Team Analysis

# ANALYSIS 1: Relationship between Ames House Prices & Area, Neighborhood

## Introduction

This analysis project is based on data about the housing market. We will build different linear regression models to find any relationship between House Prices and Living Area of the Houses in three different neighborhoods. There are 25 neighborhoods in this dataset. In this analysis we are interseted in following three neighborhoods: NAmes, Edwards and BrkSide.

## Data Description

Data set contains information from the Ames Assessor’s Office used in computing assessed values for individual residential properties sold in Ames, IA from 1872 to 2010. The Ames dataset is available on [kaggle](https://www.kaggle.com/datasets/prevek18/ames-housing-dataset) and the description of the data is available at this [link](http://jse.amstat.org/v19n3/decock/DataDocumentation.txt). The dataset is quite big with 2930 observations and 80 variables. The variable used for this analysis are SalePrice, GrLIvArea and Neighborhood. We will be exploring the relationship between SalePrice and GrLIvArea in following neighborhoods: NAmes, Edwards and BrkSide. The dataset has been partitioned equally at random into training and testing data set . The training set contains 1460 observations and testing set contains 1459 observations.

library(tidyverse)  
library(car)  
library(leaps)  
library(MASS)  
train <- read.csv('train.csv')  
test <- read.csv('test.csv')  
dim(train)

## [1] 1460 81

dim(train)

## [1] 1460 81

dim(test)

## [1] 1459 80

train %>%  
 count(Neighborhood)

## Neighborhood n  
## 1 Blmngtn 17  
## 2 Blueste 2  
## 3 BrDale 16  
## 4 BrkSide 58  
## 5 ClearCr 28  
## 6 CollgCr 150  
## 7 Crawfor 51  
## 8 Edwards 100  
## 9 Gilbert 79  
## 10 IDOTRR 37  
## 11 MeadowV 17  
## 12 Mitchel 49  
## 13 NAmes 225  
## 14 NoRidge 41  
## 15 NPkVill 9  
## 16 NridgHt 77  
## 17 NWAmes 73  
## 18 OldTown 113  
## 19 Sawyer 74  
## 20 SawyerW 59  
## 21 Somerst 86  
## 22 StoneBr 25  
## 23 SWISU 25  
## 24 Timber 38  
## 25 Veenker 11

train\_set <- train %>% filter(Neighborhood %in% c('NAmes', 'Edwards', 'BrkSide'))  
  
test\_set <- test %>% filter(Neighborhood %in% c('NAmes', 'Edwards', 'BrkSide'))

## Analysis Question 1:

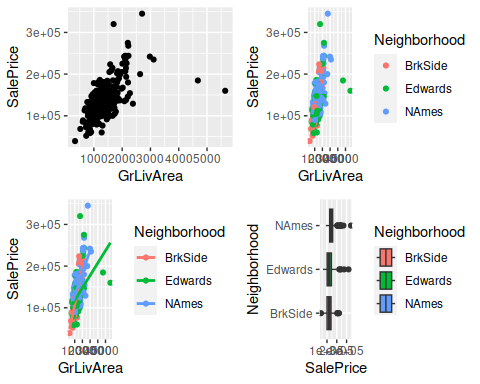
In this analysis we will be doing explanatory data analysis(EDA) and building different regression models to answer the following question: Is there are a relationship between Sale Price of the Ames house and square footage of the living area of the house in the three neighborhoods of NAmes, Edwards and BrkSide?

## Exploratory Data Analysis (EDA)

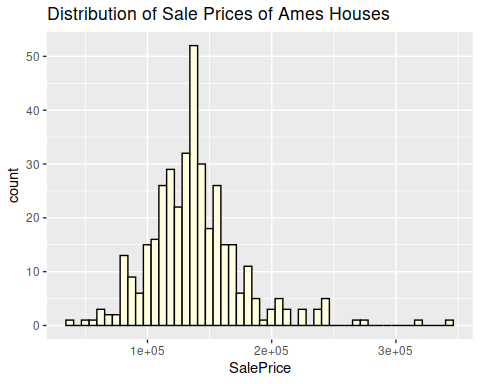
We will perform exploratory data analysis to answer the above question. The scatterplot of Sales Price and Living Area suggest a linear positive relationship between these two variables. The boxplot between Sale price and those three neighborhood suggest that NAmes has higher Sale prices compare to other two neighborhoods. The neighborhood BrkSide has lowest Sale price than other two neighborhoods. The oulier sale prices are visible in all three neighborhoods. The histogram of SalePrice suggest that the distrubtion is slightly right skewed with some extreme sale prices observations.

p1 <-ggplot(data = train\_set, aes(GrLivArea, SalePrice)) +  
 geom\_point()   
  
p2 <- ggplot(data = train\_set, aes(GrLivArea, SalePrice, color = Neighborhood)) +  
 geom\_point()   
  
  
p3 <- ggplot(data = train\_set, aes(GrLivArea, SalePrice, color = Neighborhood)) +  
 geom\_point() +  
 geom\_smooth(method = 'lm', se = FALSE)  
  
p4 <- ggplot(data = train\_set, aes(SalePrice, Neighborhood, fill = Neighborhood)) +  
 geom\_boxplot()   
  
gridExtra::grid.arrange(p1, p2, p3, p4 , nrow = 2, ncol = 2)

## `geom\_smooth()` using formula = 'y ~ x'



ggplot(train\_set, aes(SalePrice))+  
 geom\_histogram(bins = 50, fill = 'lightyellow', color = 'black') +  
 ggtitle('Distribution of Sale Prices of Ames Houses')



The coefficient of correlation between sale price and living area is 0.585 which suggest relationship is positive. We will perform cor test see whether this relationship is significant or not. The results from the cor.test are statistically significant which means there is a relationship between sale prices and living area.

cor(train\_set$SalePrice, train\_set$GrLivArea)

## [1] 0.585059

cor.test(train\_set$SalePrice, train\_set$GrLivArea)

##   
## Pearson's product-moment correlation  
##   
## data: train\_set$SalePrice and train\_set$GrLivArea  
## t = 14.081, df = 381, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5150480 0.6473156  
## sample estimates:  
## cor   
## 0.585059

## Building Linear Regression model

In our first model we fitted a very simple model between Sale Price and Living area. The coefficient of GrLivArea is statistically significant and ajdusted r-squared is 0.34 which means 34% varibility of SalePrice is explained by GrLivArea.

price\_fit1 <- lm(SalePrice ~ GrLivArea, data = train\_set)  
summary(price\_fit1)

##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea, data = train\_set)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -177619 -17918 919 15227 163722   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 78205.578 4536.054 17.24 <2e-16 \*\*\*  
## GrLivArea 45.979 3.265 14.08 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 30980 on 381 degrees of freedom  
## Multiple R-squared: 0.3423, Adjusted R-squared: 0.3406   
## F-statistic: 198.3 on 1 and 381 DF, p-value: < 2.2e-16

In our next model we added those three neighborhoods to see whether they have any impact on the sale prices of the house. From the summary of this model we can see that coefficient of Neighborhood are statistically significant and adjusted r-squared is improved. We will perform anova test to see whether our second model is better fit than first. The results from the anova test are statistically significant which means there is relationship between sale prices and different neighborhoods.

price\_fit2 <- lm(SalePrice ~ GrLivArea + Neighborhood, data = train\_set)  
summary(price\_fit2)

##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea + Neighborhood, data = train\_set)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -165078 -16215 281 13578 175400   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69781.538 5442.400 12.822 < 2e-16 \*\*\*  
## GrLivArea 45.760 3.149 14.533 < 2e-16 \*\*\*  
## NeighborhoodEdwards -2882.155 4930.632 -0.585 0.559204   
## NeighborhoodNAmes 16105.621 4395.352 3.664 0.000283 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29760 on 379 degrees of freedom  
## Multiple R-squared: 0.3965, Adjusted R-squared: 0.3917   
## F-statistic: 83 on 3 and 379 DF, p-value: < 2.2e-16

anova(price\_fit1, price\_fit2)

## Analysis of Variance Table  
##   
## Model 1: SalePrice ~ GrLivArea  
## Model 2: SalePrice ~ GrLivArea + Neighborhood  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 381 3.6578e+11   
## 2 379 3.3564e+11 2 3.014e+10 17.017 8.375e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Finally we built a model with an interaction between GrLivArea and Neighborhood see any combined effect of these two variables on sale prices. Summary of the model suggest that interaction has significant effect on sales prices. The adjusted r-squared has improved in our final model to 0.447 which means 44.7% variability of the response SalePrice is explained these two variables. We will perform anova test whether our final model is better than our previous model. The results from the anova test are statistically significant which means our final model with interaction is best fit.

price\_fit3 <- lm(SalePrice ~ GrLivArea\*Neighborhood, data = train\_set)  
summary(price\_fit3)

##   
## Call:  
## lm(formula = SalePrice ~ GrLivArea \* Neighborhood, data = train\_set)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -96204 -14568 -310 12601 181131   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19971.514 12351.125 1.617 0.10672   
## GrLivArea 87.163 9.782 8.911 < 2e-16 \*\*\*  
## NeighborhoodEdwards 68381.591 13969.511 4.895 1.46e-06 \*\*\*  
## NeighborhoodNAmes 54704.888 13882.334 3.941 9.69e-05 \*\*\*  
## GrLivArea:NeighborhoodEdwards -57.412 10.718 -5.357 1.48e-07 \*\*\*  
## GrLivArea:NeighborhoodNAmes -32.847 10.815 -3.037 0.00256 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 28550 on 377 degrees of freedom  
## Multiple R-squared: 0.4474, Adjusted R-squared: 0.44   
## F-statistic: 61.04 on 5 and 377 DF, p-value: < 2.2e-16

