Project 1

the BRICS

3/5/2019

1. Loading and pre-processing of Data

1.1 Data introduction

There are 1460 observations in the dataset. Within each observation, we have numerous features of one particular house (such as the building class, first-floor square feet, and the number of bedrooms above basement level) and its sale price. The data is from a kaggle competition.

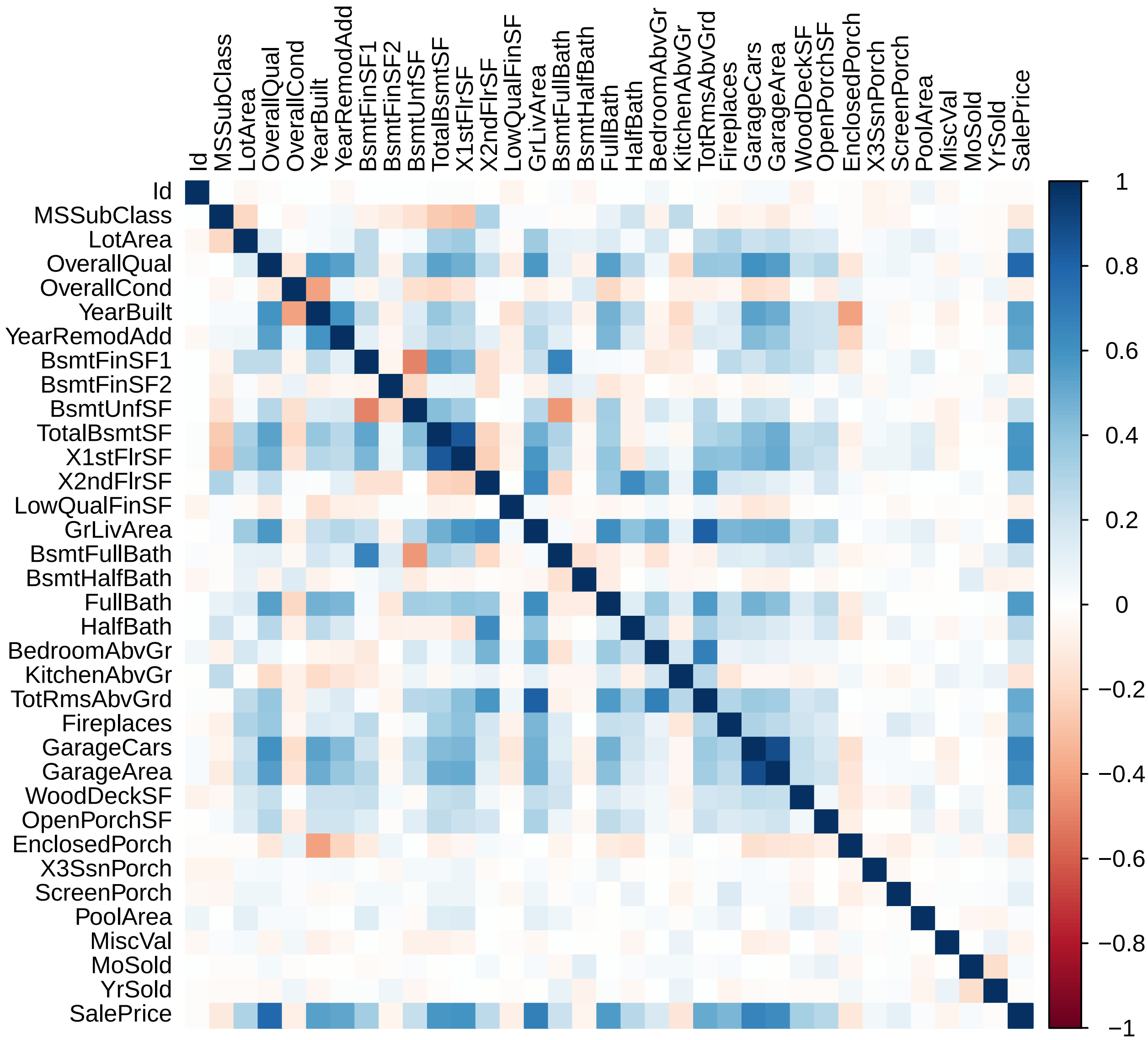
1.2 Treatment of missing values

There are 80 variables in total. Since 19 variables have missing values, we decide to remove them. After this first data processing step, we have 61 variables left.

2. Feature Selection

2.1 Correlation coefficient

The corrplot is a graphical display of correlation matrix of all features. It is important to identify the hidden structure and pattern in the matrix. From the corrplot, we can see that SalePrice is related to a lot of variables obviously.
Next, we focus on the relationship between SalePrice and other variables. To be specific, we find the variables that have a high correlation with SalePrice through corrplot by filtering them with the standard: correlation coefficient (variable, SalePrice) >0.4.
With this method, we get the following variables: *OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, X1stFlrSF, GrLivArea, FullBath, TotRms, AbvGrd, Fireplaces, GarageCars, GarageArea.*



2.2 Random forest model

Since categorical features are not considered in correlation matrix, we then use the random forest model to rank features by their importance from all the 61 variables. Top 15 variables are: *GrLivArea, OverallQual, BsmtFinSF1, GarageArea, TotalBsmtSF, LotArea, X1stFlrSF, GarageCars, X2ndFlrSF, YearBuilt, Fireplaces, YearRemodAdd, OverallCond, MSSubClass, ExterQual*.

