STAT797: House Price

Naby Diop

11/19/2019

Introduction

Many people believe that one of the best achievements in someone's life is to have their own house. Some people buy a house for living in, some others for business. However, house buyers have to go through a long and difficult process. The first and most important step of this process is to evaluate the finances. In fact, buying a house requires a lot of money. If one would like to buy a house, one should make sure that he/she have consistent income and a good amount of cash for a down payment. However, even after determining the financial ability and knowing where and which house to buy, it is always a challenge to decide the worth of a house. The dataset House Price collect the information of 1460 houses in Ames, Iowa. This dataset has 81 variables that describe all the specificities of a house, including its price. Thus, for this project we seek to predict the price of a house in Ames, Iowa given the 79 explanatory variables. The Id variable will be excluded because it does not affect the study. The main reasons for this study are to help the house sellers to fix a reasonable price; at same the time, to help the buyer check whether he/she is not overpaying. Therefore, the data will be fitted first with linear regression model followed by the random forests model.

Data Preparation

```
## 'data.frame':    1460 obs. of  80 variables:
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
```

```
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

The dataset HousePrice was collected by Dean De Cock, a professor of statistics and Director of assessment at Iowa State University. It has 1460 observations, 79 explanatory variables and one response variable (SalePrice).

We first examine the data by looking at its structure. The first thing that we notice is the data is a mix of numerical and categorical variables. However, when we look at the data description given by the collectors, we realize that fourteen variables (Alley, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinTye2, FirelaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PoolQc, Fence and MIscFeature) in the dataset have NA as output indicating the absence of a feature in a house. This doesn't mean the values are missing. Therefore, we must replace them with an actual value that can be interpreted differently by R. There are also six variables (MiscFeature, MssubClass, OverallQual, OverallCond, Utilities YrSold) that are interpreted numerical variables, but they are not. These variables need to be modified so they can be used correctly in the analysis. The code to do such transformations can be found on the appendix of this document.

As in any data analysis, the first step is to look at some graphical and numerical displays of the data. A good numerical overview is the summary of the data. We use maxsum=10 to print the ten most frequent levels within a variable, because many of them have their number of levels less than 10. For those with more than 10 levels, the number of observations of the remaining levels will be automatically combined and stored in a new level called Other in R.

```
##    MSSubClass     MSZoning     LotFrontage        LotArea
##  20     :536   C (all):  10   Min.   : 21.00   Min.   :  1300
##  60     :299   FV     :  65   1st Qu.: 59.00   1st Qu.:  7554
##  50     :144   RH     :  16   Median : 69.00   Median :  9478
##  120    : 87   RL     :1151   Mean   : 70.05   Mean   : 10517
##  30     : 69   RM     : 218   3rd Qu.: 80.00   3rd Qu.: 11602
##  160    : 63                  Max.   :313.00   Max.   :215245
##  (Other):262                  NA's   :259
##   Street     Alley      LotShape  LandContour  Utilities      LotConfig
##  Grvl:   6   1  :  50   IR1:484   Bnk:  63    AllPub:1459   Corner : 263
##  Pave:1454   2  :  41   IR2: 41   HLS:  50    NoSeWa:   1   CulDSac:  94
##              NOA:1369   IR3: 10   Low:  36                  FR2    :  47
##                         Reg:925   Lvl:1311                  FR3    :   4
##                                                             Inside :1052
##
##
##  LandSlope   Neighborhood   Condition1     Condition2     BldgType
##  Gtl:1382   NAmes  :225   Norm   :1260   Norm   :1445   1Fam  :1220
##  Mod:  65   CollgCr:150   Feedr  :  81   Feedr  :   6   2fmCon:  31
##  Sev:  13   OldTown:113   Artery :  48   Artery :   2   Duplex:  52
##             Edwards:100   RRAn   :  26   PosN   :   2   Twnhs :  43
##             Somerst: 86   PosN   :  19   RRNn   :   2   TwnhsE: 114
##             Gilbert: 79   RRAe   :  11   PosA   :   1
##             (Other):707   (Other):  15   (Other):   2
##    HouseStyle   OverallQual   OverallCond    YearBuilt     YearRemodAdd
##  1Story :726   5      :397   5      :821   Min.   :1872   Min.   :1950
##  2Story :445   6      :374   6      :252   1st Qu.:1954   1st Qu.:1967
##  1.5Fin :154   7      :319   7      :205   Median :1973   Median :1994
##  SLvl   : 65   8      :168   8      : 72   Mean   :1971   Mean   :1985
##  SFoyer : 37   4      :116   4      : 57   3rd Qu.:2000   3rd Qu.:2004
##  1.5Unf : 14   9      : 43   3      : 25   Max.   :2010   Max.   :2010
##  (Other): 19   (Other): 43   (Other): 28
##    RoofStyle       RoofMatl     Exterior1st   Exterior2nd    MasVnrType
##  Flat   :  13   CompShg:1434   VinylSd:515   VinylSd:504   BrkCmn : 15
##  Gable  :1141   Tar&Grv:  11   HdBoard:222   MetalSd:214   BrkFace:445
##  Gambrel:  11   WdShngl:   6   MetalSd:220   HdBoard:207   None   :864
##  Hip    : 286   WdShake:   5   Wd Sdng:206   Wd Sdng:197   Stone  :128
##  Mansard:   7   ClyTile:   1   Plywood:108   Plywood:142   NA's   :  8
```

```
##  Shed   :   2   Membran:   1   CemntBd: 61   CmentBd: 60
##                 (Other):   2   (Other):128   (Other):136
##    MasVnrArea     ExterQual ExterCond  Foundation  BsmtQual  BsmtCond
##  Min.   :   0.0   Ex: 52    Ex:   3   BrkTil:146   1  :121   1  :  45
##  1st Qu.:   0.0   Fa: 14    Fa:  28   CBlock:634   2  : 35   2  :  65
##  Median :   0.0   Gd:488    Gd: 146   PConc :647   3  :618   3  :   2
##  Mean   : 103.7   TA:906    Po:   1   Slab  : 24   4  :649   4  :1311
##  3rd Qu.: 166.0             TA:1282   Stone :  6   NOB: 37   NOB:  37
##  Max.   :1600.0                       Wood  :  3
##  NA's   :8
##  BsmtExposure BsmtFinType1   BsmtFinSF1     BsmtFinType2   BsmtFinSF2
##  1  :221      1  :220      Min.   :   0.0   1  :  19     Min.   :   0.00
##  2  :134      2  :148      1st Qu.:   0.0   2  :  33     1st Qu.:   0.00
##  3  :114      3  :418      Median : 383.5   3  :  14     Median :   0.00
##  4  :953      4  : 74      Mean   : 443.6   4  :  46     Mean   :  46.55
##  NOB: 38      5  :133      3rd Qu.: 712.2   5  :  54     3rd Qu.:   0.00
##               6  :430      Max.   :5644.0   6  :1256     Max.   :1474.00
##               NOB: 37                       NOB:  38
##    BsmtUnfSF       TotalBsmtSF      Heating     HeatingQC CentralAir
##  Min.   :   0.0   Min.   :   0.0   Floor:   1   Ex:741    N:  95
##  1st Qu.: 223.0   1st Qu.: 795.8   GasA :1428   Fa: 49    Y:1365
##  Median : 477.5   Median : 991.5   GasW :  18   Gd:241
##  Mean   : 567.2   Mean   :1057.4   Grav :   7   Po:  1
##  3rd Qu.: 808.0   3rd Qu.:1298.2   OthW :   2   TA:428
##  Max.   :2336.0   Max.   :6110.0   Wall :   4
##
##  Electrical     X1stFlrSF      X2ndFlrSF     LowQualFinSF
##  FuseA:  94   Min.   : 334   Min.   :   0   Min.   :  0.000
##  FuseF:  27   1st Qu.: 882   1st Qu.:   0   1st Qu.:  0.000
##  FuseP:   3   Median :1087   Median :   0   Median :  0.000
##  Mix  :   1   Mean   :1163   Mean   : 347   Mean   :  5.845
##  SBrkr:1334   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000
##  NA's :   1   Max.   :4692   Max.   :2065   Max.   :572.000
##
##    GrLivArea     BsmtFullBath     BsmtHalfBath        FullBath
##  Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:1130   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000
##  Median :1464   Median :0.0000   Median :0.00000   Median :2.000
##  Mean   :1515   Mean   :0.4253   Mean   :0.05753   Mean   :1.565
##  3rd Qu.:1777   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000
##  Max.   :5642   Max.   :3.0000   Max.   :2.00000   Max.   :3.000
##
##     HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39
##  Median :0.0000   Median :3.000   Median :1.000   Gd:586
##  Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :2.0000   Max.   :8.000   Max.   :3.000
##
##   TotRmsAbvGrd    Functional    Fireplaces    FireplaceQu GarageType
##  Min.   : 2.000   Maj1:  14   Min.   :0.000   1  : 24     1  :  6
##  1st Qu.: 5.000   Maj2:   5   1st Qu.:0.000   2  : 33     2  :870
##  Median : 6.000   Min1:  31   Median :1.000   3  :380     3  : 19
```

```
##  Mean   : 6.518   Min2:  34   Mean   :0.613   4  : 20     4  : 88
##  3rd Qu.: 7.000   Mod :  15   3rd Qu.:1.000   5  :313     5  :  9
##  Max.   :14.000   Sev :   1   Max.   :3.000   NOF:690     6  :387
##                   Typ :1360                               NOG: 81
##   GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual
##  Min.   :1900   1  :352      Min.   :0.000   Min.   :   0.0   1  :   3
##  1st Qu.:1961   2  :422      1st Qu.:1.000   1st Qu.: 334.5   2  :  48
##  Median :1980   3  :605      Median :2.000   Median : 480.0   3  :  14
##  Mean   :1979   NOG: 81      Mean   :1.767   Mean   : 473.0   4  :   3
##  3rd Qu.:2002                3rd Qu.:2.000   3rd Qu.: 576.0   5  :1311
##  Max.   :2010                Max.   :4.000   Max.   :1418.0   NOG:  81
##  NA's   :81
##  GarageCond PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch
##  1  :   2   N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  2  :  35   P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
##  3  :   9   Y:1340     Median :  0.00   Median : 25.00   Median :  0.00
##  4  :   7              Mean   : 94.24   Mean   : 46.66   Mean   : 21.95
##  5  :1326              3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00
##  NOG:  81              Max.   :857.00   Max.   :547.00   Max.   :552.00
##
##    X3SsnPorch      ScreenPorch        PoolArea       PoolQC     Fence
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   1  :   2   1  :  59
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   2  :   2   2  :  54
##  Median :  0.00   Median :  0.00   Median :  0.000   3  :   3   3  : 157
##  Mean   :  3.41   Mean   : 15.06   Mean   :  2.759   NOP:1453   4  :  11
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000              NOF:1179
##  Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##  MiscFeature    MiscVal             MoSold        YrSold       SaleType
##  1   :   2   Min.   :    0.00   Min.   : 1.000   2006:314   WD     :1267
##  2   :   2   1st Qu.:    0.00   1st Qu.: 5.000   2007:329   New    : 122
##  3   :  49   Median :    0.00   Median : 6.000   2008:304   COD    :  43
##  4   :   1   Mean   :   43.49   Mean   : 6.322   2009:338   ConLD  :   9
##  NONE:1406   3rd Qu.:    0.00   3rd Qu.: 8.000   2010:175   ConLI  :   5
##              Max.   :15500.00   Max.   :12.000              ConLw  :   5
##                                                             (Other):   9
##  SaleCondition    SalePrice
##  Abnorml: 101   Min.   : 34900
##  AdjLand:   4   1st Qu.:129975
##  Alloca :  12   Median :163000
##  Family :  20   Mean   :180921
##  Normal :1198   3rd Qu.:214000
##  Partial: 125   Max.   :755000
##
```

The frequency of some of the levels is so small that it makes it difficult to estimate their effect on the analysis. Thus, we will collapse many of them in the same level based on their similarity and frequency. The appendix of the document has the R code for such transformations. For example the variable Utilities have only 1 in NoSeWa, 1459 AllPub and 0 on any other else levels. This means Utilities is not important for the analyse, so we can drop it from the dataset. The variable LotShape has four levels. However, the levels IR1, IR2, IR3 are not too frequent and they all represent a type of irregularity. Therefore, we will collapse these three levels to one and call it IREG meaning irregular. The frequency of Bnk, HLS and Low for variable LanContour is little, plus they all represent a degree of flatness of the land. Thus, it makes sens to group them in only one category called NotFlat. We will use the same process as above to gather levels of categorical variable in the

data whenever it is possible.

After removing all the NA that were meant to identify the absence of an existing feature, we must delete all the other NA which represent missing values. The actual dimension of the data is 1074 overvations and 79 variables.