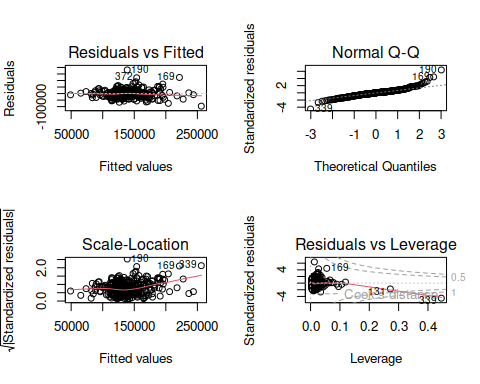
confint(price\_fit3)

## 2.5 % 97.5 %  
## (Intercept) -4314.21151 44257.23910  
## GrLivArea 67.92850 106.39657  
## NeighborhoodEdwards 40913.67039 95849.51158  
## NeighborhoodNAmes 27408.38290 82001.39258  
## GrLivArea:NeighborhoodEdwards -78.48612 -36.33834  
## GrLivArea:NeighborhoodNAmes -54.11269 -11.58065

## Diagnostics

The fitted vs residuals plot doesn’t show any patterns. The residuals look normally distributed. The studentized residuals vs fitted values suggest there are couple of observations are outliers.

par(mfrow = c(2, 2))  
plot(price\_fit3)



dev.off()

## null device   
## 1

We have performed ncv Test to check the variance of the residuals. The result of test are significant which means there variance of the residuals are not constant.

ncvTest(price\_fit3)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 69.34626, Df = 1, p = < 2.22e-16

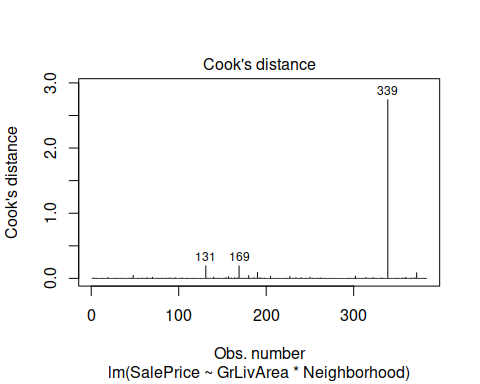
Finally we performed an outlier test to check the presence of any outliers in the residuals. The results from the test show that the observation 190, 229, 169 and 372 are outliers.

outlierTest(price\_fit3)

## rstudent unadjusted p-value Bonferroni p  
## 190 6.752661 5.5249e-11 2.1161e-08  
## 339 -4.646540 4.6764e-06 1.7911e-03  
## 169 4.563657 6.8174e-06 2.6111e-03  
## 372 4.405852 1.3761e-05 5.2704e-03

There are presence of influential observations. The plot below suggest that the observation 131, 169 and 339 are influential.

plot(price\_fit3, 4)



## Conclusion

We fitted multiple linear regression model to observe the relationship between the Sale Prices of Ames housing and Living Area in three neighborhoods. We have noticed that the relationship between sale prices and living area of the house is significant. The multiple linear regression model was able to explain 44% variability of the sale prices. The sales prices of the housing are related to area and neighborhood. Names have significantly higher Sale prices compare to other 2 neighborhoods. Overall the model was not a good fit. The variance was not constant and there presence of outlier in the residuals and there are couple of influential points.

# ANALYSIS 2: Predicting Sales Prices of Ames Iowa Houses

## Analysis Question 2

We will build multiple linear regression models to answer the following question: What are the sales prices of ames houses using different relevant parameters.

## Cleaning Data

The data set is huge and contains lot missing values for the variables. I’ll remove those variable from the data set first.

train <- train %>%  
 select\_if(~ !any(is.na(.)))

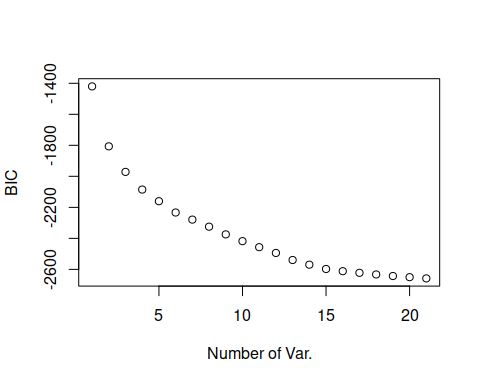
## Forward Selection Model

Our first model is forward selection model using regsubsets method. We will choose the best model based on Bayesian information criterion(BIC). The model with lowest BIC value will be the best model here. Our model based on forward subset selection method contains 18 predictors. Adjusted r-squared value of the model is 0.8226 which is pretty good which means 82% variability of the Sale Price is explained by this model.

set.seed(123)  
forward\_model <- regsubsets(SalePrice ~ ., data = train, nvmax = 20, method = "forward")

## Reordering variables and trying again:

plot(summary(forward\_model)$bic, xlab="Number of Var.", ylab = "BIC")



coef(forward\_model, which.min(summary(forward\_model)$bic))

## (Intercept) MSSubClass LotArea   
## -7.662740e+05 -3.113764e+02 1.117732e+00   
## LotShapeIR3 NeighborhoodCrawfor NeighborhoodNoRidge   
## -2.689683e+04 3.839663e+04 8.168489e+04   
## NeighborhoodNridgHt NeighborhoodStoneBr Condition1Norm   
## 6.588392e+04 7.825019e+04 3.723630e+03   
## Condition2PosN BldgType2fmCon OverallQual   
## -8.632146e+04 3.977891e+04 2.656121e+04   
## OverallCond YearBuilt RoofMatlWdShngl   
## 5.228468e+03 3.546448e+02 1.122659e+05   
## Exterior2ndHdBoard BsmtFinSF2 FullBath   
## -9.573793e+03 1.301091e+01 2.754496e+04   
## HalfBath KitchenQualGd PavedDriveY   
## 1.706036e+04 -8.868306e+03 5.384316e+03   
## SaleConditionAdjLand   
## -1.389608e+04

fit\_forward <- lm(SalePrice ~ MSSubClass + LotArea + LotShape + Neighborhood +   
 Condition1 + Condition2 + BldgType + OverallQual +  
 OverallCond + YearBuilt + RoofMatl + Exterior2nd + BsmtFinSF2 +  
 FullBath + FullBath + KitchenQual + PavedDrive + SaleCondition,   
 data = train)  
  
summary(fit\_forward)