2.3 Decision on relevant and important variables

By combining the two methods above, we decide to use these variables: *OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, X1stFlrSF, GrLivArea, FullBath, AbvGrd, Fireplaces, GarageCars, GarageArea, MSSubClass* The detalied variable descriptions are shown below.

Determinants	Description	Reason	Expected Effect
OverallQual	Overall material and finish quality	Overall material and finish quality shows the degree of excellence of the house and thus influences the price	The higher the overall quality, the higher the price
YearBuilt	Original construction date	Original construction date shows the degree of oldness and thus influences the price	The earlier the original construction date, the lower the price
YearRemodAdd	Remodel date	Remodel date suggests the degree of oldness and therefore determines the price	The earlier the remodel date, the lower the price
TotalBsmtSF	Total square feet of basement area	Total square feet of basement area implies the size of the storage area and thus influences the price	The more the total square feet of the basement area, the higher the price
X1stFlrSF	First-Floor square feet	The first-floor square feet imply the size of the living area in the first floor and thus influences the price	The more the first-floor square feet, the higher the price
GrLivArea	Above grade (ground) living area square feet	Above grade living area square feet indicates the size of the total living area above ground and thus determines the price	The more the above grade living area square feet, the higher the price
FullBath	Full bathrooms above grade	Full bathrooms above grade indicate the size and functionality of the house and therefore could influence the price	The more the full bathrooms above grade, the higher the price
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)	Total rooms above grade indicate the size and functionality of the house and therefore may impact the price	The more the total rooms above grade, the higher the price
Fireplaces	Number of fireplaces	Fireplaces bring a luxurious feel to a house and increase the value of the house	The higher the number of fireplaces, the higher the price
GarageCars	Size of the garage in car capacity	Size of the garage in car capacity measures the garage dimension and size and therefore impact the price	The bigger the size of the garage in car capacity, the higher the price
GarageArea	Size of the garage in square feet	Size of the garage in square feet measures the area to accommodate vehicles and store stuff and thus influence the price	The bigger the size of the garage in square feet, the higher the price
MSSubClass	The building class	The building class signals the oldness and quality of the house and thus influences the price	The higher the building class, the higher the price
SalePrice	the property's sale price in dollars. This is the target variable that we are trying to predict.		

3. Descriptive Statistics

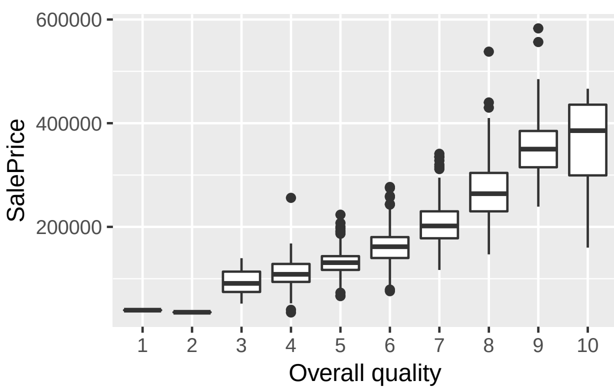
3.1 Descriptive Statistics

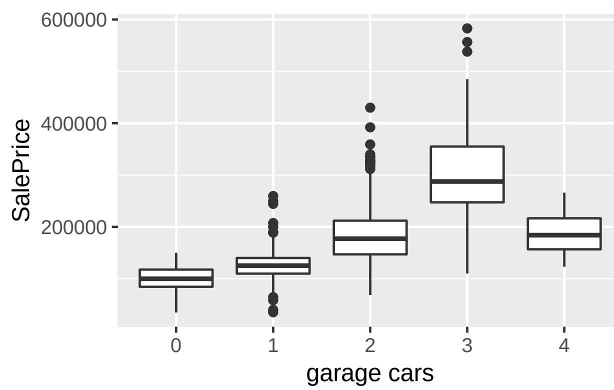
After selecting the relevant and important variables, we have created a summary table with the minimum, average, median, standard deviation and maximum.

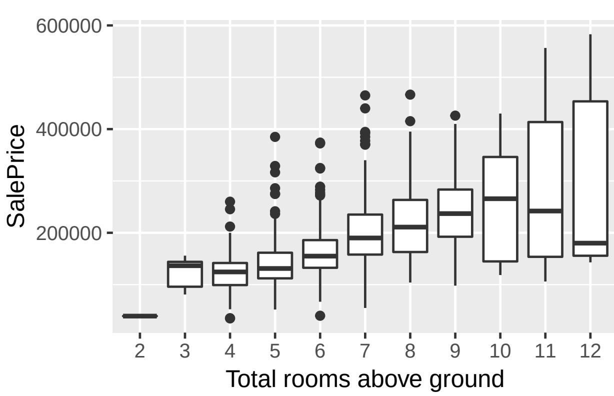
	n	min	mean	median	sd	max
OverallQual	876	1	6.13	6.0	1.37	10
YearBuilt	876	1875	1971.59	1973.5	30.45	2009
YearRemodAdd	876	1950	1985.25	1994.0	20.56	2010
TotalBsmtSF	876	0	1061.40	982.5	445.73	6110
X1stFlrSF	876	334	1158.38	1077.0	401.71	4692
GrLivArea	876	334	1503.02	1456.0	509.56	5642
FullBath	876	0	1.56	2.0	0.54	3
TotRmsAbvGrd	876	2	6.45	6.0	1.59	12
Fireplaces	876	0	0.60	1.0	0.64	3
GarageCars	876	0	1.78	2.0	0.74	4
GarageArea	876	0	476.00	480.0	213.51	1418
MSSubClass	876	20	57.46	50.0	42.34	190
SalePrice	876	34900	179801.83	163945.0	74717.48	582933

