0.02
## Utilities*         0    NaN      NaN   0.00
## LotConfig*         4  -1.40     0.09   0.05
## LandSlope*         2   4.55    20.22   0.01
## Neighborhood*     24  -0.07    -0.98   0.17
## Condition1*        8   2.81    14.55   0.02
## Condition2*        4   1.33   129.65   0.00
## BldgType*          4   2.13     2.83   0.04
## HouseStyle*        7   0.35    -0.90   0.06
## OverallQual        8   0.30    -0.19   0.04
```

```
## OverallCond        8   0.85     1.17   0.03
## YearBuilt        131   0.57    -0.68   0.92
## YearRemodAdd      60   0.46    -1.37   0.64
## RoofStyle*         5   1.61     0.94   0.02
## RoofMatl*          6  12.04   146.56   0.01
## Exterior1st*      14  -0.73    -0.27   0.10
## Exterior2nd*      15  -0.68    -0.59   0.11
## MasVnrType*        3   0.01    -0.25   0.02
## MasVnrArea      1290   2.41     7.60   5.43
## ExterQual          3   0.78    -0.09   0.02
## ExterCond          4   1.52     6.33   0.01
## Foundation*        5  -0.28     0.22   0.02
## BsmtQual           3   0.27    -0.41   0.02
## BsmtCond           3  -0.23    10.59   0.01
## BsmtExposure       3   1.21    -0.11   0.03
## BsmtFinType1       5  -0.15    -1.62   0.06
## BsmtFinSF1      4010   1.27     3.24  13.96
## BsmtFinType2       5   3.10     8.94   0.03
## BsmtFinSF2      1393   3.81    15.32   5.25
## BsmtUnfSF       2140   0.92     0.30  13.06
## TotalBsmtSF     4935   1.44     7.41  12.76
## Heating*           1  12.44   153.01   0.00
## HeatingQC          4  -0.63    -1.07   0.03
## CentralAir         1  -4.48    18.09   0.01
## Electrical*        4  -3.17     8.17   0.03
## X1stFlrSF       4688   1.58     8.22  12.21
## X2ndFlrSF       1862   0.87    -0.34  12.33
## LowQualFinSF    1064  17.14   335.46   1.36
## GrLivArea       4688   1.23     3.60  14.44
## BsmtFullBath       2   0.50    -1.13   0.02
## BsmtHalfBath       2   3.73    12.82   0.01
## FullBath           4   0.11    -0.83   0.02
## HalfBath           2   0.67    -1.21   0.01
## BedroomAbvGr       6   0.18     1.11   0.02
## KitchenAbvGr       1   6.16    35.99   0.00
## KitchenQual        3   0.46    -0.35   0.02
## TotRmsAbvGrd      12   0.88     1.67   0.04
## Functional         3  -5.39    33.21   0.01
## Fireplaces         4   0.83     0.57   0.02
## GarageType*        5   0.63    -1.48   0.06
## GarageYrBlt      114   0.63    -0.56   0.79
## GarageFinish       2   0.37    -1.44   0.02
## GarageCars         4   0.37    -0.12   0.02
## GarageArea      1388   0.89     1.15   5.92
## GarageQual         3  -2.68    12.57   0.01
## GarageCond         4  -4.35    41.62   0.01
## PavedDrive         2  -3.23     8.77   0.02
## WoodDeckSF       870   1.59     3.47   3.71
## OpenPorchSF      570   2.30     7.95   2.03
## EnclosedPorch   1012   5.11    48.76   2.01
## X3SsnPorch       360  12.98   185.10   0.61
## ScreenPorch      576   3.61    15.50   1.72
## PoolArea         800  20.59   471.12   0.92
## MiscVal        17000  18.26   379.82  21.38
```

```
## MoSold            11   0.19    -0.54   0.08
## YrSold             4  -0.18    -1.14   0.04
## SaleType*          8  -3.54    12.30   0.05
## SaleCondition*     5  -2.93     8.53   0.03
```

Visualization

From an economic and humanics point of view, the major factors that affect the prices of houses are the exterior qualities and condition and the interior qualities and condition. The proxy for these measures in our dataset sets are: OverallQual, OverallCond, YearBuilt, ExterQual,YearBuilt. ExtCond,KitchenQual, GrLivArea,Functional

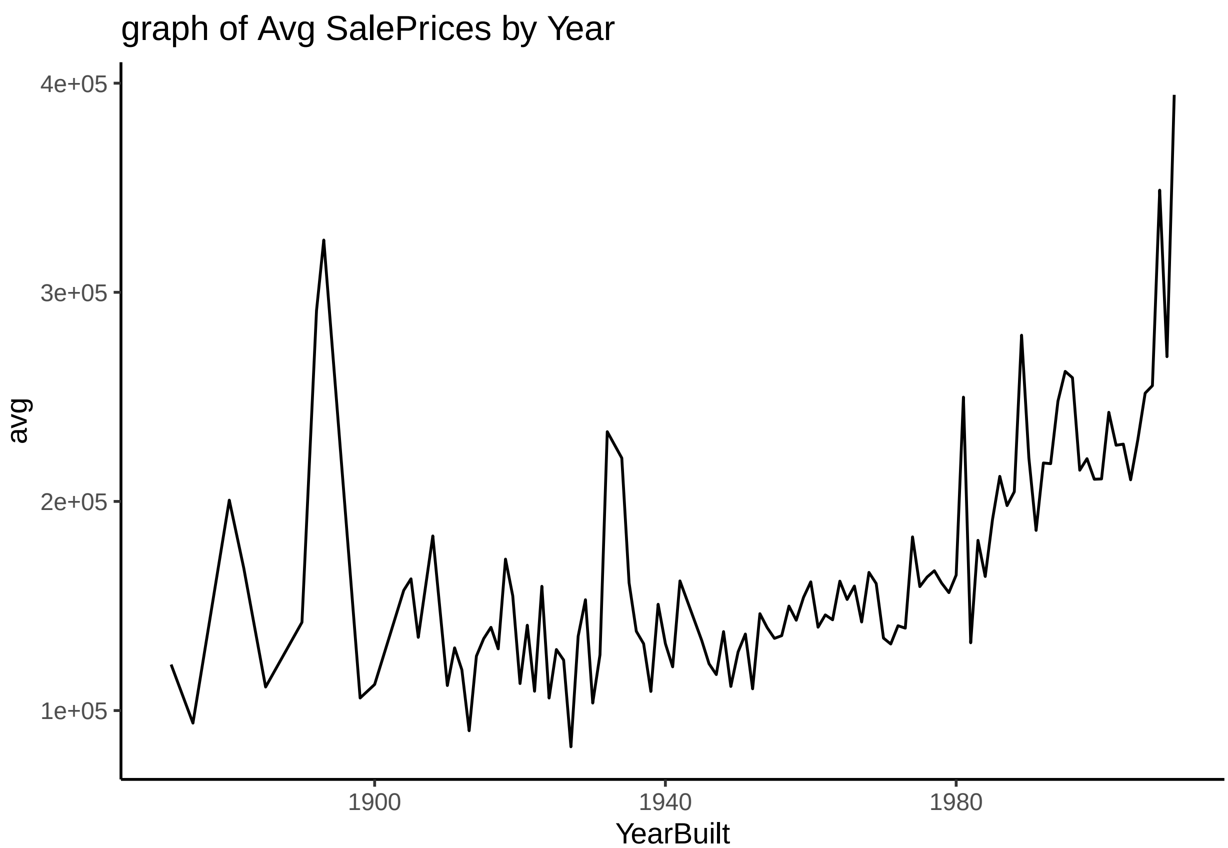
```r
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1

## -- Attaching packages --------------------------------------------- tidyverse 1.2.1 --

## v tibble  2.1.1     v purrr   0.3.2
## v tidyr   1.0.0     v dplyr   0.8.3
## v readr   1.3.1     v stringr 1.4.0
## v tibble  2.1.1     v forcats 0.4.0

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.1

## Warning: package 'dplyr' was built under R version 3.6.1

## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x purrr::%||%()    masks lemon::%||%()
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x tidyr::expand()  masks Matrix::expand()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x tidyr::pack()    masks Matrix::pack()
## x tidyr::unpack()  masks Matrix::unpack()
```

```r
avg_price_by_YearBuilt<- group_by(hptrain,YearBuilt) %>%summarise(avg = mean(SalePrice))

avg_price_Kitchenqual<- group_by(M_train,KitchenQual) %>%summarise(avg = mean(SalePrice))


ggplot() +
geom_line(avg_price_by_YearBuilt, mapping = aes(x = YearBuilt, y = avg),stat = 'identity') + theme_bw()
```




```
ggplot() +
geom_point(M_train, mapping = aes(x = YearBuilt, y = SalePrice,color =factor( KitchenQual))) + theme_bw
```

* The first Graph:This shows the relationship between the sales prices and a proxy for exterior condition and quality the year it was built.

- The Second Graph: his shows the relationship between the sales prices and a proxy for both interior and exterior condition and quality.The scatterplot shows that houses built recently and of high kitchen quality are the most expensive houses in this market.

7. Modelling

Regression Model

we shall test our economic intuition by first modelling the attributes we think are pivotal to discerning house prices against the SalePrice.

```
lm_model1 <- lm(SalePrice ~OverallQual + OverallCond
                + YearBuilt + ExterQual
                + YearBuilt + ExterCond + KitchenQual + GrLivArea + Functional, M_train)
summary(lm_model1)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + OverallCond + YearBuilt +
##     ExterQual + YearBuilt + ExterCond + KitchenQual + GrLivArea +
##     Functional, data = M_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -412831  -22835    -998   16887  265707
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.488e+05  1.981e+04  -7.512 1.22e-13 ***
## OverallQual  1.851e+04  1.682e+03  11.004  < 2e-16 ***
## OverallCond  5.207e+03  1.423e+03   3.658 0.000266 ***
## YearBuilt   -4.060e+02  6.091e+01  -6.665 4.21e-11 ***
## ExterQual    1.564e+04  3.703e+03   4.224 2.60e-05 ***
## ExterCond   -3.583e+03  4.071e+03  -0.880 0.379043
## KitchenQual  1.530e+04  2.920e+03   5.240 1.93e-07 ***
## GrLivArea    6.263e+01  3.169e+00  19.765  < 2e-16 ***
## Functional   9.804e+03  2.988e+03   3.281 0.001066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41150 on 1085 degrees of freedom
## Multiple R-squared:  0.7569, Adjusted R-squared:  0.7551
## F-statistic: 422.3 on 8 and 1085 DF,  p-value: < 2.2e-16
```

The model is able to explain just 76 percent in the variation of the SalePrices. Now lets Build a cor Matrix

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.2

## corrplot 0.84 loaded
```

```r
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.6.1

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

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

## The following object is masked from 'package:psych':
##
##     describe

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

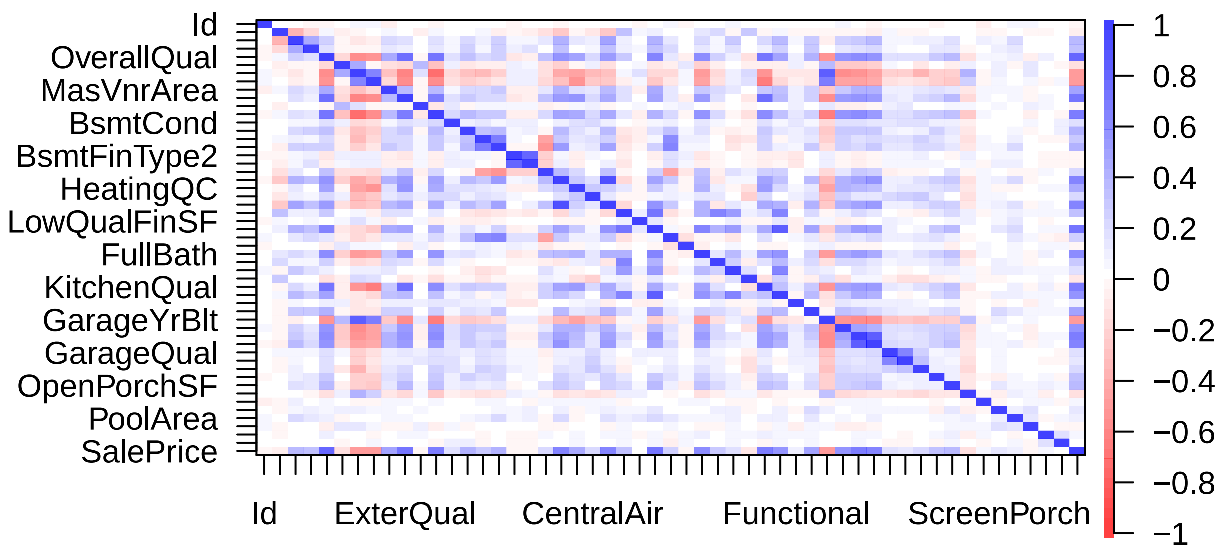
```r
nums <- unlist(lapply(M_train, is.numeric))
cordata <-M_train[, nums]
corr <-cor(cordata)
head(corr)
```

```
##                      Id  MSSubClass LotFrontage     LotArea OverallQual
## Id          1.000000000  0.01553996 -0.01447934 -0.04231498 -0.05837115
## MSSubClass  0.015539961  1.00000000 -0.38946624 -0.19790310  0.03163944
## LotFrontage -0.014479340 -0.38946624  1.00000000  0.41971402  0.24116867
```

```
## LotArea     -0.042314983 -0.19790310  0.41971402  1.00000000  0.16987639
## OverallQual -0.058371151  0.03163944  0.24116867  0.16987639  1.00000000
## OverallCond  0.008627076 -0.08555275 -0.04713215 -0.03311332 -0.18958707
##              OverallCond   YearBuilt YearRemodAdd  MasVnrArea   ExterQual
## Id           0.008627076  0.02261005   0.03023948 -0.07234409 -0.01119374
## MSSubClass  -0.085552753 -0.02160453  -0.01017785  0.04000907  0.01201779
## LotFrontage -0.047132146 -0.10795764  -0.08293758  0.18976867  0.16691218
## LotArea     -0.033113316 -0.02895353  -0.02430774  0.10659974  0.08844387
## OverallQual -0.189587068 -0.59076066  -0.56858172  0.41975578  0.74710714
## OverallCond  1.000000000  0.43764707  -0.02442673 -0.17458084 -0.20554309
##                ExterCond    BsmtQual    BsmtCond BsmtExposure
## Id          -0.001751016 -0.04796861 0.005112179  0.002239084
## MSSubClass  -0.061406823  0.05497957 0.004560655  0.001118447
## LotFrontage -0.012625586  0.17416568 0.044170154  0.196846191
## LotArea     -0.004828784  0.12400097 0.028296866  0.224879388
## OverallQual -0.030494098  0.69510092 0.166661757  0.309218113
## OverallCond  0.370193399 -0.31535638 0.084845922 -0.106991215
##              BsmtFinType1  BsmtFinSF1 BsmtFinType2  BsmtFinSF2   BsmtUnfSF
## Id           0.0006906238 -0.01323430  -0.02132908  0.01496371 -0.01431555
## MSSubClass   0.0238148137 -0.06943875  -0.04122682 -0.07383437 -0.14715525
## LotFrontage  0.0759519261  0.23973406   0.02147790  0.04692768  0.11136780
## LotArea      0.0532978394  0.23234130   0.06320433  0.13861504  0.00892374
## OverallQual  0.1991211208  0.23043768  -0.09898820 -0.08134187  0.29738366
## OverallCond -0.0607524944 -0.06828454   0.08391371  0.04059757 -0.16974268
##             TotalBsmtSF    HeatingQC  CentralAir    X1stFlrSF    X2ndFlrSF
## Id          -0.02454075 -0.003332991  0.01381400 -0.007491547 -0.005996772
## MSSubClass  -0.26427719 -0.046819983 -0.10933748 -0.258207290  0.319175589
## LotFrontage  0.40756576  0.102119078  0.07561217  0.453035137  0.074953308
## LotArea      0.32447561  0.017153951  0.03265787  0.331295090  0.075310601
## OverallQual  0.54744836  0.488505508  0.21310864  0.527908193  0.265906325
## OverallCond -0.24341873 -0.057427695  0.07424269 -0.166190772  0.004046500
##             LowQualFinSF   GrLivArea BsmtFullBath BsmtHalfBath    FullBath
## Id           -0.04055278 -0.01377187   0.02726453 -0.027414835  0.00360078
## MSSubClass    0.02493546  0.07821301  -0.01304034  0.012508925  0.11949492
## LotFrontage   0.01074777  0.39725992   0.11515085 -0.000491143  0.18969162
## LotArea       0.01995628  0.30859024   0.17987387 -0.014596636  0.13285990
## OverallQual  -0.01118637  0.61010179   0.10713753 -0.060774950  0.59788087
## OverallCond   0.04786495 -0.11525010  -0.07277768  0.121421245 -0.22599458
##                HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
## Id          -0.01540270   0.03932012  0.013251690 -0.01395709   0.01383151
## MSSubClass   0.20625892  -0.04462799  0.258401357 -0.02041618   0.03818036
## LotFrontage  0.04341389   0.27713568  0.007411095  0.17680818   0.35471401
## LotArea      0.04397656   0.14142789 -0.010854737  0.09113260   0.24184882
## OverallQual  0.23989343   0.09146247 -0.141071258  0.68718609   0.46573304
## OverallCond -0.08962479   0.01230047 -0.070659761 -0.08959193  -0.09330853
##               Functional  Fireplaces  GarageYrBlt GarageFinish
## Id          -0.002372092 -0.01579797  0.003820353   0.02718551
## MSSubClass  -0.050432050 -0.02957546 -0.051224848  -0.02935880
## LotFrontage  0.045318125  0.26029272 -0.067253989   0.22243387
## LotArea     -0.005497714  0.25584152 -0.012870504   0.10568710
## OverallQual  0.090073902  0.40972493 -0.562405490   0.55692549
## OverallCond  0.072343561 -0.03073112  0.353290553  -0.26261331
##               GarageCars  GarageArea  GarageQual  GarageCond   PavedDrive
## Id          -0.009568429 -0.02328980 -0.01117384  0.00539024  0.011751310
```