```
## [1] 1074   79
```

Now we have modified our dataset to a much more meaningful one, we can do some preliminary variable selection. It is important to select a subset of variables that best predict the response variable. For this purpose, we will be using the function Boruta from Boruta package in R. "Boruta is an all relevant feature selection wrapper algorithm, capable of working with any classification method that output variable importance measure (VIM)" (R description of Boruta). This package will help us to identify the variables that are best for predicting SalePrice.

```
## Boruta performed 199 iterations in 3.423605 mins.
##  45 attributes confirmed important: BedroomAbvGr, BldgType,
## BsmtExposure, BsmtFinSF1, BsmtFinType1 and 40 more;
##  26 attributes confirmed unimportant: Alley, BsmtCond, BsmtFinSF2,
## BsmtFinType2, BsmtHalfBath and 21 more;
##  7 tentative attributes left: Electrical, Fence, Functional,
## GarageCond, PavedDrive and 2 more;
```

The preliminary variable selection indicates 45 variables that are meaningful for this analysis, 26 variables can be excluded, and only seven variables were left undecided by boruta algorithem. However, the function TentativeRoughFix from the same package (boruta) allow a method that decides which variables among these seven we must keep.

```
## Boruta performed 199 iterations in 3.423605 mins.
## Tentatives roughfixed over the last 199 iterations.
##  46 attributes confirmed important: BedroomAbvGr, BldgType,
## BsmtExposure, BsmtFinSF1, BsmtFinType1 and 41 more;
##  32 attributes confirmed unimportant: Alley, BsmtCond, BsmtFinSF2,
## BsmtFinType2, BsmtHalfBath and 27 more;
```

After appying this function we have 46 variables confirmed important; the rest will not be used anymore. We can use the function getNonRejecdFormula (from boruta package) to have a look at the variables that will be used for fitting the models.

```
## SalePrice ~ MSSubClass + MSZoning + LotFrontage + LotArea + LotShape +
##     Neighborhood + BldgType + HouseStyle + OverallQual + OverallCond +
##     YearBuilt + YearRemodAdd + Exterior1st + Exterior2nd + MasVnrType +
##     MasVnrArea + ExterQual + Foundation + BsmtQual + BsmtExposure +
##     BsmtFinType1 + BsmtFinSF1 + BsmtUnfSF + TotalBsmtSF + HeatingQC +
##     CentralAir + X1stFlrSF + X2ndFlrSF + GrLivArea + BsmtFullBath +
##     FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual +
##     TotRmsAbvGrd + Fireplaces + FireplaceQu + GarageType + GarageYrBlt +
##     GarageFinish + GarageCars + GarageArea + PavedDrive + WoodDeckSF +
##     OpenPorchSF
## <environment: 0x7ff396b31658>
```

Data Analysis

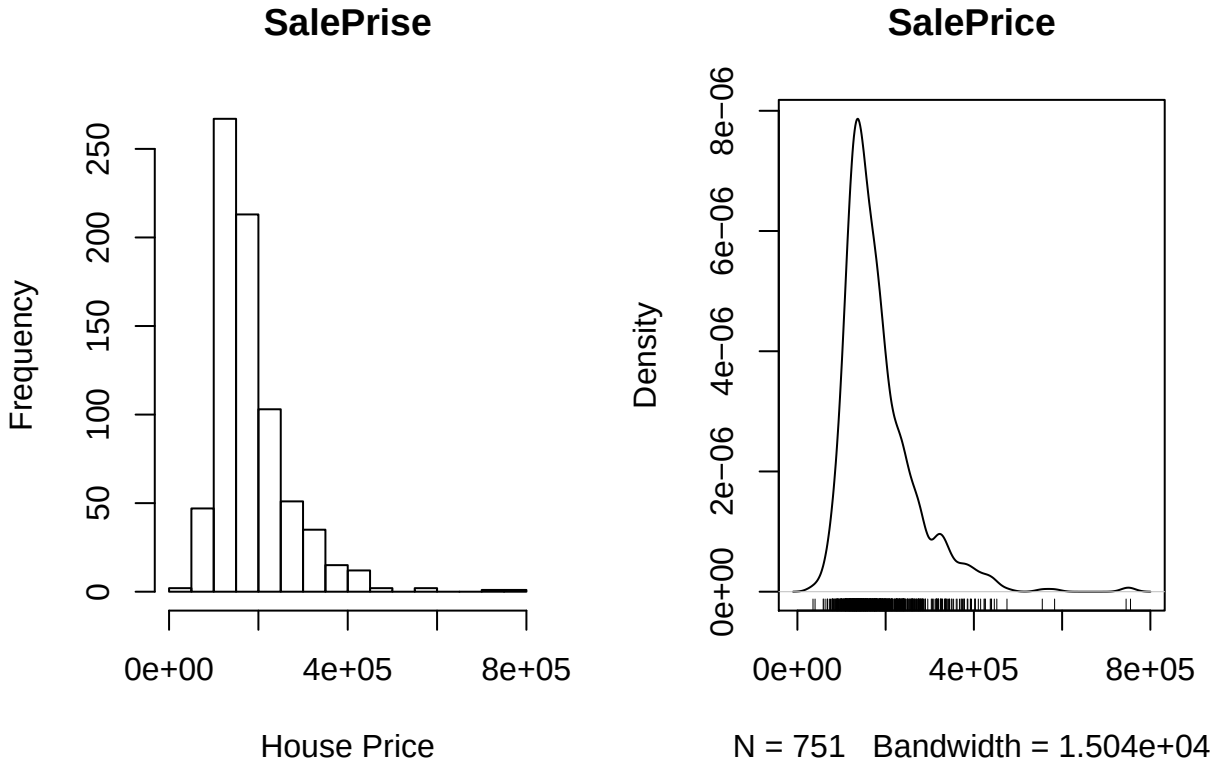
Preliminary analysis

The first step in any data analysis is to split the data into training and test set. Here, we will be using 70% of the data as train set and 30% as test set.

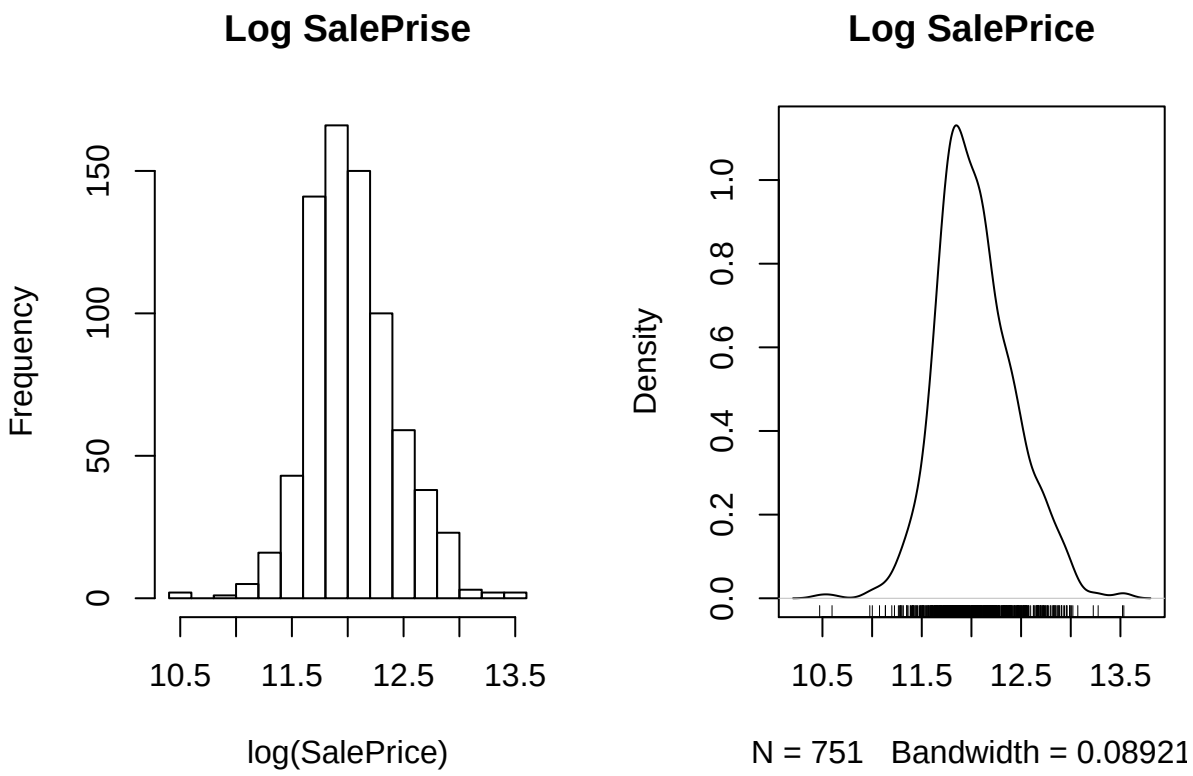
Considering first just the response variable, we can look at its distribution. Thus, we will look at the its histogram and its summary.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35311  130000  163000  183764  214200  755000
```

We see that the SalePrice of a house ranges from 35311 to 755000. The mean price of a house in Aimes Iowa is 183764.



From these two plots, the histogram and the kernel density of estimate which is a smooth version of the histogram, we see that SalePrice distribution is slightly right skewed. Such plot does not help much in detecting outliers in the data. However transforming the SalePrice with the log distribution makes it normally distributed. In another words the log of SalePrice is normal. This can be visualized by the bellow curve and histogram.



We will be using the log of SalePrice for fowarder analysis.

Linear Regresion

Linear regression is a useful tool for predicting a quantitative response. Thus, we can describe House Price data with a linear model which takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon, p = 1, 2, \cdots, 45$$

.

Where Y = $\log(SalePrice)$ with n=1460 $y = \left(y_1, \cdots, y_n\right)^T$, $\epsilon = \left(\epsilon_1, \cdots, \epsilon_n\right)^T$, $\beta = \left(\beta_0, \cdots, \beta_n\right)^T$and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,46} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,46} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,46} \end{pmatrix}$$

