Kaggle - House Price Prediction

MSDS6371 Class Project

By Mingyang Nick Yu & Allen Miller

I. Introduction

This project is focused on the Ames Housing dataset on Kaggle. We will first explore how the sales price of a home is influenced by the home's square footage in three different neighborhoods. We will then build different models aimed at accurately predicts a home's sales price based off of many different features of the home.

II. Data Description

The data comes from the Ames Housing dataset which describes the sale price of property in Ames, Iowa from the years 2006 to 2010. This includes 80 distinct variables, all of which help describe a property and may influence the price of a home. We looked at all the observations included in the dataset (2930) and used a mixture of these variables to build models that can aid in predicting a home's sale price.

III. Analysis 1

a. Restate the problem

The Century 21 Ames real estate company has hired us to analyze how square footage of living area of house(GrLivArea) is related to its sales price in the three neighborhoods they sell homes in (NAmes, BrkSide, and Edward neighborhoods).

b. Build and Fit

i. Model 1

1. $SalesPrice = \beta_0 + \beta_1 SquareFootage$
   a. $\boldsymbol{model1} = lm(SalePrice \sim GrlivArea, data = data1)$
2. Check Assumptions *(see Appendix 1.1, 1.2, 1.3, 1.4)*
   a. Linearity: *Plot 1.1* has some outliers even the general trend follows linear relationship between Square Foot of Living Area and Sale Price
   b. Normality: Q-Q *Plot(1.4)* shows skewness at either end of the plot – some outliers
   c. Equal SD: Observations 169 and 190 both have standardized residuals greater than 4 - outliers
   d. Independence: We will assume observations are independent
   e. Outliers: Observation 339 has a Cook's Distance larger than 5.6 while observation 131 has a Cook's Distance larger than 1.
      i. These may be due to some unique cases

3. Decision: Since our sample is sufficiently large, removing these four outliers should not affect our results and will be removed for further analysis. (Adjusted R-squared = 0.3406)

ii. Model 2
    1. $SalesPrice = \beta_0 + \beta_1 SquareFootage$
        a. $\boldsymbol{model2} = lm(SalePrice \sim GrlivArea, data = data2)$
        b. Model2 = model1 run without outliers
    2. Check Assumptions *(Appendix 1.5, 1.6, 1.7, 1.8)*
        a. Linearity: *Plot 1.5* shows linear relationship between square foot of living area and sales price
        b. Normality: Q-Q Plot shows normal distribution
        c. Equal SD: standardized residuals show all data within 2.5 range, so this assumption is met
        d. Independence: Assume our observations are independent.
        e. Outliers: All observations have a Cooks D less than 0.01, so no high leverage and high residual point.
    3. All assumptions are met with this model, however we are interested in adding in the neighborhood factors. So we will go on building out Model 3 and see if the Neighborhood is significant in our analysis. (Adjusted R-squared= 0.449)

iii. Model 3
    1. $SalesPrice = \beta_0 + \beta_1 SquareFootage + \beta_2(N1) + \beta_3(N2) + \beta_4(N1)SquareFootage + \beta_5(N2)SquareFootage$
        a. $\boldsymbol{model3} = lm(SalePrice \sim GrlivArea + Neighborhood + GrlivArea * Neighborhood, data = data2)$
        b. Model3 is run with neighborhood data added in.
    2. Check Assumptions *(Appendix 1.9, 1.10, 1.11, 1.12)*
        a. Linearity: As we seen from *Plot 1.9*, each neighborhood follows a linear relationship between square foot of living area and sales price
        b. Normality: QQ-Plot shows normal distribution of our model
        c. Equal SD: Standardized residual plot shows about 5% of data is beyond residual value of 2. So this assumption is met.
        d. Independence: We will assume all observations are independent.
        e. Outliers: observations have a Cook's D less than 0.20, so no major high leverage or high residual point.
    3. Conclusion: Adjusted R-squared = 0.5165, which is higher than model 2, by adding in Neighborhoods, our model is better explained. So we will move forward to interpret Model 3.

c. **The Analysis**
    i. Using model 3 we can generate separate models(one for each neighborhood)
        1. Overall Model:

a. $SalesPrice = 19971.514 + 87.163SquareFootage + 17128.908(Edwards) + 60354.199(NAmes) + -17.004(Edwards)SquareFootage + -37.601(Ames)SquareFootage$

2. BrkSide Model:
   a. $SalesPrice = 19971.514 + 87.163 * SquareFootage$
3. Edwards Model:
   a. $SalesPrice = 37100.422 + 70.159 * SquareFootage$
4. NAmes Model:
   a. $SalesPrice = 80325.739 + 49.562 * SquareFootage$

ii. Analyze the plot and assumptions
   1. Linearity, Normality, Equal SD, Independence and outliers are checked with section above.
   2. There is no evidence to suggest any major outliers that will need to be accounted for as the residuals appear in a random cloud. All assumptions are met, we will move on to interpret our findings.

d. **Conclusion**
   i. There is sufficient evidence to suggest that Model 3 is a good fit for the data (p-value < 0.0001).
   ii. We can interpret each sub-model of model 3 as follows
      1. Given that the Neighborhood is BrkSide, it is predicted that the Sales price of the house will increase by $8716.3 for every 100 square feet added to the house. We are 95% confident this increase will be between 7152.22 and 10380.29.
      2. Given that the Neighborhood is Edwards, it is predicted that the Sales Price of the house will increase by $7015.9 for every 100 square feet added to the house. We are 95% confident this increase will be between 5618.25 and 8413.43
      3. Given that the neighborhood is NAmes, it is predicted that the Sales Price of the house will increase by $4956.2 for every 10 square feet added to the house. We are 95% confident this increase will be between 4150.50 and 5761.75
   iii. Scope: Because this an observational study we cannot draw any causal inference. Not knowing if this data set is randomly drawn from a bigger population from the entire sales data from 2006-2010, any inference to the population needs to remain speculative.

IV. Analysis 2

a. **Restate the Problem**
   i. Select from all the variables available to us, and build a model that can accurately predict the sales price of a home in Ames, Iowa between 2006 and 2010.
   ii. We will first explore the Stepwise, Forward, and Backward models and use our findings to create a more accurate model.

b. **Clean-up and selection**
    i. We will first look at each variable and convert specific variables that are levels to factors so we can use them as categorical variable for linear regression.
    ii. We plotted each continuous variable vs the sales price of the house to look for correlation and independence. From there we will begin to assemble a list of variables with strong correlation that may be good predictors in our regression model. (See *plot 2.1 – 2.12*)
    iii. 31 of the 80 variables were selected for our models
        1. We used the pool area variable to us to create a new variable (poolYN) that said Yes or No for a house having a pool.
    iv. After selection we decided to look closely at our selected variables that contained N/A and transform those values to useable factors such as None for quality rankings. We then replotted these variables to confirm that there was still strong correlation.
    v. After plotting the residuals, we decided to use a log transform on the sales price to help normalize or data better.
c. **Build and Fit Models**
    i. Stepwise
        1. Using R and our selected variables we created a model using stepwise AIC to choose the optimal variables/model among the variables we have narrowed down above, it is further verified by internal 10-fold cross validation.
        2. Checking Assumptions
            a. Linearity: this has been checked by pair wise *plots 2.1-2.12*
            b. Normality: see Q-Q *plot 2.13*, normality is roughly met, although there are some outliers at the ends
            c. Equal SD: see Standardized Residuals *plot 2.14*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
            d. Independence: Assume all of our observations are independent
            e. Outliers: Looking at the Cook's D *plot 2.15*, there is one data point went over 1.5, comparing to our sample size, it should not have a huge impact on our model. So we will keep this observation and move on.
        3. Conclusions: The Stepwise model has selected following predictors (see *plot 2.16*): MSSubClass, MSZoning, LotArea, LotConfig, Neighborhood, HouseStyle, OverallQual, YearBuilt, YearRemodAdd, ExterCond, Foundation, BsmtQual, BsmtCond, TotalBsmtSF, Heating, CentralAir, GrLivArea, FullBath, KitchenQual, Fireplaces, GarageType, GarageCars, PollYN, MoSold, YrSold, among the 31 variables we feed to the model. And it give us RMSE of 0.149 and follow performance:
            a. Final Results
                i. Kaggle Score = 0.15372

ii. CV Press = 0.1494
iii. Adjusted R-squared = 0.8975

ii. Forward Selection
    1. Using R and our selected variables we created another model using Forward selection by AIC to choose the optimal variables/model among the variables we have narrowed down above, it is further verified by creating our own cross validation.
    2. Checking Assumptions:
        a. Linearity: this has been checked by pair wise *plotting 2.1-2.12*
        b. Normality: see Q-Q *plot 2.17*, normality is roughly met, although there are some outliers at the ends
        c. Equal SD: see Standardized Residuals *plot 2.18*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
        d. Independence: Assume all of our observations are independent
        e. Outliers: Looking at the Cook's D *plot 2.19*, there is one data point went over 1.25, comparing to our sample size, it should not have a huge impact on our model. So we will keep this observation and move on.
    3. Conclusion: The Forward Selection has chosen the following predictors(see *plots 2.20, 2.21*): OverallQual, Neighborhood, GrLivArea, MSSubClass, OverallCond, GarageCars, YearBuilt, Fireplaces, BsmtQual, MSZoning, Heating, LotArea, YearRemodAdd, CentralAir, KitchenQual, GarageType, TotalBsmtSF, PoolYN, BsmtCond, LotConfig, FullBath among the 31 variables we fed to the automatic model selection. It gives us the following performance for prediction and Cross validation.
        a. Kaggle Score = 0.15432
        b. CV Press = 0.14199
        c. Adjusted R-Squared = 0.89575

iii. Backward Selection
    1. Using R and our selected variables we created the next model using Backward selection by AIC to choose the optimal variables/model among the variables we have narrowed down above, it is further verified by creating our own cross validation.
    2. Checking Assumptions:
        a. Linearity: this has been checked by pair wise *plotting 2.1-2.12*
        b. Normality: see Q-Q *plot 2.22*, normality is roughly met, although there are some outliers at the ends
        c. Equal SD: see Standardized Residuals *plot 2.23*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
        d. Independence: Assume all of our observations are independent

e. Outliers: Looking at the Cook's D *plot 2.24*, there is one data point went over 1.25, comparing to our sample size, it should not have a huge impact on our model. So we will keep this observation and move on.

