Predict House Price of Residential Homes in Iowa

Linear Regression Model

Deepanshu Goyal | 9th March 2020

Predict sale price of residential homes in Ames, Iowa using linear regression

14 features out of 80 are used to predict house price with 88.02% accuracy

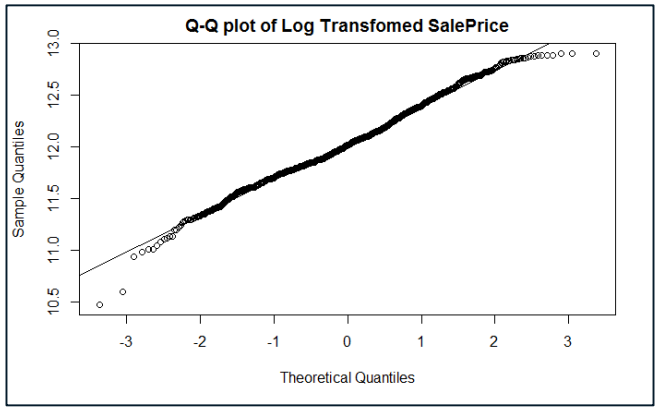
1. Overall material & finish quality of the house
2. Remodel Year
3. Number of fireplace in the house
4. Living area above ground (In sq. feet)
5. Total basement area (In sq. feet)
6. Garage area (In sq. feet)
7. Finished basement area (In sq. feet)
8. Lot area (In sq. feet)
9. Original construction date of the house
10. Wood deck area (In sq. feet)
11. Open porch area (In sq. feet)
12. Class of building
13. Heating condition & quality
14. Kitchen quality

are significant parameters in predicting house prices in Ames, Iowa

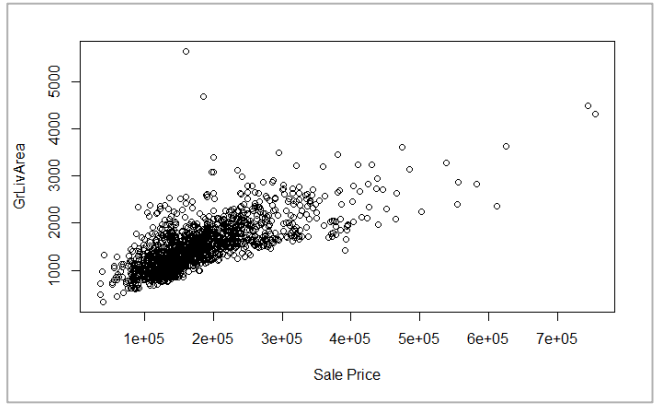
```
Call:
lm(formula = SalePrice ~ ., data = train_dataset1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82010 -0.05953  0.00782  0.07169  0.49117

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.596e+00  4.736e-01   9.705  < 2e-16 ***
OverallQual   7.599e-02  4.494e-03  16.909  < 2e-16 ***
YearRemodAdd  1.985e-03  2.389e-04   8.307 2.42e-16 ***
Fireplaces    4.296e-02  6.434e-03   6.676 3.60e-11 ***
GrLivArea     2.351e-04  1.015e-05  23.158  < 2e-16 ***
TotalBsmtSF   9.758e-05  1.160e-05   8.410  < 2e-16 ***
GarageArea    1.556e-04  2.468e-05   6.304 3.95e-10 ***
BsmtFinSF1    1.040e-04  9.213e-06  11.289  < 2e-16 ***
LotArea       2.471e-06  3.597e-07   6.870 9.87e-12 ***
YearBuilt     1.052e-03  1.823e-04   5.772 9.74e-09 ***
WoodDeckSF    9.706e-05  2.876e-05   3.376 0.000758 ***
OpenPorchSF   1.980e-04  5.827e-05   3.398 0.000699 ***
MSZoningFV    4.451e-01  4.796e-02   9.281  < 2e-16 ***
MSZoningRH    3.571e-01  5.657e-02   6.314 3.72e-10 ***
MSZoningRL    4.331e-01  4.459e-02   9.714  < 2e-16 ***
MSZoningRM    3.356e-01  4.503e-02   7.452 1.66e-13 ***
HeatingQC    -1.262e-02  2.343e-03  -5.389 8.40e-08 ***
KitchenQual  -2.258e-02  5.575e-03  -4.049 5.44e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1229 on 1314 degrees of freedom
Multiple R-squared:  0.8817,    Adjusted R-squared:  0.8802
F-statistic: 576.3 on 17 and 1314 DF,  p-value: < 2.2e-16
```

Model is validated against all the assumptions of linear regression & engineered to reduce the impact of outliers

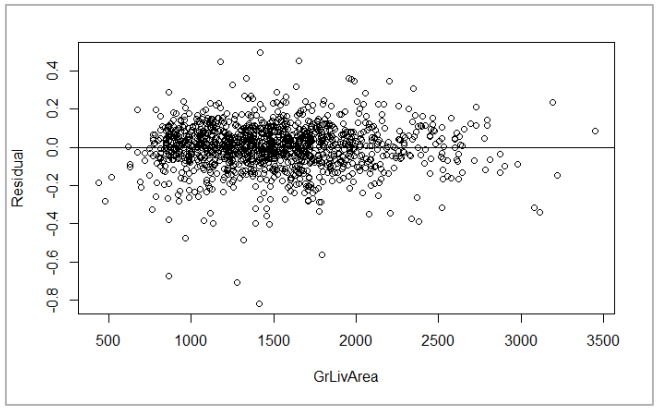
Multivariate Normality



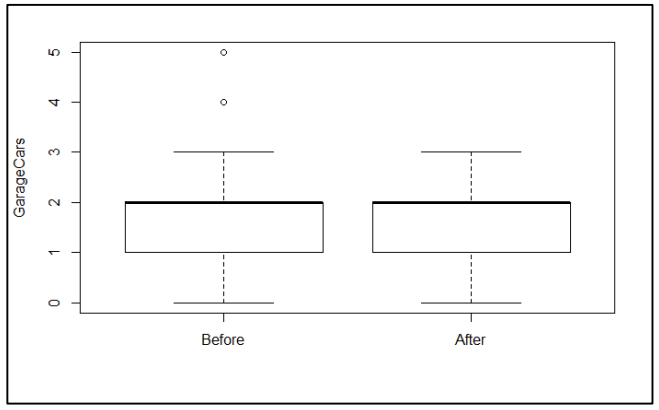
Linearity



Homoscedasticity



Outliers (Removed)



Besides all the good work there are some improvement areas

- Not all the outliers are handled effectively in the model
- Feature Engineering can be done in a better way to reduce the number of columns by merging some of the features together
- Advance regression techniques can be used to improve the feature selection & in-turn overall accuracy of the model
- Any change in the dataset structure will require change in the code. Additional efforts are required for maintaining coding standards

Our high level approach to linear regression

Divide the dataset into numerical continuous, numerical discrete & categorical features