lm model fit

```
##
## Call:
## lm(formula = Train_set2$SalePrice ~ MSSubClass + MSZoning + LotFrontage +
##     LotArea + LotShape + Neighborhood + BldgType + HouseStyle +
##     OverallQual + OverallCond + YearBuilt + YearRemodAdd + Exterior1st +
##     Exterior2nd + MasVnrType + MasVnrArea + ExterQual + Foundation +
##     BsmtQual + BsmtExposure + BsmtFinType1 + BsmtFinSF1 + BsmtUnfSF +
##     TotalBsmtSF + HeatingQC + CentralAir + X1stFlrSF + X2ndFlrSF +
##     GrLivArea + BsmtFullBath + FullBath + HalfBath + BedroomAbvGr +
##     KitchenAbvGr + KitchenQual + TotRmsAbvGrd + Fireplaces +
```

```
##     FireplaceQu + GarageType + GarageYrBlt + GarageFinish + GarageCars +
##     GarageArea + PavedDrive + WoodDeckSF + OpenPorchSF, data = Train_set2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04055 -0.06118  0.00424  0.06302  0.45491
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.156e+01  1.682e-01  68.710  < 2e-16 ***
## MSSubClassOneHfStoty -2.163e-02  4.445e-02  -0.487 0.626641
## MSSubClassTwoStory    9.973e-03  4.815e-02   0.207 0.835989
## MSSubClassTwoHfStory  4.535e-02  7.066e-02   0.642 0.521228
## MSSubClassSplit      -5.399e-02  4.895e-02  -1.103 0.270441
## MSSubClassDuplex     -1.609e-01  1.460e-01  -1.102 0.270872
## MSSubClassPud        -2.127e-01  1.390e-01  -1.530 0.126479
## MSSubClassPudM       -3.159e-01  1.481e-01  -2.133 0.033284 *
## MSSubClassTwoFam     -1.193e-01  1.432e-01  -0.833 0.405236
## MSZoningRl            2.951e-03  2.318e-02   0.127 0.898758
## LotFrontage          -5.890e-04  2.726e-04  -2.160 0.031114 *
## LotArea              -1.739e-06  1.650e-06  -1.054 0.292272
## LotShapeREG          -1.884e-03  1.220e-02  -0.154 0.877278
## NeighborhoodBlueste  -8.367e-02  1.447e-01  -0.578 0.563433
## NeighborhoodBrDale    3.704e-02  7.913e-02   0.468 0.639868
## NeighborhoodBrkSide   1.417e-02  6.938e-02   0.204 0.838237
## NeighborhoodClearCr   1.441e-01  7.829e-02   1.841 0.066071 .
## NeighborhoodCollgCr   4.847e-02  5.417e-02   0.895 0.371224
## NeighborhoodCrawfor   1.864e-01  6.198e-02   3.007 0.002747 **
## NeighborhoodEdwards  -3.733e-02  5.870e-02  -0.636 0.525020
## NeighborhoodGilbert   2.755e-02  5.728e-02   0.481 0.630708
## NeighborhoodIDOTRR   -1.556e-01  7.692e-02  -2.023 0.043544 *
## NeighborhoodMeadowV  -1.011e-01  9.553e-02  -1.059 0.290225
## NeighborhoodMitchel   2.662e-03  6.028e-02   0.044 0.964787
## NeighborhoodNAmes    -1.593e-02  5.723e-02  -0.278 0.780903
## NeighborhoodNoRidge   2.040e-01  6.403e-02   3.187 0.001511 **
## NeighborhoodNPkVill  -1.787e-01  1.166e-01  -1.533 0.125798
## NeighborhoodNridgHt   1.380e-01  5.500e-02   2.509 0.012355 *
## NeighborhoodNWAmes   -1.518e-02  6.056e-02  -0.251 0.802216
## NeighborhoodOldTown  -1.182e-01  6.726e-02  -1.758 0.079315 .
## NeighborhoodSawyer    6.303e-03  6.097e-02   0.103 0.917693
## NeighborhoodSawyerW   2.265e-02  5.918e-02   0.383 0.702058
## NeighborhoodSomerst   1.335e-01  6.023e-02   2.216 0.027062 *
## NeighborhoodStoneBr   2.092e-01  5.869e-02   3.564 0.000393 ***
## NeighborhoodSWISU    -7.014e-02  7.278e-02  -0.964 0.335561
## NeighborhoodTimber    3.848e-02  6.250e-02   0.616 0.538371
## NeighborhoodVeenker   4.700e-02  9.166e-02   0.513 0.608247
## BldgTypeOTHERS        1.410e-01  1.364e-01   1.034 0.301684
## HouseStyleOneStory    1.556e-03  4.350e-02   0.036 0.971477
## HouseStyleTwoStory   -9.091e-02  4.361e-02  -2.085 0.037511 *
## OverallQualavg        9.398e-02  2.196e-02   4.280 2.17e-05 ***
## OverallQualGood       1.714e-01  2.713e-02   6.317 5.06e-10 ***
## OverallQualExc        2.580e-01  4.168e-02   6.190 1.09e-09 ***
## OverallCondavg        1.384e-01  2.742e-02   5.049 5.83e-07 ***
## OverallCondGood       1.967e-01  2.953e-02   6.663 5.88e-11 ***
```

```
## OverallCondExc        2.558e-01  5.433e-02   4.708 3.08e-06 ***
## YearBuilt2nd20s      -7.775e-02  6.871e-02  -1.132 0.258251
## YearBuilt3nd20s      -1.029e-01  6.771e-02  -1.520 0.129030
## YearBuilt4th20s      -6.498e-02  7.242e-02  -0.897 0.369924
## YearBuilt5th20s      -8.025e-02  7.469e-02  -1.074 0.283043
## YearBuilt6th20s      -3.081e-02  7.682e-02  -0.401 0.688551
## YearBuilt7th20s      -4.307e-02  9.179e-02  -0.469 0.639044
## YearRemodAdd2nd20s    2.288e-03  2.013e-02   0.114 0.909552
## YearRemodAdd3rd20s    8.325e-04  1.885e-02   0.044 0.964791
## Exterior1stBrkFace    1.433e-01  5.997e-02   2.390 0.017139 *
## Exterior1stCemntBd   -7.423e-02  1.587e-01  -0.468 0.640210
## Exterior1stHdBoard    1.145e-01  5.795e-02   1.975 0.048653 *
## Exterior1stMetalSd    1.204e-01  8.752e-02   1.376 0.169416
## Exterior1stPlywood    1.080e-01  5.849e-02   1.847 0.065206 .
## Exterior1stStucco     8.578e-02  8.316e-02   1.031 0.302705
## Exterior1stVinylSd    5.030e-02  7.510e-02   0.670 0.503317
## Exterior1stWdSdng     1.177e-02  6.147e-02   0.191 0.848271
## Exterior1stWdShing    8.177e-02  7.055e-02   1.159 0.246873
## Exterior2ndBrk Cmn    2.297e-01  1.050e-01   2.186 0.029153 *
## Exterior2ndBrkFace   -1.138e-01  6.596e-02  -1.725 0.085058 .
## Exterior2ndCmentBd    1.628e-01  1.536e-01   1.059 0.289846
## Exterior2ndHdBoard   -5.131e-02  4.641e-02  -1.105 0.269425
## Exterior2ndMetalSd   -3.114e-02  8.201e-02  -0.380 0.704284
## Exterior2ndPlywood   -1.609e-02  4.446e-02  -0.362 0.717514
## Exterior2ndStucco    -1.671e-01  7.755e-02  -2.155 0.031538 *
## Exterior2ndVinylSd    1.606e-02  6.509e-02   0.247 0.805144
## Exterior2ndWdSdng     7.365e-02  5.174e-02   1.424 0.155070
## Exterior2ndWd Shng   -8.740e-02  5.827e-02  -1.500 0.134149
## MasVnrTypeBrkFace    -4.981e-03  5.957e-02  -0.084 0.933384
## MasVnrTypeNone       -7.641e-03  5.998e-02  -0.127 0.898667
## MasVnrTypeStone       2.255e-02  6.230e-02   0.362 0.717501
## MasVnrArea            1.738e-05  4.165e-05   0.417 0.676583
## ExterQualGood         1.852e-02  1.838e-02   1.008 0.313865
## FoundationCBlock      4.954e-02  2.299e-02   2.155 0.031570 *
## FoundationPConc       3.629e-02  2.459e-02   1.476 0.140538
## BsmtQual2            -3.819e-02  2.061e-02  -1.853 0.064355 .
## BsmtQual1            -3.687e-02  2.293e-02  -1.608 0.108376
## BsmtExposure2         4.704e-02  2.300e-02   2.046 0.041210 *
## BsmtExposure3        -3.370e-02  2.197e-02  -1.534 0.125583
## BsmtExposure4        -2.548e-02  1.550e-02  -1.645 0.100564
## BsmtExposureNOB      -1.692e-01  4.985e-02  -3.394 0.000733 ***
## BsmtFinType1Other    -4.431e-02  1.329e-02  -3.334 0.000905 ***
## BsmtFinSF1           -7.915e-05  3.379e-05  -2.343 0.019461 *
## BsmtUnfSF            -8.565e-05  3.530e-05  -2.426 0.015531 *
## TotalBsmtSF           5.129e-05  4.153e-05   1.235 0.217302
## HeatingQCGood        -3.282e-02  1.418e-02  -2.315 0.020916 *
## CentralAirY           8.896e-02  2.739e-02   3.247 0.001228 **
## X1stFlrSF             1.036e-04  1.342e-04   0.772 0.440134
## X2ndFlrSF             8.344e-05  1.305e-04   0.639 0.522801
## GrLivArea             1.163e-04  1.307e-04   0.890 0.373850
## BsmtFullBath          3.996e-02  1.395e-02   2.865 0.004316 **
## FullBath              7.439e-02  1.688e-02   4.408 1.23e-05 ***
## HalfBath              5.806e-02  1.602e-02   3.624 0.000313 ***
## BedroomAbvGr          1.163e-02  1.060e-02   1.097 0.273036
```

```
## KitchenAbvGr         -1.545e-01  5.137e-02  -3.007 0.002742 **
## KitchenQualFa        -1.691e-01  4.800e-02  -3.522 0.000459 ***
## KitchenQualGd        -8.025e-02  2.387e-02  -3.362 0.000821 ***
## KitchenQualTA        -1.078e-01  2.781e-02  -3.875 0.000118 ***
## TotRmsAbvGrd          7.790e-03  7.190e-03   1.083 0.279029
## Fireplaces           -1.926e-03  2.033e-02  -0.095 0.924544
## FireplaceQu2         -3.621e-02  4.808e-02  -0.753 0.451707
## FireplaceQu3         -6.780e-02  3.579e-02  -1.894 0.058648 .
## FireplaceQu4         -7.985e-02  5.743e-02  -1.390 0.164903
## FireplaceQu5         -6.996e-02  3.774e-02  -1.854 0.064207 .
## FireplaceQuNOF       -1.152e-01  4.414e-02  -2.610 0.009279 **
## GarageType2           1.260e-02  2.098e-02   0.601 0.548177
## GarageType1          -1.795e-02  2.471e-02  -0.727 0.467778
## GarageYrBlt2nd20s    -5.932e-02  5.797e-02  -1.023 0.306533
## GarageYrBlt3rd20s    -1.002e-01  5.470e-02  -1.831 0.067530 .
## GarageYrBlt4th20s    -1.109e-01  5.495e-02  -2.019 0.043939 *
## GarageYrBlt5th20s    -1.255e-01  5.487e-02  -2.288 0.022473 *
## GarageYrBlt6ths      -8.004e-02  6.995e-02  -1.144 0.252971
## GarageFinish2        -3.284e-03  1.508e-02  -0.218 0.827713
## GarageFinish3        -3.208e-02  1.785e-02  -1.797 0.072848 .
## GarageCars            7.640e-02  1.590e-02   4.804 1.95e-06 ***
## GarageArea            2.324e-05  5.575e-05   0.417 0.676951
## PavedDriveP          -2.503e-02  4.286e-02  -0.584 0.559385
## PavedDriveY          -1.170e-02  2.680e-02  -0.437 0.662549
## WoodDeckSF            2.232e-05  4.614e-05   0.484 0.628789
## OpenPorchSF           1.642e-04  8.628e-05   1.903 0.057467 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.123 on 626 degrees of freedom
## Multiple R-squared:  0.9178, Adjusted R-squared:  0.9015
## F-statistic: 56.34 on 124 and 626 DF,  p-value: < 2.2e-16
```