3.2 box-and-whisker plots

Also, we draw several box-and-whisker plots. We can tell from the plot that overall quality, garages cars, total rooms above ground do influence sale price of a house

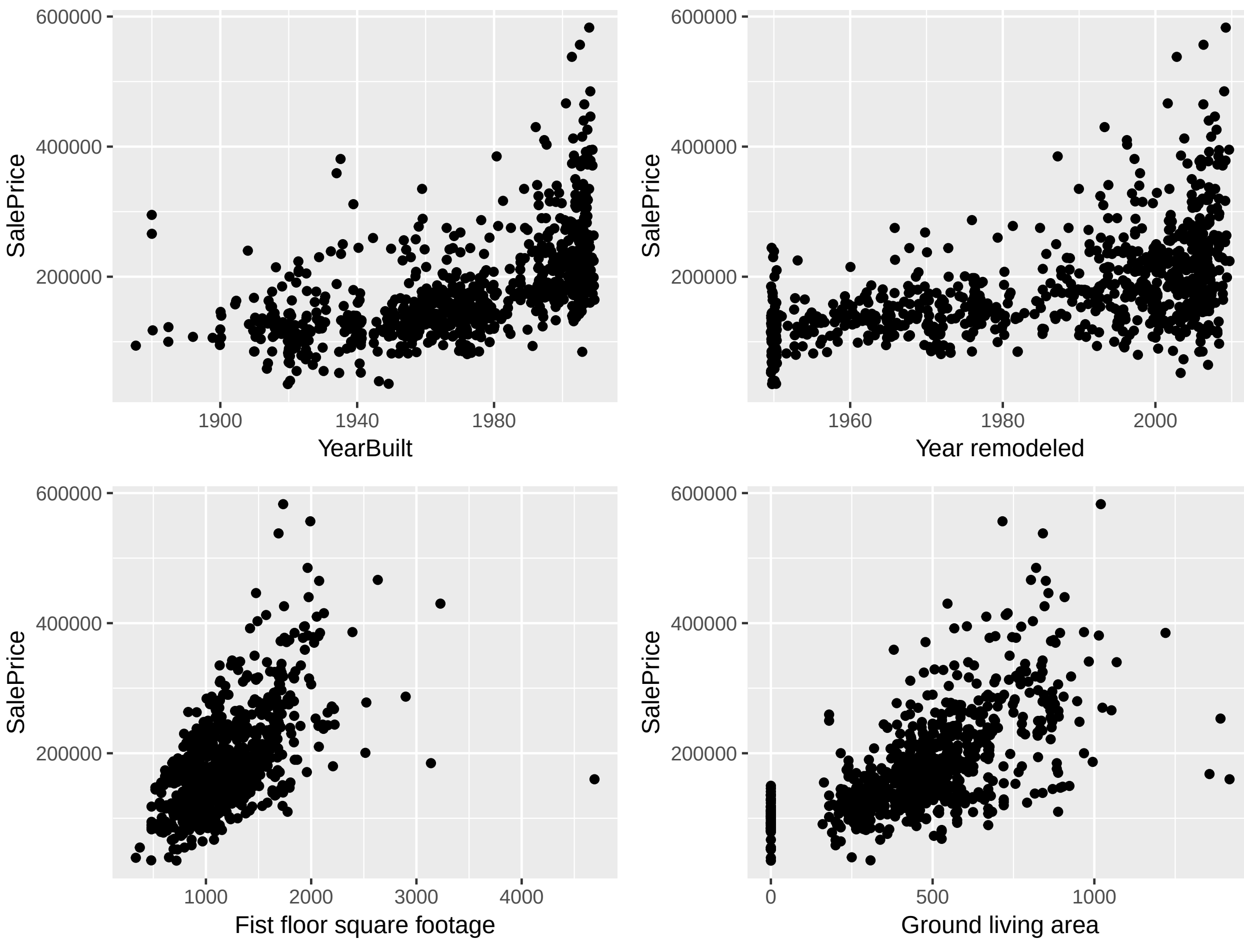






3.3 Scatterplots

Then, we create scatter plots. We can tell from the plot that year built, year remodeled, fist floor square footage, ground living area all might have linear relationship with saleprice.