1 Features Exploration & Elimination using EDA

- Check the distribution on target feature using QQ plot and histogram
- EDA on Numerical Continuous Features (Scatter Plots & Correlation Matrix)
- EDA on Numerical Discrete Features (Bar Graphs)
- EDA on Categorical Features (Bar Graphs)

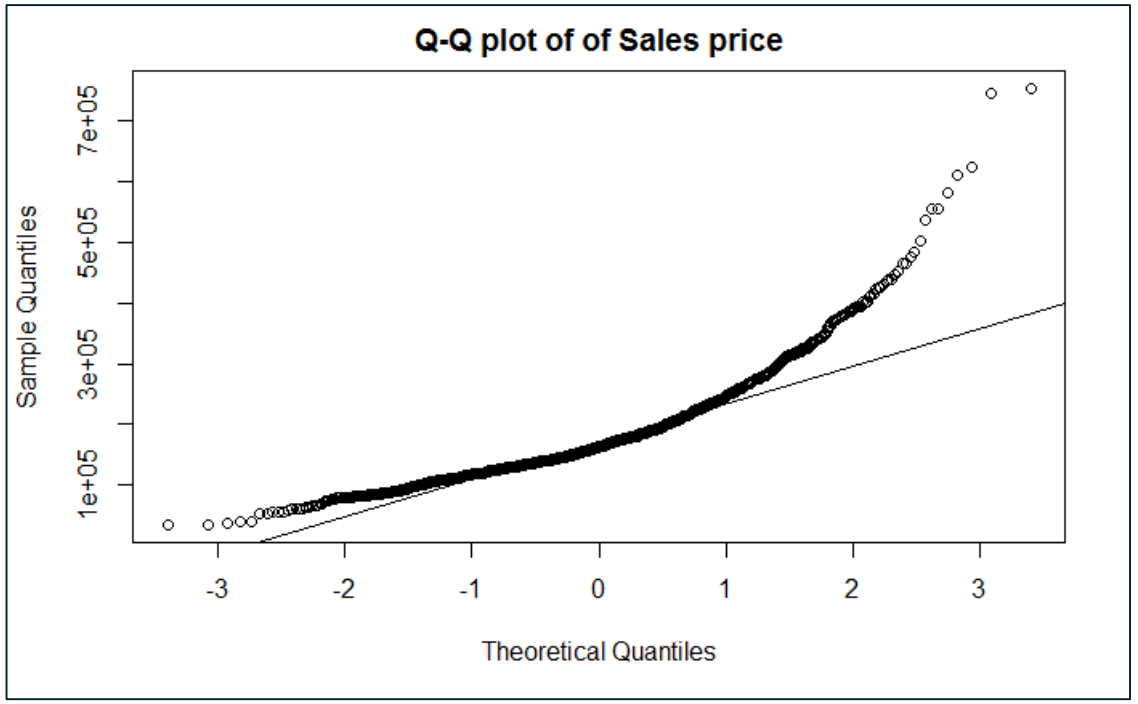
2 Feature Engineering & Data Engineering

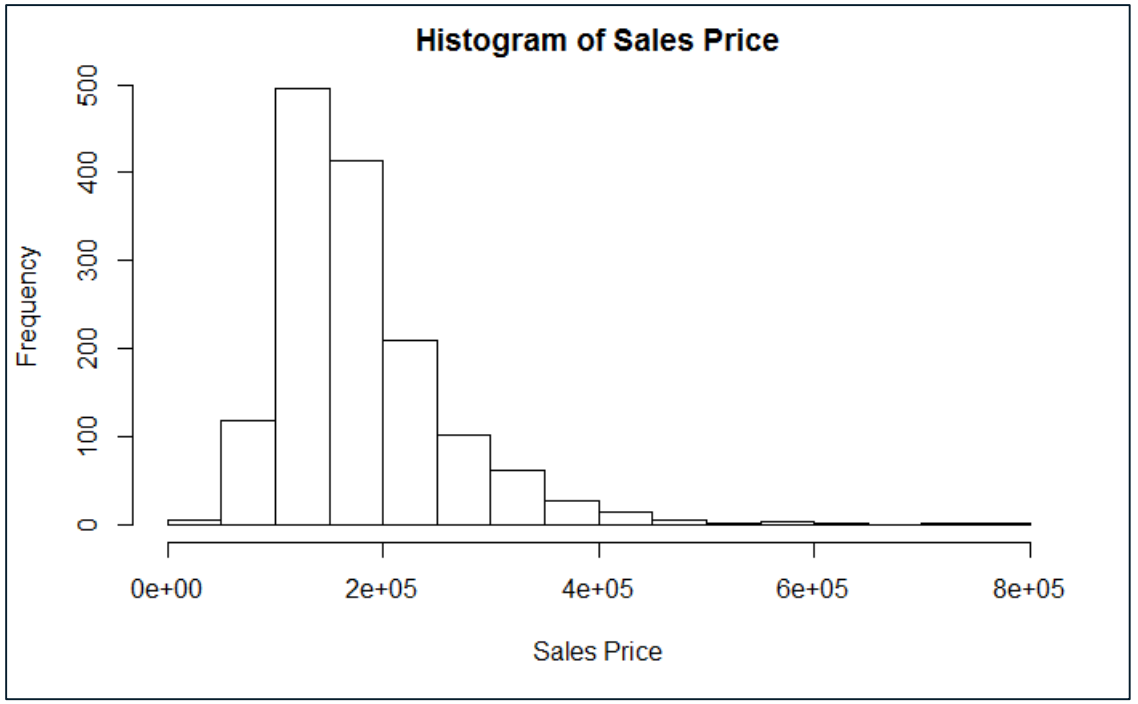
- For numeric continuous data, remove outliers using box plots
- For categorical data, replace ordinal values with numerical values to make it discrete feature
- Remove null values for numerical data & add “Missing”/0 values for categorical/discrete data
- Transform target variable using log function to make normally distributed

3 Modelling, Backward Elimination & Validate Assumptions

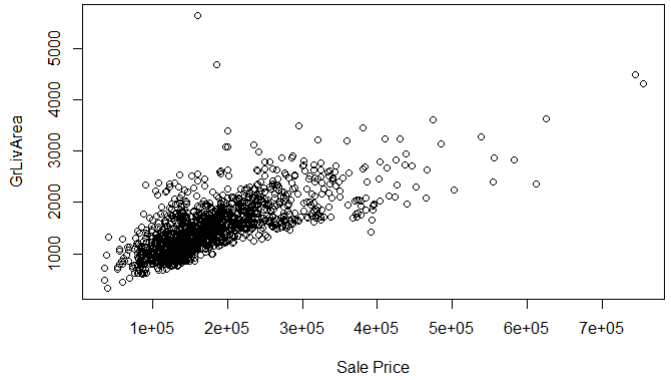
- Fit a linear model and eliminate features that are not significant one at a time (addition & removal both)
- Validate linear regression assumptions i.e. Linearity, Multivariate Normality, Homoscedasticity, Independence of Errors & Multi Collinearity in the model
- Once everything looks good, predict the house price with final set of 14 features

