Pooja Deshpande

Modeling Assignment 3
MSDS – 410 Data Modeling for Supervised Learning,
Summer 2020
Northwestern University

Contents:

Q.1

Y, response variable = *SalePrice*

Ames dataset consists of several categorical variables. Out of these, the following were considered to be predictive of *SalePrice.*

Zoning, Street, Alley, Utilities, LotConfig, Neighborhood, BldgType, HouseStyle, BsmtQual, BsmtCond, CentralAir, KitchenQual, GarageFinish, PoolQC, MiscFeature, OverallQual, OverallCond.

To narrow down this list, regression models were built with each of the above categorical variables and the ones which had the highest r-squared values were chosen.

	cat_var	r_squared
16	OverallQual	0.623
14	PoolQC	0.576
6	Neighborhood	0.571
9	BsmtQual	0.445
12	KitchenQual	0.415
3	Alley	0.293
13	GarageFinish	0.268
1	Zoning	0.118
15	MiscFeature	0.073
8	HouseStyle	0.072
11	CentralAir	0.058
10	BsmtCond	0.025
5	LotConfig	0.016
17	OverallCond	0.016
7	BldgType	0.013
2	Street	0.000
4	Utilities	0.000

Based on the r-squared values, the top 5 categorical variables were chosen for further analysis – *OverallQual, PoolQC, Neighborhood, BsmtQual, KitchenQual.*

Mean *SalePrice* by the levels of the categorical variables:

```
OverallQual
   Category         x
1         1  50150.00
2         2  58667.11
3         3  87819.79
4         4 112305.82
5         5 136426.11
6         6 163884.23
7         7 205216.69
8         8 264823.31
9         9 338116.85
10       10 344687.50

PoolQC
  Category      x
1       Ex 315000
2       Fa 215500
3       Gd 215500
4       TA 170500
```

It is evident from the above table that the mean sale price increases as the *OverallQual* of the home increases. Since *OverallQual* is a numerical variable and has several levels, it can also be considered as a continuous variable.

For *PoolQC,* two levels have the same mean value of *SalePrice.* An important point here is the missing category, ‘No pools’. These are denoted by NA values. After further inspection of this column, it was found that about 99.6% of the values are NA. Hence, a binary variable which indicates the presence or absence of a pool, was thought to be more appropriate for this situation.

```
> final_df$Pool <- ifelse(is.na(final_df$PoolQC),0,1)
> aggregate(SalePrice~Pool,final_df,mean)
  Pool SalePrice
1    0  176778.1
2    1  211062.5
```

```
BsmtQual
  Category        x
1       Ex 296228.6
2       Fa 115471.9
3       Gd 202731.2
4       TA 142645.7

KitchenQual
  Category        x
1       Ex 297740.7
2       Fa 111061.5
3       Gd 207891.4
4       Po 107500.0
5       TA 143723.9
```

It can be observed from the above table that the mean price of a home with high quality basement and kitchen is much higher than mean price of a home with lower quality basement and kitchen.

It is important to note that *BsmtQual* also consisted of null values which indicates the absence of a basement. This new category was added, and 5 dummy variables were created. The default category for *BsmtQual* is the 'no basement' category. Similarly, 4 dummy variables were created for *KitchenQual.* The default category for *KitchenQual is* the poor-quality category, 'Po'. Another point to note is that the *KitchenQual* column consists of only one instance in the 'Po' category.

***BsmtQual* dummy coded variables:**

	BsmtQual	BsmtQual_Ex	BsmtQual_Ta	BsmtQual_Po	BsmtQual_Gd	BsmtQual_Fa
1	TA	0	1	0	0	0
2	TA	0	1	0	0	0
3	TA	0	1	0	0	0
4	TA	0	1	0	0	0
5	Gd	0	0	0	1	0

***KitchenQual* dummy coded variables:**

	KitchenQual	KitchenQual_Ex	KitchenQual_Ta	KitchenQual_Gd	KitchenQual_Fa
1	TA	0	1	0	0
2	TA	0	1	0	0
3	Gd	0	0	1	0
4	Ex	1	0	0	0
5	TA	0	1	0	0