3. Conclusion: The Backward Selection has chosen the following predictors(see *plots 2.25, 2.26*): MSSubClass, MSZoning, LotArea, LotConfig, Neighborhood, OverallQual, YearRemodAdd, BsmtQual, TotalBsmtSF, Heating, CentralAir, GrLivArea, FullBath, KitchenQual, Fireplaces, GarageType, GarageCars, PoolYN, among the 31 variables we fed to the automatic model selection. It gives us the following performance for prediction and Cross validation.
    a. Kaggle Score = 0.15432
    b. CV Press = 0.148
    c. Adjusted R-Squared = 0.89575

iv. Custom Model:

1. To design our optimal custom model, we used following steps:
    a. Reimport the training dataset, use predictive mean matching to fill in missing continuous variables (use best estimate and keep original distribution of each variable.)
    b. Divide all 80 variables available to us into different subgroups
    c. Run best subset selection to pick out the best predictors within each subgroup
    d. Log transformation on SalePrice and GrLivArea to increase linearity relationship.
    e. Convert Categorical variables with NA to None or Others.
    f. Use all predictors selected under each subgroup to run a stepwise AIC to choose the optimal variables/model
    g. Further verified model by internal 10-fold cross validation.
2. Checking Assumptions:
    a. Linearity: this has been checked by pair wise *plotting 2.1-2.12*
    b. Normality: see Q-Q *plot 2.27*, normality is roughly met, although there are some outliers at the ends
    c. Equal SD: see Standardized Residuals *plot 2.28*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
    d. Independence: Assume all of our observations are independent
    e. Outliers: Looking at the Cook's D *plot 2.29*, there are two data points went around 0.6. This should not be a problem considering our sample size.
3. Conclusion: The Custom model has chosen the following predictors(see *plots 2.30, 2.31*): MSSubClass, MSZoning, LotArea, LotConfig, Neighborhood, Condition2, BldgType, OverallQual, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, Exterior1st, Exterior2nd,

MasVnrType, Foundation, BsmtQual, BsmtFinType1, TotalBsmt, Heating, HeatingQC, CentralAir, Electrical, X2ndFlr, GrLivArea, BsmtFullBath, FullBath, BedroomAbvGr, KitchenAbvGr, KitchenQual, Firepplaces, GarageCars, GarageArea, PavedDrive, WoodDeckSF, ScreenPorch, MoSold, YrSold. It gives us the following performance for prediction and Cross validation.

a. Kaggle Score = 0.14773
b. CV Press = 0.1649
c. Adjusted R-Squared = 0.9207

d. **Overall Conclusion:**

After running and comparing the four models we built, see table below, the Custom model gives us the best Adjusted R-Squared score (highest) and best Kaggle Score(Lowest). After all, we choose custom model as our best model.

Predictive Models	**Adjusted R2**	**CV PRESS**	**Kaggle Score**
Forward	0.89575	0.14199	0.15432
Backward	0.89575	0.148	0.15432
Stepwise	0.8975	0.1494	0.15372
CUSTOM	0.9207	0.1649	0.14773

V. Reference

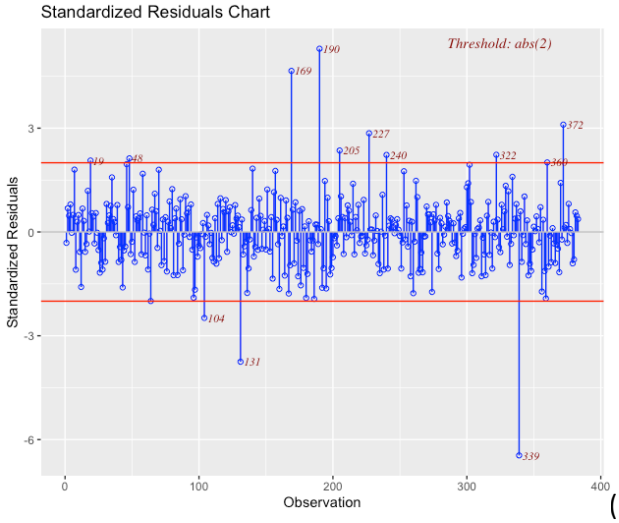
a. MSDS 6371 Project Description.docx – see details at https://github.com/nickmingyang/MSDS6371Project
b. Kaggle Competition – Data description, training and testing Data source, and other data details: https://www.kaggle.com/c/house-prices-advanced-regression-techniques
c. Project source code used for analysis and results: MSDS6371Project.Rmd under: https://github.com/nickmingyang/MSDS6371Project
d. Custom Model test set prediction—custom_model_Miller_YU.csv under: https://github.com/nickmingyang/MSDS6371Project
e. Stepwise AIC model: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/
f. Other models result labeled csv files under: https://github.com/nickmingyang/MSDS6371Project

VI. Appendix

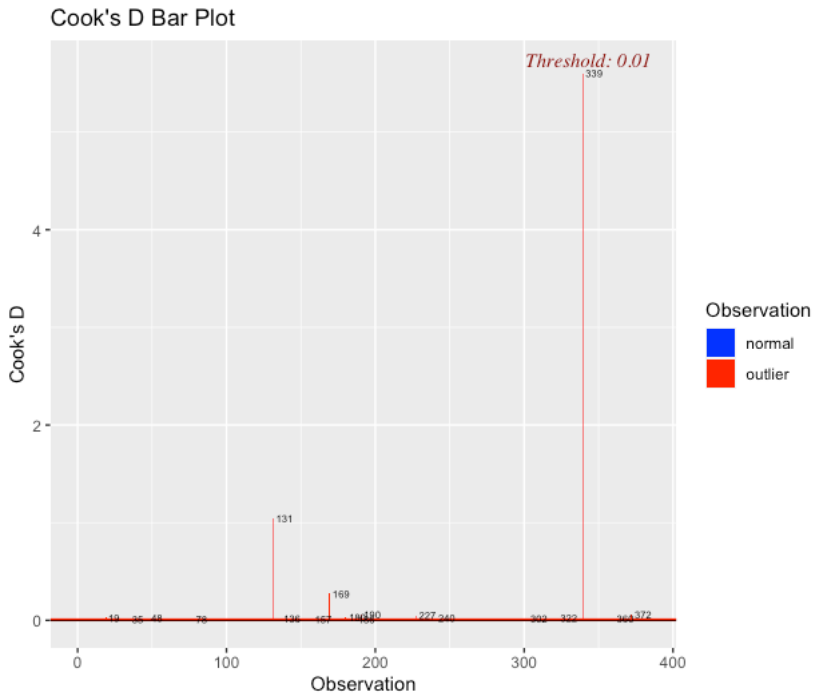
a. Analysis 1

i. Model 1

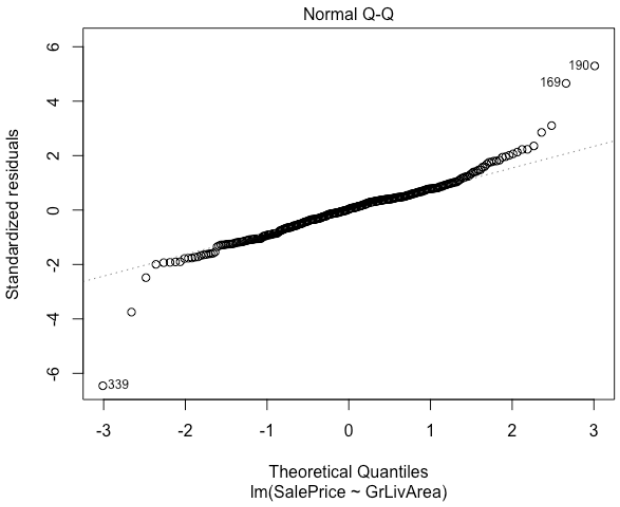
(1.1)



(1.2)

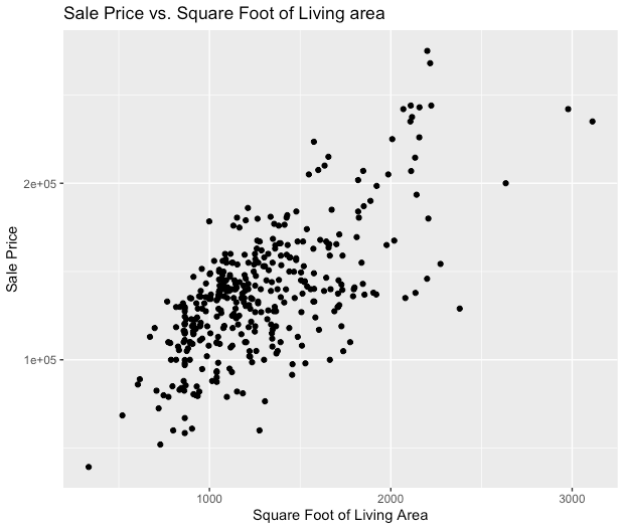


(1.3)

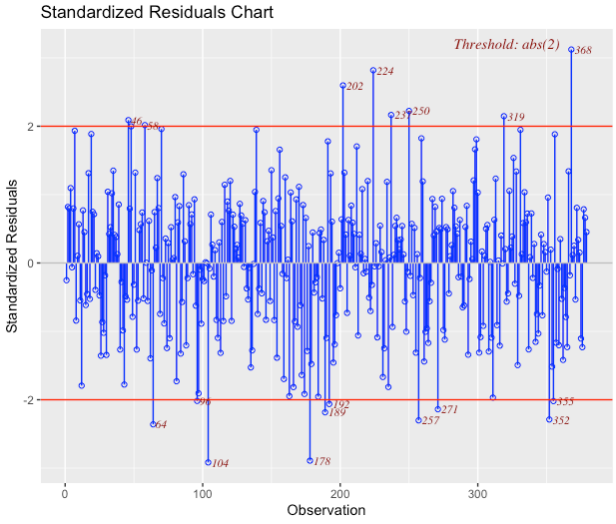


(1.4)

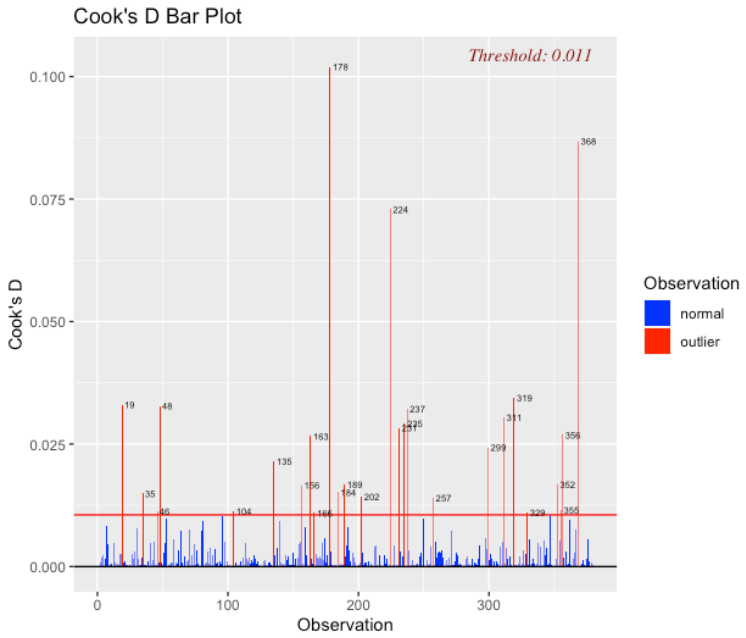
ii. Model 2


Sale Price vs. Square Foot of Living area
Sale Price
2e+05
1e+05
1000
2000
3000
Square Foot of Living Area

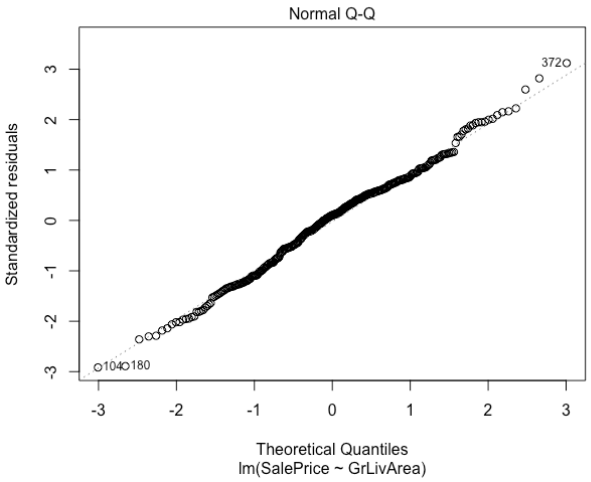
(1.5)