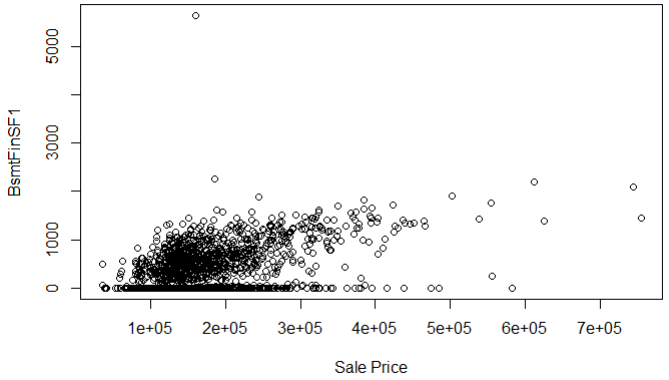
Sale Price distribution is skewed, we need to fix it before fitting it into our linear model

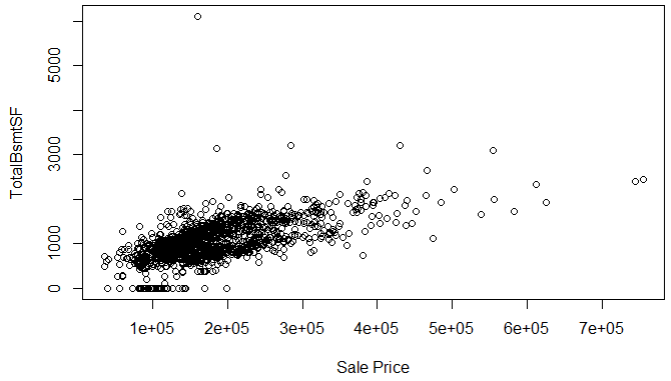


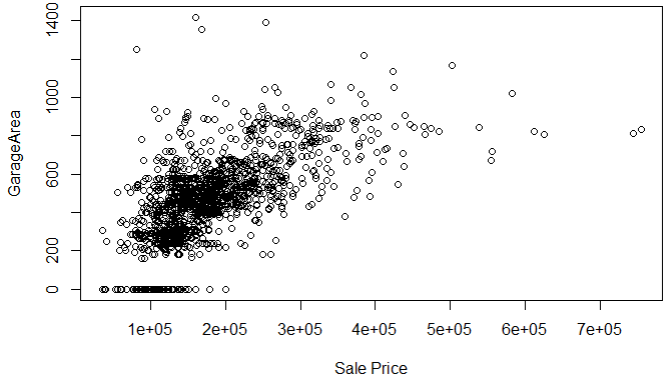


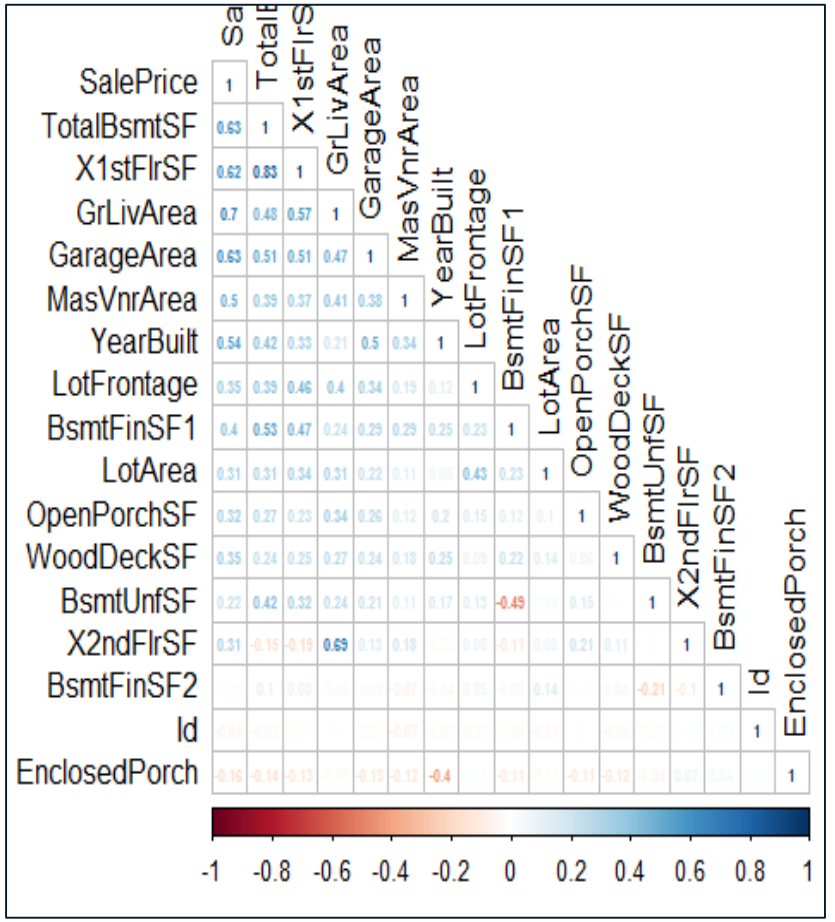
Living area, basement area, wood deck area, open porch area, basement finished area etc showing strong linear relationship with sale price