Neighborhood

	Category	x
1	Blmngtn	195852.9
2	Blueste	143590.0
3	BrDale	107359.6
4	BrkSide	125959.7
5	ClearCr	217245.3
6	CollgCr	196501.9
7	Crawfor	203912.5
8	Edwards	131036.2
9	Gilbert	189209.6
10	Greens	193531.2
11	GrnHill	280000.0
12	IDOTRR	121872.4
13	Landmrk	137000.0
14	MeadowV	100626.7
15	Mitchel	165918.5
16	NAmes	147499.5
17	NoRidge	308615.5
18	NPkVill	140743.2
19	NridgHt	284173.3
20	NWAmes	194384.1
21	OldTown	123668.3
22	Sawyer	137326.1
23	SawyerW	186755.9
24	Somerst	221740.3
25	StoneBr	262467.3
26	SWISU	129885.5
27	Timber	242024.8
28	Veenker	255865.9

There is a difference in the mean sale price of homes in different neighborhoods. However, it may be harder to handle this feature since it has too many categories. It is imperative to analyze the difference between the mean sale price of homes by the level of category in order to understand if the group membership plays an important role in determining the sale price of homes in Ames.

Q.2

Dataset	No. of observations
Train data	1540
Test data	652
Complete dataset	2192

Q.3

List of predictor variables used in the automated variable selection models:

TotalSqftCalc	Fireplaces	BsmtQual_Ta	KitchenQual_Gd
QualityIndex	Remodel (0 – no remodeling done, 1-remodeling done)	BsmtQual_Gd	KitchenQual_Ta
Age = (YearSold – YearBuilt)	BsmtQual_Po	BsmtQual_Ex	KitchenQual_Fa
Pool	BsmtQual_Fa	KitchenQual_Ex	TotRmsAbvGrd

Total number of predictors = 16

Forward Selection model:

```
Call:
lm(formula = SalePrice ~ TotalSqftCalc + Age + QualityIndex +
    BsmtQual_Ex + TotRmsAbvGrd + KitchenQual_Ta + Fireplaces +
    KitchenQual_Ex + BsmtQual_Gd + KitchenQual_Gd + BsmtQual_Ta,
    data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-88250 -14873  -1487  13161 106644

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)
(Intercept)    13407.852   6132.954   2.186              0.0290 *
TotalSqftCalc     38.618      1.272  30.369 < 0.0000000000000002 ***
Age             -434.899     34.410 -12.639 < 0.0000000000000002 ***
QualityIndex    1569.596     86.264  18.195 < 0.0000000000000002 ***
BsmtQual_Ex    48200.361   4649.492  10.367 < 0.0000000000000002 ***
TotRmsAbvGrd    6145.453    545.973  11.256 < 0.0000000000000002 ***
KitchenQual_Ta -5453.624   4213.388  -1.294              0.1957
Fireplaces      9079.812   1122.162   8.091  0.00000000000000119 ***
KitchenQual_Ex 33683.356   5559.058   6.059  0.00000000171987650 ***
BsmtQual_Gd     7425.487   3530.103   2.103              0.0356 *
KitchenQual_Gd  8351.685   4521.272   1.847              0.0649 .
BsmtQual_Ta    -4882.472   3024.164  -1.614              0.1066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24210 on 1528 degrees of freedom
Multiple R-squared:  0.8613,    Adjusted R-squared:  0.8603
F-statistic: 862.4 on 11 and 1528 DF,  p-value: < 0.00000000000000022
```

Backward selection model:

```
Call:
lm(formula = SalePrice ~ TotRmsAbvGrd + Fireplaces + BsmtQual_Ex +
    BsmtQual_Ta + BsmtQual_Gd + KitchenQual_Ex + KitchenQual_Gd +
    QualityIndex + TotalSqftCalc + Age, data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-88337 -14994  -1466  13304 106661

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)     8439.544   4784.309   1.764               0.0779 .
TotRmsAbvGrd    6121.739    545.786  11.216 < 0.0000000000000002 ***
Fireplaces      9121.418   1121.950   8.130 0.000000000000000878 ***
BsmtQual_Ex    48403.330   4647.873  10.414 < 0.0000000000000002 ***
BsmtQual_Ta    -4941.585   3024.487  -1.634               0.1025
BsmtQual_Gd     7480.154   3530.630   2.119               0.0343 *
KitchenQual_Ex 39177.248   3590.690  10.911 < 0.0000000000000002 ***
KitchenQual_Gd 13800.529   1649.716   8.365 < 0.0000000000000002 ***
QualityIndex    1558.776     85.877  18.151 < 0.0000000000000002 ***
TotalSqftCalc     38.560      1.271  30.336 < 0.0000000000000002 ***
Age             -428.277     34.035 -12.583 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24220 on 1529 degrees of freedom
Multiple R-squared:  0.8611,    Adjusted R-squared:  0.8602
F-statistic: 948.1 on 10 and 1529 DF,  p-value: < 0.00000000000000022
```