3.4 Target variable

The SalePrice is the target variable and we are trying to predict it.
First, predicting the sale price makes business sense. Generally speaking, when valuing a house, we need to focus on its features. For example, how many full bathrooms above grade (ground)? One bathroom is just the minimum requirement. If there are two to three bathrooms, the house has a bigger size and experiences more functionality. It could be labeled as a "luxury" house and thus has a higher sale price. The logic is that SalePrice might be a function of other variables.
Second, we can observe some important relationships from the box-and-whisker plots and scatter plots. There are some linear relationships between other variables and the SalePrice. For instance, by looking at the scatter plot of "Ground living area and Sale Price", we can find that as the ground living area increases, the sale price ascends. More ground living area means a spacious and large-scale house. It is easy to understand that the house will be sold at a higher price.

4. Model Selection

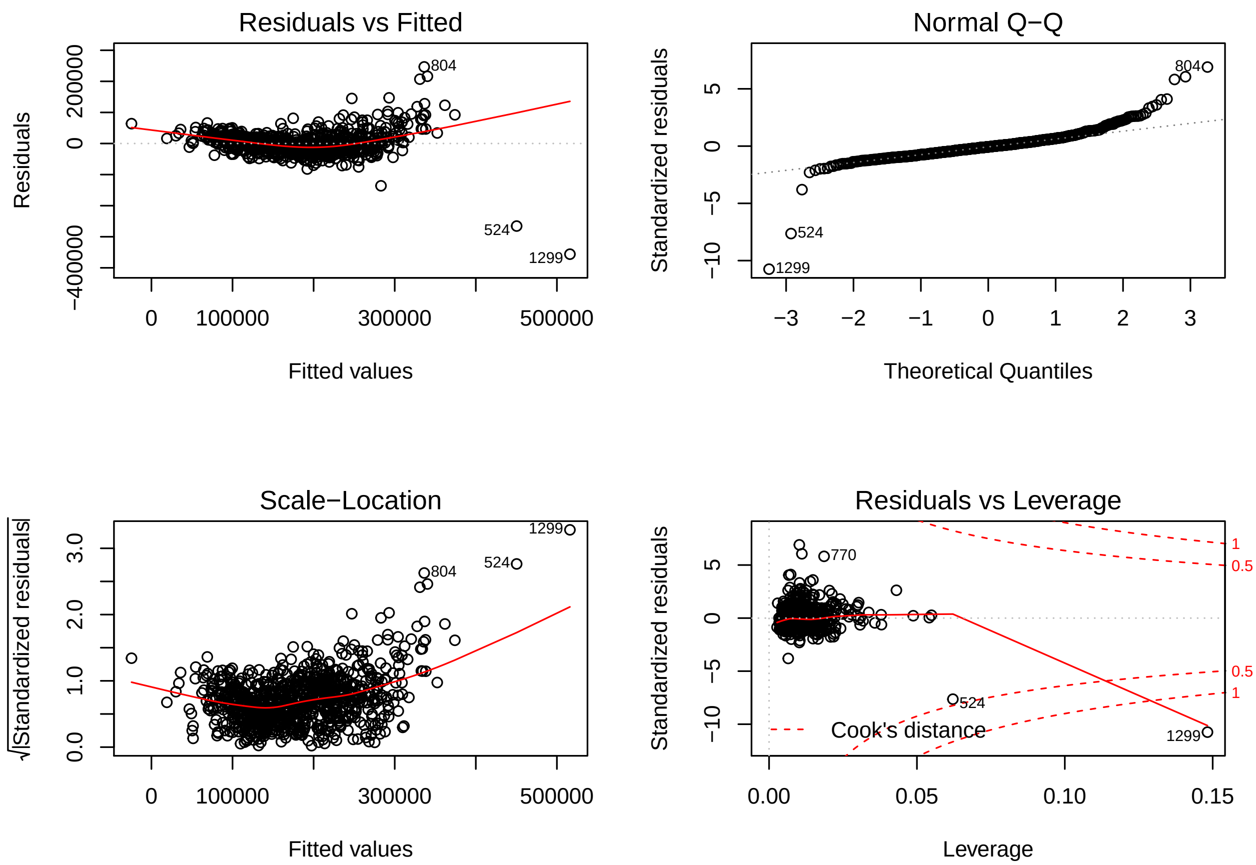
Based on the analysis above, we choose Sale Price as the target variable for the regression modeling process. In order to find the most appropriate dependent variables for the model, we use forward selection, backward selection, and forward-and-backward selection to narrow down our choices of variables. We use the 12 variables with the highest correlations with sale price or with most importance to the model we identify before as the original set of variables and let the computer select for us. After conducting these three selection methods, we find the results were all the same, which suggest a drop of 4 variables (*TotalBsmtSF, GarageArea, TotRmsAbvGrd, FullBath*). According to the output in R, the selections are based on the AIC of each model. R only keeps the model with lowest AIC. We then use the 8 remaining variables as the determinants of House Sale Price and run a multiple regression against them. The primary model is called lm.stepf, and the results are shown below.