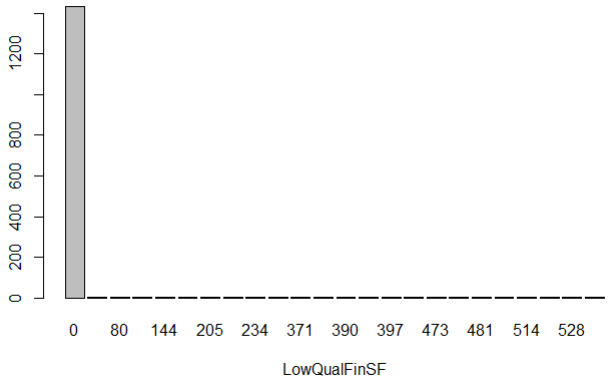


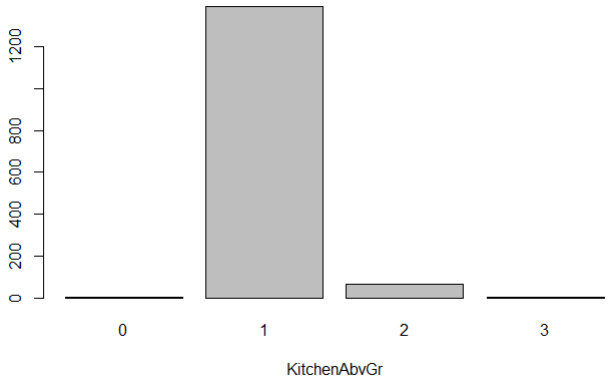


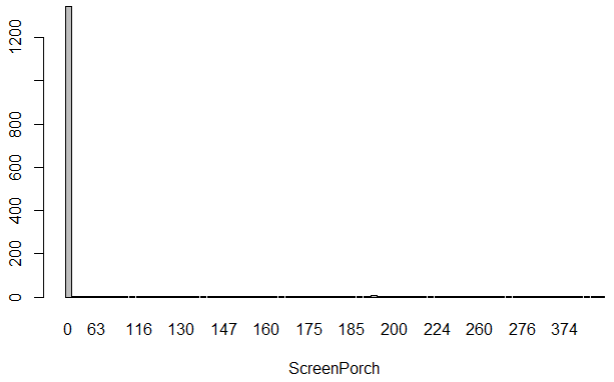


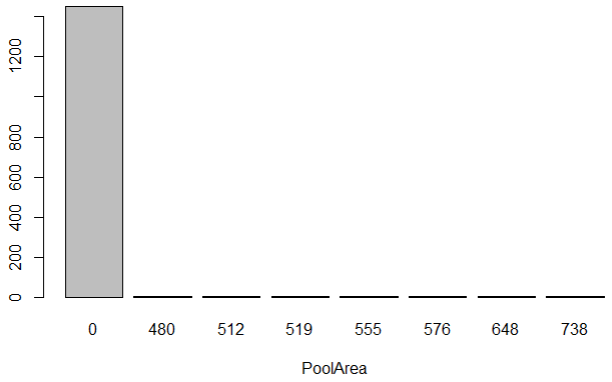


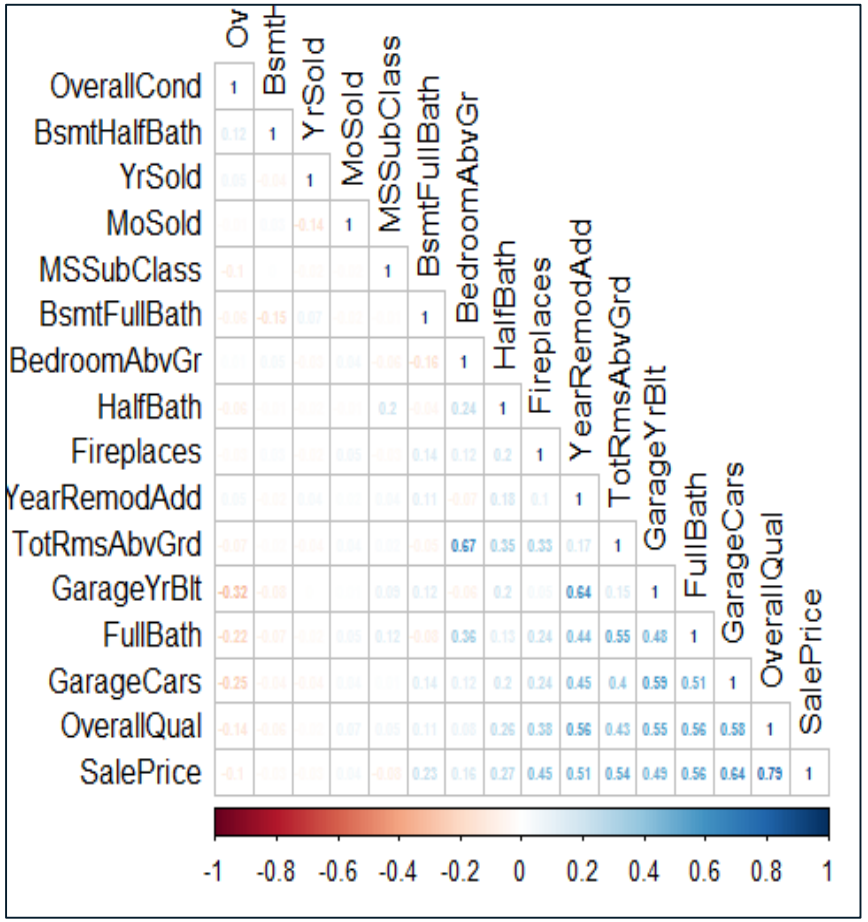
Number of kitchen, pool area, screen porch area etc showing low variance and can be easily drop whereas overall quality, garage capacity, total rooms & bathroom are strongly related with sale price