```
## MSSubClass  -0.031638414 -0.09537427  0.02535720 -0.04637963 -0.022592067
## LotFrontage  0.285748432  0.35703044  0.05600541  0.04341539  0.080518306
## LotArea      0.173524545  0.21310386  0.02332850  0.01503685  0.007308909
## OverallQual  0.605466005  0.55531450  0.16012874  0.13908616  0.168885488
## OverallCond -0.269616198 -0.23358487  0.02587940  0.01934072 -0.114669295
##               WoodDeckSF   OpenPorchSF EnclosedPorch  X3SsnPorch
## Id          -0.02759694 -0.0009871324    0.01179582 -0.06168827
## MSSubClass  -0.01851432  0.0067991938   -0.01931261 -0.03585470
## LotFrontage  0.08133784  0.1608617646    0.01605769  0.07300357
## LotArea      0.13399466  0.0980508673   -0.02278860  0.01334258
## OverallQual  0.27365228  0.3358837610   -0.15507999  0.02008128
## OverallCond -0.01885647 -0.0844047727    0.06712386 -0.01088108
##             ScreenPorch     PoolArea       MiscVal       MoSold
## Id           0.01501915  0.048486922  0.0509537414  0.007486117
## MSSubClass  -0.02185369  0.003220667 -0.0432989576 -0.025393383
## LotFrontage  0.03493750  0.211958692  0.0007892619  0.014951413
## LotArea      0.07241255  0.109293650  0.0124828851  0.006270273
## OverallQual  0.04928603  0.080037438 -0.0629438619  0.082994654
## OverallCond  0.08441611 -0.024918641  0.1214068321 -0.009660504
##                   YrSold   SalePrice
## Id          -0.005306732 -0.04759501
## MSSubClass   0.012346675 -0.08947768
## LotFrontage -0.013365896  0.34397763
## LotArea      0.006412434  0.30226803
## OverallQual  0.003529133  0.79543682
## OverallCond -0.046775170 -0.13851095
```

```
corPlot(corr)
```



The correllation Matrix would have been useful if we have less attributes but with attributes of thse amount it is just too congested. Let go ahead and run a Lm Model using all the attributes and then eliminating the attributes with zero coefficients.

- let Store the predictors in a design matrix `x` and the outcome in a vector `y`. `model.matrix` is a useful function that automatically transforms any qualitative variables into dummy variables. This is important because `glmnet()` can only use quantitative inputs.

```
x <- model.matrix(SalePrice~.,M_train)[,-c(1,2)]
y <- M_train$SalePrice
test <- data.matrix(M_test)
```

Linear Regression

```
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 3.6.2
```

```
lm_model2 <- lm(y~ x)

lm_model2$coefficients
```

```
##       (Intercept)       xMSSubClass      xMSZoningFV
##     -9.412771e+05     -6.287232e+01     3.640778e+04
##      xMSZoningRH       xMSZoningRL      xMSZoningRM
##      3.391285e+04      2.554681e+04     2.447313e+04
##     xLotFrontage          xLotArea      xStreetPave
##      8.785663e+01      7.651189e-01     3.167619e+04
```

```
##           xLotShapeIR2           xLotShapeIR3           xLotShapeReg
##           1.382747e+04           7.695083e+03           4.079627e+03
##        xLandContourHLS        xLandContourLow        xLandContourLvl
##           7.993425e+03          -2.439906e+04           3.724359e+03
##       xUtilitiesNoSeWa       xLotConfigCulDSac          xLotConfigFR2
##                     NA           1.536777e+04          -8.022298e+03
##          xLotConfigFR3       xLotConfigInside          xLandSlopeMod
##          -1.608880e+04          -1.797836e+02           5.186327e+03
##          xLandSlopeSev  xNeighborhoodBlueste   xNeighborhoodBrDale
##          -2.993163e+04           4.269238e+03           1.837330e+04
##   xNeighborhoodBrkSide  xNeighborhoodClearCr  xNeighborhoodCollgCr
##           8.368114e+03          -4.461545e+03          -9.009479e+03
##   xNeighborhoodCrawfor  xNeighborhoodEdwards  xNeighborhoodGilbert
##           1.617930e+04          -1.218979e+04          -7.708812e+03
##    xNeighborhoodIDOTRR  xNeighborhoodMeadowV  xNeighborhoodMitchel
##           6.435720e+03          -1.660349e+03          -5.977398e+03
##     xNeighborhoodNAmes  xNeighborhoodNoRidge  xNeighborhoodNPkVill
##          -8.186467e+03           2.115142e+04           1.497367e+04
##   xNeighborhoodNridgHt    xNeighborhoodNWAmes  xNeighborhoodOldTown
##           2.927539e+04          -1.470913e+04          -1.043846e+03
##    xNeighborhoodSawyer  xNeighborhoodSawyerW  xNeighborhoodSomerst
##           4.050754e+03          -1.021496e+03          -4.325727e+03
##   xNeighborhoodStoneBr     xNeighborhoodSWISU   xNeighborhoodTimber
##           4.583845e+04          -6.467621e+02          -9.176851e+03
##   xNeighborhoodVeenker       xCondition1Feedr        xCondition1Norm
##           8.111651e+03           5.874754e+02           1.261829e+04
##        xCondition1PosA        xCondition1PosN        xCondition1RRAe
##           6.554238e+03          -5.641011e+02          -8.333827e+03
##        xCondition1RRAn        xCondition1RRNe        xCondition1RRNn
##           1.254606e+04           1.338890e+04           1.218899e+04
##       xCondition2Feedr        xCondition2Norm        xCondition2PosA
##          -1.592408e+04          -7.602757e+03           3.499346e+04
##        xCondition2PosN        xCondition2RRAe        xCondition2RRAn
##          -2.313643e+05                     NA                     NA
##        xCondition2RRNn        xBldgType2fmCon        xBldgTypeDuplex
##           5.745462e+03          -1.143275e+03          -1.281874e+04
##         xBldgTypeTwnhs        xBldgTypeTwnhsE     xHouseStyle1.5Unf
##          -2.381886e+04          -1.753373e+04           1.126771e+04
##     xHouseStyle1Story      xHouseStyle2.5Fin      xHouseStyle2.5Unf
##           1.677118e+04          -2.295498e+04          -1.505830e+04
##     xHouseStyle2Story      xHouseStyleSFoyer        xHouseStyleSLvl
##          -4.570083e+03           1.085875e+04           1.310219e+04
##           xOverallQual           xOverallCond             xYearBuilt
##           8.976876e+03           5.957116e+03          -2.259700e+02
##          xYearRemodAdd        xRoofStyleGable     xRoofStyleGambrel
##           1.901837e+01           3.115274e+04           3.264687e+04
##          xRoofStyleHip      xRoofStyleMansard         xRoofStyleShed
##           3.310293e+04           4.860564e+04                     NA
##       xRoofMatlCompShg       xRoofMatlMembran         xRoofMatlMetal
##           6.783677e+05           7.907588e+05                     NA
##          xRoofMatlRoll       xRoofMatlTar&Grv       xRoofMatlWdShake
##           6.697871e+05           6.882430e+05           6.515377e+05
##       xRoofMatlWdShngl    xExterior1stAsphShn    xExterior1stBrkComm
##           7.524297e+05                     NA          -6.097332e+04
```

```
##    xExterior1stBrkFace    xExterior1stCBlock   xExterior1stCemntBd
##          -2.747016e+03          5.236071e+03         -3.043173e+04
##    xExterior1stHdBoard   xExterior1stImStucc   xExterior1stMetalSd
##          -2.041168e+04         -6.816953e+04          6.661302e+02
##    xExterior1stPlywood     xExterior1stStone    xExterior1stStucco
##          -2.867435e+04          2.351389e+04         -1.075551e+04
##    xExterior1stVinylSd   xExterior1stWd Sdng   xExterior1stWdShing
##          -2.082042e+04         -1.467408e+04         -1.272659e+04
##    xExterior2ndAsphShn   xExterior2ndBrk Cmn   xExterior2ndBrkFace
##           1.749101e+04          3.536973e+04          7.459222e+03
##     xExterior2ndCBlock   xExterior2ndCmentBd   xExterior2ndHdBoard
##                     NA          4.367043e+04          1.784041e+04
##    xExterior2ndImStucc   xExterior2ndMetalSd     xExterior2ndOther
##           4.872045e+04          8.132065e+03         -1.169323e+04
##    xExterior2ndPlywood     xExterior2ndStone    xExterior2ndStucco
##           1.928871e+04         -1.176026e+03          1.016026e+04
##    xExterior2ndVinylSd   xExterior2ndWd Sdng   xExterior2ndWd Shng
##           2.173908e+04          1.893158e+04          1.179167e+04
##     xMasVnrTypeBrkFace       xMasVnrTypeNone      xMasVnrTypeStone
##           6.179358e+03          1.351513e+04          1.554972e+04
##            xMasVnrArea            xExterQual            xExterCond
##           2.945108e+01          6.109790e+03         -3.998059e+03
##      xFoundationCBlock      xFoundationPConc       xFoundationSlab
##           5.325294e+03          4.611153e+03                    NA
##       xFoundationStone       xFoundationWood             xBsmtQual
##           3.863210e+03         -4.636248e+04          6.195667e+03
##              xBsmtCond         xBsmtExposure         xBsmtFinType1
##          -1.033680e+03          5.329080e+03          6.533059e+01
##            xBsmtFinSF1         xBsmtFinType2           xBsmtFinSF2
##           4.868539e+01         -3.219986e+02          3.537276e+01
##             xBsmtUnfSF          xTotalBsmtSF         xHeatingGasA
##           2.623227e+01                    NA          2.503166e+04
##           xHeatingGasW          xHeatingGrav          xHeatingOthW
##           2.124117e+04          3.945005e+04                    NA
##           xHeatingWall            xHeatingQC           xCentralAir
##                     NA          5.194131e+01          6.063417e+03
##       xElectricalFuseF      xElectricalFuseP        xElectricalMix
##          -2.858137e+03          9.992639e+03          1.469492e+03
##       xElectricalSBrkr            xX1stFlrSF            xX2ndFlrSF
##           1.938797e+02          4.066097e+01          7.387083e+01
##          xLowQualFinSF            xGrLivArea         xBsmtFullBath
##           2.978786e+01                    NA          2.524994e+02
##          xBsmtHalfBath             xFullBath             xHalfBath
##          -8.628838e+02          2.589723e+03          2.977699e+03
##          xBedroomAbvGr         xKitchenAbvGr          xKitchenQual
##          -5.860661e+03         -1.393706e+04          6.439366e+03
##          xTotRmsAbvGrd           xFunctional           xFireplaces
##           2.207829e+03          6.863408e+03          2.963812e+03
##      xGarageTypeAttchd    xGarageTypeBasment    xGarageTypeBuiltIn
##           6.141289e+03          2.082933e+04          4.964627e+03
##     xGarageTypeCarPort     xGarageTypeDetchd          xGarageYrBlt
##           1.915511e+04          1.175559e+04         -4.754837e+01
##          xGarageFinish           xGarageCars           xGarageArea
##           5.948869e+02          4.695204e+03          1.270784e+01
```

```
##             xGarageQual             xGarageCond             xPavedDrive
##            5.369091e+03           -3.494141e+03           -1.368681e+02
##             xWoodDeckSF            xOpenPorchSF          xEnclosedPorch
##            4.629100e+00           -2.774113e+00           -8.735321e+00
##             xX3SsnPorch            xScreenPorch               xPoolArea
##            4.978890e+01            2.643828e+01            6.404514e+01
##                xMiscVal                 xMoSold                 xYrSold
##           -2.522598e+00           -6.204934e+02           -3.742298e+02
##            xSaleTypeCon          xSaleTypeConLD          xSaleTypeConLI
##            2.551757e+04            2.034551e+04            2.127872e+03
##          xSaleTypeConLw            xSaleTypeCWD            xSaleTypeNew
##           -3.509827e+03            1.418502e+04            1.524877e+04
##            xSaleTypeOth             xSaleTypeWD xSaleConditionAdjLand
##            3.271351e+04           -1.160277e+03            3.280475e+04
##  xSaleConditionAlloca  xSaleConditionFamily  xSaleConditionNormal
##            4.099142e+03           -2.527770e+03            5.032501e+03
## xSaleConditionPartial
##            5.676453e+03
```

```r
# Let exclude them from the model
LM_train <-subset(M_train,  select=-c(Condition2,RoofMatl, Exterior1st,Exterior2nd,Foundation,

x_ols <- model.matrix(SalePrice~.,LM_train)[,-c(1,2)]
lm_model3 <- lm(y~ x_ols)
summary(lm_model3)
```

```
##
## Call:
## lm(formula = y ~ x_ols)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -287951  -13324    -924   12821  250223
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.545e+05  4.515e+04  -5.636 2.29e-08 ***
## x_olsMSSubClass            -2.408e+02  1.486e+02  -1.620 0.105546
## x_olsMSZoningFV             3.396e+04  1.741e+04   1.950 0.051463 .
## x_olsMSZoningRH             3.008e+04  1.895e+04   1.587 0.112753
## x_olsMSZoningRL             2.531e+04  1.516e+04   1.669 0.095418 .
## x_olsMSZoningRM             2.547e+04  1.407e+04   1.811 0.070503 .
## x_olsLotFrontage           -1.728e+02  6.574e+01  -2.628 0.008715 **
## x_olsLotArea                8.059e-01  1.785e-01   4.515 7.13e-06 ***
## x_olsStreetPave             2.550e+04  2.066e+04   1.235 0.217286
## x_olsLotShapeIR2            1.098e+04  7.040e+03   1.559 0.119321
## x_olsLotShapeIR3           -4.313e+04  1.518e+04  -2.842 0.004578 **
## x_olsLotShapeReg            4.498e+03  2.558e+03   1.759 0.078960 .
## x_olsLandContourHLS         2.420e+04  7.691e+03   3.147 0.001701 **
## x_olsLandContourLow         1.247e+04  1.179e+04   1.058 0.290356
## x_olsLandContourLvl         1.914e+04  5.726e+03   3.343 0.000861 ***
## x_olsLotConfigCulDSac       1.265e+04  6.241e+03   2.027 0.042947 *
## x_olsLotConfigFR2          -1.374e+04  6.765e+03  -2.031 0.042494 *
## x_olsLotConfigFR3          -2.490e+04  1.691e+04  -1.472 0.141335
## x_olsLotConfigInside       -2.510e+03  2.793e+03  -0.899 0.369131
```