Standardized Residuals Chart
Threshold: abs(2)
Standardized Residuals
2
0
-2
0
100
200
300
Observation
368
224
202
250
319
271
257
352
355
192
189
178
104
64

(1.6)



(1.7)

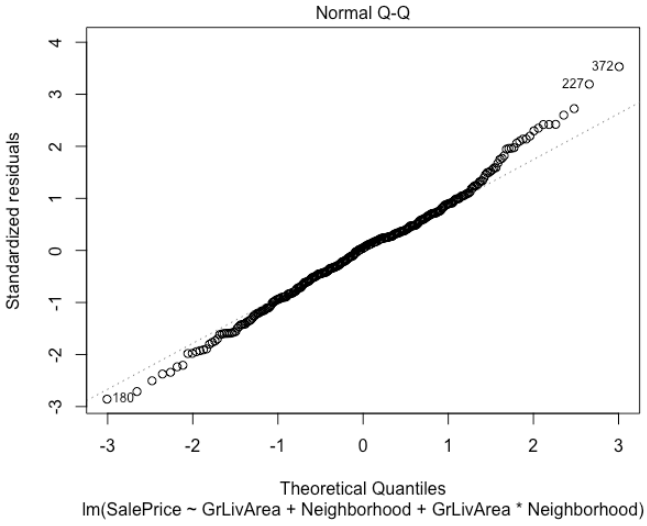


(1.8)

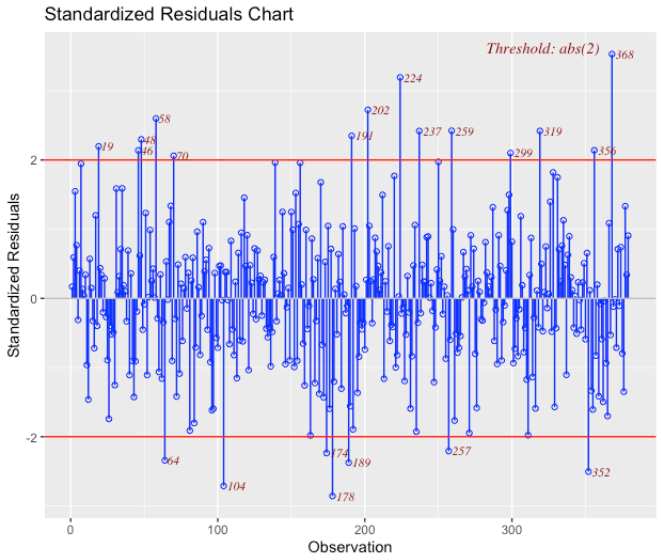
iii. Model 3


Sale Price vs. Square Foot of Living area
Sale Price
2e+05
1e+05
1000
2000
3000
Square Foot of Living Area
Neighborhood
BrkSide
Edwards
NAmes

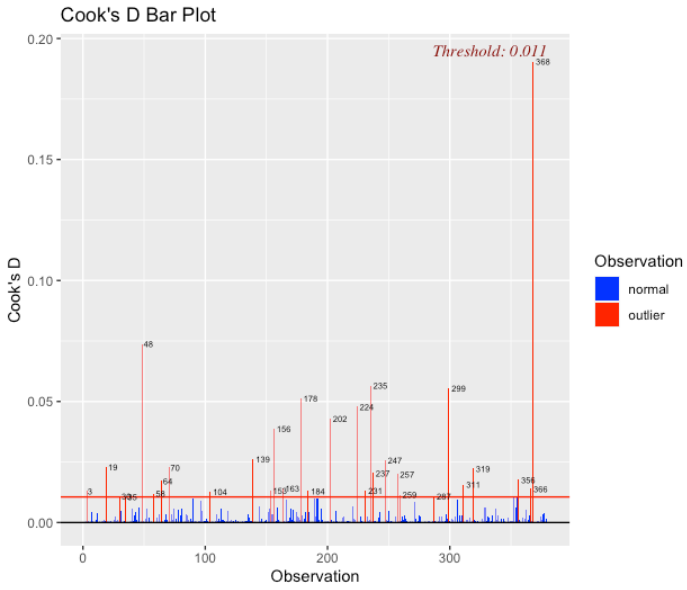
(1.9)


Normal Q-Q
Standardized residuals
372
227
180
-3
-2
-1
0
1
2
3
4
Theoretical Quantiles
lm(SalePrice ~ GrLivArea + Neighborhood + GrLivArea * Neighborhood)

(1.10)

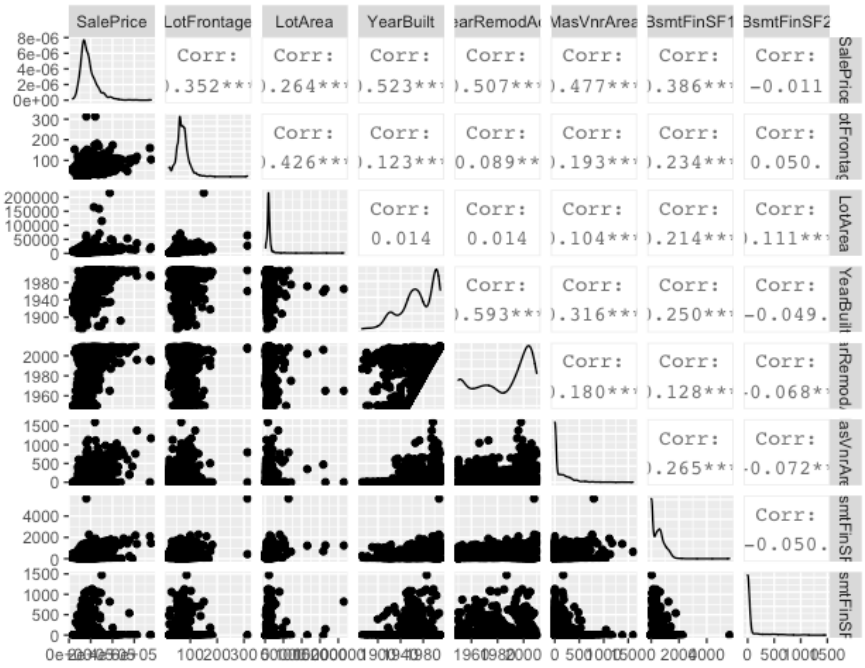


(1.11)

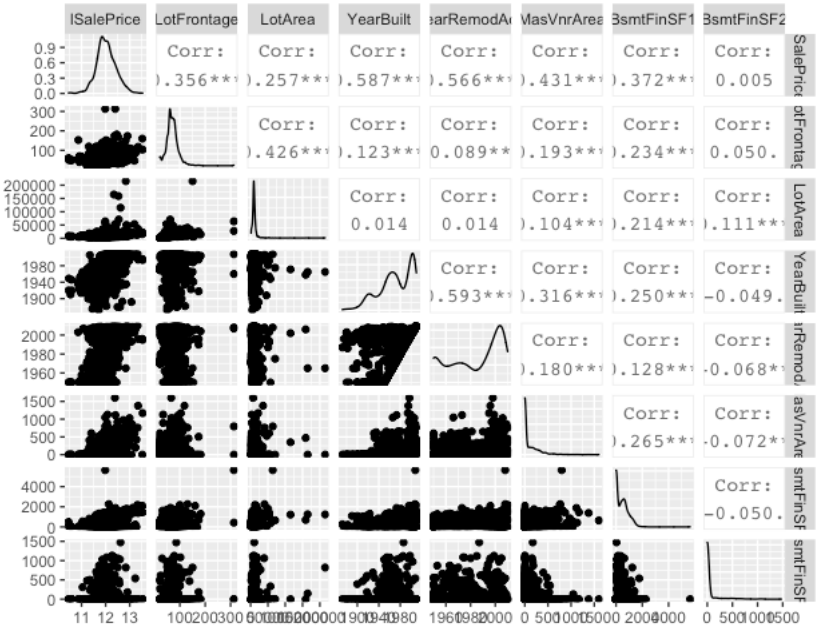


(1.12)

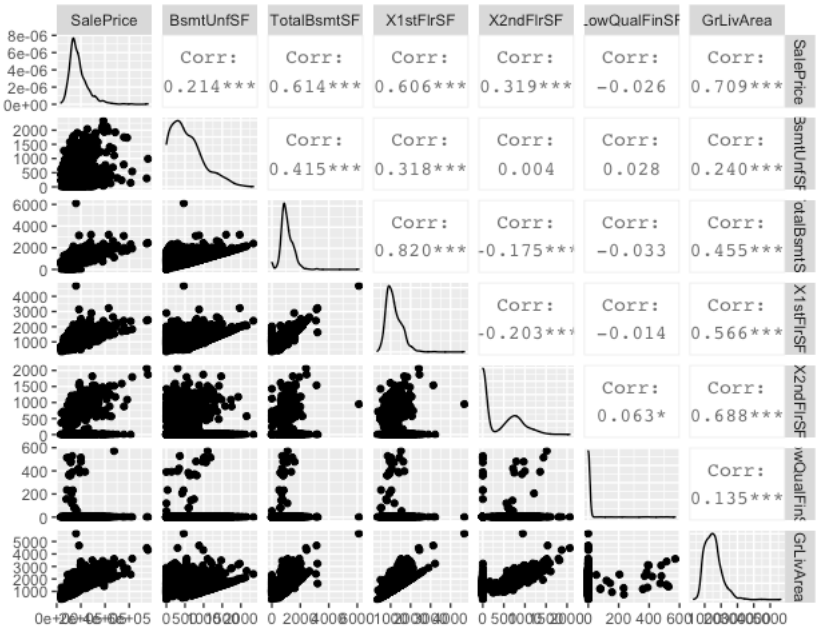
b. Analysis 2


SalePrice
LotFrontage
LotArea
YearBuilt
MasVnrArea
BsmtFinSF1
BsmtFinSF2
Corr:

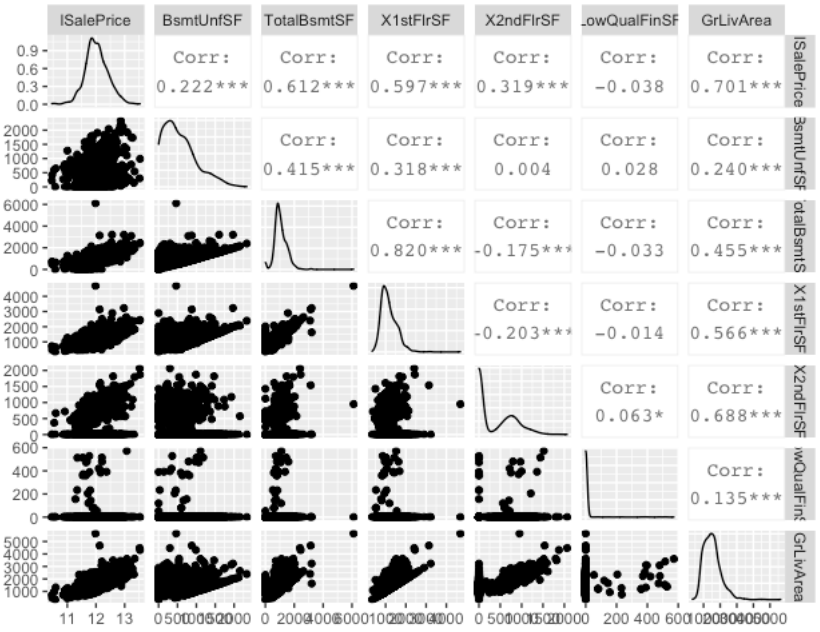
(2.1)