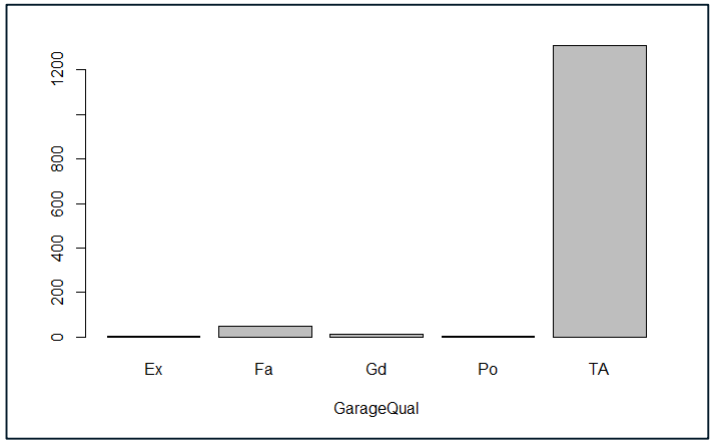


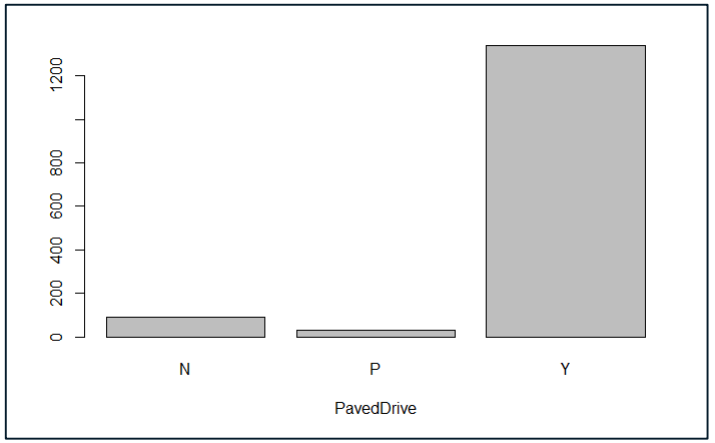


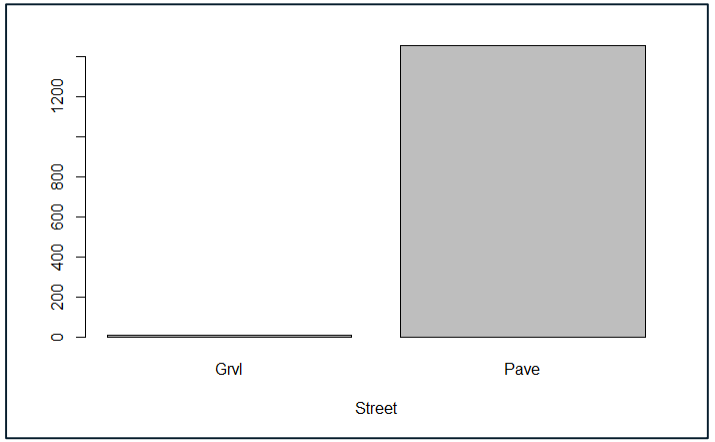


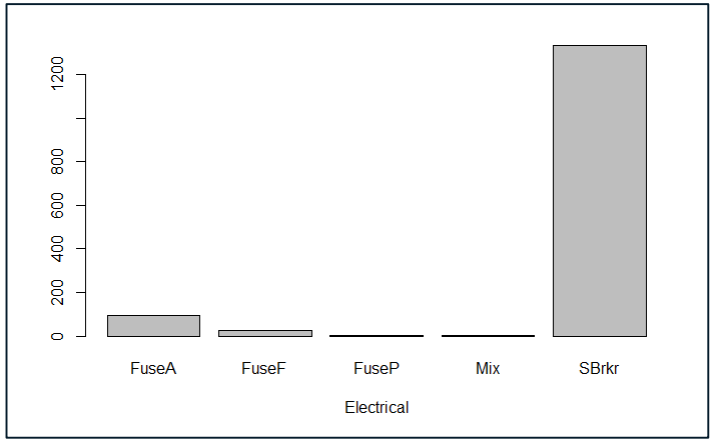


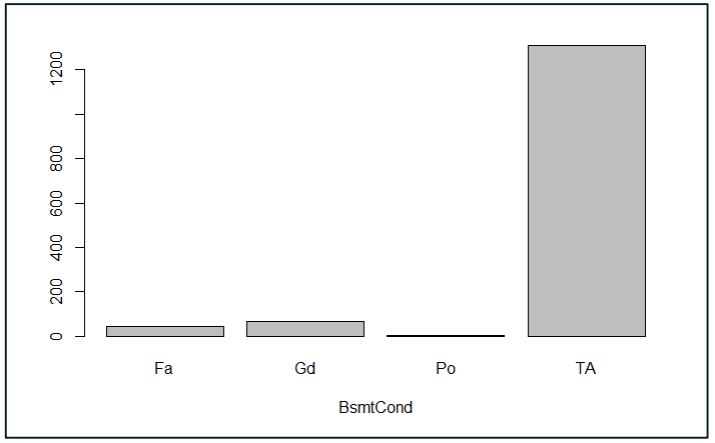
Garage quality, garage condition, pool quality, fence quality, electric functions, alley access, slope of house, roof material etc having low variance and be easily drop. There are 24 such feature