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + GarageCars +
##     MSSubClass + YearBuilt + YearRemodAdd + Fireplaces + X1stFlrSF,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -356061  -18823   -2417   14134  246566
##
## Coefficients:
##                  Estimate   Std. Error t value            Pr(>|t|)
## (Intercept)  -1261589.078   149581.871  -8.434 < 0.0000000000000002 ***
## OverallQual     19341.747     1433.036  13.497 < 0.0000000000000002 ***
## GrLivArea          39.678        3.538  11.214 < 0.0000000000000002 ***
## GarageCars      15490.426     2238.941   6.919     0.00000000000886 ***
## MSSubClass       -177.162       31.277  -5.664     0.00000002008074 ***
## YearBuilt         241.637       58.111   4.158     0.00003526680759 ***
## YearRemodAdd      375.689       77.472   4.849     0.00000146773952 ***
## Fireplaces       9311.305     2192.662   4.247     0.00002405130567 ***
## X1stFlrSF          15.451        4.235   3.648              0.00028 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35900 on 867 degrees of freedom
## Multiple R-squared:  0.7713, Adjusted R-squared:  0.7692
## F-statistic: 365.5 on 8 and 867 DF,  p-value: < 0.00000000000000022
```

5. Model improvement

5.1 Model diagnose

To ensure that the assumptions of regression (OLS) are not being violated, we run several diagnoses to check for the model validation. The VIF of the variables are all smaller than 5 which indicates that there's no problem of collinearity. We then closely exam the residual plots of our model. The somewhat curvy Residual vs. Fitted plot shows that with the increase of fitted value, the residuals decrease at first and then increase. It may indicate that the true relationship between sale price and all the the house determinants is not linear. The QQ plot shows a fat tail, suggesting a non-normal distribution of the residuals.



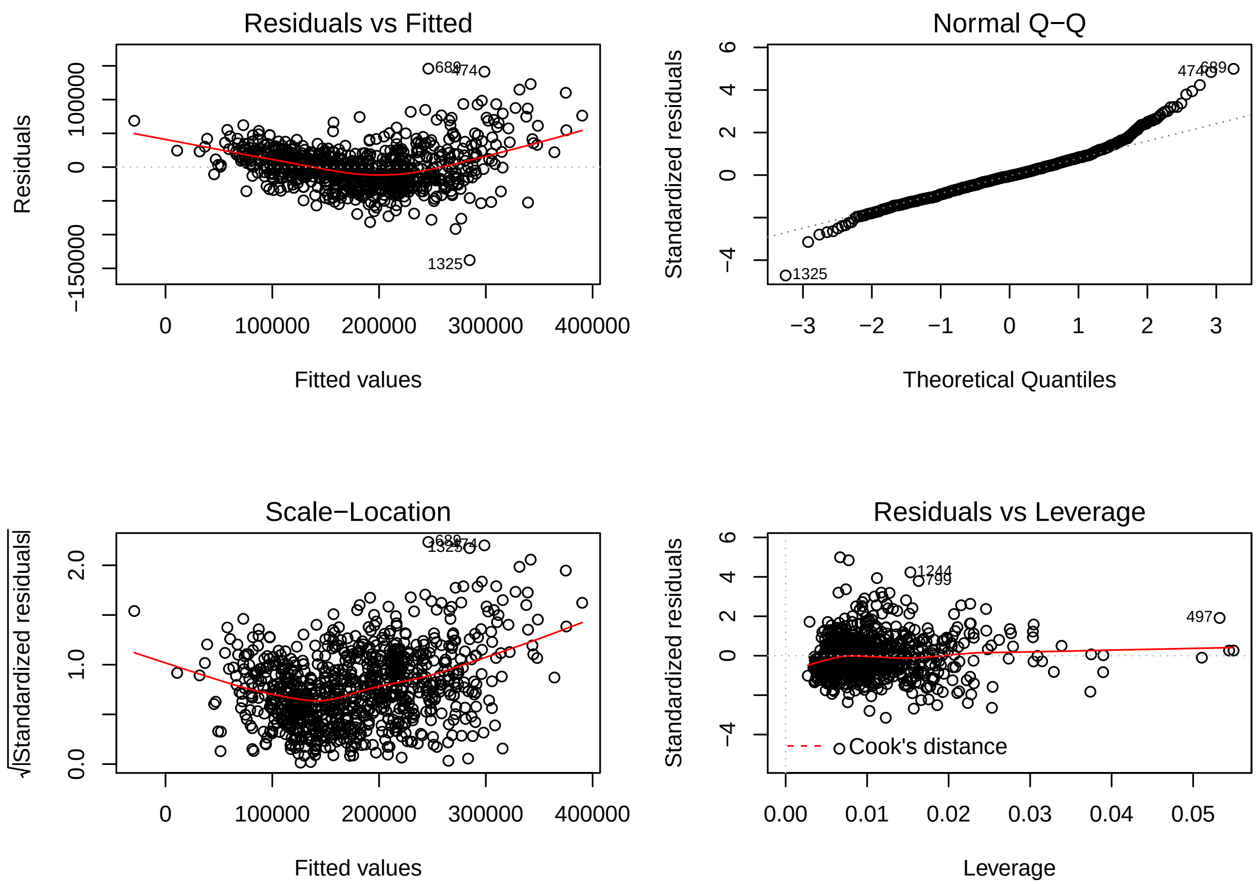