SalePrice
LotFrontage
LotArea
YearBuilt
MasVnrArea
BsmtFinSF1
BsmtFinSF2
Corr:

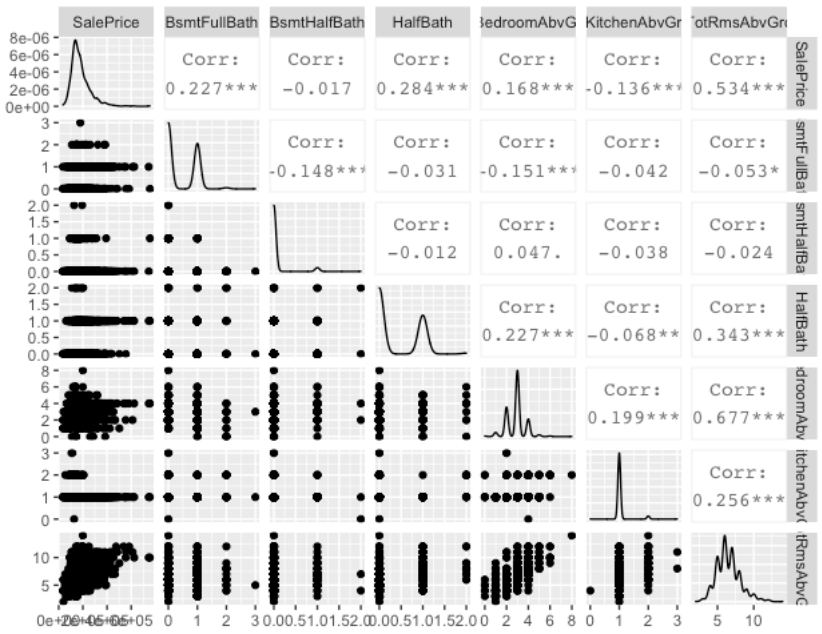
(2.2)


SalePrice
BsmtUnfSF
TotalBsmtSF
X1stFlrSF
X2ndFlrSF
LowQualFinSF
GrLivArea
Corr: 0.214***
Corr: 0.614***
Corr: 0.606***
Corr: 0.319***
Corr: −0.026
Corr: 0.709***
Corr: 0.415***
Corr: 0.318***
Corr: 0.004
Corr: 0.028
Corr: 0.240***
Corr: 0.820***
Corr: -0.175***
Corr: −0.033
Corr: 0.455***
Corr: -0.203***
Corr: −0.014
Corr: 0.566***
Corr: 0.063*
Corr: 0.688***
Corr: 0.135***

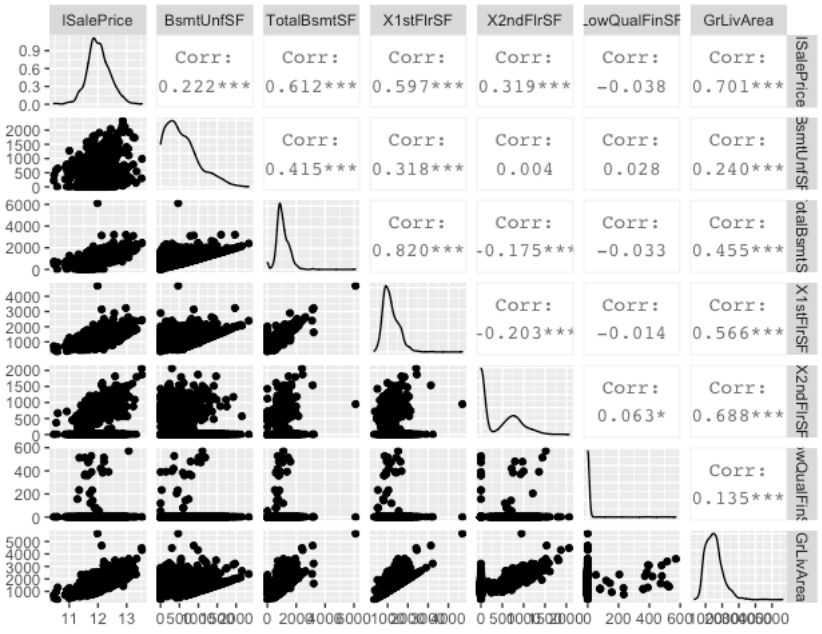
(2.3)


lSalePrice
BsmtUnfSF
TotalBsmtSF
X1stFlrSF
X2ndFlrSF
LowQualFinSF
GrLivArea
Corr: 0.222***
Corr: 0.612***
Corr: 0.597***
Corr: 0.319***
Corr: −0.038
Corr: 0.701***
Corr: 0.415***
Corr: 0.318***
Corr: 0.004
Corr: 0.028
Corr: 0.240***
Corr: 0.820***
Corr: -0.175***
Corr: −0.033
Corr: 0.455***
Corr: -0.203***
Corr: −0.014
Corr: 0.566***
Corr: 0.063*
Corr: 0.688***
Corr: 0.135***

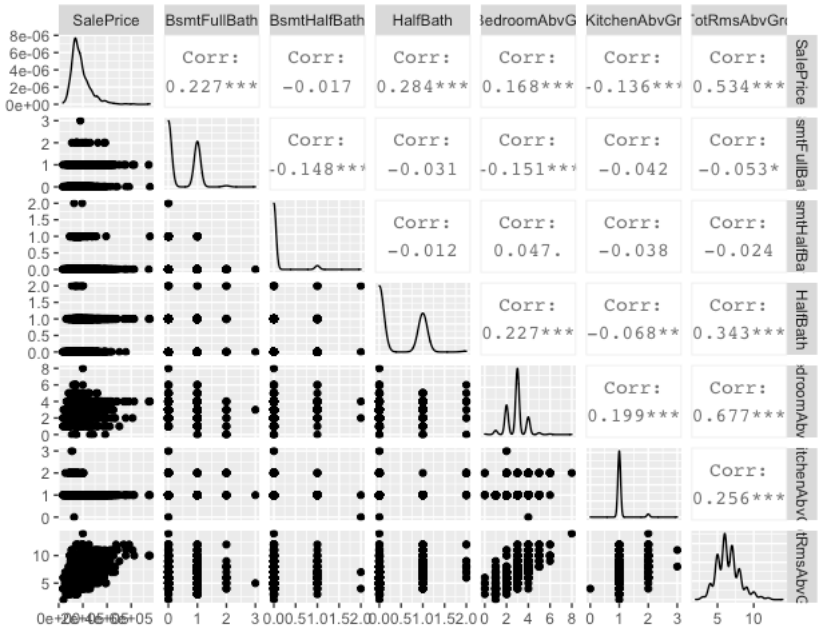
(2.4)


SalePrice
BsmtFullBath
BsmtHalfBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Corr: 0.227***
Corr: -0.017
Corr: 0.284***
Corr: 0.168***
Corr: -0.136***
Corr: 0.534***
Corr: -0.148***
Corr: -0.031
Corr: -0.151***
Corr: -0.042
Corr: -0.053*
Corr: -0.012
Corr: 0.047.
Corr: -0.038
Corr: -0.024
Corr: 0.227***
Corr: -0.068**
Corr: 0.343***
Corr: 0.199***
Corr: 0.677***
Corr: 0.256***

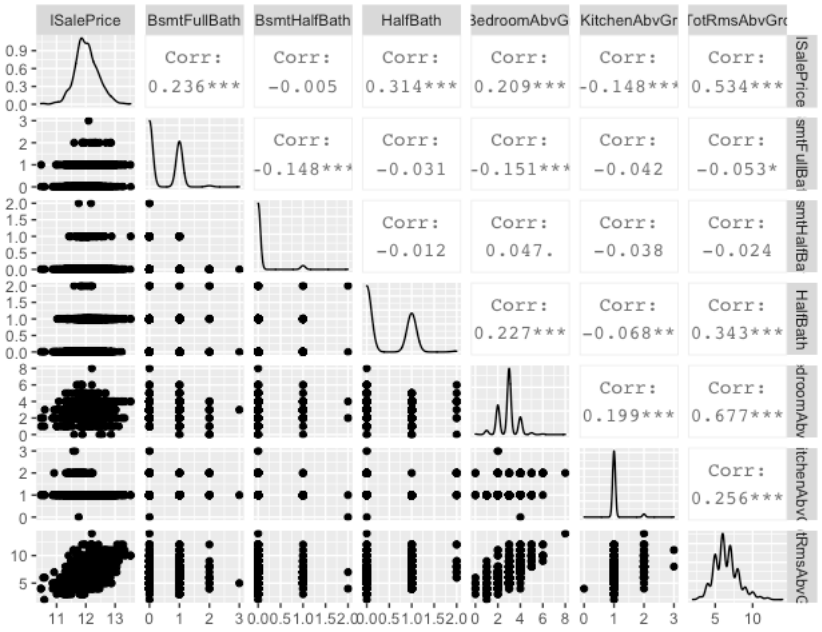
(2.5)


lSalePrice
BsmtUnfSF
TotalBsmtSF
X1stFlrSF
X2ndFlrSF
LowQualFinSF
GrLivArea
Corr: 0.222***
Corr: 0.612***
Corr: 0.597***
Corr: 0.319***
Corr: -0.038
Corr: 0.701***
Corr: 0.415***
Corr: 0.318***
Corr: 0.004
Corr: 0.028
Corr: 0.240***
Corr: 0.820***
Corr: -0.175***
Corr: -0.033
Corr: 0.455***
Corr: -0.203***
Corr: -0.014
Corr: 0.566***
Corr: 0.063*
Corr: 0.688***
Corr: 0.135***

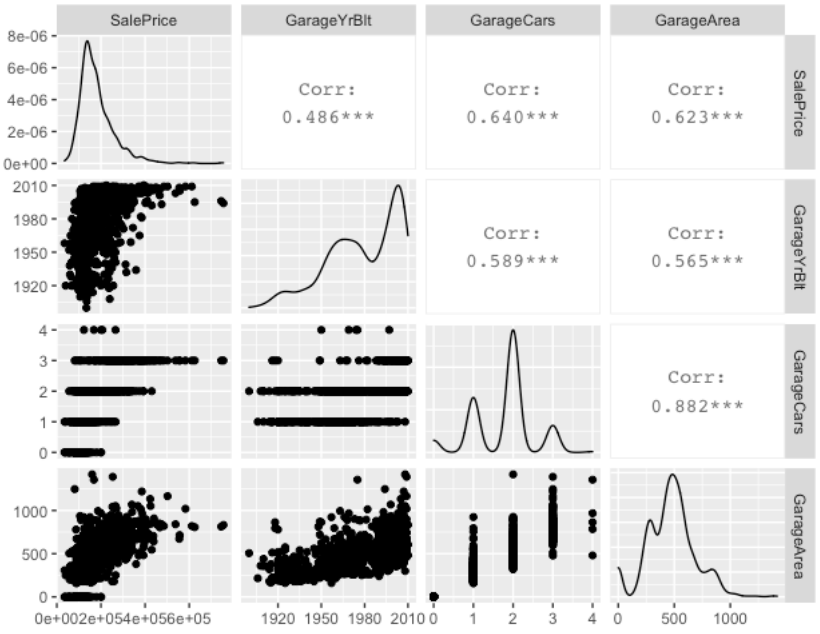
(2.6)