```
## x_olsLandSlopeMod          5.493e+03  6.262e+03   0.877 0.380651    
## x_olsLandSlopeSev         -1.794e+04  1.794e+04  -1.000 0.317538    
## x_olsNeighborhoodBlueste   8.177e+03  2.569e+04   0.318 0.750287    
## x_olsNeighborhoodBrDale    2.819e+04  1.550e+04   1.819 0.069268 .  
## x_olsNeighborhoodBrkSide   1.117e+04  1.403e+04   0.796 0.426149    
## x_olsNeighborhoodClearCr   9.325e+03  1.474e+04   0.632 0.527242    
## x_olsNeighborhoodCollgCr   2.262e+02  1.070e+04   0.021 0.983142    
## x_olsNeighborhoodCrawfor   2.554e+04  1.257e+04   2.031 0.042559 *  
## x_olsNeighborhoodEdwards  -1.136e+04  1.175e+04  -0.967 0.333947    
## x_olsNeighborhoodGilbert   2.076e+03  1.157e+04   0.179 0.857692    
## x_olsNeighborhoodIDOTRR    9.743e+03  1.603e+04   0.608 0.543457    
## x_olsNeighborhoodMeadowV   1.345e+04  1.598e+04   0.841 0.400422    
## x_olsNeighborhoodMitchel   5.930e+02  1.238e+04   0.048 0.961795    
## x_olsNeighborhoodNAmes     1.920e+03  1.134e+04   0.169 0.865566    
## x_olsNeighborhoodNoRidge   4.673e+04  1.225e+04   3.816 0.000144 ***
## x_olsNeighborhoodNPkVill   1.995e+04  1.597e+04   1.249 0.212061    
## x_olsNeighborhoodNridgHt   5.014e+04  1.066e+04   4.703 2.94e-06 ***
## x_olsNeighborhoodNWAmes   -7.581e+03  1.182e+04  -0.642 0.521245    
## x_olsNeighborhoodOldTown  -5.843e+02  1.421e+04  -0.041 0.967198    
## x_olsNeighborhoodSawyer    8.136e+03  1.207e+04   0.674 0.500534    
## x_olsNeighborhoodSawyerW   4.698e+03  1.133e+04   0.415 0.678574    
## x_olsNeighborhoodSomerst   1.461e+04  1.279e+04   1.142 0.253852    
## x_olsNeighborhoodStoneBr   6.074e+04  1.200e+04   5.062 4.97e-07 ***
## x_olsNeighborhoodSWISU     1.258e+03  1.434e+04   0.088 0.930127    
## x_olsNeighborhoodTimber    3.924e+03  1.188e+04   0.330 0.741232    
## x_olsNeighborhoodVeenker   2.453e+04  1.618e+04   1.516 0.129850    
## x_olsCondition1Feedr      -9.881e+03  7.375e+03  -1.340 0.180655    
## x_olsCondition1Norm        7.708e+03  5.795e+03   1.330 0.183830    
## x_olsCondition1PosA        1.037e+04  1.799e+04   0.577 0.564219    
## x_olsCondition1PosN       -3.123e+04  1.234e+04  -2.530 0.011559 *  
## x_olsCondition1RRAe       -1.524e+04  1.382e+04  -1.102 0.270564    
## x_olsCondition1RRAn        7.476e+03  9.244e+03   0.809 0.418858    
## x_olsCondition1RRNe        2.418e+03  3.223e+04   0.075 0.940194    
## x_olsCondition1RRNn        1.237e+04  1.913e+04   0.647 0.518005    
## x_olsBldgType2fmCon        2.820e+04  2.171e+04   1.299 0.194306    
## x_olsBldgTypeDuplex        4.731e+03  1.227e+04   0.386 0.699865    
## x_olsBldgTypeTwnhs        -1.650e+04  1.698e+04  -0.972 0.331416    
## x_olsBldgTypeTwnhsE       -9.517e+03  1.587e+04  -0.600 0.548942    
## x_olsHouseStyle1.5Unf      1.404e+04  1.175e+04   1.195 0.232190    
## x_olsHouseStyle1Story      2.645e+04  6.898e+03   3.834 0.000134 ***
## x_olsHouseStyle2.5Fin     -1.748e+04  1.969e+04  -0.888 0.374825    
## x_olsHouseStyle2.5Unf     -9.390e+03  1.299e+04  -0.723 0.470000    
## x_olsHouseStyle2Story     -8.576e+03  5.301e+03  -1.618 0.106050    
## x_olsHouseStyleSFoyer      2.358e+04  1.043e+04   2.261 0.024010 *  
## x_olsHouseStyleSLvl        2.384e+04  9.006e+03   2.647 0.008258 ** 
## x_olsOverallQual           9.910e+03  1.544e+03   6.418 2.17e-10 ***
## x_olsOverallCond           4.828e+03  1.386e+03   3.485 0.000515 ***
## x_olsYearBuilt            -8.914e+01  1.097e+02  -0.812 0.416820    
## x_olsYearRemodAdd          1.087e+02  8.641e+01   1.258 0.208636    
## x_olsMasVnrTypeBrkFace     5.703e+03  1.132e+04   0.504 0.614533    
## x_olsMasVnrTypeNone        1.308e+04  1.133e+04   1.154 0.248811    
## x_olsMasVnrTypeStone       1.379e+04  1.175e+04   1.174 0.240513    
## x_olsMasVnrArea            2.338e+01  8.188e+00   2.855 0.004395 ** 
## x_olsExterQual             7.134e+03  3.163e+03   2.255 0.024353 *  
```

```
## x_olsExterCond          -1.988e+03  3.337e+03  -0.596 0.551488
## x_olsBsmtQual            9.054e+03  2.667e+03   3.395 0.000714 ***
## x_olsBsmtCond           -2.044e+03  3.739e+03  -0.547 0.584611
## x_olsBsmtExposure        5.913e+03  1.279e+03   4.623 4.29e-06 ***
## x_olsBsmtFinType1        1.975e+03  7.560e+02   2.613 0.009122 **
## x_olsBsmtFinSF1          4.810e+00  7.621e+00   0.631 0.528071
## x_olsBsmtFinType2        2.230e+02  1.933e+03   0.115 0.908206
## x_olsBsmtFinSF2          6.602e+00  1.267e+01   0.521 0.602318
## x_olsBsmtUnfSF          -5.526e-01  7.495e+00  -0.074 0.941246
## x_olsHeatingQC           2.174e+02  1.450e+03   0.150 0.880852
## x_olsCentralAir          9.118e+03  5.768e+03   1.581 0.114220
## x_olsElectricalFuseF     3.663e+03  9.879e+03   0.371 0.710836
## x_olsElectricalFuseP     4.893e+04  2.537e+04   1.929 0.054027 .
## x_olsElectricalMix      -2.328e+03  3.557e+04  -0.065 0.947841
## x_olsElectricalSBrkr    -3.468e+01  4.478e+03  -0.008 0.993823
## x_olsX1stFlrSF           4.494e+01  8.901e+00   5.049 5.30e-07 ***
## x_olsX2ndFlrSF           7.550e+01  8.406e+00   8.982  < 2e-16 ***
## x_olsLowQualFinSF        4.969e+01  3.212e+01   1.547 0.122257
## x_olsBsmtFullBath        4.184e+03  2.929e+03   1.429 0.153418
## x_olsBsmtHalfBath        2.255e+03  4.521e+03   0.499 0.618033
## x_olsFullBath            6.117e+03  3.396e+03   1.801 0.071974 .
## x_olsHalfBath            6.744e+03  3.198e+03   2.109 0.035189 *
## x_olsBedroomAbvGr       -4.632e+03  2.131e+03  -2.173 0.029989 *
## x_olsKitchenAbvGr       -2.047e+04  9.433e+03  -2.170 0.030215 *
## x_olsKitchenQual         8.822e+03  2.527e+03   3.492 0.000502 ***
## x_olsTotRmsAbvGrd        3.314e+03  1.422e+03   2.330 0.020030 *
## x_olsFunctional          5.207e+03  2.555e+03   2.039 0.041774 *
## x_olsFireplaces          3.033e+03  2.065e+03   1.469 0.142168
## x_olsGarageTypeAttchd    7.530e+03  1.614e+04   0.467 0.640845
## x_olsGarageTypeBasment   2.169e+04  1.863e+04   1.164 0.244699
## x_olsGarageTypeBuiltIn   3.373e+03  1.704e+04   0.198 0.843152
## x_olsGarageTypeCarPort   2.025e+04  2.130e+04   0.951 0.342030
## x_olsGarageTypeDetchd    1.186e+04  1.608e+04   0.737 0.461233
## x_olsGarageYrBlt         6.772e+01  8.803e+01   0.769 0.441969
## x_olsGarageFinish        2.858e+03  1.811e+03   1.578 0.114860
## x_olsGarageCars          1.424e+04  3.246e+03   4.388 1.27e-05 ***
## x_olsGarageArea         -1.054e+01  1.155e+01  -0.912 0.361839
## x_olsGarageQual          1.244e+04  5.191e+03   2.396 0.016772 *
## x_olsGarageCond         -4.743e+03  5.571e+03  -0.851 0.394775
## x_olsPavedDrive          1.752e+03  2.738e+03   0.640 0.522372
## x_olsWoodDeckSF          1.005e+01  9.171e+00   1.096 0.273180
## x_olsOpenPorchSF        -3.868e+00  1.805e+01  -0.214 0.830364
## x_olsEnclosedPorch      -5.135e+00  1.869e+01  -0.275 0.783627
## x_olsX3SsnPorch          4.867e+01  3.276e+01   1.486 0.137706
## x_olsScreenPorch         3.710e+01  1.801e+01   2.061 0.039605 *
## x_olsPoolArea           -7.127e+00  2.686e+01  -0.265 0.790828
## x_olsMiscVal            -1.843e+00  6.210e+00  -0.297 0.766760
## x_olsMoSold             -6.233e+02  3.767e+02  -1.655 0.098325 .
## x_olsYrSold              2.526e+01  7.894e+02   0.032 0.974477
## x_olsSaleTypeCon         2.502e+04  2.418e+04   1.035 0.301008
## x_olsSaleTypeConLD       2.549e+04  1.778e+04   1.433 0.152151
## x_olsSaleTypeConLI       1.897e+04  2.032e+04   0.934 0.350754
## x_olsSaleTypeConLw       1.291e+03  1.800e+04   0.072 0.942847
## x_olsSaleTypeCWD         9.828e+03  1.752e+04   0.561 0.575035
```

```
## x_olsSaleTypeNew           3.209e+04  2.155e+04   1.489 0.136818
## x_olsSaleTypeOth           3.627e+04  3.183e+04   1.140 0.254737
## x_olsSaleTypeWD            1.999e+03  6.484e+03   0.308 0.757958
## x_olsSaleConditionAdjLand  2.397e+04  3.344e+04   0.717 0.473648
## x_olsSaleConditionAlloca   3.972e+03  1.601e+04   0.248 0.804090
## x_olsSaleConditionFamily  -9.718e+01  8.891e+03  -0.011 0.991282
## x_olsSaleConditionNormal   3.692e+03  4.472e+03   0.826 0.409224
## x_olsSaleConditionPartial -1.364e+04  2.068e+04  -0.660 0.509648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30840 on 959 degrees of freedom
## Multiple R-squared:  0.8794, Adjusted R-squared:  0.8625
## F-statistic: 52.17 on 134 and 959 DF,  p-value: < 2.2e-16
```

In the first model :about 12 variable coefficients was Na due to singularities,This is because the information given by these variables is already contained in the other variables and thus redundant.

In the end the model the third model is better than the first one as it explains about (& percent in the variation of saleprices. Now let move further and compare it to other models.

Ridge Regression

First set up a grid of possible values of lambda.

```
grid <- 10^seq(10,-2,length=100)
```

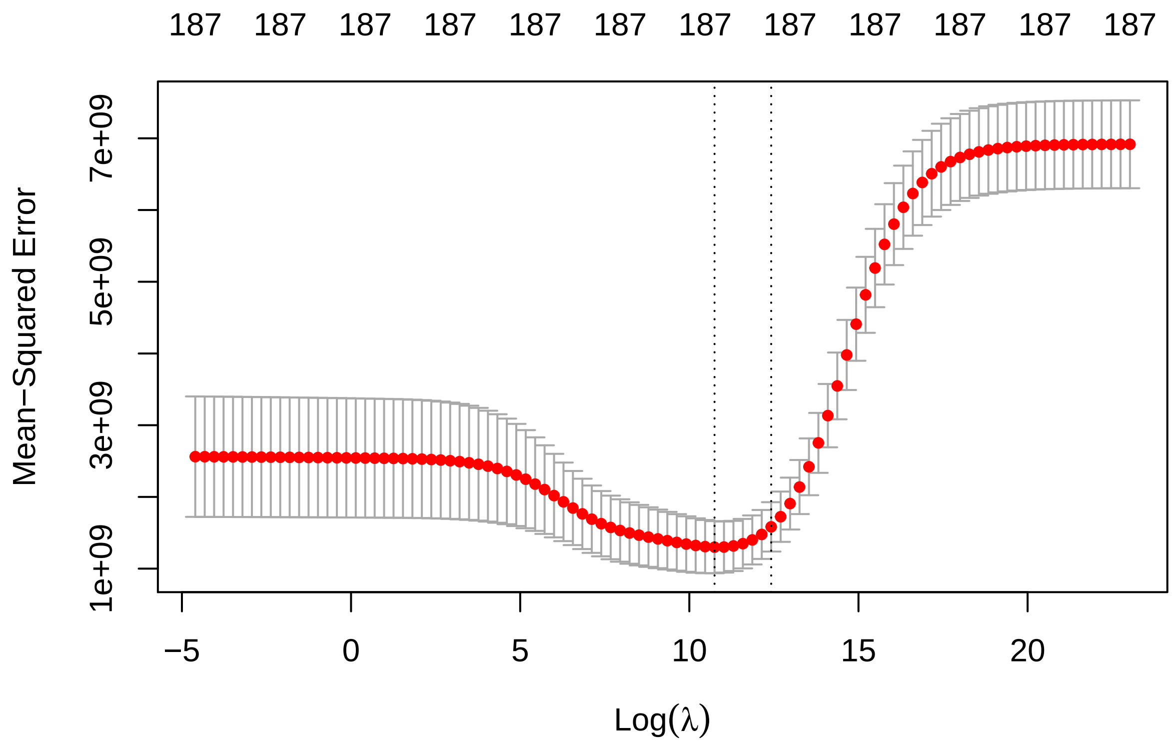
```
ridg_mode <- glmnet(x,y,alpha=0,lambda=grid)#Now estimating the model using the lambda grid.
summary(ridg_mode)
```

```
##           Length Class     Mode
## a0          100  -none-    numeric
## beta      19500  dgCMatrix S4
## df          100  -none-    numeric
## dim           2  -none-    numeric
## lambda      100  -none-    numeric
## dev.ratio   100  -none-    numeric
## nulldev       1  -none-    numeric
## npasses       1  -none-    numeric
## jerr          1  -none-    numeric
## offset        1  -none-    logical
## call          5  -none-    call
## nobs          1  -none-    numeric
```

Choosing Optimal Lambda Value

The glmnet function trains the model multiple times for all the different values oflambda which we pass as a sequence of vector to the lambda = argument in the glmnet function. The next task is to identify the optimal value of lambda which results into minimum error. This can be achieved automatically by using cv.glmnet() function.

```
ridge_cv <- cv.glmnet(x, y, alpha = 0, lambda = grid) # Using cross validation glmnet
best_lambda <- ridge_cv$lambda.min # Best lambda value
plot(ridge_cv)
```




```r
best_lambda
```

```
## [1] 46415.89
```

Building the final model

```r
best_ridge <- glmnet(x, y, alpha = 0, lambda = best_lambda) #46415.89
coef(best_ridge)
```

```
## 196 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## (Intercept)          -1.231713e+05
## MSSubClass           -6.776005e+01
## MSZoningFV            3.504390e+03
## MSZoningRH            9.296813e+02
## MSZoningRL            2.172952e+03
## MSZoningRM           -1.413998e+03
## LotFrontage           3.711968e+01
## LotArea               4.359014e-01
## StreetPave            2.271574e+04
## LotShapeIR2           8.799375e+03
## LotShapeIR3          -3.122544e+04
## LotShapeReg          -1.547669e+03
## LandContourHLS        7.546740e+03
## LandContourLow       -3.080692e+03
## LandContourLvl        3.050696e+03
```

```
## UtilitiesNoSeWa          .
## LotConfigCulDSac      1.371179e+04
## LotConfigFR2         -6.497809e+03
## LotConfigFR3         -9.496980e+03
## LotConfigInside      -1.152355e+03
## LandSlopeMod          4.277189e+03
## LandSlopeSev         -5.214543e+03
## NeighborhoodBlueste  -4.259441e+03
## NeighborhoodBrDale    1.149464e+03
## NeighborhoodBrkSide   4.241114e+03
## NeighborhoodClearCr  -2.690120e+03
## NeighborhoodCollgCr  -6.171089e+03
## NeighborhoodCrawfor   1.352176e+04
## NeighborhoodEdwards  -1.248806e+04
## NeighborhoodGilbert  -8.820287e+03
## NeighborhoodIDOTRR   -4.146089e+03
## NeighborhoodMeadowV  -1.665919e+04
## NeighborhoodMitchel  -5.072821e+03
## NeighborhoodNAmes    -4.474075e+03
## NeighborhoodNoRidge   2.861130e+04
## NeighborhoodNPkVill   1.379732e+03
## NeighborhoodNridgHt   2.298131e+04
## NeighborhoodNWAmes   -7.801588e+03
## NeighborhoodOldTown  -2.783990e+03
## NeighborhoodSawyer   -8.108692e+02
## NeighborhoodSawyerW  -2.230821e+03
## NeighborhoodSomerst   2.094351e+03
## NeighborhoodStoneBr   3.343692e+04
## NeighborhoodSWISU    -4.958002e+03
## NeighborhoodTimber   -1.391042e+03
## NeighborhoodVeenker   1.285153e+04
## Condition1Feedr      -8.729074e+03
## Condition1Norm        5.261474e+03
## Condition1PosA        6.073223e+03
## Condition1PosN       -1.253522e+04
## Condition1RRAe       -7.392642e+03
## Condition1RRAn        3.372722e+03
## Condition1RRNe       -4.056769e+03
## Condition1RRNn        1.511007e+03
## Condition2Feedr       1.179435e+03
## Condition2Norm        7.581728e+03
## Condition2PosA        4.536447e+04
## Condition2PosN       -8.499715e+04
## Condition2RRAe        .
## Condition2RRAn        .
## Condition2RRNn        1.413664e+04
## BldgType2fmCon       -2.009085e+03
## BldgTypeDuplex       -6.065732e+03
## BldgTypeTwnhs        -1.144331e+04
## BldgTypeTwnhsE       -8.523250e+03
## HouseStyle1.5Unf      5.393754e+02
## HouseStyle1Story      9.292122e+01
## HouseStyle2.5Fin     -2.259268e+01
## HouseStyle2.5Unf     -3.534718e+03
```