Stepwise selection model:

```
Call:
lm(formula = SalePrice ~ TotalSqftCalc + Age + QualityIndex +
    BsmtQual_Ex + TotRmsAbvGrd + Fireplaces + KitchenQual_Ex +
    BsmtQual_Gd + KitchenQual_Gd + BsmtQual_Ta, data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-88337 -14994  -1466  13304 106661

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)     8439.544   4784.309   1.764               0.0779 .
TotalSqftCalc     38.560      1.271  30.336 < 0.0000000000000002 ***
Age             -428.277     34.035 -12.583 < 0.0000000000000002 ***
QualityIndex    1558.776     85.877  18.151 < 0.0000000000000002 ***
BsmtQual_Ex    48403.330   4647.873  10.414 < 0.0000000000000002 ***
TotRmsAbvGrd    6121.739    545.786  11.216 < 0.0000000000000002 ***
Fireplaces      9121.418   1121.950   8.130 0.000000000000000878 ***
KitchenQual_Ex 39177.248   3590.690  10.911 < 0.0000000000000002 ***
BsmtQual_Gd     7480.154   3530.630   2.119               0.0343 *
KitchenQual_Gd 13800.529   1649.716   8.365 < 0.0000000000000002 ***
BsmtQual_Ta    -4941.585   3024.487  -1.634               0.1025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24220 on 1529 degrees of freedom
Multiple R-squared:  0.8611,    Adjusted R-squared:  0.8602
F-statistic: 948.1 on 10 and 1529 DF,  p-value: < 0.00000000000000022
```

Junk model:

```
Call:
lm(formula = SalePrice ~ OverallQual + OverallCond + QualityIndex +
    GrLivArea + TotalSqftCalc, data = train_df)

Residuals:
   Min     1Q Median     3Q    Max
-84264 -15450   -956  14645 117573

Coefficients:
                Estimate  Std. Error t value             Pr(>|t|)
(Intercept)  -167955.806   14824.944 -11.329 < 0.0000000000000002 ***
OverallQual    38589.506    2505.487  15.402 < 0.0000000000000002 ***
OverallCond    15517.710    2689.734   5.769        0.00000000962 ***
QualityIndex   -2673.670     465.959  -5.738        0.00000001152 ***
GrLivArea         28.310       2.341  12.094 < 0.0000000000000002 ***
TotalSqftCalc     37.064       1.531  24.216 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26030 on 1534 degrees of freedom
Multiple R-squared:  0.839,     Adjusted R-squared:  0.8385
F-statistic:  1599 on 5 and 1534 DF,  p-value: < 0.00000000000000022
```

VIF values for the variables:

```
> sort(vif(forward.lm),decreasing=TRUE)
KitchenQual_Gd KitchenQual_Ta    BsmtQual_Gd    BsmtQual_Ta KitchenQual_Ex    BsmtQual_Ex 
     13.037114      11.649956       8.065082       5.943345       3.665980       3.351843 
           Age  TotalSqftCalc    TotRmsAbvGrd   QualityIndex     Fireplaces 
      2.548017       1.915738       1.478827       1.400037       1.316319 
> sort(vif(backward.lm),decreasing=TRUE)
   BsmtQual_Gd    BsmtQual_Ta    BsmtQual_Ex            Age  TotalSqftCalc KitchenQual_Gd 
      8.063928       5.941990       3.348030       2.491695       1.913334       1.734950 
KitchenQual_Ex   TotRmsAbvGrd   QualityIndex     Fireplaces 
      1.528804       1.477162       1.386891       1.315239 
> sort(vif(stepwise.lm),decreasing=TRUE)
   BsmtQual_Gd    BsmtQual_Ta    BsmtQual_Ex            Age  TotalSqftCalc KitchenQual_Gd 
      8.063928       5.941990       3.348030       2.491695       1.913334       1.734950 
KitchenQual_Ex   TotRmsAbvGrd   QualityIndex     Fireplaces 
      1.528804       1.477162       1.386891       1.315239 
> 
```

None of the predictor variables have high values of VIF. Only a few indicator variables exceed the VIF collinearity threshold. This is common when the proportion of cases that belong to the reference category is small. In such cases, the indicator variables will have high VIF even if the categorical variable is not correlated with other predictors. In our model, the higher VIF can be safely ignored, and if need be then the reference category could be set to the one that has a higher proportion of cases.