SalePrice
BsmtFullBath
BsmtHalfBath
HalfBath
KitchenAbvGr
Corr: 0.227***
Corr: -0.017
Corr: 0.284***
Corr: 0.168***
Corr: -0.136***
Corr: 0.534***
Corr: -0.148***
Corr: -0.031
Corr: -0.151***
Corr: -0.042
Corr: -0.053*
Corr: -0.012
Corr: 0.047.
Corr: -0.038
Corr: -0.024
Corr: 0.227***
Corr: -0.068**
Corr: 0.343***
Corr: 0.199***
Corr: 0.677***
Corr: 0.256***

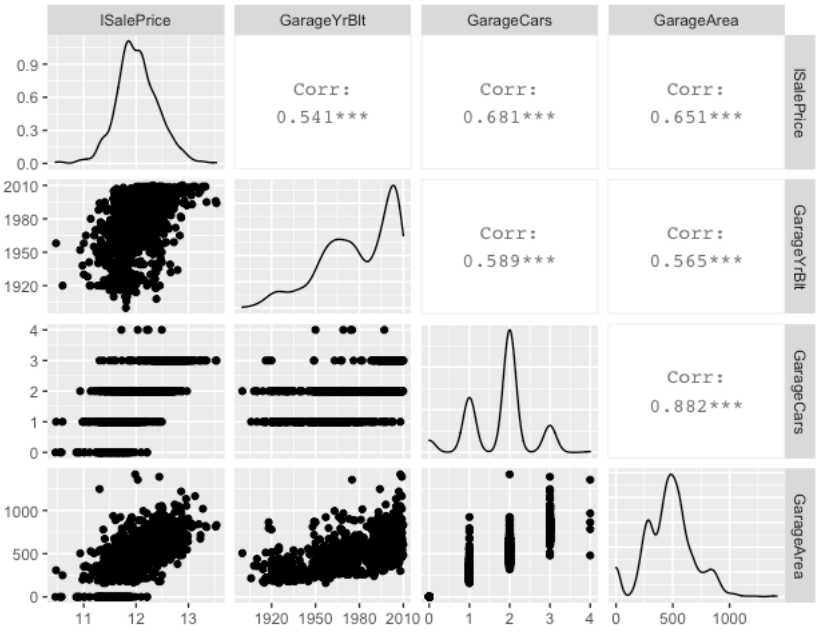
(2.7)


lSalePrice
BsmtFullBath
BsmtHalfBath
HalfBath
KitchenAbvGr
Corr: 0.236***
Corr: -0.005
Corr: 0.314***
Corr: 0.209***
Corr: -0.148***
Corr: 0.534***
Corr: -0.148***
Corr: -0.031
Corr: -0.151***
Corr: -0.042
Corr: -0.053*
Corr: -0.012
Corr: 0.047.
Corr: -0.038
Corr: -0.024
Corr: 0.227***
Corr: -0.068**
Corr: 0.343***
Corr: 0.199***
Corr: 0.677***
Corr: 0.256***

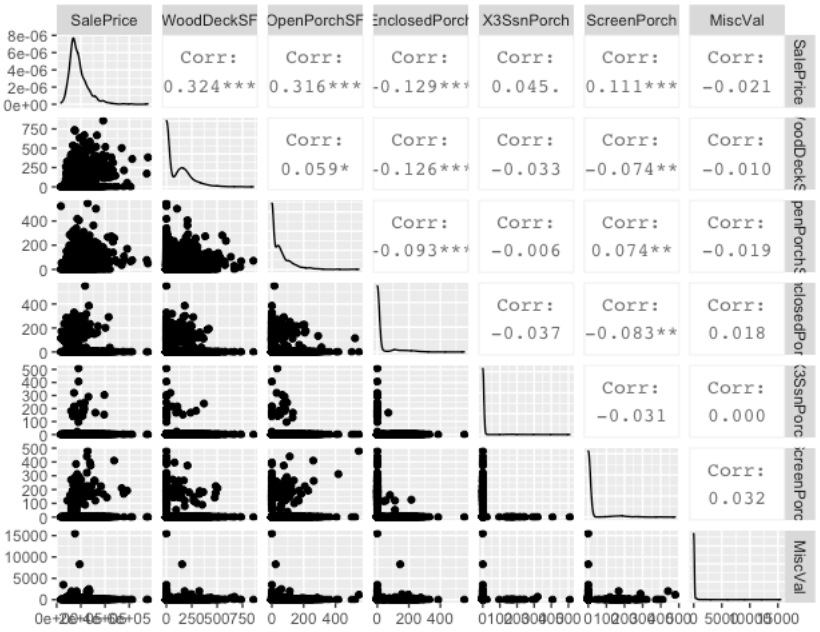
(2.8)


SalePrice
GarageYrBlt
GarageCars
GarageArea
Corr:
0.486***
Corr:
0.640***
Corr:
0.623***
Corr:
0.589***
Corr:
0.565***
Corr:
0.882***

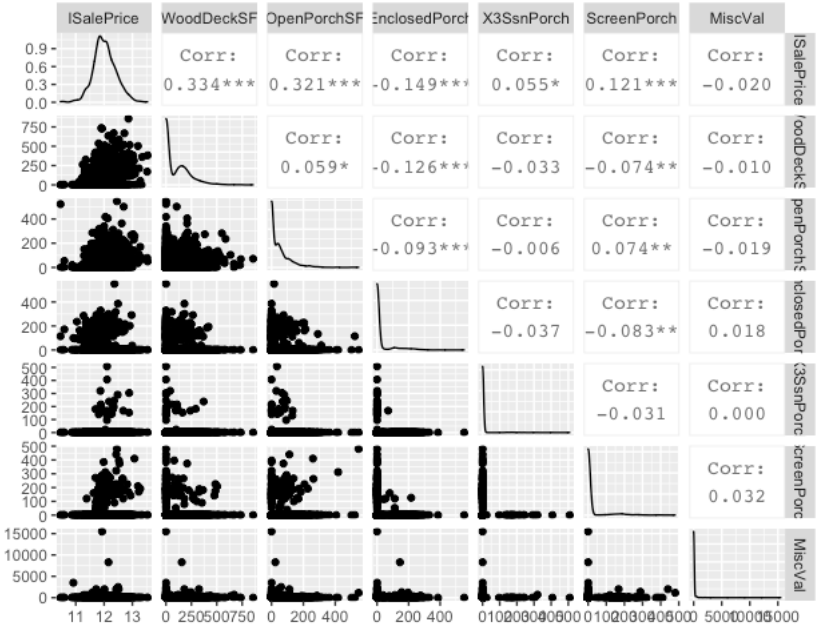
(2.9)


lSalePrice
GarageYrBlt
GarageCars
GarageArea
Corr:
0.541***
Corr:
0.681***
Corr:
0.651***
Corr:
0.589***
Corr:
0.565***
Corr:
0.882***

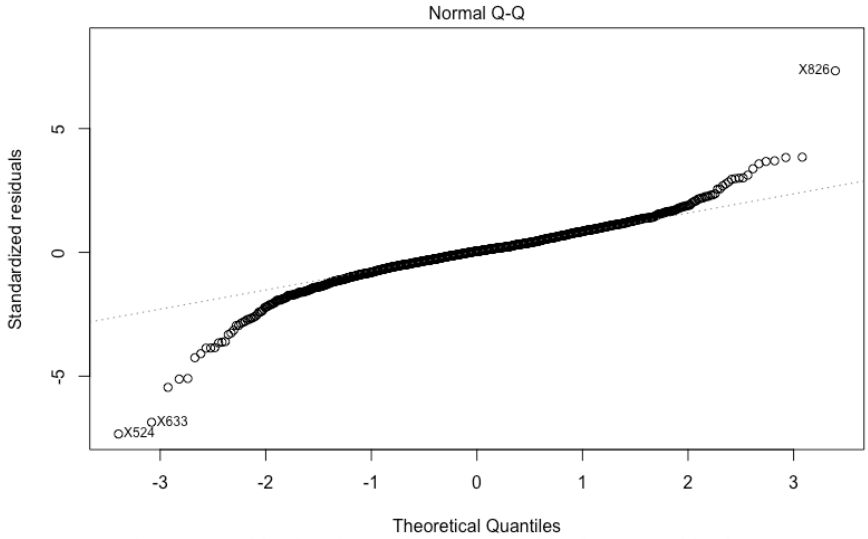
(2.10)


SalePrice
WoodDeckSF
OpenPorchSF
EnclosedPorch
X3SsnPorch
ScreenPorch
MiscVal
Corr: 0.324***
Corr: 0.316***
Corr: -0.129***
Corr: 0.045.
Corr: 0.111***
Corr: -0.021
Corr: 0.059*
Corr: -0.126***
Corr: -0.033
Corr: -0.074**
Corr: -0.010
Corr: -0.093***
Corr: -0.006
Corr: 0.074**
Corr: -0.019
Corr: -0.037
Corr: -0.083**
Corr: 0.018
Corr: -0.031
Corr: 0.000
Corr: 0.032

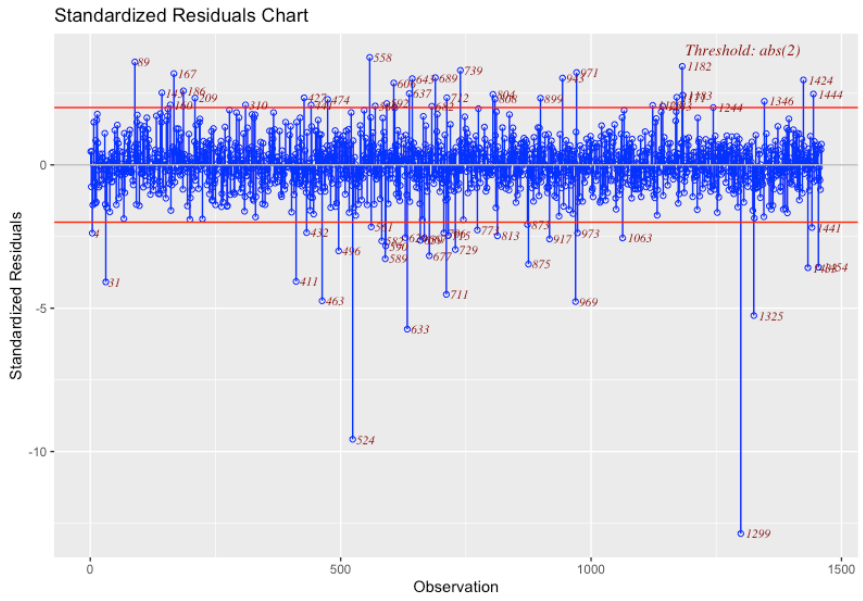
(2.11)


lSalePrice
WoodDeckSF
OpenPorchSF
EnclosedPorch
X3SsnPorch
ScreenPorch
MiscVal
Corr: 0.334***
Corr: 0.321***
Corr: -0.149***
Corr: 0.055*
Corr: 0.121***
Corr: -0.020
Corr: 0.059*
Corr: -0.126***
Corr: -0.033
Corr: -0.074**
Corr: -0.010
Corr: -0.093***
Corr: -0.006
Corr: 0.074**
Corr: -0.019
Corr: -0.037
Corr: -0.083**
Corr: 0.018
Corr: -0.031
Corr: 0.000
Corr: 0.032

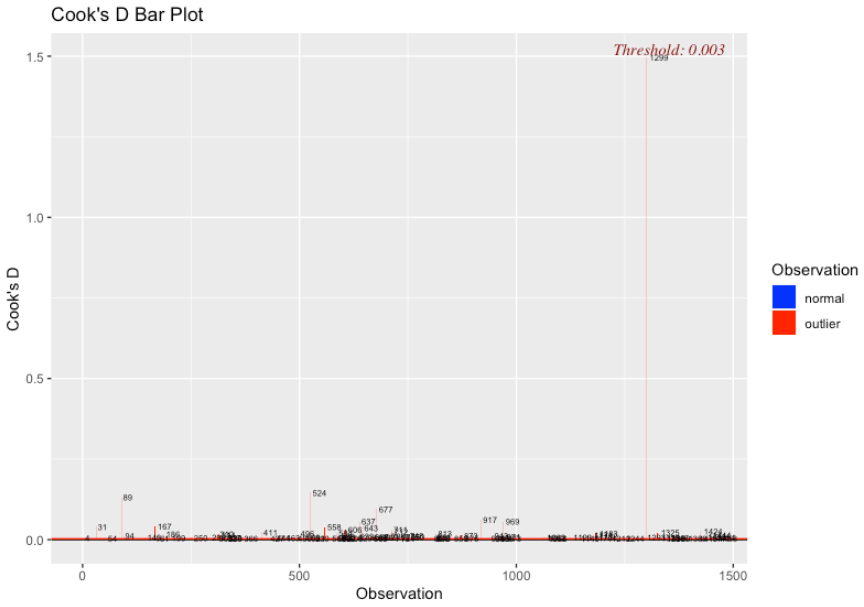
(2.12)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(.outcome ~ MSSubClass30 + MSSubClass50 + MSSubClass60 + MSSubClass70 + M ...
X826
X633
X524

(2.13)


Standardized Residuals Chart
Threshold: abs(2)
Standardized Residuals
Observation

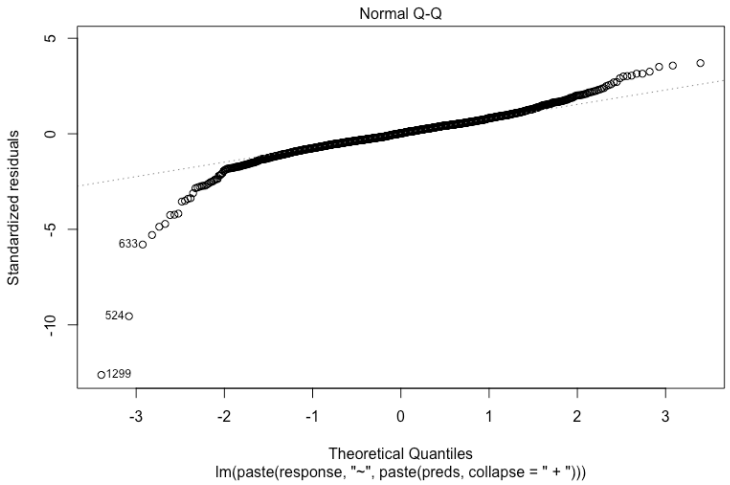
(2.14)



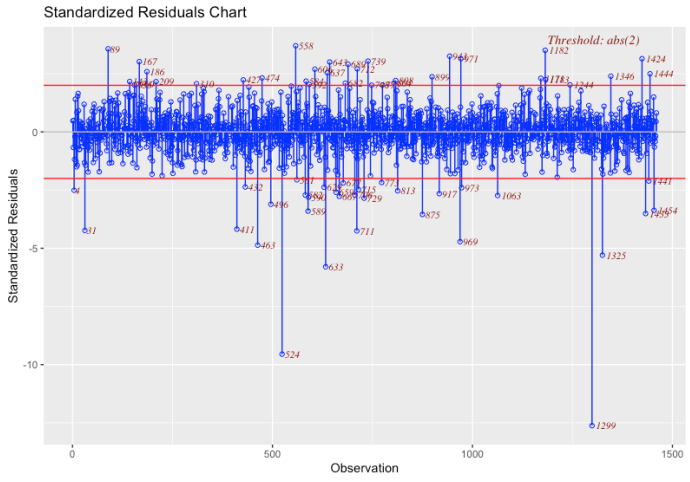
(2.15)

```
Call:
lm(formula = .outcome ~ MSSubClass30 + MSSubClass75 + MSSubClass90 +
    MSSubClass120 + MSSubClass160 + MSZoningFV + MSZoningRH +
    MSZoningRL + MSZoningRM + LotArea + LotConfigCulDSac + NeighborhoodBrkSide +
    NeighborhoodClearCr + NeighborhoodCollgCr + NeighborhoodCrawfor +
    NeighborhoodIDOTRR + NeighborhoodMeadowV + NeighborhoodNAmes +
    NeighborhoodNoRidge + NeighborhoodNridgHt + NeighborhoodSawyerW +
    NeighborhoodSomerst + NeighborhoodStoneBr + NeighborhoodSWISU +
    NeighborhoodTimber + NeighborhoodVeenker + HouseStyle1Story +
    HouseStyleSFoyer + HouseStyleSLvl + OverallQual2 + OverallQual3 +
    OverallQual4 + OverallQual5 + OverallQual6 + OverallQual7 +
    OverallQual8 + OverallQual9 + OverallQual10 + OverallCond2 +
    OverallCond3 + OverallCond4 + OverallCond5 + OverallCond6 +
    OverallCond7 + OverallCond8 + OverallCond9 + YearBuilt +
    YearRemodAdd + ExterCondFa + ExterCondGd + ExterCondPo +
    ExterCondTA + FoundationPConc + FoundationStone + BsmtQualFa +
    BsmtQualGd + BsmtQualNONE + BsmtQualTA + BsmtCondGd + BsmtCondPo +
    BsmtCondTA + TotalBsmtSF + HeatingGasA + HeatingGasW + HeatingWall +
    CentralAirY + GrLivArea + FullBath1 + KitchenQualFa + KitchenQualGd +
    KitchenQualTA + Fireplaces1 + Fireplaces2 + Fireplaces3 +
    GarageTypeAttchd + GarageTypeBasment + GarageTypeBuiltIn +
    GarageTypeDetchd + GarageTypeNONE + GarageCars + PoolYNYES +
    MoSold5 + MoSold7 + YrSold2007 + YrSold2009, data = dat)
```

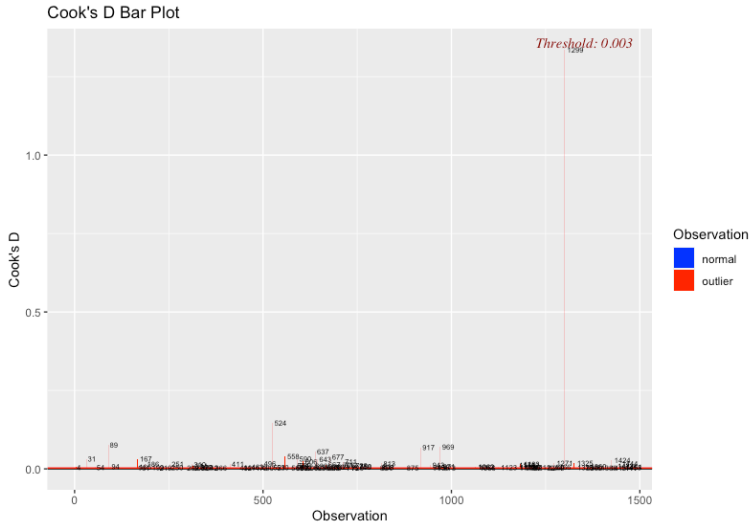
(2.16)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(paste(response, "~", paste(preds, collapse = " + ")))
633
524
1299

(2.17)


Standardized Residuals Chart
Threshold: abs(2)
Standardized Residuals
Observation

(2.18)


Cook's D Bar Plot
Threshold: 0.003
Cook's D
Observation
Observation
normal
outlier

(2.19)

```
> forward.model
```

Selection Summary

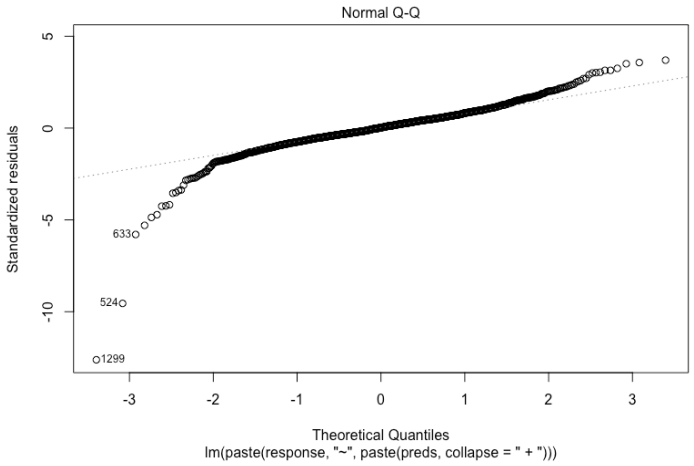
Variable	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
OverallQual	-148.980	156.765	76.036	0.67339	0.67136
Neighborhood	-561.285	177.326	55.475	0.76171	0.75619
GrLivArea	-1007.859	192.001	40.800	0.82474	0.82056
MSSubClass	-1147.798	196.434	36.367	0.84379	0.83847
OverallCond	-1260.060	199.492	33.308	0.85692	0.85121
GarageCars	-1383.757	202.240	30.561	0.86873	0.86339
YearBuilt	-1462.629	203.887	28.914	0.87580	0.87066
Fireplaces	-1545.292	205.590	27.210	0.88312	0.87802
BsmtQual	-1596.682	206.675	26.126	0.88778	0.88254
MSZoning	-1649.430	207.740	25.061	0.89235	0.88701
Heating	-1665.340	208.181	24.620	0.89424	0.88859
LotArea	-1679.711	208.455	24.346	0.89542	0.88976
YearRemodAdd	-1694.524	208.734	24.067	0.89662	0.89094
CentralAir	-1704.251	208.926	23.874	0.89745	0.89173
KitchenQual	-1713.070	209.168	23.633	0.89848	0.89259
GarageType	-1723.201	209.523	23.278	0.90001	0.89375
TotalBsmtSF	-1728.437	209.638	23.163	0.90050	0.89420
PoolYN	-1732.705	209.737	23.063	0.90093	0.89457
BsmtCond	-1735.000	209.899	22.901	0.90163	0.89508
LotConfig	-1737.196	210.059	22.742	0.90231	0.89551
FullBath	-1737.841	210.162	22.639	0.90276	0.89575