##   
## Call:  
## lm(formula = SalePrice ~ MSSubClass + LotArea + LotShape + Neighborhood +   
## Condition1 + Condition2 + BldgType + OverallQual + OverallCond +   
## YearBuilt + RoofMatl + Exterior2nd + BsmtFinSF2 + FullBath +   
## FullBath + KitchenQual + PavedDrive + SaleCondition, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -109998 -18011 -1243 15277 273254   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.953e+05 1.726e+05 -5.186 2.47e-07 \*\*\*  
## MSSubClass 6.082e+01 4.805e+01 1.266 0.205879   
## LotArea 9.744e-01 1.131e-01 8.612 < 2e-16 \*\*\*  
## LotShapeIR2 5.947e+03 5.916e+03 1.005 0.314968   
## LotShapeIR3 5.585e+03 1.234e+04 0.452 0.651020   
## LotShapeReg -3.977e+03 2.207e+03 -1.802 0.071737 .   
## NeighborhoodBlueste -1.143e+03 2.643e+04 -0.043 0.965508   
## NeighborhoodBrDale -4.812e+03 1.387e+04 -0.347 0.728760   
## NeighborhoodBrkSide -1.002e+04 1.166e+04 -0.860 0.390076   
## NeighborhoodClearCr -5.307e+03 1.239e+04 -0.428 0.668536   
## NeighborhoodCollgCr -1.962e+04 9.763e+03 -2.010 0.044647 \*   
## NeighborhoodCrawfor 2.429e+04 1.142e+04 2.127 0.033613 \*   
## NeighborhoodEdwards -1.814e+04 1.060e+04 -1.711 0.087253 .   
## NeighborhoodGilbert -3.130e+04 1.049e+04 -2.985 0.002886 \*\*   
## NeighborhoodIDOTRR -2.156e+04 1.224e+04 -1.762 0.078331 .   
## NeighborhoodMeadowV -6.161e+03 1.446e+04 -0.426 0.670189   
## NeighborhoodMitchel -1.617e+04 1.094e+04 -1.477 0.139876   
## NeighborhoodNAmes -1.233e+04 1.038e+04 -1.188 0.235122   
## NeighborhoodNoRidge 6.890e+04 1.114e+04 6.184 8.21e-10 \*\*\*  
## NeighborhoodNPkVill -4.619e+03 1.772e+04 -0.261 0.794422   
## NeighborhoodNridgHt 3.347e+04 9.987e+03 3.352 0.000824 \*\*\*  
## NeighborhoodNWAmes -1.910e+04 1.081e+04 -1.766 0.077540 .   
## NeighborhoodOldTown -2.128e+04 1.125e+04 -1.892 0.058686 .   
## NeighborhoodSawyer -1.315e+04 1.097e+04 -1.199 0.230822   
## NeighborhoodSawyerW -9.949e+03 1.057e+04 -0.941 0.346907   
## NeighborhoodSomerst -1.266e+04 9.850e+03 -1.285 0.198905   
## NeighborhoodStoneBr 4.740e+04 1.148e+04 4.129 3.86e-05 \*\*\*  
## NeighborhoodSWISU -1.072e+04 1.276e+04 -0.840 0.401030   
## NeighborhoodTimber -1.079e+04 1.118e+04 -0.965 0.334702   
## NeighborhoodVeenker 1.533e+04 1.448e+04 1.058 0.290105   
## Condition1Feedr 6.886e+03 6.926e+03 0.994 0.320300   
## Condition1Norm 8.965e+03 5.679e+03 1.579 0.114649   
## Condition1PosA 1.568e+04 1.400e+04 1.120 0.262798   
## Condition1PosN 2.062e+04 1.034e+04 1.994 0.046362 \*   
## Condition1RRAe -1.030e+04 1.230e+04 -0.837 0.402487   
## Condition1RRAn 9.427e+03 9.465e+03 0.996 0.319433   
## Condition1RRNe -9.345e+03 2.567e+04 -0.364 0.715881   
## Condition1RRNn 2.688e+03 1.727e+04 0.156 0.876352   
## Condition2Feedr -3.362e+04 3.037e+04 -1.107 0.268549   
## Condition2Norm -1.880e+04 2.623e+04 -0.717 0.473643   
## Condition2PosA 3.949e+04 4.376e+04 0.903 0.366933   
## Condition2PosN -1.642e+05 3.775e+04 -4.351 1.46e-05 \*\*\*  
## Condition2RRAe 2.135e+04 4.446e+04 0.480 0.631215   
## Condition2RRAn -2.575e+04 4.414e+04 -0.583 0.559751   
## Condition2RRNn -3.863e+04 3.637e+04 -1.062 0.288292   
## BldgType2fmCon -1.146e+04 9.621e+03 -1.191 0.233776   
## BldgTypeDuplex -1.727e+04 6.317e+03 -2.733 0.006349 \*\*   
## BldgTypeTwnhs -5.695e+04 8.901e+03 -6.398 2.15e-10 \*\*\*  
## BldgTypeTwnhsE -4.501e+04 6.431e+03 -6.999 4.01e-12 \*\*\*  
## OverallQual 2.162e+04 1.183e+03 18.274 < 2e-16 \*\*\*  
## OverallCond 3.372e+03 9.919e+02 3.400 0.000694 \*\*\*  
## YearBuilt 3.480e+02 8.241e+01 4.223 2.57e-05 \*\*\*  
## RoofMatlCompShg 2.610e+05 3.855e+04 6.769 1.92e-11 \*\*\*  
## RoofMatlMembran 2.739e+05 5.278e+04 5.190 2.42e-07 \*\*\*  
## RoofMatlMetal 2.626e+05 5.243e+04 5.009 6.18e-07 \*\*\*  
## RoofMatlRoll 2.575e+05 5.246e+04 4.909 1.02e-06 \*\*\*  
## RoofMatlTar&Grv 2.600e+05 4.008e+04 6.488 1.21e-10 \*\*\*  
## RoofMatlWdShake 2.839e+05 4.187e+04 6.780 1.77e-11 \*\*\*  
## RoofMatlWdShngl 3.679e+05 4.091e+04 8.992 < 2e-16 \*\*\*  
## Exterior2ndAsphShn -1.302e+04 2.176e+04 -0.598 0.549742   
## Exterior2ndBrk Cmn -4.887e+03 1.921e+04 -0.254 0.799266   
## Exterior2ndBrkFace -2.875e+03 1.096e+04 -0.262 0.793180   
## Exterior2ndCBlock 2.147e+04 3.595e+04 0.597 0.550342   
## Exterior2ndCmentBd 3.952e+03 1.030e+04 0.384 0.701355   
## Exterior2ndHdBoard -1.121e+04 8.730e+03 -1.285 0.199171   
## Exterior2ndImStucc 2.330e+04 1.413e+04 1.649 0.099309 .   
## Exterior2ndMetalSd -5.569e+03 8.519e+03 -0.654 0.513362   
## Exterior2ndOther -1.005e+04 3.647e+04 -0.276 0.782923   
## Exterior2ndPlywood -7.141e+03 8.948e+03 -0.798 0.424944   
## Exterior2ndStone -2.889e+04 1.801e+04 -1.603 0.109061   
## Exterior2ndStucco 1.241e+03 1.084e+04 0.114 0.908908   
## Exterior2ndVinylSd -7.078e+03 8.643e+03 -0.819 0.412997   
## Exterior2ndWd Sdng -8.695e+03 8.506e+03 -1.022 0.306831   
## Exterior2ndWd Shng -2.214e+04 9.925e+03 -2.230 0.025878 \*   
## BsmtFinSF2 1.272e+01 6.041e+00 2.106 0.035417 \*   
## FullBath 2.348e+04 2.341e+03 10.031 < 2e-16 \*\*\*  
## KitchenQualFa -6.267e+04 8.018e+03 -7.816 1.08e-14 \*\*\*  
## KitchenQualGd -5.477e+04 4.474e+03 -12.242 < 2e-16 \*\*\*  
## KitchenQualTA -6.117e+04 5.072e+03 -12.060 < 2e-16 \*\*\*  
## PavedDriveP 2.114e+03 7.493e+03 0.282 0.777912   
## PavedDriveY 6.745e+03 4.537e+03 1.487 0.137364   
## SaleConditionAdjLand -3.380e+03 1.852e+04 -0.183 0.855212   
## SaleConditionAlloca 2.326e+04 1.134e+04 2.050 0.040526 \*   
## SaleConditionFamily 6.838e+02 8.651e+03 0.079 0.937014   
## SaleConditionNormal 8.357e+03 3.717e+03 2.248 0.024719 \*   
## SaleConditionPartial 2.710e+04 5.273e+03 5.140 3.14e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 34480 on 1374 degrees of freedom  
## Multiple R-squared: 0.8226, Adjusted R-squared: 0.8116   
## F-statistic: 74.96 on 85 and 1374 DF, p-value: < 2.2e-16

confint(fit\_forward)

## 2.5 % 97.5 %  
## (Intercept) -1.233920e+06 -5.566403e+05  
## MSSubClass -3.345067e+01 1.550813e+02  
## LotArea 7.524319e-01 1.196338e+00  
## LotShapeIR2 -5.658147e+03 1.755119e+04  
## LotShapeIR3 -1.862938e+04 2.979886e+04  
## LotShapeReg -8.306440e+03 3.520365e+02  
## NeighborhoodBlueste -5.299256e+04 5.070622e+04  
## NeighborhoodBrDale -3.202663e+04 2.240297e+04  
## NeighborhoodBrkSide -3.288681e+04 1.284426e+04  
## NeighborhoodClearCr -2.961753e+04 1.900327e+04  
## NeighborhoodCollgCr -3.877278e+04 -4.697479e+02  
## NeighborhoodCrawfor 1.885631e+03 4.668741e+04  
## NeighborhoodEdwards -3.892549e+04 2.653704e+03  
## NeighborhoodGilbert -5.187626e+04 -1.073101e+04  
## NeighborhoodIDOTRR -4.556760e+04 2.446649e+03  
## NeighborhoodMeadowV -3.453362e+04 2.221146e+04  
## NeighborhoodMitchel -3.763339e+04 5.303173e+03  
## NeighborhoodNAmes -3.268467e+04 8.031399e+03  
## NeighborhoodNoRidge 4.704707e+04 9.076085e+04  
## NeighborhoodNPkVill -3.938244e+04 3.014524e+04  
## NeighborhoodNridgHt 1.388361e+04 5.306521e+04  
## NeighborhoodNWAmes -4.030877e+04 2.110767e+03  
## NeighborhoodOldTown -4.334286e+04 7.825476e+02  
## NeighborhoodSawyer -3.468099e+04 8.371777e+03  
## NeighborhoodSawyerW -3.069128e+04 1.079320e+04  
## NeighborhoodSomerst -3.198236e+04 6.662395e+03  
## NeighborhoodStoneBr 2.488456e+04 6.992260e+04  
## NeighborhoodSWISU -3.574003e+04 1.430819e+04  
## NeighborhoodTimber -3.272923e+04 1.114564e+04  
## NeighborhoodVeenker -1.308496e+04 4.374235e+04  
## Condition1Feedr -6.701280e+03 2.047388e+04  
## Condition1Norm -2.175304e+03 2.010496e+04  
## Condition1PosA -1.178000e+04 4.314728e+04  
## Condition1PosN 3.328211e+02 4.089792e+04  
## Condition1RRAe -3.443715e+04 1.383131e+04  
## Condition1RRAn -9.140060e+03 2.799333e+04  
## Condition1RRNe -5.970057e+04 4.101086e+04  
## Condition1RRNn -3.118929e+04 3.656429e+04  
## Condition2Feedr -9.319894e+04 2.596327e+04  
## Condition2Norm -7.026024e+04 3.265660e+04  
## Condition2PosA -4.634810e+04 1.253363e+05  
## Condition2PosN -2.383062e+05 -9.018892e+04  
## Condition2RRAe -6.587171e+04 1.085653e+05  
## Condition2RRAn -1.123503e+05 6.084633e+04  
## Condition2RRNn -1.099676e+05 3.270700e+04  
## BldgType2fmCon -3.033342e+04 7.412609e+03  
## BldgTypeDuplex -2.965846e+04 -4.874771e+03  
## BldgTypeTwnhs -7.441602e+04 -3.949293e+04  
## BldgTypeTwnhsE -5.762596e+04 -3.239593e+04  
## OverallQual 1.929698e+04 2.393834e+04  
## OverallCond 1.426510e+03 5.318188e+03  
## YearBuilt 1.863195e+02 5.096500e+02  
## RoofMatlCompShg 1.853438e+05 3.366084e+05  
## RoofMatlMembran 1.703856e+05 3.774496e+05  
## RoofMatlMetal 1.597781e+05 3.654851e+05  
## RoofMatlRoll 1.546093e+05 3.604204e+05  
## RoofMatlTar&Grv 1.814099e+05 3.386533e+05  
## RoofMatlWdShake 2.017628e+05 3.660382e+05  
## RoofMatlWdShngl 2.876079e+05 4.481052e+05  
## Exterior2ndAsphShn -5.569947e+04 2.966443e+04  
## Exterior2ndBrk Cmn -4.257583e+04 3.280253e+04  
## Exterior2ndBrkFace -2.438384e+04 1.863341e+04  
## Exterior2ndCBlock -4.904136e+04 9.198956e+04  
## Exterior2ndCmentBd -1.626067e+04 2.416520e+04  
## Exterior2ndHdBoard -2.834084e+04 5.911846e+03  
## Exterior2ndImStucc -4.412613e+03 5.101127e+04  
## Exterior2ndMetalSd -2.228010e+04 1.114152e+04  
## Exterior2ndOther -8.159801e+04 6.149698e+04  
## Exterior2ndPlywood -2.469412e+04 1.041147e+04  
## Exterior2ndStone -6.422443e+04 6.453168e+03  
## Exterior2ndStucco -2.002962e+04 2.251130e+04  
## Exterior2ndVinylSd -2.403351e+04 9.877759e+03  
## Exterior2ndWd Sdng -2.538125e+04 7.990568e+03  
## Exterior2ndWd Shng -4.160578e+04 -2.667479e+03  
## BsmtFinSF2 8.695797e-01 2.457054e+01  
## FullBath 1.889100e+04 2.807603e+04  
## KitchenQualFa -7.840199e+04 -4.694312e+04  
## KitchenQualGd -6.354442e+04 -4.599177e+04  
## KitchenQualTA -7.111947e+04 -5.122025e+04  
## PavedDriveP -1.258569e+04 1.681335e+04  
## PavedDriveY -2.155896e+03 1.564583e+04  
## SaleConditionAdjLand -3.971177e+04 3.295152e+04  
## SaleConditionAlloca 1.004932e+03 4.551118e+04  
## SaleConditionFamily -1.628713e+04 1.765466e+04  
## SaleConditionNormal 1.065091e+03 1.564861e+04  
## SaleConditionPartial 1.675898e+04 3.744631e+04