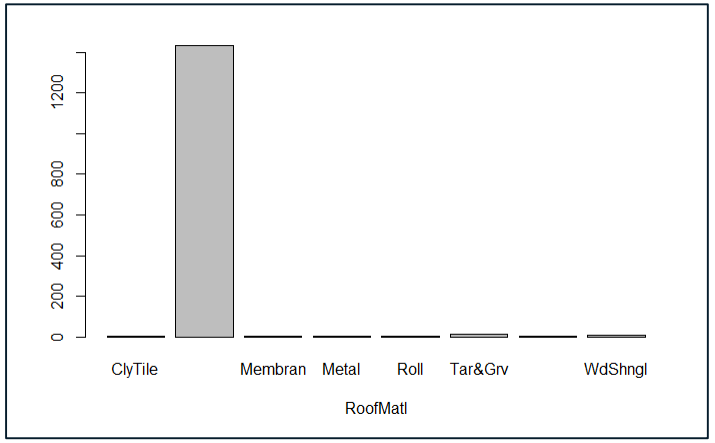




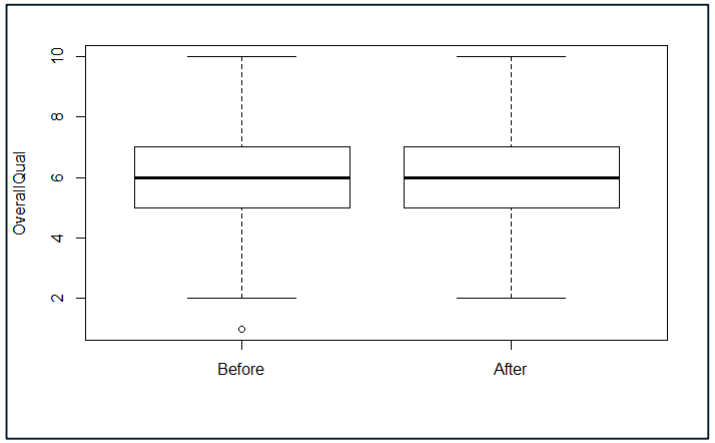


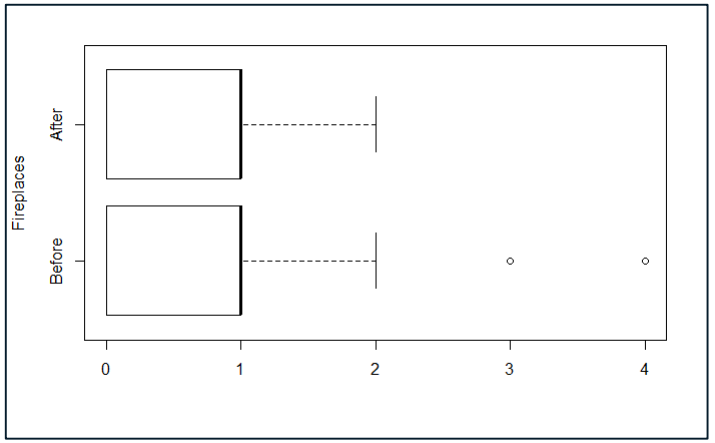


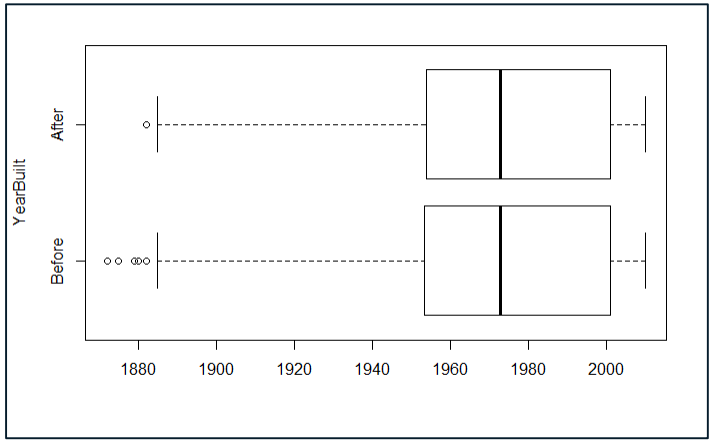


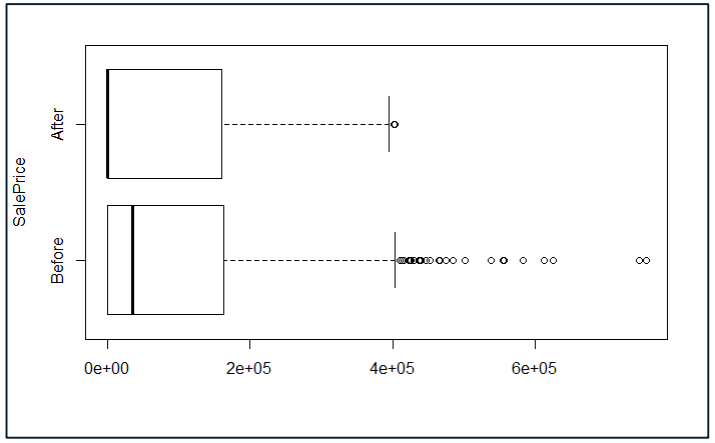


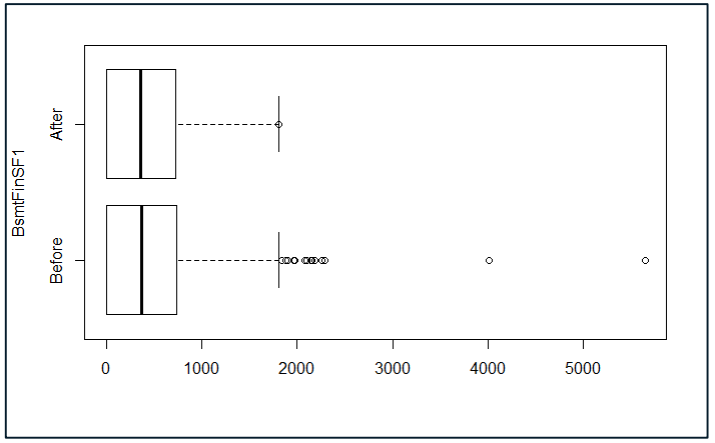
Outliers in sale price, number of fireplaces, original construction date, overall quality etc are removed to minimize their impact on the linear regression model

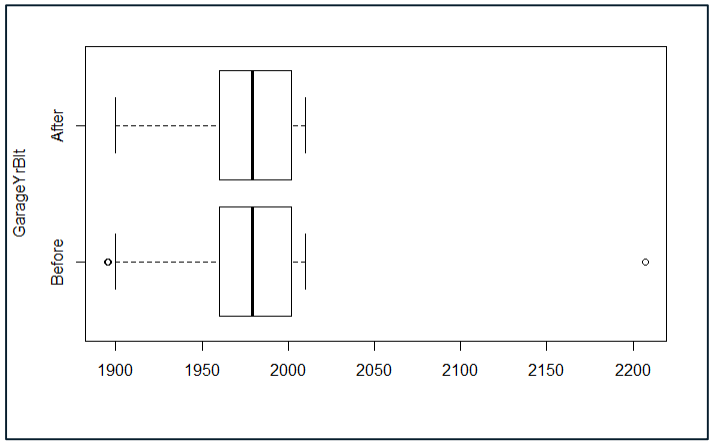




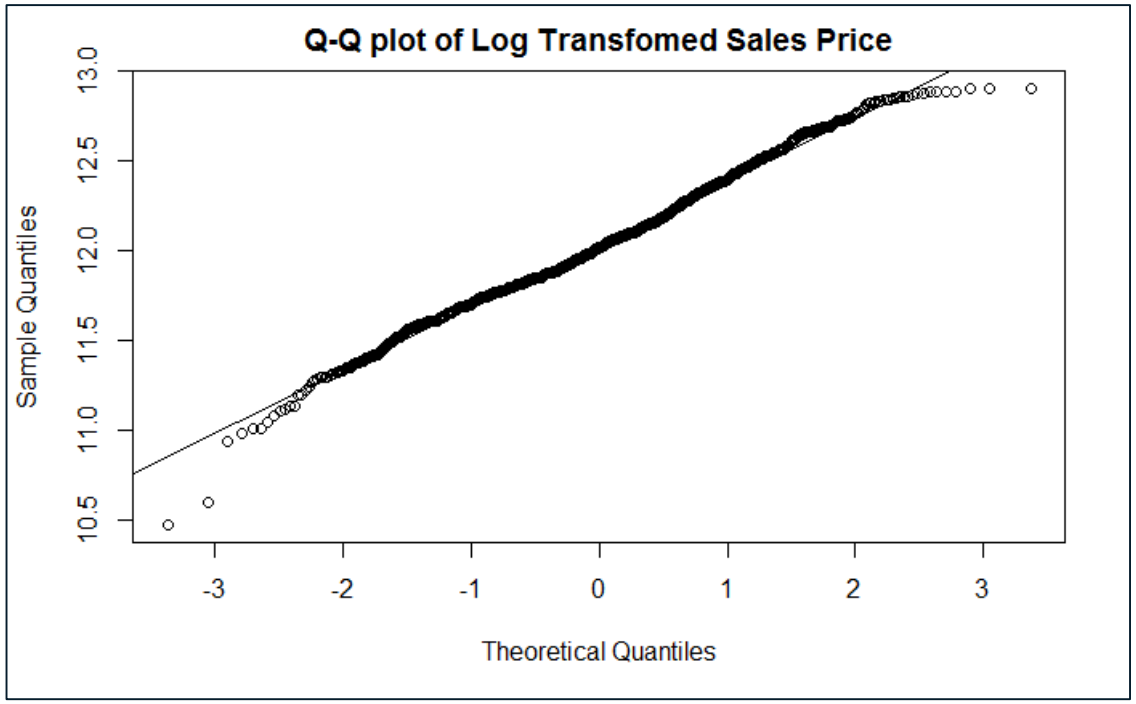


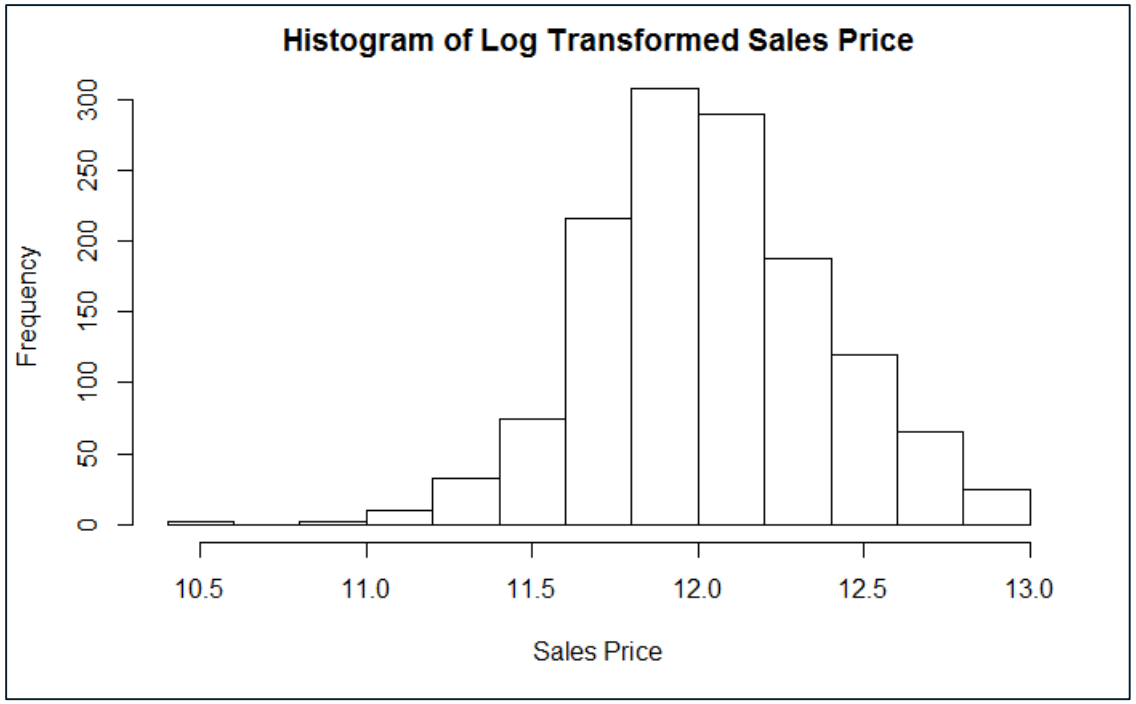






Sale Price is LOG transformed to follow normal distribution





Fit a linear model, using backward elimination remove features that are not significant, one at a time (addition & removal). Final model has 14 features

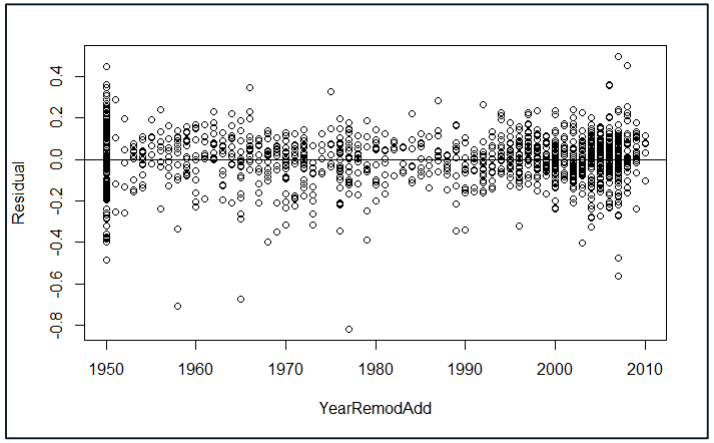
```
Call:
lm(formula = SalePrice ~ ., data = train_dataset1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82010 -0.05953  0.00782  0.07169  0.49117

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.596e+00  4.736e-01   9.705  < 2e-16 ***
OverallQual  7.599e-02  4.494e-03  16.909  < 2e-16 ***
YearRemodAdd 1.985e-03  2.389e-04   8.307 2.42e-16 ***
Fireplaces   4.296e-02  6.434e-03   6.676 3.60e-11 ***
GrLivArea    2.351e-04  1.015e-05  23.158  < 2e-16 ***
TotalBsmtSF  9.758e-05  1.160e-05   8.410  < 2e-16 ***
GarageArea   1.556e-04  2.468e-05   6.304 3.95e-10 ***
BsmtFinSF1   1.040e-04  9.213e-06  11.289  < 2e-16 ***
LotArea      2.471e-06  3.597e-07   6.870 9.87e-12 ***
YearBuilt    1.052e-03  1.823e-04   5.772 9.74e-09 ***
WoodDeckSF   9.706e-05  2.876e-05   3.376 0.000758 ***
OpenPorchSF  1.980e-04  5.827e-05   3.398 0.000699 ***
MSZoningFV   4.451e-01  4.796e-02   9.281  < 2e-16 ***
MSZoningRH   3.571e-01  5.657e-02   6.314 3.72e-10 ***
MSZoningRL   4.331e-01  4.459e-02   9.714  < 2e-16 ***
MSZoningRM   3.356e-01  4.503e-02   7.452 1.66e-13 ***
HeatingQC   -1.262e-02  2.343e-03  -5.389 8.40e-08 ***
KitchenQual -2.258e-02  5.575e-03  -4.049 5.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1229 on 1314 degrees of freedom
Multiple R-squared:  0.8817,    Adjusted R-squared:  0.8802
F-statistic: 576.3 on 17 and 1314 DF,  p-value: < 2.2e-16
```

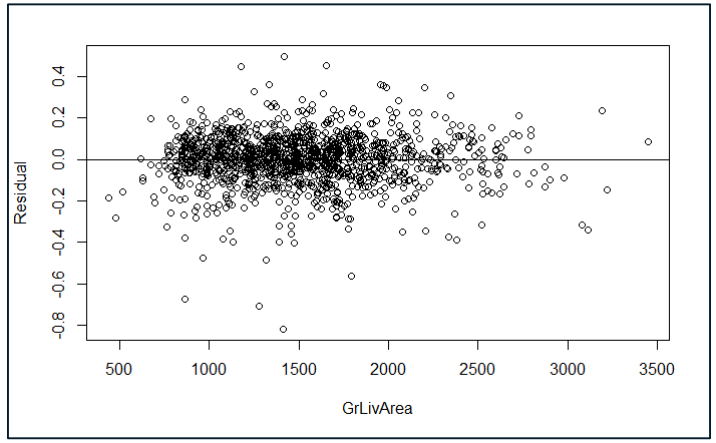
14 features out of 80 are used to predict house price with 88.02% accuracy

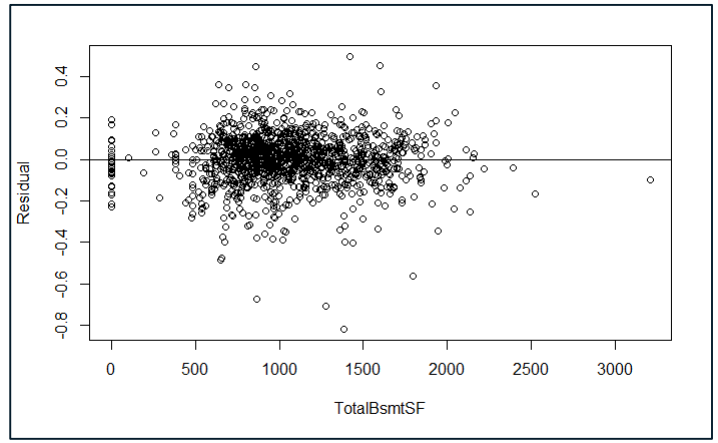
1. Overall material & finish quality of the house
2. Remodel Year
3. Number of fireplace in the house
4. Living area above ground (In sq. feet)
5. Total basement area (In sq. feet)
6. Garage area (In sq. feet)
7. Finished basement area (In sq. feet)
8. Lot area (In sq. feet)
9. Original construction date of the house
10. Wood deck area (In sq. feet)
11. Open porch area (In sq. feet)
12. Class of building
13. Heating condition & quality
14. Kitchen quality