(2.20)

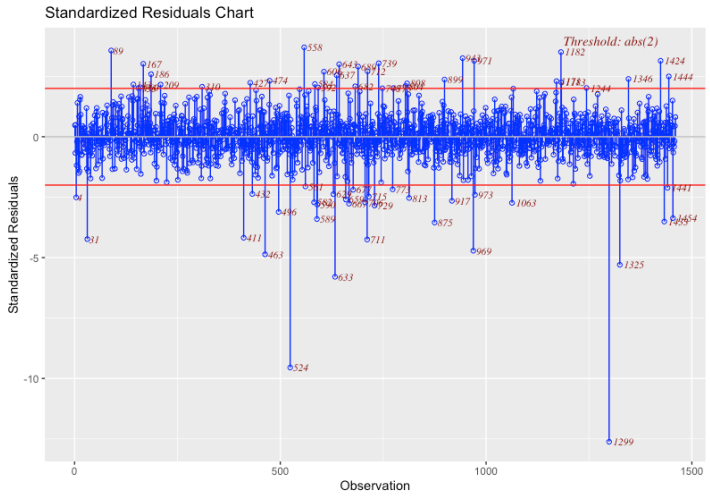
```
Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
        (Intercept)         OverallQual2         OverallQual3         OverallQual4         OverallQual5         OverallQual6
          5.649e+00            2.645e-01            3.628e-01            4.272e-01            4.612e-01            5.009e-01
       OverallQual7         OverallQual8         OverallQual9        OverallQual10  NeighborhoodBlueste   NeighborhoodBrDale
          5.595e-01            6.336e-01            7.236e-01            6.408e-01           -5.196e-03            9.901e-03
NeighborhoodBrkSide  NeighborhoodClearCr  NeighborhoodCollgCr  NeighborhoodCrawfor  NeighborhoodEdwards  NeighborhoodGilbert
          5.949e-02            1.189e-01            4.747e-02            1.553e-01           -2.326e-02            1.872e-02
 NeighborhoodIDOTRR  NeighborhoodMeadowV  NeighborhoodMitchel     NeighborhoodNAmes  NeighborhoodNoRidge  NeighborhoodNPkVill
          2.647e-02           -1.026e-01            2.140e-03            1.177e-02            1.555e-01            4.217e-03
NeighborhoodNridgHt  NeighborhoodNWAmes  NeighborhoodOldTown  NeighborhoodSawyer  NeighborhoodSawyerW  NeighborhoodSomerst
          1.464e-01           -1.521e-02           -3.865e-02           -1.230e-02            2.994e-02            6.066e-02
NeighborhoodStoneBr  NeighborhoodSWISU  NeighborhoodTimber  NeighborhoodVeenker            GrLivArea          MSSubClass30
          1.904e-01            3.167e-02            5.488e-02            1.268e-01            2.186e-04           -1.096e-01
       MSSubClass40         MSSubClass45         MSSubClass50         MSSubClass60         MSSubClass70         MSSubClass75
          -5.785e-02           -6.695e-02           -3.817e-02           -3.119e-02           -2.327e-02            2.353e-02
       MSSubClass80         MSSubClass85         MSSubClass90        MSSubClass120        MSSubClass160        MSSubClass180
          6.449e-03           -5.114e-03           -4.499e-02           -5.784e-02           -1.632e-01           -6.969e-02
      MSSubClass190        OverallCond2         OverallCond3         OverallCond4         OverallCond5         OverallCond6
          -1.020e-02           -5.522e-01           -7.127e-01           -5.766e-01           -5.563e-01           -4.980e-01
       OverallCond7         OverallCond8         OverallCond9           GarageCars            YearBuilt          Fireplaces1
          -4.640e-01           -4.582e-01           -4.183e-01            6.669e-02            1.732e-03            4.782e-02
        Fireplaces2          Fireplaces3           BsmtQualFa           BsmtQualGd         BsmtQualNONE           BsmtQualTA
          1.096e-01           -2.226e-01           -8.312e-02           -6.225e-02           -1.477e-01           -8.411e-02
         MSZoningFV           MSZoningRH           MSZoningRL           MSZoningRM           HeatingGasA          HeatingGasW
          4.646e-01            3.969e-01            3.990e-01            3.585e-01            1.461e-01            2.409e-01
        HeatingGrav          HeatingOthW          HeatingWall               LotArea          YearRemodAdd          CentralAirY
          2.396e-03            1.252e-01            2.416e-01            1.322e-06            8.993e-04            7.172e-02
      KitchenQualFa        KitchenQualGd        KitchenQualTA     GarageTypeAttchd     GarageTypeBasment    GarageTypeBuiltIn
          -1.012e-01           -6.480e-02           -8.604e-02            1.519e-01            1.234e-01            1.239e-01
  GarageTypeCarPort     GarageTypeDetchd      GarageTypeNONE          TotalBsmtSF           PoolYNYES          BsmtCondGd
          4.918e-02            1.223e-01            1.127e-01            5.080e-05           -1.268e-01            5.486e-02
       BsmtCondNONE          BsmtCondPo          BsmtCondTA    LotConfigCulDSac        LotConfigFR2         LotConfigFR3
                 NA           -3.020e-01            5.092e-02            4.251e-02           -2.469e-02           -3.641e-02
   LotConfigInside            FullBath1            FullBath2            FullBath3
          2.258e-03           -3.228e-02           -9.387e-03            3.234e-02
```

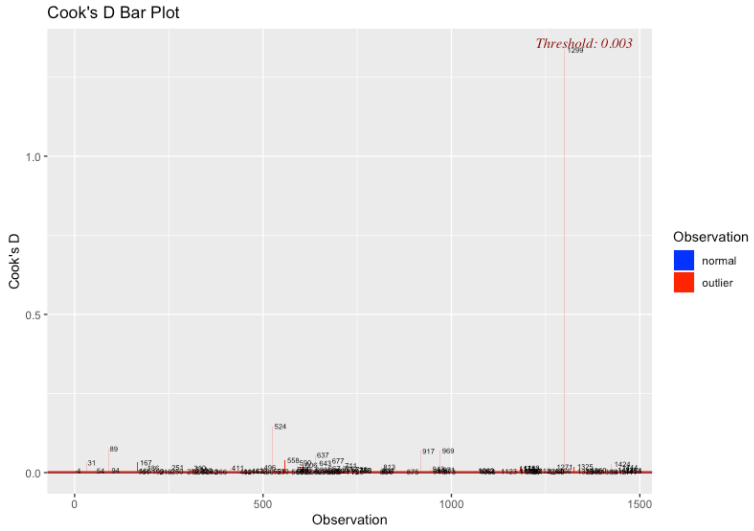
(2.21)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(paste(response, "~", paste(preds, collapse = " + ")))
633
524
1299

(2.22)


Standardized Residuals Chart
Threshold: abs(2)
Standardized Residuals
Observation

(2.23)


Cook's D Bar Plot
Threshold: 0.003
Cook's D
Observation
Observation
normal
outlier

(2.24)

```
> backward.model
```

Backward Elimination Summary

Variable	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
Full Model	-1700.261	22.081	210.720	0.90515	0.89548
MoSold	-1712.833	22.224	210.577	0.90454	0.89567
HouseStyle	-1720.795	22.316	210.485	0.90414	0.89578
ExterQual	-1726.111	22.326	210.474	0.90410	0.89597
Foundation	-1730.239	22.416	210.384	0.90371	0.89594
MasVnrArea	-1732.233	22.417	210.384	0.90371	0.89601
X1stFlrSF	-1734.123	22.418	210.382	0.90370	0.89608
TotRmsAbvGrd	-1735.787	22.423	210.377	0.90368	0.89613
YrSold	-1737.090	22.526	210.274	0.90324	0.89596
ExterCond	-1737.841	22.639	210.162	0.90276	0.89575

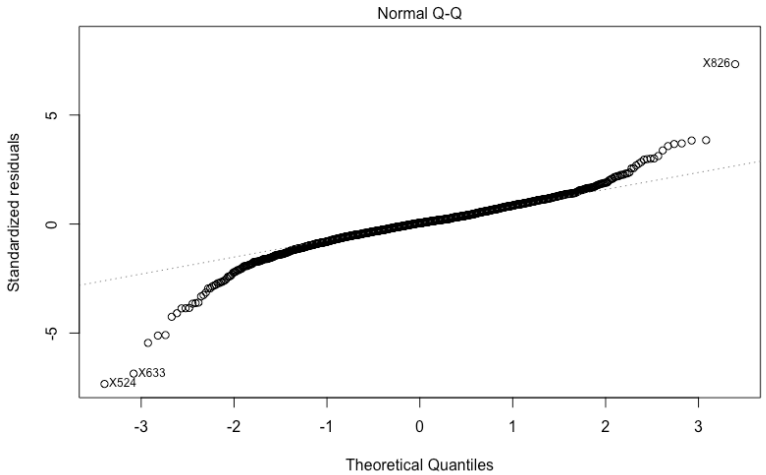
(2.25)

```
> backward.model$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
        (Intercept)         MSSubClass30         MSSubClass40         MSSubClass45         MSSubClass50         MSSubClass60
          5.649e+00           -1.096e-01           -5.785e-02           -6.695e-02           -3.817e-02           -3.119e-02
       MSSubClass70         MSSubClass75         MSSubClass80         MSSubClass85         MSSubClass90        MSSubClass120
          -2.327e-02           2.353e-02           6.449e-03           -5.114e-03           -4.499e-02           -5.784e-02
      MSSubClass160        MSSubClass180        MSSubClass190          MSZoningFV          MSZoningRH          MSZoningRL
          -1.632e-01           -6.969e-02           -1.020e-02           4.646e-01           3.969e-01           3.990e-01
         MSZoningRM              LotArea    LotConfigCulDSac          LotConfigFR2        LotConfigFR3       LotConfigInside
          3.585e-01           1.322e-06           4.251e-02           -2.469e-02           -3.641e-02           2.258e-03
NeighborhoodBlueste  NeighborhoodBrDale  NeighborhoodBrkSide  NeighborhoodClearCr  NeighborhoodCollgCr  NeighborhoodCrawfor
          -5.196e-03           9.901e-03           5.949e-02           1.189e-01           4.747e-02           1.553e-01
NeighborhoodEdwards  NeighborhoodGilbert   NeighborhoodIDOTRR  NeighborhoodMeadowV  NeighborhoodMitchel    NeighborhoodNAmes
          -2.326e-02           1.872e-02           2.647e-02           -1.026e-01           2.140e-03           1.177e-02
NeighborhoodNoRidge  NeighborhoodNPkVill  NeighborhoodNridgHt  NeighborhoodNWAmes  NeighborhoodOldTown  NeighborhoodSawyer
          1.555e-01           4.217e-03           1.464e-01           -1.521e-02           -3.865e-02           -1.230e-02
NeighborhoodSawyerW  NeighborhoodSomerst  NeighborhoodStoneBr  NeighborhoodSWISU  NeighborhoodTimber  NeighborhoodVeenker
          2.994e-02           6.066e-02           1.904e-01           3.167e-02           5.488e-02           1.268e-01
       OverallQual2         OverallQual3         OverallQual4         OverallQual5         OverallQual6         OverallQual7
          2.645e-01           3.628e-01           4.272e-01           4.612e-01           5.009e-01           5.595e-01
       OverallQual8         OverallQual9        OverallQual10         OverallCond2         OverallCond3         OverallCond4
          6.336e-01           7.236e-01           6.408e-01           -5.522e-01           -7.127e-01           -5.766e-01
       OverallCond5         OverallCond6         OverallCond7         OverallCond8         OverallCond9            YearBuilt
          -5.563e-01           -4.980e-01           -4.640e-01           -4.582e-01           -4.183e-01           1.732e-03
       YearRemodAdd           BsmtQualFa           BsmtQualGd         BsmtQualNONE           BsmtQualTA           BsmtCondGd
          8.993e-04           -8.312e-02           -6.225e-02           -1.477e-01           -8.411e-02           5.486e-02
       BsmtCondNONE           BsmtCondPo           BsmtCondTA          TotalBsmtSF          HeatingGasA          HeatingGasW
                 NA           -3.020e-01           5.092e-02           5.080e-05           1.461e-01           2.409e-01
        HeatingGrav          HeatingOthW          HeatingWall          CentralAirY            GrLivArea            FullBath1
          2.396e-03           1.252e-01           2.416e-01           7.172e-02           2.186e-04           -3.228e-02
          FullBath2            FullBath3       KitchenQualFa        KitchenQualGd        KitchenQualTA          Fireplaces1
          -9.387e-03           3.234e-02           -1.012e-01          -6.480e-02           -8.604e-02           4.782e-02
         Fireplaces2          Fireplaces3    GarageTypeAttchd    GarageTypeBasment    GarageTypeBuiltIn    GarageTypeCarPort
          1.096e-01           -2.226e-01           1.519e-01           1.234e-01           1.239e-01           4.918e-02
  GarageTypeDetchd      GarageTypeNONE          GarageCars           PoolYNYES
          1.223e-01           1.127e-01           6.669e-02           -1.268e-01
```

(2.26)

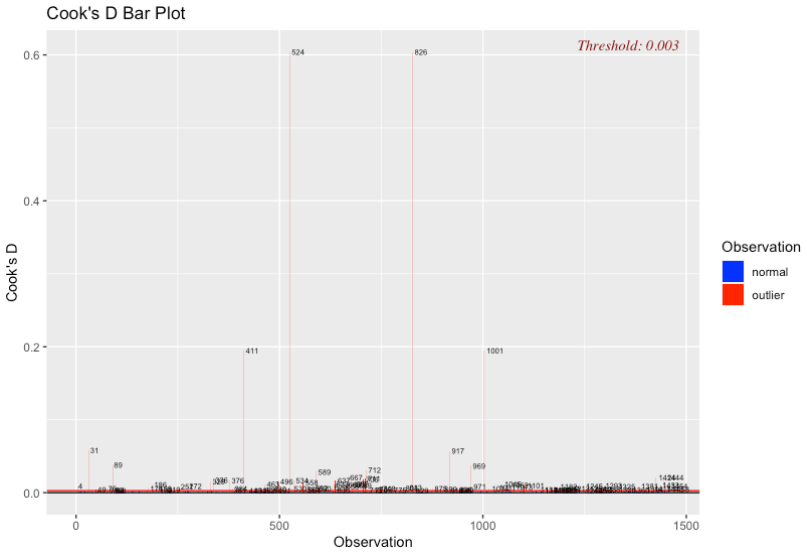


lm(.outcome ~ MSSubClass30 + MSSubClass50 + MSSubClass60 + MSSubClass70 + M ...

(2.27)


Standardized Residuals Chart
Threshold: abs(2)
Standardized Residuals
Observation
8
4
0
-4
-8
0
500
1000
1500

(2.28)


Cook's D Bar Plot
Threshold: 0.003
Cook's D
Observation
0.6
0.4
0.2
0.0
0
500
1000
1500
Observation
normal
outlier

(2.29)

```
> custom.model$finalModel

Call:
lm(formula = .outcome ~ MSSubClass30 + MSSubClass50 + MSSubClass60 +
    MSSubClass70 + MSSubClass120 + MSSubClass160 + MSZoningFV +
    MSZoningRH + MSZoningRL + MSZoningRM + LotArea + LotConfigCulDSac +
    LotConfigFR2 + LotConfigFR3 + LotConfigInside + NeighborhoodBrkSide +
    NeighborhoodCrawfor + NeighborhoodEdwards + NeighborhoodMeadowV +
    NeighborhoodMitchel + NeighborhoodNAmes + NeighborhoodNoRidge +
    NeighborhoodNridgHt + NeighborhoodOldTown + NeighborhoodSawyer +
    NeighborhoodStoneBr + NeighborhoodVeenker + Condition2PosN +
    BldgTypeTwnhs + OverallQual2 + OverallQual3 + OverallQual4 +
    OverallQual5 + OverallQual6 + OverallQual7 + OverallQual8 +
    OverallQual9 + OverallQual10 + YearBuilt + YearRemodAdd +
    RoofStyleGable + RoofMatlCompShg + RoofMatlMembran + RoofMatlMetal +
    RoofMatlRoll + `RoofMatlTar&Grv` + RoofMatlWdShake + RoofMatlWdShngl +
    Exterior1stBrkComm + Exterior1stBrkFace + Exterior1stStucco +
    `Exterior1stWd Sdng` + `Exterior2ndBrk Cmn` + Exterior2ndBrkFace +
    Exterior2ndCmentBd + Exterior2ndMetalSd + Exterior2ndStucco +
    `Exterior2ndWd Sdng` + `Exterior2ndWd Shng` + MasVnrTypeBrkFace +
    MasVnrTypeNone + MasVnrTypeStone + FoundationPConc + FoundationStone +
    BsmtQualFa + BsmtQualGd + BsmtQualTA + BsmtFinType1LwQ +
    BsmtFinType1Unf + TotalBsmtSF + HeatingGasA + HeatingGasW +
    HeatingWall + HeatingQCFa + HeatingQCGd + HeatingQCTA + CentralAirY +
    ElectricalMix + X2ndFlrSF + GrLivArea + BsmtFullBath + FullBath2 +
    FullBath3 + HalfBath + BedroomAbvGr + KitchenAbvGr + KitchenQualFa +
    KitchenQualGd + KitchenQualTA + Fireplaces1 + Fireplaces2 +
    Fireplaces3 + GarageCars + GarageArea + PavedDriveP + PavedDriveY +
    WoodDeckSF + ScreenPorch + MoSold5 + MoSold6 + MoSold7 +
    YrSold2009, data = dat)
```

(2.30)

```
Coefficients:
       (Intercept)         MSSubClass30         MSSubClass50         MSSubClass60         MSSubClass70        MSSubClass120
         1.278e+00           -7.853e-02           -5.284e-02           -5.970e-02           -3.530e-02           -6.273e-02
       MSSubClass160          MSZoningFV          MSZoningRH          MSZoningRL          MSZoningRM              LotArea
         -1.881e-01           4.924e-01            4.100e-01            4.212e-01            3.842e-01            2.184e-06
    LotConfigCulDSac        LotConfigFR2        LotConfigFR3       LotConfigInside   NeighborhoodBrkSide   NeighborhoodCrawfor
          3.638e-02           -3.969e-02           -9.211e-02           -1.161e-02            5.490e-02            1.497e-01
 NeighborhoodEdwards  NeighborhoodMeadowV   NeighborhoodMitchel      NeighborhoodNAmes   NeighborhoodNoRidge   NeighborhoodNridgHt
         -6.039e-02           -1.337e-01           -3.791e-02           -2.354e-02            7.730e-02            8.145e-02
NeighborhoodOldTown   NeighborhoodSawyer   NeighborhoodStoneBr   NeighborhoodVeenker       Condition2PosN        BldgTypeTwnhs
         -3.898e-02           -3.065e-02            1.195e-01            8.597e-02           -7.737e-01           -3.992e-02
        OverallQual2         OverallQual3         OverallQual4         OverallQual5         OverallQual6         OverallQual7
          2.162e-01            2.813e-01            3.804e-01            4.258e-01            4.670e-01            5.112e-01
        OverallQual8         OverallQual9        OverallQual10            YearBuilt         YearRemodAdd        RoofStyleGable
          5.758e-01            6.494e-01            7.429e-01            1.012e-03            1.721e-03           -1.473e-02
     RoofMatlCompShg      RoofMatlMembran       RoofMatlMetal         RoofMatlRoll     `RoofMatlTar&Grv`      RoofMatlWdShake
          1.965e+00            2.133e+00            2.067e+00            1.867e+00            2.002e+00            2.021e+00
     RoofMatlWdShngl   Exterior1stBrkComm   Exterior1stBrkFace   Exterior1stStucco `Exterior1stWd Sdng`  `Exterior2ndBrk Cmn`
          2.066e+00           -4.717e-01            8.557e-02            6.966e-02           -3.493e-02            1.039e-01
 Exterior2ndBrkFace   Exterior2ndCmentBd   Exterior2ndMetalSd   Exterior2ndStucco `Exterior2ndWd Sdng`  `Exterior2ndWd Shng`
         -4.911e-02            3.236e-02            1.788e-02           -6.451e-02            3.934e-02           -3.657e-02
   MasVnrTypeBrkFace       MasVnrTypeNone      MasVnrTypeStone      FoundationPConc      FoundationStone            BsmtQualFa
          7.977e-02            7.325e-02            1.039e-01            1.705e-02            1.515e-01           -4.066e-02
         BsmtQualGd           BsmtQualTA       BsmtFinType1LwQ      BsmtFinType1Unf         TotalBsmtSF            HeatingGasA
         -4.308e-02           -4.360e-02           -3.800e-02           -5.903e-02            1.212e-04            1.346e-01
         HeatingGasW          HeatingWall          HeatingQCFa          HeatingQCGd          HeatingQCTA          CentralAirY
          1.982e-01            2.326e-01           -3.391e-02           -1.525e-02           -3.306e-02            7.757e-02
      ElectricalMix           X2ndFlrSF            GrLivArea         BsmtFullBath           FullBath2            FullBath3
         -3.881e-01            9.193e-05            2.944e-01            2.975e-02            2.517e-02            7.007e-02
           HalfBath         BedroomAbvGr         KitchenAbvGr         KitchenQualFa        KitchenQualGd        KitchenQualTA
          2.283e-02           -1.016e-02           -1.037e-01           -7.199e-02           -6.003e-02           -7.625e-02
       Fireplaces1          Fireplaces2          Fireplaces3           GarageCars           GarageArea          PavedDriveP
          2.785e-02            4.386e-02            8.510e-02            2.131e-02            1.233e-04            3.696e-02
       PavedDriveY          WoodDeckSF          ScreenPorch              MoSold5              MoSold6              MoSold7
          2.603e-02            1.108e-04            2.145e-04            2.052e-02            1.471e-02            1.733e-02
        YrSold2009
         -2.744e-02
```

(2.31)

```
> print(custom.model)
Linear Regression with Stepwise Selection

1460 samples
  49 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1313, 1313, 1315, 1316, 1313, 1314, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.1649202  0.8265013  0.09613509
```

(2.32)