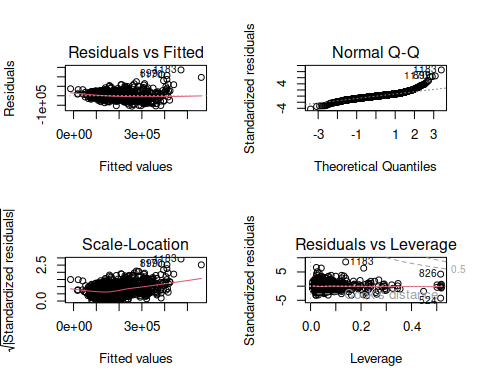
Forward\_Adjusted\_R2 <- 0.812

## Diagnostics of Selection Model

The diagnostics of this model shows that residuals are normally distributed. The ncv test results show that variance is not constant. VIF suggest that multicolinearity is present in the model. The outlier test shows that there are some observations that are outliers. The observation 524, 826 and 1183 are highly influential in this model. So our model is not a good fit and needs improvement.

par(mfrow = c(2, 2))  
plot(fit\_forward)

## Warning: not plotting observations with leverage one:  
## 121, 272, 584, 596, 1004, 1231, 1276, 1299, 1371



dev.off()

## null device   
## 1

ncvTest(fit\_forward)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 986.1211, Df = 1, p = < 2.22e-16

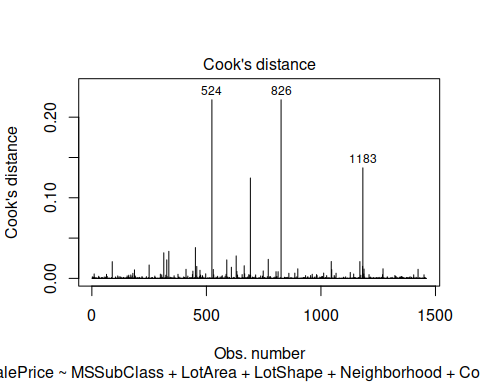
vif(fit\_forward)

## GVIF Df GVIF^(1/(2\*Df))  
## MSSubClass 5.071101 1 2.251911  
## LotArea 1.565287 1 1.251114  
## LotShape 1.864085 3 1.109373  
## Neighborhood 577.952343 24 1.141667  
## Condition1 3.156006 8 1.074475  
## Condition2 2.305334 7 1.061475  
## BldgType 16.150575 4 1.415870  
## OverallQual 3.285278 1 1.812534  
## OverallCond 1.495366 1 1.222852  
## YearBuilt 7.603801 1 2.757499  
## RoofMatl 2.167505 7 1.056811  
## Exterior2nd 22.937318 15 1.110073  
## BsmtFinSF2 1.165586 1 1.079623  
## FullBath 2.041604 1 1.428847  
## KitchenQual 3.811692 3 1.249836  
## PavedDrive 1.589166 2 1.122774  
## SaleCondition 2.143455 5 1.079224

outlierTest(fit\_forward)

## rstudent unadjusted p-value Bonferroni p  
## 1183 8.774990 4.9728e-18 7.2155e-15  
## 899 6.684378 3.3630e-11 4.8797e-08  
## 1170 6.463254 1.4201e-10 2.0606e-07  
## 692 6.428361 1.7754e-10 2.5762e-07  
## 804 5.133320 3.2556e-07 4.7239e-04  
## 441 4.750599 2.2409e-06 3.2515e-03  
## 1143 4.401285 1.1594e-05 1.6823e-02  
## 524 -4.259017 2.1930e-05 3.1820e-02  
## 826 4.259017 2.1930e-05 3.1820e-02

plot(fit\_forward, 4)



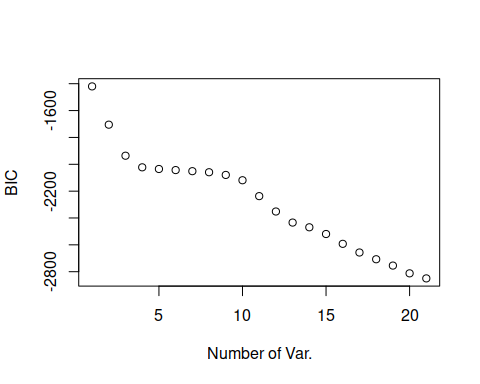
## Backward Elimination Model

We have built this model using backward Elimination technique using regsubsets function. The predictors are chosen for the final model based on Bayesian information criterion(BIC). Based on lowest BIC value the final model contains 13 predictors. The Adjusted r-squared of this model is 0.786 which is good but slightly lower than our previous model.

set.seed(123)  
backward\_model <- regsubsets(SalePrice ~ ., data = train, nvmax = 20, method = "backward")

## Reordering variables and trying again:

plot(summary(backward\_model)$bic, xlab="Number of Var.", ylab = "BIC")



coef(backward\_model, which.min(summary(backward\_model)$bic))

## (Intercept) LotArea NeighborhoodNoRidge   
## -8.061159e+05 1.402461e+00 7.738422e+04   
## NeighborhoodNridgHt Condition2PosN OverallQual   
## 5.453875e+04 -9.268434e+04 3.310640e+04   
## OverallCond YearBuilt RoofMatlCompShg   
## 2.996439e+03 2.500979e+02 2.557807e+05   
## RoofMatlMembran RoofMatlMetal RoofMatlRoll   
## 2.456086e+05 2.489145e+05 2.725861e+05   
## RoofMatlTar&Grv RoofMatlWdShake RoofMatlWdShngl   
## 2.631555e+05 2.639996e+05 3.434576e+05   
## BsmtFinSF2 LowQualFinSF BsmtFullBath   
## 4.424551e+00 2.558645e+01 1.413706e+04   
## KitchenQualTA FunctionalMaj2 FunctionalMin1   
## -1.308605e+04 -3.285878e+04 2.630804e+03   
## SaleConditionAdjLand   
## -1.853191e+04

fit\_backward <- lm(SalePrice ~ LotArea + Neighborhood + Condition2 + OverallQual + RoofMatl +   
 OverallCond + YearBuilt + BsmtFinSF2 +LowQualFinSF+  
 BsmtFullBath + KitchenQual + Functional + SaleCondition, data = train)  
summary(fit\_backward)