```
## HouseStyle2Story      8.596223e+02
## HouseStyleSFoyer     -2.985677e+03
## HouseStyleSLvl       -4.316925e+03
## OverallQual           5.914698e+03
## OverallCond           2.478144e+03
## YearBuilt            -3.512039e+01
## YearRemodAdd         -9.074422e+01
## RoofStyleGable       -4.119492e+03
## RoofStyleGambrel      8.487228e+02
## RoofStyleHip          4.316499e+03
## RoofStyleMansard      6.908563e+03
## RoofStyleShed         .
## RoofMatlCompShg       1.082669e+04
## RoofMatlMembran       3.151750e+04
## RoofMatlMetal         .
## RoofMatlRoll          3.524196e+03
## RoofMatlTar&Grv      -6.094048e+03
## RoofMatlWdShake       2.271523e+03
## RoofMatlWdShngl       7.276795e+04
## Exterior1stAsphShn    .
## Exterior1stBrkComm   -2.074870e+04
## Exterior1stBrkFace    8.480370e+03
## Exterior1stCBlock    -2.121646e+03
## Exterior1stCemntBd    7.804696e+03
## Exterior1stHdBoard   -1.980073e+03
## Exterior1stImStucc   -1.951845e+04
## Exterior1stMetalSd    1.340977e+03
## Exterior1stPlywood   -1.467179e+03
## Exterior1stStone      9.864642e+03
## Exterior1stStucco    -6.273109e+03
## Exterior1stVinylSd   -3.202190e+02
## Exterior1stWd Sdng   -2.220580e+02
## Exterior1stWdShing   -2.274192e+03
## Exterior2ndAsphShn    7.937147e+02
## Exterior2ndBrk Cmn   -4.865954e+02
## Exterior2ndBrkFace   -1.730263e+03
## Exterior2ndCBlock    -2.088221e+03
## Exterior2ndCmentBd    7.574818e+03
## Exterior2ndHdBoard   -1.013174e+03
## Exterior2ndImStucc    2.525099e+04
## Exterior2ndMetalSd    7.523235e+02
## Exterior2ndOther     -3.543300e+03
## Exterior2ndPlywood   -2.683773e+03
## Exterior2ndStone      2.082516e+03
## Exterior2ndStucco    -1.238042e+04
## Exterior2ndVinylSd    3.024473e+02
## Exterior2ndWd Sdng    1.135878e+03
## Exterior2ndWd Shng   -5.312137e+03
## MasVnrTypeBrkFace    -1.739953e+03
## MasVnrTypeNone        4.538950e+02
## MasVnrTypeStone       4.573698e+03
## MasVnrArea            2.252380e+01
## ExterQual             7.713792e+03
## ExterCond            -1.179318e+03
```

```
## FoundationCBlock     -1.101075e+03
## FoundationPConc       2.252367e+03
## FoundationSlab        .
## FoundationStone       3.766279e+02
## FoundationWood       -2.468595e+04
## BsmtQual              7.056377e+03
## BsmtCond              9.836992e+02
## BsmtExposure          4.253972e+03
## BsmtFinType1          1.459824e+03
## BsmtFinSF1            6.912746e+00
## BsmtFinType2          8.771303e+01
## BsmtFinSF2            3.484467e+00
## BsmtUnfSF            -4.684460e-01
## TotalBsmtSF           8.747060e+00
## HeatingGasA          -3.523359e+03
## HeatingGasW           4.784253e+03
## HeatingGrav           2.280724e+03
## HeatingOthW          -1.415787e+04
## HeatingWall           .
## HeatingQC             1.507491e+03
## CentralAir            5.295171e+03
## ElectricalFuseF      -1.949478e+03
## ElectricalFuseP       1.854105e+04
## ElectricalMix         4.955847e+02
## ElectricalSBrkr      -1.824850e+02
## X1stFlrSF             1.230196e+01
## X2ndFlrSF             1.054679e+01
## LowQualFinSF          1.126831e+00
## GrLivArea             1.402504e+01
## BsmtFullBath          3.739673e+03
## BsmtHalfBath         -7.932428e+02
## FullBath              6.366950e+03
## HalfBath              4.348333e+03
## BedroomAbvGr          3.591924e+02
## KitchenAbvGr         -1.079394e+04
## KitchenQual           7.352781e+03
## TotRmsAbvGrd          3.213693e+03
## Functional            3.650994e+03
## Fireplaces            6.182683e+03
## GarageTypeAttchd     -4.863024e+02
## GarageTypeBasment     3.715371e+03
## GarageTypeBuiltIn     6.466776e+03
## GarageTypeCarPort    -5.508710e+02
## GarageTypeDetchd     -7.191334e+02
## GarageYrBlt          -1.093609e+01
## GarageFinish          2.179419e+03
## GarageCars            7.200952e+03
## GarageArea            1.713925e+01
## GarageQual            3.734938e+03
## GarageCond            2.946602e+02
## PavedDrive            9.337044e+02
## WoodDeckSF            1.554664e+01
## OpenPorchSF           9.999360e+00
## EnclosedPorch        -1.979777e+00
```

```
## X3SsnPorch           3.435153e+01
## ScreenPorch          2.977282e+01
## PoolArea             4.460921e+00
## MiscVal             -1.829391e+00
## MoSold              -1.659160e+02
## YrSold              -2.040861e+02
## SaleTypeCon          1.845478e+04
## SaleTypeConLD        6.775807e+03
## SaleTypeConLI        4.713647e+02
## SaleTypeConLw       -3.099478e+03
## SaleTypeCWD          1.170887e+04
## SaleTypeNew          6.257280e+03
## SaleTypeOth          1.300726e+04
## SaleTypeWD          -2.798435e+03
## SaleConditionAdjLand  1.336588e+04
## SaleConditionAlloca   3.773978e+03
## SaleConditionFamily  -6.242111e+03
## SaleConditionNormal   2.557717e+02
## SaleConditionPartial  5.702062e+03
```

Now let test our model by spliting the training dataset into test and trining sets

```r
set.seed(1)
train <- sample(1:nrow(x), nrow(x)/2)
x_test <- x[-train,]
y_test <-  y[-train]
```

Prediction

```r
#colSums(is.na(test))
ridge_mode <- glmnet(x[train,],y[train],alpha= 0,lambda=best_lambda)
ridg_pred <- predict(ridg_mode,s= best_lambda,newx= x_test)
mean((ridg_pred-y_test)^2)
```

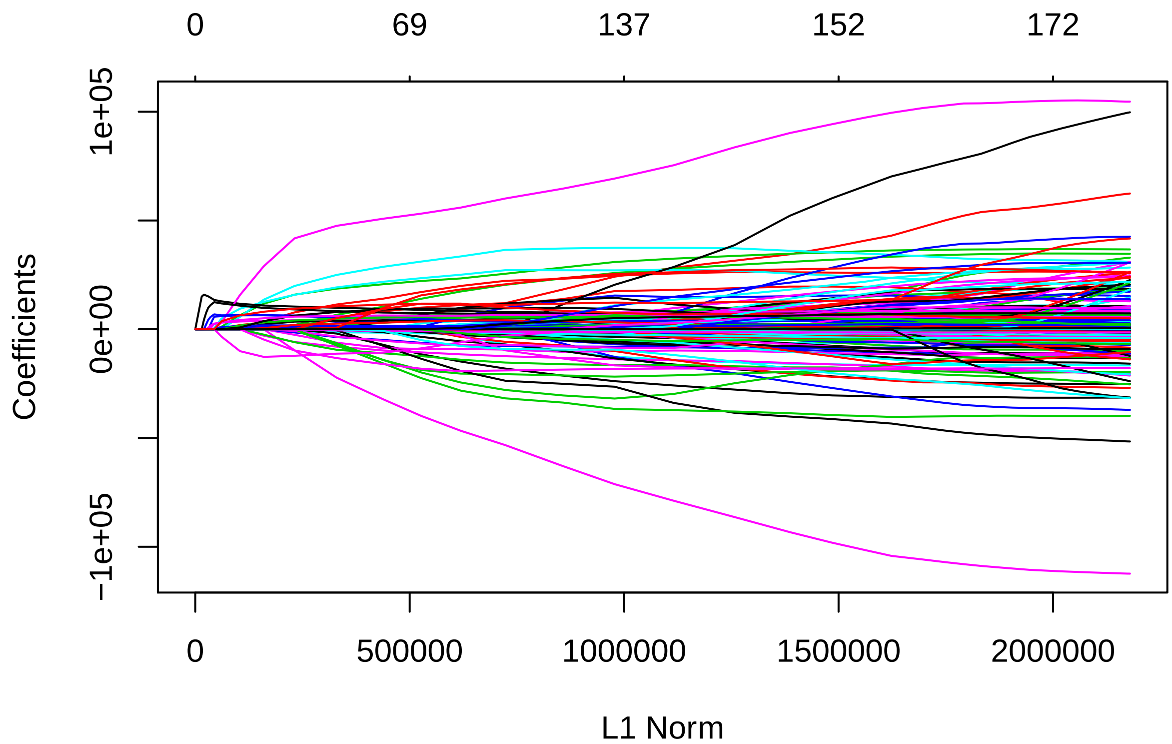
```
## [1] 1018467617
```

The Lasso

The ridge regression shrinks our coefficients but does not perform variable selection. Let's try the lasso which can also be done using `glmnet()` but now with the option `alpha = 1`.

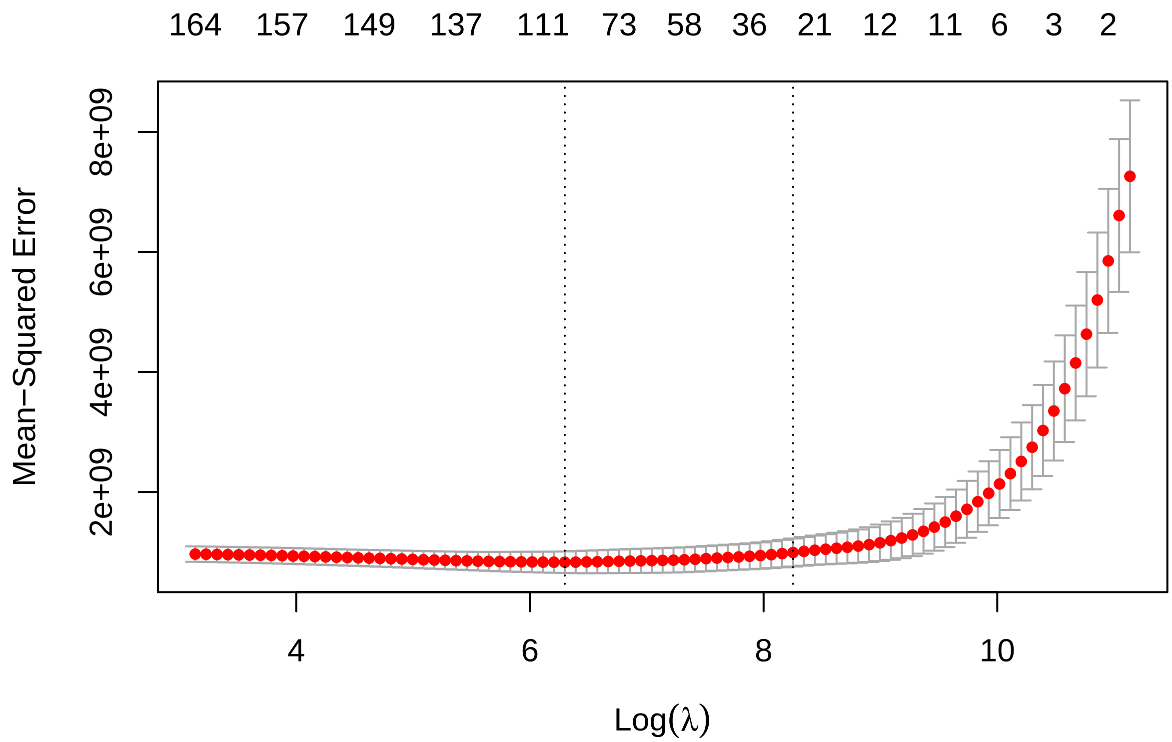
```r
lasso_mod <- glmnet(x[train,],y[train],alpha=1,lambda=grid)
plot(lasso_mod)
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values
```



We now perform cross validation to find out the optimal lambda for the lasso.

```
set.seed(1)
cv_out <- cv.glmnet(x[train,],y[train],alpha=1)
plot(cv_out)
```




```
bestlam <- cv_out$lambda.min
lasso_pred <- predict(lasso_mod,s=bestlam,newx=x_test)
mean((lasso_pred-y_test)^2)
```

```
## [1] 2094012917
```

```
out <- glmnet(x,y,alpha=1,lambda=grid)
lasso_coef <- predict(out,type="coefficients",s=bestlam)[1:196,]
lasso_coef[lasso_coef!=0]
```

```
##          (Intercept)           MSSubClass           MSZoningFV
##        -3.342937e+05        -2.162911e+02         3.674940e+03
##              LotArea           StreetPave          LotShapeIR2
##         5.456427e-01         1.585530e+04         7.712916e+03
##          LotShapeIR3      LandContourHLS       LandContourLow
##        -2.341909e+04         6.357568e+03        -2.866966e+03
##       LandContourLvl     LotConfigCulDSac          LotConfigFR2
##         4.636402e+03         1.567472e+04        -5.292392e+03
##         LotConfigFR3         LandSlopeMod  NeighborhoodBrDale
##        -9.976253e+03         1.216695e+02         5.313177e+03
##  NeighborhoodBrkSide NeighborhoodCollgCr NeighborhoodCrawfor
##         5.304543e+03        -4.687682e+01         1.913064e+04
##  NeighborhoodEdwards NeighborhoodMeadowV NeighborhoodNoRidge
##        -6.510859e+03        -2.282451e+01         3.529132e+04
##  NeighborhoodNPkVill NeighborhoodNridgHt   NeighborhoodNWAmes
##         3.778088e+03         3.915711e+04        -7.440074e+03
##  NeighborhoodOldTown  NeighborhoodSawyer  NeighborhoodSomerst
```

```
##        -2.643010e+03         2.994687e+03         8.260096e+03
##  NeighborhoodStoneBr  NeighborhoodVeenker      Condition1Feedr
##         5.067308e+04         8.312948e+03        -4.984120e+03
##       Condition1Norm       Condition1PosA       Condition1PosN
##         6.640908e+03         5.532603e+03        -1.028664e+03
##       Condition1RRAn       Condition2PosA       Condition2PosN
##         3.135127e+03         2.807981e+04        -1.729406e+05
##       Condition2RRNn       BldgType2fmCon        BldgTypeTwnhs
##         4.633167e+03         1.489029e+04        -3.393348e+03
##       BldgTypeTwnhsE     HouseStyle1Story     HouseStyle2.5Fin
##        -1.724210e+02         2.191727e+03        -4.190551e+03
##     HouseStyle2.5Unf          OverallQual          OverallCond
##        -6.374944e+03         1.063180e+04         4.152662e+03
##            YearBuilt       RoofStyleGable     RoofStyleMansard
##        -4.005052e+01        -4.749420e+03         7.048278e+03
##      RoofMatlCompShg      RoofMatlMembran         RoofMatlRoll
##         1.969219e+05         1.925244e+05         1.716867e+05
##      RoofMatlTar&Grv      RoofMatlWdShake      RoofMatlWdShngl
##         1.591756e+05         1.606732e+05         2.691141e+05
##   Exterior1stBrkComm   Exterior1stBrkFace   Exterior1stCemntBd
##        -1.218320e+04         6.233022e+03         6.279314e+03
##   Exterior1stImStucc   Exterior1stMetalSd   Exterior2ndCmentBd
##        -2.709546e+04         3.446405e+03         5.283685e+03
##   Exterior2ndImStucc     Exterior2ndOther   Exterior2ndPlywood
##         2.610797e+04        -9.388783e+03        -2.883992e+01
##    Exterior2ndStucco   Exterior2ndWd Shng    MasVnrTypeBrkFace
##        -1.318042e+04        -9.424204e+03        -4.974812e+03
##           MasVnrArea            ExterQual            ExterCond
##         2.422800e+01         6.300100e+03        -9.846125e+02
##     FoundationCBlock       FoundationWood             BsmtQual
##         6.617495e+01        -2.822540e+04         9.201198e+03
##         BsmtExposure         BsmtFinType1           BsmtFinSF1
##         6.130396e+03         1.462046e+03         1.184664e+01
##           BsmtFinSF2         HeatingOthW            HeatingQC
##         7.169435e+00        -2.031657e+04         1.825080e+02
##           CentralAir     ElectricalFuseP         LowQualFinSF
##         4.885233e+03         1.577054e+04        -1.935289e+01
##            GrLivArea         BsmtFullBath             FullBath
##         5.158925e+01         2.373056e+03         2.878284e+03
##             HalfBath         BedroomAbvGr         KitchenAbvGr
##         1.420007e+03        -1.933965e+03        -1.205347e+04
##          KitchenQual         TotRmsAbvGrd           Functional
##         7.904069e+03         1.437718e+03         4.550845e+03
##           Fireplaces     GarageTypeAttchd    GarageTypeBasment
##         3.069883e+03        -1.005623e+03         7.343665e+02
##    GarageTypeBuiltIn         GarageFinish           GarageCars
##         1.280528e+03         9.741130e+02         9.964081e+03
##           GarageArea           GarageQual           WoodDeckSF
##         2.222826e+00         2.287017e+03         9.544718e-01
##        EnclosedPorch           X3SsnPorch          ScreenPorch
##        -1.696802e+00         2.407308e+01         2.420487e+01
##               MoSold          SaleTypeCon        SaleTypeConLD
##        -2.886205e+02         1.830989e+04         7.716762e+03
##        SaleTypeConLw          SaleTypeNew          SaleTypeOth
```

```
##        -9.257052e+00         1.555559e+04         1.329819e+04
##  SaleConditionFamily SaleConditionPartial
##        -3.565096e+03         3.586534e+02
```

The ridge regression model gives us a better MSE than the lasso Reg model, this might be due to CV, may be for the lasso we were not able to capture the real optimal lambda or it might be that the ridge model was simply better suited for this dataset. Moving on let us build some tree models

Tree

```r
library(tree)
```

```
## Warning: package 'tree' was built under R version 3.6.2

## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```r
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.6.1
```

```r
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.6.1
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```r
library(rpart)
library(rpart.plot)
M_train <- M_train[,-1]# to exclude the id column
tree_hp <- rpart(SalePrice~.-SalePrice, M_train[,-1])
summary(tree_hp)
```

```
## Call:
## rpart(formula = SalePrice ~ . - SalePrice, data = M_train[, -1])
##   n= 1094
##
##            CP nsplit rel error    xerror       xstd
## 1  0.48566437      0 1.0000000 1.0014704 0.08769331
## 2  0.10851669      1 0.5143356 0.5172981 0.04762325
## 3  0.06359187      2 0.4058189 0.4090612 0.04609972
## 4  0.02561745      3 0.3422271 0.3633335 0.03396268
## 5  0.02408726      4 0.3166096 0.3815696 0.03996001
## 6  0.02242106      5 0.2925224 0.3747882 0.03938278
## 7  0.02049337      6 0.2701013 0.3581093 0.03570380
## 8  0.01113014      7 0.2496079 0.3248501 0.03397578
## 9  0.01071599      8 0.2384778 0.3170078 0.03338860
## 10 0.01000000      9 0.2277618 0.3085471 0.03306403
##
## Variable importance
```

```
##  OverallQual Neighborhood  TotalBsmtSF   GarageCars     BsmtQual 
##           31           11            8            8            8 
##   GarageArea    GrLivArea    ExterQual    X2ndFlrSF    YearBuilt 
##            7            4            4            3            2 
## TotRmsAbvGrd  KitchenQual  GarageYrBlt   HouseStyle    X1stFlrSF 
##            2            2            2            2            1 
##      LotArea  Exterior2nd   GarageType     FullBath 
##            1            1            1            1 
## 
## Node number 1: 1094 observations,    complexity param=0.4856644
##   mean=187033.3, MSE=6.91015e+09 
##   left son=2 (897 obs) right son=3 (197 obs)
##   Primary splits:
##       OverallQual  < 7.5    to the left,  improve=0.4856644, (0 missing)
##       GarageCars   < 2.5    to the left,  improve=0.3905763, (0 missing)
##       ExterQual    < 2.5    to the left,  improve=0.3864520, (0 missing)
##       Neighborhood splits as  LLLLLLLLLLLLLRLRLLLLRRLRR, improve=0.3820238, (0 missing)
##       YearBuilt    < 34.5   to the right, improve=0.3483743, (0 missing)
##   Surrogate splits:
##       Neighborhood splits as  LLLLLLLLLLLLLRLRLLLLLRLRR, agree=0.887, adj=0.371, (0 split)
##       GarageCars   < 2.5    to the left,  agree=0.885, adj=0.360, (0 split)
##       BsmtQual     < 3.5    to the left,  agree=0.879, adj=0.330, (0 split)
##       TotalBsmtSF  < 1560.5 to the left,  agree=0.878, adj=0.325, (0 split)
##       GarageArea   < 679    to the left,  agree=0.878, adj=0.320, (0 split)
## 
## Node number 2: 897 observations,    complexity param=0.1085167
##   mean=159884.6, MSE=2.337453e+09 
##   left son=4 (638 obs) right son=5 (259 obs)
##   Primary splits:
##       OverallQual  < 6.5    to the left,  improve=0.3912605, (0 missing)
##       FullBath     < 1.5    to the left,  improve=0.3551330, (0 missing)
##       Neighborhood splits as  RLLLRRRLRLLLLRLRRLLRRRLRR, improve=0.3509965, (0 missing)
##       GrLivArea    < 1413   to the left,  improve=0.3166767, (0 missing)
##       YearBuilt    < 35.5   to the right, improve=0.3164938, (0 missing)
##   Surrogate splits:
##       ExterQual    < 2.5    to the left,  agree=0.846, adj=0.467, (0 split)
##       YearBuilt    < 34.5   to the right, agree=0.836, adj=0.432, (0 split)
##       Neighborhood splits as  RLLLLRLLRLLLLRLRLLLLRRLRL, agree=0.816, adj=0.363, (0 split)
##       GarageYrBlt  < 22.5   to the right, agree=0.812, adj=0.347, (0 split)
##       KitchenQual  < 2.5    to the left,  agree=0.781, adj=0.243, (0 split)
## 
## Node number 3: 197 observations,    complexity param=0.06359187
##   mean=310649.3, MSE=9.094061e+09 
##   left son=6 (139 obs) right son=7 (58 obs)
##   Primary splits:
##       OverallQual  < 8.5    to the left,  improve=0.2683381, (0 missing)
##       TotRmsAbvGrd < 9.5    to the left,  improve=0.2454988, (0 missing)
##       GrLivArea    < 1971.5 to the left,  improve=0.2353446, (0 missing)
##       TotalBsmtSF  < 1846   to the left,  improve=0.2289771, (0 missing)
##       X1stFlrSF    < 1685   to the left,  improve=0.2243715, (0 missing)
##   Surrogate splits:
##       ExterQual   < 3.5    to the left,  agree=0.853, adj=0.500, (0 split)
##       KitchenQual < 3.5    to the left,  agree=0.802, adj=0.328, (0 split)
##       BsmtQual    < 3.5    to the left,  agree=0.772, adj=0.224, (0 split)
```

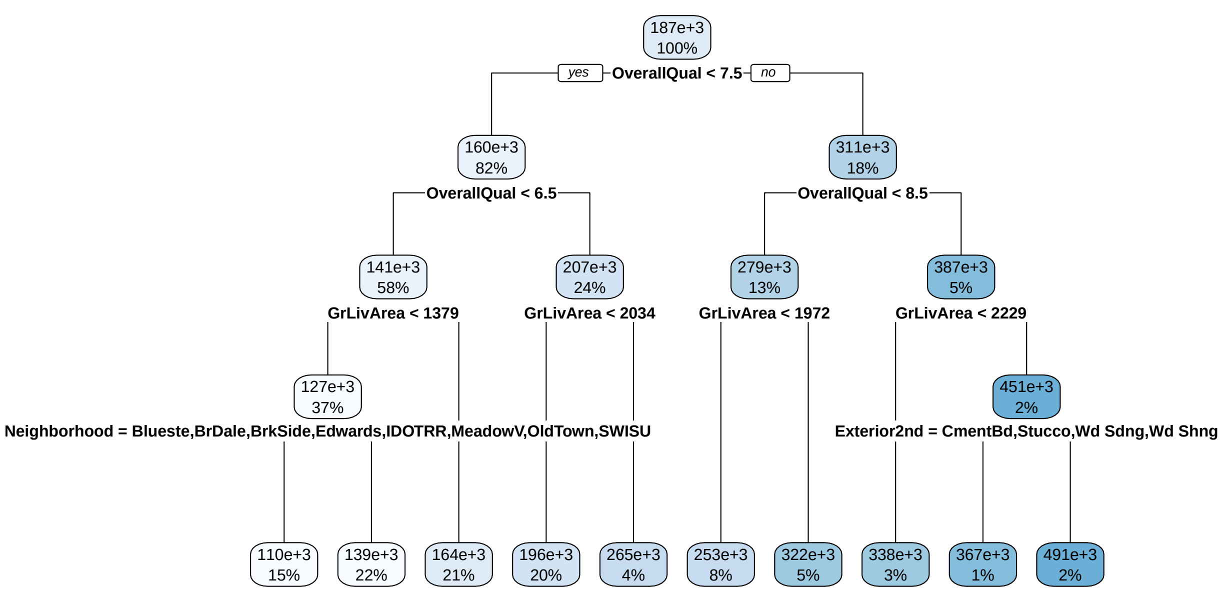
```
##       TotalBsmtSF < 1720.5 to the left,  agree=0.772, adj=0.224, (0 split)
##       X1stFlrSF   < 1723.5 to the left,  agree=0.766, adj=0.207, (0 split)
##
## Node number 4: 638 observations,    complexity param=0.02561745
##   mean=140616.3, MSE=1.162121e+09
##   left son=8 (404 obs) right son=9 (234 obs)
##   Primary splits:
##       GrLivArea    < 1378.5 to the left,  improve=0.2611974, (0 missing)
##       FullBath     < 1.5    to the left,  improve=0.2331100, (0 missing)
##       Neighborhood splits as  -LLLRRRLRLLLL-LRRLLRR-LRR, improve=0.2020836, (0 missing)
##       OverallQual  < 5.5    to the left,  improve=0.1998806, (0 missing)
##       GarageCars   < 1.5    to the left,  improve=0.1896599, (0 missing)
##   Surrogate splits:
##       TotRmsAbvGrd < 6.5    to the left,  agree=0.856, adj=0.607, (0 split)
##       X2ndFlrSF    < 567.5  to the left,  agree=0.823, adj=0.517, (0 split)
##       FullBath     < 1.5    to the left,  agree=0.790, adj=0.427, (0 split)
##       HouseStyle   splits as  RLLRRRLL,   agree=0.755, adj=0.333, (0 split)
##       X1stFlrSF    < 1366.5 to the left,  agree=0.754, adj=0.329, (0 split)
##
## Node number 5: 259 observations,    complexity param=0.02242106
##   mean=207348.6, MSE=2.065283e+09
##   left son=10 (217 obs) right son=11 (42 obs)
##   Primary splits:
##       GrLivArea   < 2033.5 to the left,  improve=0.3168704, (0 missing)
##       X2ndFlrSF   < 947.5  to the left,  improve=0.2445362, (0 missing)
##       BsmtFinSF1  < 955.5  to the left,  improve=0.2191863, (0 missing)
##       LotFrontage < 65.5   to the left,  improve=0.1896303, (0 missing)
##       LotArea     < 9637.5 to the left,  improve=0.1814999, (0 missing)
##   Surrogate splits:
##       X2ndFlrSF    < 976.5  to the left,  agree=0.927, adj=0.548, (0 split)
##       TotRmsAbvGrd < 8.5    to the left,  agree=0.880, adj=0.262, (0 split)
##       Neighborhood splits as  L--LLLLLLL-LRR-LLL-RLLLL-, agree=0.861, adj=0.143, (0 split)
##       Exterior2nd  splits as  R--R-LLRL-L-LLLL, agree=0.857, adj=0.119, (0 split)
##       X1stFlrSF    < 1997.5 to the left,  agree=0.857, adj=0.119, (0 split)
##
## Node number 6: 139 observations,    complexity param=0.02049337
##   mean=278739.4, MSE=3.978058e+09
##   left son=12 (87 obs) right son=13 (52 obs)
##   Primary splits:
##       GrLivArea  < 1971.5 to the left,  improve=0.2801768, (0 missing)
##       BsmtFinSF1 < 1325   to the left,  improve=0.2066180, (0 missing)
##       X1stFlrSF  < 1677   to the left,  improve=0.1728489, (0 missing)
##       WoodDeckSF < 238.5  to the left,  improve=0.1637139, (0 missing)
##       GarageCars < 2.5    to the left,  improve=0.1546595, (0 missing)
##   Surrogate splits:
##       X2ndFlrSF    < 874.5  to the left,  agree=0.827, adj=0.538, (0 split)
##       BedroomAbvGr < 3.5    to the left,  agree=0.813, adj=0.500, (0 split)
##       TotRmsAbvGrd < 7.5    to the left,  agree=0.806, adj=0.481, (0 split)
##       Neighborhood splits as  L----LL-L---RR-LLL-LLL-LL, agree=0.755, adj=0.346, (0 split)
##       HouseStyle   splits as  R-L--R-R, agree=0.755, adj=0.346, (0 split)
##
## Node number 7: 58 observations,    complexity param=0.02408726
##   mean=387123.3, MSE=1.306628e+10
##   left son=14 (33 obs) right son=15 (25 obs)
```

```
##   Primary splits:
##       GrLivArea    < 2229   to the left,  improve=0.2402770, (0 missing)
##       BedroomAbvGr < 3.5    to the left,  improve=0.2296375, (0 missing)
##       Neighborhood splits as  -----L-LL----R-L-L--LR-LL, improve=0.2223822, (0 missing)
##       FullBath     < 2.5    to the left,  improve=0.1881310, (0 missing)
##       TotRmsAbvGrd < 9.5    to the left,  improve=0.1851180, (0 missing)
##   Surrogate splits:
##       HouseStyle   splits as  --LRRR--,   agree=0.879, adj=0.72, (0 split)
##       X2ndFlrSF    < 284    to the left,  agree=0.879, adj=0.72, (0 split)
##       TotRmsAbvGrd < 8.5    to the left,  agree=0.879, adj=0.72, (0 split)
##       GarageType   splits as  -L-R-R,     agree=0.810, adj=0.56, (0 split)
##       LotArea      < 13379  to the left,  agree=0.776, adj=0.48, (0 split)
##
## Node number 8: 404 observations,    complexity param=0.01071599
##   mean=127356.8, MSE=6.960165e+08
##   left son=16 (161 obs) right son=17 (243 obs)
##   Primary splits:
##       Neighborhood splits as  -LLLRRRLRLLRR-RRRLRRR-LRR, improve=0.2880954, (0 missing)
##       X1stFlrSF    < 1051   to the left,  improve=0.2734242, (0 missing)
##       TotalBsmtSF  < 1007.5 to the left,  improve=0.2639977, (0 missing)
##       YearBuilt    < 70.5   to the right, improve=0.2227211, (0 missing)
##       MSZoning     splits as  LRLRL, improve=0.1885122, (0 missing)
##   Surrogate splits:
##       MSZoning    splits as  LRRRL,      agree=0.847, adj=0.615, (0 split)
##       YearBuilt   < 71.5   to the right, agree=0.809, adj=0.522, (0 split)
##       TotalBsmtSF < 813.5  to the left,  agree=0.748, adj=0.366, (0 split)
##       LotFrontage < 60.5   to the left,  agree=0.745, adj=0.360, (0 split)
##       LotArea     < 6510   to the left,  agree=0.745, adj=0.360, (0 split)
##
## Node number 9: 234 observations
##   mean=163508.7, MSE=1.139239e+09
##
## Node number 10: 217 observations
##   mean=196094.2, MSE=1.275193e+09
##
## Node number 11: 42 observations
##   mean=265496.8, MSE=2.111781e+09
##
## Node number 12: 87 observations
##   mean=252929, MSE=2.400348e+09
##
## Node number 13: 52 observations
##   mean=321922, MSE=3.638386e+09
##
## Node number 14: 33 observations
##   mean=338354.2, MSE=1.855883e+09
##
## Node number 15: 25 observations,    complexity param=0.01113014
##   mean=451498.6, MSE=2.05803e+10
##   left son=30 (8 obs) right son=31 (17 obs)
##   Primary splits:
##       Exterior2nd  splits as  -----LRR----LRLL, improve=0.1635361, (0 missing)
##       Neighborhood splits as  -------L-----R-L-L---R---, improve=0.1596074, (0 missing)
##       OpenPorchSF  < 121    to the right, improve=0.1573872, (0 missing)
```

```
##       GarageArea   < 836    to the right, improve=0.1475578, (0 missing)
##       TotalBsmtSF  < 1702   to the left,  improve=0.1350297, (0 missing)
##   Surrogate splits:
##       OpenPorchSF  < 215    to the right, agree=0.88, adj=0.625, (0 split)
##       Neighborhood splits as  -------L-----R-R-L---R---, agree=0.84, adj=0.500, (0 split)
##       Condition1   splits as  LLR-L----, agree=0.84, adj=0.500, (0 split)
##       MSZoning     splits as  ---RL, agree=0.76, adj=0.250, (0 split)
##       LotArea      < 18927  to the right, agree=0.76, adj=0.250, (0 split)
##
## Node number 16: 161 observations
##   mean=109960, MSE=6.108352e+08
##
## Node number 17: 243 observations
##   mean=138883, MSE=4.190801e+08
##
## Node number 30: 8 observations
##   mean=366929.4, MSE=1.725688e+10
##
## Node number 31: 17 observations
##   mean=491295.8, MSE=1.719481e+10
```