Comparison of automated variable selection models:

It can be observed that the stepwise selection and the backward selection methods have resulted in the exact same model. The forward selection model differs from these two by an extra variable.

Following are the results of the evaluation on the training dataset:

Model Name	Number of variables	Adjusted R – square	AIC	BIC	MSE	MAE
Forward Selection	11	**0.8603 #1**	35475.6 #2	35545.02 #2	**581630650 #1**	**18125.65 #1**
Backward Selection	10	0.8602 #2	**35475.3 #1**	**35539.37 #1**	582268372 #2	18134.73 #2
Stepwise selection	10	0.8602 #2	**35475.3 #1**	**35539.37 #1**	582268372 #2	18134.73 #2
Junk	5	0.8385 #3	35693.1 #3	35730.46 #3	675090185 #3	19514.46 #3

It can be seen from the above table that the forward selection, backward selection and stepwise selection have produced similar results. AIC and BIC values of forward selection is slightly higher due to the inclusion of an additional dummy variable. However, the MSE for forward selection is lower and the adjusted R-square slightly higher.

Q.4

MSE and MAE on the test data:

Model Name	MSE	MAE

Forward Selection	619853500	18848.15
Backward Selection	623778178	18921.46
Stepwise Selection	623778178	18921.46
Junk	815573022	21079.02

It can be observed from the test data evaluation that the forward selection model performed better on both training and test datasets. The junk model performed the worst on both the datasets. When a model has better predictive accuracy on training than on test data, it means that it is overfitting. In this case, the forward selection model has a lower training data MSE and MAE.

Q.5

Prediction grades for test data:

Forward Selection:

```
forward.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
           0.5352761           0.2039877           0.1809816           0.0797546
```

Backward Selection:

```
backward.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
          0.53680982          0.20245399          0.17638037          0.08435583
```

Stepwise Selection:

```
stepwise.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
         0.53680982          0.20245399          0.17638037          0.08435583
```

Junk:

```
junk.testPredictionGrade
   Grade 1: [0.0.10] Grade 2: (0.10,0.15] Grade 3: (0.15,0.25]     Grade 4: (0.25+]
          0.5276074           0.1641104           0.1886503           0.1196319
```

All the above models are ‘underwriting quality’ because their predictions are within ten percent of the true value more than fifty percent of the time. However, we see that the backward and stepwise models have a minor edge over the forward selection model. It was also observed that these models have higher predictive accuracy on the training sets than the test sets.

Q. 6

After observing the results from the predictive grades on the test dataset, the backward selection model was chosen. It consists of 10 variables:

```
Call:
lm(formula = SalePrice ~ TotRmsAbvGrd + Fireplaces + BsmtQual_Ex +
    BsmtQual_Ta + BsmtQual_Gd + KitchenQual_Ex + KitchenQual_Gd +
    QualityIndex + TotalSqftCalc + Age, data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-88337 -14994  -1466  13304 106661

Coefficients:
               Estimate Std. Error t value            Pr(>|t|)
(Intercept)    8439.544   4784.309   1.764              0.0779 .
TotRmsAbvGrd   6121.739    545.786  11.216 < 0.0000000000000002 ***
Fireplaces     9121.418   1121.950   8.130 0.000000000000000878 ***
BsmtQual_Ex   48403.330   4647.873  10.414 < 0.0000000000000002 ***
BsmtQual_Ta   -4941.585   3024.487  -1.634              0.1025
BsmtQual_Gd    7480.154   3530.630   2.119              0.0343 *
KitchenQual_Ex 39177.248   3590.690  10.911 < 0.0000000000000002 ***
KitchenQual_Gd 13800.529   1649.716   8.365 < 0.0000000000000002 ***
QualityIndex   1558.776     85.877  18.151 < 0.0000000000000002 ***
TotalSqftCalc    38.560      1.271  30.336 < 0.0000000000000002 ***
Age            -428.277     34.035 -12.583 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24220 on 1529 degrees of freedom
Multiple R-squared:  0.8611,	Adjusted R-squared:  0.8602
F-statistic: 948.1 on 10 and 1529 DF,  p-value: < 0.00000000000000022
```

All the model coefficients and their signs seem logically correct. There is no evidence of multicollinearity.