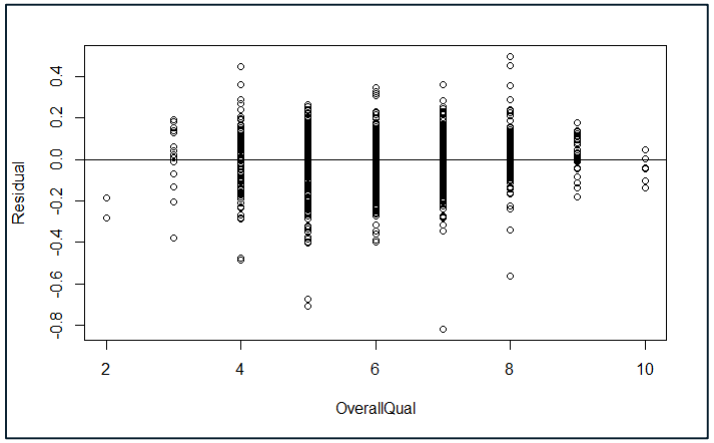
are significant parameters in predicting house prices in Ames, Iowa

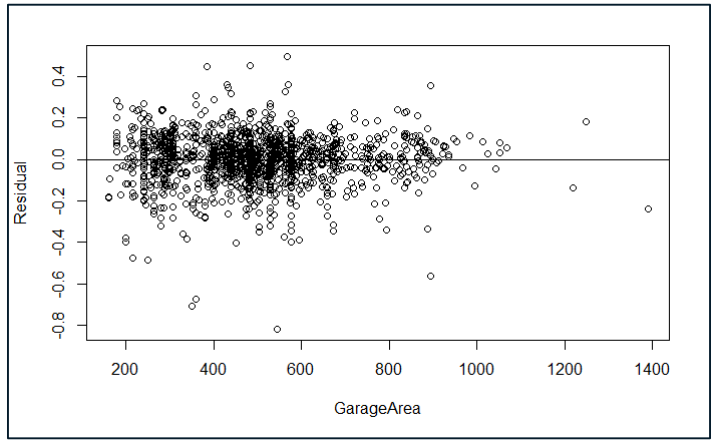
Validate Homoscedasticity (Constant Error Variance) & independence of errors using residual plots. All looks good !!

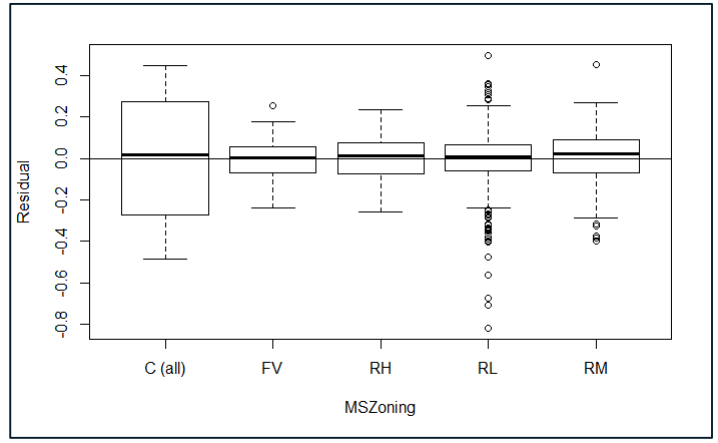




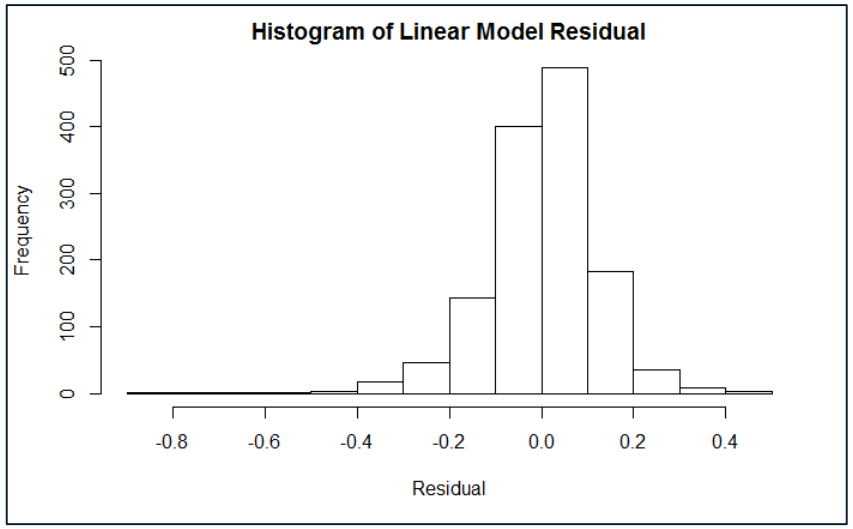


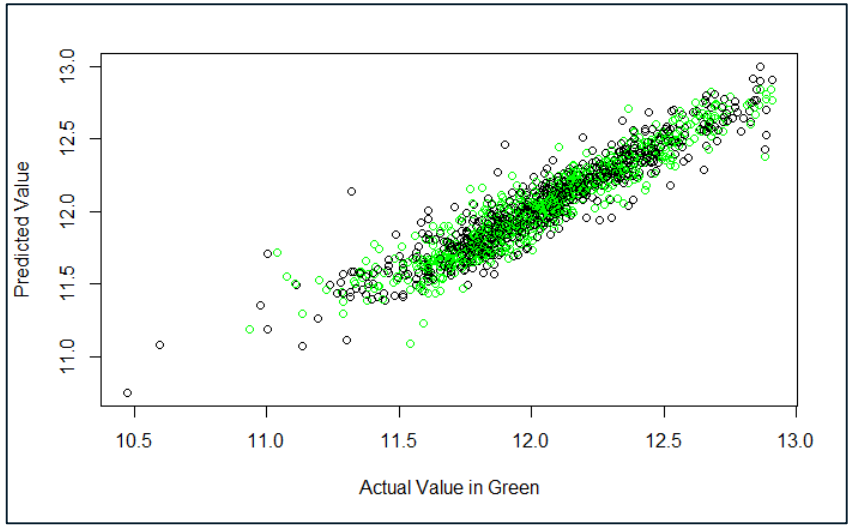






Residuals follow normal distribution with mean close to zero, predicted values (in green) are very close to actual values, VIF is within range so there is no/little multicollinearity . All looks good!!





VIF(Variance Inflation factor)

	GVIF	Df	GVIF^(1/(2*Df))
OverallQual	2.931072	1	1.712037
YearRemodAdd	2.117289	1	1.455091
Fireplaces	1.421079	1	1.192090
GrLivArea	1.902848	1	1.379437
TotalBsmtSF	1.779301	1	1.333905
GarageArea	1.682654	1	1.297172
BsmtFinSF1	1.298571	1	1.139549
LotArea	1.163635	1	1.078719
YearBuilt	2.453721	1	1.566436
WoodDeckSF	1.140169	1	1.067787
OpenPorchSF	1.203130	1	1.096873
MSZoning	1.565414	4	1.057618
HeatingQC	1.465502	1	1.210579
KitchenQual	1.650866	1	1.284860

Everything looks good, let’s predict house prices

Thank You !