##   
## Call:  
## lm(formula = SalePrice ~ LotArea + Neighborhood + Condition2 +   
## OverallQual + RoofMatl + OverallCond + YearBuilt + BsmtFinSF2 +   
## LowQualFinSF + BsmtFullBath + KitchenQual + Functional +   
## SaleCondition, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -146740 -19944 -1872 16414 311056   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.071e+05 1.631e+05 -4.947 8.44e-07 \*\*\*  
## LotArea 1.192e+00 1.109e-01 10.749 < 2e-16 \*\*\*  
## NeighborhoodBlueste -6.719e+03 2.771e+04 -0.242 0.808470   
## NeighborhoodBrDale -2.021e+04 1.325e+04 -1.525 0.127559   
## NeighborhoodBrkSide 1.566e+04 1.159e+04 1.351 0.176796   
## NeighborhoodClearCr 2.535e+04 1.235e+04 2.053 0.040258 \*   
## NeighborhoodCollgCr 1.377e+04 9.498e+03 1.450 0.147266   
## NeighborhoodCrawfor 5.270e+04 1.134e+04 4.649 3.65e-06 \*\*\*  
## NeighborhoodEdwards 6.089e+03 1.056e+04 0.577 0.564364   
## NeighborhoodGilbert 1.519e+04 9.943e+03 1.528 0.126689   
## NeighborhoodIDOTRR 2.776e+03 1.218e+04 0.228 0.819743   
## NeighborhoodMeadowV -6.680e+03 1.325e+04 -0.504 0.614200   
## NeighborhoodMitchel 9.437e+03 1.080e+04 0.874 0.382400   
## NeighborhoodNAmes 1.334e+04 1.003e+04 1.330 0.183710   
## NeighborhoodNoRidge 1.088e+05 1.083e+04 10.047 < 2e-16 \*\*\*  
## NeighborhoodNPkVill 5.121e+03 1.543e+04 0.332 0.740079   
## NeighborhoodNridgHt 5.216e+04 1.011e+04 5.160 2.83e-07 \*\*\*  
## NeighborhoodNWAmes 2.375e+04 1.033e+04 2.299 0.021651 \*   
## NeighborhoodOldTown 5.699e+03 1.123e+04 0.507 0.611961   
## NeighborhoodSawyer 1.087e+04 1.065e+04 1.021 0.307510   
## NeighborhoodSawyerW 1.806e+04 1.032e+04 1.750 0.080378 .   
## NeighborhoodSomerst 1.394e+04 9.776e+03 1.426 0.154138   
## NeighborhoodStoneBr 5.727e+04 1.176e+04 4.871 1.24e-06 \*\*\*  
## NeighborhoodSWISU 2.182e+04 1.307e+04 1.670 0.095201 .   
## NeighborhoodTimber 2.074e+04 1.106e+04 1.874 0.061104 .   
## NeighborhoodVeenker 2.923e+04 1.464e+04 1.997 0.046068 \*   
## Condition2Feedr -2.282e+04 3.030e+04 -0.753 0.451556   
## Condition2Norm -1.449e+04 2.631e+04 -0.551 0.581891   
## Condition2PosA 5.783e+04 4.575e+04 1.264 0.206400   
## Condition2PosN -1.383e+05 3.753e+04 -3.684 0.000238 \*\*\*  
## Condition2RRAe 3.572e+04 4.547e+04 0.786 0.432240   
## Condition2RRAn -2.244e+04 4.540e+04 -0.494 0.621123   
## Condition2RRNn -3.057e+04 3.706e+04 -0.825 0.409619   
## OverallQual 2.585e+04 1.181e+03 21.889 < 2e-16 \*\*\*  
## RoofMatlCompShg 2.807e+05 3.814e+04 7.361 3.10e-13 \*\*\*  
## RoofMatlMembran 2.535e+05 5.478e+04 4.627 4.06e-06 \*\*\*  
## RoofMatlMetal 2.700e+05 5.354e+04 5.042 5.20e-07 \*\*\*  
## RoofMatlRoll 2.941e+05 5.309e+04 5.540 3.61e-08 \*\*\*  
## RoofMatlTar&Grv 2.888e+05 3.978e+04 7.260 6.41e-13 \*\*\*  
## RoofMatlWdShake 3.017e+05 4.168e+04 7.238 7.49e-13 \*\*\*  
## RoofMatlWdShngl 3.715e+05 4.067e+04 9.133 < 2e-16 \*\*\*  
## OverallCond 2.810e+03 1.038e+03 2.708 0.006843 \*\*   
## YearBuilt 2.809e+02 7.670e+01 3.663 0.000259 \*\*\*  
## BsmtFinSF2 7.246e+00 6.435e+00 1.126 0.260360   
## LowQualFinSF 2.440e+01 2.157e+01 1.131 0.258275   
## BsmtFullBath 1.137e+04 2.041e+03 5.573 3.00e-08 \*\*\*  
## KitchenQualFa -6.703e+04 8.406e+03 -7.975 3.15e-15 \*\*\*  
## KitchenQualGd -5.672e+04 4.627e+03 -12.258 < 2e-16 \*\*\*  
## KitchenQualTA -6.583e+04 5.285e+03 -12.457 < 2e-16 \*\*\*  
## FunctionalMaj2 -2.937e+04 1.949e+04 -1.507 0.132051   
## FunctionalMin1 6.697e+03 1.221e+04 0.549 0.583349   
## FunctionalMin2 1.524e+04 1.206e+04 1.264 0.206370   
## FunctionalMod 2.469e+04 1.413e+04 1.748 0.080641 .   
## FunctionalSev -3.416e+04 4.041e+04 -0.845 0.398068   
## FunctionalTyp 2.693e+03 1.027e+04 0.262 0.793176   
## SaleConditionAdjLand -1.110e+03 1.936e+04 -0.057 0.954284   
## SaleConditionAlloca 2.144e+04 1.158e+04 1.851 0.064309 .   
## SaleConditionFamily 4.533e+03 9.103e+03 0.498 0.618556   
## SaleConditionNormal 1.013e+04 3.929e+03 2.579 0.010020 \*   
## SaleConditionPartial 3.720e+04 5.495e+03 6.769 1.90e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36710 on 1400 degrees of freedom  
## Multiple R-squared: 0.7951, Adjusted R-squared: 0.7865   
## F-statistic: 92.08 on 59 and 1400 DF, p-value: < 2.2e-16

confint(fit\_backward)

## 2.5 % 97.5 %  
## (Intercept) -1.127131e+06 -4.870680e+05  
## LotArea 9.744272e-01 1.409496e+00  
## NeighborhoodBlueste -6.108523e+04 4.764681e+04  
## NeighborhoodBrDale -4.620764e+04 5.791373e+03  
## NeighborhoodBrkSide -7.074453e+03 3.840423e+04  
## NeighborhoodClearCr 1.127614e+03 4.956282e+04  
## NeighborhoodCollgCr -4.859188e+03 3.240466e+04  
## NeighborhoodCrawfor 3.046443e+04 7.494243e+04  
## NeighborhoodEdwards -1.462950e+04 2.680728e+04  
## NeighborhoodGilbert -4.309759e+03 3.469840e+04  
## NeighborhoodIDOTRR -2.111375e+04 2.666501e+04  
## NeighborhoodMeadowV -3.267058e+04 1.931012e+04  
## NeighborhoodMitchel -1.175011e+04 3.062450e+04  
## NeighborhoodNAmes -6.336747e+03 3.302625e+04  
## NeighborhoodNoRidge 8.753605e+04 1.300124e+05  
## NeighborhoodNPkVill -2.515209e+04 3.539332e+04  
## NeighborhoodNridgHt 3.233046e+04 7.199535e+04  
## NeighborhoodNWAmes 3.485312e+03 4.401829e+04  
## NeighborhoodOldTown -1.633479e+04 2.773311e+04  
## NeighborhoodSawyer -1.001862e+04 3.175978e+04  
## NeighborhoodSawyerW -2.187566e+03 3.831600e+04  
## NeighborhoodSomerst -5.238385e+03 3.311740e+04  
## NeighborhoodStoneBr 3.420296e+04 8.032816e+04  
## NeighborhoodSWISU -3.814563e+03 4.744642e+04  
## NeighborhoodTimber -9.670997e+02 4.244169e+04  
## NeighborhoodVeenker 5.106755e+02 5.794431e+04  
## Condition2Feedr -8.225186e+04 3.661989e+04  
## Condition2Norm -6.609324e+04 3.711630e+04  
## Condition2PosA -3.191294e+04 1.475800e+05  
## Condition2PosN -2.119215e+05 -6.466087e+04  
## Condition2RRAe -5.347912e+04 1.249252e+05  
## Condition2RRAn -1.115067e+05 6.661678e+04  
## Condition2RRNn -1.032770e+05 4.213585e+04  
## OverallQual 2.353496e+04 2.816846e+04  
## RoofMatlCompShg 2.059160e+05 3.555431e+05  
## RoofMatlMembran 1.459956e+05 3.609253e+05  
## RoofMatlMetal 1.649397e+05 3.749835e+05  
## RoofMatlRoll 1.899857e+05 3.982876e+05  
## RoofMatlTar&Grv 2.107547e+05 3.668215e+05  
## RoofMatlWdShake 2.199022e+05 3.834192e+05  
## RoofMatlWdShngl 2.916685e+05 4.512389e+05  
## OverallCond 7.748509e+02 4.845696e+03  
## YearBuilt 1.304919e+02 4.314015e+02  
## BsmtFinSF2 -5.377733e+00 1.986996e+01  
## LowQualFinSF -1.792223e+01 6.672065e+01  
## BsmtFullBath 7.370865e+03 1.537837e+04  
## KitchenQualFa -8.352283e+04 -5.054372e+04  
## KitchenQualGd -6.579762e+04 -4.764291e+04  
## KitchenQualTA -7.620235e+04 -5.546706e+04  
## FunctionalMaj2 -6.759676e+04 8.861629e+03  
## FunctionalMin1 -1.724929e+04 3.064380e+04  
## FunctionalMin2 -8.408607e+03 3.889083e+04  
## FunctionalMod -3.014621e+03 5.240398e+04  
## FunctionalSev -1.134396e+05 4.511418e+04  
## FunctionalTyp -1.745496e+04 2.284193e+04  
## SaleConditionAdjLand -3.909531e+04 3.687474e+04  
## SaleConditionAlloca -1.275476e+03 4.414780e+04  
## SaleConditionFamily -1.332308e+04 2.238947e+04  
## SaleConditionNormal 2.423960e+03 1.783781e+04  
## SaleConditionPartial 2.641714e+04 4.797498e+04

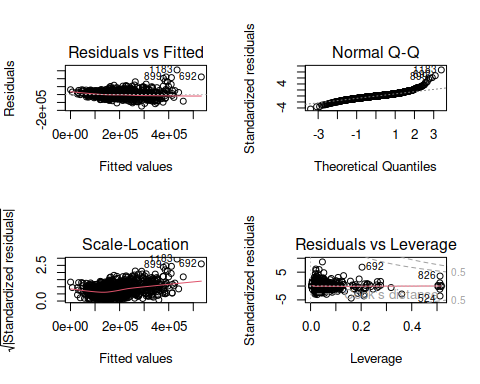
Backward\_Adjusted\_R2 <- 0.7865

## Diagnostics

The residuals looks normal and there is no obvious pattern in residuals vs fitted plot. The variance of the residuals is not constant based on the results of ncvTest. The variance inflation factor suggest Neighborhood shows multicolinearity. The outliear test shows that there couple of observations that are outlirers. The observation 524, 692 and 826 are influentials. So we need further imporvements to our model to get better predictions.

par(mfrow = c(2, 2))  
plot(fit\_backward)

## Warning: not plotting observations with leverage one:  
## 121, 272, 584, 667, 1004, 1231, 1276, 1299



dev.off()

## null device   
## 1

ncvTest(fit\_backward)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 987.3706, Df = 1, p = < 2.22e-16

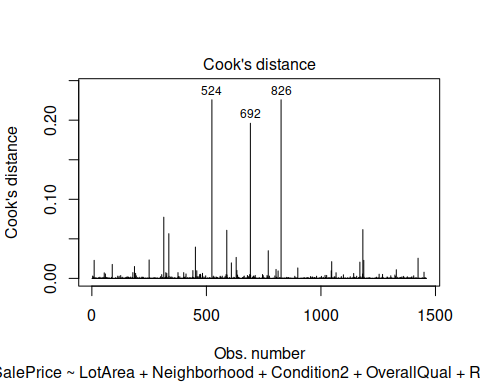
vif(fit\_backward)