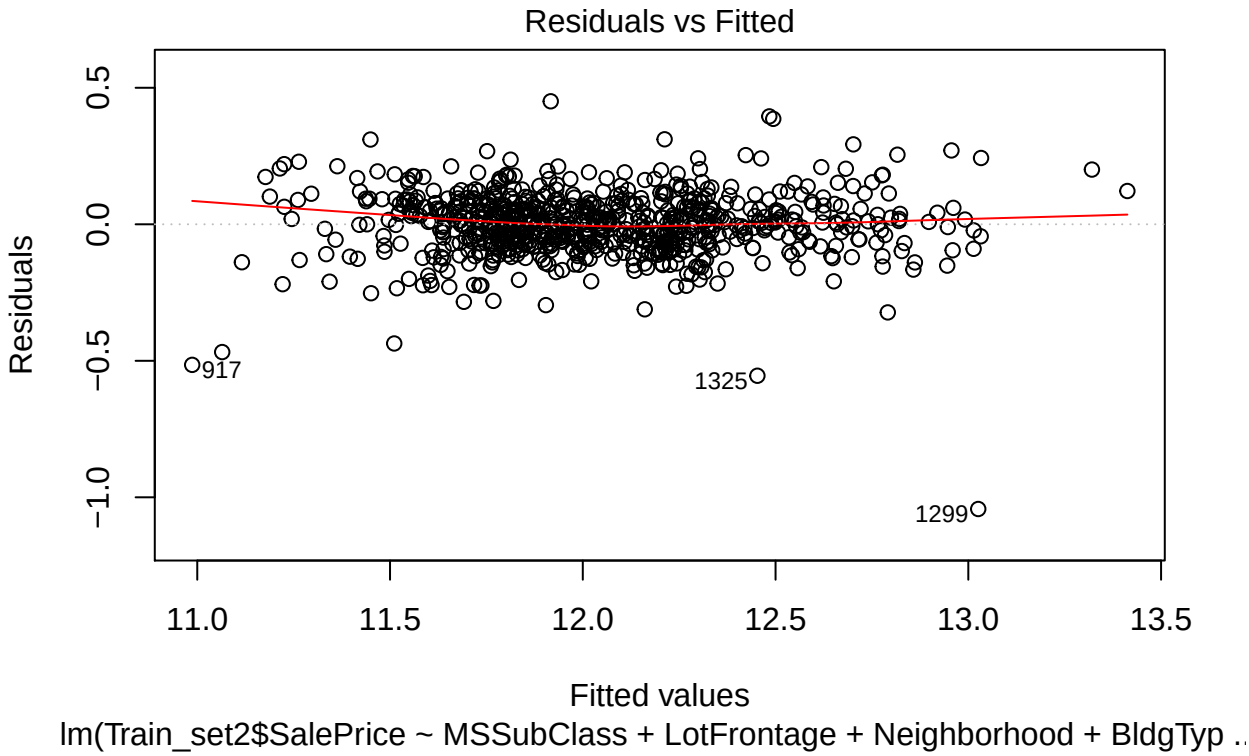
The output of lm_fit is much too large to allow an objective interpretation on each variable. However, we notice that R-squared is 0.92, this mains that aproximatively 92% of the variation in SalePrice is explained by the model which is a good indicator. We also have the ajusted R-squared that is 0.90. The adjusted R-squared is a penalizer coefficient. In fact, when we add a new veriable to our model R-squared will allways increase whereas the ajusted will not when the added variable does not increase the accuracy of the model. We can do model selection to reduce the number of predictors and build a much more simple and accurate model. The Akaike Information Criterion (AIC) will be used to perform this task. It can be performed using step() function in R.

```
##
## Call:
## lm(formula = Train_set2$SalePrice ~ MSSubClass + LotFrontage +
##     Neighborhood + BldgType + HouseStyle + OverallQual + OverallCond +
##     Exterior2nd + Foundation + BsmtExposure + BsmtFinType1 +
##     BsmtFinSF1 + BsmtUnfSF + HeatingQC + CentralAir + X1stFlrSF +
##     X2ndFlrSF + BsmtFullBath + FullBath + HalfBath + KitchenAbvGr +
##     KitchenQual + TotRmsAbvGrd + FireplaceQu + GarageType + GarageYrBlt +
##     GarageFinish + GarageCars + OpenPorchSF, data = Train_set2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04281 -0.06324  0.00489  0.06556  0.45048
```

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.142e+01  1.181e-01  96.670  < 2e-16 ***
## MSSubClassOneHfStoty -2.211e-02  4.195e-02  -0.527 0.598381
## MSSubClassTwoStory   -6.714e-03  4.590e-02  -0.146 0.883749
## MSSubClassTwoHfStory  3.798e-02  6.482e-02   0.586 0.558102
## MSSubClassSplit      -5.447e-02  4.626e-02  -1.178 0.239343
## MSSubClassDuplex     -2.303e-01  1.398e-01  -1.647 0.099943 .
## MSSubClassPud        -2.771e-01  1.330e-01  -2.083 0.037665 *
## MSSubClassPudM       -3.849e-01  1.421e-01  -2.709 0.006927 **
## MSSubClassTwoFam     -1.836e-01  1.369e-01  -1.341 0.180323
## LotFrontage          -7.329e-04  2.463e-04  -2.976 0.003028 **
## NeighborhoodBlueste  -3.216e-02  1.391e-01  -0.231 0.817172
## NeighborhoodBrDale    5.541e-02  7.201e-02   0.769 0.441908
## NeighborhoodBrkSide   1.376e-02  6.010e-02   0.229 0.819046
## NeighborhoodClearCr   1.814e-01  7.330e-02   2.474 0.013608 *
## NeighborhoodCollgCr   7.604e-02  5.090e-02   1.494 0.135699
## NeighborhoodCrawfor   1.991e-01  5.708e-02   3.488 0.000519 ***
## NeighborhoodEdwards  -7.411e-03  5.482e-02  -0.135 0.892507
## NeighborhoodGilbert   4.384e-02  5.318e-02   0.824 0.410011
## NeighborhoodIDOTRR   -1.436e-01  6.405e-02  -2.242 0.025274 *
## NeighborhoodMeadowV  -8.255e-02  8.792e-02  -0.939 0.348088
## NeighborhoodMitchel   4.229e-02  5.571e-02   0.759 0.448009
## NeighborhoodNAmes     1.843e-02  5.364e-02   0.344 0.731281
## NeighborhoodNoRidge   2.285e-01  5.993e-02   3.813 0.000150 ***
## NeighborhoodNPkVill  -1.200e-01  1.094e-01  -1.097 0.272910
## NeighborhoodNridgHt   1.859e-01  5.087e-02   3.655 0.000278 ***
## NeighborhoodNWAmes    1.497e-02  5.664e-02   0.264 0.791661
## NeighborhoodOldTown  -9.298e-02  5.656e-02  -1.644 0.100703
## NeighborhoodSawyer    3.116e-02  5.724e-02   0.544 0.586340
## NeighborhoodSawyerW   5.547e-02  5.503e-02   1.008 0.313845
## NeighborhoodSomerst   1.584e-01  5.262e-02   3.011 0.002702 **
## NeighborhoodStoneBr   2.351e-01  5.571e-02   4.219 2.79e-05 ***
## NeighborhoodSWISU    -6.633e-02  6.615e-02  -1.003 0.316395
## NeighborhoodTimber    7.714e-02  5.738e-02   1.344 0.179349
## NeighborhoodVeenker   7.482e-02  8.915e-02   0.839 0.401649
## BldgTypeOTHERS        2.086e-01  1.305e-01   1.599 0.110390
## HouseStyleOneStory    5.701e-03  4.075e-02   0.140 0.888783
## HouseStyleTwoStory   -8.124e-02  4.151e-02  -1.957 0.050740 .
## OverallQualavg        9.531e-02  2.072e-02   4.601 5.05e-06 ***
## OverallQualGood       1.727e-01  2.585e-02   6.681 5.03e-11 ***
## OverallQualExc        2.747e-01  3.957e-02   6.942 9.25e-12 ***
## OverallCondavg        1.488e-01  2.633e-02   5.652 2.36e-08 ***
## OverallCondGood       2.042e-01  2.802e-02   7.286 9.11e-13 ***
## OverallCondExc        2.616e-01  5.025e-02   5.206 2.58e-07 ***
## Exterior2ndBrk Cmn    2.264e-01  1.030e-01   2.197 0.028351 *
## Exterior2ndBrkFace   -2.281e-02  5.121e-02  -0.446 0.656099
## Exterior2ndCmentBd    5.142e-02  4.071e-02   1.263 0.207029
## Exterior2ndHdBoard    1.775e-02  3.160e-02   0.562 0.574508
## Exterior2ndMetalSd    3.792e-02  3.033e-02   1.250 0.211711
## Exterior2ndPlywood    3.491e-02  3.380e-02   1.033 0.302018
## Exterior2ndStucco    -1.303e-01  4.941e-02  -2.637 0.008551 **
## Exterior2ndVinylSd    2.481e-02  3.106e-02   0.799 0.424681
```

```
## Exterior2ndWdSdng     4.519e-02  3.056e-02   1.479 0.139692
## Exterior2ndWd Shng   -7.484e-02  4.019e-02  -1.862 0.063003 .
## FoundationCBlock      4.923e-02  2.147e-02   2.293 0.022157 *
## FoundationPConc       3.820e-02  2.343e-02   1.630 0.103526
## BsmtExposure2         5.959e-02  2.190e-02   2.721 0.006677 **
## BsmtExposure3        -3.319e-02  2.139e-02  -1.552 0.121141
## BsmtExposure4        -2.543e-02  1.513e-02  -1.680 0.093339 .
## BsmtExposureNOB      -1.439e-01  4.243e-02  -3.393 0.000733 ***
## BsmtFinType1Other    -4.453e-02  1.257e-02  -3.541 0.000426 ***
## BsmtFinSF1           -4.444e-05  2.321e-05  -1.915 0.055918 .
## BsmtUnfSF            -4.537e-05  2.391e-05  -1.897 0.058202 .
## HeatingQCGood        -2.938e-02  1.322e-02  -2.223 0.026583 *
## CentralAirY           9.526e-02  2.536e-02   3.756 0.000188 ***
## X1stFlrSF             2.338e-04  3.281e-05   7.125 2.72e-12 ***
## X2ndFlrSF             2.067e-04  3.596e-05   5.749 1.37e-08 ***
## BsmtFullBath          4.446e-02  1.341e-02   3.317 0.000960 ***
## FullBath              8.303e-02  1.597e-02   5.200 2.66e-07 ***
## HalfBath              6.071e-02  1.507e-02   4.030 6.24e-05 ***
## KitchenAbvGr         -1.578e-01  4.759e-02  -3.315 0.000966 ***
## KitchenQualFa        -1.635e-01  4.610e-02  -3.547 0.000417 ***
## KitchenQualGd        -8.620e-02  2.291e-02  -3.762 0.000184 ***
## KitchenQualTA        -1.074e-01  2.608e-02  -4.119 4.28e-05 ***
## TotRmsAbvGrd          1.196e-02  6.200e-03   1.930 0.054081 .
## FireplaceQu2         -1.939e-02  4.682e-02  -0.414 0.678941
## FireplaceQu3         -6.054e-02  3.482e-02  -1.739 0.082543 .
## FireplaceQu4         -7.594e-02  5.561e-02  -1.365 0.172578
## FireplaceQu5         -5.816e-02  3.677e-02  -1.582 0.114181
## FireplaceQuNOF       -1.015e-01  3.655e-02  -2.777 0.005642 **
## GarageType2           1.740e-02  2.038e-02   0.854 0.393571
## GarageType1          -1.381e-02  2.339e-02  -0.590 0.555096
## GarageYrBlt2nd20s    -5.211e-02  4.785e-02  -1.089 0.276562
## GarageYrBlt3rd20s    -7.868e-02  4.758e-02  -1.654 0.098692 .
## GarageYrBlt4th20s    -9.973e-02  4.787e-02  -2.083 0.037627 *
## GarageYrBlt5th20s    -9.116e-02  4.818e-02  -1.892 0.058919 .
## GarageYrBlt6ths      -4.770e-02  5.049e-02  -0.945 0.345202
## GarageFinish2        -1.186e-02  1.457e-02  -0.814 0.416044
## GarageFinish3        -3.606e-02  1.717e-02  -2.100 0.036081 *
## GarageCars            7.977e-02  1.157e-02   6.896 1.25e-11 ***
## OpenPorchSF           1.854e-04  8.292e-05   2.235 0.025726 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1224 on 661 degrees of freedom
## Multiple R-squared:  0.9139, Adjusted R-squared:  0.9024
## F-statistic: 78.88 on 89 and 661 DF,  p-value: < 2.2e-16
```

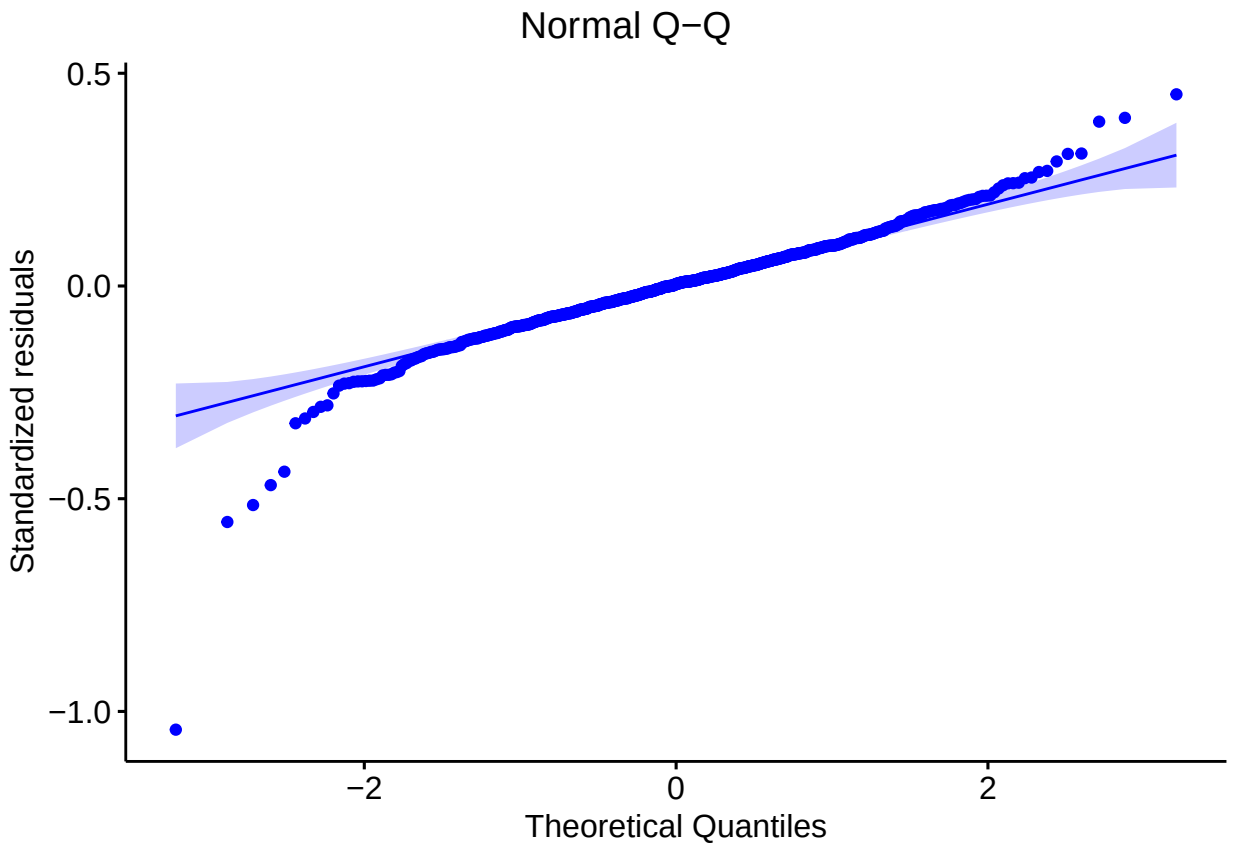
With the AIC criterion we are aible to drop many predictors and endup with 26 variables. Also we notice that the R-squared and Adjusted R-squared are roughly the same as the full model builded above. This essentialy means that the models have the same accuracy eventhough Best_lm has less varibles. The plot below is the Residuals vs Fitted.



Generaly this plot is used to check the constance error variance and linearity assumptions of the model. We can see that the plot does show some tendency of lack of fit. The red line in the middle is slightly curved. The plot also indicates that the variance is not constant. In fact, the second half of the plot indicate a slitely bigger variance than the first half. In other words, the residuals increase as the fitted values increase. So, the inference here is, heteroscedasticity exists. We can test constant variance assumption using the bptest function in r.