```r
rpart.plot(tree_hp)
```

```
## Warning: Bad 'data' field in model 'call' (expected a data.frame or a matrix).
## To silence this warning:
##     Call rpart.plot with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```



```r
tree_hp$variable.importance
```

```
##  OverallQual Neighborhood  TotalBsmtSF   GarageCars     BsmtQual
## 4.972569e+12 1.859153e+12 1.330203e+12 1.323223e+12 1.319153e+12
##   GarageArea    GrLivArea    ExterQual    X2ndFlrSF    YearBuilt
## 1.174128e+12 7.001733e+11 6.236221e+11 4.074873e+11 3.970137e+11
## TotRmsAbvGrd  KitchenQual  GarageYrBlt   HouseStyle    X1stFlrSF
## 3.675016e+11 3.570280e+11 2.850651e+11 2.492876e+11 1.833666e+11
##      LotArea  Exterior2nd   GarageType     FullBath BedroomAbvGr
## 1.376232e+11 1.043187e+11 1.019718e+11 8.276084e+10 7.746191e+10
##     MSZoning  OpenPorchSF   Condition1  LotFrontage
## 7.084858e+10 5.258785e+10 4.207028e+10 2.918363e+10
```

```r
set.seed(2)
sp <- sample(1:nrow(M_train), 700)
ttrain <- M_train[sp,]
ttest <- M_train[-sp,]
y_test <- ttest$SalePrice
ttest <- M_train[-sp,-75]
tree_hp2 <- tree(SalePrice~.-SalePrice,ttrain)
tree_pred=predict(tree_hp2,ttest)
tree_pred
```

```
##        2        5        6        7       12       20       21       22
## 126755.5 331038.7 135207.0 242718.4 331038.7 135207.0 331038.7 182149.4
##       23       24       28       30       34       37       38       39
## 331391.7 101136.9 331391.7 135207.0 144292.7 135207.0 135207.0 135207.0
##       41       53       59       61       63       75       78       80
## 135207.0 101136.9 331038.7 135207.0 242718.4 144292.7 135207.0 101136.9
##       83       86       93       94      104      106      113      116
## 242718.4 331038.7 135207.0 144292.7 182149.4 242718.4 256471.3 182936.0
##      119      131      135      136      140      143      152      159
## 256471.3 256471.3 144292.7 182149.4 182936.0 135207.0 331391.7 182149.4
##      160      162      168      169      172      173      175      180
## 256471.3 331038.7 331038.7 182149.4 144292.7 182149.4 182936.0 101136.9
##      185      189      212      217      221      223      234      239
## 135207.0 135207.0 135207.0 182149.4 234390.7 182936.0 135207.0 242718.4
##      253      254      256      258      259      264      265      267
## 182936.0 135207.0 256471.3 234390.7 182149.4 101136.9 101136.9 182936.0
##      268      269      278      280      282      293      295      298
## 144292.7 101136.9 135207.0 256471.3 135207.0 144292.7 144292.7 182149.4
##      303      305      307      315      324      325      326      328
## 234390.7 256471.3 256471.3 182149.4 101136.9 256471.3 101136.9 135207.0
##      338      339      340      346      349      353      359      366
## 234390.7 182149.4 135207.0 182936.0 182149.4 135207.0 126755.5 101136.9
##      369      373      375      379      382      388      389      403
## 135207.0 135207.0 182149.4 331391.7 182149.4 135207.0 234390.7 135207.0
##      410      411      412      418      419      420      421      424
## 242718.4 135207.0 135207.0 182936.0 135207.0 135207.0 182149.4 331038.7
##      425      428      431      438      440      441      445      447
## 135207.0 135207.0 101136.9 135207.0 135207.0 543206.0 182149.4 144292.7
##      449      450      451      456      467      470      473      476
## 101136.9 101136.9 101136.9 182149.4 182149.4 182936.0 135207.0 135207.0
##      478      479      480      483      486      490      494      498
```

```
## 284859.6 331391.7 101136.9 182149.4 135207.0 101136.9 135207.0 182149.4
##      499      500      501      502      503      504      507      509
## 135207.0 135207.0 101136.9 182149.4 135207.0 234390.7 242718.4 182149.4
##      514      515      518      526      527      535      541      545
## 135207.0 135207.0 256471.3 182149.4 135207.0 242718.4 331391.7 182149.4
##      552      555      557      558      559      562      564      573
## 101136.9 256471.3 135207.0 101136.9 182149.4 135207.0 144292.7 182149.4
##      576      578      588      590      591      596      598      601
## 135207.0 135207.0 135207.0 135207.0 182149.4 331391.7 182149.4 242718.4
##      602      606      608      610      619      620      626      628
## 101136.9 256471.3 144292.7 135207.0 331391.7 331038.7 135207.0 144292.7
##      633      634      640      641      643      648      649      652
## 182149.4 135207.0 242718.4 242718.4 475077.3 135207.0 144292.7 144292.7
##      654      655      657      659      668      672      678      679
## 144292.7 543206.0 135207.0 144292.7 182936.0 135207.0 101136.9 331391.7
##      689      690      692      695      696      697      698      700
## 242718.4 135207.0 543206.0 182936.0 135207.0 135207.0 135207.0 182149.4
##      705      708      709      716      718      719      720      728
## 234390.7 242718.4 182149.4 135207.0 135207.0 256471.3 135207.0 182149.4
##      732      733      734      736      738      740      744      754
## 182149.4 256471.3 135207.0 182149.4 242718.4 182149.4 144292.7 331038.7
##      760      761      762      764      767      772      773      778
## 331038.7 135207.0 135207.0 331038.7 182149.4 135207.0 135207.0 135207.0
##      787      788      794      796      798      801      803      809
## 144292.7 256471.3 242718.4 182936.0 135207.0 182936.0 182149.4 135207.0
##      811      815      819      820      821      822      824      825
## 135207.0 135207.0 126755.5 182149.4 182149.4 101136.9 144292.7 242718.4
##      826      828      831      832      847      849      850      855
## 331391.7 234390.7 135207.0 182149.4 182149.4 182936.0 182936.0 144292.7
##      864      865      868      877      884      885      889      890
## 135207.0 182149.4 135207.0 135207.0 144292.7 135207.0 284859.6 144292.7
##      902      904      915      919      921      931      933      936
## 135207.0 234390.7 135207.0 256471.3 182936.0 242718.4 331391.7 135207.0
##      939      953      956      957      964      966      970      979
## 182149.4 135207.0 182936.0 126755.5 331391.7 182936.0 135207.0 135207.0
##      990      993      994     1000     1005     1006     1009     1014
## 182149.4 144292.7 182936.0 182149.4 182149.4 135207.0 234390.7 101136.9
##     1016     1017     1020     1021     1022     1024     1027     1028
## 242718.4 234390.7 182149.4 135207.0 182149.4 182149.4 135207.0 242718.4
##     1035     1040     1053     1056     1062     1064     1067     1068
## 135207.0 101136.9 144292.7 182936.0 101136.9 101136.9 182936.0 144292.7
##     1069     1077     1083     1084     1086     1088     1092     1093
## 144292.7 144292.7 234390.7 135207.0 135207.0 242718.4 182149.4 144292.7
##     1095     1106     1107     1110     1120     1126     1128     1129
## 135207.0 331038.7 182149.4 331391.7 135207.0 135207.0 234390.7 182149.4
##     1135     1145     1151     1153     1159     1169     1171     1176
## 182936.0 135207.0 135207.0 182936.0 242718.4 144292.7 135207.0 543206.0
##     1177     1179     1182     1185     1186     1187     1192     1193
## 135207.0 135207.0 242718.4 182936.0 135207.0 144292.7 242718.4 144292.7
##     1195     1203     1204     1205     1208     1223     1232     1236
## 135207.0 135207.0 234390.7 135207.0 182936.0 144292.7 135207.0 182936.0
##     1246     1250     1255     1264     1268     1282     1285     1289
## 182936.0 135207.0 182149.4 182936.0 331391.7 182149.4 144292.7 242718.4
##     1292     1293     1298     1300     1306     1308     1312     1314
```

```
## 101136.9 144292.7 135207.0 135207.0 242718.4 135207.0 182149.4 331038.7
##     1315     1316     1318     1329     1330     1331     1336     1339
## 135207.0 144292.7 182149.4 144292.7 182149.4 242718.4 135207.0 182149.4
##     1342     1346     1352     1356     1362     1364     1367     1370
## 135207.0 101136.9 144292.7 182149.4 234390.7 182936.0 182149.4 242718.4
##     1373     1375     1377     1378     1385     1386     1387     1388
## 256471.3 182149.4 101136.9 144292.7 135207.0 101136.9 256471.3 144292.7
##     1389     1393     1401     1404     1405     1406     1407     1411
## 331391.7 135207.0 135207.0 242718.4 101136.9 242718.4 135207.0 182149.4
##     1414     1419     1423     1428     1431     1434     1437     1438
## 331391.7 135207.0 135207.0 144292.7 182936.0 182936.0 135207.0 331391.7
##     1439     1440     1441     1449     1452     1453     1455     1458
## 101136.9 182149.4 182936.0 135207.0 242718.4 135207.0 182149.4 256471.3
##     1459     1460
## 135207.0 135207.0
```

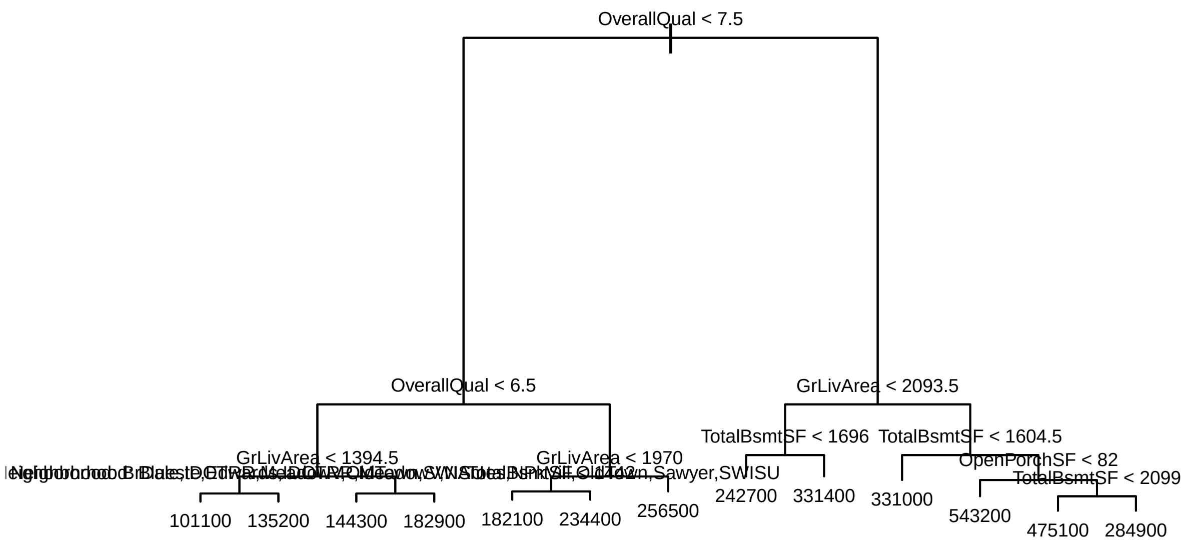
```r
set.seed(3)
cv_hp <- cv.tree(tree_hp2)
names(cv_hp)
```

```
## [1] "size"   "dev"    "k"      "method"
```

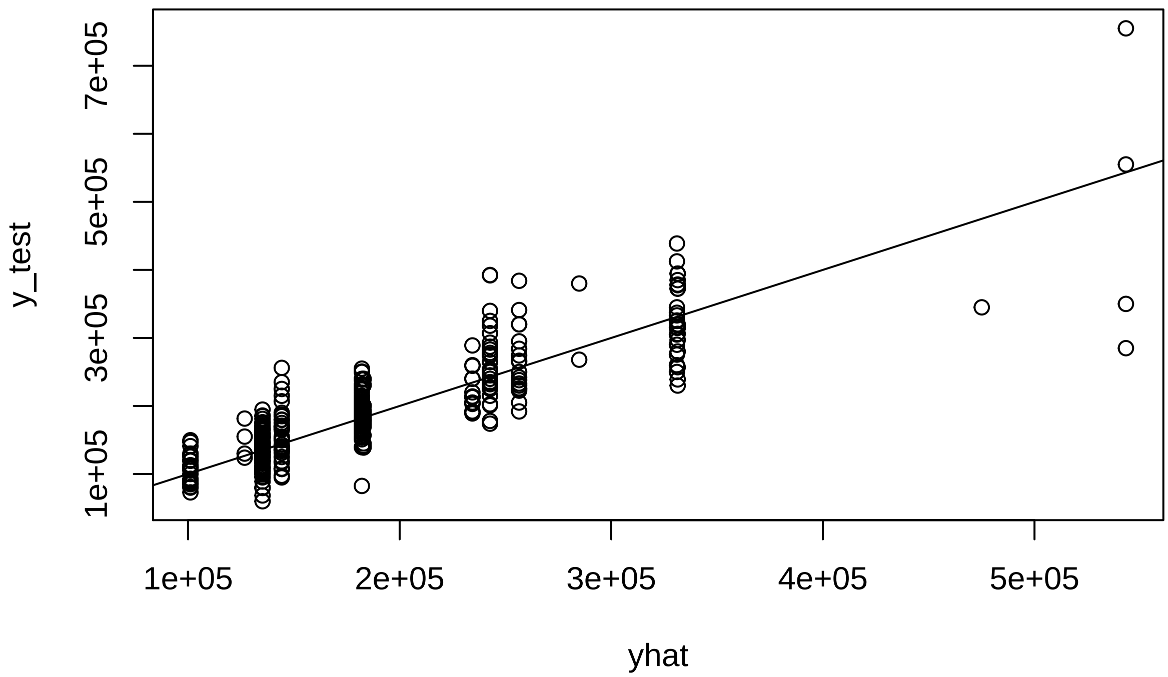
```r
cv_hp
```

```
## $size
##  [1] 13 12 11 10  9  8  7  6  5  4  3  2  1
##
## $dev
##  [1] 1.702244e+12 1.780216e+12 1.780216e+12 1.780216e+12 1.838748e+12
##  [6] 1.931682e+12 1.984517e+12 2.073495e+12 2.176634e+12 2.148031e+12
## [11] 2.174935e+12 2.615880e+12 5.110679e+12
##
## $k
##  [1]         -Inf 5.586099e+10 5.611700e+10 5.676630e+10 9.868033e+10
##  [6] 1.021228e+11 1.106876e+11 1.160007e+11 1.449851e+11 1.703217e+11
## [11] 3.460922e+11 4.918912e+11 2.500518e+12
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```r
best_level <- cv_hp$size[which.min(cv_hp$dev)]
prune_tree <- prune.tree(tree_hp2,best=best_level)
plot(prune_tree)
text(prune_tree,pretty=0, cex = 0.6)
```




```r
yhat <- predict(tree_hp2,newdata=ttest)
plot(yhat,y_test)
abline(0,1)
```




```r
mean((yhat-y_test)^2)
```

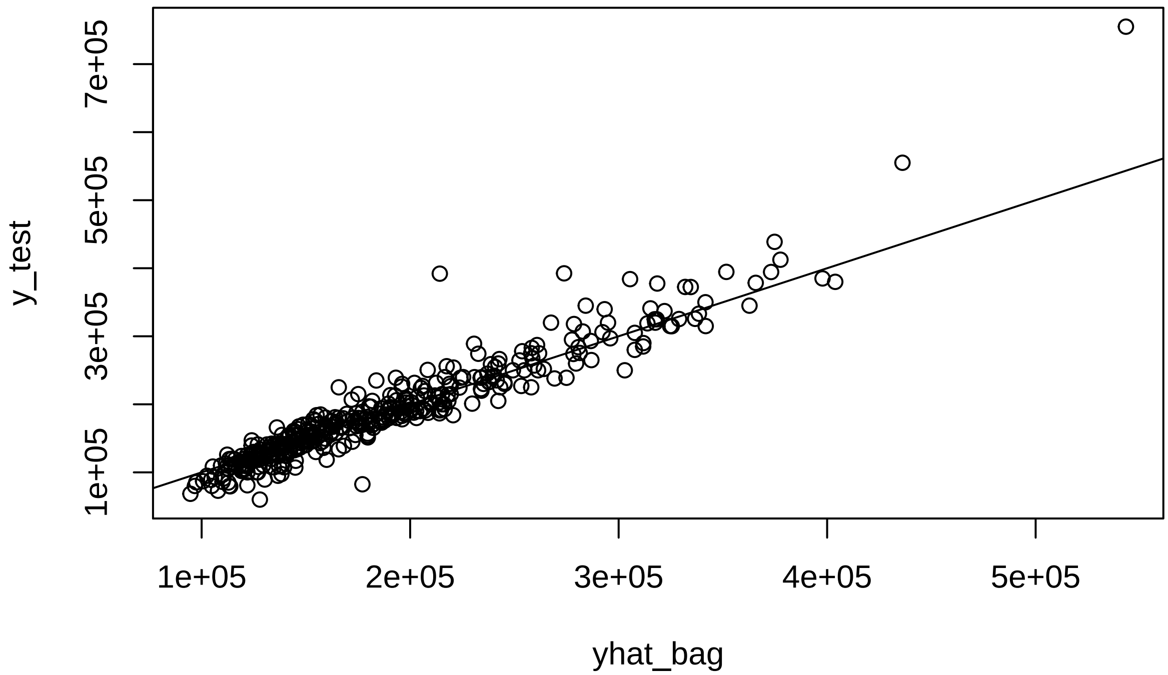
```
## [1] 1656281752
```

Bagging and Random forest

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.6.1
## 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
## The following object is masked from 'package:psych':
##
##     outlier
```

```r
set.seed(1)
bag_hp <- randomForest(SalePrice~.-SalePrice, ttrain,mtry=13,importance=TRUE)
yhat_bag <- predict(bag_hp,newdata=ttest)
plot(yhat_bag, y_test)
abline(0,1)
```




```r
mean((yhat_bag-y_test)^2)
```

```
## [1] 653229764
```

```r
#decrease the number of trees using ntree
bag_hp2 <- randomForest(SalePrice~.-SalePrice, ttrain,mtry=13,ntree=25)
yhat_bag <- predict(bag_hp,newdata=ttest)
mean((yhat_bag-y_test)^2)
```

```
## [1] 653229764
```

```r
#now estimate random forest
set.seed(1)
rf_hp <- randomForest(SalePrice~.-SalePrice, ttrain,mtry= 6,importance=TRUE)
yhat_rf <- predict(bag_hp,newdata=ttest)
mean((yhat_rf-y_test)^2)
```

```
## [1] 653229764
```

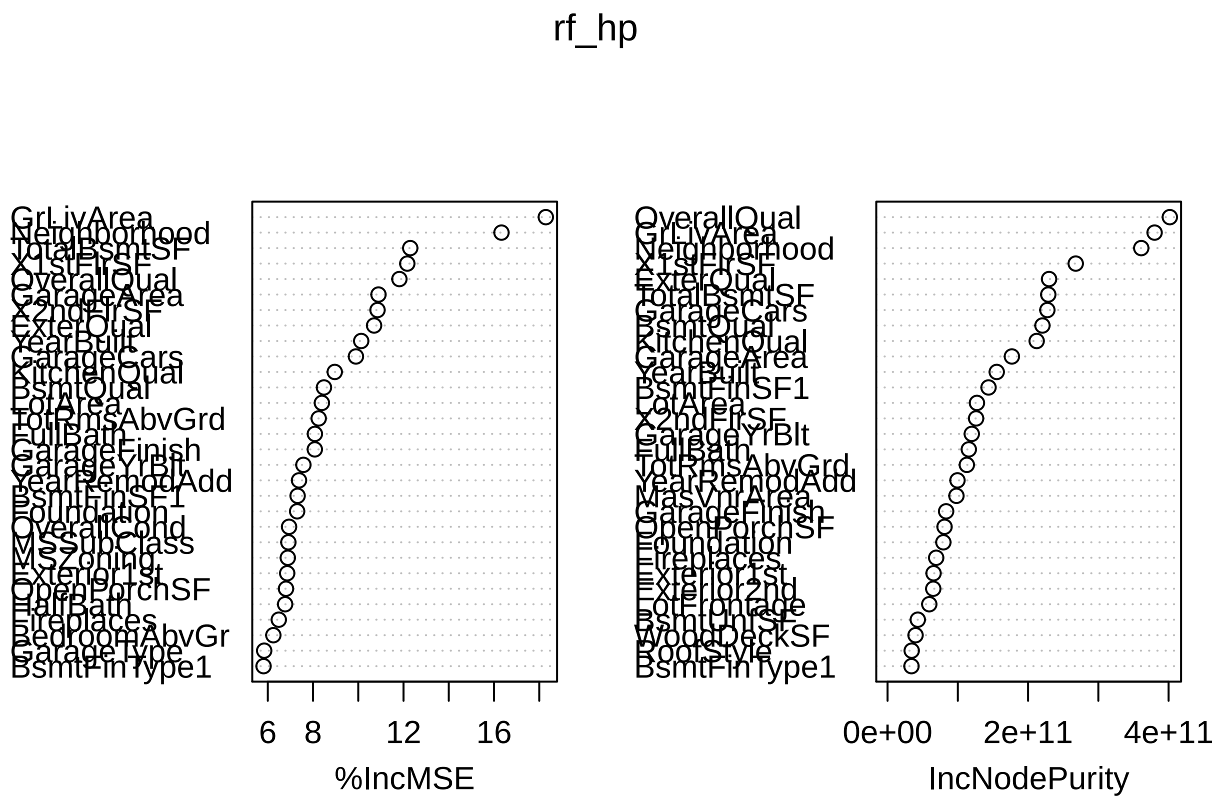
```r
importance(rf_hp)
```

```
##                    %IncMSE IncNodePurity
## MSSubClass      6.91033048   19537964196
```

```
## MSZoning       6.88495235   22645255470
## LotFrontage    4.63417776   59389755255
## LotArea        8.39084615  127226798943
## Street         0.00000000     193435300
## LotShape       1.71670248   21335555427
## LandContour    2.54825263   13419182321
## Utilities      0.00000000             0
## LotConfig      1.31014993   13078177115
## LandSlope      0.28943001    5823885891
## Neighborhood  16.33049235  361077721745
## Condition1     2.54137507   12407158115
## Condition2     1.42835809    2388473643
## BldgType       4.14057050   10722223074
## HouseStyle     5.20458180   21418098448
## OverallQual   11.81821245  401712805619
## OverallCond    6.94059879   15581859801
## YearBuilt     10.12989658  155366482564
## YearRemodAdd   7.38100825   99505065922
## RoofStyle      2.36420393   34368946517
## RoofMatl       1.55730254    6697163629
## Exterior1st    6.85878107   65479746696
## Exterior2nd    5.05004404   65030525000
## MasVnrType     2.98358124   21194017130
## MasVnrArea     3.23960877   98040513005
## ExterQual     10.69925703  229862405900
## ExterCond     -0.03589102    5557448565
## Foundation     7.29820491   79414005810
## BsmtQual       8.48273179  220415165362
## BsmtCond       1.34413719    3131642796
## BsmtExposure   1.68907522   30162731713
## BsmtFinType1   5.81312771   34020500946
## BsmtFinSF1     7.31878578  143723069589
## BsmtFinType2  -1.19107087    3699633574
## BsmtFinSF2     0.71938541    4671740881
## BsmtUnfSF      5.15439232   43045004332
## TotalBsmtSF   12.29733596  228723836037
## Heating        0.73336671    1876933394
## HeatingQC      3.93928769   23131759445
## CentralAir     4.72480590    5786034215
## Electrical     1.82922083    3109662289
## X1stFlrSF     12.16657608  267831543464
## X2ndFlrSF     10.85155502  126027739323
## LowQualFinSF   0.44182619    2930989610
## GrLivArea     18.28989226  380038575867
## BsmtFullBath   5.14587401   16491899656
## BsmtHalfBath   1.77670562     998457134
## FullBath       8.08244618  115628077910
## HalfBath       6.76679620   21223854956
## BedroomAbvGr   6.24524043   32927461315
## KitchenAbvGr   2.68093429    2125221940
## KitchenQual    8.96025201  212231046920
## TotRmsAbvGrd   8.25078955  112829342720
## Functional     2.21734471    3801733850
## Fireplaces     6.48611834   69233461700
```

```
## GarageType      5.84148295   31947132885
## GarageYrBlt     7.57397664  119836082114
## GarageFinish    8.07826069   83358581904
## GarageCars      9.89567045  227436321902
## GarageArea     10.89333749  176939251093
## GarageQual      2.41681087    7040201576
## GarageCond      2.86533941    1645012036
## PavedDrive      2.32168206    2695858693
## WoodDeckSF      4.11861128   39851519088
## OpenPorchSF     6.80527151   81080187714
## EnclosedPorch   1.59785617    5763036161
## X3SsnPorch      1.99730824     811393739
## ScreenPorch     2.20953764   10551363025
## PoolArea       -3.33702184   29706514707
## MiscVal        -0.89984157     361720967
## MoSold          1.70611952   23109938227
## YrSold          0.85497330   13016628529
## SaleType        3.35684080   31134701313
## SaleCondition   1.60438968   26838952236
```

```r
varImpPlot(rf_hp)
```



Boosting

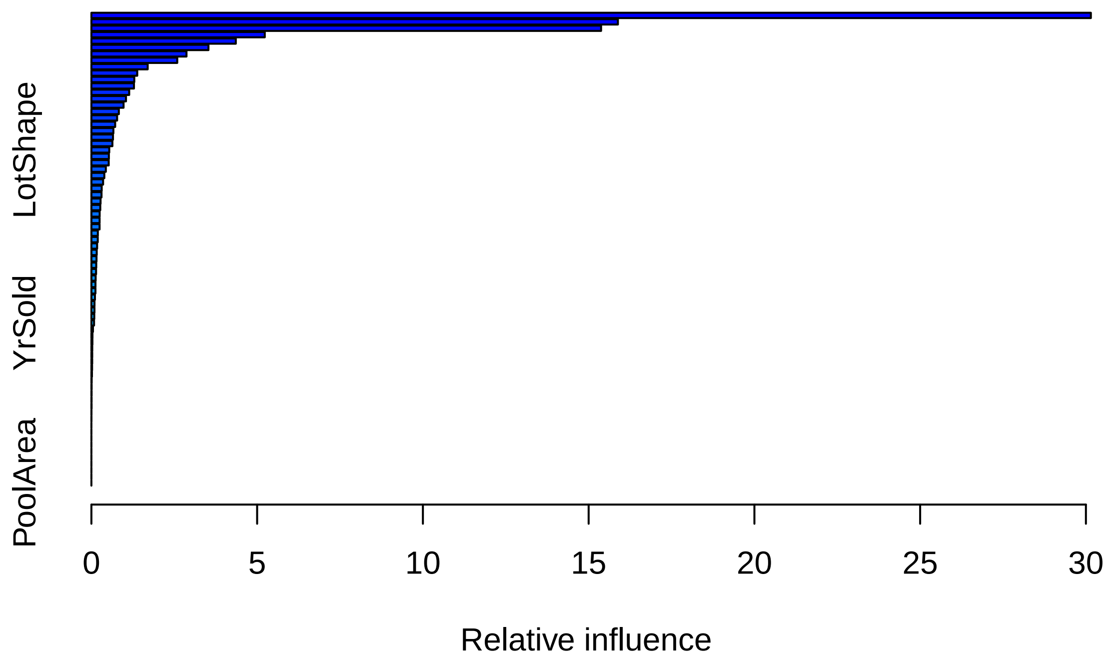
```r
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 3.6.2
## Loaded gbm 2.1.5
```

```r
set.seed(1)
boost_hp <- gbm(SalePrice~.- c(SalePrice), ttrain,distribution="gaussian",
n.trees=5000,interaction.depth=4)
```

```
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution =
## distribution, : variable 8: Utilities has no variation.
```

```r
summary(boost_hp)
```




```
##                         var      rel.inf
## OverallQual     OverallQual 3.014976e+01
## GrLivArea         GrLivArea 1.588282e+01
## Neighborhood   Neighborhood 1.537520e+01
## TotalBsmtSF     TotalBsmtSF 5.230696e+00
## X1stFlrSF         X1stFlrSF 4.356184e+00
## BsmtFinSF1       BsmtFinSF1 3.531566e+00
## GarageCars       GarageCars 2.868541e+00
## LotArea             LotArea 2.589759e+00
## GarageArea       GarageArea 1.698470e+00
## TotRmsAbvGrd   TotRmsAbvGrd 1.382904e+00
## Exterior2nd     Exterior2nd 1.293483e+00
## YearBuilt         YearBuilt 1.284456e+00
## MasVnrArea       MasVnrArea 1.140458e+00
## X2ndFlrSF         X2ndFlrSF 1.043898e+00
```

```
## LotFrontage     LotFrontage 9.708286e-01
## OpenPorchSF     OpenPorchSF 8.278779e-01
## BsmtUnfSF         BsmtUnfSF 7.758584e-01
## BsmtQual           BsmtQual 7.176064e-01
## Exterior1st     Exterior1st 6.631842e-01
## YearRemodAdd   YearRemodAdd 6.511265e-01
## KitchenQual     KitchenQual 6.334542e-01
## LotShape           LotShape 5.388925e-01
## GarageFinish   GarageFinish 5.278195e-01
## OverallCond     OverallCond 5.222401e-01
## LandContour     LandContour 4.358984e-01
## WoodDeckSF       WoodDeckSF 3.899062e-01
## BsmtExposure   BsmtExposure 3.532313e-01
## BsmtFinType1   BsmtFinType1 3.114828e-01
## SaleCondition SaleCondition 3.055070e-01
## Fireplaces       Fireplaces 2.767828e-01
## SaleType           SaleType 2.675098e-01
## Condition1       Condition1 2.494329e-01
## MSZoning           MSZoning 2.480108e-01
## MoSold               MoSold 2.470524e-01
## FullBath           FullBath 1.944766e-01
## ScreenPorch     ScreenPorch 1.918805e-01
## GarageType       GarageType 1.730673e-01
## GarageYrBlt     GarageYrBlt 1.606074e-01
## CentralAir       CentralAir 1.546267e-01
## EnclosedPorch EnclosedPorch 1.463497e-01
## BedroomAbvGr   BedroomAbvGr 1.413233e-01
## BsmtFullBath   BsmtFullBath 1.290234e-01
## LotConfig         LotConfig 1.228148e-01
## LandSlope         LandSlope 1.222527e-01
## BldgType           BldgType 1.065568e-01
## ExterQual         ExterQual 9.336498e-02
## HalfBath           HalfBath 9.280395e-02
## MSSubClass       MSSubClass 8.686156e-02
## YrSold               YrSold 8.054725e-02
## MasVnrType       MasVnrType 4.758908e-02
## HouseStyle       HouseStyle 3.182902e-02
## Functional       Functional 3.044998e-02
## BsmtFinSF2       BsmtFinSF2 2.572815e-02
## GarageQual       GarageQual 2.507401e-02
## RoofStyle         RoofStyle 2.429258e-02
## Foundation       Foundation 2.123782e-02
## GarageCond       GarageCond 1.658660e-02
## KitchenAbvGr   KitchenAbvGr 8.804354e-03
## BsmtCond           BsmtCond 7.025152e-03
## BsmtFinType2   BsmtFinType2 6.766673e-03
## HeatingQC         HeatingQC 5.951091e-03
## Electrical       Electrical 5.390071e-03
## ExterCond         ExterCond 2.768227e-03
## PavedDrive       PavedDrive 1.913491e-03
## MiscVal             MiscVal 7.871904e-05
## BsmtHalfBath   BsmtHalfBath 6.099420e-05
## Street               Street 0.000000e+00
## Utilities         Utilities 0.000000e+00
```

```
## Condition2       Condition2 0.000000e+00
## RoofMatl           RoofMatl 0.000000e+00
## Heating             Heating 0.000000e+00
## LowQualFinSF   LowQualFinSF 0.000000e+00
## X3SsnPorch       X3SsnPorch 0.000000e+00
## PoolArea           PoolArea 0.000000e+00
```

```r
par(mfrow=c(1,2))
yhat_boost=predict(boost_hp,newdata=ttest,n.trees=5000)
mean((yhat_boost-y_test)^2)
```

```
## [1] 666643285
```

After our Exploration, of all the models the random forest model has the least MSE(653229764), so we shall use it to predict the house prices for our Major test dataset.

Predicting Our Test data

```r
rf_hp <- randomForest(SalePrice~.-SalePrice,
            M_train,mtry= 6,importance=TRUE)

M_test$Price <- predict(rf_hp,newdata=M_test)
head(M_test$Price)
```

```
## [1] 128379.1 153856.6 184967.7 191093.2 197638.6 184247.2
```

A variable having no variation means it does not add any value as a predictor thus it does not affect our prediction.

8. Classification

Titanic Survival Prediction

```r
Titanic <- read.csv('Titrain.csv')

str(Titanic)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```r
colSums(is.na(Titanic))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
```

```
##           0           0           0           0           0           0
```

```
# The age column has missing values we sahre replace them by mean imputaion method
library(tidyverse)
Avg_sex_class <- group_by (Titanic,Sex, Pclass) %>%
                  summarise(Avg = mean(Age, na.rm = TRUE))

Avg_sex_class
```

```
## # A tibble: 6 x 3
## # Groups:   Sex [2]
##   Sex    Pclass   Avg
##   <fct>   <int> <dbl>
## 1 female      1  34.6
## 2 female      2  28.7
## 3 female      3  21.8
## 4 male        1  41.3
## 5 male        2  30.7
## 6 male        3  26.5
```

```
train <- Titanic

# Using the average age by gender 6 Pclass to impute the missing age values
train[which(train$Sex =='female' & train$Pclass == 1 & is.na(train$Age)),'Age'] = 34.61

train[which(train$Sex =='female' & train$Pclass == 2 & is.na(train$Age)),'Age'] = 28.72

train[which(train$Sex =='female' & train$Pclass == 3 & is.na(train$Age)),'Age'] = 21.75

train[which(train$Sex =='male' & train$Pclass == 1 & is.na(train$Age)),'Age'] = 41.28

train[which(train$Sex =='male' & train$Pclass == 2 & is.na(train$Age)),'Age'] = 30.74

train[which(train$Sex =='male' & train$Pclass == 3 & is.na(train$Age)),'Age'] = 26.51

colSums(is.na(train))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

```
train1 <- train
```

Classifcation Tree

the variables: Pclass,Sex,Age , SibSp ,Parch,Fare,Embarked are variables that may likely affect the chances of survival, this is from my peronal know-how based on some economic and psychological intuition.

```
train1$Survived <- as.factor(train1$Survived)
tree_tit <- rpart(Survived~ Pclass+Sex+Age + SibSp +Parch+Fare+Embarked, train1)
summary(tree_tit)
```

```
## Call:
## rpart(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare + Embarked, data = train1)
##   n= 891
```

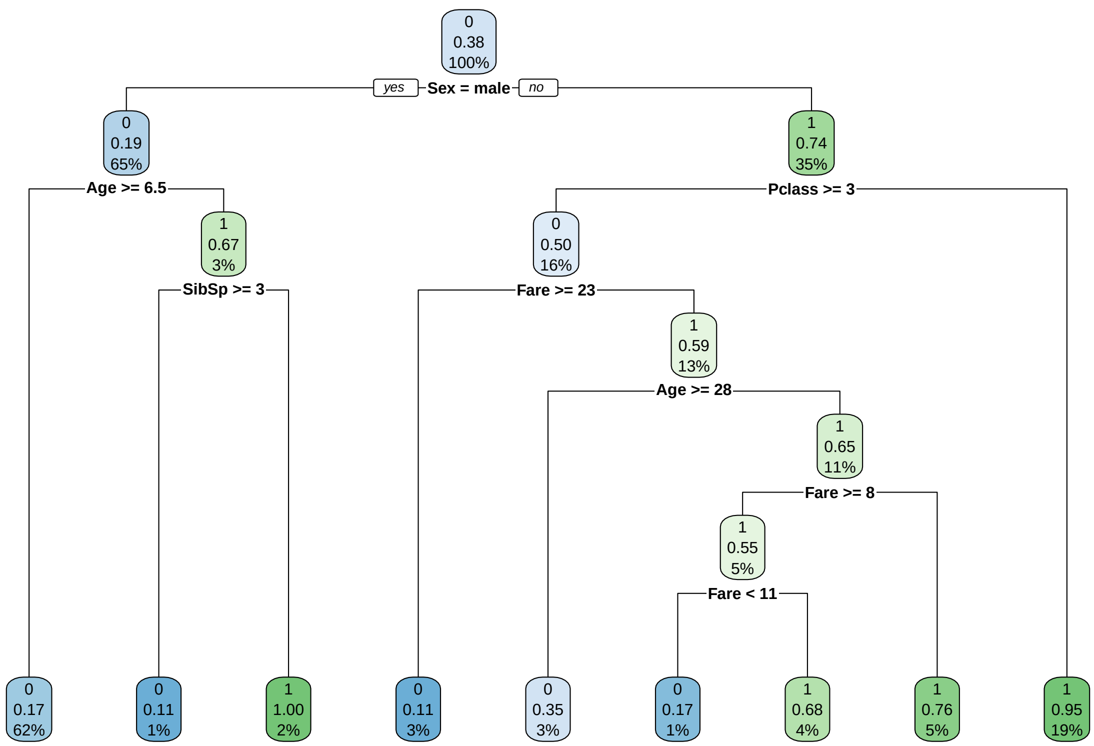
```
##
##           CP nsplit rel error    xerror       xstd
## 1 0.44444444      0 1.0000000 1.0000000 0.04244576
## 2 0.03070175      1 0.5555556 0.5555556 0.03574957
## 3 0.02339181      3 0.4941520 0.5263158 0.03504339
## 4 0.02046784      4 0.4707602 0.5029240 0.03444798
## 5 0.01169591      6 0.4298246 0.4853801 0.03398272
## 6 0.01000000      8 0.4064327 0.4853801 0.03398272
##
## Variable importance
##      Sex     Fare   Pclass      Age    SibSp    Parch Embarked
##       44       17       12       11        6        6        4
##
## Node number 1: 891 observations,    complexity param=0.4444444
##   predicted class=0  expected loss=0.3838384  P(node) =1
##     class counts:   549   342
##    probabilities: 0.616 0.384
##   left son=2 (577 obs) right son=3 (314 obs)
##   Primary splits:
##       Sex      splits as  RL,           improve=124.42630, (0 missing)
##       Pclass   < 2.5      to the right, improve= 43.78183, (0 missing)
##       Fare     < 10.48125 to the left,  improve= 37.94194, (0 missing)
##       Embarked splits as  RRLL,         improve= 12.86541, (0 missing)
##       Age      < 6.5      to the right, improve= 10.05326, (0 missing)
##   Surrogate splits:
##       Fare     < 77.6229  to the left,  agree=0.679, adj=0.089, (0 split)
##       Parch    < 0.5      to the left,  agree=0.678, adj=0.086, (0 split)
##       Age      < 21.875   to the right, agree=0.654, adj=0.019, (0 split)
##       Embarked splits as  RLLL,         agree=0.650, adj=0.006, (0 split)
##
## Node number 2: 577 observations,    complexity param=0.02339181
##   predicted class=0  expected loss=0.1889081  P(node) =0.647587
##     class counts:   468   109
##    probabilities: 0.811 0.189
##   left son=4 (553 obs) right son=5 (24 obs)
##   Primary splits:
##       Age      < 6.5      to the right, improve=11.431650, (0 missing)
##       Fare     < 26.26875 to the left,  improve=10.216720, (0 missing)
##       Pclass   < 1.5      to the right, improve=10.019140, (0 missing)
##       Parch    < 0.5      to the left,  improve= 3.350327, (0 missing)
##       Embarked splits as  -RLL,         improve= 3.079304, (0 missing)
##
## Node number 3: 314 observations,    complexity param=0.03070175
##   predicted class=1  expected loss=0.2579618  P(node) =0.352413
##     class counts:    81   233
##    probabilities: 0.258 0.742
##   left son=6 (144 obs) right son=7 (170 obs)
##   Primary splits:
##       Pclass   < 2.5      to the right, improve=31.163130, (0 missing)
##       Fare     < 48.2     to the left,  improve=10.114210, (0 missing)
##       SibSp    < 2.5      to the right, improve= 9.372551, (0 missing)
##       Parch    < 3.5      to the right, improve= 5.140857, (0 missing)
##       Embarked splits as  RRLL,         improve= 3.750944, (0 missing)
##   Surrogate splits:
```

```
##       Fare     < 25.69795 to the left,  agree=0.799, adj=0.563, (0 split)
##       Age      < 21.875   to the left,  agree=0.732, adj=0.417, (0 split)
##       Embarked splits as  RRLR,         agree=0.637, adj=0.208, (0 split)
##       SibSp    < 1.5      to the right, agree=0.592, adj=0.111, (0 split)
##       Parch    < 1.5      to the right, agree=0.567, adj=0.056, (0 split)
##
## Node number 4: 553 observations
##   predicted class=0  expected loss=0.1681736  P(node) =0.620651
##     class counts:   460    93
##    probabilities: 0.832 0.168
##
## Node number 5: 24 observations,    complexity param=0.02046784
##   predicted class=1  expected loss=0.3333333  P(node) =0.02693603
##     class counts:     8    16
##    probabilities: 0.333 0.667
##   left son=10 (9 obs) right son=11 (15 obs)
##   Primary splits:
##       SibSp  < 2.5      to the right, improve=8.8888890, (0 missing)
##       Pclass < 2.5      to the right, improve=3.8095240, (0 missing)
##       Fare   < 20.825   to the right, improve=2.6666670, (0 missing)
##       Age    < 1.5      to the right, improve=0.6095238, (0 missing)
##   Surrogate splits:
##       Pclass   < 2.5      to the right, agree=0.792, adj=0.444, (0 split)
##       Fare     < 26.95    to the right, agree=0.750, adj=0.333, (0 split)
##       Embarked splits as  -RLR,         agree=0.708, adj=0.222, (0 split)
##
## Node number 6: 144 observations,    complexity param=0.03070175
##   predicted class=0  expected loss=0.5  P(node) =0.1616162
##     class counts:    72    72
##    probabilities: 0.500 0.500
##   left son=12 (27 obs) right son=13 (117 obs)
##   Primary splits:
##       Fare     < 23.35    to the right, improve=10.051280, (0 missing)
##       Embarked splits as  -RRL,         improve= 7.071429, (0 missing)
##       SibSp    < 2.5      to the right, improve= 4.571429, (0 missing)
##       Age      < 38.5     to the right, improve= 4.545455, (0 missing)
##       Parch    < 1.5      to the right, improve= 3.773262, (0 missing)
##   Surrogate splits:
##       SibSp < 2.5      to the right, agree=0.882, adj=0.370, (0 split)
##       Parch < 1.5      to the right, agree=0.882, adj=0.370, (0 split)
##       Age   < 37.5     to the right, agree=0.819, adj=0.037, (0 split)
##
## Node number 7: 170 observations
##   predicted class=1  expected loss=0.05294118  P(node) =0.1907969
##     class counts:     9   161
##    probabilities: 0.053 0.947
##
## Node number 10: 9 observations
##   predicted class=0  expected loss=0.1111111  P(node) =0.01010101
##     class counts:     8     1
##    probabilities: 0.889 0.111
##
## Node number 11: 15 observations
##   predicted class=1  expected loss=0  P(node) =0.01683502
```

```
##     class counts:     0    15
##    probabilities: 0.000 1.000
##
## Node number 12: 27 observations
##   predicted class=0  expected loss=0.1111111  P(node) =0.03030303
##     class counts:    24     3
##    probabilities: 0.889 0.111
##
## Node number 13: 117 observations,    complexity param=0.02046784
##   predicted class=1  expected loss=0.4102564  P(node) =0.1313131
##     class counts:    48    69
##    probabilities: 0.410 0.590
##   left son=26 (23 obs) right son=27 (94 obs)
##   Primary splits:
##       Age      < 27.5     to the right, improve=3.3508150, (0 missing)
##       Embarked splits as  -RRL,         improve=2.6048030, (0 missing)
##       Fare     < 7.8875   to the right, improve=2.0325270, (0 missing)
##       SibSp    < 0.5      to the right, improve=0.3076923, (0 missing)
##       Parch    < 1.5      to the left,  improve=0.1582418, (0 missing)
##
## Node number 26: 23 observations
##   predicted class=0  expected loss=0.3478261  P(node) =0.02581369
##     class counts:    15     8
##    probabilities: 0.652 0.348
##
## Node number 27: 94 observations,    complexity param=0.01169591
##   predicted class=1  expected loss=0.3510638  P(node) =0.1054994
##     class counts:    33    61
##    probabilities: 0.351 0.649
##   left son=54 (49 obs) right son=55 (45 obs)
##   Primary splits:
##       Fare     < 8.0396   to the right, improve=1.9626670, (0 missing)
##       Embarked splits as  -RRL,         improve=1.7716050, (0 missing)
##       Age      < 6.5      to the right, improve=0.9354783, (0 missing)
##       Parch    < 1.5      to the left,  improve=0.3304408, (0 missing)
##       SibSp    < 0.5      to the right, improve=0.3300525, (0 missing)
##   Surrogate splits:
##       SibSp    < 0.5      to the right, agree=0.723, adj=0.422, (0 split)
##       Embarked splits as  -LRL,         agree=0.723, adj=0.422, (0 split)
##       Parch    < 0.5      to the right, agree=0.691, adj=0.356, (0 split)
##       Age      < 11       to the left,  agree=0.628, adj=0.222, (0 split)
##
## Node number 54: 49 observations,    complexity param=0.01169591
##   predicted class=1  expected loss=0.4489796  P(node) =0.05499439
##     class counts:    22    27
##    probabilities: 0.449 0.551
##   left son=108 (12 obs) right son=109 (37 obs)
##   Primary splits:
##       Fare     < 10.825   to the left,  improve=4.6953480, (0 missing)
##       Age      < 6.5      to the right, improve=2.5331860, (0 missing)
##       Parch    < 0.5      to the left,  improve=2.4805880, (0 missing)
##       Embarked splits as  -RRL,         improve=0.5815983, (0 missing)
##       SibSp    < 0.5      to the left,  improve=0.0743294, (0 missing)
##
```

```
## Node number 55: 45 observations
##    predicted class=1  expected loss=0.2444444  P(node) =0.05050505
##      class counts:    11    34
##     probabilities: 0.244 0.756
##
## Node number 108: 12 observations
##    predicted class=0  expected loss=0.1666667  P(node) =0.01346801
##      class counts:    10     2
##     probabilities: 0.833 0.167
##
## Node number 109: 37 observations
##    predicted class=1  expected loss=0.3243243  P(node) =0.04152637
##      class counts:    12    25
##     probabilities: 0.324 0.676
```

```r
rpart.plot(tree_tit)
```




```r
set.seed(2)

sp <- sample(1:nrow(train1), 600)
ttrain <- train1[sp,]
ttest <- train1[-sp,]
y_test <- ttest$Survived

ttest <- subset( train1, select = c(Pclass,Sex,Age , SibSp ,Parch,Fare,Embarked))
ttest <- ttest[-sp,]
tree_tit1 <- tree(Survived~ Pclass+Sex+Age + SibSp +Parch+Fare+Embarked,ttrain)
```

```r
tree_pred=predict(tree_tit1,ttest,type="class")
table(tree_pred,y_test)
```

```
##          y_test
## tree_pred   0   1
##         0 161  27
##         1  22  81
```

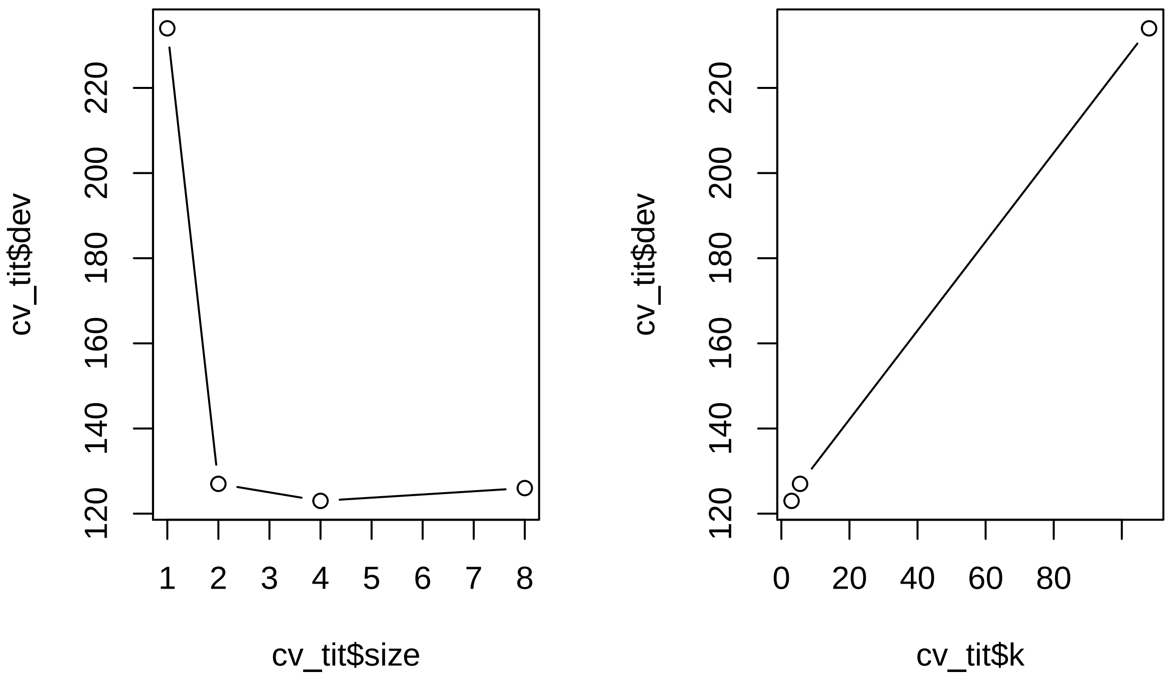
```r
set.seed(3)
cv_tit <- cv.tree(tree_tit1,FUN=prune.misclass)
names(cv_tit)
```

```
## [1] "size"   "dev"    "k"      "method"
```

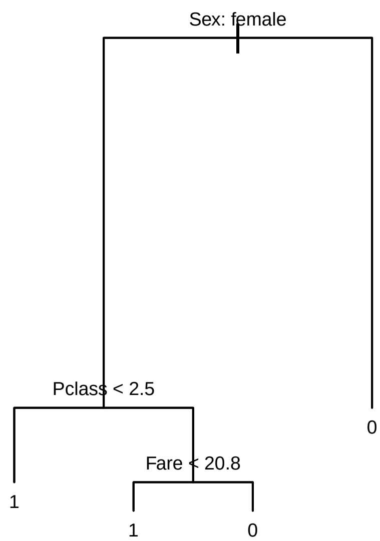
```r
cv_tit
```

```
## $size
## [1] 8 4 2 1
##
## $dev
## [1] 126 123 127 234
##
## $k
## [1]  -Inf   3.0   5.5 108.0
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```r
best_level <- which.min(cv_tit$dev)
par(mfrow=c(1,2))
plot(cv_tit$size,cv_tit$dev,type="b")
plot(cv_tit$k,cv_tit$dev,type="b")
```




```
prune_tit <- prune.misclass(tree_tit1,best=4)
plot(prune_tit)
text(prune_tit,pretty=0, cex = 0.6)
```




```
Titanic2 <- read.csv('Titest.csv')
Avg_sex_class <- group_by (Titanic2,Sex, Pclass) %>%
                  summarise(Avg = mean(Age, na.rm = TRUE))

(Avg_sex_class)
```

```
## # A tibble: 6 x 3
## # Groups:   Sex [2]
##   Sex    Pclass   Avg
##   <fct>   <int> <dbl>
## 1 female      1  41.3
## 2 female      2  24.4
## 3 female      3  23.1
## 4 male        1  40.5
## 5 male        2  30.9
## 6 male        3  24.5
```

```
train <- Titanic2

# Using the average age by gender 6 Pclass to impute the missing age values
train[which(train$Sex =='female' & train$Pclass == 1 & is.na(train$Age)),'Age'] = 41.33

train[which(train$Sex =='female' & train$Pclass == 2 & is.na(train$Age)),'Age'] = 24.38

train[which(train$Sex =='female' & train$Pclass == 3 & is.na(train$Age)),'Age'] = 23.07

train[which(train$Sex =='male' & train$Pclass == 1 & is.na(train$Age)),'Age'] = 4052
```

```r
train[which(train$Sex =='male' & train$Pclass == 2 & is.na(train$Age)),'Age'] = 30.94
train[which(train$Sex =='male' & train$Pclass == 3 & is.na(train$Age)),'Age'] = 24.52

colSums(is.na(train))
```

```
## PassengerId      Pclass        Name         Sex         Age       SibSp
##           0           0           0           0           0           0
##       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           1           0           0
```

```r
ttest <- subset( train, select = c(Pclass,Sex,Age , SibSp ,Parch,Fare,Embarked))
```

Predicting the survival class for our test dataset

```r
Titanic2$Survived=predict(tree_tit,ttest,type="class")

head(Titanic2)
```

```
##   PassengerId Pclass                                         Name    Sex
## 1         892      3                             Kelly, Mr. James   male
## 2         893      3             Wilkes, Mrs. James (Ellen Needs) female
## 3         894      2                    Myles, Mr. Thomas Francis   male
## 4         895      3                             Wirz, Mr. Albert   male
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female
## 6         897      3                   Svensson, Mr. Johan Cervin   male
##    Age SibSp Parch  Ticket    Fare Cabin Embarked Survived
## 1 34.5     0     0  330911  7.8292              Q        0
## 2 47.0     1     0  363272  7.0000              S        0
## 3 62.0     0     0  240276  9.6875              Q        0
## 4 27.0     0     0  315154  8.6625              S        0
## 5 22.0     1     1 3101298 12.2875              S        1
## 6 14.0     0     0    7538  9.2250              S        0
```