## GVIF Df GVIF^(1/(2\*Df))  
## LotArea 1.326381 1 1.151686  
## Neighborhood 28.191307 24 1.072039  
## Condition2 1.294373 7 1.018601  
## OverallQual 2.888295 1 1.699499  
## RoofMatl 1.770769 7 1.041660  
## OverallCond 1.443390 1 1.201412  
## YearBuilt 5.809656 1 2.410323  
## BsmtFinSF2 1.166801 1 1.080186  
## LowQualFinSF 1.191370 1 1.091499  
## BsmtFullBath 1.214397 1 1.101997  
## KitchenQual 3.312330 3 1.220925  
## Functional 1.599519 6 1.039918  
## SaleCondition 1.721596 5 1.055828

outlierTest(fit\_backward)

## rstudent unadjusted p-value Bonferroni p  
## 1183 8.919143 1.4362e-18 2.0854e-15  
## 692 6.896715 8.0345e-12 1.1666e-08  
## 899 6.207569 7.0773e-10 1.0276e-06  
## 1170 6.002381 2.4734e-09 3.5914e-06  
## 804 5.688194 1.5612e-08 2.2668e-05  
## 1047 4.827028 1.5379e-06 2.2330e-03  
## 441 4.467253 8.5611e-06 1.2431e-02  
## 1143 4.459566 8.8695e-06 1.2878e-02  
## 589 -4.391850 1.2085e-05 1.7548e-02

plot(fit\_backward, 4)



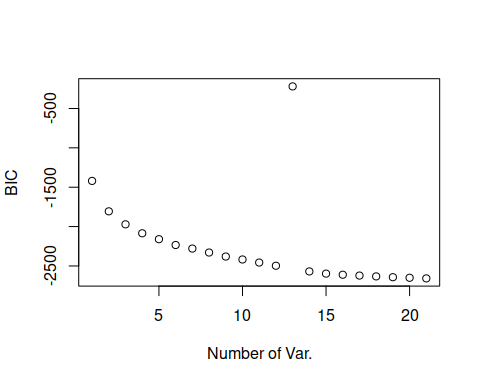
## Best Subset Model

We build this model using best subset selction method using regsubsets. The best coefficients based on lowest BIC values are extracted. The final best model contains 18 vairables. The model performs better than previous model with adjusted r-squared value of 0.8171.

set.seed(123)  
best\_model <- regsubsets(SalePrice ~ . , data = train, nvmax = 20, method = "seqrep")

## Reordering variables and trying again:

plot(summary(best\_model)$bic, xlab="Number of Var.", ylab = "BIC")



coef(best\_model, which.min(summary(best\_model)$bic))

## (Intercept) MSSubClass LotArea   
## -7.662740e+05 -3.113764e+02 1.117732e+00   
## LotShapeIR3 NeighborhoodCrawfor NeighborhoodNoRidge   
## -2.689683e+04 3.839663e+04 8.168489e+04   
## NeighborhoodNridgHt NeighborhoodStoneBr Condition1Norm   
## 6.588392e+04 7.825019e+04 3.723630e+03   
## Condition2PosN BldgType2fmCon OverallQual   
## -8.632146e+04 3.977891e+04 2.656121e+04   
## OverallCond YearBuilt RoofMatlWdShngl   
## 5.228468e+03 3.546448e+02 1.122659e+05   
## Exterior2ndHdBoard BsmtFinSF2 FullBath   
## -9.573793e+03 1.301091e+01 2.754496e+04   
## HalfBath KitchenQualGd PavedDriveY   
## 1.706036e+04 -8.868306e+03 5.384316e+03   
## SaleConditionAdjLand   
## -1.389608e+04

fit\_best <- lm(SalePrice ~ MSSubClass + LotArea + LotShape + Neighborhood + Condition1 +  
 Condition2 + BldgType + OverallQual + OverallCond + YearBuilt +  
 RoofMatl + Exterior2nd + BsmtFinSF2 + FullBath + HalfBath +  
 KitchenQual + PavedDrive + SaleCondition, data = train)  
summary(fit\_best)

##   
## Call:  
## lm(formula = SalePrice ~ MSSubClass + LotArea + LotShape + Neighborhood +   
## Condition1 + Condition2 + BldgType + OverallQual + OverallCond +   
## YearBuilt + RoofMatl + Exterior2nd + BsmtFinSF2 + FullBath +   
## HalfBath + KitchenQual + PavedDrive + SaleCondition, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -116889 -18325 -1233 14141 269800   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.180e+05 1.705e+05 -4.797 1.79e-06 \*\*\*  
## MSSubClass -7.714e+01 5.194e+01 -1.485 0.137710   
## LotArea 9.693e-01 1.115e-01 8.693 < 2e-16 \*\*\*  
## LotShapeIR2 5.264e+03 5.831e+03 0.903 0.366800   
## LotShapeIR3 5.891e+03 1.216e+04 0.484 0.628245   
## LotShapeReg -3.757e+03 2.175e+03 -1.727 0.084335 .   
## NeighborhoodBlueste -9.779e+03 2.608e+04 -0.375 0.707757   
## NeighborhoodBrDale -1.372e+04 1.374e+04 -0.999 0.318210   
## NeighborhoodBrkSide -8.491e+03 1.149e+04 -0.739 0.459988   
## NeighborhoodClearCr -4.408e+03 1.221e+04 -0.361 0.718190   
## NeighborhoodCollgCr -2.030e+04 9.621e+03 -2.110 0.035009 \*   
## NeighborhoodCrawfor 2.243e+04 1.126e+04 1.993 0.046468 \*   
## NeighborhoodEdwards -1.737e+04 1.044e+04 -1.663 0.096491 .   
## NeighborhoodGilbert -3.601e+04 1.036e+04 -3.476 0.000525 \*\*\*  
## NeighborhoodIDOTRR -2.072e+04 1.206e+04 -1.718 0.086040 .   
## NeighborhoodMeadowV -8.834e+03 1.426e+04 -0.620 0.535676   
## NeighborhoodMitchel -1.434e+04 1.079e+04 -1.329 0.184025   
## NeighborhoodNAmes -1.324e+04 1.023e+04 -1.294 0.195847   
## NeighborhoodNoRidge 6.320e+04 1.102e+04 5.738 1.18e-08 \*\*\*  
## NeighborhoodNPkVill -7.894e+03 1.747e+04 -0.452 0.651456   
## NeighborhoodNridgHt 3.317e+04 9.842e+03 3.370 0.000771 \*\*\*  
## NeighborhoodNWAmes -2.365e+04 1.068e+04 -2.215 0.026957 \*   
## NeighborhoodOldTown -2.108e+04 1.108e+04 -1.902 0.057429 .   
## NeighborhoodSawyer -1.347e+04 1.081e+04 -1.245 0.213269   
## NeighborhoodSawyerW -1.153e+04 1.042e+04 -1.106 0.268860   
## NeighborhoodSomerst -1.550e+04 9.717e+03 -1.595 0.110928   
## NeighborhoodStoneBr 4.655e+04 1.131e+04 4.114 4.11e-05 \*\*\*  
## NeighborhoodSWISU -1.082e+04 1.257e+04 -0.861 0.389592   
## NeighborhoodTimber -1.097e+04 1.102e+04 -0.996 0.319512   
## NeighborhoodVeenker 1.074e+04 1.429e+04 0.752 0.452386   
## Condition1Feedr 4.412e+03 6.836e+03 0.645 0.518844   
## Condition1Norm 7.843e+03 5.599e+03 1.401 0.161507   
## Condition1PosA 1.349e+04 1.380e+04 0.977 0.328659   
## Condition1PosN 1.923e+04 1.019e+04 1.887 0.059318 .   
## Condition1RRAe -1.354e+04 1.213e+04 -1.116 0.264614   
## Condition1RRAn 8.640e+03 9.328e+03 0.926 0.354463   
## Condition1RRNe -1.606e+04 2.532e+04 -0.635 0.525841   
## Condition1RRNn 4.758e+03 1.702e+04 0.280 0.779897   
## Condition2Feedr -2.877e+04 2.994e+04 -0.961 0.336699   
## Condition2Norm -1.704e+04 2.585e+04 -0.659 0.509834   
## Condition2PosA 2.952e+04 4.315e+04 0.684 0.493956   
## Condition2PosN -1.634e+05 3.720e+04 -4.392 1.21e-05 \*\*\*  
## Condition2RRAe -6.156e+02 4.395e+04 -0.014 0.988825   
## Condition2RRAn -1.557e+04 4.353e+04 -0.358 0.720662   
## Condition2RRNn -3.482e+04 3.584e+04 -0.971 0.331527   
## BldgType2fmCon 9.065e+03 9.998e+03 0.907 0.364712   
## BldgTypeDuplex -1.077e+04 6.306e+03 -1.707 0.087988 .   
## BldgTypeTwnhs -4.226e+04 9.061e+03 -4.663 3.42e-06 \*\*\*  
## BldgTypeTwnhsE -2.972e+04 6.764e+03 -4.394 1.20e-05 \*\*\*  
## OverallQual 2.095e+04 1.170e+03 17.895 < 2e-16 \*\*\*  
## OverallCond 3.717e+03 9.790e+02 3.797 0.000153 \*\*\*  
## YearBuilt 3.080e+02 8.145e+01 3.782 0.000162 \*\*\*  
## RoofMatlCompShg 2.618e+05 3.799e+04 6.890 8.46e-12 \*\*\*  
## RoofMatlMembran 2.768e+05 5.201e+04 5.322 1.20e-07 \*\*\*  
## RoofMatlMetal 2.742e+05 5.170e+04 5.303 1.32e-07 \*\*\*  
## RoofMatlRoll 2.610e+05 5.170e+04 5.048 5.05e-07 \*\*\*  
## RoofMatlTar&Grv 2.633e+05 3.950e+04 6.667 3.78e-11 \*\*\*  
## RoofMatlWdShake 2.801e+05 4.127e+04 6.788 1.69e-11 \*\*\*  
## RoofMatlWdShngl 3.664e+05 4.031e+04 9.088 < 2e-16 \*\*\*  
## Exterior2ndAsphShn -1.574e+04 2.145e+04 -0.734 0.463121   
## Exterior2ndBrk Cmn -1.103e+04 1.896e+04 -0.582 0.560697   
## Exterior2ndBrkFace -5.298e+03 1.081e+04 -0.490 0.624177   
## Exterior2ndCBlock 2.201e+04 3.542e+04 0.621 0.534494   
## Exterior2ndCmentBd 4.656e+03 1.015e+04 0.459 0.646662   
## Exterior2ndHdBoard -1.244e+04 8.606e+03 -1.445 0.148564   
## Exterior2ndImStucc 2.462e+04 1.392e+04 1.769 0.077181 .   
## Exterior2ndMetalSd -6.888e+03 8.397e+03 -0.820 0.412223   
## Exterior2ndOther -6.674e+03 3.595e+04 -0.186 0.852722   
## Exterior2ndPlywood -8.842e+03 8.822e+03 -1.002 0.316356   
## Exterior2ndStone -2.987e+04 1.775e+04 -1.682 0.092724 .   
## Exterior2ndStucco -6.874e+02 1.069e+04 -0.064 0.948736   
## Exterior2ndVinylSd -8.203e+03 8.519e+03 -0.963 0.335770   
## Exterior2ndWd Sdng -1.013e+04 8.385e+03 -1.208 0.227096   
## Exterior2ndWd Shng -2.287e+04 9.781e+03 -2.338 0.019544 \*   
## BsmtFinSF2 1.260e+01 5.953e+00 2.117 0.034428 \*   
## FullBath 2.611e+04 2.343e+03 11.147 < 2e-16 \*\*\*  
## HalfBath 1.443e+04 2.231e+03 6.468 1.37e-10 \*\*\*  
## KitchenQualFa -6.161e+04 7.903e+03 -7.796 1.26e-14 \*\*\*  
## KitchenQualGd -5.313e+04 4.416e+03 -12.032 < 2e-16 \*\*\*  
## KitchenQualTA -5.895e+04 5.010e+03 -11.766 < 2e-16 \*\*\*  
## PavedDriveP 1.344e+03 7.385e+03 0.182 0.855577   
## PavedDriveY 6.670e+03 4.471e+03 1.492 0.135992   
## SaleConditionAdjLand -5.219e+03 1.825e+04 -0.286 0.774996   
## SaleConditionAlloca 2.065e+04 1.119e+04 1.846 0.065055 .   
## SaleConditionFamily 1.713e+03 8.527e+03 0.201 0.840810   
## SaleConditionNormal 7.069e+03 3.668e+03 1.927 0.054180 .   
## SaleConditionPartial 2.565e+04 5.201e+03 4.931 9.17e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 33980 on 1373 degrees of freedom  
## Multiple R-squared: 0.8279, Adjusted R-squared: 0.8171   
## F-statistic: 76.78 on 86 and 1373 DF, p-value: < 2.2e-16