<u>Reexamination of variables to include in the model:</u>

- Removal of *TotRmsAbvGrd* causes a 2% change in the R2 value. Hence it is retained.
- Removal of *Fireplaces* causes less than 1% change in the R2 value. It is removed from the final model.
- Removal of *Age* causes less than 1% change in the R2 value. It is removed from the final model.

All the indicator variables for BsmtQual and KitchenQual are included in the final model.

Set of variables in the model: *TotalSqftCalc, QualityIndex, TotRmsAbvGrd, BstmQual And KitchenQual.*

Two of them are categorical variables. Some interaction terms were considered, and it was found that the interaction between BsmtQual and QualityIndex is significant. It also caused a change in the R2 value. This term was included in the final model.

<u>Final model:</u>

The final model consists of 6 terms.

```
Call:
lm(formula = SalePrice ~ TotRmsAbvGrd + BsmtQual + KitchenQual +
    QualityIndex + TotalSqftCalc + BsmtQual * QualityIndex, data = train.clean)

Residuals:
   Min     1Q Median     3Q    Max
-91449 -15835   -896  14271 108881

Coefficients:
                         Estimate  Std. Error t value
(Intercept)           -126702.599   24904.753  -5.087
TotRmsAbvGrd             6119.994     562.896  10.872
BsmtQualFa             131069.126   24796.770   5.286
BsmtQualGd             134605.869   23476.920   5.734
BsmtQualTA             146548.029   23315.773   6.285
BsmtQualNo             134345.415   26962.565   4.983
KitchenQualFa          -29094.139    5853.373  -4.970
KitchenQualGd           -9042.033    3817.292  -2.369
KitchenQualTA          -26638.629    4021.061  -6.625
QualityIndex             6631.578     577.001  11.493
TotalSqftCalc              42.962       1.236  34.753
BsmtQualFa:QualityIndex  -5315.062     635.633  -8.362
BsmtQualGd:QualityIndex  -4440.347     586.223  -7.574
BsmtQualTA:QualityIndex  -5596.024     582.394  -9.609
BsmtQualNo:QualityIndex  -4952.075     778.903  -6.358
```

```
                               Pr(>|t|)
(Intercept)                 0.0000004078276754 ***
TotRmsAbvGrd              < 0.0000000000000002 ***
BsmtQualFa                  0.0000001433517160 ***
BsmtQualGd                  0.0000000118365163 ***
BsmtQualTA                  0.0000000004259917 ***
BsmtQualNo                  0.0000006986102569 ***
KitchenQualFa               0.0000007431868126 ***
KitchenQualGd                            0.018 *
KitchenQualTA               0.0000000000480791 ***
QualityIndex              < 0.0000000000000002 ***
TotalSqftCalc             < 0.0000000000000002 ***
BsmtQualFa:QualityIndex   < 0.0000000000000002 ***
BsmtQualGd:QualityIndex     0.0000000000000622 ***
BsmtQualTA:QualityIndex   < 0.0000000000000002 ***
BsmtQualNo:QualityIndex     0.0000000002698292 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24980 on 1525 degrees of freedom
Multiple R-squared:  0.8526,    Adjusted R-squared:  0.8512
F-statistic: 629.9 on 14 and 1525 DF,  p-value: < 0.00000000000000022
```

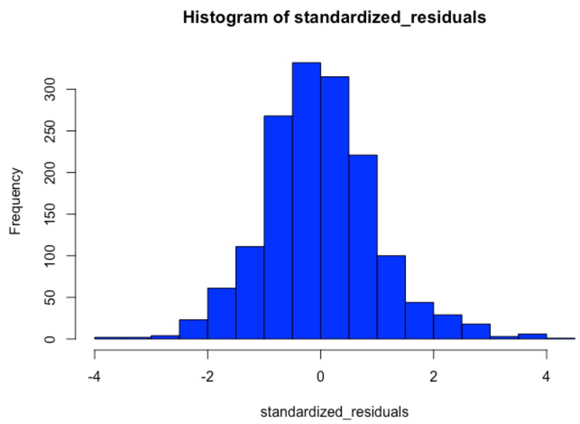
R-square = 0.8526, Adjusted R-square = 0.8512