```
##   OverallQual    GrLivArea    GarageCars    MSSubClass     YearBuilt
##      2.603233     2.207502      1.874873      1.191147      2.126483
## YearRemodAdd   Fireplaces     X1stFlrSF
##      1.722516     1.345336      1.965705
```

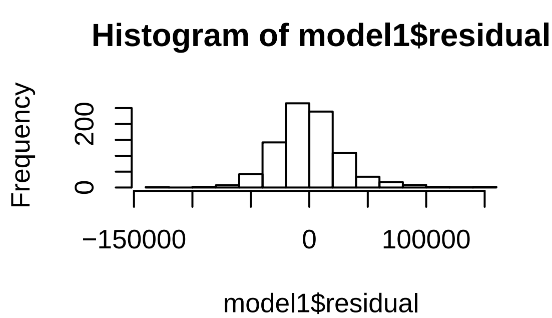
5.2 Deal with Outlier

By looking at all four plots, we notice that there are some outlier problems for our model. The outliers with index of 524,770, 804, 1047, and 1299 prevail in all four plots and are therefore removed from the dataset. After the removal of outliers, the new model, model1, has an improved adjusted R-Squared, lower AIC, and all statistically significant coefficients. The residual plots also improve a lot after the removal of the outliers. The histogram of residuals has a bell shape, which indicates a normal distribution of the residuals. Finally, we check for heteroskedasticity and fix the problem by robusting the standard errors. We would then conclude that the model 1 is the final best linear regression model to estimate the house sale price. The final model is shown below. The adjusted r-squared is 0.8322 and the RMSE using validation data is 39203.

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -137828  -17332   -1301   14882  145822
##
## Coefficients:
##                    Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)   -1316783.414   122218.099 -10.774 < 0.0000000000000002 ***
```

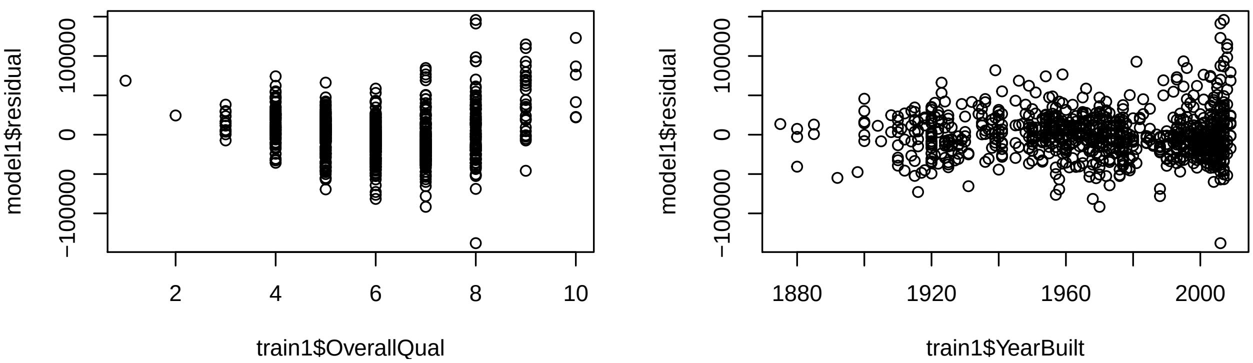
```
## OverallQual     18006.845     1173.593  15.343 < 0.0000000000000002 ***
## YearBuilt         279.472       47.569   5.875 0.000000006035496444 ***
## YearRemodAdd      360.522       63.258   5.699 0.000000016519515638 ***
## X1stFlrSF          29.593        3.570   8.290 0.000000000000000432 ***
## GrLivArea          46.940        3.018  15.556 < 0.0000000000000002 ***
## Fireplaces       8478.929     1800.389   4.709 0.000002893137214814 ***
## GarageCars      10152.849     1857.722   5.465 0.000000060548712390 ***
## MSSubClass       -147.347       25.595  -5.757 0.000000011908547014 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29300 on 862 degrees of freedom
## Multiple R-squared:  0.8337, Adjusted R-squared:  0.8322
## F-statistic: 540.3 on 8 and 862 DF,  p-value: < 0.000000000000000022
```