confint(fit\_best)

## 2.5 % 97.5 %  
## (Intercept) -1.152541e+06 -4.834642e+05  
## MSSubClass -1.790186e+02 2.474500e+01  
## LotArea 7.505843e-01 1.188048e+00  
## LotShapeIR2 -6.174083e+03 1.670158e+04  
## LotShapeIR3 -1.797110e+04 2.975345e+04  
## LotShapeReg -8.023864e+03 5.097792e+02  
## NeighborhoodBlueste -6.094166e+04 4.138369e+04  
## NeighborhoodBrDale -4.067498e+04 1.323477e+04  
## NeighborhoodBrkSide -3.102912e+04 1.404667e+04  
## NeighborhoodClearCr -2.836692e+04 1.955013e+04  
## NeighborhoodCollgCr -3.917876e+04 -1.430301e+03  
## NeighborhoodCrawfor 3.514644e+02 4.451621e+04  
## NeighborhoodEdwards -3.786027e+04 3.117072e+03  
## NeighborhoodGilbert -5.633434e+04 -1.568688e+04  
## NeighborhoodIDOTRR -4.437865e+04 2.940316e+03  
## NeighborhoodMeadowV -3.680557e+04 1.913804e+04  
## NeighborhoodMitchel -3.550258e+04 6.824268e+03  
## NeighborhoodNAmes -3.329990e+04 6.828032e+03  
## NeighborhoodNoRidge 4.159185e+04 8.480881e+04  
## NeighborhoodNPkVill -4.216694e+04 2.637874e+04  
## NeighborhoodNridgHt 1.386424e+04 5.247666e+04  
## NeighborhoodNWAmes -4.459332e+04 -2.699579e+03  
## NeighborhoodOldTown -4.281833e+04 6.657605e+02  
## NeighborhoodSawyer -3.467921e+04 7.748069e+03  
## NeighborhoodSawyerW -3.197535e+04 8.917258e+03  
## NeighborhoodSomerst -3.455946e+04 3.562379e+03  
## NeighborhoodStoneBr 2.435376e+04 6.874008e+04  
## NeighborhoodSWISU -3.547926e+04 1.384139e+04  
## NeighborhoodTimber -3.259291e+04 1.064425e+04  
## NeighborhoodVeenker -1.729295e+04 3.877724e+04  
## Condition1Feedr -8.999525e+03 1.782260e+04  
## Condition1Norm -3.140546e+03 1.882635e+04  
## Condition1PosA -1.358709e+04 4.055807e+04  
## Condition1PosN -7.571948e+02 3.922694e+04  
## Condition1RRAe -3.734552e+04 1.026175e+04  
## Condition1RRAn -9.658141e+03 2.693851e+04  
## Condition1RRNe -6.573026e+04 3.360068e+04  
## Condition1RRNn -2.863265e+04 3.814774e+04  
## Condition2Feedr -8.750707e+04 2.995951e+04  
## Condition2Norm -6.775634e+04 3.366989e+04  
## Condition2PosA -5.512405e+04 1.141724e+05  
## Condition2PosN -2.363655e+05 -9.040059e+04  
## Condition2RRAe -8.682381e+04 8.559256e+04  
## Condition2RRAn -1.009641e+05 6.982630e+04  
## Condition2RRNn -1.051255e+05 3.549389e+04  
## BldgType2fmCon -1.054742e+04 2.867798e+04  
## BldgTypeDuplex -2.313565e+04 1.603969e+03  
## BldgTypeTwnhs -6.003080e+04 -2.447928e+04  
## BldgTypeTwnhsE -4.298975e+04 -1.645333e+04  
## OverallQual 1.864916e+04 2.324119e+04  
## OverallCond 1.797029e+03 5.637837e+03  
## YearBuilt 1.482393e+02 4.677901e+02  
## RoofMatlCompShg 1.872571e+05 3.363235e+05  
## RoofMatlMembran 1.747599e+05 3.788210e+05  
## RoofMatlMetal 1.727667e+05 3.756043e+05  
## RoofMatlRoll 1.595799e+05 3.624099e+05  
## RoofMatlTar&Grv 1.858420e+05 3.408123e+05  
## RoofMatlWdShake 1.991693e+05 3.610726e+05  
## RoofMatlWdShngl 2.872796e+05 4.454462e+05  
## Exterior2ndAsphShn -5.780888e+04 2.633019e+04  
## Exterior2ndBrk Cmn -4.822002e+04 2.615595e+04  
## Exterior2ndBrkFace -2.650679e+04 1.591054e+04  
## Exterior2ndCBlock -4.748110e+04 9.149987e+04  
## Exterior2ndCmentBd -1.526428e+04 2.457616e+04  
## Exterior2ndHdBoard -2.932020e+04 4.442688e+03  
## Exterior2ndImStucc -2.688252e+03 5.193578e+04  
## Exterior2ndMetalSd -2.336041e+04 9.585029e+03  
## Exterior2ndOther -7.718924e+04 6.384029e+04  
## Exterior2ndPlywood -2.614756e+04 8.463049e+03  
## Exterior2ndStone -6.469382e+04 4.958804e+03  
## Exterior2ndStucco -2.165677e+04 2.028199e+04  
## Exterior2ndVinylSd -2.491599e+04 8.509247e+03  
## Exterior2ndWd Sdng -2.658185e+04 6.316357e+03  
## Exterior2ndWd Shng -4.205281e+04 -3.678054e+03  
## BsmtFinSF2 9.253563e-01 2.428186e+01  
## FullBath 2.151741e+04 3.070833e+04  
## HalfBath 1.005268e+04 1.880411e+04  
## KitchenQualFa -7.711640e+04 -4.610822e+04  
## KitchenQualGd -6.179607e+04 -4.447019e+04  
## KitchenQualTA -6.877524e+04 -4.911902e+04  
## PavedDriveP -1.314326e+04 1.583213e+04  
## PavedDriveY -2.101256e+03 1.544173e+04  
## SaleConditionAdjLand -4.102651e+04 3.058907e+04  
## SaleConditionAlloca -1.290113e+03 4.259753e+04  
## SaleConditionFamily -1.501408e+04 1.844009e+04  
## SaleConditionNormal -1.270897e+02 1.426561e+04  
## SaleConditionPartial 1.544412e+04 3.584981e+04

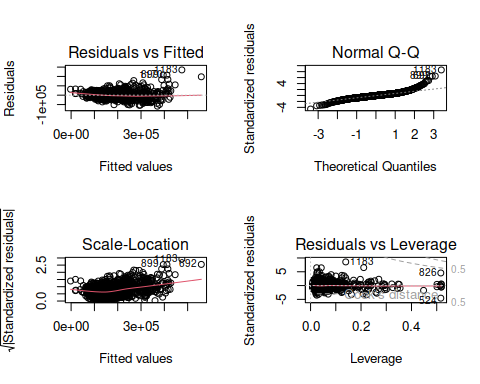
best\_Ajdusted\_R2 <- 0.817

## Diagnostics

The residuals look normal and there are presence of outliers and some influential points. So this model can further be improved.