```
##
##  studentized Breusch-Pagan test
##
## data:  Best_lm
## BP = 371.32, df = 89, p-value < 2.2e-16
```

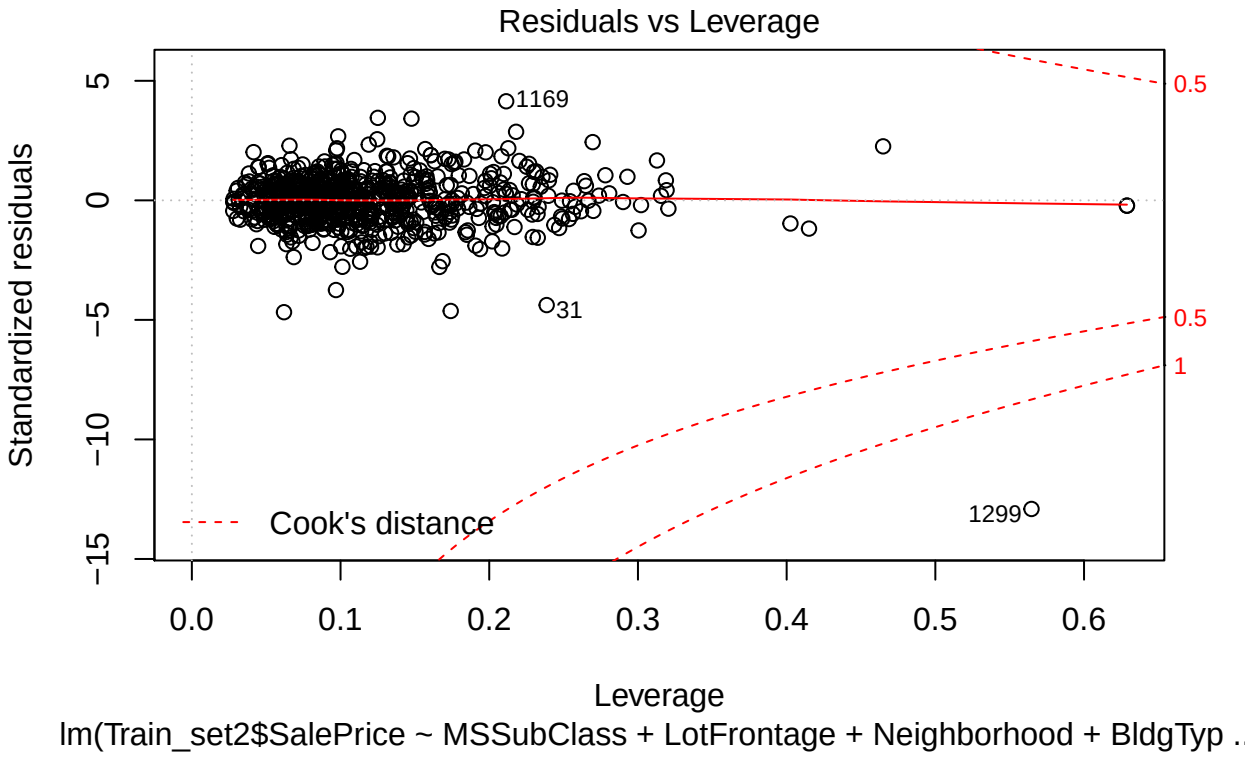
From the output, the p-value < 0.05 this imply that the variance of the residuals is not constant and infer that heteroscedasticity is indeed present, which confirm the graphical inference we made above. One way to fix this lack of fit is to build the model with some other variables or do some variable transformation using functions such as log, sqrt, etc. Another way is to use Box-Cox transformation. Box-Cox is a mathematical transformation of the variable to make it approximate to a normal distribution. Often, doing a box-cox transformation of the response variable solves the issue. Here we will not be fixing this issue insted we will build some other models and compare one model to another in order to chose the best one. Now we can check whether the model satisfies the normality assumption by plotting the QQ_Normal plot.



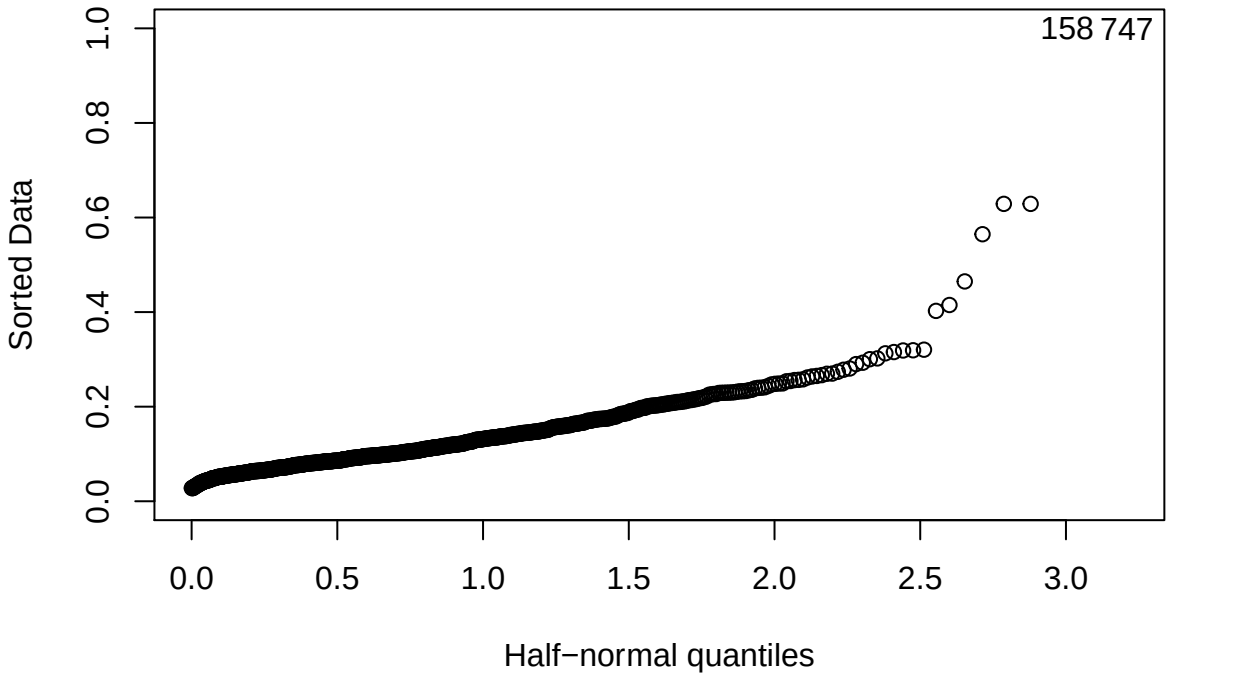
The major portion of the observations follow a line with few exceptions. However, we can see at the beginning and at the end of the plot that some observation deviate from the line. This indicates a long-tailed error. This suggest that we should consider robust fitting. We can use Shapiro-Wilk test to test this normality assumption.

```
##
##  Shapiro-Wilk normality test
##
## data:  Best_lm$residuals
## W = 0.92666, p-value < 2.2e-16
```

From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution. In other words, we can not assume the normality. However, for large dataset such as this one the normality assumption is not crusial, as the inference will be approximately correct in spite of the nonnormality. Here the deviation from normality is acceptable, therefore, we won't be changing the model because of nonnormality of the residuals. We can also check the presence or absence of outliers by plotting Residuals vs Leverage.

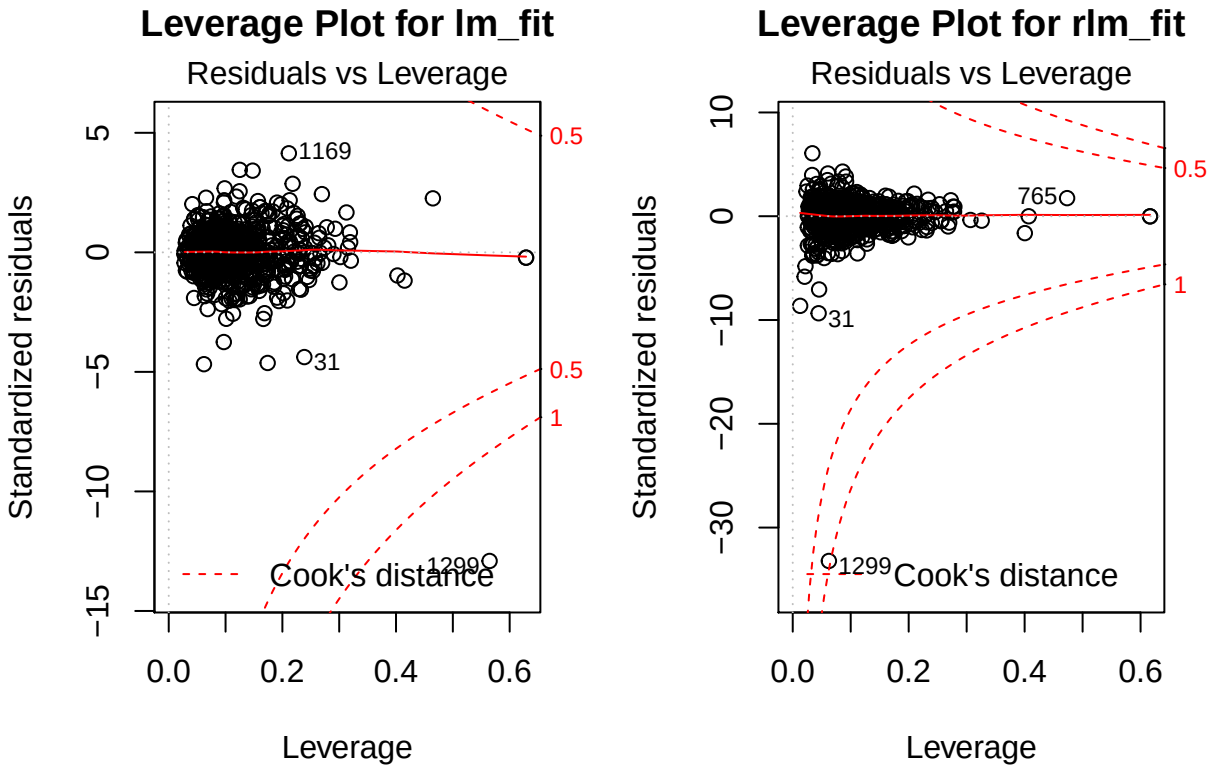


There are few observations that lie out of the contour lines for Cook statistics. These observations represent the outliers. We must study them closly so, we can identify there effect on the model. For more detail about the outliers we can plot the half-normal plot for the leverage.

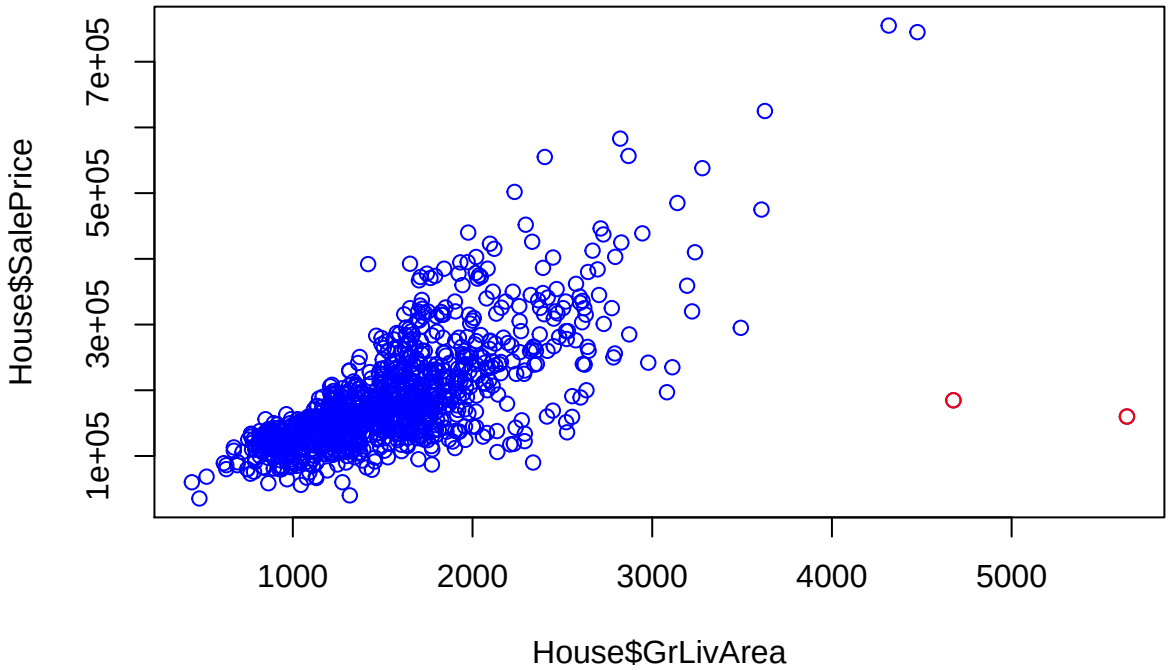


This plot shows six points with much higher leverage than the rest. One way to deal with the outliers is to delete them fom the data, but this leads in generale to other ouliers. So we will use robust linear model to take care of these outliers. Robust regression is an alternative to the least square (linear model) approch that downweights the effect of larger errors. The function rlm from the MASS package is used to fit robust

regression in R.



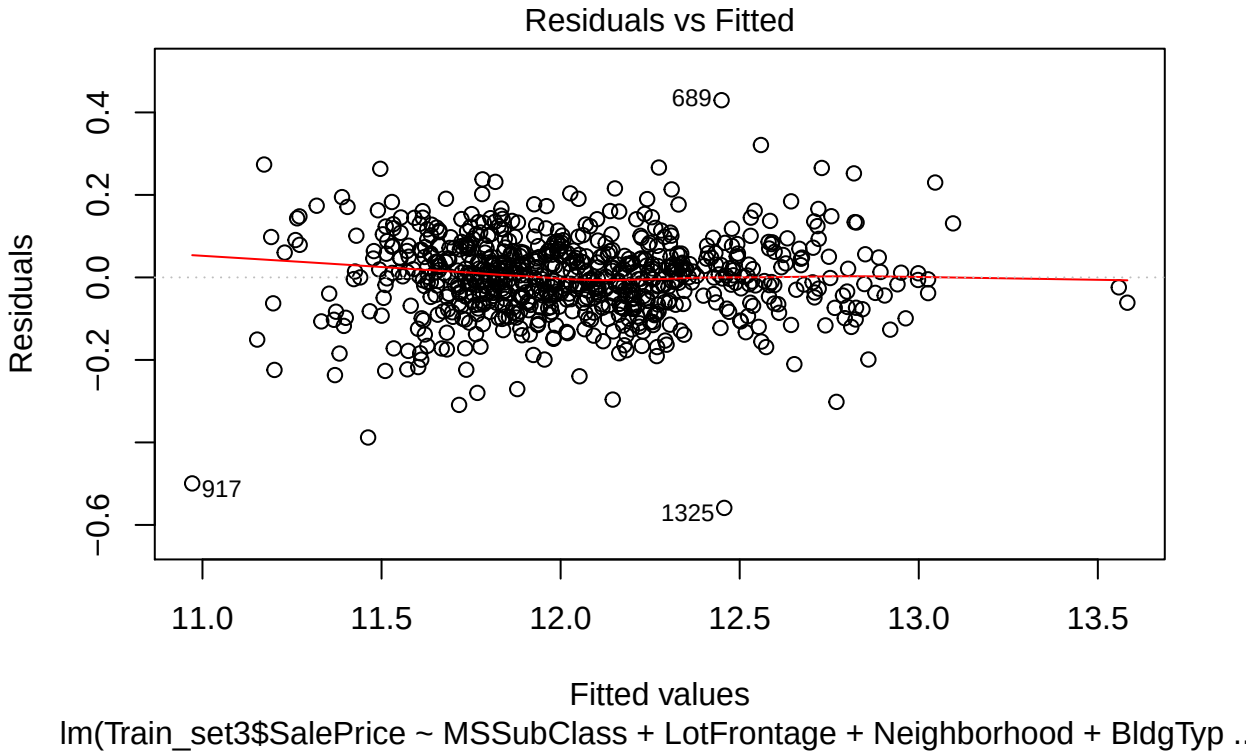
As shown from the plots robust linear method reduces the effect of the outliers on the model. But still the house id 1299 remains influential. The houses 31, 765 are also influential, but they are contained within the limits of cooks distance. We can take look at the position of these houses by plotting SalePrice against GrLivArea whith determine the geographical position of a house.



This plot shows two houses with a verry low price highled with red color, these houses doesn't fallow the natural progression of SalePrice. So we will go back to the data and delete them. There are two other points

on the top of the plot that seems influential. However, we won't remove them because they fallow the natural progress of SalePrice.

After we remove the outliers from the original data, we can fit again linear model with 26 variables, selected by the Best_lm model and check our assumptions.




```
##
##  studentized Breusch-Pagan test
##
## data:  New.lm_fit
## BP = 121.93, df = 89, p-value = 0.01175
```

As seen on the plot above, New.lm_fit is a better model than the sofar fitted models. We can assum that the error has constanstant variance with a α level of 0.01 with was not the case with previous models. New.lm_fit will than be used for the rest of the analysis.

Test Error for the sofor fitted models

Here we will calculate the errors on the test set. The total error of a model is composed of three different terms: *bias*, *variance* and the *the irreducible error*. One task of a statistician is to minimize the bias and the variance, the goal is to reduce this two terms to zero. The irreducible error, is the noise term in the true relationship (the fitted model vs a model that exactly predict SalePrice) and it cannot be reduced by any model. The term $bais^2 + variance + irr\_error$ is call the mean squared error (MSE).

Robust regression test MSE, Bias Square and Variance

MSE

```
## [1] 0.01815682
```

Variance

The variance is define as the variability of a model prediction for a given data point.

```
## [1] 0.1431293
```

Bias

The bias is define as the mean of difference between the expected prediction of our model and the correct value which we are trying to predict.

```
## [1] 0.01087349
```

Linear regression test error, bias square and variance (New.lm_fit)

MSE

```
## [1] 0.01819873
```

Variance

```
## [1] 0.1398916
```

Bias

```
## [1] 0.003113556
```

New.lm_fit performs better on the test set than the rlm_fit, therefore, we be using **New.lm_fit** for forwarder analysis. In fact the lover the MSE the better the model. This means that the lower the bias and the variance the better the model. In this case the linear model produices smaller MSE, variance and bias.

Random Forest

Random Forest model is developed by aggregeting trees. In insted of building one tree, we create a lot of decision trees and aggregate all the results. We can use *Random Forest* for *classification* and *regression*. In this case it will be use for regresion because salePrice is a numerical variable. *Random Forest Model* gives a lot of advantages. Among them there is the possibility that it provides for variable selection based on their importance.

Fitting random forest model

We use the function randomForest from randomForest packege. As stated above random forest first aggregate a results given by many decision trees, the default value for the number of trees is 500. Than it select randomly a sample from the data. Each sample use a fix number of variables. For classification model it uses the square root of the number of feattures, and for regresion it uses the third of the number of variables. Here, we will be fitting random forest model on the 46 variables obtained from the preliminary variable selection using Boruta function and we will be using the data in which we have already remove the two outliers discribes above.

```
##
## Call:
##  randomForest(formula = SalePrice ~ MSSubClass + MSZoning + LotFrontage +      LotArea + LotShape + L
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 15
##
##           Mean of squared residuals: 0.01905326
##                     % Var explained: 87.57
```

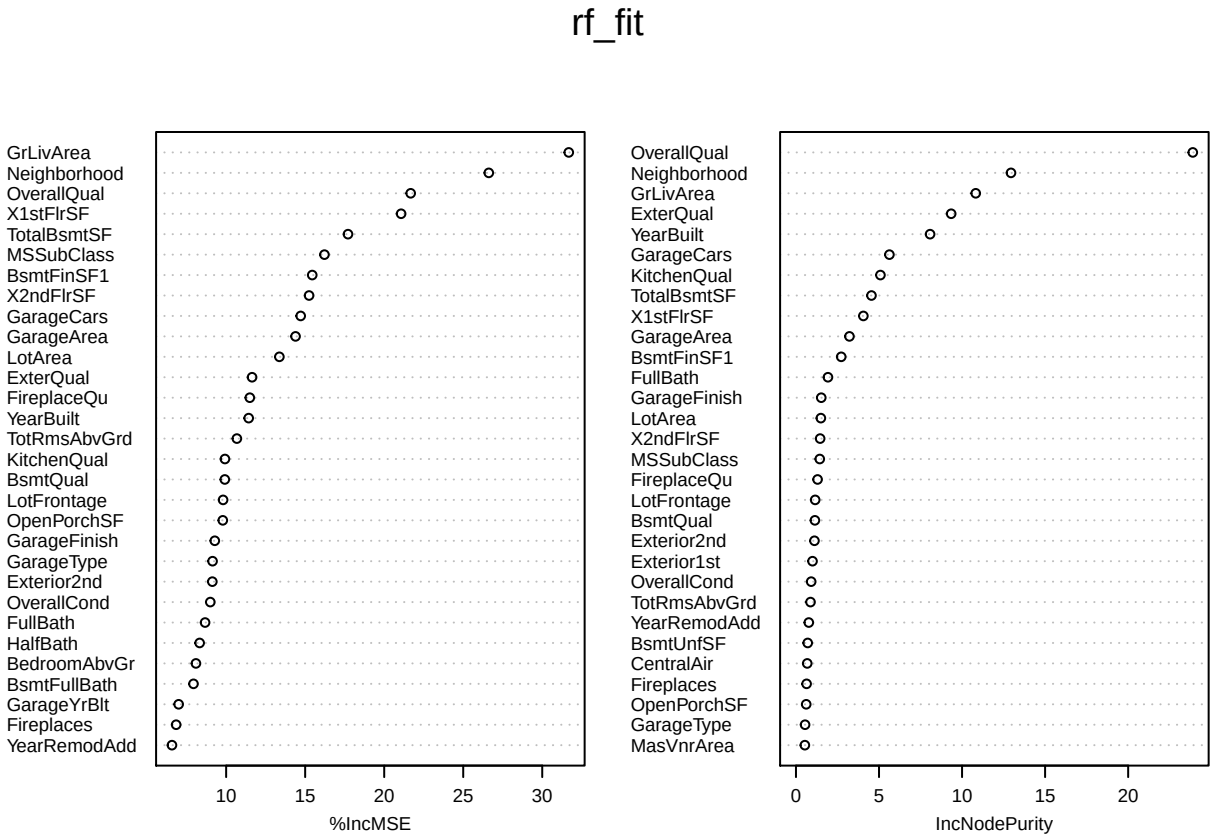
We can see that the model is 87.57% accurate. Whith is not far from the what we have since in the previous model (about 90%)

Variable importance

	%IncMSE	IncNodePurity
MSSubClass	16.2185735	1.4328181
MSZoning	5.1724973	0.1524946
LotFrontage	9.8058493	1.1566072
LotArea	13.3681972	1.4976714
LotShape	3.0142906	0.0665056
Neighborhood	26.6181129	12.9460780
BldgType	4.4529989	0.1097161
HouseStyle	5.7751631	0.1603062
OverallQual	21.6767374	23.8906583
OverallCond	8.9972739	0.9152110
YearBuilt	11.4251858	8.0748405
YearRemodAdd	6.5704828	0.7715588
Exterior1st	6.3675967	0.9949213
Exterior2nd	9.1244421	1.1131201
MasVnrType	5.4365324	0.1616016
MasVnrArea	5.7915805	0.5353842
ExterQual	11.6399132	9.3446528
Foundation	5.4317812	0.5279260
BsmtQual	9.9152642	1.1401165
BsmtExposure	5.1233859	0.3530570
BsmtFinType1	6.1361287	0.1160176
BsmtFinSF1	15.4471159	2.7240017
BsmtUnfSF	6.0162761	0.7058044
TotalBsmtSF	17.7100607	4.5449762
HeatingQC	3.6297846	0.1266642
CentralAir	5.2204985	0.6776148
X1stFlrSF	21.0711379	4.0572846
X2ndFlrSF	15.2467484	1.4552260
GrLivArea	31.6792522	10.8234213
BsmtFullBath	7.9324114	0.2339080
FullBath	8.6664165	1.9255104
HalfBath	8.3219522	0.2872985
BedroomAbvGr	8.0867058	0.4891262
KitchenAbvGr	4.8601528	0.0955229
KitchenQual	9.9226164	5.0845491
TotRmsAbvGrd	10.6778916	0.8749161
Fireplaces	6.8350663	0.6388713
FireplaceQu	11.4969604	1.3002975
GarageType	9.1382510	0.5470921
GarageYrBlt	6.9935541	0.4878482
GarageFinish	9.2808299	1.5223976
GarageCars	14.7156216	5.6258811
GarageArea	14.3894012	3.2190892
PavedDrive	0.2859949	0.1150035
WoodDeckSF	3.2332380	0.3312830
OpenPorchSF	9.7864355	0.6159312

The %InMSE indicate the mean decrease in accuracy of the model when we remove a variable, and the IncNodePurity indicate the total mean decrease in node impurity that result from splitting over variables. In another words, these two parameters measure how important is a given variable. For example if the variable Neighborhood is drop from the model the MSE (Mean Square Error) will increase by 21% and the node impurity will also drasticly increase. This mean that Neiborhood is very important for the analysis. We can

also notice the verry low %InMSE of the variable PavedDrive which means that the variable is not important for the model. Removing it will also result in an unsignificant increase of the node impurity. The following plot is the graphical display of variables' importance.



This plot shows that the variables Neiborhood, OverallQual, GrLivArea and GarageCars are the top four importante variable for predincting SalePrice. Whereas, the varibles Fireplaces, BedroomAbvGr, YearRemodAdd and WoodDeckSF are the least four important variables.

Test errors

MSE

```
## [1] 0.01894265
```

Viance

```
## [1] 0.1133001
```

Bias

```
## [1] 0.005779385
```

The test error is smaller than the MSE of train set, which is a good indicator of the accuracy of the model.

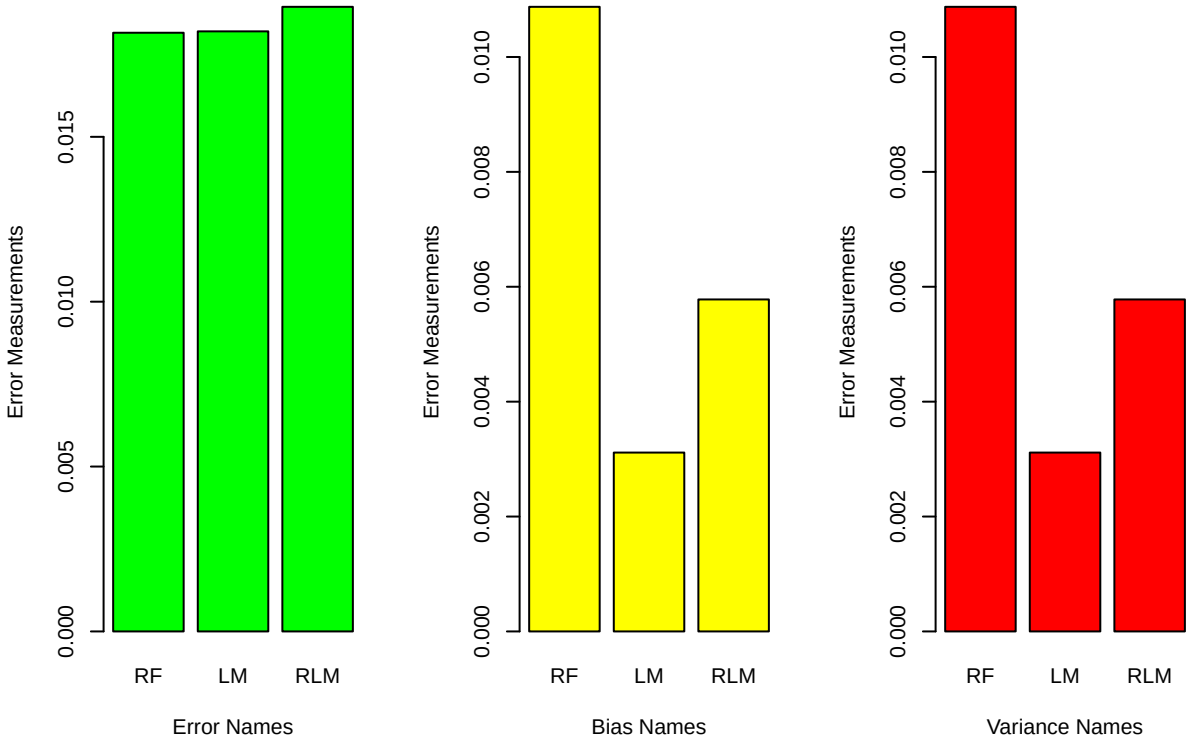
Conclusion

Durring the process of this analysis we have fitted three different models for predicting SalePrice. Here, are looking for a comparative way that allow to choose the best possible model. As we stated in the body of this paper, to have an insight of the best model we must look at their MSE, Bias and Variance. The model with the lower test errors will be the best model. The table bellow print test errors of the three different models.

Table 2: Errors Tatble

	MSE	BIAS	VARIANCE
rlm	0.0189427	0.0057794	0.1133001
lm	0.0181987	0.0031136	0.1398916
rf	0.0181568	0.0108735	0.1431293

A visual over look the table (MSE, BIAS, VARIANCE) is given.



A comparative analysis of the *linear regression* (lm) model and *random forest* (rf) model show that linear model is a better fit for this data. We can look at the 10 first predicted values by the linear model and compare them to the first 10 actual values of SalePrice.

Table 3: Comparative Table

log(SalePrice)	PredictedValue	Difference
12.90669	12.85058	0.056
12.46844	12.53741	0.069
11.66993	11.78308	0.113
12.34583	12.26937	0.076
12.00151	12.02270	0.021
12.96219	12.95077	0.011
12.23077	12.17867	0.052
12.13619	11.67860	0.037
11.60824	11.58847	0.020
11.71587	12.18197	0.046

We can see that the difference between the predicted and the actual values is very small for most of the

entries. In fact the highest difference in absolute value is 0.069 which is less than 1% of the actual value that coorespond to it.

Reference

Robert Gentleman, Kurt Hornik, and Giovanni Parmigiani, *An Introduction to Applied Multivariate Analysis with R*, Springer, 2011.

Gareth James, Daniela Witten, and Trevor Hastie *An Introduction to Statistical Learning*, Springer+Business Media, 2013.

Jay L. Devore, and Kenneth N. Berk, *Modern Mathematical Statistics With Applications*, Springer, Second Edition, 2012.

Julian J. Faraway *Extending the Linear Model with R*: Generalized Linear, Mixed Effect and Nonparametric Regression Models, CRC Press, Second edition, 2016.

Kutner, Nachtsheim, Neter and Li, *Applied Linear Statistical Models*, McGraw-Hill!Irwin, 5th edition, 2015.

Appendix

R packages

```
library(faraway)
library(lattice)
library(caret)
library(dummies)
library(forcats)
library(magrittr)
library(dplyr)
library(ranger)
library(Boruta)
library(randomForest)
library(MASS)
library(glmnet)
library(tree)
library(pls)
library(ISLR)
library(lmtest)
library(ggplot2)
library(ggpubr)
library(car)
library(sandwich)
library(knitr)
library(sjPlot)
library(jtools)
```

Loading the Data

```
House1 = read.csv("train.csv")
House = House1[-1]
attach(House)
```

Structure and summary of the Data

```
str(House)
summary(House)
```

Data transformation

```
House$Alley = as.factor(ifelse(is.na(House$Alley), "NOA", House$Alley))
House$BsmtQual = as.factor(ifelse(is.na(House$BsmtQual), "NOB", House$BsmtQual))

House$BsmtCond = as.factor(ifelse(is.na(House$BsmtCond), "NOB", House$BsmtCond))

House$BsmtExposure = as.factor(ifelse(is.na(House$BsmtExposure), "NOB", House$BsmtExposure))

House$BsmtFinType1 = as.factor(ifelse(is.na(House$BsmtFinType1), "NOB", House$BsmtFinType1))
House$BsmtFinType2 = as.factor(ifelse(is.na(House$BsmtFinType2), "NOB", House$BsmtFinType2))

House$FireplaceQu = as.factor(ifelse(is.na(House$FireplaceQu), "NOF", House$FireplaceQu))

House$GarageType = as.factor(ifelse(is.na(House$GarageType), "NOG", House$GarageType))

House$GarageFinish = as.factor(ifelse(is.na(House$GarageFinish), "NOG", House$GarageFinish))

House$GarageQual = as.factor(ifelse(is.na(House$GarageQual), "NOG", House$GarageQual))

House$GarageCond = as.factor(ifelse(is.na(House$GarageCond), "NOG", House$GarageCond))

House$PoolQC = as.factor(ifelse(is.na(House$PoolQC), "NOP", House$PoolQC))

House$Fence = as.factor(ifelse(is.na(House$Fence), "NOF", House$Fence))

House$MiscFeature = as.factor(ifelse(is.na(House$MiscFeature), "NONE", House$MiscFeature))
House$MiscFeature = as.factor(House$MiscFeature)
House$MSSubClass = as.factor(House$MSSubClass)
House$OverallQual = as.factor(House$OverallQual)
House$OverallCond = as.factor(House$OverallCond)
House$Utilities = as.factor(House$Utilities)
House$YrSold = as.factor(House$YrSold)
```

```
House = House[-9]
House$LotShape = fct_collapse(House$LotShape, REG = "Reg", IREG =
c("IR1", "IR2", "IR3"))

House$LandContour = fct_collapse(House$LandContour, Lvl = "Lvl", NotFlat =
c("Bnk", "HLS", "Low"))

House$MSZoning = fct_collapse(House$MSZoning, Rl = "RL",
OTHERS= c("C (all)", "FV", "RH", "RM"))

House$Alley = fct_collapse(House$Alley, NOA = "NOA",
OTHERS= c("1", "2"))

House$LotConfig = fct_collapse(House$LotConfig, Inside = "Inside",
FR = c("FR2", "FR3"), Corner  = "Corner", CulDSas = "CulDSac")

House$LandSlope = fct_collapse(House$LandSlope, Gtl = "Gtl",
OTHERS= c("Mod", "Sev"))

House$Condition1 = fct_collapse(House$Condition1, Norm = "Norm",
AbNorm = c("Feedr", "Artery", "PosN", "PosA", "RRNn", "RRNe", "RRAe", "RRAn"))
```

```
House$Condition2 = fct_collapse(House$Condition2, Norm = "Norm",
AbNorm = c("Feedr", "Artery", "PosN", "PosA", "RRNn", "RRNe", "RRAe", "RRAn"))

House$BldgType = fct_collapse(House$BldgType, "1Fam"="1Fam",
OTHERS= c("2fmCon", "Duplex", "Twnhs", "TwnhsE"))

House$HouseStyle = fct_collapse(House$HouseStyle, OneStory = "1Story", TwoStory
= "2Story", Others = c("1.5Unf", "1.5Fin", "2.5Fin", "2.5Unf", "SFoyer", "SLvl"))

House$OverallCond = fct_collapse(House$OverallCond, Exc = c("10", "9"), Good
= c("8", "7"),  avg = c("6", "5"), notGood = c("4", "3", "2", "1"))

House$OverallQual = fct_collapse(House$OverallQual, Exc = c("10", "9"),
Good= c("8", "7"),  avg = c("6", "5"), notGood = c("4", "3", "2", "1"))

House$RoofStyle = fct_collapse(House$RoofStyle, Gable = "Gable", Hit= "Hit",
Other = c("Gambrel", "Flat", "Shed", "Mansard"))

House$RoofMatl = fct_collapse(House$RoofMatl, CompShg = "CompShg", Other =
c("ClyTile", "Membran", "Metal", "Roll", "Tar&Gry", "WdShake", "WdShngl"))

House$Exterior1st = fct_collapse(House$Exterior1st, VinylSd = "VinylSd",
MetalSd= "MetalSd",  HdBoard = "HdBoard", WdSdng = "Wd Sdng", Plywood=
"Plywood", CemntBd ="CemntBd", BrkFace="BrkFace", WdShing ="WdShing",
Stucco = "Stucco", Other =c("AsbShng", "AsphShn", "BrkComm", "CBlock",
"ImStucc", "Other", "PreCast", "Stone"))

House$Exterior2nd = fct_collapse(House$Exterior2nd, VinylSd = "VinylSd",
MetalSd= "MetalSd",  HdBoard = "HdBoard", WdSdng = "Wd Sdng", Plywood=
"Plywood", CemntBd ="CemntBd", BrkFace="BrkFace", WdShing ="WdShing",
Stucco = "Stucco", Other =c("AsbShng", "AsphShn", "BrkComm", "CBlock",
"ImStucc", "Other", "PreCast", "Stone"))

House$ExterQual = fct_collapse(House$ExterQual, Exc = c("Ex", "Gd"), Good=
c("Fa", "TA"))

House$ExterCond = fct_collapse(House$ExterCond, Exc = c("Ex", "Gd"), Good=
c("Fa", "TA", "Po"))

House$Foundation = fct_collapse(House$Foundation, CBlock="CBlock",
PConc="PConc", Other= c("Slab", "Stone", "Wood", "BrkTil"))

House$BsmtQual = fct_collapse(House$BsmtQual, "1" = "4", "2" = "3",
OTHERS= c("1", "2", "NOB"))

House$BsmtFinType2 = fct_collapse(House$BsmtFinType2, "1" = "6",
OTHERS= c("1", "2", "3", "4", "5", "NOB"))

House$BsmtCond = fct_collapse(House$BsmtCond, "1" = "4",
OTHERS= c("1", "2", "3", "NOB"))

House$Heating = fct_collapse(House$Heating, Gas = c("GasA", "GasW"),
Other= c("Floor", "Grav", "OthW", "Wall"))
```

```
House$HeatingQC = fct_collapse(House$HeatingQC, Exc = c("Ex", "Gd"),
Good= c("Fa", "TA", "Po"))

House$Electrical = fct_collapse(House$Electrical, SBrkr = "Sbrkr",
Other= c("FuseA", "FuseF", "FuseP", "Mix"))

House$Functional = fct_collapse(House$Functional, Typ = "Typ",
Deduction= c("Maj1", "Maj2", "Min1", "Min2", "Mod", "Sev"))

House$GarageType = fct_collapse(House$GarageType, "1" = "6", "2" ="2",
Other= c("1", "3", "4", "5", "NOG"))

House$BsmtFinType1 = fct_collapse(House$BsmtFinType1, "AccQuarters" =
c("3", "1", "4"),  Other= c("NOB", "5", "2", "6"))

House$GarageFinish = fct_collapse(House$GarageFinish, "1" = "1", "2" ="2",
"3" =c("3", "NOG"))

House$GarageQual = fct_collapse(House$GarageQual, "1" = "5", "2"= c("1", "2",
"3", "4", "NOG"))

House$GarageCond = fct_collapse(House$GarageCond, "1" = "5", "2"= c("1", "2",
"3", "4", "NOG"))

House$PoolQC = fct_collapse(House$PoolQC, "1" = "NOP", "2" = c("1", "2", "3"))

House$Fence = fct_collapse(House$Fence, "1" = "NOP", "2" = c("1", "2", "3", "4"))

House$MiscFeature = fct_collapse(House$MiscFeature, "1" = "NONE", "2"=
c("1", "2", "3", "4"))

House$SaleType = fct_collapse(House$SaleType, WD = "WD", New = "New",
Other = c("CWD", "VWD", "COD", "Con", "ConLw", "ConLD", "Oth"))

House$SaleCondition = fct_collapse(House$SaleCondition, Normal = "Normal",
Abnorml = "Abnorml", Partial ="Partial", Other = c("AdjLand", "Family", "Alloca"))

House$MoSold = cut(House$MoSold, breaks = c(1, 6, 12), labels =
c("1stFyear", "2ndFyear"), right = FALSE)

House$YearBuilt = cut(House$YearBuilt, breaks = c(1880, 1900, 1920, 1940,
1960, 1980, 2000, 2010), labels = c("1st20s", "2nd20s",  "3nd20s",
"4th20s", "5th20s", "6th20s", "7th20s"), right = FALSE)

House$YearRemodAdd = cut(House$YearRemodAdd, breaks = c(1950, 1970, 1990,
2010), labels = c( "1st20s", "2nd20s", "3rd20s"), right = FALSE)

House$MSSubClass = fct_collapse(House$MSSubClass, OneStory =
c("20", "30", "40"), OneHfStoty = c("45", "50"), TwoStory =
c("60", "70"), TwoHfStory = "75",  Split = c("80", "85"), Duplex = "90",
Pud = c("120", "150"), PudM = c("160", "180"), TwoFam = "190")
```

```
House$GarageYrBlt = cut(House$GarageYrBlt, breaks = c(1900,  1920,  1940,
1960,  1980,  2000, 2010), labels = c("1st20s", "2nd20s",  "3rd20s",
"4th20s", "5th20s",  "6ths"), right = FALSE)
```

remove NAs

```
House = na.omit(House)
sum(is.na(House))
dim(House)
```

Preliminary Variable Selection

```
set.seed(100)
H2 = Boruta(SalePrice ~ ., data = House, doTrace = 0, maxRuns = 200)
print(H2)
H3 = TentativeRoughFix(H2)
getNonRejectedFormula(H3)
```

Splitting the Data

```
set.seed(200)
Train_index = sample(dim(House)[1], dim(House)*0.70)
Train_set = House[Train_index, ]
Test_set = House[-Train_index, ]
```

Linear model

```
set.seed(300)
lm_fit =  lm(Train_set$SalePrice ~ MSSubClass + MSZoning + LotFrontage + LotArea + LotShape +
    Neighborhood + BldgType + HouseStyle + OverallQual + OverallCond +
    YearBuilt + YearRemodAdd + Exterior1st + Exterior2nd + MasVnrType +
    MasVnrArea + ExterQual + Foundation + BsmtQual + BsmtExposure +
    BsmtFinType1 + BsmtFinSF1 + BsmtUnfSF + TotalBsmtSF + HeatingQC +
    CentralAir + X1stFlrSF + X2ndFlrSF + GrLivArea + BsmtFullBath +
    FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQual +
    TotRmsAbvGrd + Fireplaces + FireplaceQu + GarageType + GarageYrBlt +
    GarageFinish + GarageCars + GarageArea + PavedDrive + WoodDeckSF +
    OpenPorchSF, data = Train_set)
```

Best model selection & inferences

```
Best_lm = step(lm_fit, trace = 0)
```

Removing rows

```
Train_set3 = Train_set2[-c(722,373), ]
```

Robust Regresion

```
set.seed(111)
par(mfrow = c(1,2))
rlm_fit = rlm(Train_set$SalePrice ~ Train_set$LotFrontage + Train_set$LotArea
+ Train_set$Neighborhood + Train_set$BldgType + Train_set$HouseStyle +
Train_set$OverallQual + Train_set$OverallCond + Train_set$Exterior2nd +Train_set$BsmtQual + Train_set$Bs
+ Train_set$BsmtUnfSF + Train_set$TotalBsmtSF + Train_set$CentralAir +
Train_set$X1stFlrSF + Train_set$GrLivArea + Train_set$BsmtFullBath +
Train_set$FullBath + Train_set$HalfBath + Train_set$KitchenAbvGr +
Train_set$KitchenQual + Train_set$TotRmsAbvGrd + Train_set$Fireplaces +
Train_set$FireplaceQu + Train_set$GarageCars + Train_set$SaleCondition + Train_set$BsmtFinType1,
```

```
    data = Train_set)
plot(Best_lm, 5, main = "Leverage Plot for lm_fit")
plot(rlm_fit, 5, main = "Leverage Plot for rlm_fit")
```

lm Test Errors

```
set.seed(500)
p2 = predict(Best_lm, Test_set)
error.lm = mean((Test_set$SalePrice - p2)^2)
```

New.lm_fit

```
New.lm_fit =  lm(formula = Train_set3$SalePrice ~ MSSubClass + LotFrontage +
    Neighborhood + BldgType + HouseStyle + OverallQual + OverallCond +
    Exterior2nd + Foundation + BsmtExposure + BsmtFinType1 +
    BsmtFinSF1 + BsmtUnfSF + HeatingQC + CentralAir + X1stFlrSF +
    X2ndFlrSF + BsmtFullBath + FullBath + HalfBath + KitchenAbvGr +
    KitchenQual + TotRmsAbvGrd + FireplaceQu + GarageType + GarageYrBlt +
    GarageFinish + GarageCars + OpenPorchSF, data = Train_set3)
```

```
plot(New.lm_fit, 1)
```

```
bptest(New.lm_fit)
```

Random Forest Model

rf Test Error

```
pp =predict(rf_fit, Test_set)
error.rf = mean((Test_set$SalePrice - pp)^2)
```

My tables

```
table = matrix(NA, nrow = 3, ncol = 3)
rownames(table) = c("rlm", "lm", "rf")
colnames(table) = c("MSE", "BIAS", "VARIANCE")
table[3,]= c(error.rlm, rlm.bias, rlm.var)

table[2, ]= c(error.lm, lm.bias, lm.var)

table[1,]= c(error.rf, rf.bias, rf.var)
kable(table, caption = "Errors Tatble")
```

```
table = matrix(NA, nrow = 10, ncol = 3)
colnames(table) = c("log(SalePrice)", "PredictedValue", "Difference")
table[,1]= c(Train_set11$SalePrice[1], Train_set11$SalePrice[2], Train_set11$SalePrice[3], Train_set11$S

table[, 2]= c(prd[1], prd[2],prd[3], prd[4], prd[5], prd[6], prd[7], prd[8], prd[9], prd[10])

table[,3]= c(
round(abs(Train_set11$SalePrice[1]-prd[1]),3), round(abs(Train_set11$SalePrice[2]-prd[2]),3), round(abs
kable(table, caption = "Comparative Table")
```