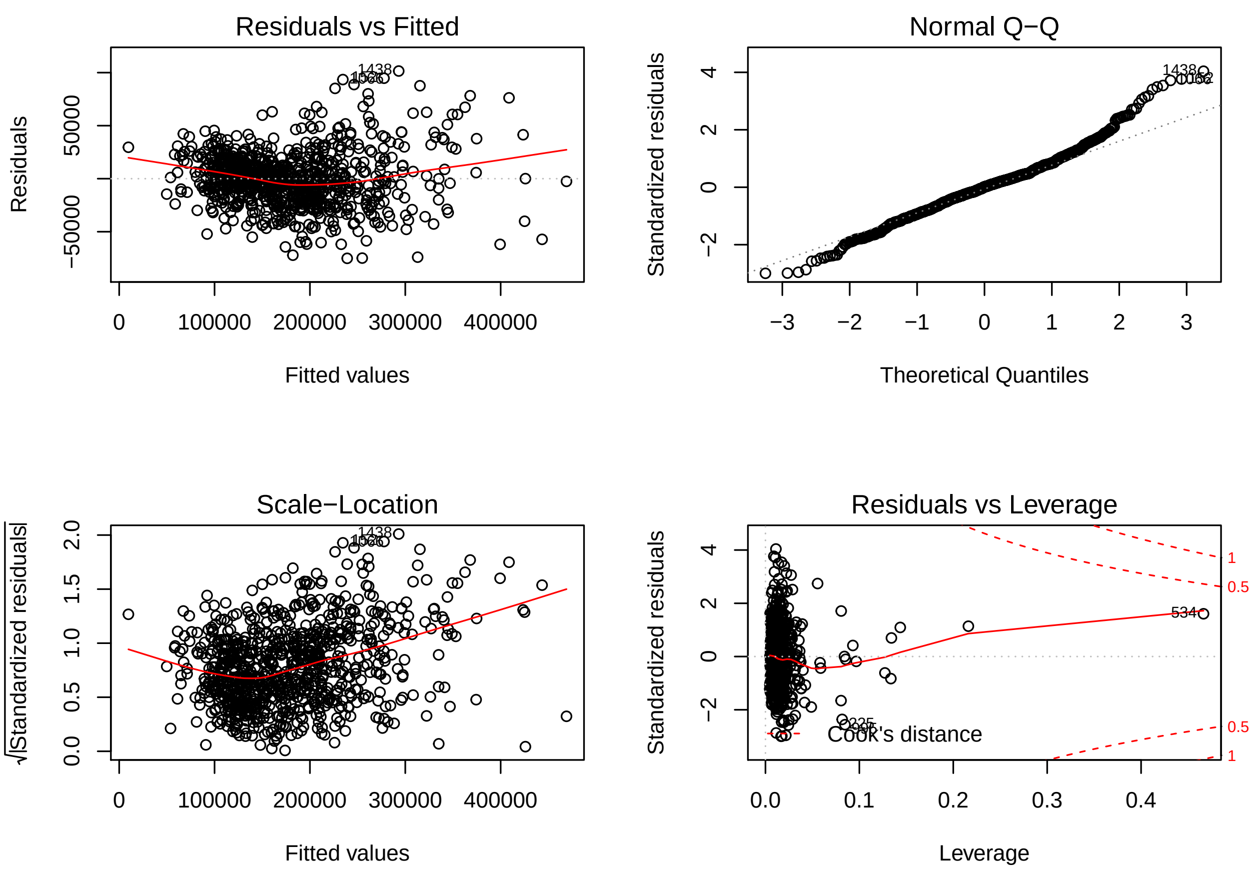


5.3 Add Polynomial Term

When we draw the residual independent varialble plots, we notice that there are high powers for OverallQual and YearBuilt (these plots have trends).



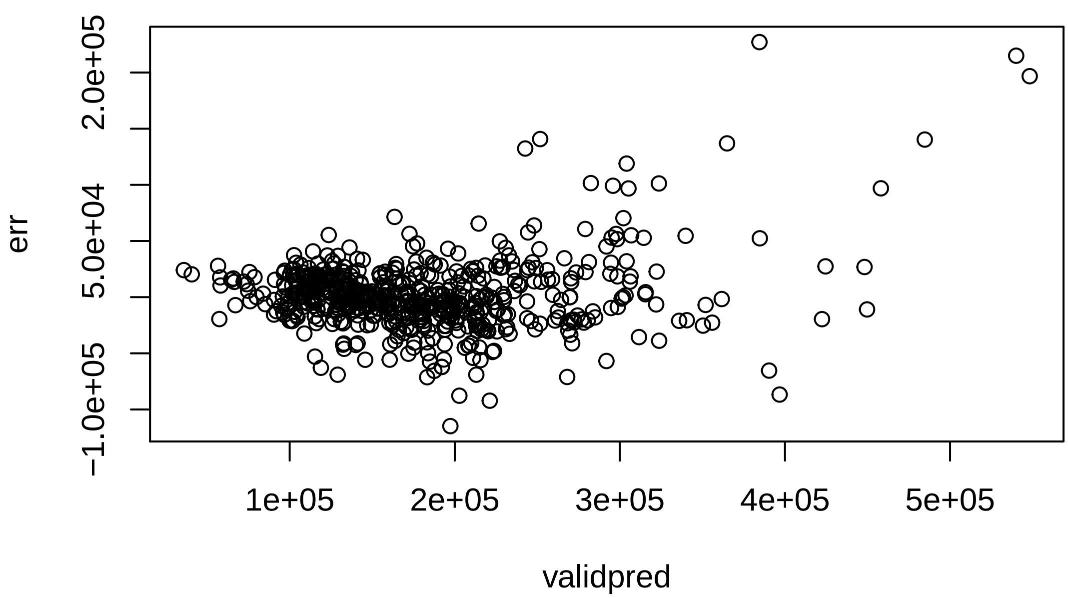
Therefore, we try the quadratic and cubic terms for these variables. We don't include powers higher than 3 in order to avoid overfit. After adding the polynomial terms and removal of outliers, the new model is called polymodel2, with imporved adjusted R-squared and smaller RMSE. Our final adjusted R-squared is 0.8723, and the RMSE for test data is 32911. The residual plots also improve a lot after add higher order terms and the removal of the outliers. The residual plots as well as the results for the final model are shown below.




```
##
## Call:
## lm(formula = SalePrice ~ . + I(YearBuilt^2e+00) + I(YearBuilt^3e+00) +
```

```
##     I(OverallQual^3e+00) + I(OverallQual^2e+00), data = train[-outlier_index2,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.5e+04 -1.6e+04  1.9e+02  1.3e+04  1.0e+05
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      -9.9e+07    2.1e+08 -5.0e-01    6e-01
## OverallQual       4.2e+04    1.5e+04  2.7e+00    7e-03 **
## YearBuilt         1.5e+05    3.2e+05  5.0e-01    6e-01
## YearRemodAdd      4.0e+02    5.9e+01  6.8e+00    2e-11 ***
## X1stFlrSF         1.9e+01    3.2e+00  5.7e+00    1e-08 ***
## GrLivArea         5.0e+01    2.7e+00  1.9e+01   <2e-16 ***
## Fireplaces        9.9e+03    1.6e+03  6.2e+00    1e-09 ***
## GarageCars        1.0e+04    1.6e+03  6.3e+00    6e-10 ***
## MSSubClass       -1.3e+02    2.2e+01 -5.9e+00    6e-09 ***
## I(YearBuilt^2)   -7.3e+01    1.6e+02 -4.0e-01    7e-01
## I(YearBuilt^3)    1.2e-02    2.8e-02  4.0e-01    7e-01
## I(OverallQual^3)  7.7e+02    1.4e+02  5.5e+00    5e-08 ***
## I(OverallQual^2) -9.4e+03    2.6e+03 -3.6e+00    3e-04 ***
## ---
## Signif. codes:  0e+00 '***' 1e-03 '**' 1e-02 '*' 5e-02 '.' 1e-01 ' ' 1e+00
##
## Residual standard error: 2.5e+04 on 855 degrees of freedom
## Multiple R-squared:  0.87,   Adjusted R-squared:  0.87
## F-statistic: 4.9e+02 on 1.2e+01 and 8.55e+02 DF,  p-value: <2e-16
```