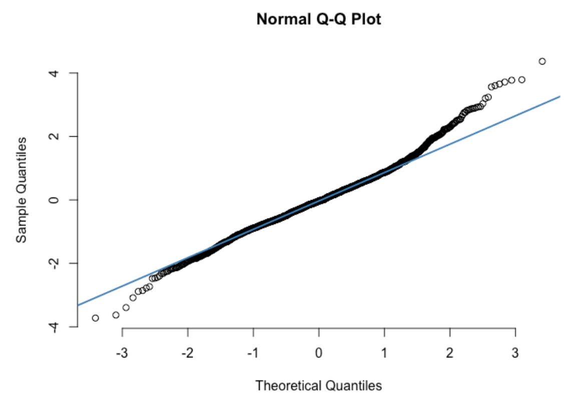
All the variables are statistically significant. F-test is also statistically significant which means all the predictors have a significant relationship with the response variable, *SalePrice.*

ANOVA test:

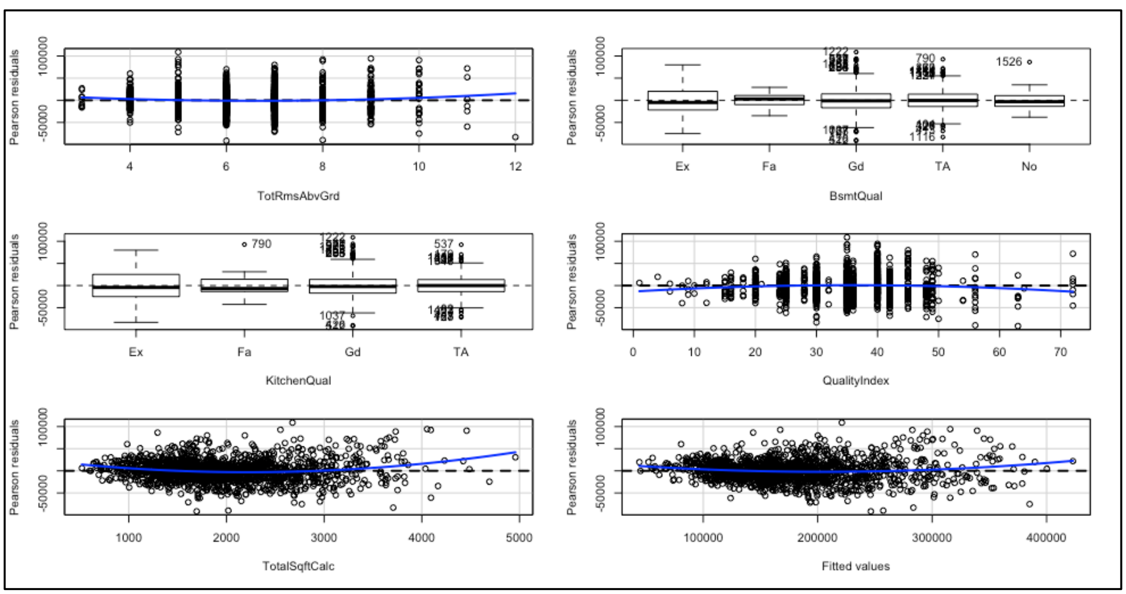
```
> anova(m6)
Analysis of Variance Table

Response: SalePrice
                        Df        Sum Sq       Mean Sq  F value                Pr(>F)
TotRmsAbvGrd             1 2008332400030 2008332400030 3217.330 < 0.00000000000000022 ***
BsmtQual                 4 1976988426630  494247106658  791.779 < 0.00000000000000022 ***
KitchenQual              3  405299029858  135099676619  216.428 < 0.00000000000000022 ***
QualityIndex             1  254881056143  254881056143  408.317 < 0.00000000000000022 ***
TotalSqftCalc            1  786582170844  786582170844 1260.097 < 0.00000000000000022 ***
BsmtQual:QualityIndex    4   72891103015   18222775754   29.193 < 0.00000000000000022 ***
Residuals             1525  951940559843     624223318
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual plots:

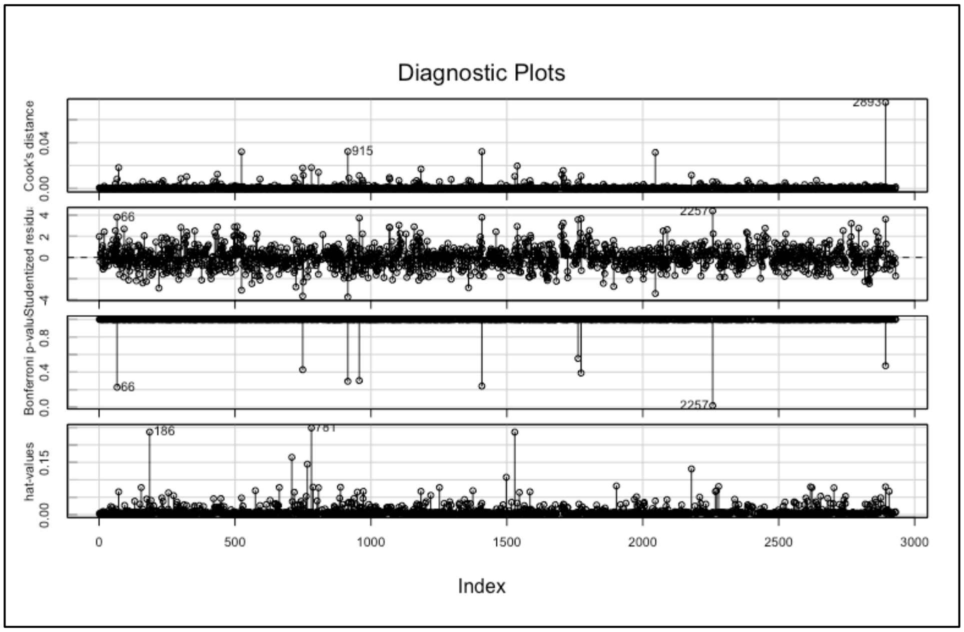




Residuals are approximately normally distributed. There is very mild right skewness.



Residual plots of the predictors are not really funnel shaped. There may be mild heteroscedasticity, which can be ignored.



Diagnostic plots reveal that there is one influential point/outlier, instance number 2893. The cook's distance value for this instance is as follows:

```
> cooks.distance(m6)[names(cooks.distance(m6))=="2893"]
      2893
0.07498959
```

<u>Two things were considered here –</u>

Cook's distance cut off value of 1: The value in this case is way below one and hence, the observation need not be removed from the dataset.

Cook's distance cutoff value of 4/n: The cut off value in this case would be 0.00259. In that case, this observation needs to be removed from the dataset.

The model built on a dataset where the instance was removed resulted in a very minor change in the R2 value.

```
Residual standard error: 24890 on 1524 degrees of freedom
Multiple R-squared:  0.8538,    Adjusted R-squared:  0.8524
F-statistic: 635.5 on 14 and 1524 DF,  p-value: < 0.00000000000000022
```

Diagnostic plots of this new model showed a few other influential points. Hence for simplicity sake, I decided to retain this instance in the final model.

Q.7

Some of the challenges that were presented by this dataset were-

- Collinearity
- Heteroscedasticity
- Too many variables that may seem to be correlated with the sale price of a home in Ames.

To further improve accuracy, other variables in the dataset should be considered. The initial step in this assignment involved the usage of intuition to make a decision about the choice of categorical variables that are to be included in the model. However, other variables such as *Neighborhood, Zoning, SubClass* should be included in the stepwise selection process to make sure we don't miss out on the right predictors for the model. Simpler models are in fact better than complicated. A model with interaction terms, categorical variables with several levels is difficult to interpret. However, these models might result in better R2 and predictive accuracy. The decision about a simpler or a complex model would totally depend upon the end goal of building a model. If the model is mainly built for understanding the relationships between the variables, then a simpler/more interpretable model is better. If it is being used as a part of a larger pipeline in a business context, then a more complex model with better predictive accuracy would be appropriate.