par(mfrow = c(2, 2))  
plot(fit\_best)

## Warning: not plotting observations with leverage one:  
## 121, 272, 584, 596, 1004, 1231, 1276, 1299, 1371



dev.off()

## null device   
## 1

ncvTest(fit\_best)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 1015.275, Df = 1, p = < 2.22e-16

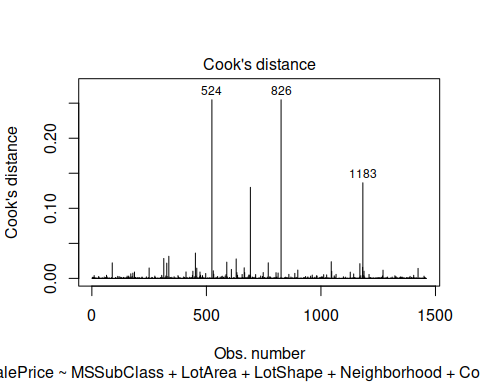
vif(fit\_best)

## GVIF Df GVIF^(1/(2\*Df))  
## MSSubClass 6.099664 1 2.469750  
## LotArea 1.565365 1 1.251145  
## LotShape 1.865408 3 1.109504  
## Neighborhood 627.936223 24 1.143642  
## Condition1 3.177080 8 1.074922  
## Condition2 2.343452 7 1.062719  
## BldgType 18.903718 4 1.444004  
## OverallQual 3.311406 1 1.819727  
## OverallCond 1.499820 1 1.224672  
## YearBuilt 7.647821 1 2.765469  
## RoofMatl 2.181393 7 1.057293  
## Exterior2nd 23.126439 15 1.110377  
## BsmtFinSF2 1.165597 1 1.079628  
## FullBath 2.104983 1 1.450856  
## HalfBath 1.590216 1 1.261038  
## KitchenQual 3.833219 3 1.251010  
## PavedDrive 1.589627 2 1.122855  
## SaleCondition 2.155019 5 1.079804

outlierTest(fit\_best)

## rstudent unadjusted p-value Bonferroni p  
## 1183 8.794335 4.2284e-18 6.1354e-15  
## 692 6.606250 5.6238e-11 8.1601e-08  
## 899 6.541040 8.6002e-11 1.2479e-07  
## 1170 6.524249 9.5884e-11 1.3913e-07  
## 804 5.047347 5.0808e-07 7.3723e-04  
## 441 5.031627 5.5075e-07 7.9914e-04  
## 1143 4.683814 3.0952e-06 4.4911e-03  
## 524 -4.591180 4.8112e-06 6.9810e-03  
## 826 4.591180 4.8112e-06 6.9810e-03

plot(fit\_best, 4)



## Custom Model

Finally, we built our last custom model. I chose following 10 predictors for this custom model: OverallCond, YearBuilt, YearRemodAdd, BedroomAbvGr, GrLivArea, KitchenAbvGr, TotRmsAbvGrd, GarageCars, PoolArea and LotArea. The summary of this model shows that this performed the worst compare to other models with adjusted r-squared value of 0.737.

custom\_fit <- lm(SalePrice ~ OverallCond + YearBuilt + YearRemodAdd + BedroomAbvGr + GrLivArea + KitchenAbvGr + TotRmsAbvGrd + GarageCars + PoolArea + LotArea, data = train)  
  
summary(custom\_fit)

##   
## Call:  
## lm(formula = SalePrice ~ OverallCond + YearBuilt + YearRemodAdd +   
## BedroomAbvGr + GrLivArea + KitchenAbvGr + TotRmsAbvGrd +   
## GarageCars + PoolArea + LotArea, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -449279 -21631 -3228 17607 287987   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.806e+06 1.219e+05 -14.820 < 2e-16 \*\*\*  
## OverallCond 7.138e+03 1.148e+03 6.219 6.52e-10 \*\*\*  
## YearBuilt 7.232e+02 5.603e+01 12.906 < 2e-16 \*\*\*  
## YearRemodAdd 2.057e+02 7.324e+01 2.809 0.00504 \*\*   
## BedroomAbvGr -1.760e+04 1.855e+03 -9.489 < 2e-16 \*\*\*  
## GrLivArea 8.201e+01 4.001e+00 20.499 < 2e-16 \*\*\*  
## KitchenAbvGr -3.965e+04 5.269e+03 -7.525 9.24e-14 \*\*\*  
## TotRmsAbvGrd 5.975e+03 1.419e+03 4.210 2.71e-05 \*\*\*  
## GarageCars 1.997e+04 1.912e+03 10.444 < 2e-16 \*\*\*  
## PoolArea -2.178e+01 2.717e+01 -0.801 0.42301   
## LotArea 6.784e-01 1.116e-01 6.078 1.55e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 40720 on 1449 degrees of freedom  
## Multiple R-squared: 0.7391, Adjusted R-squared: 0.7373   
## F-statistic: 410.5 on 10 and 1449 DF, p-value: < 2.2e-16

confint(custom\_fit)

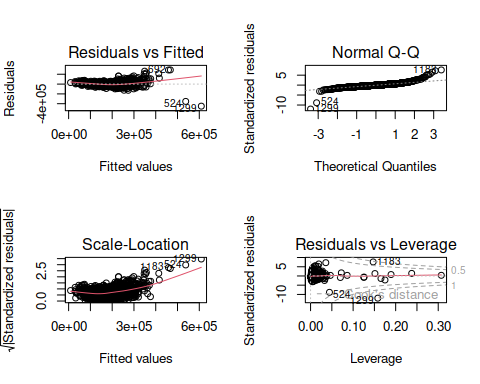
## 2.5 % 97.5 %  
## (Intercept) -2.045540e+06 -1.567328e+06  
## OverallCond 4.886535e+03 9.389196e+03  
## YearBuilt 6.132901e+02 8.331261e+02  
## YearRemodAdd 6.202857e+01 3.493568e+02  
## BedroomAbvGr -2.124017e+04 -1.396307e+04  
## GrLivArea 7.416379e+01 8.985940e+01  
## KitchenAbvGr -4.998605e+04 -2.931343e+04  
## TotRmsAbvGrd 3.191226e+03 8.759741e+03  
## GarageCars 1.622112e+04 2.372316e+04  
## PoolArea -7.507878e+01 3.152473e+01  
## LotArea 4.594348e-01 8.972829e-01

custom\_Ajdsuted\_R2 <- 0.737

## Diagnostics

Diagnostic plots suggest that residuals are normally distributed. There are nvc test suggest that variance is not constant. There are dozens of observations that are ouliers. There are some influential observations that should be taking care of to improve the model performance.

par(mfrow = c(2, 2))  
plot(custom\_fit)



dev.off()

## null device   
## 1

ncvTest(custom\_fit)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 2473.03, Df = 1, p = < 2.22e-16

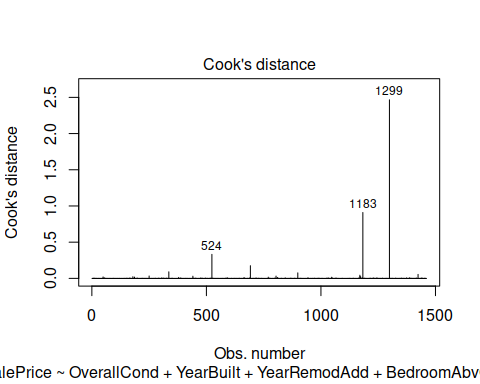
vif(custom\_fit)

## OverallCond YearBuilt YearRemodAdd BedroomAbvGr GrLivArea KitchenAbvGr   
## 1.435351 2.520468 2.011821 2.014869 3.889160 1.186196   
## TotRmsAbvGrd GarageCars PoolArea LotArea   
## 4.683619 1.797015 1.048795 1.091953

outlierTest(custom\_fit)

## rstudent unadjusted p-value Bonferroni p  
## 1299 -12.668985 5.8379e-35 8.5233e-32  
## 524 -9.113197 2.5899e-19 3.7813e-16  
## 1183 7.771990 1.4554e-14 2.1249e-11  
## 692 7.332827 3.7349e-13 5.4529e-10  
## 899 6.607988 5.4593e-11 7.9706e-08  
## 804 5.635761 2.0916e-08 3.0537e-05  
## 1170 5.277021 1.5127e-07 2.2086e-04  
## 441 5.192279 2.3730e-07 3.4645e-04  
## 1047 5.044967 5.1099e-07 7.4604e-04

plot(custom\_fit, 4)



## Comaring Results

Finally, we compared the results from each model. The table below suggests that the stepwise best subset model resulted with highest adjusted r-squared value of 0.817. The forward model results are pretty close to the best subset model with adjusted r-squared value of 0.812. The custom model with 10 predictors performed the worst here.

data.frame(`Adjusted R2` = c(custom\_Ajdsuted\_R2, best\_Ajdusted\_R2, Backward\_Adjusted\_R2, Forward\_Adjusted\_R2),   
 row.names = c('Custom', 'Stepwise', 'Backward', 'Forward')) %>%  
 knitr::kable()

|  | Adjusted.R2 |
| --- | --- |
| Custom | 0.7370 |
| Stepwise | 0.8170 |
| Backward | 0.7865 |
| Forward | 0.8120 |

## Conclusion

We fitted 4 multiple linear regression models. The results suggested that the best subset model performed the best with highest adjuseted r-squared value. The custom model performed the worst with lowest r-squared value. Overall some of the assumptions for our models were fulfilled. There were outliers in the residuals and variance of the residuals were not constant. There were presence of influential observatoins and multi-colinearity. So our models need further improvement to predict the Sale Prices of Ames houses.