We also draw a residual ~ predicted price plot in the in validation data. We can tell from the plot that the model is good since residuals are symmetrically distributed around 0 except for the 3 outliers.



6. Outliers

Among all, eight of our observations were removed outliers. The outliers accounted for about 0.9% of our observations. It's an acceptable number which would not lead to an obvious drop of total observation number.

7. Model Interpretation and reflection

Our final model is different because we have added higher-order terms including the square of original construction date, cube of original construction date, square of Overall Quality and cube of Overall Quality. The following table provides a simple intrepretation of the coefficients.

Variable Name	Coefficient	Interpretation	Process in terms of marketing
YearRemodAdd	402.479	With the date of the remodel one year later, the sales price of the house would increase by $402.479 on average.	When listing the house in the market, the remodel should be highlighted if the remodel date of the house is pretty close and the this could give the potential customer a feeling that this house rather new and functionable when comparing with other houses that built in the same year.
X1stFlrSF	18.512	With one more square feet in the first floor, the sales price of the house would increase by $18.512 on average.	The first-floor area suggests the living area or the common area of the house, so for the family with kids, the common area for them is very important.
GrLivArea	50.133	With one square feet of the house, the sales price would increase by $50.133 on average.	The above ground area suggests mostly the living area for bedrooms. When listing in the market, different customer has different requirements, so the bigger family the family is, the larger above ground living area is more preferred.
Fireplaces	9857.507	With one more fireplace in the house, the sales price would increase by $9857.507 on average.	Fireplace is not a very frequent choice in the house market these years, while for some customers who prefer Medieval European style, the fireplace maybe a good add-on item for the house. While, the cost on the fireplace is also expensive and the costs for the usage and maintaining are also expensive, so the sales price of the house increased a lot when the

			fireplaces are included.
GarageCars	10107.446	With one more car available in the garage, the sales price of the house would increase by $10107.466 on average.	Most family owns at least one car, so the garage is a must-have item for the house to be attractive in the markets. Also, the larger the garage is, the more attractive the house is in the market since many families own more than one car or have the intend to purchase more cars in the future.
MSSubClass	-131.993	With one level of the class increases in the building class of the house, the sales price decrease by $131.993 on average.	The class of the house suggests the quality of the house overall, and in extreme weather, the better the class is the more attractive the house is in the market. So, with high level of class, the house could be more preferred.
YearBuilt **I(YearBuilt^2)** **I(YearBuilt^3)**	146717.559 -73.273 0.012	When year of the house-built increases by one year, the house price increases by 0.036*year of the house built^2-146*year of the house built+146645 dollars.	The year the house is built determines the first impression of the functionable ability of the house when customers look at it.
OverallQual **I(OverallQual^3)** **I(OverallQual^2)**	41875.654 769.683 -9377.121	When overall quality rating of house increases by 1 unit, the house price will increase by 2310*overall quality rating of house^2-16444*overall quality rating of house+33269 dollars.	The quality and the area of the house are two main considerations most customers have when searching for a potential house to purchase. So, the houses with higher quality are generally more preferable, while the cost performance between the quality and the house price is also important. In the market, the house with the best cost performance between the quality and the house price is the most attractive.

Figure 1: