EDX_Project_2

Kanak Tyagi

13 November 2019

Introduction

Imagine that you are a statistical consultant who has recently been hired by a real estate investment firm based in Ames, Iowa. They have had an intern collect and collate all of the recent house sales in Ames and have put together a large spreadsheet that contains the sale price of each house along with many of its physical features.

Your employers want you to take this data and develop a model to predict the selling price of a given home. They hope to use this information to help assess whether the asking price of a house is higher or lower than the true value of the house. If the home is undervalued, it may be a good investment for the firm.

To better assess the quality of our model, the whole data have been randomly divided into three separate data sets: a training data set, a test data set, and a validation data set. Initially we will use the training data set; the others will be used later for comparison purposes.

set working directory

```
setwd("C:/Users/Tyagi/Downloads")
getwd()
```

```
## [1] "C:/Users/Tyagi/Downloads"
```

Load the relevant packages.

```
if(!require(MASS)){install.packages('MASS')}
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 3.5.3
```

```
if(!require(dplyr)){install.packages('dplyr')}
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

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

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

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

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

```
if(!require(ggplot2)){install.packages('ggplot2')}
```

```
## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
if(!require(BAS)){install.packages('BAS')}
```

```
## Loading required package: BAS

## Warning: package 'BAS' was built under R version 3.5.3
```

```
if(!require(GGally)){install.packages('GGally')}
```

```
## Loading required package: GGally

## Warning: package 'GGally' was built under R version 3.5.3

##
## Attaching package: 'GGally'

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

```
if(!require(car)){install.packages('car')}
```

```
## Loading required package: car

## Warning: package 'car' was built under R version 3.5.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.5.2

##
## Attaching package: 'car'

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

```r
if(!require(conover.test)){install.packages('conover.test')}
```

```
## Loading required package: conover.test
```

```
## Warning: package 'conover.test' was built under R version 3.5.2
```

```r
if(!require(kableExtra)){install.packages('kableExtra')}
```

```
## Loading required package: kableExtra
```

```
## Warning: package 'kableExtra' was built under R version 3.5.3
```

```
##
## Attaching package: 'kableExtra'
```

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

```r
if(!require(gridExtra)){install.packages('gridExtra')}
```

```
## Loading required package: gridExtra
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```

```
##
## Attaching package: 'gridExtra'
```

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

```r
if(!require(ggpubr)){install.packages('ggpubr')}
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 3.5.3
```

```
## Loading required package: magrittr
```

```r
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse
```

```
## v tibble  2.0.1     v purrr   0.2.5
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.3
```

```
## Warning: package 'forcats' was built under R version 3.5.3
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_conflic
## x gridExtra::combine()     masks dplyr::combine()
## x tidyr::extract()         masks magrittr::extract()
## x dplyr::filter()          masks stats::filter()
## x kableExtra::group_rows() masks dplyr::group_rows()
## x dplyr::lag()             masks stats::lag()
## x car::recode()            masks dplyr::recode()
## x dplyr::select()          masks MASS::select()
## x purrr::set_names()       masks magrittr::set_names()
## x purrr::some()            masks car::some()
```

```
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: caret
```

```
## Warning: package 'caret' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 3.5.3
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

Load Dataset

```
load('ames_train.RData')
```

Exploratory Data Analysis

When you first get your data, it's very tempting to immediately begin fitting models and assessing how they perform. However, before you begin modeling, it's absolutely essential to explore the structure of the data and the relationships between the variables in the data set.

the dataset and its basic summary statistics

Dimensions of the data

```
dim(ames_train)
```

```
## [1] 1000   81
```

intial 7 rows with header

```
head(ames_train)
```

```
## # A tibble: 6 x 81
##      PID  area  price MS.SubClass MS.Zoning Lot.Frontage Lot.Area Street Alley
##    <int> <int>  <int>       <int> <fct>            <int>    <int> <fct>  <fct>
## 1 9.09e8   856 126000          30 RL                  NA     7890 Pave   <NA>
## 2 9.05e8  1049 139500         120 RL                  42     4235 Pave   <NA>
## 3 9.11e8  1001 124900          30 C (all)             60     6060 Pave   <NA>
## 4 5.35e8  1039 114000          70 RL                  80     8146 Pave   <NA>
## 5 5.34e8  1665 227000          60 RL                  70     8400 Pave   <NA>
## 6 9.08e8  1922 198500          85 RL                  64     7301 Pave   <NA>
## # ... with 72 more variables: Lot.Shape <fct>, Land.Contour <fct>,
## #   Utilities <fct>, Lot.Config <fct>, Land.Slope <fct>, Neighborhood <fct>,
## #   Condition.1 <fct>, Condition.2 <fct>, Bldg.Type <fct>, House.Style <fct>,
## #   Overall.Qual <int>, Overall.Cond <int>, Year.Built <int>,
## #   Year.Remod.Add <int>, Roof.Style <fct>, Roof.Matl <fct>,
## #   Exterior.1st <fct>, Exterior.2nd <fct>, Mas.Vnr.Type <fct>,
## #   Mas.Vnr.Area <int>, Exter.Qual <fct>, Exter.Cond <fct>, Foundation <fct>,
## #   Bsmt.Qual <fct>, Bsmt.Cond <fct>, Bsmt.Exposure <fct>,
## #   BsmtFin.Type.1 <fct>, BsmtFin.SF.1 <int>, BsmtFin.Type.2 <fct>,
## #   BsmtFin.SF.2 <int>, Bsmt.Unf.SF <int>, Total.Bsmt.SF <int>, Heating <fct>,
## #   Heating.QC <fct>, Central.Air <fct>, Electrical <fct>, X1st.Flr.SF <int>,
## #   X2nd.Flr.SF <int>, Low.Qual.Fin.SF <int>, Bsmt.Full.Bath <int>,
## #   Bsmt.Half.Bath <int>, Full.Bath <int>, Half.Bath <int>,
## #   Bedroom.AbvGr <int>, Kitchen.AbvGr <int>, Kitchen.Qual <fct>,
## #   TotRms.AbvGrd <int>, Functional <fct>, Fireplaces <int>,
## #   Fireplace.Qu <fct>, Garage.Type <fct>, Garage.Yr.Blt <int>,
## #   Garage.Finish <fct>, Garage.Cars <int>, Garage.Area <int>,
## #   Garage.Qual <fct>, Garage.Cond <fct>, Paved.Drive <fct>,
## #   Wood.Deck.SF <int>, Open.Porch.SF <int>, Enclosed.Porch <int>,
## #   X3Ssn.Porch <int>, Screen.Porch <int>, Pool.Area <int>, Pool.QC <fct>,
## #   Fence <fct>, Misc.Feature <fct>, Misc.Val <int>, Mo.Sold <int>,
## #   Yr.Sold <int>, Sale.Type <fct>, Sale.Condition <fct>
```

basic summary statistics

```
summary(ames_train)
```

```
##       PID                 area          price          MS.SubClass
##  Min.   :5.263e+08   Min.   : 334   Min.   : 12789   Min.   : 20.00
##  1st Qu.:5.285e+08   1st Qu.:1092   1st Qu.:129763   1st Qu.: 20.00
##  Median :5.354e+08   Median :1411   Median :159467   Median : 50.00
##  Mean   :7.059e+08   Mean   :1477   Mean   :181190   Mean   : 57.15
##  3rd Qu.:9.071e+08   3rd Qu.:1743   3rd Qu.:213000   3rd Qu.: 70.00
##  Max.   :1.007e+09   Max.   :4676   Max.   :615000   Max.   :190.00
##
##    MS.Zoning    Lot.Frontage       Lot.Area       Street     Alley
##  A (agr):  0   Min.   : 21.00   Min.   :  1470   Grvl:  3   Grvl: 33
##  C (all):  9   1st Qu.: 57.00   1st Qu.:  7314   Pave:997   Pave: 34
##  FV     : 56   Median : 69.00   Median :  9317              NA's:933
##  I (all):  1   Mean   : 69.21   Mean   : 10352
##  RH     :  7   3rd Qu.: 80.00   3rd Qu.: 11650
##  RL     :772   Max.   :313.00   Max.   :215245
##  RM     :155   NA's   :167
##  Lot.Shape Land.Contour  Utilities      Lot.Config  Land.Slope  Neighborhood
##  IR1:338   Bnk: 33      AllPub:1000   Corner :173   Gtl:962    NAmes  :155
##  IR2: 30   HLS: 38      NoSeWa:   0   CulDSac: 76   Mod: 33    CollgCr: 85
##  IR3:  3   Low: 20      NoSewr:   0   FR2    : 36   Sev:  5    Somerst: 74
##  Reg:629   Lvl:909                    FR3    :  5              OldTown: 71
##                                       Inside :710              Sawyer : 61
##                                                                Edwards: 60
##                                                                (Other):494
##   Condition.1   Condition.2   Bldg.Type    House.Style   Overall.Qual
##  Norm   :875   Norm   :988   1Fam  :823   1Story :521   Min.   : 1.000
##  Feedr  : 53   Feedr  :  6   2fmCon: 20   2Story :286   1st Qu.: 5.000
##  Artery : 23   Artery :  2   Duplex: 35   1.5Fin : 98   Median : 6.000
##  RRAn   : 14   PosN   :  2   Twnhs : 38   SLvl   : 41   Mean   : 6.095
##  PosN   : 11   PosA   :  1   TwnhsE: 84   SFoyer : 36   3rd Qu.: 7.000
##  RRAe   : 11   RRNn   :  1                2.5Unf : 10   Max.   :10.000
##  (Other): 13   (Other):  0                (Other):  8
##   Overall.Cond     Year.Built   Year.Remod.Add   Roof.Style    Roof.Matl
##  Min.   :1.000   Min.   :1872   Min.   :1950   Flat   :  9   CompShg:984
##  1st Qu.:5.000   1st Qu.:1955   1st Qu.:1966   Gable  :775   Tar&Grv: 11
##  Median :5.000   Median :1975   Median :1992   Gambrel:  8   WdShake:  2
##  Mean   :5.559   Mean   :1972   Mean   :1984   Hip    :204   WdShngl:  2
##  3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004   Mansard:  4   Metal  :  1
##  Max.   :9.000   Max.   :2010   Max.   :2010   Shed   :  0   ClyTile:  0
##                                                             (Other):  0
##   Exterior.1st  Exterior.2nd  Mas.Vnr.Type  Mas.Vnr.Area    Exter.Qual
##  VinylSd:349   VinylSd:345          :  7   Min.   :   0.0   Ex: 39
##  HdBoard:164   HdBoard:150   BrkCmn :  8   1st Qu.:   0.0   Fa: 11
##  MetalSd:147   MetalSd:148   BrkFace:317   Median :   0.0   Gd:337
##  Wd Sdng:138   Wd Sdng:130   CBlock :  0   Mean   : 104.1   TA:613
##  Plywood: 74   Plywood: 96   None   :593   3rd Qu.: 160.0
##  CemntBd: 40   CmentBd: 40   Stone  : 75   Max.   :1290.0
##  (Other): 88   (Other): 91                 NA's   :7
##  Exter.Cond  Foundation  Bsmt.Qual  Bsmt.Cond  Bsmt.Exposure BsmtFin.Type.1
##  Ex:  4     BrkTil:102       :  1       :  1       :  2      GLQ    :294
##  Fa: 19     CBlock:430   Ex  : 87   Ex  :  2   Av  :157      Unf    :279
```

```
##  Gd:116     PConc :453   Fa  : 28   Fa  : 23   Gd  : 98      ALQ    :163
##  Po:  0     Slab  : 12   Gd  :424   Gd  : 44   Mn  : 87      Rec    :107
##  TA:861     Stone :  3   Po  :  1   Po  :  1   No  :635      BLQ    : 87
##             Wood  :  0   TA  :438   TA  :908   NA's: 21      (Other): 49
##                          NA's: 21   NA's: 21                 NA's   : 21
##   BsmtFin.SF.1    BsmtFin.Type.2  BsmtFin.SF.2      Bsmt.Unf.SF
##  Min.   :   0.0   Unf    :863    Min.   :   0.00   Min.   :   0.0
##  1st Qu.:   0.0   LwQ    : 31    1st Qu.:   0.00   1st Qu.: 223.5
##  Median : 400.0   Rec    : 29    Median :   0.00   Median : 461.0
##  Mean   : 464.1   BLQ    : 24    Mean   :  48.07   Mean   : 547.0
##  3rd Qu.: 773.0   ALQ    : 20    3rd Qu.:   0.00   3rd Qu.: 783.0
##  Max.   :2260.0   (Other): 12    Max.   :1526.00   Max.   :2336.0
##  NA's   :1        NA's   : 21    NA's   :1         NA's   :1
##  Total.Bsmt.SF     Heating    Heating.QC Central.Air Electrical
##  Min.   :   0.0   Floor:  0   Ex:516     N: 55            :  0
##  1st Qu.: 797.5   GasA :988   Fa: 22     Y:945       FuseA: 54
##  Median : 998.0   GasW :  8   Gd:157                 FuseF: 12
##  Mean   :1059.2   Grav :  2   Po:  1                 FuseP:  2
##  3rd Qu.:1301.0   OthW :  1   TA:304                 Mix  :  0
##  Max.   :3138.0   Wall :  1                          SBrkr:932
##  NA's   :1
##   X1st.Flr.SF      X2nd.Flr.SF     Low.Qual.Fin.SF   Bsmt.Full.Bath
##  Min.   : 334.0   Min.   :   0.0   Min.   :   0.00   Min.   :0.0000
##  1st Qu.: 876.2   1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.:0.0000
##  Median :1080.5   Median :   0.0   Median :   0.00   Median :0.0000
##  Mean   :1157.1   Mean   : 315.2   Mean   :   4.32   Mean   :0.4474
##  3rd Qu.:1376.2   3rd Qu.: 688.2   3rd Qu.:   0.00   3rd Qu.:1.0000
##  Max.   :3138.0   Max.   :1836.0   Max.   :1064.00   Max.   :3.0000
##                                                      NA's   :1
##  Bsmt.Half.Bath      Full.Bath       Half.Bath     Bedroom.AbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.000   Median :3.000
##  Mean   :0.06106   Mean   :1.541   Mean   :0.378   Mean   :2.806
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:3.000
##  Max.   :2.00000   Max.   :4.000   Max.   :2.000   Max.   :6.000
##  NA's   :1
##  Kitchen.AbvGr   Kitchen.Qual TotRms.AbvGrd     Functional    Fireplaces
##  Min.   :0.000   Ex: 67       Min.   : 2.00   Typ    :935   Min.   :0.000
##  1st Qu.:1.000   Fa: 20       1st Qu.: 5.00   Min2   : 24   1st Qu.:0.000
##  Median :1.000   Gd:403       Median : 6.00   Min1   : 18   Median :1.000
##  Mean   :1.039   Po:  1       Mean   : 6.34   Mod    : 16   Mean   :0.597
##  3rd Qu.:1.000   TA:509       3rd Qu.: 7.00   Maj1   :  4   3rd Qu.:1.000
##  Max.   :2.000                Max.   :13.00   Maj2   :  2   Max.   :4.000
##                                               (Other):  1
##  Fireplace.Qu  Garage.Type  Garage.Yr.Blt  Garage.Finish  Garage.Cars
##  Ex  : 16     2Types : 10   Min.   :1900        :  2      Min.   :0.000
##  Fa  : 24     Attchd :610   1st Qu.:1961   Fin :247      1st Qu.:1.000
##  Gd  :232     Basment: 11   Median :1979   RFn :278      Median :2.000
##  Po  : 18     BuiltIn: 56   Mean   :1978   Unf :427      Mean   :1.767
##  TA  :219     CarPort:  1   3rd Qu.:2002   NA's: 46      3rd Qu.:2.000
##  NA's:491     Detchd :266   Max.   :2010                 Max.   :5.000
##               NA's   : 46   NA's   :48                   NA's   :1
##   Garage.Area     Garage.Qual Garage.Cond Paved.Drive  Wood.Deck.SF
```

```
##  Min.   :   0.0       :  1        :  1    N: 67       Min.   :  0.00
##  1st Qu.: 312.0   Ex  :  1    Ex  :  1    P: 29       1st Qu.:  0.00
##  Median : 480.0   Fa  : 37    Fa  : 21    Y:904       Median :  0.00
##  Mean   : 475.4   Gd  :  7    Gd  :  6                Mean   : 93.84
##  3rd Qu.: 576.0   Po  :  3    Po  :  6                3rd Qu.:168.00
##  Max.   :1390.0   TA  :904    TA  :918                Max.   :857.00
##  NA's   :1        NA's: 47    NA's: 47
##  Open.Porch.SF    Enclosed.Porch    X3Ssn.Porch       Screen.Porch
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.00
##  Median : 28.00   Median :  0.00   Median :  0.000   Median :  0.00
##  Mean   : 48.93   Mean   : 23.48   Mean   :  3.118   Mean   : 14.77
##  3rd Qu.: 74.00   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.00
##  Max.   :742.00   Max.   :432.00   Max.   :508.000   Max.   :440.00
##
##    Pool.Area       Pool.QC      Fence     Misc.Feature    Misc.Val
##  Min.   :  0.000   Ex  :  1   GdPrv: 43   Elev:  0     Min.   :    0.00
##  1st Qu.:  0.000   Fa  :  1   GdWo : 37   Gar2:  2     1st Qu.:    0.00
##  Median :  0.000   Gd  :  1   MnPrv:120   Othr:  1     Median :    0.00
##  Mean   :  1.463   TA  :  0   MnWw :  2   Shed: 25     Mean   :   45.81
##  3rd Qu.:  0.000   NA's:997   NA's :798   TenC:  1     3rd Qu.:    0.00
##  Max.   :800.000                          NA's:971     Max.   :15500.00
##
##     Mo.Sold          Yr.Sold       Sale.Type   Sale.Condition
##  Min.   : 1.000   Min.   :2006   WD     :863   Abnorml: 61
##  1st Qu.: 4.000   1st Qu.:2007   New    : 79   AdjLand:  2
##  Median : 6.000   Median :2008   COD    : 27   Alloca :  4
##  Mean   : 6.243   Mean   :2008   ConLD  :  7   Family : 17
##  3rd Qu.: 8.000   3rd Qu.:2009   ConLw  :  6   Normal :834
##  Max.   :12.000   Max.   :2010   Con    :  5   Partial: 82
##                                  (Other): 13
```

Cleaning the Data

Glimpsing the summary, some continuous variables such as Lot.Frontage have a great number of NA's that truly corresponde to missing data while the categorical variables such as Fence, Garage.Qual and Garage.Cond have NA's corresponding not to missing data but to another category such as "Not having a fence"/"Not having a garage". Before embarking on EDA thus, it makes sense to transform those NA's in a new category otherwise we may risk to incur a bias in the data and the modelling by discarding so many rows of data. I will also create a new variable as in the first peer assessment, Years.Old that shows how many years old each house is. Some of the categorical variables: MS.SubClass, Overall.Cond and Overall.Qual are also incorrectly coded as having type int so I will also convert them. Also, according to one of the past assessments, we should filter the dataset to contain only the Sale conditions that were normal, as the houses with non-normal selling conditions exhibit atypical behavior and can disproportionately influence the model. Finally, we will also had a log of the Lot.Area and of the price due to their non-linear relationships as seen in other assessments.

```
ames_train <- ames_train %>%
  mutate(Alley = if_else(is.na(Alley), 'No Alley Access', as.character(Alley)),
         Bsmt.Qual = if_else(is.na(Bsmt.Qual), 'No Basement', as.character(Bsmt.Qual)),
         Bsmt.Cond = if_else(is.na(Bsmt.Cond), 'No Basement', as.character(Bsmt.Cond)),
         Bsmt.Exposure = if_else(is.na(Bsmt.Exposure), 'No Basement', as.character(Bsmt.Cond)),
         BsmtFin.Type.1 = if_else(is.na(BsmtFin.Type.1), 'No Basement', as.character(BsmtFin.Type.1)),
         BsmtFin.Type.2 = if_else(is.na(BsmtFin.Type.2), 'No Basement', as.character(BsmtFin.Type.2)),
```

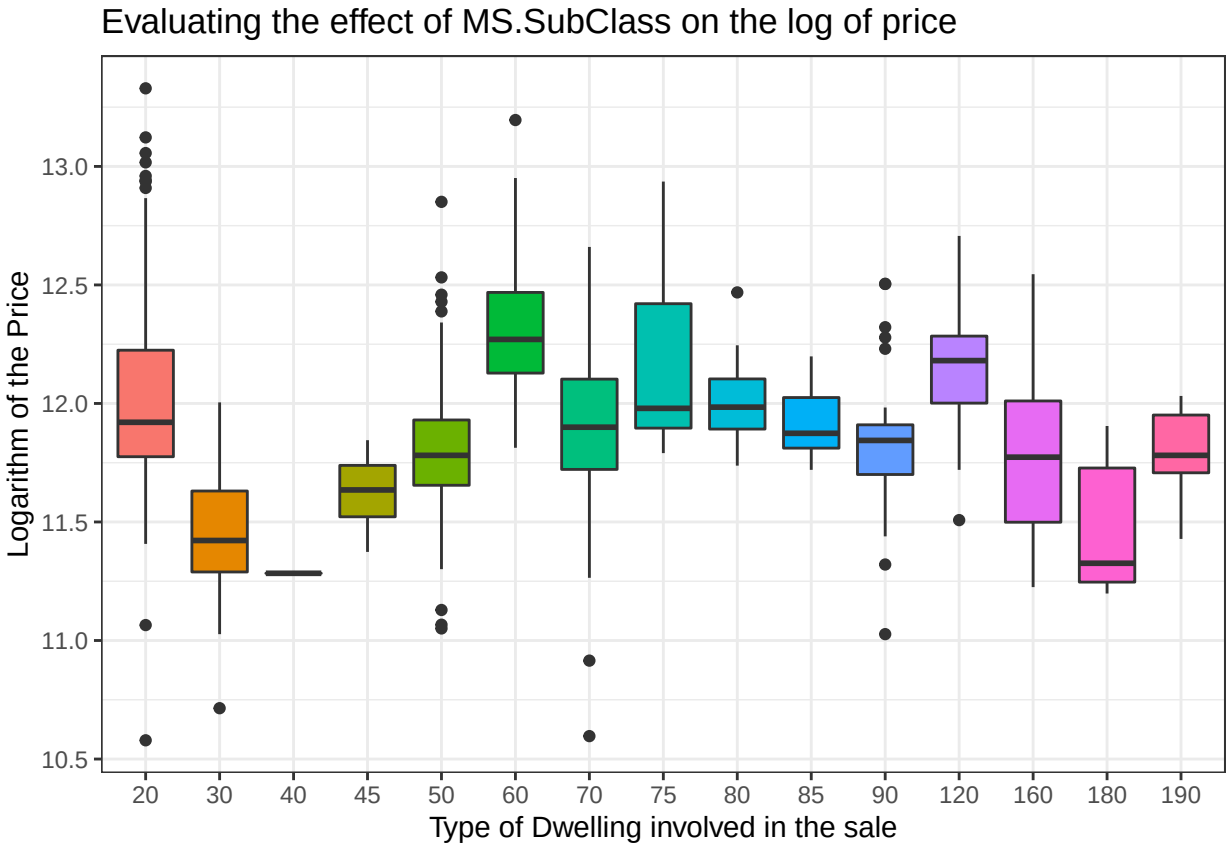
```
         Fireplace.Qu = if_else(is.na(Fireplace.Qu), 'No Fireplace', as.character(Fireplace.Qu)),
         Garage.Type = if_else(is.na(Garage.Type), 'No Garage', as.character(Garage.Type)),
         Garage.Finish = if_else(is.na(Garage.Finish), 'No Garage', as.character(Garage.Finish)),
         Garage.Qual = if_else(is.na(Garage.Qual), 'No Garage', as.character(Garage.Qual)),
         Garage.Cond = if_else(is.na(Garage.Cond), 'No Garage', as.character(Garage.Cond)),
         Pool.QC = if_else(is.na(Pool.QC), 'No Pool', as.character(Pool.QC)),
         Fence = if_else(is.na(Fence), 'No Fence', as.character(Fence)),
         Misc.Feature = if_else(is.na(Misc.Feature), 'No Misc Features', as.character(Misc.Feature)),
         Years.Old = 2018 - Year.Built,
         MS.SubClass = as.factor(MS.SubClass),
         Overall.Qual = as.factor(Overall.Qual),
         Overall.Cond = as.factor(Overall.Cond),
         log.price = log(price),
         log.Lot.Area = log(Lot.Area))

ames_train <- ames_train %>%
  filter(Sale.Condition == 'Normal')
```

Plots

As many of the variables such as Neighbourhoods, Lot.Area and the such were already explored in past assessments, for this one, I will be focusing on variables which still weren't explored and find their effect on price. As Lot.Area has been shown to be a good variable to build the model and as there is a need to use its log as well as the log of price to get a proper linear relation, all variables explored in these graphics will be against the log of price.

Plot-1 : Evaluating the effect of MS.SubClass on price

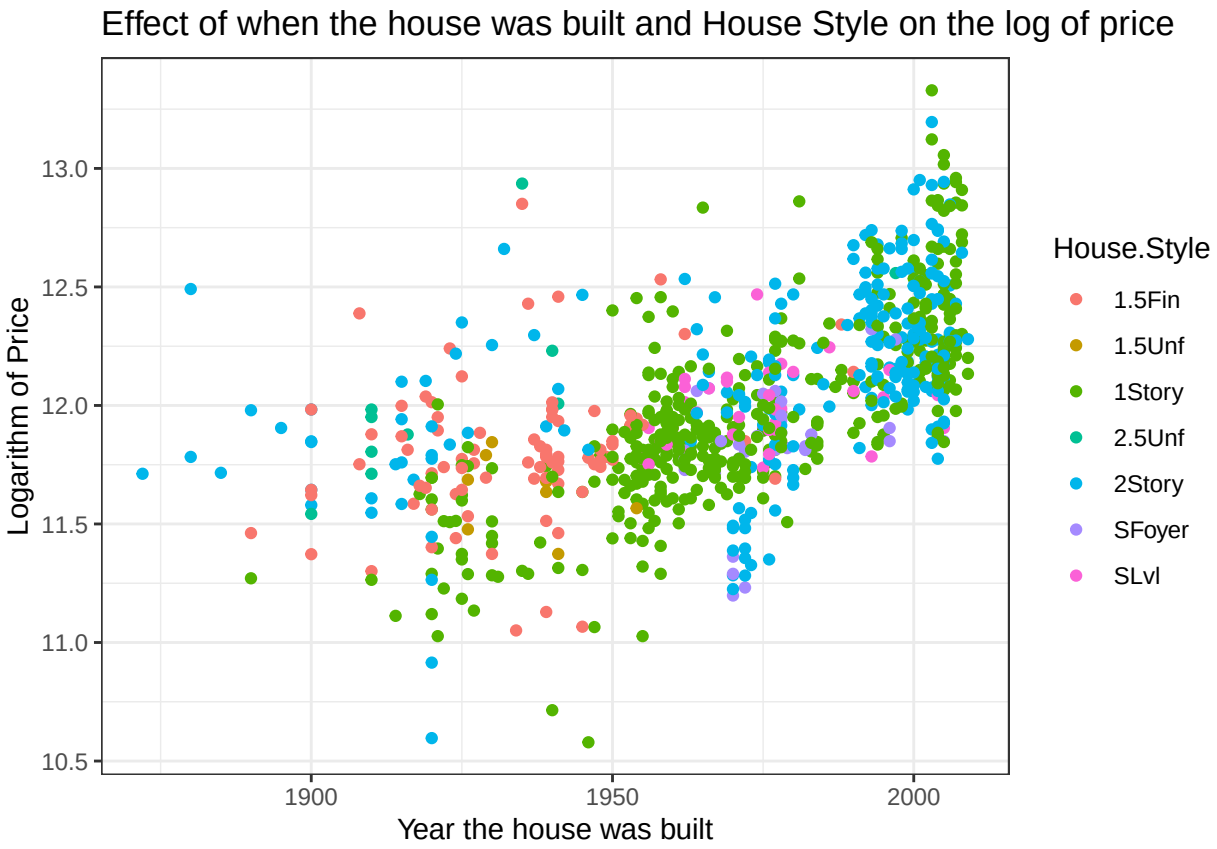
```
ggplot(ames_train, aes(x = MS.SubClass, y = log.price, fill = MS.SubClass)) + geom_boxplot() + theme_bw
```



According to this graphic, there exists some variability in the prices of the houses according to the types of Dwelling involved in the sale (only exception seems to be the 40 - 1-STORY W/FINISHED ATTIC ALL AGES dwelling). The types with highest price median are the 60 (2-STORY 1946 & NEWER) and the 120 (1-STORY PUD (Planned Unit Development) - 1946 & NEWER) subtypes.

Plot-2 : Effect of when the house was built (Year.Built) and House Style (House.Style) on log of price

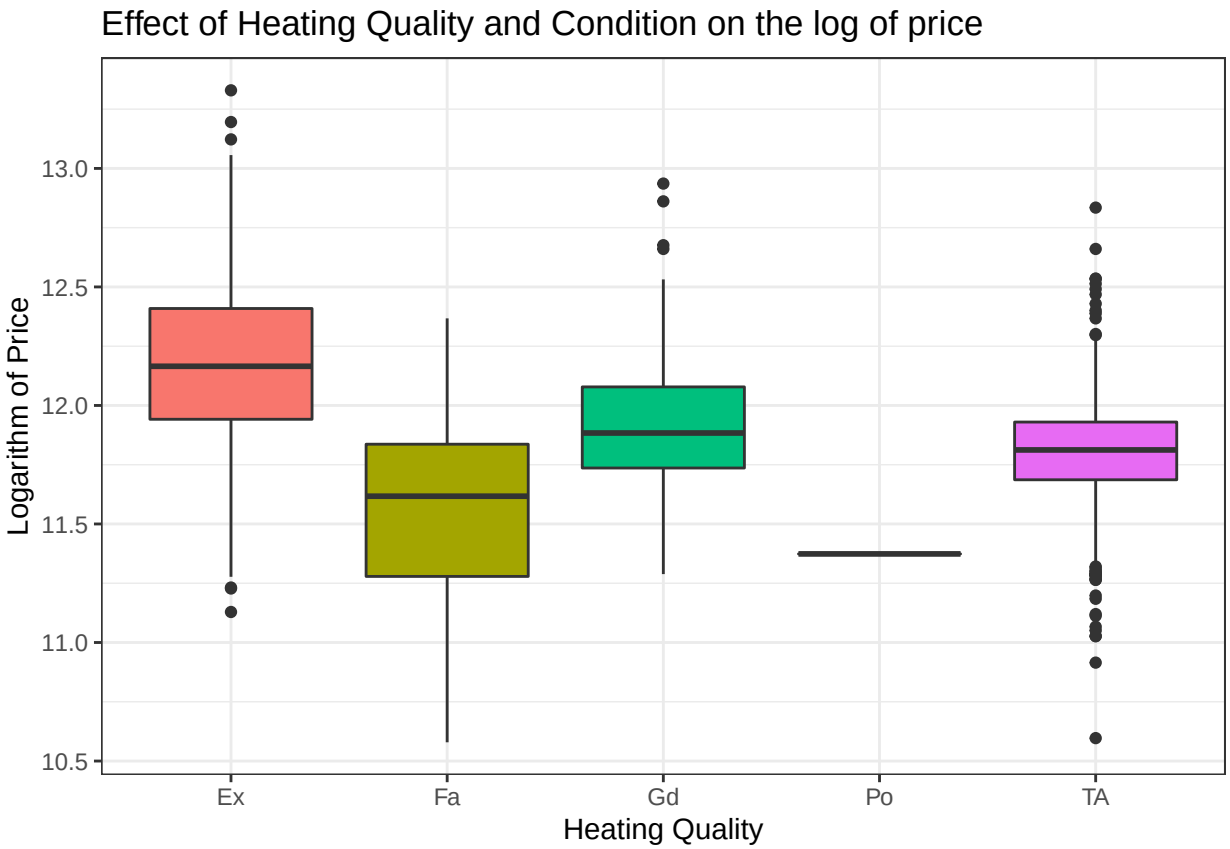
```
ggplot(ames_train, aes(x = Year.Built, y = log.price, col=House.Style)) + geom_point() + theme_bw() + la
```



As expected, more modern houses are sold for higher prices than the older ones. Also, in the later years, 2 Story houses and Two and one-half story: 2nd level unfinished (2.5Unf) seem to sell for highest prices. We also notice that 1 Story Houses and 1.5Fin (One and one-half story: 2nd level finished) and 1.5Unf(One and one-half story: 2nd level unfinished) were more common in older years than recently.

Plot-3 : Effect of Heating Quality and condition on the Log Price

```
ggplot(ames_train, aes(x = Heating.QC, y = log.price, fill=Heating.QC)) + geom_boxplot() + theme_bw() +
```



expected houses with excellent heating quality tend to be higher priced than the other houses, followed by houses with a good quality. Houses with fair quality show lower prices.

Development and assessment of an initial model, following a semi-guided process of analysis

An Initial Model

In building a model, it is often useful to start by creating a simple, intuitive initial model based on the results of the exploratory data analysis. (Note: The goal at this stage is not to identify the "best" possible model but rather to choose a reasonable and understandable starting point. Later you will expand and revise this model to create your final model.

```
initial_model <- lm(log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual + Overall.Cond + Heating.QC

summary(initial_model)
```

```
##
## Call:
## lm(formula = log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual +
##     Overall.Cond + Heating.QC + Year.Built + House.Style + Neighborhood +
##     Exterior.1st + X1st.Flr.SF, data = ames_train)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.81978 -0.06178  0.00189  0.06245  0.46222
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.343e-01  9.516e-01   0.982 0.326518
## log(Lot.Area)        1.200e-01  1.404e-02   8.550  < 2e-16 ***
## MS.SubClass30       -3.314e-02  2.828e-02  -1.172 0.241576
## MS.SubClass40       -2.447e-01  1.189e-01  -2.058 0.039956 *
## MS.SubClass45       -5.956e-03  1.317e-01  -0.045 0.963951
## MS.SubClass50       -2.844e-02  5.803e-02  -0.490 0.624273
## MS.SubClass60       -7.022e-02  5.027e-02  -1.397 0.162917
## MS.SubClass70       -6.679e-02  5.531e-02  -1.208 0.227543
## MS.SubClass75        9.773e-03  7.274e-02   0.134 0.893159
## MS.SubClass80       -1.155e-01  9.489e-02  -1.218 0.223733
## MS.SubClass85       -5.471e-02  5.595e-02  -0.978 0.328454
## MS.SubClass90       -1.457e-01  3.162e-02  -4.609 4.76e-06 ***
## MS.SubClass120      -9.528e-03  2.504e-02  -0.380 0.703712
## MS.SubClass160      -1.450e-01  5.944e-02  -2.440 0.014936 *
## MS.SubClass180      -1.466e-01  7.505e-02  -1.954 0.051110 .
## MS.SubClass190      -1.647e-02  4.340e-02  -0.379 0.704436
## Overall.Qual2        1.170e-01  1.423e-01   0.822 0.411309
## Overall.Qual3        3.034e-01  1.328e-01   2.285 0.022584 *
## Overall.Qual4        3.537e-01  1.286e-01   2.751 0.006092 **
## Overall.Qual5        3.998e-01  1.281e-01   3.121 0.001870 **
## Overall.Qual6        4.667e-01  1.286e-01   3.629 0.000303 ***
## Overall.Qual7        5.491e-01  1.293e-01   4.247 2.44e-05 ***
## Overall.Qual8        6.119e-01  1.305e-01   4.688 3.27e-06 ***
## Overall.Qual9        7.672e-01  1.333e-01   5.753 1.27e-08 ***
## Overall.Qual10       8.339e-01  1.441e-01   5.787 1.05e-08 ***
## Overall.Cond2        2.776e-01  1.544e-01   1.798 0.072646 .
## Overall.Cond3        1.183e-01  1.053e-01   1.123 0.261596
## Overall.Cond4        2.838e-01  9.957e-02   2.850 0.004491 **
## Overall.Cond5        3.582e-01  9.977e-02   3.590 0.000352 ***
## Overall.Cond6        4.020e-01  9.992e-02   4.023 6.33e-05 ***
## Overall.Cond7        4.586e-01  9.975e-02   4.597 5.02e-06 ***
## Overall.Cond8        4.901e-01  1.004e-01   4.882 1.28e-06 ***
## Overall.Cond9        5.062e-01  1.056e-01   4.795 1.96e-06 ***
## Heating.QCFa        -1.167e-01  3.120e-02  -3.741 0.000197 ***
## Heating.QCGd        -6.009e-03  1.305e-02  -0.460 0.645355
## Heating.QCPo        -5.746e-02  1.261e-01  -0.456 0.648759
## Heating.QCTA        -3.608e-02  1.217e-02  -2.964 0.003130 **
## Year.Built           4.439e-03  4.646e-04   9.554  < 2e-16 ***
## House.Style1.5Unf   -1.981e-01  1.257e-01  -1.576 0.115357
## House.Style1Story   -1.763e-01  5.706e-02  -3.089 0.002081 **
## House.Style2.5Unf    9.971e-02  6.827e-02   1.461 0.144567
## House.Style2Story    1.257e-01  5.243e-02   2.397 0.016791 *
## House.StyleSFoyer   -2.674e-02  6.722e-02  -0.398 0.690918
## House.StyleSLvl     -2.226e-02  1.026e-01  -0.217 0.828253
## NeighborhoodBlueste  1.325e-02  8.735e-02   0.152 0.879435
## NeighborhoodBrDale  -1.109e-01  7.140e-02  -1.553 0.120780
## NeighborhoodBrkSide -3.104e-02  5.910e-02  -0.525 0.599558
## NeighborhoodClearCr  3.162e-02  6.512e-02   0.485 0.627488
## NeighborhoodCollgCr -7.803e-02  5.097e-02  -1.531 0.126188
```

```
## NeighborhoodCrawfor  9.644e-02  5.814e-02   1.659 0.097599 .
## NeighborhoodEdwards -1.036e-01  5.413e-02  -1.913 0.056107 .
## NeighborhoodGilbert -7.055e-02  5.421e-02  -1.301 0.193518
## NeighborhoodGreens   1.643e-01  7.547e-02   2.177 0.029756 *
## NeighborhoodGrnHill  2.828e-01  9.360e-02   3.022 0.002599 **
## NeighborhoodIDOTRR  -1.369e-01  5.980e-02  -2.289 0.022360 *
## NeighborhoodMeadowV -2.236e-01  7.235e-02  -3.090 0.002073 **
## NeighborhoodMitchel -3.589e-02  5.352e-02  -0.671 0.502650
## NeighborhoodNAmes   -4.928e-02  5.342e-02  -0.923 0.356521
## NeighborhoodNoRidge  6.119e-02  5.465e-02   1.120 0.263227
## NeighborhoodNPkVill  4.832e-02  7.644e-02   0.632 0.527505
## NeighborhoodNridgHt  4.819e-02  5.091e-02   0.947 0.344106
## NeighborhoodNWAmes  -5.763e-02  5.511e-02  -1.046 0.296002
## NeighborhoodOldTown -8.509e-02  5.768e-02  -1.475 0.140567
## NeighborhoodSawyer  -5.194e-02  5.449e-02  -0.953 0.340835
## NeighborhoodSawyerW -1.038e-01  5.298e-02  -1.959 0.050487 .
## NeighborhoodSomerst  2.628e-02  5.111e-02   0.514 0.607312
## NeighborhoodStoneBr  4.483e-02  5.826e-02   0.770 0.441819
## NeighborhoodSWISU   -3.622e-02  6.601e-02  -0.549 0.583366
## NeighborhoodTimber  -2.924e-02  5.717e-02  -0.512 0.609150
## NeighborhoodVeenker  4.546e-02  6.370e-02   0.714 0.475627
## Exterior.1stBrkComm  2.812e-01  1.255e-01   2.241 0.025347 *
## Exterior.1stBrkFace  8.171e-02  4.947e-02   1.652 0.099010 .
## Exterior.1stCemntBd  1.336e-01  5.565e-02   2.401 0.016605 *
## Exterior.1stHdBoard  3.596e-02  4.622e-02   0.778 0.436847
## Exterior.1stImStucc  3.155e-02  1.249e-01   0.253 0.800613
## Exterior.1stMetalSd  6.570e-02  4.526e-02   1.452 0.147028
## Exterior.1stPlywood  2.172e-02  4.753e-02   0.457 0.647850
## Exterior.1stStucco   1.335e-01  5.399e-02   2.473 0.013626 *
## Exterior.1stVinylSd  5.207e-02  4.608e-02   1.130 0.258840
## Exterior.1stWd Sdng  7.285e-02  4.524e-02   1.610 0.107765
## Exterior.1stWdShing  2.790e-02  5.212e-02   0.535 0.592623
## X1st.Flr.SF          4.142e-04  1.938e-05  21.374  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1134 on 752 degrees of freedom
## Multiple R-squared:  0.9204, Adjusted R-squared:  0.9118
## F-statistic: 107.3 on 81 and 752 DF,  p-value: < 2.2e-16
```

I chose those ten variables with a combination of past assessments, the plotted graphics in this assessment and a bit of general intuition and expert knowledge. From past assessments, the Lot Area was highly correlated with price (albeit needing both to be transformed into their log scale), as well as Overall.Qual and Overall.Cond, Neighbourhood and Year.Built, which makes sense as since the prime mantra of selling houses is location, both the condition and quality of the house, as well as their neighbourhood and age will probably be correlated with selling price. Also, with the plots created in this assessement, Heating Quality and the Subclass of the Dwelling as well as the House Style seemed to be correlated with selling prices. Finally, it made intuitive sense that the exterior covering on house (Exterior.1st) and the square feet area of the first floor (didn't choose 2nd floor as well as some of the houses don't have 2nd floors) would also have an effect on selling prices.

According to the model results, all variables chosen seem to be important predictors and each one should be interpreted holding all the others constant. Some variables such as exterior covering, 1st floor square feet area, age of the house, the logarithm of the lot area and the overall quality and condition raise the selling price as their value increases, holding all the other variables constant; others such as MS.SubClass,

House Style and Neighbourhood either decrease or increase the selling price of the houses according to their categories.

The adjusted R^2 for this model is 91.2%, which means that these variables explain 91.2% of the variance in the logarithm of the selling prices of the houses in this training set which is very good.

Model Selection

From the initial model as the starting point, I will use a backwards stepwise approach using both AIC and BIC as criteria to choose the better model.

AIC

```
initial_model_AIC <- stepAIC(initial_model, direction = 'backward', trace = FALSE)
summary(initial_model_AIC)
```

```
##
## Call:
## lm(formula = log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual +
##     Overall.Cond + Heating.QC + Year.Built + House.Style + Neighborhood +
##     Exterior.1st + X1st.Flr.SF, data = ames_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81978 -0.06178  0.00189  0.06245  0.46222
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.343e-01  9.516e-01   0.982 0.326518
## log(Lot.Area)        1.200e-01  1.404e-02   8.550  < 2e-16 ***
## MS.SubClass30       -3.314e-02  2.828e-02  -1.172 0.241576
## MS.SubClass40       -2.447e-01  1.189e-01  -2.058 0.039956 *
## MS.SubClass45       -5.956e-03  1.317e-01  -0.045 0.963951
## MS.SubClass50       -2.844e-02  5.803e-02  -0.490 0.624273
## MS.SubClass60       -7.022e-02  5.027e-02  -1.397 0.162917
## MS.SubClass70       -6.679e-02  5.531e-02  -1.208 0.227543
## MS.SubClass75        9.773e-03  7.274e-02   0.134 0.893159
## MS.SubClass80       -1.155e-01  9.489e-02  -1.218 0.223733
## MS.SubClass85       -5.471e-02  5.595e-02  -0.978 0.328454
## MS.SubClass90       -1.457e-01  3.162e-02  -4.609 4.76e-06 ***
## MS.SubClass120      -9.528e-03  2.504e-02  -0.380 0.703712
## MS.SubClass160      -1.450e-01  5.944e-02  -2.440 0.014936 *
## MS.SubClass180      -1.466e-01  7.505e-02  -1.954 0.051110 .
## MS.SubClass190      -1.647e-02  4.340e-02  -0.379 0.704436
## Overall.Qual2        1.170e-01  1.423e-01   0.822 0.411309
## Overall.Qual3        3.034e-01  1.328e-01   2.285 0.022584 *
## Overall.Qual4        3.537e-01  1.286e-01   2.751 0.006092 **
## Overall.Qual5        3.998e-01  1.281e-01   3.121 0.001870 **
## Overall.Qual6        4.667e-01  1.286e-01   3.629 0.000303 ***
## Overall.Qual7        5.491e-01  1.293e-01   4.247 2.44e-05 ***
## Overall.Qual8        6.119e-01  1.305e-01   4.688 3.27e-06 ***
```

```
## Overall.Qual9        7.672e-01  1.333e-01   5.753 1.27e-08 ***
## Overall.Qual10       8.339e-01  1.441e-01   5.787 1.05e-08 ***
## Overall.Cond2        2.776e-01  1.544e-01   1.798 0.072646 .
## Overall.Cond3        1.183e-01  1.053e-01   1.123 0.261596
## Overall.Cond4        2.838e-01  9.957e-02   2.850 0.004491 **
## Overall.Cond5        3.582e-01  9.977e-02   3.590 0.000352 ***
## Overall.Cond6        4.020e-01  9.992e-02   4.023 6.33e-05 ***
## Overall.Cond7        4.586e-01  9.975e-02   4.597 5.02e-06 ***
## Overall.Cond8        4.901e-01  1.004e-01   4.882 1.28e-06 ***
## Overall.Cond9        5.062e-01  1.056e-01   4.795 1.96e-06 ***
## Heating.QCFa        -1.167e-01  3.120e-02  -3.741 0.000197 ***
## Heating.QCGd        -6.009e-03  1.305e-02  -0.460 0.645355
## Heating.QCPo        -5.746e-02  1.261e-01  -0.456 0.648759
## Heating.QCTA        -3.608e-02  1.217e-02  -2.964 0.003130 **
## Year.Built           4.439e-03  4.646e-04   9.554  < 2e-16 ***
## House.Style1.5Unf   -1.981e-01  1.257e-01  -1.576 0.115357
## House.Style1Story   -1.763e-01  5.706e-02  -3.089 0.002081 **
## House.Style2.5Unf    9.971e-02  6.827e-02   1.461 0.144567
## House.Style2Story    1.257e-01  5.243e-02   2.397 0.016791 *
## House.StyleSFoyer   -2.674e-02  6.722e-02  -0.398 0.690918
## House.StyleSLvl     -2.226e-02  1.026e-01  -0.217 0.828253
## NeighborhoodBlueste  1.325e-02  8.735e-02   0.152 0.879435
## NeighborhoodBrDale  -1.109e-01  7.140e-02  -1.553 0.120780
## NeighborhoodBrkSide -3.104e-02  5.910e-02  -0.525 0.599558
## NeighborhoodClearCr  3.162e-02  6.512e-02   0.485 0.627488
## NeighborhoodCollgCr -7.803e-02  5.097e-02  -1.531 0.126188
## NeighborhoodCrawfor  9.644e-02  5.814e-02   1.659 0.097599 .
## NeighborhoodEdwards -1.036e-01  5.413e-02  -1.913 0.056107 .
## NeighborhoodGilbert -7.055e-02  5.421e-02  -1.301 0.193518
## NeighborhoodGreens   1.643e-01  7.547e-02   2.177 0.029756 *
## NeighborhoodGrnHill  2.828e-01  9.360e-02   3.022 0.002599 **
## NeighborhoodIDOTRR  -1.369e-01  5.980e-02  -2.289 0.022360 *
## NeighborhoodMeadowV -2.236e-01  7.235e-02  -3.090 0.002073 **
## NeighborhoodMitchel -3.589e-02  5.352e-02  -0.671 0.502650
## NeighborhoodNAmes   -4.928e-02  5.342e-02  -0.923 0.356521
## NeighborhoodNoRidge  6.119e-02  5.465e-02   1.120 0.263227
## NeighborhoodNPkVill  4.832e-02  7.644e-02   0.632 0.527505
## NeighborhoodNridgHt  4.819e-02  5.091e-02   0.947 0.344106
## NeighborhoodNWAmes  -5.763e-02  5.511e-02  -1.046 0.296002
## NeighborhoodOldTown -8.509e-02  5.768e-02  -1.475 0.140567
## NeighborhoodSawyer  -5.194e-02  5.449e-02  -0.953 0.340835
## NeighborhoodSawyerW -1.038e-01  5.298e-02  -1.959 0.050487 .
## NeighborhoodSomerst  2.628e-02  5.111e-02   0.514 0.607312
## NeighborhoodStoneBr  4.483e-02  5.826e-02   0.770 0.441819
## NeighborhoodSWISU   -3.622e-02  6.601e-02  -0.549 0.583366
## NeighborhoodTimber  -2.924e-02  5.717e-02  -0.512 0.609150
## NeighborhoodVeenker  4.546e-02  6.370e-02   0.714 0.475627
## Exterior.1stBrkComm  2.812e-01  1.255e-01   2.241 0.025347 *
## Exterior.1stBrkFace  8.171e-02  4.947e-02   1.652 0.099010 .
## Exterior.1stCemntBd  1.336e-01  5.565e-02   2.401 0.016605 *
## Exterior.1stHdBoard  3.596e-02  4.622e-02   0.778 0.436847
## Exterior.1stImStucc  3.155e-02  1.249e-01   0.253 0.800613
## Exterior.1stMetalSd  6.570e-02  4.526e-02   1.452 0.147028
## Exterior.1stPlywood  2.172e-02  4.753e-02   0.457 0.647850
```

```
## Exterior.1stStucco   1.335e-01  5.399e-02   2.473 0.013626 *
## Exterior.1stVinylSd  5.207e-02  4.608e-02   1.130 0.258840
## Exterior.1stWd Sdng  7.285e-02  4.524e-02   1.610 0.107765
## Exterior.1stWdShing  2.790e-02  5.212e-02   0.535 0.592623
## X1st.Flr.SF          4.142e-04  1.938e-05  21.374  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1134 on 752 degrees of freedom
## Multiple R-squared:  0.9204, Adjusted R-squared:  0.9118
## F-statistic: 107.3 on 81 and 752 DF,  p-value: < 2.2e-16
```

BIC

```
initial_model_BIC <- stepAIC(initial_model, direction='backward', k = log(nrow(ames_train)), trace = FA
summary(initial_model_BIC)
```

```
##
## Call:
## lm(formula = log(price) ~ log(Lot.Area) + Overall.Qual + Overall.Cond +
##     Heating.QC + Year.Built + House.Style + X1st.Flr.SF, data = ames_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78931 -0.07107  0.00531  0.07199  0.46587
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.749e+00  5.313e-01   3.292 0.001038 **
## log(Lot.Area)     1.418e-01  9.636e-03  14.718  < 2e-16 ***
## Overall.Qual2     3.798e-02  1.512e-01   0.251 0.801649
## Overall.Qual3     2.129e-01  1.413e-01   1.507 0.132157
## Overall.Qual4     2.710e-01  1.375e-01   1.971 0.049058 *
## Overall.Qual5     3.607e-01  1.370e-01   2.633 0.008622 **
## Overall.Qual6     4.449e-01  1.377e-01   3.231 0.001285 **
## Overall.Qual7     5.461e-01  1.383e-01   3.950 8.51e-05 ***
## Overall.Qual8     6.842e-01  1.392e-01   4.915 1.07e-06 ***
## Overall.Qual9     8.413e-01  1.420e-01   5.924 4.67e-09 ***
## Overall.Qual10    9.166e-01  1.537e-01   5.962 3.72e-09 ***
## Overall.Cond2     4.316e-01  1.620e-01   2.665 0.007854 **
## Overall.Cond3     2.035e-01  1.072e-01   1.899 0.057975 .
## Overall.Cond4     3.599e-01  1.004e-01   3.586 0.000356 ***
## Overall.Cond5     4.583e-01  1.001e-01   4.579 5.40e-06 ***
## Overall.Cond6     4.965e-01  9.988e-02   4.971 8.15e-07 ***
## Overall.Cond7     5.608e-01  1.001e-01   5.600 2.95e-08 ***
## Overall.Cond8     5.803e-01  1.009e-01   5.749 1.27e-08 ***
## Overall.Cond9     5.863e-01  1.076e-01   5.449 6.72e-08 ***
## Heating.QCFa     -1.242e-01  3.346e-02  -3.713 0.000219 ***
## Heating.QCGd     -7.129e-03  1.324e-02  -0.538 0.590440
## Heating.QCPo     -1.292e-01  1.351e-01  -0.956 0.339301
## Heating.QCTA     -6.051e-02  1.181e-02  -5.121 3.80e-07 ***
```

```
## Year.Built         3.880e-03  2.535e-04  15.308  < 2e-16 ***
## House.Style1.5Unf -1.699e-01  4.995e-02  -3.402 0.000702 ***
## House.Style1Story -1.443e-01  1.752e-02  -8.236 7.18e-16 ***
## House.Style2.5Unf  9.087e-02  4.434e-02   2.049 0.040742 *
## House.Style2Story  6.890e-02  1.818e-02   3.790 0.000162 ***
## House.StyleSFoyer -7.202e-02  2.841e-02  -2.535 0.011428 *
## House.StyleSLvl   -1.106e-01  2.619e-02  -4.224 2.68e-05 ***
## X1st.Flr.SF        4.072e-04  1.883e-05  21.630  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1257 on 803 degrees of freedom
## Multiple R-squared:  0.8955, Adjusted R-squared:  0.8916
## F-statistic: 229.4 on 30 and 803 DF,  p-value: < 2.2e-16
```

According to the results, both approaches do not arrive at the same model. The model using BIC as criteria arrives to a model with a lower adjusted R^2 but with lesser predictor variables resulting in a more parsimonious model which is excellent for interpretation and fits with BIC objective which is to allow consistent estimation of the underlying data generating process.

The model using AIC as criteria arrives to a model with higher adjusted R^2 and using more predictor variables, which is the initial model without any changes. This fits with AIC objective which is better for prediction as it is asymptotically equivalent to cross-validation, at the cost of a more parsimonious explanation.

The differences between the criteria explain why they disagree. As the main objective of this assessment is predicting the selling prices of houses, I will stick with the AIC model as it fulfill that objective better.

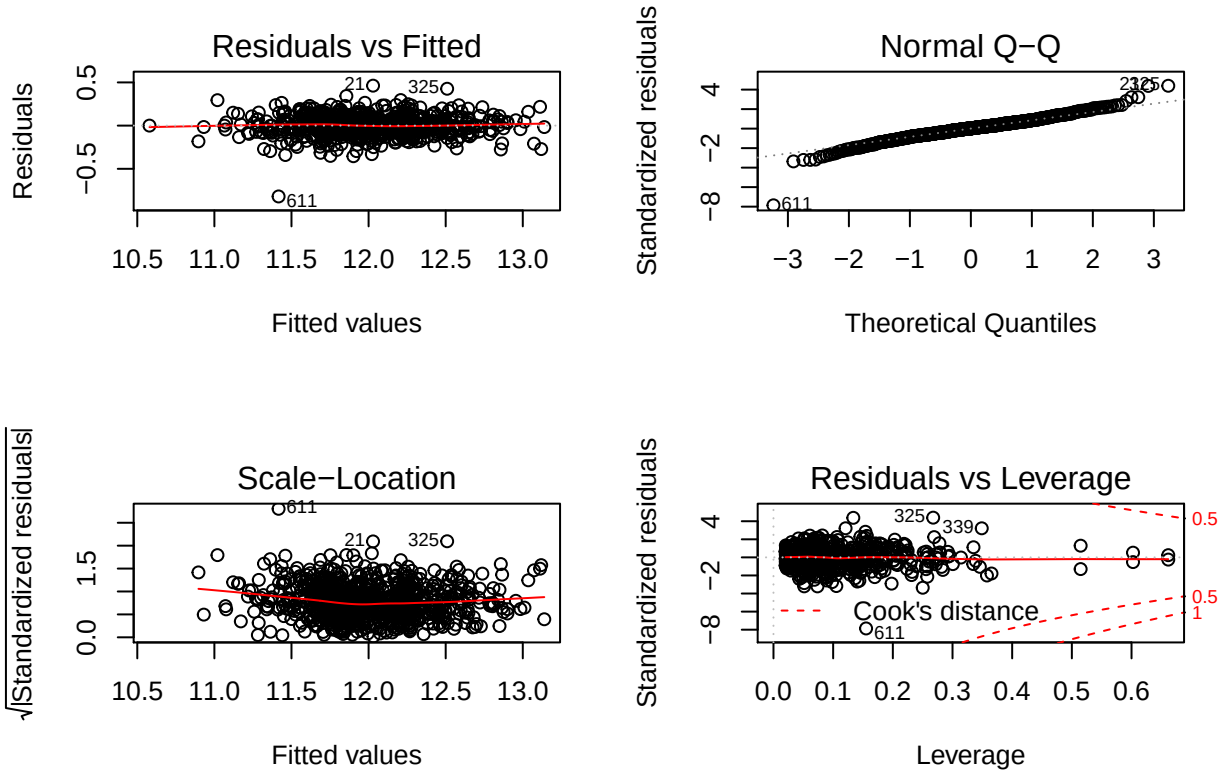
Initial Model Residuals

One way to assess the performance of a model is to examine the model's residuals. Here, we are creating a residual plot for your preferred model from above and using it to assess whether your model appears to fit the data well.

```
par(mfrow=c(2,2))
plot(initial_model_AIC)
```

```
## Warning: not plotting observations with leverage one:
##   114, 127, 151, 176, 405, 655, 763
```

```
## Warning: not plotting observations with leverage one:
##   114, 127, 151, 176, 405, 655, 763
```



By examining the residual plots, there appears to be no major problem with the residuals of the model apart from some high leverage outliers (rows 325, 339 and 611). This is expected when using categorical variables for building a regression model as it's hard for a subject to be a serious outlier in terms of a predictor if that predictor only has a few possible levels. I am expecting no serious implications in my model inference or predictions with such residual plots. The heavier tails of the distribution could be a problem, however the sample is so big that it won't be a problem due to the Central Limit Theorem. Even so, it would only impact our estimation/inference capacity. As the main goal here is prediction, this does not seem to be a problem.

Initial Model RMSE

I have calculated it directly based on the model output.

Extract Predictions

```
predictions_initial <- exp(predict(initial_model_AIC, ames_train))
```

Extract Residuals

```
residuals_initial <- ames_train$price - predictions_initial
```

Calculate RMSE

```
rmse_initial <- sqrt(mean(residuals_initial^2))
rmse_initial
```

```
## [1] 20376.42
```

The RMSE (root mean square error) for this initial model is 20376.42 dollars.

Overfitting

The process of building a model generally involves starting with an initial model (as I have done above), identifying its shortcomings, and adapting the model accordingly. This process may be repeated several times until the model fits the data reasonably well. However, the model may do well on training data but perform poorly out-of-sample (meaning, on a dataset other than the original training data) because the model is overly-tuned to specifically fit the training data. This is called "overfitting." To determine whether overfitting is occurring on a model, I have compared the performance of a model on both in-sample and out-of-sample data sets. To look at performance of my initial model on out-of-sample data, I will use the data set ames_test.

For testing the performance of the initial model against the test data and comparing it to the performance of the initial model on the training data, I will calculate the RMSE for the model predictions using the test data and then compare it to the RMSE obtained in the previous question. Due to the conversions made in the original dataset, we also need to convert the same categories in the test data.

```
load("ames_test.Rdata")

ames_test <- ames_test %>%
  mutate(Alley = if_else(is.na(Alley), 'No Alley Access', as.character(Alley)),
         Bsmt.Qual = if_else(is.na(Bsmt.Qual), 'No Basement', as.character(Bsmt.Qual)),
         Bsmt.Cond = if_else(is.na(Bsmt.Cond), 'No Basement', as.character(Bsmt.Cond)),
         Bsmt.Exposure = if_else(is.na(Bsmt.Exposure), 'No Basement', as.character(Bsmt.Cond)),
         BsmtFin.Type.1 = if_else(is.na(BsmtFin.Type.1), 'No Basement', as.character(BsmtFin.Type.1)),
         BsmtFin.Type.2 = if_else(is.na(BsmtFin.Type.2), 'No Basement', as.character(BsmtFin.Type.2)),
         Fireplace.Qu = if_else(is.na(Fireplace.Qu), 'No Fireplace', as.character(Fireplace.Qu)),
         Garage.Type = if_else(is.na(Garage.Type), 'No Garage', as.character(Garage.Type)),
         Garage.Finish = if_else(is.na(Garage.Finish), 'No Garage', as.character(Garage.Finish)),
         Garage.Qual = if_else(is.na(Garage.Qual), 'No Garage', as.character(Garage.Qual)),
         Garage.Cond = if_else(is.na(Garage.Cond), 'No Garage', as.character(Garage.Cond)),
         Pool.QC = if_else(is.na(Pool.QC), 'No Pool', as.character(Pool.QC)),
         Fence = if_else(is.na(Fence), 'No Fence', as.character(Fence)),
         Misc.Feature = if_else(is.na(Misc.Feature), 'No Misc Features', as.character(Misc.Feature)),
         Years.Old = 2018 - Year.Built,
         MS.SubClass = as.factor(MS.SubClass),
         Overall.Qual = as.factor(Overall.Qual),
         Overall.Cond = as.factor(Overall.Cond),
         log.price = log(price),
         log.Lot.Area = log(Lot.Area))
```

There is one problem to solve yet, as I used the House.Style variable to build the model. The test data has 2 houses with the House.Style 2.5Fin where none existed in the training data. Calculating the predictions in the new test data will thus result in an error. The only solution I found was to remove the 2 houses with this problem from the test dataset. The same problem happens with the predictor Neighborhood as the test data has rows with the level Landmark and the predictor Exterior.1st. The same solution was used to eliminate the rows with this problem.

```
ames_test <- ames_test %>%
  filter(House.Style != '2.5Fin') %>%
  filter(Neighborhood != 'Landmrk') %>%
  filter(Exterior.1st != 'AsphShn')
```

```
predictions_test <- exp(predict(initial_model_AIC,ames_test))
residuals_test <- ames_test$price - predictions_test
```

```
rmse_test <- sqrt(mean(residuals_test^2))
rmse_test
```

```
## [1] 23613.98
```

As the RMSE rises with the predictions in the test data, we can conclude,as expected, that this model fits the training data better than out of sample data. A way of simplifying the model as suggested would be perhaps to use the model built with the BIC instead of the AIC.

Development of a Final Model

Now that I have developed an initial model to use as a baseline, I am creating a final model with at most 20 variables to predict housing prices in Ames, IA, selecting from the full array of variables in the dataset and using any of the tools that I introduced in this specialization.

Final Model

As the initial model already showed a good predictive power with a low RMSE in the test data, I'm not going to do many alterations on it other than try to improve its accuracy with a few more variables to differentiate lower quality houses from higher quality houses. Thus, I will try to add the Year.Remod.Add, Garage.Area, Bsmt.Qual and Pool.QC variables to the initial model.

```
final_model <- lm(log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual +
                    Overall.Cond + Heating.QC + Year.Built + House.Style + Neighborhood +
                    Exterior.1st + X1st.Flr.SF + Year.Remod.Add + Bsmt.Qual + Garage.Area + Pool.QC, da
```

Now to compare if the AIC criteria also chooses the same model using backward step selection:

```
final_model_AIC <- step(final_model, direction='backward', trace = FALSE)
final_model_AIC
```

```
##
## Call:
## lm(formula = log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual +
##     Overall.Cond + Heating.QC + Year.Built + House.Style + Neighborhood +
##     Exterior.1st + X1st.Flr.SF + Year.Remod.Add + Bsmt.Qual +
##     Garage.Area + Pool.QC, data = ames_train)
##
## Coefficients:
##          (Intercept)         log(Lot.Area)         MS.SubClass30
##            0.7703524             0.1034355            -0.0494407
##        MS.SubClass40         MS.SubClass45         MS.SubClass50
##           -0.2519347            -0.0700470            -0.0330923
##        MS.SubClass60         MS.SubClass70         MS.SubClass75
##           -0.0344992            -0.0466195             0.0261264
##        MS.SubClass80         MS.SubClass85         MS.SubClass90
##           -0.0479479             0.0043472            -0.0889491
##       MS.SubClass120        MS.SubClass160        MS.SubClass180
##           -0.0144470            -0.1211142            -0.1144631
##       MS.SubClass190         Overall.Qual2         Overall.Qual3
##           -0.0248707             0.0658041             0.1388422
##        Overall.Qual4         Overall.Qual5         Overall.Qual6
##            0.2006287             0.2442514             0.3039198
##        Overall.Qual7         Overall.Qual8         Overall.Qual9
##            0.3833148             0.4355305             0.5501241
##       Overall.Qual10         Overall.Cond2         Overall.Cond3
##            0.5283166             0.4153341             0.1891467
##        Overall.Cond4         Overall.Cond5         Overall.Cond6
##            0.3173963             0.3919867             0.4333619
##        Overall.Cond7         Overall.Cond8         Overall.Cond9
##            0.4814563             0.4982292             0.5087935
##         Heating.QCFa          Heating.QCGd          Heating.QCPo
##           -0.0922195            -0.0027556            -0.0524307
##         Heating.QCTA            Year.Built     House.Style1.5Unf
##           -0.0271476             0.0035633            -0.1415186
##    House.Style1Story     House.Style2.5Unf     House.Style2Story
##           -0.1676994             0.0772414             0.0828086
##    House.StyleSFoyer       House.StyleSLvl   NeighborhoodBlueste
##           -0.0797133            -0.0897766            -0.0288367
##   NeighborhoodBrDale   NeighborhoodBrkSide   NeighborhoodClearCr
##           -0.1236434            -0.0373261             0.0317162
##  NeighborhoodCollgCr   NeighborhoodCrawfor   NeighborhoodEdwards
##           -0.0882559             0.1065171            -0.0876648
##  NeighborhoodGilbert    NeighborhoodGreens   NeighborhoodGrnHill
##           -0.0728531             0.1445898             0.3892030
##   NeighborhoodIDOTRR   NeighborhoodMeadowV   NeighborhoodMitchel
##           -0.1341729            -0.2199825            -0.0598681
##    NeighborhoodNAmes   NeighborhoodNoRidge   NeighborhoodNPkVill
##           -0.0527558             0.0454349             0.0256766
##  NeighborhoodNridgHt    NeighborhoodNWAmes   NeighborhoodOldTown
##            0.0249297            -0.0636705            -0.0972129
##   NeighborhoodSawyer   NeighborhoodSawyerW   NeighborhoodSomerst
##           -0.0573920            -0.1065980             0.0037567
##  NeighborhoodStoneBr     NeighborhoodSWISU    NeighborhoodTimber
##            0.0472626            -0.0084792            -0.0446161
```

```
##  NeighborhoodVeenker   Exterior.1stBrkComm   Exterior.1stBrkFace
##            0.0441231             0.2707707             0.1010640
##  Exterior.1stCemntBd   Exterior.1stHdBoard   Exterior.1stImStucc
##            0.1149466             0.0258691             0.0155796
##  Exterior.1stMetalSd   Exterior.1stPlywood    Exterior.1stStucco
##            0.0505972             0.0129309             0.1116177
##  Exterior.1stVinylSd   Exterior.1stWd Sdng   Exterior.1stWdShing
##            0.0277315             0.0574819             0.0162194
##          X1st.Flr.SF        Year.Remod.Add          Bsmt.QualFa
##            0.0003683             0.0012895           -0.0829082
##          Bsmt.QualGd  Bsmt.QualNo Basement          Bsmt.QualPo
##           -0.0635837            -0.2422905           -0.1951916
##          Bsmt.QualTA           Garage.Area            Pool.QCFa
##           -0.0693317             0.0001567           -0.2287028
##            Pool.QCGd        Pool.QCNo Pool
##           -0.3064662            -0.3213074
```

The AIC criteria also chooses the same model with the 14 selected predictor variables chosen.

Now to compare this model with the initial one and check if it truly is better than the initial.

```
anova(initial_model_AIC, final_model_AIC)
```

```
## Analysis of Variance Table
##
## Model 1: log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual + Overall.Cond +
##     Heating.QC + Year.Built + House.Style + Neighborhood + Exterior.1st +
##     X1st.Flr.SF
## Model 2: log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual + Overall.Cond +
##     Heating.QC + Year.Built + House.Style + Neighborhood + Exterior.1st +
##     X1st.Flr.SF + Year.Remod.Add + Bsmt.Qual + Garage.Area +
##     Pool.QC
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    752 9.676
## 2    742 8.544 10     1.132 9.8305 1.483e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results show that the new variables increase the predictive power of this new model compared to the older one. The final model summary thus is:

```r
summary(final_model_AIC)
```

```
##
## Call:
## lm(formula = log(price) ~ log(Lot.Area) + MS.SubClass + Overall.Qual +
##     Overall.Cond + Heating.QC + Year.Built + House.Style + Neighborhood +
##     Exterior.1st + X1st.Flr.SF + Year.Remod.Add + Bsmt.Qual +
##     Garage.Area + Pool.QC, data = ames_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84705 -0.05676 -0.00044  0.06029  0.42586
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.704e-01  1.054e+00   0.731 0.465216
## log(Lot.Area)         1.034e-01  1.371e-02   7.543 1.35e-13 ***
## MS.SubClass30        -4.944e-02  2.708e-02  -1.826 0.068277 .
## MS.SubClass40        -2.519e-01  1.125e-01  -2.239 0.025465 *
## MS.SubClass45        -7.005e-02  1.250e-01  -0.560 0.575537
## MS.SubClass50        -3.309e-02  5.546e-02  -0.597 0.550930
## MS.SubClass60        -3.450e-02  4.820e-02  -0.716 0.474347
## MS.SubClass70        -4.662e-02  5.283e-02  -0.882 0.377854
## MS.SubClass75         2.613e-02  6.905e-02   0.378 0.705267
## MS.SubClass80        -4.795e-02  9.017e-02  -0.532 0.595065
## MS.SubClass85         4.347e-03  5.353e-02   0.081 0.935299
## MS.SubClass90        -8.895e-02  3.105e-02  -2.865 0.004291 **
## MS.SubClass120       -1.445e-02  2.416e-02  -0.598 0.549993
## MS.SubClass160       -1.211e-01  5.705e-02  -2.123 0.034088 *
## MS.SubClass180       -1.145e-01  7.164e-02  -1.598 0.110543
## MS.SubClass190       -2.487e-02  4.125e-02  -0.603 0.546770
## Overall.Qual2         6.580e-02  1.349e-01   0.488 0.625941
## Overall.Qual3         1.388e-01  1.277e-01   1.087 0.277179
## Overall.Qual4         2.006e-01  1.238e-01   1.620 0.105583
## Overall.Qual5         2.443e-01  1.234e-01   1.979 0.048140 *
## Overall.Qual6         3.039e-01  1.240e-01   2.451 0.014463 *
## Overall.Qual7         3.833e-01  1.246e-01   3.076 0.002175 **
## Overall.Qual8         4.355e-01  1.258e-01   3.462 0.000568 ***
## Overall.Qual9         5.501e-01  1.296e-01   4.246 2.46e-05 ***
## Overall.Qual10        5.283e-01  1.426e-01   3.706 0.000226 ***
## Overall.Cond2         4.153e-01  1.483e-01   2.801 0.005223 **
## Overall.Cond3         1.891e-01  1.018e-01   1.858 0.063619 .
## Overall.Cond4         3.174e-01  9.555e-02   3.322 0.000938 ***
## Overall.Cond5         3.920e-01  9.581e-02   4.091 4.76e-05 ***
## Overall.Cond6         4.334e-01  9.599e-02   4.515 7.38e-06 ***
## Overall.Cond7         4.815e-01  9.616e-02   5.007 6.92e-07 ***
## Overall.Cond8         4.982e-01  9.656e-02   5.160 3.17e-07 ***
## Overall.Cond9         5.088e-01  1.019e-01   4.993 7.41e-07 ***
## Heating.QCFa         -9.222e-02  2.977e-02  -3.097 0.002027 **
## Heating.QCGd         -2.756e-03  1.245e-02  -0.221 0.824903
## Heating.QCPo         -5.243e-02  1.194e-01  -0.439 0.660786
## Heating.QCTA         -2.715e-02  1.176e-02  -2.309 0.021241 *
## Year.Built            3.563e-03  4.663e-04   7.641 6.66e-14 ***
```

```
## House.Style1.5Unf    -1.415e-01  1.195e-01  -1.184 0.236778
## House.Style1Story     -1.677e-01  5.466e-02  -3.068 0.002231 **
## House.Style2.5Unf      7.724e-02  6.486e-02   1.191 0.234052
## House.Style2Story      8.281e-02  5.027e-02   1.647 0.099893 .
## House.StyleSFoyer     -7.971e-02  6.443e-02  -1.237 0.216373
## House.StyleSLvl       -8.978e-02  9.783e-02  -0.918 0.359076
## NeighborhoodBlueste   -2.884e-02  8.327e-02  -0.346 0.729226
## NeighborhoodBrDale    -1.236e-01  6.804e-02  -1.817 0.069571 .
## NeighborhoodBrkSide   -3.733e-02  5.608e-02  -0.666 0.505848
## NeighborhoodClearCr    3.172e-02  6.183e-02   0.513 0.608121
## NeighborhoodCollgCr   -8.826e-02  4.831e-02  -1.827 0.068147 .
## NeighborhoodCrawfor    1.065e-01  5.520e-02   1.930 0.054021 .
## NeighborhoodEdwards   -8.766e-02  5.144e-02  -1.704 0.088775 .
## NeighborhoodGilbert   -7.285e-02  5.166e-02  -1.410 0.158874
## NeighborhoodGreens     1.446e-01  7.158e-02   2.020 0.043733 *
## NeighborhoodGrnHill    3.892e-01  8.986e-02   4.331 1.69e-05 ***
## NeighborhoodIDOTRR    -1.342e-01  5.686e-02  -2.360 0.018543 *
## NeighborhoodMeadowV   -2.200e-01  6.880e-02  -3.197 0.001446 **
## NeighborhoodMitchel   -5.987e-02  5.086e-02  -1.177 0.239531
## NeighborhoodNAmes     -5.276e-02  5.073e-02  -1.040 0.298744
## NeighborhoodNoRidge    4.543e-02  5.187e-02   0.876 0.381306
## NeighborhoodNPkVill    2.568e-02  7.260e-02   0.354 0.723669
## NeighborhoodNridgHt    2.493e-02  4.834e-02   0.516 0.606198
## NeighborhoodNWAmes    -6.367e-02  5.234e-02  -1.216 0.224186
## NeighborhoodOldTown   -9.721e-02  5.473e-02  -1.776 0.076094 .
## NeighborhoodSawyer    -5.739e-02  5.173e-02  -1.109 0.267582
## NeighborhoodSawyerW   -1.066e-01  5.030e-02  -2.119 0.034403 *
## NeighborhoodSomerst    3.757e-03  4.856e-02   0.077 0.938358
## NeighborhoodStoneBr    4.726e-02  5.521e-02   0.856 0.392267
## NeighborhoodSWISU     -8.479e-03  6.280e-02  -0.135 0.892628
## NeighborhoodTimber    -4.462e-02  5.467e-02  -0.816 0.414746
## NeighborhoodVeenker    4.412e-02  6.177e-02   0.714 0.475275
## Exterior.1stBrkComm    2.708e-01  1.192e-01   2.271 0.023423 *
## Exterior.1stBrkFace    1.011e-01  4.731e-02   2.136 0.032997 *
## Exterior.1stCemntBd    1.149e-01  5.305e-02   2.167 0.030564 *
## Exterior.1stHdBoard    2.587e-02  4.423e-02   0.585 0.558820
## Exterior.1stImStucc    1.558e-02  1.185e-01   0.131 0.895421
## Exterior.1stMetalSd    5.060e-02  4.315e-02   1.173 0.241357
## Exterior.1stPlywood    1.293e-02  4.554e-02   0.284 0.776540
## Exterior.1stStucco     1.116e-01  5.127e-02   2.177 0.029787 *
## Exterior.1stVinylSd    2.773e-02  4.430e-02   0.626 0.531480
## Exterior.1stWd Sdng    5.748e-02  4.326e-02   1.329 0.184369
## Exterior.1stWdShing    1.622e-02  5.008e-02   0.324 0.746113
## X1st.Flr.SF            3.683e-04  1.935e-05  19.032  < 2e-16 ***
## Year.Remod.Add         1.290e-03  3.140e-04   4.107 4.45e-05 ***
## Bsmt.QualFa           -8.291e-02  3.817e-02  -2.172 0.030173 *
## Bsmt.QualGd           -6.358e-02  2.253e-02  -2.823 0.004890 **
## Bsmt.QualNo Basement -2.423e-01  3.841e-02  -6.308 4.86e-10 ***
## Bsmt.QualPo           -1.952e-01  1.175e-01  -1.662 0.096974 .
## Bsmt.QualTA           -6.933e-02  2.659e-02  -2.607 0.009307 **
## Garage.Area            1.567e-04  2.648e-05   5.919 4.96e-09 ***
## Pool.QCFa             -2.287e-01  1.673e-01  -1.367 0.171966
## Pool.QCGd             -3.065e-01  1.745e-01  -1.756 0.079483 .
## Pool.QCNo Pool        -3.213e-01  1.257e-01  -2.557 0.010751 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1073 on 742 degrees of freedom
## Multiple R-squared:  0.9297, Adjusted R-squared:  0.921
## F-statistic: 107.8 on 91 and 742 DF,  p-value: < 2.2e-16
```

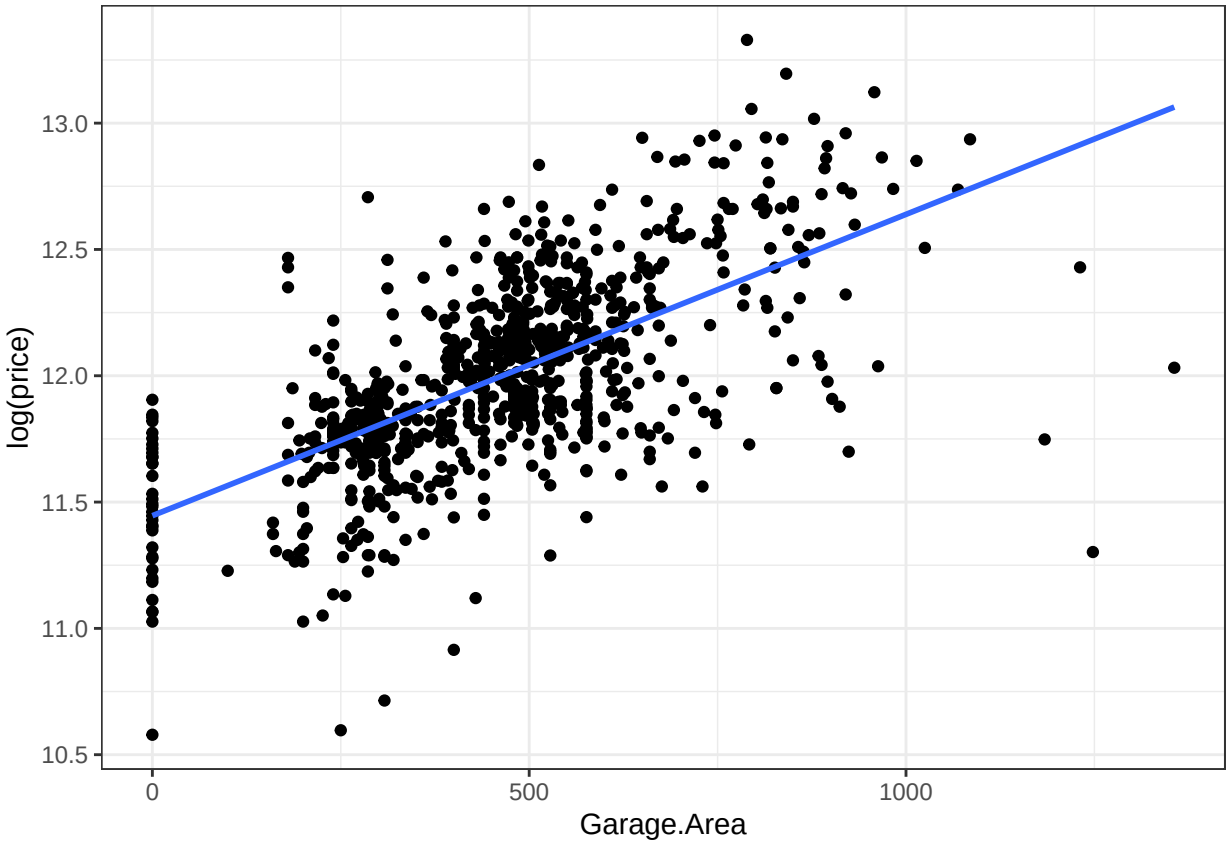
All the new variables chosen to add to the previous model are shown to be significant and the adjusted R^2 of the model also rises to 92.1%, explaning 92.1% of the variance of the log of the selling price of the houses.

Transformation

Lot.Area needs to be log transformed as well as price due to their skeweness. The same doesn't happen with any other other variables used in the model so there is no need for transformations.

To prove this is right, plotting the other continuous predictor variables used in the model.

```
ggplot(ames_train, aes(x = Garage.Area, y = log(price))) + geom_point() + theme_bw() + geom_smooth(metho
```



Model Testing

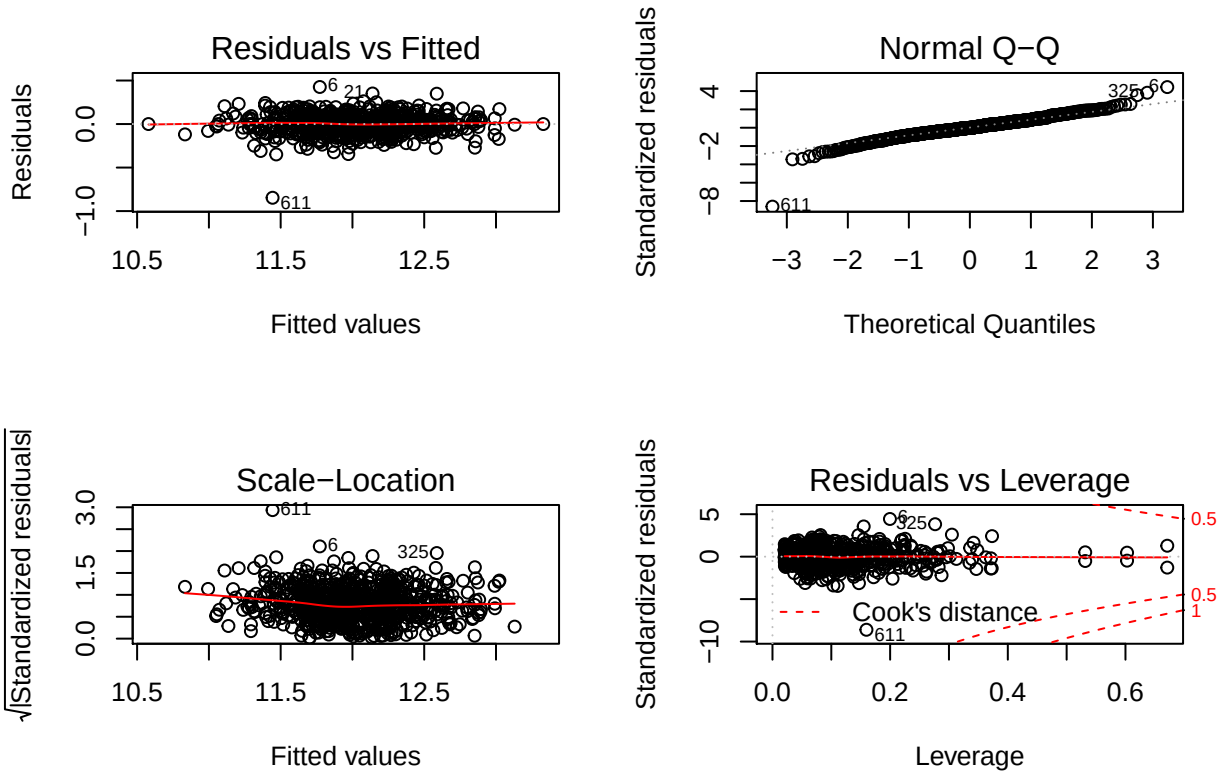
I was very happy with the results from the initial model and its low RSME on the test data, even though it was a bit higher than the RMSE in the training data as expected. The final model resulted only from a try to add some more variables in order to improve predictive power.

Final Model Assessment

For final model, creating and briefly interpreting an informative plot of the residuals.

```
par(mfrow=c(2,2))
plot(final_model_AIC)
```

```
## Warning: not plotting observations with leverage one:
##   53, 105, 114, 127, 151, 176, 405, 406, 494, 655, 763

## Warning: not plotting observations with leverage one:
##   53, 105, 114, 127, 151, 176, 405, 406, 494, 655, 763
```



Once again, other than the heavy tails on the normal distribution of the residuals, there does not appear to exist any major assumption violation in the residuals plots. This may bring some problems in the inference estimation of confidence intervals but as we are mainly interested in predictive power here and the sample number is large, this does not seem to be a major problem. Once again, some points show high leverage in the Cook's plot but once again this is to be expected as the linear model uses so many categorical variables.

Final Model RMSE

```
ames_test <- ames_test %>%
  filter(Pool.QC != 'TA')
```

```
predictions_final_test <- exp(predict(final_model_AIC,ames_test))
residuals_final_test <- ames_test$price - predictions_final_test
```

```
rmse_final_test <- sqrt(mean(residuals_final_test^2))
rmse_final_test
```

```
## [1] 24639.6
```

Sadly, it seems the final model shows a bigger RMSE compared to the initial model in the test data even though it has a better Adjusted R^2 than the initial model. Thus, it probably means this model is more overfitted than the initial one and better predictions may be made with the initial simpler model.

Final Model Validation

Testing my final model on a separate, validation data set is a great way to determine how it will perform in real-life practice.

I am using the "ames_validation" dataset to do some additional assessment of my final model.

```
load("ames_validation.Rdata")

ames_validation <- ames_validation %>%
  mutate(Alley = if_else(is.na(Alley), 'No Alley Access', as.character(Alley)),
         Bsmt.Qual = if_else(is.na(Bsmt.Qual), 'No Basement', as.character(Bsmt.Qual)),
         Bsmt.Cond = if_else(is.na(Bsmt.Cond), 'No Basement', as.character(Bsmt.Cond)),
         Bsmt.Exposure = if_else(is.na(Bsmt.Exposure), 'No Basement', as.character(Bsmt.Cond)),
         BsmtFin.Type.1 = if_else(is.na(BsmtFin.Type.1), 'No Basement', as.character(BsmtFin.Type.1)),
         BsmtFin.Type.2 = if_else(is.na(BsmtFin.Type.2), 'No Basement', as.character(BsmtFin.Type.2)),
         Fireplace.Qu = if_else(is.na(Fireplace.Qu), 'No Fireplace', as.character(Fireplace.Qu)),
         Garage.Type = if_else(is.na(Garage.Type), 'No Garage', as.character(Garage.Type)),
         Garage.Finish = if_else(is.na(Garage.Finish), 'No Garage', as.character(Garage.Finish)),
         Garage.Qual = if_else(is.na(Garage.Qual), 'No Garage', as.character(Garage.Qual)),
         Garage.Cond = if_else(is.na(Garage.Cond), 'No Garage', as.character(Garage.Cond)),
         Pool.QC = if_else(is.na(Pool.QC), 'No Pool', as.character(Pool.QC)),
         Fence = if_else(is.na(Fence), 'No Fence', as.character(Fence)),
         Misc.Feature = if_else(is.na(Misc.Feature), 'No Misc Features', as.character(Misc.Feature)),
         Years.Old = 2018 - Year.Built,
         MS.SubClass = as.factor(MS.SubClass),
         Overall.Qual = as.factor(Overall.Qual),
         Overall.Cond = as.factor(Overall.Cond),
         log.price = log(price),
         log.Lot.Area = log(Lot.Area))
ames_validation <- ames_validation %>%
  filter(House.Style != '2.5Fin') %>%
  filter(Exterior.1st != 'CBlock' & Exterior.1st != 'PreCast') %>%
```

```
  filter(Pool.QC != 'TA') %>%
  filter(MS.SubClass != '150')
```

```
predictions_validation <- exp(predict(final_model_AIC,ames_validation))
residuals_validation <- ames_validation$price - predictions_validation
```

```
rmse_validation <- sqrt(mean(residuals_validation^2))
rmse_validation
```

```
## [1] 20494.79
```

Although the RMSE of the final model was higher in the test data compared to the initial model, in the validation data it achieves a lower RMSE (20494.79 dollars) than the one achieved in the test data (24639.6 dollars). This is a much better value than the one achieved in the test data and shows that perhaps the final model is not as overfitted to the training data as I originally thought.

Percentage of the 95% predictive confidence intervals that contain the true price of the house in the validation data set:

Predict prices

```
predict.full.CI <- exp(predict(final_model_AIC, ames_validation, interval = "prediction", level=0.95))
```

Calculate proportion of observations that fall within prediction intervals

```
coverage.prob.full <- mean(ames_validation$price > predict.full.CI[,"lwr"] &
                           ames_validation$price < predict.full.CI[,"upr"])
coverage.prob.full
```

```
## [1] 0.9495352
```

The coverage probability of this final model is approximately 95%, thus this model properly reflects uncertainty.

Conclusion

This dataset contains enough variables to build an interesting linear model that tries to predict the selling price for houses. Based on the reuslts from this model, it achieves a low RMSE in the validation data and quantifies uncertainty well. It also doesn't have major problems in the diagnostic plots. The variables that seem to be more important for predicting the selling price of a house according to AIC and this model are the logarithm of the Lot Area, Overall.Qual, Overall.Cond, Heating.QC, Year.Built, House.Style, Neighborhood,

Exterior.1st, X1st.Flr.SF, Year.Remod.Add, Bsmt.Qual, Garage.Area and Pool.QC. The final model has an adjusted R^2 of 92.1%, explaining 92.1% of the variance in the logarithm of house prices.

With this project I learned a lot about fitting linear models, exploring a new dataset, diagnosing problems and discovering how to choose and validate created models with test and validation data. Unfortunately, due to a lack of time I could only do this via a Frequentist Approach and I'm sure the Bayesian Approach results would also be interesting and would have the advantage of fitting priors that could quantify better some expert knowledge regarding the dataset.

In the future, collecting more variables in this dataset or using more advanced prediction methodologies than linear regression could contribute to achieve better predictive power starting with this final model as a scaffold.